Corruption-Robust Linear Bandits

Mayuree Binjolkar, Romain Camilleri, Andy Jones, Anqi Li, Andrew Wagenmaker

May 31, 2021

1 Introduction

The linear bandit problem is a generalization of the standard multi-armed bandit problem, in which the rewards (losses) exhibit a linear relationship with the arms. In particular, consider some set $\mathcal{X} \subseteq \mathbb{R}^d$ and some vector $\theta_t \in \mathbb{R}^d$. Then every arm corresponds to some vector $x \in \mathcal{X}$, and the reward (loss) for arm x is given by $\theta_t^\top x_x$.

The linear bandit problem has been extensively studied in the purely *stochastic* regime [1, 10, 16, 17], where $\theta_t = \theta$ for all time, and the observations are given by:

$$y_t = x_t^\top \theta + \eta_t$$

for $\eta_t \sim \mathcal{N}(0, \sigma^2)$. In this regime, one can obtain minimax regret scaling as $\mathcal{O}(\sqrt{T})$, as well as instance-dependent regret scaling as $\mathcal{O}(\log T)$. The other primary regime of interest is the purely *adversarial* regime [2, 4], where the observations and losses are given by

$$y_t = x_t^\top \theta_t$$

and now θ_t is chosen by an adversary. Here the goal is to simply obtain regret that scales as $\mathcal{O}(\sqrt{T})$.

A third regime exists in-between these two, the *corrupted stochastic* regime, where $\theta_t = \theta + c_t$ for some corruption c_t. Now the observations are given by

$$y_t = x_t^\top (\theta + c_t) + \eta_t$$

and the goal is to obtain a regret bound which scales in terms of the total corruption level, $C = \sum_{t=1}^T \max_{x \in \mathcal{X}} |\langle x, c_t \rangle|$.

A recent line of work in the multi-armed bandit setting has sought to merge these regimes and obtain algorithms that are simultaneously optimal in every regime, without prior knowledge of which regime they are operating in. While a long line of work exists here, the definitive recent work [19] proposes and analyzes the Tsallis-INF algorithm, which is simultaneously optimal in all regimes.

Despite the extensive body of literature on this problem in the multi-armed bandit setting, until recently little work has been done on this in the linear bandit setting. Indeed, until this year, the only real work on this is that of [13], which addresses the purely stochastic and corrupted stochastic regimes, and provides no guarantees in the adversarial setting. In addition, their rates are not optimal. Recently, however, [12] tackles all three regimes—proposing an algorithm that is nearly optimal simultaneously in all settings.

Related to the above problem is that of misspecified linear bandits. Here, we assume a base stochastic model but now assume that the rewards are not purely given by a linear function, but are only approximately linear. While this can be viewed as an instance of the corrupted setting, the additional structure in the problem allows for a tighter analysis. Formally, at time t, we can observe the measurement x_t through the real value

$$y_t = \mu_{x_t} + \eta_t + \zeta_t$$

with $|\zeta_t| \leq h$.

1.1 Problem Setting

In both the stochastic and corrupted stochastic regime, the regret is defined as,

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle x_t - x, \theta \rangle = \sum_{t=1}^{T} \Delta_x N_x, \tag{1}$$

where $\Delta_x = \langle x - x^*, \theta \rangle$ is the *sub-optimality gap* of arm x. We additionally denote the minimum gap $\Delta_{\min} := \min_{x \neq x^*} \Delta_x$.

In the stochastic setting, at each round, the learner observes loss $y_t = \langle x_t, \theta \rangle + \epsilon_t(x_t)$, where ϵ_t is a zero-mean noise given x_t. In the corrupted setting, the learner observes $y_t = \langle x_t, \theta + c_t \rangle + \epsilon_t(x_t)$, where c_t is a corruption vector. We define the total amount of corruption as $C = \sum_{t=1}^{T} \max_{x \in \mathcal{X}} |\langle x, c_t \rangle|$.

In the adversarial regime, the learner observes the loss $y_t = \langle x_t, \theta_t \rangle$ at each step, where θ_t can be chosen by an adversary. Now, the regret is defined as the best action in hindsight:

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle x_t - x, \theta_t \rangle. \tag{2}$$

Note that the definitions of regret in the adversarial vs corrupted stochastic and stochastic regimes are subtly different. In the stochastic regimes, the regret is only defined with respect to the true parameter θ, and the noise $\epsilon_t(x_t)$ and corruptions c_t do not change the regret. However, in the adversarial regime, the noise and corruptions are all considered part of θ_t, so the regret definition does depend on them.

Note: The stochastic setting can be viewed as a special case of the corrupted case with $C = 0$. Therefore, we will only analyze the corrupted case, and the results for the stochastic setting can be directly obtained by setting $C = 0$.

Assumptions: We assume that $\langle x, \theta \rangle$, $\langle x, c_t \rangle$, and y_t are all in $[-1, 1]$ for all t and $x \in \mathcal{X}$.

1.2 Instance-Dependence in Linear Bandits

Typically, in the best-of-both-worlds or best-of-three-worlds problems, are goal will be to obtain minimax $\mathcal{O}(\sqrt{T})$ regret in the adversarial regime, and instance-dependent $\log T$ regret in the stochastic regime (note that any adversarial minimax algorithm is also minimax in the stochastic regime so simply aiming for algorithms that are minimax in both is trivial). In the multi-armed bandit setting, the optimal instance-dependent rate is given by the familiar sum of inverse gaps:

$$\sum_x \frac{\log T}{\Delta_x}.$$

However, in the linear regime, we do not want the regret to scale with a sum over all arms but rather with the dimensionality, d. Two notions of instance-dependent regret are possible in this setting then. First, we can go after regret of

$$\frac{d \log T}{\Delta_{\min}}.$$

While this does not scale with the size of $\mathcal{X}$, it is still potentially very loose, as it is effectively saying that we are paying a $1/\Delta_{\min}$ in every direction. To remedy this, several works have shown that the correct instance-dependent scaling is given by the solution to the following optimization:

$$\begin{aligned} c(\mathcal{X}, \theta) = \min_{N_x x \in \mathcal{X}, N_x \geq 0} \quad & \sum_{x \in \mathcal{X}} N_x \Delta_x \\ \text{s.t.} \quad & \|x\|^2_{H(N)^{-1}} \leq \frac{\Delta_x^2}{2}, \quad \forall x \in \mathcal{X} \setminus \{x^*\}. \end{aligned}$$

It can be shown that this is the "correct" instance dependent scaling and matching upper and lower bounds of order $\mathcal{O}(c(\mathcal{X}, \theta) \log T)$ exist [10]. Furthermore, it will always be the case that $\mathcal{O}(c(\mathcal{X}, \theta) \leq d/\Delta_{\min})$, so this always improves on the naive instance-dependent result.

Given this, the gold standard will be to obtain algorithms that are minimax in the adversarial regime, and scale with $c(\mathcal{X}, \theta)$ in the stochastic regime. However, as we will see, this can be challenging, and at times we will have to settle for algorithms that scale with $d/\Delta_{\min}$ instead.

1.3 Organization

We first discuss the work of [13] in Section 2, which addresses the stochastic and corrupted stochastic settings. Given this baseline, we move on to the discussion of [12]. In Section 3 we discuss robust mean estimation, which is a basic primitive used in the algorithm of [12] to ensure accurate mean estimation in the corrupted setting. Then, in Section 4 we introduce and analyze the best-of-two-worlds algorithm from [12] which handles the stochastic and corrupted stochastic setting. Building off of this, in Section 5 we present and analyze the best-of-three-worlds algorithm which works in the stochastic, corrupted stochastic, and adversarial regime. Finally, in Section TODO: something we consider the misspecified linear bandits problem.

2 Stochastic Linear Optimization with Adversarial Corruption

[13] provides an algorithm for dealing with stochastic linear optimization with adversarial corruption. The proposed algorithm in this research builds on [8]. The main idea is to divide the time horizon into epochs so as to eliminate the effect that we get from corruption. Dividing into epochs increases exponentially in length and use only previous epoch's estimation for conducting exploitation in the current round. One of the main challenges that the paper explores is that the ordinary least square estimator cannot be adopted due to the correlation between the different time steps of the estimation (that impedes the concentration-inequalities' application).

2.1 Preliminaries

$D \subseteq R^d$ is a d-polytope and at each time step $t \in [T] :=$ 1,2,...,T, an action $x_t \in D$ is chosen by the algorithm. $\theta \in R^d$ is an unknown hidden vector and η_t is a sequence of sub-Gaussian random noise that has mean 0 and variance proxy 1. In a given time step, t, and for an action, x_t, the reward is defined as $r_t(x_t) = \langle x_t, \theta \rangle + \eta_t$. Here the first term is the inner product of x_t and θ. For each time step $t \leq T$, there exists an adaptive adversary that may corrupt the observed reward by choosing a corruption function c_t. First, x_t is chosen, then the corrupted reward $r_t(x_t) + c_t(x_t)$ is observed, and finally the actual reward $r_t(x_t)$ is received. The total corruption generated by the adversary is denoted by $C = \Sigma_{t=1}^{T} max_{x \in D} |c_t(x)|$ and the value of C is evaluated by $R(T) = \Sigma_{t=1}^{T} \langle x^* - x_t, \theta \rangle$. Also, given P as the set of extreme points of D and $P^- = P \setminus \{x^*\}$, the second highest reward is denoted x_2 and based on this, the expected reward is given as $\Delta = \langle x^* - x_{(2)}, \theta \rangle$.

2.2 SBE Algorithm

The Support Basis Exploration (SBE) algorithm is used for stochastic linear optimization with adversarial corruption and it runs in epoch m (length greater than 4^m). The total number of epochs M is bounded above by $\log T$. In this, the choice of the current action depends only on the information that is received from the last epoch and therefore, the level of earlier corruption will have a decreasing effect on the later epochs. Exploration and exploitation are separated so that the correlation between the vector pulls in each epoch can be decreased.

Algorithm 1: SBE: Support Basis Exploration Algorithm

Parameters: Confidence $\delta \in (0,1)$, time horizon T, decision set D.
Initialization: Exploration set $S = \{s_j\}_{j\in[d]}$.
Set $\zeta = 2^{14}d^6 \log(4d\log T/\delta)$. Set estimated gap $\hat{\Delta}^{(0)} = 1$ and exploration ratio $\gamma_0 = 1/5$.
for *epoch* $m = 1, 2, \cdots, M$ **do**
  Set $n_m = \zeta \cdot 4^m$. Let $N_m = n_m + \zeta(\hat{\Delta}^{(m-1)})^{-2}$, and $T_m = T_{m-1} + N_m$.
  for t *from* $T_{m-1}+1$ *to* T_m **do**
    if $Z = 1$ *for Bernoulli random variable* $Z \sim$ Bernoulli(γ_{m-1}) **then**
      Sample uniformly an action from the exploration set S.
    else
      Choose the best action $x_*^{(m-1)}$ according to the estimate $\hat{\theta}^{(m-1)}$.
    end
  end
  Let $\hat{\theta}^{(m)}$ be the estimate of θ in this epoch, defined later in Section 4.
  Set $\hat{\Delta}^{(m)}$ as the maximum of 2^{-m} and the difference between the expected reward for the best and second best actions given $\hat{\theta}^{(m)}$.
  Set $\gamma_m = (\hat{\Delta}^{(m)})^{-2}/((\hat{\Delta}^{(m)})^{-2} + 2^{2(m+1)})$.
end

2.2.1 Parameter Estimation

The hidden vector, θ, can be expressed according to the exploration set S, $\theta = \Sigma_{j=1}^{d} b_j s_j$. If the basis vector s_j is chosen in time step t, ξ_j^t is defined as an indicator for any $j \in [d]$. $n_e^{(m)} = \mathbb{E}\left[\Sigma_{t=T_{m-1}+1}^{T_m} \xi_j^t\right]$ is the expected number of time steps that can be used to explore each of the basis vector s_j. However, since s_j is sampled uniformly, $n_e^{(m)}$ is independent of j. Based on this, the "average reward" for exploring s_j in epoch m is

$$r_j^{(m)} = \frac{1}{n_e^{(m)}} \Sigma_{t=T_{m-1}+1}^{T_m} \xi_j^t . (\langle s_j, \theta, \rangle + \eta_t + c_t(s_j)) \tag{3}$$

ξ_j^t is independent of the noise η_t and the amount of corruption $c_t(s_j)$. Taking the expectation over the randomness of independent variables ξ_j^t and η_t on both the sides gives

$$\mathbb{E}\left[r_j^{(m)}\right] = \langle s_j, \theta \rangle + \frac{1}{N_m} \Sigma_{t=T_{m-1}+1}^{T_m} \mathbb{E}[c_t(s_j)] \leq b_j \|s_j\|_2^2 + \frac{C_m}{N_m}, \tag{4}$$

where $C_m = \Sigma_{t=T_{m-1}+1}^{T_m} \max_{x\in D} |c_t(x)|$. Based on this, at the end of each epoch, $\hat{b}_j^{(m)} = \frac{r_j^{(m)}}{\|s_j\|_2^2}$ is the estimate of b_j and $\hat{\theta}^{(m)} = \Sigma_{j=1}^{d} \hat{b}_j^{(m)} s_j$ is the estimate of θ.

Next, an upper bound for the error of $\hat{\theta}^{(m)}$ in each dimension j has been provided.

Lemma 1. *(Lemma 4.1, [13]) With the probability at least* $1-\delta$*, the estimate of* $\hat{b}_j^m$ *is such that*

$$|\hat{b}_j^m - b_j| \|s_j\|_2^2 \leq \frac{2C_m}{N_m} + \frac{\hat{\Delta}^{(m-1)}}{32d^2}$$

Proof. Indicator ξ_j^t and the noise η_t are independent random variables, then using a form of the Chernoff-Hoeffding bound [9], for any deviation κ and any $j \in [d]$

$$\Pr\left[\left|\frac{1}{n_e^{(m)}} \Sigma_{t=T_{m-1}+1}^{T_m} \xi_j^t . (\langle s_j, \theta \rangle + \eta_t) - \langle s_j, \theta \rangle\right| \geq \frac{\kappa}{2}\right] \leq 2\exp -\frac{\kappa^2 n_e^{(m)}}{16} \tag{5}$$

Let $X_t = (\xi_j^t) - \frac{n_e^{(m)}}{N_m} c_t(s_j)$ for all t. The filtration $\left\{F_t\right\}_{t=1}^{T}$ is generated by the random variables $\left\{\xi_j^s\right\}_{j\in[d], s\leq t}$ and $\left\{\eta_s\right\}_{s\leq t+1}$. Define $Y_t = \Sigma_{s=1}^{t} X_s$. Because ξ_j^t is independent of the corruption level c_j^t

conditional on F_{t-1},$\{Y_t\}_{t=1}^{T}$ yields a martingale with respect to the filtration. The variance of X_t conditional on F_{t-1} can be bounded as

$$V = \mathbb{E}[X_t^2|F_{t-1}] \leq \Sigma_{t=T_{m-1}+1}^{T_m}|c_t(s_j)|Var([\xi_j^t]) \leq \frac{n_e^{(m)}}{N_m}\Sigma_{T_{m-1}+1}^{T_m}|c_t(s_j)|. \tag{6}$$

Both the first and second inequality hold because $|c_t(s_j)| \leq 1$ and $Var[\xi_j^t] \leq \frac{n_e^{(m)}}{N_m}$ respectively. For martingales, we will use a Freedman-type concentration inequality [3], and for any $\nu > 0$

$$\Pr\big[\frac{1}{n_e^m}\Sigma_{t=T_{m-1}+1}^{T_m}X_t \geq \frac{V + ln4/v}{n_e^{(m)}}\big] \leq \frac{\nu}{4} \tag{7}$$

Combining Equation 3 with Equation 5,

$$\Pr\big[\frac{\Sigma_{t=T_{m-1}+1}^{T_m}\xi_j^t(s_j)}{n_e^{(m)}} \geq \frac{2C_m}{N_m} + \frac{ln4\nu}{n_e^{(m)}}\big] \leq \Pr\big[\frac{1}{n_e^m}\Sigma_{t=T_{m-1}+1}^{T_m}X_t \geq \frac{V + ln4/v}{n_e^{(m)}}\big] \leq \frac{\nu}{4} \tag{8}$$

Substituting $\nu = 4exp\big\{-\frac{\kappa n_e^{(m)}}{2}\big\}$ and combining Inequalities 3 and 6, gives

$$\Pr\Big[\Big|r_j^{(m)} - \langle s_j, \theta\rangle\Big| \geq \kappa + \frac{2C_m}{N_m}\Big] \leq 4exp\Big\{-\frac{\kappa^2 n_e^{(m)}}{16}\Big\} \tag{9}$$

For the full explanation of the proof, refer to [13]. □

Lemma 2. *(Lemma 4.2, [13]) With the probability at least $1-\delta$, we have*

$$\Big|\langle x, \hat{\theta}^{(m)} - \theta\rangle\Big| \leq \frac{4d^2C_m}{N_m} + \frac{\hat{\Delta^{(m)}}}{16}$$

Proof. Based on the exploration set S which is an orthogonal set, for any context x, there are multipliers $\{\hat{a}_j\}_{j\in[d]}$ such that

$$\Big|\langle x, \hat{\theta}^{(m)} - \theta\rangle\Big| = \Sigma_{j=1}^{d}\Big|\hat{d}_j\langle s_j, \hat{\theta}^m - \theta\rangle\Big| \leq \Sigma_{j=1}^{d}|\hat{a}_j|\Big|b_j^{(m)} - b_j\Big|\|s_j\|_2^2.$$

Using Corollary 2.2 from [13] and Lemma 1, and with probability $1-\delta$,

$$\Big|\langle x, \hat{\theta}^{(m)}\rangle\Big| \leq \frac{4d^2C_m}{N_m} + \frac{\hat{\Delta^{(m-1)}}}{16}$$

□

Next, the upper and lower bounds for the estimated gap $\hat{\Delta}^{(m)}$ are provided.

Lemma 3. *(Upper Bound for $\hat{\Delta}^{(m)}$ (Lemma 4.3, [13])). Suppose that event ε happens, then for all the epochs $m \geq 1$*

$$\hat{\Delta}^{(m)} \leq \Big[\Delta + 2^{-m} + 4d^2\Sigma_{s=1}^{m}\Big(\frac{1}{8}\Big)^{m-s}\frac{C_s}{N_s}\Big]$$

Lemma 4. *(Lower Bound for $\hat{\Delta}^{(m)}$ (Lemma 4.4, [13])). Suppose that event ε happens, then for all the epochs m*

$$\hat{\Delta}^{(m)} \geq \frac{\Delta}{2} - 2^{-m-1} - 8d^2\Sigma_{s=1}^{m}\Big(\frac{1}{8}\Big)^{m-s}\frac{C_s}{N_s}$$

2.2.2 Regret Estimation

Theorem 1. *(Theorem 5.1, [13]) With probability at least $1-\delta$, the regret is bounded by*

$$R = O\Big(\frac{d^2 C \log T}{\Delta} + \frac{d^5 \log \frac{d \log T}{\delta} \log T}{\Delta^2}\Big)$$

Proof. Let $R_1^{(m)}$ and $R_2^{(m)}$ be pseudo regret for exploitation and exploration in epoch m respectively. By Lemma 2, the probability of event ε is $1-\delta$.

For exploitation, let $\Delta^{(m)} = \langle \theta, x^* - x_*^{(m-1)} \rangle$ be the pseudo regret for the action $x_*^{(m-1)}$. For a given event ε, $\Delta^{(m)} = \langle \theta - \hat{\theta}^{(m-1)}, x^* \rangle + \langle \hat{\theta}^{(m-1)}, x^* - x_*^{(m-1}\rangle + \langle \hat{\theta}^{(m-1)} - \theta, x_*^{(m-1)} \rangle \leq 2\beta_{(m-1)}$. Defining $\rho_m = d^2 \Sigma_{s=1}^m \big(\frac{1}{8}^{(m-s)}\big) \frac{C_s}{N_s}$, $\Delta^{(m} = \frac{\Delta}{4} + 2^{-m} + 8\rho_{m-1}$. If $\Delta^{(m)} = 0$ then the total regret for exploitation is 0. Else, $\Delta^{(m)} \geq \Delta$ and two different cases can be considered for this. For the first case when $\Delta \geq 2^{-m+1}$, $\frac{\Delta}{4} + 2^{-m} \leq \frac{\Delta}{4} + \frac{\Delta}{2} \leq \frac{3\Delta^{(m)}}{4}$. When $\Delta \leq 2^{-m+1}$, $\Delta^{(m)} \leq 8\rho_{m-1} + 2^{-m+1}$. Summing over all the epochs,

$$R_1 = \frac{4\zeta M}{\Delta} + 32 \Sigma_{s=1}^M C_s \Sigma_{m=s}^M \frac{4^{m-s}}{8^{m-1-s}} \leq \frac{4\zeta M}{\Delta} + 512C, \tag{10}$$

For exploration, the expected number of time steps in which exploration is conducted is $\frac{\zeta}{(\Delta^{(m)})^2}$ and the pseudo regret for these time steps each is bounded by 1.For $\Delta \leq 2^{(1-m)}$, $R_2^{(m)} \leq \frac{\zeta}{(\Delta^{(m)})^2} \leq \frac{4\zeta}{\Delta^2}$ and for $\Delta \geq 2^{(1-m)}$, two cases are considered. The first case is when $\rho_m \geq \frac{\Delta}{64}$ and the second case is when $\rho_m \leq \frac{\Delta}{64}$. For the first case, $\frac{\Delta}{64} \leq \rho_m = d^2 \Sigma_{s=1}^m \big(\frac{1}{8}\big)^{m-s} \frac{C_s}{N_s} \leq \frac{2d^2 \Sigma_{s=1}^m C_s}{N_m} \leq \frac{2d^2}{N_m}$. Based on this $R_2^{(m)} \leq N_m \leq \frac{128 d^2 C}{\Delta}$. For the second case, using Lemma 4, $R_2^{(m)} \leq \frac{\zeta}{(\hat{\Delta}^{(m)})^2} \leq \frac{64\zeta}{\Delta^2}$. Based on these cases, the total pseudo regret for exploration is

$$R_2 = \Sigma_{m=1}^M R_2^{(m)} \leq \Sigma_{m=1}^M \frac{\zeta}{(\hat{\Delta}^{(m)})^2} \leq \frac{64\zeta \log T}{\Delta^2} + \frac{128 d^2 C \log T}{\Delta} \tag{11}$$

Combining inequalities 4 and 5, the total psuedo regret is

$$R = R_1 + R_2 = O\Big(\frac{d^2 C \log T}{\Delta} + \frac{d^5 \log \frac{d \log T}{\delta} \log T}{\Delta^2}\Big)$$

□

2.2.3 Computational efficiency

Theorem 2. *(Theorem 6.1, [13]) For any bounded convex body $K \subseteq R^d$, there is a polynomial time algorithm that computes an ellipsoid E satisfies*

$$E \subseteq K \subseteq 2d^{3/2} E$$

Plug in the polynomial time algorithm for finding John's ellipsoid from [14] into the SBE algorithm and set the parameter $\zeta = 2^{14} d^6 \log \frac{4d \log T}{\delta}$ to get a computationally efficient algorithm with a regret $O\Big(\frac{d^{5/2} C \log T}{\Delta} + \frac{d^6 \log \frac{d \log T}{\delta} \log T}{\Delta^2}\Big)$.

3 Robust Estimators

3.1 Introduction

Often in statistics, we seek to estimate the expected value μ of a random variable X. The natural first choice for such an estimator is the empirical mean, however we will show that this choice of estimator performs sub-optimally when the distribution of X is heavy-tailed. While choosing an estimator $\widehat{\mu}$ equal to the empirical mean results in optimal mean-squared error $\mathbb{E}\left[(\widehat{\mu} - \mu)^2\right]$, the preferred optimality condition in many contexts is for $\widehat{\mu}$ to be close to μ with high probability. That is, the ideal optimality measure of

an estimator $\widehat{\mu}$ is related to the smallest ϵ such that $\mathbb{P}\{|\widehat{\mu} - \mu| > \epsilon\} \leq \delta$. In the following sections we will demonstrate the sub-optimality of the empirical mean, and introduce a few robust estimators that provide better guarantees by reducing the effect of variance in heavy-tailed distributions (at the cost of adding a bit of bias to the estimators).

3.2 Empirical Mean's (Lack of) Optimality

By the central limit theorem, the empirical mean $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ of i.i.d. random variables $X_1, \ldots, X_n$ can be shown to satisfy

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mu} - \mu| > \frac{\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right\} \leq \delta. \tag{12}$$

However, this is an asymptotic guarantee, and we seek to find non-asymptotic guarantees. If the distribution is such that there exists $L > 0$ such that for all $\lambda > 0$

$$\mathbb{E}\left[\exp\left(\lambda\left(X_i - \mu\right)\right)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{L^2}\right), \tag{13}$$

then by the Chernoff bound we can show that

$$\mathbb{P}\left\{|\overline{\mu} - \mu| > \frac{L\sigma\sqrt{\log(2/\delta)}}{\sqrt{n}}\right\} \leq \delta. \tag{14}$$

The issue here is that assumption (13) is very restrictive - particularly, it requires strong tail-decay (it is equivalent to $X_i - \mu$ being L-subgaussian). Without strong tail-decay, the empirical mean estimator offers significantly worse high-probability guarantees. For example, if the only assumption about the distribution is that it has finite variance, then the best guarantee is provided by Chebyshev's inequality, which implies

$$\mathbb{P}\left\{|\overline{\mu} - \mu| > \sigma\sqrt{\frac{1}{n\delta}}\right\} \leq \delta, \tag{15}$$

which is an exponentially worse bound than (14) as a function of δ. This bound is known to be very tight in the worst case, as Catoni showed that for any σ^2 there exists a distribution such that

$$\mathbb{P}\left\{|\overline{\mu} - \mu| > \sigma\sqrt{\frac{1}{n\delta}}\left(1 - \frac{e\delta}{n}\right)^{(n-1)/2}\right\} \geq \delta. \tag{16}$$

For more details on this, consult [15] (Lugosi & Mendelson, 2019), as this section is taken primarily from that source.

3.3 Median of Means

The median-of-means estimator partitions $X_1, \ldots, X_n$ into k blocks of (roughly) equal size, computes the empirical mean of each block, and takes the median of the obtained empirical means.

Lemma 1 [15] (Lugosi & Mendelson, 2019): Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $m, k \in \mathbb{Z}^+$ such that $n = mk$. Then, the median-of-means estimator $\widehat{\mu}_n$ with k blocks satisfies

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{4}{m}}\right\} \leq \exp\left(\frac{-k}{8}\right).$$

In particular, for any $\delta \in (0,1)$, if $k = \lceil 8 \log(1/\delta) \rceil$, then

$$\mathbb{P}\left\{ |\widehat{\mu}_n - \mu| > \sigma \sqrt{\frac{32 \log(1/\delta)}{n}} \right\} \leq \delta.$$

For more details on this, consult [15] (Lugosi & Mendelson, 2019), as this section is taken primarily from that source.

3.4 Trimmed Mean

The trimmed-mean estimator directly removes outliers - it removes the εn highest and lowest points for some parameter $\varepsilon \in (0,1)$, then takes the mean of the remaining points. A simple variant of the trimmed-mean estimator works as follows:

First, the data is split in two equal parts. One half is used to determine the correct truncation level. The other half is average, except for the points that fall outside of the truncation region. Assuming the data consists of $2n$ points $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ drawn i.i.d. with mean μ and variance σ^2, we define the truncation function

$$\phi_{\alpha,\beta}(x) = \begin{cases} \beta, & \text{if } x > \beta, \\ x, & \text{if } x \in [\alpha, \beta], \\ \alpha, & \text{if } x < \alpha, \end{cases}$$

for $\alpha \leq \beta$. For $x_1, \ldots, x_m \in \mathbb{R}$, let $x_1^* \leq x_2^* \leq \cdots \leq x_m^*$ be its non-decreasing rearrangement. Given $\delta \geq 8 \exp\left(\frac{-3n}{16}\right)$, set $\varepsilon = \frac{16 \log(8/\delta)}{3n}$. Let $\alpha = Y^*_{\varepsilon n}$ and $\beta = Y^*_{(1-\varepsilon)n}$.

$$\widehat{\mu}_{2n} = \frac{1}{n} \sum_{i=1}^{n} \phi_{\alpha,\beta}(X_i).$$

Lemma 2 [15] (Lugosi & Mendelson, 2019):

$$\mathbb{P}\left\{ |\widehat{\mu}_{2n} - \mu| > 9\sigma \sqrt{\frac{\log(8/\delta)}{n}} \right\} \leq \delta.$$

For more details on this, consult [15] (Lugosi & Mendelson, 2019), as this section is taken primarily from that source.

3.5 Catoni's Estimator(s)

We begin by observing that the empirical mean $\overline{\mu}$ is the unique root of the function

$$f(z) = \sum_{i=1}^{n} (X_i - z). \tag{17}$$

Catoni proposed the introduction of an antisymmetric non-decreasing function $\psi : \mathbb{R} \to \mathbb{R}$ satisfying

$$-\log\left(1 - y + y^2/2\right) \leq \psi(y) \leq \log\left(1 + y + y^2/2\right),$$

and parameter $\alpha \in \mathbb{R}$, such that the Catoni estimator parameterized by α, which we shall denote $\widehat{\mu}_\alpha$, is the unique root of the function

$$f(z) = \sum_{i=1}^{n} \psi\left(\alpha\left(X_i - z\right)\right). \tag{18}$$

The reasoning for the introduction of ψ is that if $\psi(y)$ increases much slower than y, then the effects of a heavy tail are diminished. Catoni discusses a variety of choices for ψ. The widest possible ψ is

$$\psi(y) = \begin{cases} \log\left(1 + y + y^2/2\right), & \text{if } y \geq 0, \\ -\log\left(1 - y + y^2/2\right), & \text{if } y < 0, \end{cases} \tag{19}$$

and the narrowest possible ψ is

$$\psi(y) = \begin{cases} \log(2), & \text{if } y \geq 1, \\ -\log\left(1 - y + y^2/2\right), & \text{if } 0 \leq y < 1, \\ -\log\left(1 - y + y^2/2\right), & \text{if } -1 \leq y < 0, \\ -\log(2), & \text{if } y < -1. \end{cases} \tag{20}$$

The choice (19) of ψ is the one making $\widehat{\mu}_\alpha$ the closest to $\overline{\mu}$. Since $\overline{\mu}$ is optimal for Gaussian distributions, this is often the chosen ψ.

For more details on this, consult [6] (Catoni, 2011), as this section is taken primarily from that source.

3.5.1 Concentration Inequality for Catoni's Estimator

Lemma 3 [18] (Wei et al., 2020): Let $\mathcal{F}_0 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}\left[X_i | \mathcal{F}_{i-1}\right] = \mu_i$ for some fixed μ_i, and $\sum_{i=1}^n \mathbb{E}\left[\left(X_i - \mu_i\right)^2 | \mathcal{F}_{i-1}\right] \leq V$ for some fixed V. Denote $\mu = \frac{1}{n}\sum_{i=1}^n \mu_i$ and let $\widehat{\mu}_\alpha$ be the Catoni estimator of $X_1, \ldots, X_n$ with a fixed parameter $\alpha > 0$. That is, $\widehat{\mu}_\alpha$ is the unique root of the function

$$f(z) = \sum_{i=1}^n \psi\left(\alpha\left(X_i - z\right)\right),$$

where

$$\psi(y) = \begin{cases} \log\left(1 + y + y^2/2\right), & \text{if } y \geq 0, \\ -\log\left(1 - y + y^2/2\right), & \text{if } y < 0. \end{cases}$$

Then, for any $\delta \in (0,1)$, as long as $n \geq \alpha^2\left(V + \sum_{i=1}^n \left(\mu_i - \mu\right)^2\right) + 2\log(1/\delta)$, we have with probability at least $1 - 2\delta$,

$$\mathbb{P}\left\{ |\widehat{\mu}_\alpha - \mu| > \frac{\alpha\left(V + \sum_{i=1}^n \left(\mu_i - \mu\right)^2\right)}{n} + \frac{2\log(2/\delta)}{\alpha n} \right\} \leq \delta. \tag{21}$$

In particular, if $\mu_1 = \cdots = \mu_n = \mu$, we have

$$\mathbb{P}\left\{ |\widehat{\mu}_\alpha - \mu| > \frac{\alpha V}{n} + \frac{2\log(2/\delta)}{\alpha n} \right\} \leq \delta. \tag{22}$$

Proof: Observe that $\psi(y) \leq \log\left(1 + y + y^2/2\right)$ for all $y \in \mathbb{R}$. So, for any fixed $z \in \mathbb{R}$ and any i, we have

$$\begin{aligned} \mathbb{E}_i\left[\exp(\psi(\alpha(X_i - z)))\right] &\leq \mathbb{E}_i\left[1 + \alpha(X_i - z) + \frac{\alpha^2(X_i - z)^2}{2}\right] && (\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{i-1}]) \\ &= 1 + \alpha(\mu_i - z) + \frac{\alpha^2\mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \alpha^2(\mu_i - z)^2}{2} \\ &\leq \exp\left(\alpha(\mu_i - z) + \frac{\alpha^2\mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \alpha^2(\mu_i - z)^2}{2}\right). && (1 + y \leq \exp(y)) \end{aligned}$$

Define random variables $Z_0 = 1$, and for $i \geq 1$,

$$Z_i = Z_{i-1} \exp\left(\psi(\alpha(X_i - z))\right) \exp\left(-\left(\alpha(\mu_i - z) + \frac{\alpha^2 \mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \alpha^2(\mu_i - z)^2}{2}\right)\right).$$

The previous calculation shows $\mathbb{E}[Z_i] \leq Z_{i-1}$. So, taking expectation over all random variables $X_1, \ldots, X_n$, we have

$$\mathbb{E}[Z_n] \leq \mathbb{E}[Z_{n-1}] \leq \cdots \leq \mathbb{E}[Z_0] = 1.$$

Define

$$g(z) = n\alpha(\mu - z) + \frac{1}{2}\alpha^2 \sum_{i=1}^{n}(\mu_i - z)^2 + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right).$$

If $f(z) \geq g(z)$, then, by the condition on V, we have

$$\begin{aligned}
\sum_{i=1}^{n} \psi(\alpha(X_i - z)) &\geq n\alpha(\mu - z) + \frac{1}{2}\alpha^2 \sum_{i=1}^{n}(\mu_i - z)^2 + \frac{1}{2}\alpha^2 \sum_{i=1}^{n} \mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \log\left(\frac{1}{\delta}\right) \\
&= \sum_{i=1}^{n}\left(\alpha(\mu_i - z) + \frac{\alpha^2(\mu_i - z)^2) + \alpha^2 \sum_{i=1}^{n} \mathbb{E}_i\left[(X_i - \mu_i)^2\right]}{2}\right) + \log\left(\frac{1}{\delta}\right),
\end{aligned}$$

which implies $Z_n \geq \frac{1}{\delta}$. So,

$$\begin{aligned}
\mathbb{P}\{f(z) \geq g(z)\} &\leq \mathbb{P}\left\{Z_n \geq \frac{1}{\delta}\right\} \\
&\leq \mathbb{P}\left\{Z_n \geq \frac{\mathbb{E}[Z_n]}{\delta}\right\} \\
&\leq \delta. && \text{(Markov's inequality)}
\end{aligned}$$

We can rewrite $g(z)$ as

$$\begin{aligned}
g(z) &= n\alpha(\mu - z) + \frac{1}{2}\alpha^2\left(nz^2 - 2n\mu z + \sum_{i=1}^{n}\mu_i^2\right) + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right) \\
&= n\alpha(\mu - z) + \frac{1}{2}\alpha^2\left(n(z - \mu)^2 - n\mu^2 + \sum_{i=1}^{n}\mu_i^2\right) + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right) \\
&= n\alpha(\mu - z) + \frac{1}{2}n\alpha^2(z - \mu)^2 + \frac{1}{2}\alpha^2\left(\sum_{i=1}^{n}\mu_i^2 - n\mu^2\right) + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

Pick z to be the smaller root z_0 of the quadratic function $g(z)$, that is,

$$z_0 = \mu + \frac{1}{\alpha}\left(1 - \sqrt{1 - \frac{\alpha^2\left(V + \sum_{i=1}(\mu_i - \mu)^2\right)}{n} - \frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right),$$

which exists due to the condition on n. Since f is non-increasing (since ψ is non-decreasing and $\alpha > 0$) and $f\left(\widehat{\mu}_\alpha\right) = 0$, we have

$$\begin{aligned}
\mathbb{P}\left\{\widehat{\mu}_\alpha \geq z_0\right\} &= \mathbb{P}\left\{f(z_0) \geq 0\right\} \\
&= \mathbb{P}\left\{f(z_0) \geq g(z_0)\right\} \\
&\leq \delta.
\end{aligned}$$

Thus, with probability at least $1-\delta$, we have

$$\begin{aligned}
\widehat{\mu}_\alpha - \mu &\le z_0 - \mu \\
&= \frac{1}{\alpha}\left(1 - \sqrt{1 - \frac{\alpha^2\left(V + \sum_{i=1}(\mu_i-\mu)^2\right)}{n} - \frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \\
&\le \frac{1}{\alpha}\left(\frac{\alpha^2\left(V + \sum_{i=1}(\mu_i-\mu)^2\right)}{n} + \frac{2}{n}\log\left(\frac{1}{\delta}\right)\right) \qquad \left(1-\sqrt{1-x} \le x \text{ for } x \in [0,1]\right) \\
&= \frac{\alpha\left(V + \sum_{i=1}(\mu_i-\mu)^2\right)}{n} + \frac{2\log(1/\delta)}{\alpha n}.
\end{aligned}$$

Via a symmetric argument, with probability at least $1-\delta$, we have

$$\mu - \widehat{\mu}_\alpha \le \frac{\alpha\left(V + \sum_{i=1}(\mu_i-\mu)^2\right)}{n} + \frac{2\log(1/\delta)}{\alpha n}.$$

Applying a union bound, we have

$$\mathbb{P}\left\{|\widehat{\mu}_\alpha - \mu| > \frac{\alpha\left(V + \sum_{i=1}^n (\mu_i-\mu)^2\right)}{n} + \frac{2\log(2/\delta)}{\alpha n}\right\} \le \delta,$$

as desired. Q.E.D.

3.6 In Lee et al.

Lee et al. [12] propose two algorithms: Randomized Instance-optimal Algorithm (RIA) and Best of Three Worlds (BoTW).

During each block, RIA constructs unbiased loss estimators $\widehat{\ell}_{t,x}$ to estimate $\ell_{t,x} = \langle x, \ell_t\rangle$ for each $t \in \mathcal{B}_m$. A robust estimator $\mathrm{Rob}_{m,x}$ is then constructed for each x, where

$$\mathrm{Rob}_{m,x} = \mathrm{Clip}_{[-1,1]}\left(\widehat{\mu}_{\alpha_x}\left(\left\{\widehat{\ell}_{\tau,x}\right\}_{\tau\in\mathcal{B}_m}\right)\right), \tag{23}$$

with

$$\alpha_x = \sqrt{\frac{4\log\left(2^m|\mathcal{X}|/\delta\right)}{2^m\left\|x\right\|_{S_m^{-1}}^2 + 2^m}}. \tag{24}$$

The clipping to $[-1,1]$ is performed because, by the initial problem assumptions, $\ell_{t,x} \in [-1,1]$.

The second phase of BoTW constructs loss estimators $\widehat{\ell}_{t,x}$ to estimate $\ell_{t,x} = \langle x, \ell_t\rangle$ for each $t \in \{t_0{+}1, \ldots, t_1\}$. A robust estimator $\mathrm{Rob}_{t,x}$ is constructed for each x at each time $t \in \{t_0+1, \ldots, t_1\}$, where

$$\mathrm{Rob}_{t,x} = \mathrm{Clip}_{[-1,1]}\left(\widehat{\mu}_{\alpha_x}\left(\left\{\widehat{\ell}_{\tau,x}\right\}_{\tau=t_0+1}^{t}\right)\right), \tag{25}$$

with

$$\alpha_x = \sqrt{\frac{4\log\left(t|\mathcal{X}|/\delta\right)}{t - t_0 + \sum_{\tau=t_0+1}^{t} 2\left\|x\right\|_{S_m^{-1}}^2}}. \tag{26}$$

Algorithm 1 Randomized Instance-optimal Algorithm

Input: $\delta < 0.1$

$t \leftarrow 1$.

for $m = 0, 1, 2 \ldots$ **do**

  Define block $\mathcal{B}_m = \{t,\ t+1, \ldots,\ t+2^m-1\}$.

  Find a randomized strategy $p_m = \mathbf{OP}(2^m, \widehat{\Delta}_m)$ with

$$\widehat{\Delta}_{m,x} = \begin{cases} 0 & \text{if } m = 0, \\ \mathrm{Rob}_{m-1,x} - \min_{x' \in \mathcal{X}} \mathrm{Rob}_{m-1,x'} & \text{else.} \end{cases}$$

  Compute second moment $S_m = \sum_{x \in \mathcal{X}} p_{m,x} x x^\top$.

  while $t \in \mathcal{B}_m$ **do**

    Sample $x_t \sim p_m$ and observe y_t.

    Compute for all $x \in \mathcal{X}$, $\widehat{\ell}_{t,x} = x^\top S_m^{-1} x_t y_t$.

    $t \leftarrow t+1$.

  for $x \in \mathcal{X}$ **do**

    Construct robust loss estimators

$$\mathrm{Rob}_{m,x} = \mathrm{Clip}_{[-1,1]}\left(\mathbf{Catoni}_{\alpha_x}\left(\left\{\widehat{\ell}_{\tau,x}\right\}_{\tau \in \mathcal{B}_m}\right)\right)$$

    with $\alpha_x = \sqrt{\frac{4\log(2^m|\mathcal{X}|/\delta)}{2^m \cdot \|x\|^2_{S_m^{-1}} + 2^m}}$.

$\mathbf{OP}(t, \widehat{\Delta})$: return any minimizer p^* of the following:

$$\min_{p \in \mathcal{P}_\mathcal{X}} \sum_x p_x \widehat{\Delta}_x, \tag{3}$$

$$\text{s.t. } \|x\|^2_{S(p)^{-1}} \le \frac{t\widehat{\Delta}_x^2}{\beta_t} + 4d, \ \ \forall x \in \mathcal{X}, \tag{4}$$

where $S(p) = \sum_{x \in \mathcal{X}} p_x x x^\top$ and $\beta_t = 2^{15} \log \frac{t|\mathcal{X}|}{\delta}$.

Figure 1: Near Instance-Optimal Randomized Algorithm for Stochastic and Corrupted Linear Bandits

4 Randomized Instance-optimal Algorithm

4.1 Algorithm Intuition

The near instance-optimal algorithm for stochastic and corrupted linear bandits is listed in Algorithm 1 (Figure 1). The algorithm proceeds in blocks with exponentially increasing round. Within each block, we first solve an optimization problem **OP** to get a distribution p_m (see Figure 1), which is inspired by the lower bound optimization problem:

$$\min_{N_x x \in \mathcal{X}, N_x \ge 0} \quad \sum_{x \in \mathcal{X}} N_x \Delta_x$$
$$\text{s.t.} \quad \|x\|^2_{H(N)^{-1}} \le \frac{\Delta_x^2}{2}, \quad \forall x \in \mathcal{X} \setminus \{x^*\},$$

where $H(N) = \sum_{x \in \mathcal{X}} N_x x x^\top$. Then, the arms are sampled and pulled according to the distribution p. Finally, we use an unbiased estimator $\widehat{l}_{t,x} = x^\top S_m^{-1} x_t y_t$ (same as adversarial linear bandits), and use a robust Catoni's estimator to compute the estimated gaps $\widehat{\Delta}_m$. This estimated gap is used in the optimization problem **OP** for the next round.

At the first glance, the structure of the algorithm is similar to the action elimination algorithm with G-optimal design. For the action elimination algorithm, within each block, we solve the G-optimal design problem, and pull arms deterministically to match the distribution. Then, we estimate the gap using the estimated $\widehat{\theta}$ given by the least-squares estimator.

The main differences between Algorithm 1 and the action elimination algorithm are the following:

- Instead of solving the same optimization problem (G-optimal design) on a shrinking set of arms, Algorithm 1 solves a different optimization problem in each block (given by different estimated gap $\widehat{\Delta}$) on the set of all arms. This helps the algorithm to achieve near instance-optimal regret bound. (To see why eliminating arms is bad, consider the example Kevin mentioned in class.)
- Instead of deterministically matching the counts, Algorithm 1 samples arms according to the solved distribution p_m. As a result, it uses the loss estimator for adversarial linear bandits rather than solving the least-squares problem. This is important for the corrupted setting.

- Algorithm 1 uses the Catoni estimator to establish the high probability regret bound.

4.2 Analysis of Algorithm 1

4.2.1 Main Theorem

Algorithm 1, *with high probability* achieves near instance-optimal regret bound (with an additional $\log(T|\mathcal{X}|/\delta)$ factor). This is formally stated in the theorem below (Theorem 1 in [12]):

Theorem 3 (Theorem 1, [12])**.** *In the corrupted setting, Algorithm 1 guarantees that with probability at least* $1-\delta$*,*

$$\mathrm{Reg}(T) = \mathcal{O}\left(c(\mathcal{X},\theta)\log T\log\frac{T|\mathcal{X}|}{\delta} + M^*\log^{3/2}\frac{1}{\delta} + C + d\sqrt{\frac{C}{\Delta_{\min}}\log\frac{C|\mathcal{X}|}{\Delta_{\min}\delta}}\right), \tag{27}$$

where M^* *is some constant that depends on* $\mathcal{X}$ *and* θ *only.*

Remark 1: Recall that in the stochastic setting, we have the lower bound that $\liminf_{T\to\inf}\mathbb{E}[\mathrm{Reg}(T)] \geq \Omega(c(\mathcal{X},\theta)\log T)$. Therefore, by setting $C=0$, we can see that in the stochastic setting, Algorithm 1 is near instance-optimal.

Remark 2: M^* is a constant that depends on $\mathcal{X}$ and θ only, and (importantly) does not depend on T. Hence it doe not change the rate.

4.2.2 Proof Sketch

We will prove Theorem 3 through proving the following Lemmas.

[Lemma for Robust Estimator] The first step is to show that the estimated gaps $\widehat{\Delta}_{m,x}$ with the robust estimator is close to the true gap Δ_x for each block with high probability.

Lemma 5 (Lemma 3, [12])**.** *With probability at least* $1-\delta$*, Algorithm 1 ensures for all* m *and all* x*,*

$$\Delta_x \leq 2\widehat{\Delta}_{m,x} + \sqrt{\frac{d\gamma_m}{4\cdot 2^m}} + 2\rho_{m-1}, \tag{28}$$

$$\widehat{\Delta}_{m,k} \leq 2\Delta_x + \sqrt{\frac{d\gamma_m}{4\cdot 2^m}} + 2\rho_{m-1}, \tag{29}$$

where $\rho_m = \sum_{k=0}^{m}\frac{2^k C_k}{4^{m-1}}$ ($\rho_{-1}:=0$), $C_k = \sum_{\tau\in\mathcal{B}_k}\max_{x\in\mathcal{X}}|c_{\tau,x}|$ is the amount of corruption within block k, and $\gamma_m = 2^{15}\log(2^m|\mathcal{X}|/\delta)$.

Remark: it is also important to notice that the last two terms $\sqrt{\frac{d\gamma_m}{4\cdot 2^m}}$ and $2\rho_{m-1}$ is shrinking (eventually to 0) as m increases. We will use this to show that with large enough t (and hence m), we can upper and lower bound $\widehat{\Delta}_{m,x}$ by a constant factor of Δ_x.

Proof. We will prove this lemma by induction.

- Base case: when $m=0$, by construction (line 5 in Algorithm 1), we have $\widehat{\Delta}_{0,x}=0$ for all $x\in\mathcal{X}$. Therefore, (29) holds trivially. By the assumption that $\langle x,\theta\rangle\in[-1,1]$ for all x, we also have $\Delta_x\leq 2\leq\sqrt{\frac{d2^{15}}{4}} = \sqrt{\frac{d\gamma_0}{4\cdot 2^0}}$. Therefore, conditions (28) and (29) hold for $m=0$.
- Induction: suppose that the conditions (28) and (29) hold for m, we need to show that they also hold for $m+1$.

  First, we need to consider the statistics of $\widehat{l}_{\tau,x}$, which will be used for analyzing the Catoni's estimator.

$$\begin{aligned}
\mathbb{E}[\widehat{l}_{\tau,x}] &= \mathbb{E}[x^\top S_m^{-1}x_\tau(\langle x_\tau,\theta+c_\tau\rangle+\epsilon_\tau(x_\tau))]\\
\text{(Tower rule of expectation)} \quad &= \mathbb{E}_{x_\tau}[\mathbb{E}_{\epsilon_\tau}[x^\top S_m^{-1}x_\tau(\langle x_\tau,\theta+c_\tau\rangle+\epsilon_\tau(x_\tau))|x_\tau]]\\
(\mathbb{E}[\epsilon_\tau(x_\tau)|x_\tau]=0) \quad &= \mathbb{E}[x^\top S_m^{-1}x_\tau x_\tau^\top(\theta+c_\tau)]\\
\text{(Linearity of expectation)} \quad &= x^\top S_m^{-1}\mathbb{E}[x_\tau x_\tau^\top](\theta+c_\tau)\\
\mathbb{E}[x_\tau x_\tau^\top]=S_m \quad &= x^\top(\theta+c_\tau)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\widehat{l}^2_{\tau,x}] &= \mathbb{E}[(x^\top S_m^{-1} x_\tau y_\tau)^2] \\
|y_\tau| \le 1 \quad &\le \mathbb{E}[(x^\top S_m^{-1} x_\tau)^2] = \mathbb{E}[x^\top S_m^{-1} x_\tau x_\tau^\top S_m^{-1} x] \\
\text{(Linearity of expectation)} \quad &= x^\top S_m^{-1} \mathbb{E}[x_\tau x_\tau^\top] S_m^{-1} x \\
\mathbb{E}[x_\tau x_\tau^\top] = S_m \quad &= x^\top S_m^{-1} x = \|x\|^2_{S_m^{-1}}
\end{aligned}$$

Due to that p_m is a feasible solution of $\mathbf{OP}(2^m, \widehat{\Delta}_m)$, we can further bound $\|x\|^2_{S_m^{-1}}$,

$$\begin{aligned}
\|x\|^2_{S_m^{-1}} &\le \frac{2^m \widehat{\Delta}^2_{m,x}}{\gamma_m} + 4d \\
(\text{(29) holds for } m) \quad &= \frac{2^m \left(2\Delta_x + \sqrt{\frac{d\gamma_m}{4\cdot 2^m}} + 2\rho_{m-1}\right)^2}{\gamma} + 4d \\
((a+b+c)^2 \le 4a^2 + 4b^2 + 4c^2) \quad &= \frac{16\cdot 2^m \Delta_x^2}{\gamma_m} + \frac{16\cdot 2^m \rho^2_{m-1}}{\gamma_m} + \frac{4\cdot 2^m}{\gamma_m} \cdot \frac{d\gamma_m}{4\cdot 2^m} + 4d \\
&\le \frac{16\cdot 2^m \Delta_x^2}{\gamma_m} + \frac{16\cdot 2^m \rho^2_{m-1}}{\gamma_m} + 8d
\end{aligned}$$

We can use this bound on the second moment to bound the difference between the estimated gap and the true gap. (We skipped a few steps during the class.) With probability at least $1 - \frac{\delta}{4^m}$, we have, for all

$x \in \mathcal{X}$,

$$
\begin{aligned}
\Delta_x - \widehat{\Delta}_{m+1,x} &= \langle x, \theta\rangle - \langle x^*, \theta\rangle - (\mathrm{Rob}_{m,x} - \min_{x'} \mathrm{Rob}_{m,x'}) \\
(\mathrm{Rob}_{m,x^*} \geq \min_{x'} \mathrm{Rob}_{m,x'}) \quad &= \langle x, \theta\rangle - \langle x^*, \theta\rangle - \mathrm{Rob}_{m,x} + \mathrm{Rob}_{m,x^*} \\
(C_k = \sum_{\tau\in\mathcal{B}_m} \max_{x\in\mathcal{X}} |\langle x, c_\tau\rangle|) \quad &\leq \left|\mathrm{Rob}_{m,x^*} - \langle x^*, \theta\rangle - \frac{\sum_{\tau\in\mathcal{B}_m}\langle x^*, c_\tau\rangle}{2^m}\right| \\
&\quad + \left|\mathrm{Rob}_{m,x} - \langle x, \theta\rangle - \frac{\sum_{\tau\in\mathcal{B}_m}\langle x, c_\tau\rangle}{2^m}\right| + \frac{2C_m}{2^m} \\
(\text{Catoni's estimator}) \quad &\leq \frac{2C_m}{2^m} + \frac{1}{2^m}\left(\alpha_x\left(2^m\|x\|^{-1}_{S_m} + \sum_{\tau\in\mathcal{B}_m}(\langle x, \tau\rangle - \bar{c}_{m,x})^2\right) + \frac{4\log(2^m|\mathcal{X}|/\delta)}{\alpha_x}\right) \\
&\quad + \frac{1}{2^m}\left(\alpha_{x^*}\left(2^m\|x^*\|^{-1}_{S_m} + \sum_{\tau\in\mathcal{B}_m}(\langle x^*, \tau\rangle - \bar{c}_{m,x^*})^2\right) + \frac{4\log(2^m|\mathcal{X}|/\delta)}{\alpha_{x^*}}\right) \\
(\sum_{\tau\in\mathcal{B}_m}(c_{\tau,x} - \bar{c}_{m,x})^2 \quad &\leq \frac{C_m}{2^{m-1}} + \frac{1}{2^m}\left(\alpha_x\left(2^m\|x\|^{-1}_{S_m} + 2^m\right) + \frac{\gamma_m}{2^{12}\alpha_x}\right) \\
\leq c^2_{\tau,x} \leq 2^m) \quad &\quad + \frac{1}{2^m}\left(\alpha_{x^*}\left(2^m\|x^*\|^{-1}_{S_m} + 2^m\right) + \frac{\gamma_m}{2^{12}\alpha_{x^*}}\right) \\
(\text{Optimal choice of } \alpha_x) \quad &\leq \frac{2}{64\cdot 2^m}\sqrt{\left(2^m\|x\|^2_{S_m^{-1}} + 2^m\right)\gamma_m} + \frac{2}{64\cdot 2^m}\sqrt{\left(2^m\|x^*\|^2_{S_m^{-1}} + 2^m\right)\gamma_m} + \frac{C_m}{2^{m-1}} \\
(\text{Bound on } \|x\|^2_{S_m^{-1}}) \quad &\leq \frac{2}{64\cdot 2^m}\sqrt{\left(\frac{16\cdot 2^{2m}\Delta_x^2}{\gamma_m} + 16\cdot 2^m d + \frac{16\cdot 2^{2m}\rho^2_{m-1}}{\gamma_m}\right)\gamma_m} \\
&\quad + \frac{2}{64\cdot 2^m}\sqrt{\left(16\cdot 2^m d + \frac{16\cdot 2^{2m}\rho^2_{m-1}}{\gamma_m}\right)\gamma_m} + \frac{C_m}{2^{m-1}} \\
&\leq \frac{4}{64\cdot 2^m}\sqrt{\left(\frac{16\cdot 2^{2m}\Delta_x^2}{\gamma_m} + 16\cdot 2^m d + \frac{16\cdot 2^{2m}\rho^2_{m-1}}{\gamma_m}\right)\gamma_m} + \frac{C_m}{2^{m-1}} \\
(\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}) \quad &\leq \frac{\Delta_x}{2} + \sqrt{\frac{d\gamma_m}{16\cdot 2^m}} + \frac{\rho_{m-1}}{4} + \frac{C_m}{2^{m-1}} \\
&\leq \frac{\Delta_x}{2} + \sqrt{\frac{d\gamma_m}{16\cdot 2^m}} + \rho_m
\end{aligned}
$$

By rearranging the terms, we have $\Delta_x \leq 2\widehat{\Delta}_{m+1,x} + \sqrt{\frac{d\gamma_m}{4\cdot 2^m}} + 2\rho_m$. We can show the other condition through similar steps.

Finally, we can finish the proof by taking the union bound over all m. □

[Lemma for the Optimization Problem OP] We next show that the optimization problem **OP** solved by each block is always *feasible*, and show a upper bound on the *optimal value* $\sum_x p_x \widehat{\Delta}_{m,x}$. This, when combined with Lemma 5, can be used to further bound the regret.

Lemma 6 (Lemma 4, [12])**.** *Let p be the solution of* $\mathbf{OP}(t, \widehat{\Delta})$*. Then we have* $\sum_{x\in\mathcal{X}} p_x\widehat{\Delta}_x = \mathcal{O}\left(\frac{d\log(t|\mathcal{X}|/\delta)}{\sqrt{t}}\right)$.

Proof. Consider the optimization problem

$$\min_{p\in\mathcal{P}_\mathcal{X}} \sum_{x\in\mathcal{X}} p_x\widehat{\Delta}_x + \frac{2}{\xi}(-\log\det(S(p))). \tag{30}$$

Note that the second term is the objective for D-optimal design. We will construct a feasible solution of **OP** from the optimal solution of (30), which we denoted as p^*. By the KKT condition of (30), (here we

skip the step to show that p^* is also the optimal solution of the problem on the set of sub-distributions $p : \sum_x p_x \le 1, p_x \ge 0$),

$$\widehat{\Delta}_x - \frac{2}{\xi} x^\top S(p^*)^{-1} x - \lambda_x + \lambda = 0. \tag{31}$$

Multiplying by p_x^* and summing over all $x \in \mathcal{X}$,

$$\begin{aligned} 0 &= \sum_x p_x^* \widehat{\Delta}_x - \frac{2}{\xi} \sum_x p_x^* x^\top S(p^*)^{-1} x - \sum_x \lambda_x p_x^* + \lambda \\ \text{(complementary slackness)} \quad &= \sum_x p_x^* \widehat{\Delta}_x - \frac{2}{\xi} \mathrm{Tr}(S(p^*) S(p^*)^{-1}) + \lambda \\ &= \sum_x p_x^* \widehat{\Delta}_x - \frac{2d}{\xi} + \lambda. \end{aligned}$$

Since $\sum_x p_x^* \widehat{\Delta}_x \ge 0$ and $\lambda \ge 0$, we must have $\sum_x p_x^* \widehat{\Delta}_x \le \frac{2d}{\xi}$ and $\lambda \le \frac{2d}{\xi}$.

Also by (31), we have,

$$\|x\|^2_{S(p^*)^{-1}} = \frac{\xi}{2}(\widehat{\Delta}_x - \lambda + \lambda) \le \frac{\xi \widehat{\Delta}_x}{2} + 0 + \frac{\xi}{2} \cdot \frac{2d}{\xi} = \frac{\xi \widehat{\Delta}_x}{2} + d.$$

Now we construct a distribution by mixing p^* and another distribution ν (in [12] it is denoted as $q^{G,k}$), i.e. $q = \frac{1}{2}p^* + \frac{1}{2}\nu$. Let $G = \{x : \widehat{\Delta}_x \le \frac{1}{\sqrt{t}}\}$. The probability distribution ν has the property that $\sum_{x \notin G} \nu_x \le \frac{1}{\sqrt{t}}$, and $\|x\|^2_{S(\nu)^{-1}} \le 2d$ for all $x \in G$. (The existence of such a distribution is given by Lemma 11 in [12].)

Feasibility: Choose $\xi = \frac{\sqrt{t}}{\beta_t}$. Then, for all $x \notin G$, we have,

$$\|x\|^2_{S(q)^{-1}} \le \|x\|_{\left(\frac{1}{2}S(p^*)\right)^{-1}} \le 2\left(\frac{\xi \widehat{\Delta}_x}{2} + d\right) = \xi \widehat{\Delta}_x + 2d = \frac{\sqrt{t}\widehat{\Delta}_x}{\beta_t} + 2d = \frac{t\frac{1}{\sqrt{t}}\widehat{\Delta}_x}{\beta_t} + 2d \le \frac{t\widehat{\Delta}_x^2}{\beta_t} + 4d.$$

Similarly, for any $x \in G$, we have,

$$\|x\|^2_{S(q)^{-1}} \le \|x\|_{\left(\frac{1}{2}S(\nu)\right)^{-1}} \le 2 \cdot 2d = 4d \le \frac{t\widehat{\Delta}_x^2}{\beta_t} + 4d.$$

Optimal Value: Let p be the optimal solution of **OP**, then by feasibility of q,

$$\begin{aligned} \sum_x p_x \widehat{\Delta}_x \le \sum_x q_x \widehat{\Delta}_x &= \frac{1}{2}\sum_x p_x^* \widehat{\Delta}_x + \frac{1}{2}\sum_x \nu_x \widehat{\Delta}_x \\ &\le \frac{1}{2} \cdot \frac{2d}{\xi} + \frac{1}{2}\sum_{x \in G} \nu_x \widehat{\Delta}_x + \frac{1}{2}\sum_{x \notin G} \nu_x \widehat{\Delta}_x \\ &\le \frac{d\beta_t}{\sqrt{t}} + \frac{1}{2\sqrt{t}} + \frac{2}{2\sqrt{t}} = \mathcal{O}\left(\frac{d\beta_t}{\sqrt{t}}\right) = \mathcal{O}\left(\frac{d\log(t|\mathcal{X}|/\delta)}{\sqrt{t}}\right) \end{aligned}$$

□

[Anytime Regret Bound] By combining Lemma 5 and Lemma 6, we can obtain the following sublinear anytime regret bound, which is $\mathcal{O}(d\sqrt{T}\log T + C)$.

Theorem 4 (Theorem 2, [12])**.** *In the corrupted setting, Algorithm 1 guarantees that with probability at least* $1 - \delta$, $\mathrm{Reg}(T) \le \mathcal{O}(d\sqrt{T}\log(T|\mathcal{X}|/\delta) + C)$.

Proof. Step 1: Relate the regret with $\sum_t \sum_x p_{m_t,x}\Delta_x$ using a concentration inequality. Note that we did not need this for the algorithms we saw previously in class, as during each block, we were explicitly matching pulls rather than drawing from the distribution.

By Freedman's Inequality, we have,

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}\sum_{x\in\mathcal{X}} \mathbb{1}\{x_t = x\}\Delta_t \\
&\le \sum_{t=1}^{T}\sum_{x\in\mathcal{X}} p_{m_t,x}\Delta_x + 2\sqrt{\log(1/\delta)\sum_{t=1}^{T}\mathbb{E}_t\left[\left(\sum_{x\neq x^*}\mathbb{1}\{x_t = x\}\Delta_x\right)^2\right]} + \log(1/\delta) \\
&\le \sum_{t=1}^{T}\sum_{x\in\mathcal{X}} p_{m_t,x}\Delta_x + 2\sqrt{\log(1/\delta)\sum_{t=1}^{T}\mathbb{E}_t\left[\sum_{x\neq x^*}\mathbb{1}\{x_t = x\}\Delta_x\right]} + \log(1/\delta) \\
&\le \sum_{t=1}^{T}\sum_{x\in\mathcal{X}} p_{m_t,x}\Delta_x + 2\sqrt{\log(1/\delta)\sum_{t=1}^{T}\sum_{x\in\mathcal{X}} p_{m_t,x}\Delta_x} + \log(1/\delta) \\
&\le 2\sum_{t=1}^{T}\sum_{x\in\mathcal{X}} p_{m_t,x}\Delta_x + 2\log(1/\delta)
\end{aligned}$$

Step 2: Bounding $\sum_t \sum_x p_{m,x}\Delta_x$. By Lemma 6, we have, for all m,

$$\sum_{x\in\mathcal{X}} p_{m,x}\widehat{\Delta}_{m,x} \le \mathcal{O}\left(\frac{d\log(2^m|\mathcal{X}|/\delta)}{2^{m/2}}\right).$$

□

Combining this with Lemma 5 (relating Δ with $\widehat{\Delta}$), we have,

$$\begin{aligned}
2^m\sum_{x\in\mathcal{X}} p_{m,x}\Delta_x &\le 2^m\left(2\sum_{x\in\mathcal{X}} p_{m,x}\widehat{\Delta}_{m,x} + \mathcal{O}\left(\sqrt{\frac{d\gamma_m}{2^m}} + \rho_{m-1}\right)\right) \\
&\le \mathcal{O}\left(d2^{m/2}\log(2^m/\delta)\right) + \mathcal{O}\left(\sqrt{2^m d\gamma_m} + \sum_{k=0}^{m-1}\frac{2^k C_k}{4^{m-2}}\right) \\
&\le \mathcal{O}\left(d2^{m/2}\log(2^m/\delta) + \sqrt{2^m d\gamma_m} + \sum_{k=0}^{m}\frac{C_k}{2^{m-k}}\right).
\end{aligned}$$

Step 3: Summing up till round T (block $\lfloor\log_2 T\rfloor$), we have,

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{x\in\mathcal{X}} p_{m_t,x}\Delta_x &\le \sum_{m=0}^{\lfloor\log_2 T\rfloor}\mathcal{O}\left(d2^{m/2}\log(2^m|\mathcal{X}|/\delta) + \sqrt{2^m d\gamma_m}\right) + \mathcal{O}\left(\sum_{k=0}^{\lfloor\log_2 T\rfloor} C_k\sum_{m=k}^{\lfloor\log_2 T\rfloor}\frac{1}{2^{m-k}}\right) \\
&= \mathcal{O}\left(d\sqrt{T}\log(T|\mathcal{X}|/\delta) + \sum_{k=0}^{\lfloor\log_2 T\rfloor} C_k\right) \\
&= \mathcal{O}\left(d\sqrt{T}\log(T|\mathcal{X}|/\delta) + C\right).
\end{aligned}$$

["Regret Bound" for Large t Onward]

Lemma 7 (Lemma 15, [12]). *Suppose $\widehat{\Delta}_x \in [\frac{1}{\sqrt{r}}\Delta_x, \sqrt{r}\Delta_x]$ for all $x \in \mathcal{X}$ for some $r > 1$, and $p = \mathbf{OP}(t, \widehat{\Delta})$ for some $t \ge r\beta_t M$, where $M = \sum_{x\in\mathcal{X}} N_x^*$ with N_x^* such that $\sum_{x\in\mathcal{X}} N_x\Delta_x \le 2c(\mathcal{X},\theta)$, and $\|x\|^2_{H(N)^{-1}} \le \frac{\Delta_x^2}{2}$ for all $x \in \mathcal{X}\setminus\{x^*\}$[1]. Then $\sum_{x\in\mathcal{X}} p_x\Delta_x \le \frac{r^2\beta_t}{t}c(\mathcal{X},\theta)$.*

[1]The existence of such a finite M is shown in Lemma 14 in [12]

Proof. Consider N^* defined in the Lemma, define a distribution $\tilde{p}$ as,

$$\tilde{p}_x = \begin{cases} \frac{r\beta_t N_x^*}{2t}, & x \neq x^*, \\ 1 - \sum_{x' \neq x^*} \tilde{p}_{x'}, & x = x^*. \end{cases}$$

Since $t \geq r\beta_t M$, we have,

$$\tilde{p}_{x^*} = 1 - \sum_{x' \neq x^*} \frac{r\beta_t N_X^*}{2t} \geq 1 - \frac{r\beta_t M}{2t} \geq \frac{1}{2} \geq \frac{r\beta_t M}{2t} \geq \frac{r\beta_t N_x^*}{2t}.$$

Therefore, $\tilde{p}$ is a valid distribution, and $\tilde{p}_x \geq \frac{r\beta_t N_x^*}{2t}$ for all $x \in \mathcal{X}$.
Now we show that $\tilde{p}$ is feasible for $\mathbf{OP}(t, \widehat{\Delta}_x)$. For all $x \neq x^*$,

$$\|x\|^2_{S(\tilde{p})^{-1}} \leq \|x\|^2_{(\sum_{y \in \mathcal{X}} \frac{r\beta_t N_y^*}{2t} yy^\top)^{-1}} = \frac{2t}{r\beta_t} \|x\|^2_{(\sum_{y \in \mathcal{X}} N_y^* yy^\top)^{-1}} \leq \frac{t\Delta_x^2}{r\beta_t} \leq \frac{t\widehat{\Delta}_x^2}{\beta_t} \leq \frac{4t\widehat{\Delta}_x^2}{\beta_t} + 4d.$$

For $x = x^*$, we have,

$$\|x^*\|_{S(\tilde{p})^{-1}} = \|S(\tilde{p})^{-1} x^*\|^2_{S(\tilde{p})} \geq \|S(\tilde{p})^{-1} x^*\|^2_{\frac{1}{2} x^* x^{*\top}} = \frac{1}{2} \|x^*\|^4_{S(\tilde{p})^{-1}} \quad \Longrightarrow \quad \|x^*\|^2_{S(\tilde{p})^{-1}} \leq 2.$$

Finally, by $\widehat{\Delta}_x \in [\frac{1}{\sqrt{r}} \Delta_x, \sqrt{r} \Delta_x]$, we have,

$$\begin{aligned} \sum_{x \in \mathcal{X}} p_x \Delta_x &\leq \sqrt{r} \sum_{x \in \mathcal{X}} p_x \widehat{\Delta}_x \leq \sqrt{r} \sum_{x \in \mathcal{X}} \sum_{x \in \mathcal{X}} \tilde{p}_x \widehat{\Delta}_x \leq \sqrt{r} \sum_{x \in \mathcal{X}} \frac{r\beta N_x^*}{2t} \widehat{\Delta}_x \\ &\leq \frac{r^2 \beta_t}{t} \sum_{x \in \mathcal{X}} N_x^* \Delta_x \leq \frac{r^2 \beta_t}{t} c(\mathcal{X}, \theta). \end{aligned} \tag{32}$$

□

Lemma 8 (Lemma 17, [12]). *Let $T^* := \frac{32C}{\Delta_{\min}} + 4M' \log\left(\frac{2M'|\mathcal{X}|}{\delta}\right)$, where $M' = 2^{20}\left(M + \frac{d}{\Delta_{\min}^2}\right)$, and M is defined in the previous Lemma. Then Algorithm 1 guarantees with probability at least $1 - \delta$:*

$$\sum_{t=T^*+1}^{T} \sum_{x} p_{m_t,x} \Delta_x \leq \mathcal{O}(c(\mathcal{X}, \theta) \log T \log(T/\delta)). \tag{33}$$

Proof Sketch: For some large T^*, for $t \geq T^*$, we have

$$\sqrt{\frac{d\gamma_{m_t}}{4 \cdot 2^{m_t}}} + 2\rho_{m_t} \leq \frac{1}{2}\Delta_{\min} \leq \frac{1}{2}\Delta_x, \quad \forall x \in \mathcal{X} \setminus \{x^*\}. \tag{34}$$

With this, we have that $\Delta_x \leq 4\widehat{\Delta}_{m_t,x}$, and $\widehat{\Delta}_{m_t,x} \leq 4\Delta_x$.
Therefore, by Lemma 7,

$$\sum_{x \in \mathcal{X}} p_{m_t,x} \Delta_x \leq \mathcal{O}\left(\frac{\beta_t}{2^{m_t}} c(\mathcal{X}, \theta)\right).$$

Summing over $t \geq T^* + 1$, we get, with probability at least $1 - \delta$,

$$\sum_{t=T^*+1}^{T} \sum_{x \in \mathcal{X}} p_{m_t,x} \Delta_x = \mathcal{O}(\beta_T c(\mathcal{X}, \theta) \log T) = \mathcal{O}\left(c(\mathcal{X}, \theta) \log T \log(T|\mathcal{X}|/\delta)\right).$$

□

Finally, we can prove Theorem 3 by applying Theorem 4 for $t \leq T^*$ and Lemma 8 for $t \geq T^* + 1$.

Remark: This paper additionally show a regret bound of $\mathcal{O}\left(\frac{d^2}{\Delta_{\min}} \log^2\left(\frac{T|\mathcal{X}|}{\Delta_{\min}\delta}\right) + C\right)$ (see Theorem 19 in [12]), which is independent of M^*. Though they noted that this bound could be looser than Theorem 1, as $c(\mathcal{X}, \theta) \leq \mathcal{O}\left(\frac{d}{\Delta_{\min}}\right)$.

5 Best of Three Worlds Algorithm

Algorithm 2 BOTW (Best of Three Worlds)

Input: an algorithm $\mathcal{A}$ satisfying Assumption 1.
Initialize: $L \leftarrow L_0$ (L_0 defined in Assumption 1).
while *true* **do**
  Run BOTW-SE with input L, and receive output t_0.
  $L \leftarrow 2t_0$.

Algorithm 3 BOTW-SE (BOTW – Single Epoch)

Input: L (minimum duration)
Define: $f_T = \log T$
Initialize: a new instance of $\mathcal{A}$.
// Phase 1
for $t = 1, 2, \ldots$ **do**
  Execute and update $\mathcal{A}$. Receive estimators $\{\widehat{\ell}_{t,x}\}_{x \in \mathcal{X}}$.
  if $t \geq L$ and there exists an action $\widehat{x}$ such that

$$\sum_{s=1}^{t} y_s - \sum_{s=1}^{t} \widehat{\ell}_{s,\widehat{x}} \geq -5\sqrt{f_T C_1 t}, \tag{9}$$

$$\sum_{s=1}^{t} y_s - \sum_{s=1}^{t} \widehat{\ell}_{s,x} \leq -25\sqrt{f_T C_1 t}, \quad \forall x \neq \widehat{x}, \tag{10}$$

  then $t_0 \leftarrow t$, $\widehat{\Delta}_x \leftarrow \frac{1}{t_0}\left(\sum_{s=1}^{t_0} \widehat{\ell}_{s,x} - \widehat{\ell}_{s,\widehat{x}}\right)$, **break**.
// Phase 2
for $t = t_0 + 1, \ldots$ **do**
  Let $p_t = \mathbf{OP}(t, \widehat{\Delta})$ and $\widetilde{p}_t = \frac{1}{2}e_{\widehat{x}} + \frac{1}{2}p_t$.
  Sample $x_t \sim \widetilde{p}_t$ and observe y_t.
  Calculate $\widehat{\ell}_{t,x}$ and $\widehat{\Delta}_{t,x}$ based on Eq. (8) and Eq. (13).
  if

$$\exists x \neq \widehat{x}, \ \widehat{\Delta}_{t,x} \notin \left[0.39\widehat{\Delta}_x, 1.81\widehat{\Delta}_x\right] \text{ or} \tag{11}$$

$$\sum_{s=t_0+1}^{t} \left(y_s - \widehat{\ell}_{s,\widehat{x}}\right) \geq 20\sqrt{f_T C_1 t_0}. \tag{12}$$

  then break.
Return t_0.

Figure 2: Best of Three Worlds

While the best-of-two-worlds algorithm presented above attains near-optimal rates in the corrupted and stochastic setting, it has no guarantees in the adversarial setting. Building on this algorithm, the authors present another algorithm "Best of Three Worlds" which attains near-optimal rates in all regimes.

The best-of-three-worlds algorithm proceeds in epochs of exponentially increasing length, where at each epoch the algorithm attempts to identify which arm achieves the low regret, and, given this, designs an allocation which aims to approximate the asymptotically optimal allocation. Critically, it relies on access to some black-box linear bandit algorithm $\mathcal{A}$ satisfying the following assumption.

Assumption 1. *$\mathcal{A}$ is a linear adversarial bandit algorithm that outputs a loss estimator $\widehat{\ell}_{t,x}$. There exist constants $L_0, C_1 \geq 2^{15} d \log(T|\mathcal{X}|/\delta)$ and universal constant $C_2 \geq 20$ such that for $t \geq L_0$, $\mathcal{A}$ guarantees the following with probability at least $1 - \delta$ for $\forall x \in \mathcal{X}$:*

$$\sum_{s=1}^{t} (\ell_{s,x_s} - \ell_{s,x}) \leq \sqrt{C_1 t} - C_2 \left| \sum_{s=1}^{t} (\ell_{s,x} - \widehat{\ell}_{s,x}) \right|.$$

They show that several standard adversarial linear bandit algorithms meet this assumption. In particular, there exist algorithms $\mathcal{A}$ such that $C_1 = \Theta(d \log(T|\mathcal{X}|/\delta))$ and $L_0 = \Theta(d \log^2(T|\mathcal{X}|/\delta))$. Their main result is the following.

Theorem 5 (Theorem 6 and 7)**.** *The best-of-three-worlds algorithm guarantees that with probability at least $1 - \delta$, in the stochastic setting with $C = 0$, the regret is bounded as*

$$\mathcal{O}\left(c(\mathcal{X}, \theta) \log T \log \frac{T|\mathcal{X}|}{\delta} + \frac{C_1 \sqrt{\log T}}{\Delta_{\min}} + M^* \log^{3/2} \frac{1}{\delta} + \sqrt{C_1 L_0}\right)$$

and in the stochastic corrupted setting with $C > 0$, the regret is bounded as

$$\mathcal{O}\left(\frac{C_1 \log T}{\Delta_{\min}} + C + \sqrt{C_1 L_0}\right).$$

Furthermore, in the adversarial setting, the best-of-three-worlds algorithm will have regret bounded as, with probability $1 - \delta$,

$$\mathcal{O}(\sqrt{C_1 T \log T} + \sqrt{C_1 L_0}).$$

Therefore, they hit the optimal instance-dependent rate, up to a factor of $\log T$, in the purely stochastic setting, an additive in C rate in the stochastic corrupted setting, and the minimax optimal rate in the adversarial setting all simultaneously and without requiring knowledge of which setting they are in.

5.1 Algorithm Intuition

Adversarial Case. In Phase 1, Algorithm 3 runs some algorithm $\mathcal{A}$ which is a black-box algorithm which achieves low regret for adversarial bandits. It follows, then, that the regret incurred in Phase 1 should be low.

The algorithm will only enter Phase 2 if there is a single arm, $\widehat{x}$, such that the estimated loss incurred by $\widehat{x}$ is significantly less than the loss of any other arm. Assuming this condition is met and we are in Phase 2, it is shown that, due to the first exit condition of Phase 2 (equation (11)), as long as we are in Phase 2, $\widehat{x}$ is still the optimal arm. As the second exit condition of Phase 2 (equation (12)) compares the loss incurred to the loss of $\widehat{x}$, it follows that this is a valid surrogate for the regret incurred in this phase. Thus, if the regret incurred becomes too large, the algorithm will exit Phase 2 and return to Phase 1, where it knows it is able to achieve low regret.

Combining these two facts allows one to show that the total regret incurred in the adversarial case is small.

Stochastic, Uncorrupted Case. In the stochastic, uncorrupted case, one can show that the algorithm will only leave Phase 1 once $\widehat{x} = x^*$ and all the gaps are estimated well. Furthermore, one can show that it will take at most $\mathcal{O}(1/\Delta_{\min}^2)$ steps to reach the exit condition of Phase 1, so the total regret of Phase 1 can be bounded conveniently by a constant that does not depend on T.

Once the algorithm reaches Phase 2, if the optimal arm has been properly identified and the gaps are also well-estimated, the solution to the asymptotically optimal allocation on the estimated bandit will be near the solution to the asymptotically optimal allocation on the true bandit. As such, the total regret incurred in Phase 2 can be bounded by the asymptotically optimal constant times $\operatorname{poly}\log T$. Notably, as discussed below, since the algorithm must continually check if it is truly in the stochastic regime, it must continually play the asymptotically optimal allocation even after identifying the best arm. This extra exploration incurs the extra $\log T$ factor.

Stochastic, Corrupted Case. In the stochastic, corrupted case the algorithm first behaves like in the adversarial case, and then ultimately behaves like in the stochastic, uncorrupted case. Since C is finite—there is a finite level of corruption—for T large enough, we will have learned the optimal arm and all the gaps well enough to have "solved" the problem, and allowing us to incur low-regret. However, until T is large enough, the instance behaves more like an adversarial instance.

As such, we can use the bound in the adversarial case to bound the total regret incurred before T is large enough and we have solved the problem, and once T is large enough we can use the bound in the stochastic case. It can be shown that this breaking point is roughly when Phase 1 is $\mathcal{O}(C^2 + \log^2 T/\Delta_{\min}^2)$ steps, which ultimately yields that the regret of Phase 1 is bounded as $C + \log^2 T/\Delta_{\min}$. Once the problem is solved, it will incur regret $\mathcal{O}(c(\mathcal{X}, \theta) \log^2 T)$. However, the $\log^2 T/\Delta_{\min}$ regret incurred in Phase 1 will dominate this, yielding the worse rate in the corrupted setting.

5.2 Discussion

Burn-in time for stochastic result. In some ways the stochastic bound is almost vacuous. In the stochastic case, their algorithm functions by estimating the best arm and all gaps correctly with probability $1-\delta$. At this point, they have effectively "solved" the problem and, if they knew they were in the stochastic regime, could simply play the best arm and incur 0 regret, rather than playing the allocation from **OP** at all.

While this does seem somewhat trivial, the need to protect against the adversarial makes this result still meaningful. Even in Phase 2 they need to constantly verify that their estimates of the gaps are correct. Playing the allocation **OP** is the optimal way to do this—it balances the need to explore so as to minimize regret while still guaranteeing all the gap estimates are good. Thus, by playing the solution to **OP**, they are able to ensure that there is enough exploration to account for an adversary, but that the regret is still as small as possible in the stochastic regime.

Where does the $\log^2 T$ come from? It seems the $\log(T|\mathcal{X}|/\delta)$ term comes from union bounding over all time for all arms. The black-box adversarial regret algorithm that is applied in Phase 1 needs such a union bound to guarantee that at every timestep the regret bound holds. This is then needed to guarantee that (a) the gaps are well estimated and (b) the first phase achieves low regret.

Once the algorithm moves to Phase 2, it needs to constantly verify that the estimates of the gaps remain accurate to protect against an adversary corrupting them. Solving **OP** optimally balances this need for exploration with the desire to achieve low regret (as is noted above). However, it needs to ensure that at every step for every arm the estimates remain good—so as to prevent an adversary from tricking the algorithm—and, as such, the $\log(T|\mathcal{X}|/\delta)$ is necessary in this stage as well.

The second $\log T$ term comes from upper bounding the regret incurred by playing the solution to **OP** in Phase 2 rather than just playing the optimal arm. Intuitively, then, the reason we get a $\log^2 T$ is that, unlike the stochastic regime where you can stop exploring and just play the optimal arm, in this regime you have to continually explore so as to always ensure you are not in the adversarial case. The lower bound presented below seems to imply the $\log^2 T$ is necessary.

Why can't this hit the $c(\mathcal{X}, \theta)$ rate in the corrupted setting? Note that in the purely stochastic setting, we only are in Phase 1 once, which allows us to simply apply the bound $\sqrt{C_1 t_0}$ to bound the regret in the initial stage. This ultimately scales as $\mathcal{O}(\log^{3/2} T)$, and so is lower order.

In the corrupted setting, however, we do not know how many times we will be in the first phase, as the total number of epochs depends on the value of C. As such, we can only apply the adversarial regret bound (Theorem 7), which has an additional $\sqrt{\log T}$ term. This causes the regret of the initial stages to dominate, and we get the $\mathcal{O}(d \log^2 T/\Delta_{\min})$ regret.

5.3 Lower Bound

Theorem 6 (Theorem 27)**.** *For any $\gamma \in (0,1)$, if an algorithm guarantees a regret of*

$$\frac{\gamma(1-\gamma)}{256 \log \frac{4}{\Delta_{\min}}} \cdot c(\mathcal{X}, \theta) \log^2 T$$

in stochastic environments for sufficiently large T, then there exists an adversarial environment such that with probability at least $\frac{1}{4} T^{-\gamma/4}$, the regret of the same algorithm is at least $\frac{1}{6} T^{\gamma} \Delta_{\min}$.

Intuition for lower bound. The proof of the lower bound works by constructing two instances and showing that, unless sufficient exploration is done, the learner will not be able to detect if the bandit switches from one instance to the other, in which case it will incur large regret on the second instance.

More formally, they show that if an algorithm achieves regret $\mathcal{O}(c(\mathcal{X}, \theta) \cdot \log^{2-\beta} T)$ on a fixed instance, there must be some arm x such that

$$\|x - x^*\|^2_{A_i^{-1}} \geq \Omega(\Delta_x^2 / \log^{1-\beta} T)$$

for some $\beta \in (0,1)$ and $A_i = \mathbb{E}[\sum_{t=t_i}^{t_{i+1}} x_t x_t^\top]$. In other words, the algorithm cannot achieve low regret and simultaneously distinguish between every arm well. Given this, consider an adversary switching to some instance

$$\theta' = \theta - \frac{A_i^{-1}(x - x^*)}{\|x - x^*\|^2_{A_i^{-1}}} 2\Delta_x.$$

Now, x is a better action than x^*, but the learner cannot distinguish between the two as stated above, and will thus incur large regret.

Intuitively, then, the learner must explore at a rate greater than $\log T$ in order to always be able to estimate all gaps well enough to ensure that the loss vector has not been changed by an adversary.

Caveats of lower bound. There seem to be a few shortcomings of the lower bound. First, the observation model is non-standard. Rather than observing $x_t^\top \ell_t + \eta_t$ as is standard in linear bandits, they assume the observation, y_t, is

$$y_t = \begin{cases} 1 & \text{with probability } \frac{1}{2} + \frac{1}{2} x_t^\top \theta \\ -1 & \text{with probability } \frac{1}{2} - \frac{1}{2} x_t^\top \theta \end{cases}$$

So rather than the full feedback, they're getting a binary feedback. This seems more difficult to learn in perhaps, but is essentially still sub-gaussian noise.

Second, one would ideally like to be able to say that the bad event occurs with constant probability but in their case the bad event occurs with probability $\frac{1}{4} T^{-\gamma/4}$, which will go to 0 as $T \to \infty$. In order to make the probability of the bad event a constant, you need to choose $\gamma = \mathcal{O}(1/\log T)$, which will make the regret of the bad event a constant that does not scale with T. However, with this model the expected regret will be at least $\mathcal{O}(T^{3\gamma/4})$, which is still potentially worse than the optimal rate, so this is perhaps not that big an issue.

5.4 Proof Sketch (Adversarial Case)

The proof in the adversarial setting involves three key steps:

1. Show that, since we are running a low-regret algorithm, the regret in Phase 1 is bounded as $\mathcal{O}(\sqrt{C_1 t_0})$.
2. Show that in Phase 2, the arm identified as empirically best, $\widehat{x}$, in Phase 1 remains the best arm for the duration of Phase 2.
3. Given this, we can use $\widehat{x}$ to accurately estimate the regret. Show that the conditions of Phase 2 then allow us to bound the regret in this phase as $\mathcal{O}(\sqrt{C_1 t_0 \log T})$.

The proof of the first step follows trivially by Assumption 1—since in Phase 1 we are running an algorithm with a low-regret guarantee for adversarial bandits, it follows that we will achieve low regret in this phase.

The following lemma proves the second step.

Lemma 9 (Lemma 8)**.** *With probability at least $1-\delta$, for any t in Phase 2, we have $\widehat{x} \in \arg\min_{x \in \mathcal{X}} \sum_{s=1}^{t} \ell_{s,x}$.*

Proof. We first show that the deviation in the first phase is well-bounded:

$$\begin{aligned}
\left| \sum_{s=1}^{t_0} (\ell_{s,x} - \widehat{\ell}_{s,x}) \right| &\le c \left(\sqrt{C_1 t_0} + \sum_{s=1}^{t_0} (\widehat{\ell}_{s,x} - \ell_{s,x_s}) \right) && \text{(Rearranging regret bound for Phase 1)} \\
&\le c \left(2\sqrt{C_1 t_0} + 5\sqrt{f_T C_1 t_0} + \sum_{s=1}^{t_0} (\widehat{\ell}_{s,x} - \widehat{\ell}_{s,\widehat{x}}) \right) && \text{(Definition of } \widehat{x}) \\
&\le c \left(7\sqrt{f_T C_1 t_0} + t_0 \widehat{\Delta}_x \right)
\end{aligned}$$

In Phase 2, we can use some properties of the robust estimator to get that

$$\left|(t-t_0)\text{Rob}_{t,x} - \sum_{s=t_0+1}^{t} \ell_{s,x}\right| \leq \frac{t\widehat{\Delta}_x}{16}$$

Using the above inequality, we can bound the deviation $\sum_{s=1}^{t_0}(\ell_{s,\widehat{x}} - \widehat{\ell}_{s,\widehat{x}})$. To bound the deviation for $s > t_0$, we can apply Freedman's inequality and, using that we choose $\widehat{x}$ with probability at least $1/2$ to control the variance, conclude that

$$\left|\sum_{s=1}^{t}(\ell_{s,\widehat{x}} - \widehat{\ell}_{s,\widehat{x}})\right| \leq 3\sqrt{f_T C_1 t}$$

Putting these pieces together, we have that

$$\begin{aligned}
\sum_{s=1}^{t}(\ell_{s,x} - \ell_{s,\widehat{x}}) &\geq \sum_{s=1}^{t_0}(\widehat{\ell}_{s,x} - \widehat{\ell}_{s,\widehat{x}}) + \left((t-t_0)\text{Rob}_{t,x} - \sum_{s=t_0+1}^{t}\widehat{\ell}_{s,\widehat{x}}\right) - c\sqrt{f_T C_1 t} - c' t\widehat{\Delta}_x \\
&\geq t\widehat{\Delta}_{t,x} - c\sqrt{f_T C_1 t} - c' t\widehat{\Delta}_x \\
&\geq t\widehat{\Delta}_{t,x} - 0.37 t\widehat{\Delta}_x \\
&> 0
\end{aligned}$$

where the final inequality holds since we are in Phase 2, which implies that $\widehat{\Delta}_{t,x} \in [0.39\widehat{\Delta}_x, 1.81\widehat{\Delta}_x]$. Thus, $\widehat{x}$ is still the best arm. □

Given this result, the following lemma proves the third step.

Lemma 10 (Lemma 9)**.** *With probability at least $1-\delta$, for any time t in Phase 2, we have that for any $x \in \mathcal{X}$,*

$$\sum_{s=1}^{t}(\ell_{s,x_s} - \ell_{s,x}) = \mathcal{O}(\sqrt{C_1 t_0 f_T})$$

Proof. By the previous result, we know that in Phase 2 the optimal arm is still $\widehat{x}$. It follows that

$$\sum_{s=t_0+1}^{t}(\ell_{s,x_s} - \ell_{s,\widehat{x}})$$

is an accurate measure of the regret. In practice, we only have access to the empirical quantity

$$\sum_{s=t_0+1}^{t}(y_s - \widehat{\ell}_{s,\widehat{x}})$$

The difference between these quantities is a martingale difference sequence with variance that can be bounded conveniently since we pull $\widehat{x}$ at least $1/2$ of the time, ensuring $\widehat{\ell}_{s,\widehat{x}}$ is well-behaved. It can then be shown that this deviation is $\mathcal{O}(\sqrt{t_0})$. As the exit condition of Phase 2 is that

$$\sum_{s=t_0+1}^{t}(y_s - \widehat{\ell}_{s,\widehat{x}}) \geq 20\sqrt{f_T C_1 t_0}$$

it follows that the regret of Phase 2 will be upper bounded as $\mathcal{O}(\sqrt{f_T C_1 t_0} + \sqrt{t_0})$. □

Combining these two results and using that the total number of epochs is at most $\mathcal{O}(\log T)$, an application of Jensen's inequality to combine the regret over all phases gives Theorem 7.

5.5 Proof Sketch (Corrupted Stochastic Case)

The key lemma towards proving Theorem 6 is the following.

Lemma 11 (Lemma 20)**.** *In the stochastic setting with corruptions, within a single epoch,*

1. *With probability at least* $1-4\delta$, $t_0 \le \max\{\frac{900 f_T C_1}{\Delta_{\min}^2}, \frac{900C^2}{f_T C_1}, L\}$.
2. *If* $C \le \frac{1}{30}\sqrt{f_T C_1 L}$, *then with probability at least* $1-\delta$, $\widehat{x} = x^*$.
3. *If* $C \le \frac{1}{30}\sqrt{f_T C_1 L}$, *then with probability at least* $1-2\delta$, $t_0 \ge \frac{64 f_T C_1}{\Delta_{\min}^2}$.
4. *If* $C \le \frac{1}{30}\sqrt{f_T C_1 L}$, *then with probability at least* $1-3\delta$, $\widehat{\Delta}_x \in [0.7\Delta_x, 1.3\Delta_x]$ *for all* $x \ne x^*$.

Proof. The key insight in proving this result is that for any t in Phase 1 and any x,

$$\left|\sum_{s=1}^{t}(\ell_{s,x} - \widehat{\ell}_{s,x})\right| \le \frac{1}{C_2}(\sqrt{C_1 t} + t\Delta_x + 2C) \tag{35}$$

In words, our estimate of the reward for each arm is within a factor of $\mathcal{O}(1/\sqrt{T} + C/T + \Delta_x)$ of it's true reward. this result follows directly from the guarantee of the low-regret algorithm we are running in Phase 1, which gives

$$C_2\left|\sum_{s=1}^{t}(\ell_{s,x} - \widehat{\ell}_{s,x})\right| \le \sqrt{C_1 t} + \sum_{s=1}^{t}(\ell_{s,x} - \ell_{s,x_s})$$

Furthermore, $\ell_{t,x_t} \ge \ell_{t,x^*} - \max_{x\in\mathcal{X}} |c_{t,x}|$, so

$$\sum_{s=1}^{t}\ell_{s,x_s} \ge \sum_{s=1}^{t}\ell_{s,x^*} - C$$

Combining these results and using the definition of Δ_x gives the result.

Given this, the four claims intuitively follow. In the case with low corruption, one can show that (35) and the exit condition of Phase 1 implies that we will only exit the phase after we have identified the best arm and the gaps for every other arm. That $t_0 = \Theta(\frac{f_T C_1}{\Delta_{\min}^2})$ follows intuitively since this is the amount of time necessary to identify the best arm if using a naive approach. In the corrupted case, we can bound $t_0 = \mathcal{O}(C^2)$ since in this regime, the terms scaling with t will dominate the C term in (35), which will allow the exit condition of Phase 1 to be reached. □

We then have the following result.

Lemma 12 (Lemma 21 and 22)**.** *When* $C \le \frac{1}{30}\sqrt{f_T C_1 L}$, *with high probability Phase 2 will never end.*

Proof. That the gap estimates remain accurate follows since t_0 is large enough to dominate the corruptions.

Next, it is shown that

$$\sum_{s=t_0+1}^{t}(y_s - \widehat{\ell}_{s,\widehat{x}}) \le 20\sqrt{f_T C_1 t_0} \tag{36}$$

for all remaining time. This follows by first relating $y_s - \widehat{\ell}_{s,\widehat{x}}$ to the true regret at time s, $\ell_{s,x_s} - \ell_{s,x^*}$. Again using that $\widehat{x}$ is pulled at least half the time, the variance of $(y_s - \widehat{\ell}_{s,\widehat{x}}) - (\ell_{s,x_s} - \ell_{s,x^*})$ can be controlled so this deviation can be bounded. Then, using that Lemma 20 holds in this case so the gaps are well estimated, Lemma 13 can be applied which gives a bound on the regret incurred by playing an allocation computed by **OP**. Combining these proves (36) holds for all time. □

Proof of Theorem 6. The proof of Theorem 6 is broken into two parts.

Case 1: $C = 0$**.** In the truly stochastic case, the requirement of Lemma 20 will be met at the first epoch. Thus, assuming that T is large enough, Phase 1 will only exit after it has found the best arm and estimated the gap of every arm correctly. Using the bound from Lemma 20 that $t_0 = \Theta(\frac{f_T C_1}{\Delta_{\min}^2})$, the regret in the first stage can be bounded by, using the regret bound from the black-box algorithm, as

$$\mathcal{O}(\sqrt{f_T C_1^2/\Delta_{\min}^2}) = \mathcal{O}(C_1\sqrt{\log T}/\Delta_{\min}) = \mathcal{O}(d\log^{3/2} T/\Delta_{\min})$$

Critically, as we are only in Phase 1 a single time, we can use the regret bound of the black-box algorithm which scales as $\sqrt{C_1 t_0}$, rather than applying the adversarial regret bound, which scales as $\sqrt{C_1 T\log T}$. The lack of that additional $\sqrt{\log T}$ term ensures that the regret of this initial stage is lower order.

Then, for large enough T, we can use that the gaps are well-enough estimated to guarantee that the empirical solution to **OP** is near the solution to **OP** on the true instance, so the regret of Phase 2 can be upper bounded as

$$\mathcal{O}(c(\mathcal{X},\theta)\log T\log(T|\mathcal{X}|/\delta) + M\log^{3/2} 1/\delta)$$

where here M is the time needed to guarantee that the time-constrained solution returned by **OP** achieves a regret scaling with $c(\mathcal{X},\theta)$.

Case 2: $C > 0$**.** By what was shown in Lemma 21, Phase 2 will never end once $L \geq cC^2/(C_1 f_T)$. Furthermore, by what was shown in Lemma 20, we will always have that, in the previous epochs, $t_0 \leq \max\{\frac{900 f_T C_1}{\Delta_{\min}^2}, \frac{900C^2}{f_T C_1}\}$. Applying the adversarial regret bound to all stages where $L \leq cC^2/(C_1 f_T)$, we get that the regret in these stages is bounded as

$$\mathcal{O}\left(\sqrt{C_1 f_T \max\{\frac{900 f_T C_1}{\Delta_{\min}^2}, \frac{900C^2}{f_T C_1}\}}\right) = \mathcal{O}\left(\sqrt{C_1 f_T}(\frac{\sqrt{f_T C_1}}{\Delta_{\min}} + \frac{C}{\sqrt{f_T C_1}})\right) = \mathcal{O}\left(\frac{d\log T\log(T|\mathcal{X}|/\delta)}{\Delta_{\min}} + C\right)$$

By Lemma 21, we will have that the regret in the final Phase 2 will be bounded as $\mathcal{O}(\sqrt{f_T C_1 t_0})$. By our bound on t_0 this can be shown to be order

$$\mathcal{O}\left(\frac{d\log T\log(T|\mathcal{X}|/\delta)}{\Delta_{\min}} + C\right)$$

which gives the result. □

6 Misspecified models

At time t, we can observe the measurement x_t, through the real value

$$y_t = \mu_{x_t} + \xi_t$$

with $|y| \leq B$ and $\mathbb{E}[\xi^2] \leq \sigma^2$. We assume that $\max_{x\in\mathcal{X}} |x^\top\theta^* - \mu_x| \leq h$. Note that this framework covers the (more classical) stochastic linear bandit setting where $h = 0$ (also called the well specified case).

We define $\tau(\mathcal{X}, h, \varepsilon)$ the number of samples needed to design an algorithm that returns an ε-optimal arm almost surely, when the set of measurements is $\mathcal{X} \subset \mathbb{R}^d$ and the misspecification tolerance is h.

The rest of this section investigate attempts to tackle the misspecified setting, where $h > 0$. We first discuss the lower bound provided in [11], which we then compare to the upper bounds of [11] and [5].

6.1 Lower bound, [11]

Note that this lower bound was first provided in [7]. The setting chosen is pure exploration in noiseless bandits. Namely, we assume that $\xi_t = 0$ at every time t.

Theorem 7 ([7], [11]). *For all $\varepsilon > h$ and $k \gtrsim \exp\left\{d\left(\frac{h}{\varepsilon}\right)^2\right\}$, there exists a set of measurements $\mathcal{X} \subset \mathbb{R}^d$ of size k such that the number of samples needed to design an algorithm that returns an ε-optimal arm almost surely is*

$$\tau(\mathcal{X}, h, \varepsilon) \gtrsim \exp\left\{d\left(\frac{h}{\varepsilon}\right)^2\right\}$$

The proof can be found in e.g. [11]. The set of measurements $\mathcal{X}$ used is obtained using the Johnson-Lidenstrauss lemma.
We see that the query complexity is exponential in d (curse of dimensionality) when ε is not much larger than h, but is benign when $\varepsilon = \Omega(h\sqrt{d})$.

6.2 Upper bounds, [11] and [5]

Theorem 8 ([11]). *Let $\mathcal{X} \in \mathbb{R}^{k\times d}$ and $\varepsilon > 2h(1+\sqrt{2d})$, there exists an algorithm requiring its number of samples to be at most $4d\log\log(d) + 16$ to find an ε-optimal action. Thus*

$$\tau(\mathcal{H}^h_{\mathcal{X}}, \varepsilon) \le 4d\log\log(d) + 16$$

The algorithm used goes as follows:
- Use $4d\log\log(d) + 16$ measurements to compute a near G-optimal design. That is compute λ_0 such that $g(\lambda_0) \le 2\min_{\lambda\in\triangle_{\mathcal{X}}} g(\lambda) = 2d$, where

$$g(\lambda) := \max_{a\in\mathcal{X}} \|a\|^2_{\left(\sum_{a\in\mathcal{X}}\lambda(a)aa^\top\right)^{-1}}$$

- Compute a least squares estimate with these $4d\log\log(d) + 16$ measurements
- Output the measurement that looks best according to the least squares estimate.

Proposition 1. *[Robust IPS estimator, [5]] Fix any finite sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{V} \subset \mathcal{Z}$, number of samples τ and regularization $\gamma > 0$. If the estimator is run with $\frac{\delta}{|\mathcal{V}|}$-robust mean estimator $\widehat{\mu}(\cdot)$ and if $\tau \ge c_1\log(|\mathcal{V}|/\delta)$ then with probability at least $1-\delta$, we have*

$$\max_{v\in\mathcal{V}} \frac{|W^{(v)} - \langle\theta_*, v\rangle|}{\|v\|_{(\sum_{x\in\mathcal{X}}\lambda_x xx^\top + \gamma I)^{-1}}} \le \sqrt{\gamma}\|\theta_*\|_2 + h + c\sqrt{\tfrac{(B^2+\sigma^2)}{\tau}\log(2|\mathcal{V}|/\delta)},$$

Moreover, $W^{(v)} = \widehat{\mu}(\{v^\top A^{(\gamma)}(\lambda)^{-1}x_t y_t\}_{t=1}^\tau)$ can be replaced by $\langle\widehat{\theta}, v\rangle$ by multiplying the RHS by a factor of 2.

The full proof can be found in [5], but we provide a sketch of the proof here.

Proof sketch. Due to the regularization and potential misspecification if $h > 0$, each $v^\top A^{(\gamma)}(\lambda)^{-1}x_t y_t$ is biased. Thus, we apply the guarantee of $W^{(v)} = \widehat{\mu}(\{v^\top A^{(\gamma)}(\lambda)^{-1}x_t y_t\}_{t=1}^\tau)$ to the expectation of its arguments. The triangle inequality followed by repeated applications of Cauchy-Schwartz yields

$$\begin{aligned}
|W^{(v)} - \langle v, \theta_*\rangle| &\le |W^{(v)} - \mathbb{E}[v^\top A^{(\gamma)}(\lambda)^{-1}x_1 y_1]| + |\mathbb{E}[v^\top A^{(\gamma)}(\lambda)^{-1}x_1 y_1] - \langle v, \theta_*\rangle| \\
&\le c\sqrt{\frac{\nu^2\log(1/\delta)}{\tau}} + \sqrt{\gamma}\|\theta_*\|_2 + h
\end{aligned}$$

where we obtain an upper bound on the variance ν^2 by

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(v^\top A^{(\gamma)}(\lambda)^{-1}x_1 y_1) &\le \mathbb{E}[(v^\top A^{(\gamma)}(\lambda)^{-1}x_1 y_1)^2] \\
&= \mathbb{E}\left[\left(v^\top A^{(\gamma)}(\lambda)^{-1}x_1\right)^2 \mu^2_{x_1}\right] + \mathbb{E}\left[\left(v^\top A^{(\gamma)}(\lambda)^{-1}x_1\right)^2 \xi_1^2\right] \\
&\le (B^2+\sigma^2)\|v\|^2_{A^{(\gamma)}(\lambda)^{-1}}.
\end{aligned}$$

□

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

[2] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. 2009.

[3] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[4] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41–1. JMLR Workshop and Conference Proceedings, 2012.

[5] Romain Camilleri, Julian Katz-Samuels, and Kevin Jamieson. High-dimensional experimental design and kernel bandits, 2021.

[6] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study, 2011.

[7] Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning?, 2020.

[8] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

[9] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[10] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

[11] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model, 2020.

[12] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously, 2021.

[13] Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

[14] László Lovász. *Geometric algorithms and algorithmic geometry.* American Mathematical Society, 1990.

[15] Gabor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions–a survey, 2019.

[16] Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *arXiv preprint arXiv:2010.12247*, 2020.

[17] Andrew Wagenmaker, Julian Katz-Samuels, and Kevin Jamieson. Experimental design for regret minimization in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3088–3096. PMLR, 2021.

[18] Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits?, 2020.

[19] Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.