Université de Lille
Ecole Doctorale MADIS

Thèse de Doctorat

Spécialité **Informatique**

présentée par
Achraf Azize

Privacy-Utility Trade-offs in Sequential Decision-Making under Uncertainty

---

Compromis entre confidentialité et utilité dans la prise de décision séquentielle dans l'incertain

---

sous la direction de **Philippe Preux** et de **Debabrota Basu**.

Soutenue publiquement à **Villeneuve d'Ascq**, le **26/11/2024** devant le jury composé de

M. Adam **Smith**	Professeur, Boston University	Rapporteur
M. Aurélien **Garivier**	Professeur, ENS Lyon	Rapporteur
M. Christos **Dimitrakakis**	Professeur, Université de Neuchâtel	Président du Jury
M^{me} Catuscia **Palamidessi**	Directrice de recherche, Inria Saclay	Examinatrice
M. Aurélien **Bellet**	Directeur de recherche, Inria Montpellier	Invité
M. Philippe **Preux**	Professeur, Université de Lille	Directeur de thèse
M. Debabrota **Basu**	Chargé de recherche, Inria Lille	Co-Directeur de thèse

---

Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL),
UMR 9189 Équipe Scool, 59650, Villeneuve d'Ascq, France

Remerciements

First and foremost, I would like to express my deepest gratitude to you, Deb, for your unwavering enthusiasm, availability, and kindness throughout this journey. Your guidance and encouragement have been invaluable to me, and I am truly grateful for your support. I would also like to thank you, Philippe, for supporting me, especially when random challenges arose. Your steadfast support and reliability made navigating these hurdles so much easier.

Thanks to you, Adam and Aurélien G, for your thoughtful and constructive comments. Your works on privacy and bandits have been a massive inspiration to me. Your insights have contributed significantly to improving the quality of this manuscript. I am also profoundly thankful to Catuscia, Christos, and Aurélien B for being part of my thesis jury and for their valuable feedback.

I want to acknowledge all the Scool team members at Inria for the joyful moments and camaraderie we shared. From playing pétanque and ping pong to the unforgettable team travels to EWRL, being with you made my time truly memorable. Here is a long list of wonderful people I wanted to thank, in (pseudo) order of their appearance in my Scool adventure: Timothée, Amélie, Omar, Réda, Rémi, Marc, Dorian, Emilie, Odalric, Matheus, Riccardo, Julien, Hernan, Hassan, David, Riad, Hector, Alena, Mahdi, William, Cyrille, Tuan, Lucille, Aymen, Brahim, Ayoub, Sumit, Thomas, Yann, Adrienne, Guillaume, Udvas, Mohamed Yassine, Waris, Alex, Adrien, Sabrine, Juliette and Tanguy. A heartfelt thank you to Marc, Alena, and Hector for introducing me to numerous fascinating board games and for being my favourite dynamic trio.

I am deeply grateful to all my friends from home/school, prépa, Polytechnique, and Lille. A genuine thank you to Ali, Wassim, Amine, and Claire for your unwavering friendship and support during my thesis. I would also like to thank all my friends who hosted me during my travels: Yassine (Saint-Lô), Souhail (London), Youssef (Bruxelles), Ayoub and Priyank (New York). I also want to extend a special thanks to Junya for the wonderful research visit to Kyoto.

I am profoundly grateful to my parents, brother, and sister for their constant love and support. Your encouragement and belief in me have strengthened and inspired me. I love you so much.

Finally, I would like to thank my cat Kyubi, my guitar and my ps5.

Résumé

Les thèmes abordés dans cette thèse visent à caractériser les compromis entre confidentialité et utilité dans la prise de décision séquentielle dans l'incertain. Le principal cadre adopté pour définir la confidentialité est la protection différentielle, et le principal cadre d'utilité est le problème de bandit stochastique à plusieurs bras. Tout d'abord, nous proposons différentes définitions qui étendent la définition de confidentialité à l'environnement des bandits à plusieurs bras. Ensuite, nous quantifions la difficulté des bandits avec protection différentielle en prouvant des bornes inférieures sur la performance des algorithmes de bandits confidentiels. Ces bornes suggèrent l'existence de deux régimes de difficulté en fonction du budget de confidentialité et des distributions de récompenses. Nous proposons également un plan générique pour concevoir des versions confidentielles quasi-optimales des algorithmes de bandits. Nous instancions ce schéma directeur pour concevoir des versions confidentielles de différents algorithmes de bandits dans différents contextes: bandits à bras finis, linéaires et contextuels avec le regret comme mesure d'utilité, et bandits à bras finis avec la complexité d'échantillonnage comme mesure d'utilité. L'analyse théorique et expérimentale des algorithmes proposés valide aussi l'existence de deux régimes de difficulté en fonction du budget de confidentialité.

Dans la deuxième partie de cette thèse, nous passons des défenses de la confidentialité aux attaques. Plus précisément, nous étudions les attaques par inférence d'appartenance, où un adversaire cherche à savoir si un point cible a été inclus ou pas dans l'ensemble de données d'entrée d'un algorithme. Nous définissons la fuite d'information sur un point comme l'avantage de l'adversaire *optimal* essayant de déduire l'appartenance de ce point. Nous quantifions ensuite cette fuite d'information pour la moyenne empirique et d'autes variantes en termes de la distance de Mahalanobis entre le point cible et la distribution génératrice des données. Notre analyse asymptotique repose sur une nouvelle technique de preuve qui combine une expansion de Edgeworth du test de vraisemblance et un théorème central limite de Lindeberg-Feller. Notre analyse montre que le test de vraisemblance pour la moyenne empirique est une attaque par produit scalaire mais *corrigée* pour la géométrie des données en utilisant l'inverse de la matrice de covariance. Enfin, comme conséquences de notre analyse, nous proposons un nouveau score de covariance et une nouvelle stratégie de sélection des points cible pour l'audit des algorithmes de descente de gradient dans le cadre de l'apprentissage fédéré en white-box.

Abstract

The topics addressed in this thesis aim to characterise the privacy-utility trade-offs in sequential decision-making under uncertainty. The main privacy framework adopted is Differential Privacy (DP), and the main setting for studying utility is the stochastic Multi-Armed Bandit (MAB) problem. First, we propose different definitions that extend DP to the setting of multi-armed bandits. Then, we quantify the hardness of private bandits by proving lower bounds on the performance of bandit algorithms verifying the DP constraint. These bounds suggest the existence of two hardness regimes depending on the privacy budget and the reward distributions. We further propose a generic blueprint to design near-optimal DP extensions of bandit algorithms. We instantiate the blueprint to design DP versions of different bandit algorithms under different settings: finite-armed, linear and contextual bandits under regret as a utility measure, and finite-armed bandits under sample complexity of identifying the optimal arm as a utility measure. The theoretical and experimental analysis of the proposed algorithms furthermore validates the existence of two hardness regimes depending on the privacy budget.

In the second part of this thesis, we shift the view from privacy defences to attacks. Specifically, we study *fixed-target* Membership Inference (MI) attacks, where an adversary aims to infer whether a *fixed* target point was included or not in the input dataset of an algorithm. We define the *target-dependent leakage* of a datapoint as the advantage of the *optimal* adversary trying to infer the membership of that datapoint. Then, we quantify both the target-dependent leakage and the trade-off functions for the empirical mean and variants of interest in terms of the Mahalanobis distance between the target point and the data-generating distribution. Our asymptotic analysis builds on a novel proof technique that combines an Edgeworth expansion of the Likelihood Ratio (LR) test and a Lindeberg-Feller central limit theorem. Our analysis shows that the LR test for the empirical mean is a scalar product attack but *corrected* for the geometry of the data using the inverse of the covariance matrix. Finally, as by-products of our analysis, we propose a new covariance score and a new canary selection strategy for auditing gradient descent algorithms in the white-box federated learning setting.

Keywords: Differential Privacy, Multi-armed Bandits, Regret Analysis, Best-arm Identification, Membership Inference, Privacy auditing

Table of Contents

Notation

Acronyms and Abbreviations

a.s.	almost surely
e.g.	exempli gratia, means "for example"
i.e.	id est, means "that is"
i.i.d.	independent and identically distributed
l.h.s.	left hand side
r.h.s.	right hand side
s.t.	such that
w.r.t.	with respect to
BAI	Best Arm Identification
CDF	Cumulative Distribution Function
DP	Differential Privacy
FC-BAI	Fixed Confidence Best Arm Identification
FL	Federated Learning
FTL	Follow-The-Leader algorithm
LR	Likelihood Ratio
MAB	Multi-Armed Bandits
MI	Membership Inference
UCB	Upper Confidence Bound

General Notation

$\mathfrak{B}$	Borel set
$[K]$	Set of integers $\{1, \cdots, K\}$
$X \sim \nu$	The random variable X following a distribution ν

$\Pr(E)$	Probability of an event E
$\mathbb{E}[X]$	Expectation of a random variable X
$\mathbb{E}_\nu$	Expectation under distribution ν
$\mathbb{1}(E)$	Indicator function of an event E
$\bar{X}$	Complement of a set X
o, $\mathcal{O}$, Ω and Θ	Landau's notation
$\tilde{o}$, $\tilde{\mathcal{O}}$, $\tilde{\Omega}$ and $\tilde{\Theta}$	Landau's notation up to polylogarithmic terms
ζ	Riemann ζ function, $\zeta(s) := \sum_{n=1}^{+\infty} n^{-s}$ for all $s > 1$
$\langle x, y \rangle$	Cartesian product between vectors, $\langle x, y \rangle = \sum_{i \in [d]} x_i y_i$
$\|x\|_2$	ℓ_2-norm, $\|x\|_2 = \sqrt{\langle x, x \rangle}$
$\|x\|_1$	ℓ_1-norm, $\|x\|_1 = \sum_{i \in [d]} \|x_i\|$
$\|x\|_M$	M-norm, $\|x\|_M = \left\|M^{1/2} x\right\|_2$ for M symetric positive definite matrix
$\|x\|_\infty$	ℓ_∞-norm, $\|x\|_\infty = \max_{i \in [d]} \|x_i\|$
$\{e_i\}_{i \in [d]}$	Canonical basis of $\mathbb{R}^d$, $e_i = (\mathbb{1}(j = i))_{j \in [d]}$
$\mathrm{diag}(x) \in \mathbb{R}^{d \times d}$	Diagonal matrix for a vector $x \in \mathbb{R}^d$
$\mathrm{Span}(\mathcal{A})$	Span of a set of vectors $\mathcal{A}$
$I_d \in \mathbb{R}^{d \times d}$	Identity matrix
Σ_K	$(K-1)$-dimensional simplex, $\Sigma_K := \left\{ w \in \mathbb{R}_+^K \mid w \geq 0,\ \sum_{i \in [K]} w_i = 1 \right\}$
KL	Kullback-Leibler (KL) divergence
kl	KL between Bernoulli distributions
TV	Total variation distance
D_α	Rényi divergence of order α
D_f	f-divergence
$\mathrm{Lap}(b)$	Laplace distribution centred at 0 with scale b
$\mathcal{N}(\mu, C)$	Gaussian distribution with mean μ and covariance matrix C
$\mathrm{Bernoulli}(p)$	Bernoulli distribution with parameter p
Φ	Cumulative Distribution Function (CDF) of the standard normal distribution
log	Natural logarithm function

Differential Privacy

$\mathcal{M}$	Randomised mechanism
D	Input dataset of the mechanism $\mathcal{M}$, $D = \{x_1, \dots, x_n\} \in \mathcal{X}^n$
$\mathcal{X}$	Input universe
$\mathcal{O}$	Output universe
$\mathcal{M}_D$	Output distribution
$\mathcal{M}_D(E)$	Probability of observing output event E given input dataset D to mechanism $\mathcal{M}$
d_{Ham}	Hamming distance
$D \sim D'$	Neighbouring datasets D and D'
$Z_{D,D'}^{\mathcal{M}}$	Privacy loss random variable between $\mathcal{M}_D$ and $\mathcal{M}_{D'}$
$\varepsilon, \delta, \rho$	Privacy budgets

Multi-Armed Bandits

$K \in \mathbb{N}^\star$	Number of arms
$T \in \mathbb{N}^\star$	Time horizon
$a \in [K]$	Arm or action
$a^\star \in [K]$	Optimal arm
$t \in [T]$	Step of the interaction
$a_t \in [K]$	Arm played at step t
$r_t \in \mathbb{R}$	Reward observed at step t
H_t	History of the interaction, until step t included
$\nu = (P_a : a \in \mathcal{A})$	Bandit instance or environment
$\pi = (\pi_t)_{t=1}^T$	Policy
$\mu_a \in \mathbb{R}$	Mean of rewards of arm a, $\mu_a = \mathbb{E}_{X \sim P_a}[X]$
$\mathrm{Reg}_T(\pi, \nu)$	Regret of policy π on bandit instance ν
Δ_a	Sub-optimality gap of arm a, $\Delta_a = \mu_{a^\star} - \mu_a$
π^{BAI}	FC-BAI strategy
τ	Stopping time
$N_a(T)$	Number of times the arm a is played till T, $N_a(T) = \sum_{t=1}^T \mathbb{1}(a_t = a)$
$\hat{a}$	Final recommended arm

$\top$	Halting action
δ	Confidence level
c_t	Context at step t

Membership Inference Games

$\mathcal{M}$	Randomised mechanism
$\mathcal{Z}$	Input universe
D	Unknown input dataset of the mechanism $\mathcal{M}$, $D = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$
$z^\star \in \mathcal{Z}$	Target datapoint
n	Number of input samples
d	Dimension of input samples
$\mathcal{O}$	Output universe
$o \in \mathcal{O}$	Output of the mechanism $\mathcal{M}$
b	Secret membership bit, $b = \mathbb{1}(z^\star \in D)$
$\mathcal{A}$	Adversary, $\mathcal{A} : (z^\star, o) \to \{0, 1\}$
$\mathcal{A}_{z^\star}$	Target-dependent adversary, $\mathcal{A}_{z^\star} : o \to \{0, 1\}$
Acc	Accuracy of an adversary
Adv	Advantage of an adversary
α	Type I error of an adversary
β	Type II error of an adversary
Pow	Power of an adversary, or also Trade-off function
ℓ	Likelihood ratio score

Chapter 1

Introduction

1.1 Context and Scope

Imagine you are a researcher in healthcare trying to find the most effective treatment between K candidate medicines. You decide to run a clinical trial where you sequentially allocate doses of the treatments to patients who agree to participate in your trial. At the end of the trial, you want to determine the identity of the best medicine with some statistical guarantees. As a researcher, you would also like to publish the results of your study to the scientific community. The setting is summarised in Figure 1.1. However, you fear publishing insightful statistics about your trial may compromise the participants' privacy. Thus, you wonder:

(a) Is it possible to design a clinical trial where the participants' privacy is "guaranteed"?

(b) If yes, what is the cost of guaranteeing this constraint on the statistical utility of the trial?

(c) Is it even possible to infer private information about participants by only looking at published aggregated statistics?

Privacy regulations, such as the General Data Protection Regulation (GDPR) [VVdB17] in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) [Ann03] in the United States, play a significant role in healthcare research. These frameworks impose strict controls on collecting, storing, and sharing personal health information (PHI). Cryptographic tools like functional encryption [BSW11] or digital signatures [Kat10] can enhance data security. However, these tools only prevent information leakage beyond the outputs of the computed functions of interest. In a clinical trial, these outputs could be participants' aggregate statistics or the sequence of recommended doses. If not handled with care, such outputs could potentially reveal personal information about the patients in the trial.

Differential Privacy (DP) [DR14a] is considered the gold standard for privacy-preserving data analysis. It effectively solves the challenge of gaining valuable insights about a population

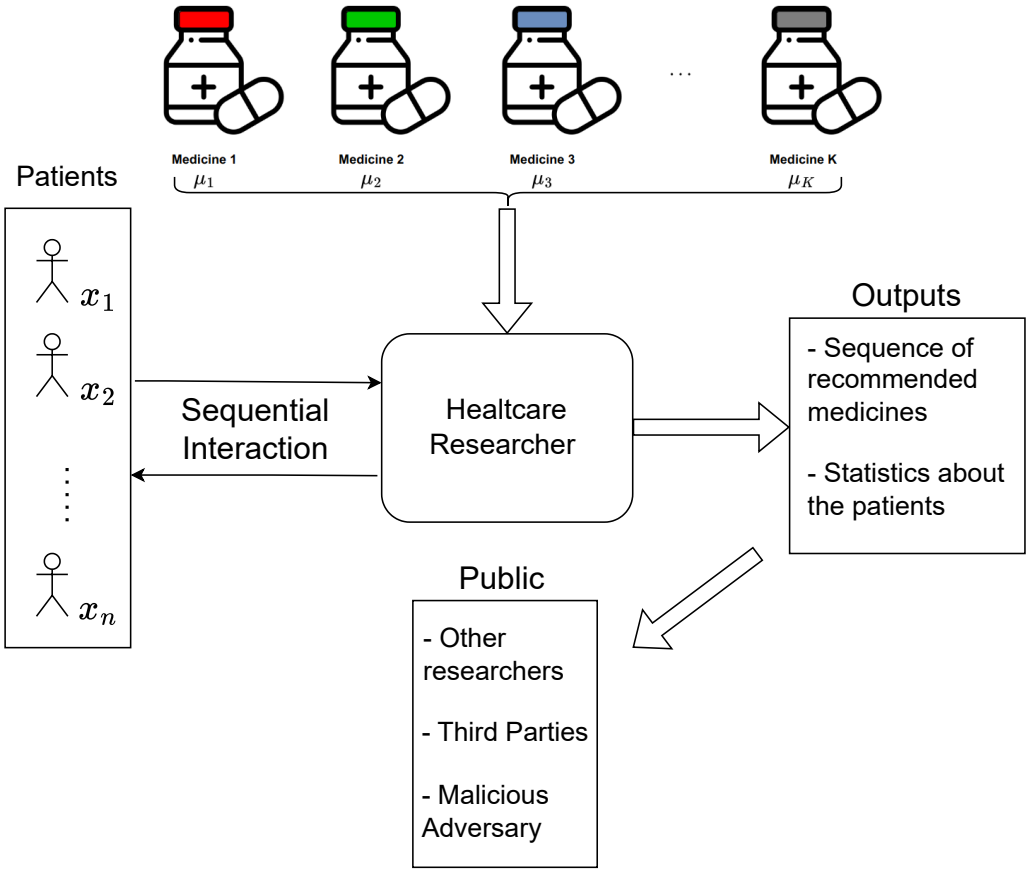


Figure 1.1 – The healthcare research wants to determine the most efficient medicine by designing a sequential interaction with patients. The researcher publishes the results of the interaction, *i.e.* the sequence of recommended medicines and other statistics about the patients to the public. The public is composed of other researchers but may also contain malicious adversaries trying to infer private information about the patients.

without compromising the privacy of individuals. It ensures that the same conclusions can be drawn regardless of whether an individual chooses to participate in the dataset. This demonstrates a commitment from the data curator (the healthcare researcher in our example) to the participants in a study, assuring them that their involvement will not lead to negative consequences from their data being used. Specifically, DP guarantees that any sequence of algorithm outputs is "essentially" equally likely to occur, regardless of the presence or absence of any individual. The probabilities are taken over random choices made by the algorithm, and "essentially" is captured by closeness parameters that we call privacy budgets. As these parameters become smaller, the privacy guarantee strengthens while typically the utility deteriorates. Since DP is a constraint on the class of algorithms, different algorithms may achieve the same DP constraint for a given task, but some will have better utility than others.

Privacy-preserving data analysis aims to design algorithms that verify the DP constraint while achieving the maximum utility possible over the class of DP algorithms. DP is an excellent framework for discussing patient privacy in clinical trials. Next, we introduce multi-armed bandits as the framework for measuring utility.

Motivated by clinical trial design, William R. Thompson introduced the problem of multi-armed bandits [Tho33a], or just bandits for short. Bandits [LS20] are a simple model for decision-making under uncertainty. Thompson's motivation for introducing and studying bandits was to design clinical trials that adapt treatment allocations on the fly as the medicines appear more or less effective. Specifically, in bandits, a learner sequentially interacts with an environment ν, which is a set of unknown distributions (or arms or actions), *i.e.* $\nu = \{P_a\}_{a=1}^{K}$. The K arms are the K candidate medicines in clinical trials. At each step t of the interaction, the learner chooses an arm a_t from $\{1, \ldots, K\}$ and the environment reveals a reward r_t from the distribution P_{a_t}. In a clinical trial, a new patient arrives at each step t of the protocol, and the researcher recommends one of the K medicines and observes the patient's reaction. It is possible to define different reward functions for a clinical trial. For simplicity, let us consider the binary reward model, where $r_t = 1$ if the patient at step t is cured, and 0 otherwise. The main challenge in bandits is to utilise the history of interactions $H_{t-1} \triangleq (a_1, r_1, \ldots, a_{t-1}, r_{t-1})$ to recommend a "good" allocation at step t. A "good" choice to recommend at step t depends on the goal of the learner, which can be of two types: (a) to maximise the reward accumulated over time, *i.e.* cure the maximum number of patients during the clinical trial, or (b) to find the reward distribution (or arm) with the highest expected reward, *i.e.* find the most effective medicine. The first problem is called the *regret minimisation* problem [ACBF02], while the second one is called the *Best Arm Identification* (*BAI*) problem [KCG16].

Under both regret minimisation and BAI settings, bandits are increasingly deployed for different applications beyond clinical trials. These applications include recommendation systems [SWS+22], online advertisement [CLK+14], crowd-sourcing [ZCL14], user studies [LEHT22], hyper-parameter tuning [LJD+17], communication networks [LPJ22], and pandemic mitigation [LVR+19] to name a few. All of these applications often involve users' sensitive and personal data, which raises serious data privacy concerns [TBD+16].

The first part of this thesis investigates the privacy-utility trade-offs for privacy-preserving algorithms in bandits. We adhere to Differential Privacy (DP) as the privacy framework, and multi-armed bandits under both regret minimisation and Best Arm Identification (BAI) to measure the utility. We aim to address two main questions:

Q1.1 *What is the fundamental hardness of differentially private bandits expressed in terms of lower bounds on the utility?*

Q1.2 *How to design an algorithmic framework that converts near-optimal bandit algorithms into near-optimal bandit algorithms satisfying DP?*

In the second part, we investigate the counterfactual question of whether it is even possible to infer private information about individuals by only examining the outputs of some aggregated statistics. Therefore, in the second part, we shift from the perspective of a health researcher aiming to safeguard the privacy of study participants, to the viewpoint of an adversary analysing the researcher's published statistics in an attempt to deduce private health-related information about the participants. This viewpoint shift is interesting for different reasons. Designing strategies for the adversary to execute is crucial in establishing fundamental limits of what is possible to protect. Thus, studying such algorithms helps formulate achievable privacy goals. For example, the adversary may aim to reconstruct 99.9% of the dataset's participants by only looking at the published statistics. Most people would agree that the success of this type of goal would constitute a colossal privacy outbreak. By the same logic, any "decent" privacy protection technique should be able to defend against such failures. This type of adversarial goal is called reconstruction [DN03], and following their study, the concept of differential privacy emerged [DMNS06].

The adversary may lower the threshold for defining "a win" and set a simpler goal: Is it possible to determine whether a specific individual's data was present or not in a given dataset, by only looking at published statistics computed on the given dataset? This type of adversarial goal is called tracing [HSR+08, DSSU17] or Membership Inference [SSSS17] in the machine learning literature, and the individual in question is called the target. Tracing is considered a privacy breach since the membership to a dataset itself can leak private information. For example, suppose that a medical dataset only contains the data of a control group with a specific disease. Suppose an adversary can prove a target individual's membership to this dataset by only looking at aggregate statistics computed on the private dataset. In that case, the adversary can conclude that the target suffers from this disease and thus violates their privacy. Tracing is a weaker privacy failure than reconstruction, making it an essential tool to provide tight lower bounds on the utility of differentially private algorithms [BUV14].

Another exciting application of tracing is privacy auditing. With the success of Differential Privacy (DP), research results have proliferated, enabling the construction of intricate data pipelines that adhere to DP. Notably, DP is now used in production by the US Census Bureau [Abo18], Google [EPK14], Apple [TVV+17] and Microsoft [DKY17], among others. Since DP is a theoretical constraint, a DP algorithm comes with a mathematical proof that yields an upper bound on the privacy budgets, and an implementation that runs in production. However, proofs may have mistakes, and implementations may have bugs. This raises the question: Is it possible to empirically certify the privacy of an algorithm? Specifically, if a company or a data curator claims their algorithm satisfies the DP constraint with some privacy budget, is it possible to empirically validate this budget? This problem is called privacy auditing. Typically, a privacy audit runs a tracing attack, then translates the adversary's Type-I/Type-II errors into a lower bound on the privacy budget. These algorithms apply different heuristics to find the

most leaking data point (aka *canary*) used as a target point for tracing. Thus, designing optimal tracing attacks can lead to optimal privacy auditing procedures.

The second part of the thesis revisits the fundamental questions of tracing in relation to privacy auditing. The study of statistical efficiency and design of tracing attacks has begun with summary statistics on genomic data [HSR+08, SOJH09, DSS+15]. [HSR+08] and [SOJH09] studied the first attack to detect an individual in the exact empirical mean statistic computed on a dataset generated by Bernoulli distributions, using Likelihood Ratio (LR) tests. [DSS+15] improves the attack by assuming access to noisy statistics and one reference sample, and develops a *scalar product attack* to understand the correlation of a target point with the marginals of noisy statistics. *However, these tracing attacks are always studied in a threat model where the target point attacked is sampled randomly, either from the input dataset, or from a data generating distribution* [HSR+08, SOJH09, DSS+15]. This means that the metrics of the attack analysed, *i.e.* the *power* or *trade-off functions between Type-I/Type-II errors*, are "averaged" over the target point's sampling. This obfuscates the target-dependent hardness of tracing attacks, which is essential to understand due to the worst-case nature of privacy [SU20]. In the Machine Learning (ML) community, the design of tracing attacks began with [SSSS17] under the name of Membership Inference (MI) attacks. Similar to tracing, the threat model of MI attacks, originally proposed in [YGFJ18], also averages out the effect of the target point. Thus, the accuracy and ROC curves generally portrayed in the MI attack literature [SSSS17, YGFJ18, CCN+22] are averaged over the training dataset, and thus hide the target-dependent hardness of MI games. We aim to fill this gap in the literature on tracing attacks. Hence, we ask the question

Q2.1 *Why are some points statistically harder to trace than others, and how can we quantify this hardness?*

Though DP bounds the worst-case privacy leakage on any attack, it is not always evident how these guarantees bound the power of specific privacy attacks, such as tracing [ZYS+20, HOT+23]. Thus, we ask:

Q2.2 *Can we quantify the target-dependent effect of privacy-preserving mechanisms?*

Finally, privacy audits typically trace specific targets that leak the most, *i.e.* canaries. Thus, understanding the target-dependent hardness of tracing can directly provide a recipe for designing canaries. Thus, we ask

Q2.3 *How can we leverage the target-dependent hardness of tracing to design an optimal canary selection strategy?*

1.2 Outline and Contributions

This thesis is organised in two parts. The first part answers the questions **Q1.1** and **Q1.2**, while the second part answers the questions **Q2.1**, **Q2.2** and **Q2.3**.

1.2.1 Outline

After this first introductory chapter, Chapter 2 introduces the necessary background for the rest of the thesis. The background includes basic facts about Differential Privacy (DP) and Multi-armed Bandits (MAB) used in Part I. It also introduces the Membership Inference (MI) game, the privacy auditing problem and the asymptotic statistics background used in Part II.

In Part I, we study the cost of Differential Privacy on the utility in bandits. Part I is composed of tree chapters: Chapter 3, Chapter 4 and Chapter 5. In Chapter 3, we extend Differential Privacy definitions to the bandit setting. In Chapter 4, we provide regret and sample complexity lower bounds of bandits satisfying DP. Finally, in Chapter 5, we design near-optimal private bandit algorithms.

In Part II, we study Membership Inference (MI) games and their effect on privacy auditing. Part II is composed of two chapters: Chapter 6 and Chapter 7. In Chapter 6, we introduce the target-dependent MI game and study the performance of its optimal adversaries for the empirical mean mechanism and variants of interest. In Chapter 7, we use the theoretical insight of Chapter 6 to provide a new covariance score and a new canary selection strategy that improves privacy auditing of gradient descent algorithms in the white-box federated learning setting.

1.2.2 Contributions

Answering the questions **Q1.1** and **Q1.2** led to the following contributions:

1. *Privacy Definitions for Bandits.* In Chapter 3, we extend the DP definition to the bandit setting. First, we discuss three main challenges for adapting DP to bandits: the online and interactive nature of the bandit interaction and partial information. We propose four extensions of DP to bandits: Table DP (Definition 3.2), View DP (Definition 3.3), Interactive DP (Definition 3.5) and DP in the adaptive continual release model (Definition 3.9). Each definition deals with the three challenges differently: Table DP and View DP are non-interactive definitions where the inputs of the policy are fixed in advance, and differ in the input considered due to partial information. Interactive DP and DP in the adaptive continual release model tackle the interactive nature of bandits by considering an adversarial analyst who adaptively chooses the inputs. Formalising and linking these

definitions is a crucial step that was missing in the private bandits literature. Our first contribution is to fill this gap.

2. *Hardness of Preserving Privacy in Bandits as Lower Bounds.* In Chapter 4, we derive lower bounds on the regret and sample complexity of bandits with DP. To prove the lower bounds, we develop a generic proof technique that relates lower bounds to a transport problem using coupling techniques. Then, we instantiate the proof to provide different flavours of lower bounds for regret and sample complexity, under different settings: a minimax (Theorem 4.15) and a problem-dependent (Theorem 4.16) regret lower bound for finite-armed bandits under ε-View DP, a minimax (Theorem 4.17) and a problem-dependent (Theorem 4.18) regret lower bound for linear bandits under ε-View DP, a minimax regret lower bounds for finite-armed (Theorem 4.20) and linear bandits (Theorem 4.21) under ρ-Interactive zCDP, and a sample complexity lower bound (Theorem 4.24) for finite-armed bandits under ε-View DP. All the lower bounds show the existence of two privacy regimes depending on the privacy budget and the reward distributions. In the low-privacy regime (large budget), bandits with DP are not harder than bandits without privacy. In the high-privacy regime (small budget), privacy has an additional cost on the utility of bandit algorithms. We characterise the additional hardness of DP in the high privacy regime using novel information-theoretic quantities based on the Total variation, such as the $\mathrm{TV}_{\mathrm{inf}}$ (Theorem 4.16) and $T^{\star}_{\mathrm{TV}}$ (Theorem 4.24). Finally, the change of regimes can be shown to happen when the privacy budget is of the same order as the sub-optimality gaps of arm rewards.

3. *Algorithm Design.* In Chapter 5, we propose a generic blueprint to design near-optimal DP extensions of bandit algorithms. The main intuition behind the blueprint is that, by running the bandit algorithm on non-overlapping sequences of input rewards, less noise should be added to satisfy DP, thanks to the parallel composition property of DP (Lemma 2.10). We instantiate the blueprint to design DP versions of different bandit algorithms under different settings: finite-armed (AdaP-UCB, AdaP-KLUCB, and AdaC-UCB), linear (AdaC-GOPE) and contextual bandits (AdaC-OFUL) for regret, and finite-armed bandits (AdaP-TT and AdaP-TT*) for sample complexity. All these algorithms run in adaptive phases and add calibrated noise to achieve Interactive DP.

4. *Regret and Sample Complexity Analysis.* In Chapter 5, we analyse the utility of each proposed algorithm, and compare the theoretical performance to the lower bounds of Chapter 4. We show that the upper bounds on the performance of our proposed algorithms match the provided lower bounds up to constants in the problem-dependent bounds, and up to logarithmic terms in the horizon T in the minimax bounds. In general, we show that Interactive DP can be preserved almost for free in terms of the minimax regrets. Specifically, for a fixed privacy budget b ($b = \varepsilon$ for pure DP and $b = \sqrt{\rho}$ for zCDP) and asymptotically

in the horizon T, the cost of Interactive DP in the regret of these algorithms exhibits an additional $\tilde{\mathcal{O}}(\log(T)/b)$, which is significantly lower than the privacy oblivious regret, *i.e.* $\tilde{\mathcal{O}}(\sqrt{T})$. The proposed algorithms' theoretical and experimental analysis further validate the existence of two hardness regimes depending on the privacy budget b.

Answering the questions **Q2.1**, **Q2.2** and **Q2.3** led to the following contributions:

1. *Defining the target-dependent leakage.* We instantiate a *fixed-target MI game* (Algorithm 12). We define the leakage of a target point as the advantage of the optimal attacker, *i.e.* the LR attacker, trying to identify this fixed target point. We also characterise the target-dependent leakage in terms of a Total Variation distance (Equation (6.1)).

2. *Explaining the target-dependent leakage using the Mahalanobis distance.* We investigate the fixed-target MI game for the empirical mean. First, we find the asymptotic distributions of the LR scores if the target datum is included in the empirical mean and also if not. Then, we recover the optimal advantage (Equation (6.2)) and trade-off functions (Equation (6.3)). This shows that the target-dependent hardness of MI games depends on the Mahalanobis distance between the target point $z^\star$ and the true data-generating distribution (Table 6.1). This insight is used to propose Algorithm 14 for optimally choosing canaries in auditing gradient descent algorithms in the white-box Federated Learning setting. Our experiments show that the Mahalanobis distance explains the target-dependent hardness of MI games on synthetic and real datasets.

3. *A new covariance attack.* We analyse the LR score for the empirical mean asymptotically. Our novel proof technique that combines an Edgeworth expansion with Lindeberg-Feller central limit theorem shows that the *LR score is asymptotically a scalar product attack, corrected by the inverse of the covariance matrix* (Equation (6.4)). This enables us to provide a novel score for attacks and improves the scalar product by correcting it for the geometry of the data. We use this "covariance score" to propose a novel white-box attack (Algorithm 13) that experimentally outperforms the scalar product attack.

4. *Tight quantification of the effects of noise addition, sub-sampling, and misspecified targets on leakage.* We further study the impact of privacy-preserving mechanisms, such as the Gaussian mechanism [DR14b] and sub-sampling, on the target-dependent leakage. As shown in Table 6.1, both reduce the leakage scores and, thus, the powers of the optimal attacks. We numerically validate them. Finally, we quantify how target misspecification affects the leakage, and how it depends on the similarity between the real and misspecified targets.

1.3 Lisf of Publications

Here is a list of publications that I co-authored during my PhD.

Publications in international conferences with proceedings

- **Achraf Azize** and Debabrota Basu. *When Privacy Meets Partial Information: A Refined Analysis of Differentially Private Bandits*. In Advances in Neural Information Processing Systems (NeurIPS), 2022 [AB22].

  This paper studies the complexity of regret in bandits under pure DP. Elements of this paper have been adapted in the regret lower bounds under pure DP in Section 4.3.1 of Chapter 4, the theoretical analysis of regret algorithms in Section 5.2 and their experimental analysis in Section 5.5 of Chapter 5.

- **Achraf Azize**, Marc Jourdan, Aymen Al Marjani and Debabrota Basu. *On the Complexity of Differentially Private Best-Arm Identification with Fixed Confidence*. In Advances in Neural Information Processing Systems (NeurIPS), 2023 [AJMB23].

  This paper studies the complexity of sample complexity in bandits under pure DP. Elements of this paper have been adapted in the sample complexity lower bounds under pure DP in Section 4.4 of Chapter 4, the theoretical analysis of best-arm identification algorithms in Section 5.4 and their experimental analysis in Section 5.5 of Chapter 5.

- **Achraf Azize** and Debabrota Basu. *Concentrated Differential Privacy for Bandits*. In IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2024 [AB24a].

  This paper studies the complexity of regret in bandits under concentrated DP. It introduces the privacy definitions of Chapter 3, the coupling techniques in Section 4.2 of Chapter, the regret lower bound under zCDP in Section 4.3.1 of Chapter 4, the theoretical analysis of regret algorithms under zCDP in Section 5.3 and their experimental analysis in Section 5.5 of Chapter 5.

- **Achraf Azize** and Debabrota Basu. *Open Problem: What is the Complexity of Joint Differential Privacy in Linear Contextual Bandits?*. In Conference on Learning Theory (COLT), 2024 [AB24c].

  We discuss this open problem in Chapter 9.

Presentations in international workshops

- **Achraf Azize** and Debabrota Basu. *Quantifying the target-dependent Membership Leakage*. In Theory and Practice of Differential Privacy (TPDP), 2024 [AB24b].

  Part II is based on this paper.

- **Achraf Azize** and Debabrota Basu. *Rényi Differentially Private Bandits*. In The Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI), 2023.

  This is a first version of [AB24a], where we study Rényi DP for bandits.

Also, a primary version of [AB22] has been presented in the European Workshop on Reinforcement Learning, (EWRL) 2022. A primary version of both [AJMB23] and [AB24a] have been presented in the European Workshop on Reinforcement Learning, (EWRL) 2023.

Preprints under review

- **Achraf Azize** and Debabrota Basu. *Quantifying the target-dependent Membership Leakage*. Under review at an international conference with proceedings.

- **Achraf Azize**, Marc Jourdan, Aymen Al Marjani and Debabrota Basu. *Differentially Private Best-Arm Identification* [AJMB24]. Under review at the Journal of Machine Learning Research (JMLR).

  This is a journal version of [AJMB23]. In this version, we improve on our previous results by proposing a new algorithm AdaP-TT* that is inspired by the lower bound. We analyse the sample complexity of AdaP-TT* and show that the upper bound on its sample complexity matches the lower bound up to multiplicative constant for **all** bandit instances. This is an improvement over AdaP-TT that only matches the lower bound for instances where the maximum and minimum sub-optimality gaps have similar magnitudes. Our experimental results also validate this improvement achieved by AdaP-TT*. In the journal version, we also study the complexity of the BAI problem under the local trust model. We provide a new sample complexity lower bound for bandits under local DP and an algorithm with a matching upper bound.

Chapter 2

Background

This chapter provides an overview of several key concepts in Differential Privacy (DP), Multi-Armed Bandits (MAB), Membership Inference (MI) games and Asymptotic Statistics. The section on Differential Privacy is based on the book [DR14a]. The section on Multi-Armed Bandits is a summary of results and definitions from the book [LS20]. The section on Asymptotic Statistics reports basic definitions and properties from the book [VdV00], while the version of the Edgeworth expansion reported comes from the book [Pet12]. Finally, the section on Membership Inference games is an adaptation and formalisation of results from different papers in the literature [SOJH09, DSS+15, YGFJ18, EMRS19, YMM+22, CCN+22, HOT+23, DRS19, JUO20, MSS22, NHS+23, SNJ23, AKO+23].

Contents

2.1 Differential Privacy (DP)

In this section, we formalise the definition of Differential Privacy (DP). We discuss different interpretations of this definition and the properties that justify its success and wide adoption. We also present two fundamental mechanisms to achieve DP: the Laplace and Gaussian mechanisms. Finally, we present the binary tree mechanism [DNPR10a, CSS11] for DP under continual observation.

2.1.1 The language of Differential Privacy

First, we define the main object of interest, a mechanism, its input and outputs.

Definition 2.1 (Mechanism, its input and output)**.** *A mechanism $\mathcal{M}$ is a randomised algorithm. $\mathcal{M}$ takes as input a dataset $D \triangleq \{x_1, \dots, x_n\} \in \mathcal{X}^n$, which is a collection of n data points from the input universe $\mathcal{X}$. $\mathcal{M}$ outputs a distribution $\mathcal{M}_D \in \mathcal{P}(\mathcal{O})$, where $\mathcal{P}(\mathcal{O})$ is the set of distributions over the probability space $(\mathcal{O}, \mathcal{F})$, and $\mathcal{O}$ is the output space. The probability space is over the coin flips of the mechanism $\mathcal{M}$. Given some measurable event A in $(\mathcal{O}, \mathcal{F})$, we note $\mathcal{M}(A|D) \triangleq \mathcal{M}_D(A)$ the probability of observing the event A given that the input of the mechanism is D.*

Second, we define the Hamming distance between two datasets, which characterises the neighbouring relation.

Definition 2.2 (The Hamming distance, and neighbouring relation)**.** *Given two datasets $D \triangleq \{x_1, \dots, x_n\}$ and $D' \triangleq \{x'_1, \dots, x'_n\}$ in $\mathcal{X}^n$, let $d_{\text{Ham}}(D, D') \triangleq \sum_{i=1}^{n} \mathbb{1}\left(D_i \neq D'_i\right)$ denote the Hamming distance between D and D',* i.e. *the number of different records between D and D'. We say that D and D' are neighbouring datasets, that we note $D \sim D'$, if and only if $d_{\text{Ham}}(D, D') \leq 1$,* i.e. *$D$ and D' differ by at most one record.*

2.1.2 Formalising Differential Privacy

We are now ready to formally define Differential Privacy (DP).

Definition 2.3 ((ε, δ)-DP [DR14a] and ρ-zCDP [BS16])**.** *A mechanism $\mathcal{M}$ satisfies*

- *(ε, δ)-DP for a given $\varepsilon \geq 0$ and $\delta \in [0, 1)$, if*

$$\sup_{A \in \mathcal{F}, D \sim D'} \mathcal{M}_D(A) - e^{\varepsilon} \mathcal{M}_{D'}(A) \leq \delta. \tag{2.1}$$

- *ρ-zCDP (zero Concentrated DP) for a given $\rho \geq 0$ if, for all $\alpha > 1$*

$$\sup_{D \sim D'} D_{\alpha}(\mathcal{M}_D \| \mathcal{M}_{D'}) \leq \rho\alpha. \tag{2.2}$$

Here, $D_\alpha(P\|Q) \triangleq \frac{1}{\alpha-1}\log \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^\alpha\right]$ denotes the Rényi divergence of order α between P and Q. We define ε-pure DP to be $(\varepsilon, 0)$-DP, or simply ε-DP.

DP is a worst-case constraint on the class of randomised mechanisms. A mechanism $\mathcal{M}$ satisfies DP if the mechanism behaves "similarly" on *all* neighbouring datasets D and D', even for very "unlikely" realisations of the mechanism $\mathcal{M}$.

By designing a mechanism $\mathcal{M}$ that satisfies the DP constraint, the curator honours the privacy "promise": whatever would have happened to any user due to their participation in a DP study, *i.e.* the world where $o \sim \mathcal{M}_D$, would likely have happened if they did not participate, *i.e.* the world where $o \sim \mathcal{M}_{D'}$, for $D' \sim D$. The worst-case nature of DP ensures that the promise is honoured, even if the user has a very unlikely data point (*e.g.* an outlier), and for any coalition of the other users. The effectiveness of the DP definition lies in its information-theoretic nature, in the sense that DP protects against any adversary with unlimited amounts of computational power and auxiliary information.

The hypothesis test interpretation. The adversary can formulate their attack as a binary hypothesis test. Specifically, based only on the observed output o, the adversary needs to determine whether

$$H_0\text{: The output was generated from the dataset } D$$
$$\text{vs}$$
$$H_1\text{: The output was generated from the neighbouring dataset } D'.$$

DP imposes a trade-off between the Type I and Type II errors of any adversary trying to conduct this hypothesis test. For a choice of a rejection region $S \subset \mathcal{O}$ *i.e.* the subset of outputs where the attacker rejects H_0, the type I error, also called the probability of False alarm, is

$$P_{\text{FA}}(D, D', \mathcal{M}, S) \triangleq \mathcal{M}_D(S),$$

and the type II error, also called the probability of missed detection is

$$P_{\text{MD}}(D, D', \mathcal{M}, S) \triangleq \mathcal{M}_{D'}(\bar{S}),$$

where $\bar{S}$ is the complement of the region S in $\mathcal{O}$.

The DP constraint on a mechanism $\mathcal{M}$ is equivalent to the following set of constraints on the probability of false alarm and missed detection

Theorem 2.4 (Hypothesis test formulation of Differential Privacy [KOV15])**.** *For any $\varepsilon > 0$ and any $\delta \in [0, 1]$, a mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if, for all neighbouring $D \sim D'$, and all rejection regions $S \subset \mathcal{O}$:*

$$P_{FA}(D, D', \mathcal{M}, S) + e^\varepsilon P_{MD}(D, D', \mathcal{M}, S) \geq 1 - \delta, \quad \textit{and,} \tag{2.3}$$

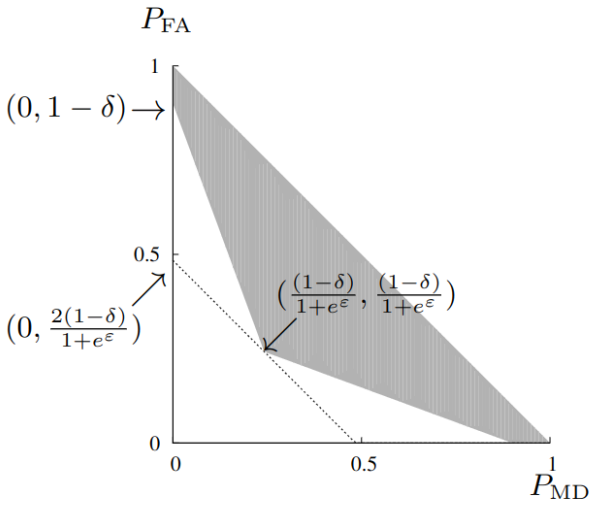


Figure 2.1 – Trade-off between Type I and Type II errors for (ε, δ)-DP mechanism. For simplicity, only the region where $P_{\text{FA}} + P_{\text{MD}} \leq 1$. The rest of the region is symmetric with respect to $P_{\text{FA}} + P_{\text{MD}} = 1$ [KOV15].

$$e^{\varepsilon} P_{FA}(D, D', \mathcal{M}, S) + P_{MD}(D, D', \mathcal{M}, S) \geq 1 - \delta \tag{2.4}$$

This shows that it is impossible for any adversary to get both Type I and Type II erros to be small when the mechanism verifies DP. This also provides an operational interpretation of the DP constraint, represented graphically in Figure 2.1.

The privacy loss interpretation. Definition 2.3 uses two notions of "similarity" between the output distributions on neighbouring datasets. To provide an intuition of the formulation of these similarities, we introduce the notion of the privacy loss random variable.

Definition 2.5 (Privacy Loss Random Variable)**.** *Let $\mathcal{M}$ be a randomised mechanism, and D and D' two datasets. We define the likelihood ratio $f(o) \triangleq \log\left(\frac{\mathcal{M}_D(o)}{\mathcal{M}_{D'}(o)}\right)$ for every output $o \in \mathcal{O}$. The privacy loss random variable between $\mathcal{M}_D$ and $\mathcal{M}_{D'}$ denoted by $Z^{\mathcal{M}}_{D,D'}$ is distributed according to $f(\mathcal{M}_D)$.*

The value of the privacy loss $Z^{\mathcal{M}}_{D,D'}$ represents how well $\mathcal{M}_D$ and $\mathcal{M}_{D'}$ are distinguishable. If $Z^{\mathcal{M}}_{D,D'} > 0$, then the observed output is more likely to have occurred under input D. The larger $Z^{\mathcal{M}}_{D,D'}$, the more likely the input is from D. Likewise, $Z^{\mathcal{M}}_{D,D'} < 0$ indicates the output is more likely under D'. If $Z^{\mathcal{M}}_{D,D'} = 0$, both the inputs D and D' explain the output equally well.

The closeness notions between $\mathcal{M}_D$ and $\mathcal{M}_{D'}$ used in Definition 2.3 could be expressed as bounds on the privacy loss $Z^{\mathcal{M}}_{D,D'}$. A mechanism is ε-pure DP if and only if $|Z^{\mathcal{M}}_{D,D'}| \leq \varepsilon$ almost surely on the randomness of the mechanism, for all $D \sim D'$. If the probability of the event $\{|Z^{\mathcal{M}}_{D,D'}| > \varepsilon\}$ is less than δ for every $D \sim D'$, then $\mathcal{M}$ is (ε, δ)-DP. The converse is also true, up to a small loss in parameters. That is why (ε, δ)-DP could be though of as "ε-DP with probability $1 - \delta$", and called approximate-DP.

The ρ-zCDP constraint bounds the moment generating function of $Z_{D,D'}^{\mathcal{M}}$ for every $D \sim D'$. Specifically, zCDP implies that the privacy loss random variable is subgaussian with a small mean, *i.e.* $Z_{D,D'}^{\mathcal{M}}$ is small with high probability, with larger deviations from zero becoming increasingly unlikely. The ρ-zCDP class of algorithms could be thought of as an intermediate class, between pure and approximate DP.

Proposition 2.6 (Relation between zCDP, pure and approximate DP [BS16])**.**

- *If $\mathcal{M}$ is ε-DP, then $\mathcal{M}$ is $(\frac{1}{2}\varepsilon^2)$-zCDP.*
- *If $\mathcal{M}$ is ρ-zCDP, then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP, for any $\delta > 0$.*

For the three definitions of DP, *i.e.* pure-DP, approximate-DP and zCDP, the "similarity" notions are controlled by parameters, *i.e.* ε, δ and ρ that we call the privacy budgets. We will think of ε and ρ as small real values in $[0, 1]$, while the failure probability δ should be "cryptographically" small, *i.e.* $\delta \ll 1/n$, where n is the size of the dataset.

2.1.3 Properties of Differential Privacy

In the following, we present three main properties of DP: post-processing, group privacy and compositions.

Proposition 2.7 (Post-processing (Proposition 2.1, [DR14a]))**.** *Let $\mathcal{M}$ be a mechanism and f be an arbitrary randomised function defined on $\mathcal{M}$'s output.*

- *If $\mathcal{M}$ is (ε, δ)-DP, then $f \circ \mathcal{M}$ is (ε, δ)-DP.*
- *If $\mathcal{M}$ is ρ-zCDP, then $f \circ \mathcal{M}$ is ρ-zCDP.*

The post-processing property ensures that any quantity that is constructed only from a private output is still private, with the same privacy budget. This property is a consequence of the data processing inequality.

Proposition 2.8 (Group Privacy)**.** *Let D and D' be two datasets in $\mathcal{X}^n$.*

- *If $\mathcal{M}$ is (ε, δ)-DP, then for any event $A \in \mathcal{F}$*

$$\mathcal{M}_D(A) \leq e^{\varepsilon d_{\mathrm{Ham}}(D,D')} \mathcal{M}_{D'}(A) + \delta d_{\mathrm{Ham}}(D, D') e^{d_{\mathrm{Ham}}(D,D')-1}. \tag{2.5}$$

- *If $\mathcal{M}$ is ρ-zCDP, then*

$$D_\alpha(\mathcal{M}_D \| \mathcal{M}_{D'}) \leq \rho \alpha d_{\mathrm{Ham}}(D, D')^2. \tag{2.6}$$

Group privacy translates the closeness of output distributions on neighbouring input datasets, to a closeness of output distributions on any two datasets D and D' that depends on

the Hamming distance $d_{\mathrm{Ham}}(D, D')$. This property will be the basis for proving lower bounds in Chapter 4.

Proposition 2.9 (Simple Composition). *Let $\mathcal{M}^1, \ldots, \mathcal{M}^k$ be k mechanisms. We define the mechanism*

$$\mathcal{G} : D \to \bigotimes_{i=1}^{k} \mathcal{M}_D^i$$

as the k composition of the mechanisms $\mathcal{M}^1, \ldots, \mathcal{M}^k$.

- *If each $\mathcal{M}^i$ is $(\varepsilon_i, \delta_i)$-DP, then $\mathcal{G}$ is $(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i)$-DP.*
- *If each $\mathcal{M}^i$ is ρ_i-zCDP, then $\mathcal{G}$ is $\sum_{i=1}^k \rho_i$-zCDP.*

Composition is a fundamental property of DP. Composition helps to analyse the privacy of sophisticated algorithms, by understanding the privacy of each building block, and summing directly the privacy budgets. Proposition 2.9 can be improved in two directions. (a) It is possible to show that the result is still true if the mechanisms are chosen adaptively, and that the mechanism at step i takes as auxiliary input the outputs of the last $i-1$ mechanisms. (b) Advanced composition theorems [KOV15] for (ε, δ)-DP improve the dependence on k the number of composed mechanisms. Specifically, if the same mechanism is composed k times, Proposition 2.9 concludes that the composed mechanism is $(k\varepsilon, k\delta)$-DP. Advanced composition [KOV15] shows that the k-fold adaptively composed mechanism is $(\varepsilon', \delta' + k\delta)$-DP for any δ' where $\varepsilon' \triangleq \sqrt{2k\log(1/\delta')}\varepsilon + k\varepsilon(e^\varepsilon - 1)$. Roughly speaking, advanced composition provides a $(\sqrt{k}\varepsilon, \delta)$-DP guarantee, improving by $\sqrt{k}$ the $(k\varepsilon, k\delta)$-DP guarantee of simple composition.

In addition to the classic composition theorems, we provide here an additional property of interest: parallel composition.

Lemma 2.10 (Parallel Composition). *Let $\mathcal{M}^1, \ldots, \mathcal{M}^k$ be k mechanisms, such that $k < n$, where n is the size of the input dataset. Let $t_1, \ldots t_k, t_{k+1}$ be indexes in $[1, n]$ such that $1 = t_1 < \cdots < t_k < t_{k+1} - 1 = n$.*
Let's define the following mechanism

$$\mathcal{G} : \{x_1, \ldots, x_n\} \to \bigotimes_{i=1}^{k} \mathcal{M}^i_{\{x_{t_i}, \ldots, x_{t_{i+1}-1}\}}$$

$\mathcal{G}$ is the mechanism that we get by applying each $\mathcal{M}^i$ to the i-th partition of the input dataset $\{x_1, \ldots, x_n\}$ according to the indexes $t_1 < \cdots < t_k < t_{k+1}$.

- *If each $\mathcal{M}^i$ is (ε, δ)-DP, then $\mathcal{G}$ is (ε, δ)-DP*
- *If each $\mathcal{M}^i$ is ρ_i-zCDP, then $\mathcal{G}$ is ρ-zCDP*

In parallel composition, the k mechanisms are applied to different "non-overlapping" parts of the input dataset. If each mechanism is DP, then the parallel composition of the k mechanisms is DP, *with the same privacy budget*. This property will be the basis for designing private bandit algorithms in Chapter 5.

2.1.4 Achieving Differential Privacy

Let us introduce two fundamental mechanisms to achieve DP. Both of these methods add calibrated independent noise to the output. The noise is either sampled from a Laplace or Gaussian distribution. The variance of the noise is calibrated to the sensitivity of the function to be made DP.

First, let us introduce the two noise distributions: Laplace and Gaussian distributions.

Definition 2.11 (The Laplace and Gaussian distributions)**.** *The Laplace distribution centred at 0 with scale b, denoted* $\mathrm{Lap}(b)$*, is the distribution with probability density function*

$$\mathrm{Lap}(x|b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right).$$

for any $x \in \mathbb{R}$*. The variance of this distribution is* $\sigma^2 = 2b^2$*.*

The Gaussian distribution on $\mathbb{R}^k$*, with mean* $\mu \in real^k$ *and covariance matrix* $\Sigma \in \mathbb{R}^{k\times k}$ *a positive definite matrix, denoted* $\mathcal{N}(\mu, \Sigma)$*, is the distribution with probability density function*

$$\mathcal{N}(x|\mu,\Sigma) \triangleq \frac{1}{(2\pi)^{k/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

for any $x \in \mathbb{R}^k$*.*

Now we introduce the ℓ_1 and ℓ_2 sensitivities of an algorithm.

Definition 2.12 (The ℓ_1 and ℓ_2 sensitivities)**.** *Let* $f : \mathcal{X} \to \mathbb{R}^k$ *be a deterministic algorithm. The* ℓ_1 *sensitivity of* f*, denoted* $s_1(f)$ *is*

$$s_1(f) \triangleq \max_{D\sim D'} \left\|f(D) - f(D')\right\|_1.$$

Similarly, the ℓ_2 *sensitivity of* f*, denoted* $s_2(f)$ *is*

$$s_2(f) \triangleq \max_{D\sim D'} \left\|f(D) - f(D')\right\|_2.$$

Here, $\|\cdot\|_1$ *and* $\|\cdot\|_2$ *denote the* L_1 *and* L_2 *norm on* $\mathbb{R}^k$ *respectively.*

Theorem 2.13 (The Laplace Mechanism (Theorem 3.6, [DR14a])). *Let $f : \mathcal{X} \to \mathbb{R}^k$ be a deterministic algorithm with ℓ_1 sensitivity $s_1(f)$. Let*

$$\mathcal{M}_L(f, \varepsilon) \triangleq f + (Y_1, \ldots, Y_k),$$

where Y_i are i.i.D from $\mathrm{Lap}\left(\frac{s_1(f)}{\varepsilon}\right)$.

The mechanism $\mathcal{M}_L(f, \varepsilon)$ is called the Laplace Mechanism, and satisfies ε-pure DP.

The Laplace mechanism achieves pure DP. To achieve approximate or zero concentrated DP, we use the Gaussian mechanism.

Theorem 2.14 (The Gaussian Mechanism [DR14a, BS16]). *Let $f : \mathcal{X} \to \mathbb{R}^k$ be a a deterministic algorithm with ℓ_2 sensitivity $s_2(f)$. Let*

$$\mathcal{M}_G(f, b) \triangleq f + Z,$$

such that $Z \sim \mathcal{N}(0, b \times s_2(f)^2 I_d)$. Here, $\mathcal{N}(\mu, \Sigma)$ denotes the Gaussian distribution with mean μ and co-variance matrix Σ, and $\|\cdot\|_2$ denotes the L_2 norm on $\mathbb{R}^k$.

The mechanism $\mathcal{M}_G(f, b)$ is called the Gaussian Mechanism.

- *For $b = \frac{2}{\varepsilon^2} \log\left(\frac{1.25}{\delta}\right)$, $\mathcal{M}_G(f, b)$ satisfies (ε, δ)-DP*
- *For $b = \frac{1}{2\rho}$, $\mathcal{M}_G(f, b)$ satisfies ρ-zCDP.*

Let us consider the empirical mean of a dataset as an example to illustrate the use of the Laplace and Gaussian mechanisms.

Example 2.15 (Empirical mean). *Let $k \in \mathbb{N}^*$ and $\mathcal{X} = \{0, 1\}^k$,* i.e. *the input universe is the binary vectors of length k. The empirical mean mechanism f_m associates to a dataset $D = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ the vector $f_m(D) = \frac{1}{n} \sum_{i=1}^n x_i$.*

The empirical mean's value changes the most if a row x_i goes from $(1, \ldots, 1)$ to zeros $(0, \ldots, 0)$. Thus, the ℓ_1 sensitivity of f_m is k/n, and the ℓ_2 sensitivity is $\sqrt{k}/n$.

1. *Adding i.i.d Laplace noise $f_m + \bigotimes^k \mathrm{Lap}(k/\varepsilon)$ achieves ε-pure DP, with L_2 error of order*

$$\left\| \bigotimes^k \mathrm{Lap}(k/\varepsilon) \right\|_2 \approx \frac{k^{3/2}}{\varepsilon n}.$$

2. *Adding i.i.d Gaussian noise $f_m + \mathcal{N}\left(0, \frac{2k}{\varepsilon^2 n^2} \log\left(\frac{1.25}{\delta}\right) I_k\right)$ achieves (ε, δ)-DP, with L_2 error of order*

$$\left\| \mathcal{N}\left(0, \frac{2k}{\varepsilon^2 n^2} \log\left(\frac{1.25}{\delta}\right) I_k\right) \right\|_2 \approx \frac{k}{\varepsilon n} \log\left(\frac{1}{\delta}\right).$$

3. *Adding i.i.d Gaussian noise* $f_m + \mathcal{N}\left(0, \frac{k}{2\rho n^2} I_k\right)$ *achieves* ρ*-zCDP, with* L_2 *error of order*

$$\left\| \mathcal{N}\left(0, \frac{k}{2\rho n^2} I_k\right) \right\|_2 \approx \frac{k}{n\sqrt{\rho}}.$$

In addition to the Laplace and Gaussian mechanisms, it is possible to achieve DP using other mechanisms like the Exponential mechanism [MT07], randomised response [War65], and the Sparse Vector Technique (SVT) [DNR+09], among others. In this thesis, we only use the Laplace and Gaussian mechanisms as building blocks, in addition to composition and post-processing, to design and analyse different sophisticated DP algorithms in Chapter 5.

2.1.5 DP under continual observation

In the continual observation setting [DNPR10a, CSS11], a mechanism receives its input dataset $D \triangleq \{x_1, \ldots, x_T\} \in \mathcal{X}^T$ as a stream. Specifically, at each step $t \in [T]$, the mechanism gets a record x_t and outputs an answer a_t. Similar to the batch model, a continual release mechanism satisfies DP if the mechanism's output sequence $(a_1, \ldots, a_T)$ is indistinguishable under two neighbouring input streams.

A fundamental problem studied in continual observation is the continual counting problem. Given an input stream $\sigma_T \in \{0,1\}^T$, a private counting mechanism outputs at each step t a private approximation $a_t \sim \mathcal{M}(\sigma)_t$ of the sum $\sum_{s=1}^{t} \sigma_s$. A counting mechanism $\mathcal{M}$ is then (α, β)-useful, if for all time-steps $t \in [T]$, with probability at least $1 - \beta$ over the randomness of $\mathcal{M}$, we have $|\mathcal{M}(\sigma)_t - \sum_{s=1}^{t} \sigma_s| \leq \alpha$.

It is possible to construct two simple mechanisms based solely on the Laplace mechanism.

Simple Mechanism I. Since the input stream is binary, the sum up to step t has L_1 sensitivity of 1. Thus, consider the simple mechanism that outputs $\left(\sum_{s=1}^{t} \sigma_s\right) + \mathrm{Lap}\,(1/\varepsilon)$ at each step t. Each released sum satisfies ε-DP. Thus, using basic composition, releasing the whole sequence of sums is $(T\varepsilon)$-DP. Equivalently, to design a mechanism that is ε-DP, each published sum should be $\left(\sum_{s=1}^{t} \sigma_s\right) + \mathrm{Lap}\,(T/\varepsilon)$. The error at each step is then $O(T/\varepsilon)$.

Simple Mechanism II. A second approach is to add to each input stream σ_t a sample $\gamma_t \sim \mathrm{Lap}\,(1/\varepsilon)$. Then, the Simple Mechanism II outputs $\sum_{s=1}^{t}(\sigma_s + \gamma_s)$. It is straightforward that the mechanism satisfies ε-DP. At each step t, the noise added is the sum of t independent $\mathrm{Lap}(1/\varepsilon)$. Hence, using the concentration of the sum of Laplace distributions, the noise level added at each step t can be shown to be of order $O(\sqrt{t}/\varepsilon)$. This already improves Simple Mechanism I's error.

Building on the intuitions from the two simple mechanisms, the binary tree mechanism [DNPR10a, CSS11] is proposed.

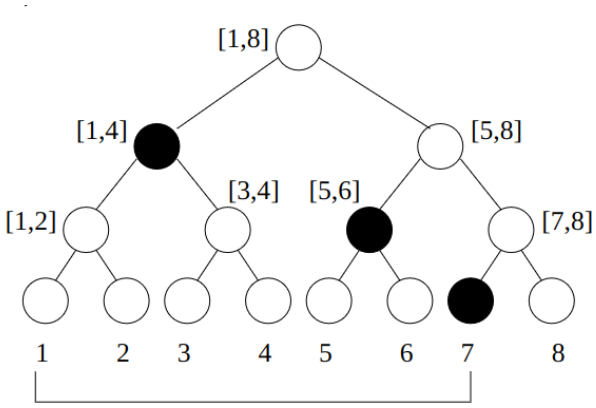


Figure 2.2 – The binary tree construction for the interval [1,8]. The sum of time steps 1 through 7 can be recovered by adding the p-sums corresponding to the black nodes [CSS11].

The binary tree mechanism [DNPR10a, CSS11]. The complexity of the counting problem lies in the fact that the input values observed early in the process impact all subsequent output values. The binary tree mechanism builds on this observation to provide an ε-DP counting algorithm, with error at each step of order $O\left(\frac{\log(T)^{3/2}}{\varepsilon}\right)$. The best way to understand the binary tree mechanism is to think in terms of the partial sum (*i.e.* p-sum) framework. A p-sum is a partial sum of consecutive items. For $1 \leq i \leq j \leq T$, the notation $\sigma[i,j] \triangleq \sum_{t=i}^{j} \sigma_t$ denotes the partial sum involving items i through j. Instead of outputting the estimated counts, the binary tree mechanism releases a sequence of noisy p-sums. This sequence of p-sums has the property of providing "sufficient" information to estimate the count at each time step t.

The intuition is best explained using a binary interval tree as shown in Figure 2.2. Each leaf node in the tree represents a time step, and each interior node represents a range. The binary tree mechanism releases the p-sums corresponding to each node in the tree. Then, to recover the sum of time steps 1 through t, finding a set of nodes in the tree to uniquely cover the range $[1,t]$ suffices. This construction has two properties:

(a) Every time step t appears in at most $\log T$ p-sums. Equivalently, this means that changing one σ_t in the input only changes at most $O(\log T)$ p-sums. Thus, using the Laplace mechanism, adding a noise of scale $\log(T)/\varepsilon$ to each p-sum in the tree makes the release of the whole tree ε-DP.

(b) Every continuous range $[1,t]$ can be represented with a set of $O(\log T)$ nodes in the tree. Equivalently, this means that each sum $\sum_{s=1}^{t} \sigma_s$ can be recovered by summing at most $O(\log T)$ p-sums from the tree.

Combining points (a) and (b) gives that the binary tree mechanism releases each sum $\sum_{s=1}^{t} \sigma_s$ with a noise that is (at most) the sum of $O(\log T)$ i.i.d Laplace variables $\mathrm{Lap}(\log(T)/\varepsilon)$.

Again, using the concentration of the sum of Laplace variables, we conclude that the scale of the noise added at each step is $O(\log(T)^{3/2}/\varepsilon)$.

The binary tree mechanism is used as a building block for many online learning algorithms. In Chapter 5, we discuss some shortcomings of adapting this mechanism to derive private bandit algorithms, and propose simpler ways to overcome these shortcomings.

2.2 Multi-Armed Bandits

In this section, we formalise the bandit problem and its two main utility measures studied in this thesis: regret minimisation and Best Arm Identification (BAI). For each utility objective, we first present state-of-the-art algorithm design ideas and analyse their utility. Then, we show that these algorithms achieve (asymptotic) *exact* optimality by providing tight lower bounds.

2.2.1 The language of bandits

A bandit problem is a sequential game between a learner and an environment. The game is played over T rounds, where $T \in \mathbb{N}^\star$ is a positive natural number called the horizon. In each round $t \in [T]$, the learner first chooses an action a_t from a given set $\mathcal{A}$. Actions are also called "arms" in the literature. Then, the environment reveals a reward $r_t \in \mathbb{R}$. The learner chooses arm a_t based only on the interaction history $H_{t-1} \triangleq (a_1, r_1, \dots, a_{t-1}, r_{t-1})$. A policy is a mapping from histories to actions. An environment is a mapping from history sequences ending in actions to rewards. Both the learner and the environment may randomise their decisions. The most common objective of the learner is to choose actions that lead to the largest possible cumulative reward over the T rounds, *i.e.* $\sum_{t=1}^{T} r_t$. The fundamental challenge in bandit problems is that the environment is unknown to the learner. All the learner might know is that the true environment lies in some set $\mathcal{E}$ called the environment class.

2.2.2 The canonical model for stochastic bandits

A simple problem setting is that of stochastic stationary bandits. In this case, the environment is restricted to generating the reward in response to each action from a distribution that is specific to that action and independent of the previous action choices and rewards. A stochastic bandit (or environment) is a collection of distributions $\nu \triangleq (P_a : a \in \mathcal{A})$, where $\mathcal{A}$ is the set of available actions. The learner and the environment interact sequentially over T rounds. In each round $t \in 1, \dots, T$, the learner chooses an action $a_t \in \mathcal{A}$, which is fed to the environment. The environment then samples a reward $r_t \in \mathbb{R}$ from distribution P_{a_t} and reveals r_t to the learner. The interaction between the learner (or policy) and environment induces a probability

Algorithm 1 Bandit interaction between a policy and an environment

Input: A policy π and an environment $\nu \triangleq (P_a : a \in [K])$
for $t = 1, \dots$ **do**
    The policy samples an action $a_t \sim \pi_t(. \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$
    The policy observes a reward $r_t \sim P_{a_t}$
end for
if Regret minimisation **then**
    The interaction ends after T steps
else FC-BAI
    The policy decides to stop the interaction at step τ and recomends the final guess $\hat{a}$
end if

measure on the sequence of outcomes $a_1, r_1, a_2, r_2, \dots, a_T, r_T$. In the following, we construct the probability space that carries these random variables.

Let $T \in \mathbb{N}^\star$ be the horizon. Let $\nu = (P_a : a \in [K])$ a bandit instance with $K \in \mathbb{N}^\star$ finite arms, and each P_a is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with $\mathfrak{B}$ being the Borel set. For each $t \in [T]$, let $\Omega_t = ([K] \times \mathbb{R})^t \subset \mathbb{R}^{2t}$ and $\mathcal{F}_t = \mathfrak{B}(\Omega_t)$. We first formalise the definition of a policy.

Definition 2.16 (The policy)**.** *A policy π is a sequence $(\pi_t)_{t=1}^T$, where π_t is a probability kernel from $(\Omega_t, \mathcal{F}_t)$ to $([K], 2^{[K]})$. Since $[K]$ is discrete, we adopt the convention that for $a \in [K]$,*

$$\pi_t(a \mid a_1, r_1, \dots, a_{t-1}, r_{t-1}) = \pi_t(\{a\} \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$$

We want to define a probability measure on $(\Omega_T, \mathcal{F}_T)$ that respects our understanding of the sequential nature of the interaction between the learner and a stationary stochastic bandit. Specifically, the sequence of outcomes should satisfy the following two assumptions:

(a) The conditional distribution of action a_t given $a_1, r_1, \dots, a_{t-1}, r_{t-1}$ is $\pi(a_t \mid H_{t-1})$ almost surely.

(b) The conditional distribution of reward r_t given $a_1, r_1, \dots, a_{t-1}, r_{t-1}, a_t$ is P_{a_t} almost surely.

The probability measure on $(\Omega_T, \mathcal{F}_T)$ depends on both the environment ν and the policy π. To construct this probability, let λ be a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for which P_a is absolutely continuous with respect to λ for all $a \in [K]$. Let $p_a = dP_a/d\lambda$ be the Radon–Nikodym derivative of P_a with respect to λ. Letting ρ be the counting measure with $\rho(B) = |B|$, the density $p_{\nu\pi} : \Omega_T \to \mathbb{R}$ can now be defined with respect to the product measure $(\rho \times \lambda)^T$ by

$$p_{\nu\pi}(a_1, r_1, \dots, a_T, r_T) \triangleq \prod_{t=1}^{T} \pi_t(a_t \mid a_1, r_1, \dots, a_{t-1}, r_{t-1}) p_{a_t}(r_t)$$

and $\mathcal{P}_{\nu\pi}$ is defined as

$$\mathbb{P}_{\nu\pi}(B) \triangleq \int_B p_{\nu\pi}(\omega)(\rho \times \lambda)^T(\,\mathrm{d}\omega) \quad \mathrm{forall} B \in \mathcal{F}_T$$

Hence $(\Omega_T, \mathcal{F}_T, \mathbb{P}_{\nu\pi})$ is a probability space over histories induced by the interaction between π and ν. We define also a marginal distribution over the sequence of actions by

$$m_{\nu\pi}(a_1, \ldots, a_T) \triangleq \int_{r_1,\ldots,r_T} p_{\nu\pi}(a_1, r_1, \ldots, a_T, r_T)\,\mathrm{d}r_1 \ldots\, \mathrm{d}r_T,$$

and $\mathrm{forall} C \in \mathcal{P}([K]^T)$,

$$\mathbb{M}_{\nu\pi}(C) \triangleq \sum_{(a_1,\ldots,a_T)\in C} m_{\nu\pi}(a_1, a_2, \ldots, a_T).$$

Hence, $([K]^T, \mathcal{P}([K]^T), \mathbb{M}_{\nu\pi})$ is a probability space over sequence of actions produced when π interacts with ν for T time-steps.

2.2.3 Utility measures in bandits as learning objectives

We discuss two learning objectives.

Regret Minimisation. The first learning objective studied in this thesis is regret minimisation. Informally, the regret of a policy is the deficit suffered by the learner relative to the optimal policy that knows the environment, and plays always the optimal arm. Let $\nu = (P_a : a \in [K])$ a bandit instance and define $\mu_a(\nu) = \int_{-\infty}^{\infty} x dP_a(x)$ the mean of reward distribution P_a. We assume throughout that $\mu_a(\nu)$ exists and is finite for all actions. Let $\mu^\star(\nu) = \max_{a\in[K]} \mu_a(\nu)$ the largest mean among all the arms. The regret of policy π on bandit instance ν is

$$\mathrm{Reg}_T(\pi, \nu) \triangleq T\mu^\star(\nu) - \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} r_t\right]. \tag{2.7}$$

The expectation is taken with respect to the probability measure $\mathbb{P}_{\nu\pi}$ on action-reward sequences induced by the interaction of π and ν. The regret can be decomposed in terms of the loss due to pulling each of the sub-optimal arms:

$$\mathrm{Reg}_T(\pi, \nu) = \sum_{a=1}^{K} \Delta_a \mathbb{E}_{\nu\pi}\left[N_a(T)\right], \tag{2.8}$$

where $N_a(T) \triangleq \sum_{t=1}^{T} \mathbb{1}\left(a_t = a\right)$ is the number of times the arm a is played till T, and $\Delta_a \triangleq \mu^\star(\nu) - \mu_a(\nu)$ is the sub-optimality gap.

So what can we hope for? A relatively weak objective is to find a policy π with sub-linear regret on all $\nu \in \mathcal{E}$. Formally, this objective is to find a policy π such that, for all $\nu \in \mathcal{E}$

$$\lim_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \nu)}{T} = 0$$

Best Arm Identification. The second learning objective is Best Arm Identification (BAI). BAI is a pure exploration problem that aims to find the arm with the highest expected reward, *i.e.* $a^\star(\nu) \triangleq \arg\max_{a\in[K]} \mu_a(\nu)$. It has been studied in two major theoretical frameworks [ABM10, GGL12, JN14, GK16]: the fixed-confidence and fixed-budget setting. In the fixed-budget setting, the objective is to minimise the probability of misidentifying a correct answer with a fixed number of samples T. In this thesis, we consider the fixed-confidence setting (FC-BAI), in which the learner aims at minimising the number of samples used to identify a correct answer with confidence $1-\delta \in (0,1)$. [1] To achieve this, the learner defines an FC-BAI strategy to interact with the bandit instance $\boldsymbol{\nu} = \{\nu_a\}_{a\in[K]}$, consisting of K arms with finite means $\{\mu_a\}_{a\in[K]} \in (0,1)^K$. We assume that there is a unique best arm $a^\star(\boldsymbol{\nu})$ defined as $a^\star(\boldsymbol{\nu}) = \arg\max_{a\in[K]} \mu_a$. We augment the action set by a *stopping action* $\top$, and write $a_t = \top$ to denote that the algorithm has stopped before step t. In the following, we define the ingredients of an FC-BAI strategy.

Definition 2.17 (FC-BAI strategy)**.** *A FC-BAI strategy π^{BAI} is composed of*

i. A pair of sampling and stopping rules $(\mathrm{S}_n : \mathcal{H}_{n-1} \to \mathcal{P}([K]\cup\{\top\}))_{n\geq 1}$. For an action $a \in [K]$, $\mathrm{S}_n(a \mid \mathcal{H}_{n-1})$ denotes the probability of playing action a given history $\mathcal{H}_{n-1}$. On the other hand, $\mathrm{S}_n(\top \mid \mathcal{H}_{n-1})$ is the probability of the algorithm halting given $\mathcal{H}_{n-1}$. For any history $\mathcal{H}_{n-1}$, a consistent sampling and stopping rule S_n satisfies $\mathrm{S}_n(\top \mid \mathcal{H}_{n-1}) = 1$ if $\top$ has been played before n.

ii. A recommendation rule $(\mathrm{Rec}_n : \mathcal{H}_{n-1} \to \mathcal{P}([K]))_{n>1}$. This rule dictates $\mathrm{Rec}_n(a \mid \mathcal{H}_{n-1})$, i.e. *the probability of returning action a as a final guess for the best action given $\mathcal{H}_{n-1}$.*

In addition to the sampling rule, which is the same as the definition of a policy in regret minimisation (*i.e.* Defintion 2.16), an FC-BAI strategy has a stopping rule, that dictates when the strategy stops sampling, and a recommendation rule that proposes a final guess of the optimal arm after stopping. To analyse the performance of an FC-BAI strategy, we define the stopping time and δ-correctness.

Definition 2.18 (Stopping time)**.** *We denote by τ_δ the **stopping time** (or **sample complexity**) of the policy π^{BAI} the first step n such that $a_n = \top$.*

The FC-BAI strategy is δ-correct if, after stopping, it recommends the optimal arm with probability $1-\delta$.

[1] We remind not to confuse risk level δ with the δ of (ε,δ)-DP. Later when studying privacy for BAI, we only consider ε-pure DP as the privacy definition, and δ always represents the risk (or probability of mistake) of the BAI strategy.

Definition 2.19 (δ-correctness)**.** *A FC-BAI strategy π^{BAI} is called δ-**correct** for a class of bandit instances $\mathcal{E}$, if for every bandit instance $\boldsymbol{\nu} \in \mathcal{E}$, π^{BAI} recommends $\widehat{a}$ as the optimal action $a^\star(\boldsymbol{\nu})$ with probability at least $1-\delta$,* i.e. $\mathbb{P}_{\pi^{\mathrm{BAI}},\boldsymbol{\nu}}(\tau_\delta < +\infty, \widehat{a} = a^\star(\boldsymbol{\nu})) \geq 1-\delta$.

The goal in FC-BAI is to design a δ-correct BAI policy, with the smallest expected sample complexity $\mathbb{E}_{\pi,\nu}[\tau(\delta)]$.

2.2.4 Concentration of measure

Before discussing the design and analysis of bandit algorithms, we need to introduce one more tool from probability theory, called concentration of measure. Since the mean rewards are initially unknown, they need to be estimated from the history. How long does it take to learn about the mean reward of an action? The main technique introduced here is the Crameŕ Chernoff exponential tail inequalities for subgaussian random variables.

Definition 2.20 (Subgaussianity)**.** *A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$$

The tails of a σ-subgaussian random variable decay approximately as fast as that of a Gaussian with zero mean and the same variance.

Lemma 2.21 (Concentration of subgaussian random variables)**.** *If X is σ-subgaussian, then for any $\varepsilon \geq 0$,*

$$\mathbb{P}\left(X \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

The proof is based on a generic approach called the Cramér Chernoff method.

Let $\lambda > 0$, then

$$\begin{aligned}\mathbb{P}\left(X \geq \varepsilon\right) &= \mathbb{P}\left(\exp\left(\lambda X\right) \geq \exp\left(\lambda\varepsilon\right)\right)\\ &\overset{(a)}{\leq} \mathbb{E}[\exp(\lambda X)]exp(-\lambda\varepsilon)\\ &\overset{(b)}{\leq} \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right)\end{aligned}$$

where (a) is due to Markov's inequality, and (b) is by definition of subgaussianity.

Taking $\lambda = \varepsilon/\sigma^2$ concludes the proof.

Lemma 2.22 (Properties of Subgaussian Random Variables)**.** *Suppose that X_1 and X_2 are independent and σ_1 and σ_2-subgaussian, respectively, then*

1. *cX is $|c|\,\sigma$-subgaussian for all $c \in \mathbb{R}$.*
2. *$X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian.*
3. *If X has mean zero and $X \in [a, b]$ almost surely, then X is $\frac{b-a}{2}$-subgaussian.*

2.2.5 Regret minimisation algorithms

To minimize regret, the optimal strategy is to always choose the optimal arms. This is obviously not feasible since the environment is unknown to the learner.

Failure of Follow The Leader (FTL). A first natural idea to deal with the unknown environment is to estimate the empirical mean of each arm, and then pull the arm with the highest estimated mean. Specifically, the Follow-The-Leader algorithm (FTL) starts by pulling each arm once, then chooses at step t the action

$$a_t \in \arg\max_{a \in [K]} \hat{\mu}_a(t-1)$$

where $\hat{\mu}_a(t) \triangleq \frac{1}{N_a(t)} \sum_{s=1}^{t} \mathbb{1}(A_s = a)\, r_s$ is the empirical mean estimate of μ_a at step t, and $N_a(t) \triangleq \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of pulls of arm a at step t.

It is easy to show that FTL achieves linear regret. Specifically, the optimal arm may appear to be worse than a sub-optimal arm after the first exploratory round and thus never be pulled again, leading to linear regret.

One potential fix for this strategy is to explore each arm for $m > 1$ rounds, then commit to the best empirical arm for the rest of the rounds. This is the Explore-then-Commit (ETC) algorithm. It is clear that the regret of ETC depends heavily on the choice of m. As for FTL, if m is too small, then there is a high probability of committing to a sub-optimal arm, leading to linear regret. On the other hand, choosing m to be big enough means that many rounds were "wasted" on sub-optimal arms, leading to linear regret. Optimising for m captures well the exploration-exploitation dilemma. We refer to Chapter 6 in [LS20] for a complete regret analysis of the algorithm. The take-away message from the analysis is that optimising for m to get sub-linear regret depends on the knowledge of the gaps between the means $\Delta_a \triangleq \mu^\star - \mu_a$, and on the knowledge of the horizon T.

Optimism in the face of uncertainty and the UCB algorithms. The principle of optimism in the face of states that one should act as if the environment is as nice as plausibly possible. Applying this principle for bandits means using the history observed so far to assign to each arm an index (*i.e.* a value), called the Upper Confidence Bound (UCB), that with high probability is an overestimate of the unknown mean. The UCB algorithm chooses at each step the arm with the highest estimated upper confidence bounds. The reason why UCB is successful is that

Algorithm 2 UCB Meta-algorithm

Input: K number of arms
Initialisation: Choose each arm one.
for $t > K$ **do**
    Chose the optimistic arm

$$A_t \in \arg\max_{a \in [K]} \mathrm{UCB}_a(t-1)$$

    Observe reward r_t and update upper confidence bounds
end for

the upper confidence bound of an arm is high if either the true mean is high and the arm is the best choice (exploitation), or if the arm has not been chosen often enough, which means that exploring this arm will provide useful information about the environment (exploration). In one step, the UCB algorithm integrates both exploration and exploitation.

The main ingredient to instantiate the UCB meta-algorithm (Algorithm 2) is to construct the high-probability upper confidence on the real means $(\mathrm{UCB}_a(t-1))_{a\in[K]}$ at each step t. To do so, we will use the concentration of measure results on subgaussian random variables introduced earlier. Let $(X_t)_{t=1}^T$ be a sequence of independent 1-subgaussian random variables, with mean μ and empirical mean $\hat{\mu}_T \triangleq \frac{1}{T}\sum_{t=1}^T X_t$. By Lemma 2.22, $\hat{\mu}_T - \mu = \sum_{t=1}^T (X_t - \mu)/n$ is $1/\sqrt{T}$-subgaussian. Using Proposition 2.21, we get that $\mathbb{P}(\hat{\mu}_T - \mu \geq \varepsilon) \leq \exp(-\frac{T\varepsilon^2}{2})$. This means that, with probability $1-\delta$ for $\delta \in (0,1)$, $\mu \leq \hat{\mu}_T + \sqrt{\frac{2\log(1/\delta)}{T}}$.

In bandits, the policy observed $N_a(t-1)$ samples from arm a at step t and thus builds an upper confidence index:

$$\mathrm{UCB}_a(t-1,\delta) \triangleq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \tag{2.9}$$

The UCB index is the sum of the empirical mean of rewards, and the exploration bonus, also known as the confidence width. We now provide an upper bound on the regret of the UCB algorithm.

Theorem 2.23. *Consider the UCB meta-algorithm of Algorithm 2, with the index defined in Eq (2.9). Let ν be a K-armed 1-subgaussian bandit instance. For any horizon T, with $\delta = 1/T^2$, then the problem-dependent regret upper bound is*

$$\mathrm{Reg}_T(\mathrm{UCB}, \nu) \leq 3\sum_{a=1}^K \Delta_a + \sum_{a:\Delta_a>0} \frac{16\log(T)}{\Delta_a} \tag{2.10}$$

This gives also the following gap-free (also called worst-case, problem free or problem independent) upper bound

$$\mathrm{Reg}_T(\mathrm{UCB}, \nu) \leq 8\sqrt{TK\log(T)} + 3\sum_{a=1}^{K} \Delta_a$$

In particular, if $\Delta_a \leq 1$, then

$$\mathrm{Reg}_T(\mathrm{UCB}, \nu) \leq 8\sqrt{TK\log(T)} + 3K \tag{2.11}$$

UCB achieves sub-linear regret. The proof can be found in Chapter 7 of [LS20]. The proof defines the "good event": all the real means are well-estimated by the empirical means, which happens with probability $1-\delta$. Under this good event, the UCB algorithm stops sampling a sub-optimal arm a once the confidence width is smaller than the mean gap. Specifically, if n_a is the number of times that UCB samples a sub-optimal arm a, then n_a verifies that $2\sqrt{\frac{2\log(1/\delta)}{n_a}} \leq \Delta_a$. Solving for n_a and replacing $\delta = 1/T$ gives the problem-dependent regret upper bound of Equation (2.10). It is called a problem-dependent bound since it depends on the gaps Δ_a, which depend on the means of the rewards distributions of the bandit instance. The gap-free upper bound of Equation (2.11) is retrieved by optimising for the worst-case instance. Another way of proving the gap-free bound is by showing that the regret of UCB algorithms is upper bounded by the sum of the confidence intervals width. In the UCB index of Equation (2.9), the width of the confidence interval is $\sqrt{\frac{4\log(1/\delta)}{N_a(t-1)}}$. Thus, an upper bound on the regret is $\sum_{t=1}^{T} 2\sqrt{\frac{2\log(1/\delta)}{N_{a_t}(t-1)}}$. To provide an intuition on upper bounding this quantity, it is possible to think of $N_{a_t}(t-1) \approx t$, and thus $\sum_t \frac{1}{\sqrt{t}} \approx \int_t \frac{1}{\sqrt{t}} dt = \sqrt{T}$.

As we will show later, for better choices of the exploration bonuses, UCB is optimal and matches exactly the regret lower bounds. For example, let us consider the following index:

$$\mathrm{UCB}_a(t-1) \triangleq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log f(t)}{N_a(t-1)}} \tag{2.12}$$

where $f(t) \triangleq 1 + t\log^2(t)$.

This index is different than the index of Equation (2.9) in the choice of the exploration bonus. In addition to providing better regret upper bound, the index in Equation (2.12) has the advantage of being independent of the risk parameter δ, thus making the algorithm independent of the apriori knowledge of the horizon T. This kind of algorithm is known as an "anytime algorithm".

Theorem 2.24. *Consider the UCB meta-algorithm of Algorithm 2, with the index defined in Equation (2.12). Let ν be a K-armed 1-subgaussian bandit instance. For any horizon T, the asymptotic*

problem-dependent regret upper bound is

$$\lim_{T\to\infty} \frac{\mathrm{Reg}_T(\mathrm{UCB}, \nu)}{\log(T)} \leq \sum_{a:\Delta_a>0} \frac{2}{\Delta_a}$$

This gives also the following gap-free upper bound

$$\mathrm{Reg}_T(\mathrm{UCB}, \nu) \leq 2\sqrt{CKT\log(T)} + C\sum_{a=1}^{K} \Delta_a$$

for some constant $C > 0$.

This asymptotic bound on the regret is not improvable in a strong sense.

Finally, for the stronger assumption that the rewards are Bernoulli, it is possible to design tighter upper confidence bounds. The Bernoulli model where rewards are in $\{0, 1\}$ is fundamental in many applications, where the environment's feedback is binary.

The Bernoulli distribution is $1/2$-subgaussian regardless of its mean. Thus, the results of the previous algorithm are applicable here. However, the additional knowledge that the rewards are Bernoulli is not being fully exploited by these algorithms. The reason is essentially that the variance of a Bernoulli random variable depends on its mean, and when the variance is small, the empirical mean concentrates faster, a fact that should be used to make the confidence intervals smaller.

We introduce the relative entropy between Bernoulli random variables to construct tighter confidence bounds on the real mean.

Definition 2.25 (Relative entropy between Bernoulli distributions)**.** . *The relative entropy between Bernoulli distributions with parameters $p, q \in [0, 1]$ is*

$$\mathrm{kl}(p, q) = p\log(p/q) + (1-p)\log((1-p)/(1-q)) \tag{2.13}$$

and singularities are defined by taking limits.

Using Chernoff's bound as the concentration of measure tool, it is possible to define the following UCB index, called KL-UCB in the literature:

$$\text{KL-UCB}_a(t-1) \triangleq \max\left\{\tilde{\mu} \in [0, 1] : \mathrm{kl}(\hat{\mu}_a(t-1), \tilde{\mu}) \leq \frac{\log f(t)}{N_a(t-1)}\right\} \tag{2.14}$$

where $f(t) \triangleq 1 + t\log^2(t)$.

Theorem 2.26. *Consider the UCB meta-algorithm of Algorithm 2, with the* KL-UCB *index defined in Equation (2.14). Let ν be a K-armed Bernoulli bandit instance. For any horizon T, the asymptotic*

problem-dependent regret upper bound is

$$\lim_{T\to\infty} \frac{\mathrm{Reg}_T(\text{KL-UCB}, \nu)}{\log(T)} \leq \sum_{a:\Delta_a>0} \frac{\Delta_a}{d(\mu_a, \mu^\star)}$$

This improves on the regret upper bound of Theorem 2.24, since by Pinsker's inequality $d(\mu_a, \mu^\star) \geq 2(\mu^\star - \mu_a)^2 = 2\Delta_a^2$.

2.2.6 Regret lower bounds

In this section, we provide two flavours of regret lower bounds for bandits.

The first type of regret lower bounds studied is the minimax regret lower bounds.

Definition 2.27 (Minimax Regret)**.** *Let $\mathcal{E}$ be a class of bandit instances. The worst case regret of a policy π on the class $\mathcal{E}$ is*

$$\mathrm{Reg}_T(\pi, \mathcal{E}) \triangleq \sup_{\nu\in\mathcal{E}} \mathrm{Reg}_T(\pi, \nu).$$

Let Π be the set of all policies. The minimax regret is

$$\mathrm{Reg}_T^\star(\mathcal{E}) \triangleq \inf_{\pi\in\Pi} \mathrm{Reg}_T(\pi, \mathcal{E}) = \inf_{\pi\in\Pi} \sup_{\nu\in\mathcal{E}} \mathrm{Reg}_T(\pi, \nu)$$

A small value of $\mathrm{Reg}_T^\star(\mathcal{E})$ indicates that the underlying bandit problem is less challenging in the worst-case sense. The main result of this part is to show that $\mathrm{Reg}_T^\star(\mathcal{E})$ is $\Omega(\sqrt{KT})$.

Theorem 2.28 (Minimax Regret Lower Bound)**.** *Let $\mathcal{E}_\mathcal{G}^K$ be the set of K-armed Gaussian bandits, with unit variance. Then, for $K > 1$ and $T \geq K - 1$,*

$$\mathrm{Reg}_T^\star(\mathcal{E}_\mathcal{G}^K) \geq \frac{1}{27}\sqrt{(K-1)T}.$$

The method used to prove Theorem 2.28 can be viewed as a generalisation and strengthening of Le Cam's method in statistics. There are two differences compared to Le Cam's method [LeC73]: (a) dealing with a sequential setting, and (b) choosing the alternative problem depends on the algorithm. In Section 4.1 of Chapter 4, we present in detail the intuition and techniques used to prove these lower bounds, and how to adapt it to the DP policy class, *i.e.* the subclass of policies that satisfy Differential Privacy.

The minimax regret lower bound serves as a useful measure of the robustness of a policy but can be excessively conservative. Instance-dependent lower bounds overcome this by capturing the optimal performance of a policy on a specific bandit instance. Since minimising regret over a class of bandit instances is a multi-objective criterion, an algorithm designer might try and

design algorithms that perform well on one kind of instance. An extreme example is the policy that always chooses $A_t = 1$, which suffers zero regret when the first arm is optimal and linear regret otherwise. Thus, it is important to define exactly what is meant by a reasonable policy. An example of such a class of reasonable policies is the consistent policies, *i.e.* policies that achieve sub-linear regret on every instance in the class $\mathcal{E}$.

Definition 2.29 (Consistent Policy)**.** *A policy π is called consistent over a class of bandits $\mathcal{E}$ if forall $\nu \in \mathcal{E}$ and $p > 0$, it holds that*

$$\lim_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \nu)}{T^p} = 0$$

The class of consistent policies over $\mathcal{E}$ is denoted by $\Pi_{\mathrm{cons}}(\mathcal{E})$.

For example, the UCB algorithm is consistent over the class of 1-subgaussian random variables. The strategy that always chooses the first action is not consistent on any class $\mathcal{E}$ unless the first arm is optimal for every $\nu \in \mathcal{E}$.

The main result of this section is a problem-dependent lower bound, for consistent policies, over the class of unstructured bandits, *i.e.* the class of bandits $\mathcal{E} \triangleq \mathcal{P}_1 \times \cdots \times \mathcal{M}_K$, with $\mathcal{P}_1, \ldots, \mathcal{M}_K$ are sets of distributions.

Definition 2.30 (The KL inf)**.** *Let $\mathcal{M}$ be a set of distributions with finite means. Let $\mu : \mathcal{P} \to \mathbb{R}$ the function that maps a distribution $P \in \mathcal{P}$ to its mean. Let $\mu^\star \in \mathbb{R}$ and $P \in \mathcal{M}$ such that $\mu(P) < \mu^\star$. Define*

$$\mathrm{KL}_{\inf}(P, \mu^\star, \mathcal{M}) \triangleq \inf_{P' \in \mathcal{M}} \{\mathrm{KL}(P, P') : \mu(P') > \mu^\star\}$$

where KL *is the Kullback-Leibler divergence,* i.e. *for two probability distributions $\mathbb{P}, \mathbb{Q}$ on $(\Omega, \mathcal{F})$, the KL divergence is $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) \triangleq \int \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega)\right) \mathrm{d}\mathbb{P}(\omega)$ when $\mathbb{P} \ll \mathbb{Q}$, and $+\infty$ otherwise.*

Theorem 2.31. *Let $\mathcal{E} \triangleq \mathcal{M}_1 \times \cdots \times \mathcal{M}_K$ and $\pi \in \Pi_{\mathrm{cons}}(\mathcal{E})$ a consistent policy over $\mathcal{E}$. Then, for any $\nu = (P_a : a \in K) \in \mathcal{E}$,*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \nu)}{\log(T)} \geq \sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathrm{KL}_{\inf}(P_a, \mu^\star, \mathcal{M}_a)} \tag{2.15}$$

where $\Delta_a \triangleq \mu^\star - \mu_a$ is the suboptimality gap of arm a.

The lower bound of Theorem 2.31 and the "KL inf" are fundamental quantities that characterise the complexity of a bandit problem. For the class of Gaussian k-armed bandit with variance 1, it is easy to show that $\mathrm{KL}_{\inf}(P_a, \mu^\star, \mathcal{M}_a) = \frac{(\mu^\star - \mu_a)^2}{2} = \frac{\Delta_a^2}{2}$, which shows the asymptotic optimality of Theorem 2.24. For the Bernoulli bandit class, $\mathrm{KL}_{\inf}(P_a, \mu^\star, \mathcal{M}_a) = d(\mu_a, \mu^\star)$, which also validates the "exact" optimality of KL-UCB.

2.2.7 Contextual and linear bandits

In many bandit problems, the learner has access to additional information that may help predict the quality of the actions. Imagine creating a movie recommendation system in which users ask for movie suggestions to watch next. It would not be wise to disregard the user's demographic information, the history of movies watched, or ratings when making these recommendations.

The setting of contextual bandits is an "augmented" bandit framework that better models real-world problems where contextual information is available. In this setting, the policy observes at each step t a context $c_t \in \mathcal{C}$, which may be random or not. Having observed the context, the policy chooses an action $a_t \in [K]$ and observes a reward r_t. In the linear contextual bandits, the reward r_t depends on both the arm a_t and the context c_t in terms of a linear structural equation, that allows for learning to transfer from one context to another.

$$r_t \triangleq \langle \theta^\star, \psi(a_t, c_t) \rangle + \eta_t. \tag{2.16}$$

Here, $\psi : [K] \times \mathcal{C} \to \mathbb{R}^d$ is the feature map, $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is the noise, which may be assumed to be conditionally 1-subgaussian. While decision-making with Equation (2.16), all that matters is the value of the feature vector. Thus, the bandit literature often studies a reduced model [LS20], where in round t, the policy is served with the decision set $\mathcal{A}_t \subset \mathbb{R}^d$, from which it chooses an action $a_t \in \mathcal{A}_t$ and receives a reward $r_t \triangleq \langle \theta^\star, a_t \rangle + \eta_t$, where η_t is 1-subgaussian given $\mathcal{A}_1, a_1, R_1, \ldots, \mathcal{A}_{t-1}, a_{t-1}, R_{t-1}, \mathcal{A}_t$, and A_t. Different choices of $\mathcal{A}_t$ lead to different settings. For example, if $\mathcal{A}_t \triangleq \{\psi(c_t, a) : a \in [K]\}$, then we have a contextual linear bandit, or if $\mathcal{A}_t \triangleq \{e_1, \ldots, e_d\}$, where $(e_i)_i$ are the unit vectors of $\mathbb{R}^d$ then the resulting bandit problem reduces to a d-finite armed bandit. For the contextual bandit setting, the contexts can be either generated stochastically, *i.e.* sampled from some distribution , or assumed to be generated arbitrarily, *i.e.* adversarial contexts.

The goal is to design a policy that minimises the regret, which is defined as

$$R_T \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle \theta^\star, a - a_t \rangle\right]. \tag{2.17}$$

To design an algorithm for the linear contextual setting, we use again optimism in the face of uncertainty, *i.e.* acting like the environment is as nice as possible. For this setting, the main quantity of interest is the regression parameter $\theta^\star \mathbb{R}^d$. The main step to adapt UCB for linear contextual bandit is to construct a confidence set $\mathcal{C}_t \subset \mathbb{R}^d$ at each step t based on the history $(a_1, r_1, \ldots, a_{t-1}, r_{t-1})$ that contains with high probability the unknown parameter $\theta^\star$. Given a

confidence set $\mathcal{C}_t$, the LinUCB index is then defined as

$$\mathrm{LinUCB}_t(a) = \max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle$$

to be an upper bound on the mean reward $\langle \theta^\star, a \rangle$ of arm a.

The LinUCB algorithm then selects at each time step the arm

$$a_t = \arg\max_{a \in \mathcal{A}_t} \mathrm{LinUCB}_t(a).$$

The main question is how to choose $\mathcal{C}_t$. First, we need an analogue of the empirical mean in UCB to estimate the unknown $\theta^\star$. A natural candidate is the regularised least-squares estimator, also called the ridge estimator, which is

$$\hat{\theta}_t \triangleq V_t^{-1} \sum_{s=1}^{t-1} a_s r_s \tag{2.18}$$

where $V_t \triangleq \lambda I_d + \sum_{s=1}^{t-1} a_s r_s$ is a $d \times d$ matrix called the design matrix and $\lambda > 0$.

Since θ_t is an estimate of $\theta^\star$, a natural candidate for $\mathcal{C}_t$ is to be an ellipsoid centred at θ_t, *i.e.*

$$\mathcal{C}_t \triangleq \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_t \right\|_{V_t^{-1}} \leq \beta_t \right\} \tag{2.19}$$

for fine-tuned β_t that we specify later.

Under some boundness assumptions (Assumption 19.1 in [LS20]) and for a choice of β_t as in Equation (19.8) of [LS20], it is possible to show that the regret of LinUCB in a contextual linear bandit instance achieves

$$R_T \leq C d \sqrt{T} \log(T),$$

where C is a universal constant. The minimax lower bounds for linear contextual bandits (Chapter 24 in [LS20]) show that LinUCB is minimax-optimal up to logarithmic factors.

2.2.8 Sample complexity lower bound

In FC-BAI, being δ-correct imposes a lower bound on the expected sample complexity on any instance.

Theorem 2.32 ([GK16])**.** *Let $\delta \in (0, 1)$. For any δ-correct FC-BAI strategy and all instances $\boldsymbol{\nu} \in \mathcal{M}$, we have that*

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \, \mathrm{kl}(\delta, 1 - \delta)$$

Algorithm 3 Generic Top Two sampling rule

Input: Mechanism to choose the Leader arm $\mathcal{L}$, Mechanism to choose the Challenger arm $\mathcal{C}$, Mechanism to choose between the Leader and Challenger $\mathcal{T}$
Output: Next arm to sample a_t
Let $H_{t-1} \triangleq (a_1, r_1, \dots, a_{t-1}, r_{t-1})$ be the history
Choose $b_t = \mathcal{L}(H_{t-1}) \in [K]$ ▷ Choose the Leader
Choose $c_t = \mathcal{C}(H_{t-1}) \in [K]/\{b_t\}$ ▷ Choose the Challenger
Sample $a_t \in \{b_t, c_t\}$ using $\mathcal{T}$
Return a_t

where

$$T^\star_{\mathrm{KL}}(\boldsymbol{\nu})^{-1} \triangleq \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a \in [K]} \omega_a \mathrm{KL}(\nu_a, \lambda_a). \tag{2.20}$$

Here, kl *is the relative entropy between Bernoullis introduced in Definition 2.25. It is possible to show that* $\mathrm{kl}(\delta, 1-\delta) \sim_{\delta \to 0} \log(1/\delta)$ *and* $\mathrm{kl}(\delta, 1-\delta) \geq \log(1/(2.4\delta))$ *for all* $\delta \in (0,1)$. $\Sigma_K \triangleq \{\omega \in [0,1]^K \mid \sum_{a=1}^K \omega_a = 1\}$ *is the probability simplex, and the set of alternative instances is denoted by* $\mathrm{Alt}(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \in \mathcal{M} \mid a^\star(\boldsymbol{\lambda}) \neq a^\star(\boldsymbol{\nu})\}$, i.e. *the bandit instances* $\boldsymbol{\lambda}$ *with a different optimal arm than* $\boldsymbol{\nu}$.

The sample complexity lower bound of Theorem 2.32 shows that the $T^\star_{\mathrm{KL}}$ quantity, named the KL characteristic time, controls the complexity of the FC-BAI problem. $T^\star_{\mathrm{KL}}$ can be thought of as the FC-BAI counterpart of the KL inf quantity of regret. In general, the expression of $T^\star_{\mathrm{KL}}$ cannot be simply written as a sum over the arms of individual complexity terms. For Gaussian FC-BAI, with variance 1, it is possible to show that

$$\sum_{a=1}^K \frac{2}{\Delta_a^2} \leq T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \leq \sum_{a=1}^K \frac{4}{\Delta_a^2}$$

for $\mu_1 > \mu_2 \geq \dots \mu_K$, $\Delta_1 \triangleq \Delta_2$ and $\Delta_a \triangleq \mu_1 - \mu_a$.

In addition, it is possible to show that any FC-BAI strategy that matches the lower bound draws each arm with respect to the proportion dictated by $w^\star$ where

$$w^\star(\boldsymbol{\nu}) \triangleq \arg\max_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a \in [K]} \omega_a \mathrm{KL}(\nu_a, \lambda_a).$$

In the following, we show that this lower bound is asymptotically tight.

2.2.9 FC-BAI algorithms

Early FC-BAI algorithms failed to reach the lower bound of Theorem 2.32, e.g. Successive Elimination (SE) based algorithms [EDMMM06] or confidence bounds based algorithms, e.g. LUCB [KTAS12] or lil'UCB [JMNB14]. On the other hand, inspired by this lower bound, many

algorithms have been designed to tackle FC-BAI. The Track-and-Stop algorithm [GK16] is the first algorithm to reach asymptotic optimality, by sequentially solving the optimisation problem $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}_n)$, where $\boldsymbol{\nu}_n$ is the empirical instance, and tracking the associated optimal weights. To reduce the computational cost of Track-and-Stop, several asymptotically optimal algorithms have been proposed recently: online optimisation-based approach, e.g. game-based algorithm [DKM19] or Frank-Wolfe-based Sampling (FWS) [WTP21], and Top Two algorithms [Rus16]. While most algorithms can be modified to achieve privacy, this thesis focuses on the Top Two family of algorithms due to their great empirical performance and easy implementation.

Before presenting the sampling rules, we present the Generalised Likelihood Ratio (GLR) stopping rule and Empirical Best (EB) recommendation rules. To design an FC-BAI strategy, the EB recommendation and GLR rules are then complemented by choosing a sampling rule from the Top Two family of sampling rules.

The Empirical Best (EB) recommendation rule. If the algorithm stops at step t, the EB recommendation rule proposes as a final guess the arm with the highest empirical mean, *i.e.*

$$\hat{a}_t \in \arg\max_{a\in[K]} \hat{\mu}_a(t-1),$$

where we remind that $\hat{\mu}_a(t) \triangleq \frac{1}{N_a(t)} \sum_{s=1}^t \mathbb{1}\left(A_s = a\right) r_s$ is the empirical mean estimate of μ_a at step t, and $N_a(t) \triangleq \sum_{s=1}^t \mathbb{1}\left(A_s = a\right)$ is the number of pulls of arm a at step t.

The Generalised Likelihood Ratio (GLR) stopping rule. An algorithm decides to stop when enough statistical evidence has been collected to certify that the final recommendation (*i.e.* the EB) is the optimal arm, with probability $1-\delta$. This can be formulated as a sequential hypothesis test. Specifically, K sequential tests are run in parallel. Each test is associated with verifying the optimality of arm $a \in [K]$, and tries to distinguish between two hypotheses: H_0^a: arm a is sub-optimal for the bandit instance $\boldsymbol{\nu}$ vs H_1^a: arm a is optimal for the bandit instance $\boldsymbol{\nu}$. To solve each test, a Generalised Likelihood Ratio (GLR) statistic is computed. Specifically

$$\mathrm{GLR}_t(a) \triangleq \inf_{\boldsymbol{\lambda}\in\mathcal{M}:a^\star(\lambda)\neq a} \ell_t(\boldsymbol{\nu}_t, \boldsymbol{\lambda}),$$

where $\boldsymbol{\nu}_t$ is the empirical bandit instance at step t, and ℓ_t is the log-likelihood ratio of the of the rewards collected before time t. A high value of $\mathrm{GLR}_t(a)$ indicates that the hypothesis H_0^a should be rejected, and that arm a is optimal. The GLR stopping rule stops as soon as one of these $\mathrm{GLR}_t(a)$ statistics is big enough. Specifically, the GLR stopping rule stops at step

$$\tau_\delta \triangleq \inf\left\{ t : \max_{a\in[K]} \mathrm{GLR}_t(a) > c(t-1,\delta) \right\},$$

where $c(t, \delta)$ is a threshold, fine-tuned using time-uniform concentration inequalities to ensure δ-correctness.

For Gaussian bandits with unit variance, it is possible to show that the GLR can be simplified to $\min_{a \neq \hat{a}_t} W_t(\hat{a}_t, a)$, where $W_t(a, b)$ is the empirical transportation cost between arm a and arm b defined by

$$W_t(a, b) \triangleq \mathbb{1}\left(\hat{\mu}_t(a) > \hat{\mu}_t(b)\right) \frac{\left(\hat{\mu}_t(a) - \hat{\mu}_t(b)\right)^2}{2\left(1/N_a(t) + 1/N_b(t)\right)}.$$

The stopping rule is then

$$\tau_\delta \triangleq \inf\left\{t : \min_{a \neq \hat{a}_t} W_t(\hat{a}_t, a) > c(t-1, \delta)\right\},$$

and the threshold $c(t, \delta)$ is asymptotically in $\log(1/\delta)$ as $\delta \to 0$. The power of the GLR rule ensures δ-correctness, independently from the sampling rule.

The Top Two family of sampling rules. At every step, a Top Two sampling rule samples the arm a_t between two candidates: a leader arm b_t and a challenger arm c_t. The generic top two sampling rule is presented in Algorithm 3. In recent years, numerous variants of Top Two algorithms have been analysed and shown to be asymptotically optimal [Rus16, QKR17, SHM+20, JDB+22, YQWY23, JDK24]. We refer to Chapter 2 in [Jou24] for a extensive review of different possible ingredients to instantiate Algorithm 3 while achieving asymptotic optimality.

For the sake of simplicity, we only consider in this thesis one particular instance of the generic top two sampling rule, called the TTUCB algorithm [JD24].

The TTUCB algorithm uses the following ingredients:

The leader. TTUCB chooses a UCB leader

$$b_t \triangleq \arg\max_{a \in K} \mathrm{UCB}_a(t).$$

The challenger. TTUCB chooses a Transportation Cost (TC) challenger

$$c_t \triangleq \arg\min_{a \neq b_t} W_t(b_t, a).$$

Choosing between the leader and challenger. For a hyper-parameter $\beta \in [0, 1]$ called the target allocation, TTUCB uses a Tracking approach to choose between the leader and challenger. Let $N_{b,a}(t)$ denote the number of times arm b was pulled when a was the leader up to step t, and $L_a(t)$ denotes the number of times arm a was the leader up to step t. In order to select the next arm to sample a_t, TTUCB sets $a_t = b_t$ if $N_{b_t,b_t}(t) \leq \beta L_{b_t}(t+1)$, else $a_t = c_t$. This tracking ensures that the optimal arm is sampled β-fraction of the time.

For Gaussian bandits with unit variance, the FC-BAI strategy combining an EB recommendation rule, with a GLR with $c(t,\delta) \sim_{\delta\to 0} \log(1/\delta)$ and the TTUCB sampling rule achieves

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star_\beta(\boldsymbol{\nu}),$$

where $\beta \in [0,1]$ is the target allocation, and

$$T^\star_\beta(\boldsymbol{\nu})^{-1} \triangleq \sup_{\omega\in\Sigma_K : \omega_{a^\star}=\beta} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a\in[K]} \omega_a \mathrm{KL}(\nu_a, \lambda_a).$$

The characteristic time $T^\star_\beta$ can be shown to be the lower bound of the sample complexity over the sub-class of FC-BAI strategies which allocate β fraction of their samples to the optimal arm. This quantity is closely related to $T^\star_{\mathrm{KL}}$ as the latter can be expressed as

$$T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) = \min_{\beta\in[0,1]} T^\star_\beta(\boldsymbol{\nu}).$$

For $\beta = 1/2$, we also have the "worst-case" inequality $T^\star_{1/2}(\boldsymbol{\nu}) \leq 2T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$. This shows that, for a fixed $\beta = 1/2$, TTUCB achieves the lower bound $T^\star_{\mathrm{KL}}$ up to a constant 2. It is possible to improve on this 2 constant, and achieve the lower bound exactly. The idea is to track a clever choice of the target allocation β, which depends on the round t and the identity of the leader and challenger, *i.e.* $\beta_t(b_t, c_t)$. To track the allocation $\beta_t(b_t, c_t)$, the tracking procedure of TTUCB also needs to be adapted. We refer the curious reader to Chapter 2 of [Jou24] for the exact expressions of the target allocation $\beta_t(b_t, c_t)$ and the tracking procedures. For these two changes, the TTUCB algorithm is then shown to achieve

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star_{\mathrm{KL}}(\boldsymbol{\nu}).$$

2.3 Membership Inference Games and Privacy Auditing

In this section, we formalise the Membership Inference (MI) game, based on the formalism introduced by [YGFJ18]. Then, we instantiate the MI game with the empirical mean mechanism, and present two strategies to trace the empirical mean: the likelihood ratio (LR) test [SOJH09] and the scalar product attack [DSS+15]. We also explore the effects of Differential Privacy (DP) on controlling the power of an adversary in MI games. Finally, we show how these effects of DP on MI games can be utilised in privacy audit procedures.

Algorithm 4 The Crafter

Input: Mechanism $\mathcal{M}$, Data-generating distribution $\mathcal{D}$, Number of samples n
Output: $(z^\star, o, b)$, where $z^\star \in \mathcal{Z}$, $o \in \mathcal{O}$ and $b \in \{0, 1\}$
Build a dataset $D \sim \bigotimes_{i=1}^{n} \mathcal{D}$
Sample $b \sim \text{Bernoulli}\left(\frac{1}{2}\right)$
if $b = 0$ **then**
    Sample $z^\star \sim \mathcal{D}$ independent of D
else
    Sample $i \sim \mathcal{U}[n]$
    Assign $z^\star$ to be the i-th element of D, *i.e.* $D_i \leftarrow z^\star$
end if
Sample $o \sim \mathcal{M}(D)$
Return $(z^\star, o, b)$

2.3.1 Definition of a MI game and threat model

Let $\mathcal{M}$ be a randomised mechanism that takes as input a dataset D of n points each belonging to $\mathcal{Z}$ and outputs $o \in \mathcal{O}$. In a Membership Inference (MI) game, an adversary attempts to infer whether a given target point $z^\star \in \mathcal{Z}$ was included in the input dataset of $\mathcal{M}$. Given access to an output $o \sim \mathcal{M}(D)$, the adversary tries to infer whether $z^\star \in D$, where D is the input dataset that generated the output o.

The MI game is presented in Algorithm 5. It is a game between two entities: the crafter (Algorithm 4) and the adversary $\mathcal{A}$. The MI game runs in multiple rounds. At each round t, the crafter samples a tuple $(z_t^\star, o_t, b_t)$, where $z_t^\star$ is a target point, o_t is an output of the mechanism and b_t is the secret binary membership of $z^\star$. To generate the tuple $(z_t^\star, o_t, b_t)$ at step t, the crafter(Algorithm 4) takes as input the mechanism $\mathcal{M}$, the data-generating distribution $\mathcal{D}$, the number of samples n and a target datum $z^\star$. The crafter starts by sampling a dataset $D = \{z_1, \ldots, z_n\}$ of size n independently from the data-generating distribution. Then, the crafter flips a fair coin $b_t \sim \text{Bern}(1/2)$. If $b_t = 0$, the crafter samples a target $z_t^\star \sim \mathcal{D}$ independent from the input dataset. This corresponds to the case where $z_t^\star$ is not included in the input dataset, and thus the output o_t is completely independent from $z_t^\star$. Otherwise if $b_t = 1$, $z_t^\star$ is included at a random position i in the dataset D, then the output $o_t \sim \mathcal{M}(D \cup \{z_t^\star\} \setminus \{z_i\})$ depends on $z_t^\star$. Thus, the secret binary $b_t \in \{0, 1\}$ encodes the membership of $z_t^\star$ in the input dataset, which generated o_t, *i.e.* $b_t = 0$ corresponds to the OUT case and $b_t = 1$ corresponds to the IN case. Then, at each step t of the game, the adversary $\mathcal{A}$ takes as input only $(z_t^\star, o_t)$ and outputs $\hat{b}_t$, trying to reconstruct b_t.

The MI game can also be seen as a hypothesis test. Here, the adversary tries to test the hypothesis "H_0: *The output o observed was generated from a dataset sampled i.i.d. from $\mathcal{D}$*", *i.e.* $b = 0$, versus "H_1: *The target point $z^\star$ was included in the input dataset producing the output o*", *i.e.* $b = 1$.

Algorithm 5 The Membership Inference (MI) game

Input: Mechanism $\mathcal{M}$, Data-generating distribution $\mathcal{D}$, Number of samples n, Adversary $\mathcal{A}$, Rounds T
Output: A list $L \in \{0,1\}^T$, where $L_t = 1$ if the adversary succeeds at step t.
Initialise a empty list L of length T
for t = 1, . . . , T **do**
    Sample $(z_t^\star, o_t, b_t) \sim$ Crafter, *i.e.* Algorithm 4 with inputs ($\mathcal{M}$, $\mathcal{D}$, n)
    Sample $\hat{b}_t \sim \mathcal{A}(z_t^\star, o_t)$
    Set $L_t \leftarrow \mathbb{1}\left(b_t = \hat{b}_t\right)$
end for
Return L

We denote by $p_{\text{out}}(o, z^\star)$ and $p_{\text{in}}(o, z^\star)$ the joint distributions of the pair output-target $(o, z^\star)$ under H_0 and H_1, respectively.

2.3.2 Performance metrics in an MI game

An adversary $\mathcal{A}$ is a (possibly randomised) algorithm that takes as input the pair $(z^\star, o)$ generated by the crafter (Algorithm 4) and outputs a guess $\hat{b} \sim \mathcal{A}(z^\star, o)$ trying to infer b. The adversary wins if $\hat{b} = b$ and loses otherwise. The performance of $\mathcal{A}$ can be assessed either with aggregated metrics like the accuracy and the advantage, or with test-based metrics like Type I error, Type II error, and trade-off functions.

The accuracy of $\mathcal{A}$ is defined as

$$\text{Acc}_n(\mathcal{A}) \triangleq \Pr[\mathcal{A}(z^\star, o) = b], \tag{2.21}$$

where the probability is over the generation of $(z^\star, o, b)$ using Algorithm 4 with input $(\mathcal{M}, \mathcal{D}, n)$.

The advantage of an adversary is the re-centred accuracy

$$\text{Adv}_n(\mathcal{A}) \triangleq 2\text{Acc}_n(\mathcal{A}) - 1. \tag{2.22}$$

We can also define two errors from the hypothesis testing formulation. *The Type I error*, also called the False Positive Rate, is

$$\alpha_n(\mathcal{A}) \triangleq \Pr\left[\mathcal{A}(z^\star, o) = 1 \mid b = 0\right]. \tag{2.23}$$

The Type II error, also called the False Negative Rate, is

$$\beta_n(\mathcal{A}) \triangleq \Pr\left[\mathcal{A}(z^\star, o) = 0 \mid b = 1\right]. \tag{2.24}$$

The power of the test is $1 - \beta_n(\mathcal{A})$. Using the prior $b \sim \mathrm{Bern}(1/2)$, we can show that

$$\mathrm{Adv}_n(\mathcal{A}) = 1 - \alpha_n(\mathcal{A}) - \beta_n(\mathcal{A}).$$

The advantage is a quantity always between $[-1, 1]$. The random adversary guesser, *i.e.* the adversary $\mathcal{A}_{\mathrm{rd}} \sim \mathrm{Bern}(1/2)$ that guesses 0 or 1 with probability $1/2$ oblivious from the game, has an advantage of 0, *i.e.* $\mathrm{Adv}_n(\mathcal{A}_{\mathrm{rd}})$. This means that the advantage of an adversary measures how much better the adversary is doing compared to the random guesser. If $\mathrm{Adv}_n(\mathcal{A}) > 0$, then the adversary $\mathcal{A}$ is better than random guessing.

An adversary can use a threshold over a score function to conduct the MI games, *i.e.* for $\mathcal{A}_{s,\tau}(z^\star, o) \triangleq \mathbb{1}\left(s(o; z^\star) > \tau\right)$ where s is a scoring function and τ is a threshold. We want to design scores that maximise the power under a fixed significance level α, *i.e.*

$$\mathrm{Pow}_n(s, \alpha) \triangleq \max_{\tau \in T_\alpha} 1 - \beta_n(\mathcal{A}_{s,\tau}), \tag{2.25}$$

where $T_\alpha \triangleq \{\tau \in \mathbb{R} : \alpha_n(\mathcal{A}_{s,\tau}) \leq \alpha\}$. $\mathrm{Pow}_n(s, \alpha)$ is also called a trade-off function.

2.3.3 The Neyman-Pearson lemma and optimal MI adversaries

It is a fundamental result of statistics that given two data-generating distributions p_0 and p_1 under hypotheses H_0 and H_1, respectively; no test can achieve better power than the Likelihood Ratio (LR) test, *i.e.* The Neyman Pearson Lemma [NP33].

By recalling the hypothesis testing formulation of the MI game, where $p_n^{\mathrm{out}}(o, z^\star)$ is the distribution of the pair output-target $(o, z^\star)$ under H_0 and $p_n^{\mathrm{in}}(o, z^\star)$ is the distribution of the pair output-target $(o, z^\star)$ under H_1. Then, the *log-Likelihood Ratio* (LR) score for the MI game is

$$\ell_n(o, z^\star) \triangleq \log\left(\frac{p_n^{\mathrm{in}}(o, z^\star)}{p_n^{\mathrm{out}}(o, z^\star)}\right).$$

A direct consequence of the Neaman-Pearson lemma is that the LR score ℓ_n maximises the power under significance α for every $\alpha \in (0, 1)$.

2.3.4 The Likelihood Ratio test for Bernoulli empirical mean MI games

In this section, we revisit results from [SOJH09]. In [SOJH09], the MI game is instantiated with the empirical mean mechanism denoted by $\mathcal{M}_n^{\mathrm{emp}}$. The mechanism $\mathcal{M}_n^{\mathrm{emp}}$ takes as input a dataset of size n of d-dimensional points, *i.e.* $D = \{Z_1, \ldots, Z_n\} \in (\mathbb{R}^d)^n$, and outputs the exact empirical mean $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n Z_i \in \mathbb{R}^d$.

Assumptions on the data generating distribution and asymptotic regime. [SOJH09] supposes that the data-generating distribution $\mathcal{D}$ is column-wise independent Bernoulli distributions, *i.e.* $\mathcal{D} \triangleq \bigotimes_{j=1}^{d} \text{Bernoulli}(\mu_j)$, with $\mu_j \in [a, 1-a]$ for some $a \in (0, 1/2)$. We denote by $\rightsquigarrow$ convergence in distribution, *i.e.* A sequence of random variables $X_n \rightsquigarrow X$ if and only if $\Pr(X_n \leq x) \to \Pr(X \leq x)$ for all x. Let Φ represent the Cumulative Distribution Function (CDF) of the standard normal distribution, *i.e.* $\Phi(\alpha) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-t^2/2} \, \mathrm{d}t$ for $\alpha \in \mathbb{R}$. [SOJH09] studies the *asymptotic behaviour of the LR test*, when both the sample size n and the dimension d tend to infinity such that $d/n = \tau > 0$.

The analysis of [SOJH09] starts by showing that the exact formula of the LR score, at output $o = \hat{\mu}_n$ and target $z^\star$ is

$$\ell_n(\hat{\mu}_n, z^\star) = \sum_{j=1}^{d} z_j^\star \log\left(\frac{\hat{\mu}_{n,j}}{\mu_j}\right) + (1 - z_j^\star) \log\left(\frac{1-\hat{\mu}_{n,j}}{1-\mu_j}\right). \tag{2.26}$$

As d and n tend to infinity such that $d/n = \tau$, [SOJH09] shows that the LR score converges in distribution to

$$\ell_n(\hat{\mu}_n, z^\star) \rightsquigarrow^{H_0} \mathcal{N}\left(-\frac{1}{2}\tau, \tau\right)$$

under H_0 and converges to

$$\ell_n(\hat{\mu}_n, z^\star) \rightsquigarrow^{H_1} \mathcal{N}\left(\frac{1}{2}\tau, \tau\right)$$

under H_1.

The asymptotic distribution of the LR score helps to provide the asymptotic trade-off of the optimal LR attacker. Specifically, the main result (Section T2.1 in [SOJH09]) is that

$$\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta) \approx \sqrt{d/n}$$

where α is the Type I error, β is the Type II error and Φ represents the Cumulative Distribution Function (CDF) of the standard normal distribution, i.e. $\Phi(\alpha) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-t^2/2} \, \mathrm{d}t$ for $\alpha \in \mathbb{R}$. This trade-off between α and β shows that the MI game gets easier, as d/n gets bigger.

The LR score of Equation (2.26) assumes the knowledge of the real mean μ. To derive a "realistic" adversary for this attack, [SOJH09] proposes to estimate μ using reference samples $D_{n_0}^{\text{ref}} \triangleq \{Z_1^{\text{ref}}, \ldots, Z_{n_0}^{\text{ref}}\}$ sampled independently from the input dataset. Using the reference samples, $\hat{\mu}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} Z_i^{\text{ref}}$ is estimated and plugged in the LR score of Equation (2.26). This leads to the empirical LR score

$$\ell_n^{\text{emp}}(\hat{\mu}_n, z^\star; D_{n_0}^{\text{ref}}) \triangleq \sum_{j=1}^{d} z_j^\star \log\left(\frac{\hat{\mu}_{n,j}}{\hat{\mu}_{0,j}}\right) + (1 - z_j^\star) \log\left(\frac{1-\hat{\mu}_{n,j}}{1-\hat{\mu}_{0,j}}\right). \tag{2.27}$$

If the number of reference points is $n_0 = \lambda n$, then [SOJH09] shows that the number of samples used to get the same power as the optimal adversary who knows the real means is increased by a factor $(1+\lambda)/\lambda$.

2.3.5 The scalar product score for MI games

[DSS+15] proposes a scalar product attack for tracing the empirical mean that thresholds over the score

$$s^{\text{scal}}(\hat{\mu}_n, z^\star; z^{\text{ref}}) \triangleq (z^\star - z^{\text{ref}})^T \hat{\mu}_n,$$

where z^{ref} is one reference point. The intuition behind this attack is to compare the target-output correlation $(z^\star)^T\hat{\mu}_n$ with a reference-output correlation $(z^{\text{ref}})^T\hat{\mu}_n$. The analysis of [DSS+15] shows that with only one reference point $z^{\text{ref}} \sim \mathcal{D}$, and even for noisy estimates of the mean, the attack is able to trace the data of some individuals in the regime $d \sim n^2$.

Informally, the analysis of [DSS+15] considers a scalar product attack, taking as input any $1/2$-*accurate estimate* $\hat{\mu}$ of a dataset $D \in (\{-1,1\}^d)^n$ of dimension $d = O\left(n^2 \log(1/\delta)\right)$ *i.e.* the estimate $\hat{\mu}$ is close to the empirical mean of the dataset up to $1/2$, a target point $z^\star \in \{-1,1\}^d$ and *only a single reference sample* $z^{\text{ref}} \in \{-1,1\}^d$. The data-generating distribution is assumed to be chosen from a strong class of distributions $\mathcal{P}$. Then,

- If $z^\star$ is IN the dataset D, then

$$\Pr\left\{s^{\text{scal}}(\hat{\mu}_n, z^\star; z^{\text{ref}}) > \tau\right\} \geq \Omega(1/n).$$

- If $z^\star$ is Out of the dataset D, then

$$\Pr\left\{s^{\text{scal}}(\hat{\mu}_n, z^\star; z^{\text{ref}}) < \tau\right\} \geq 1 - \delta.$$

for a carefully chosen threshold $\tau = O(\sqrt{d\log(1/\delta)})$.

The condition of $1/2$-accuracy is a weak condition, compared to the exact empirical mean attack [SOJH09]. The price of the weak notion of accuracy is that the attack is only guaranteed for $d \gtrapprox n^2$, whereas the exact attack of [SOJH09] is able to trace for $d \approx n$.

For more accurate mechanisms, and a larger number of reference samples, [DSS+15] also shows that it is possible to trace for intermediate values of d. Specifically, if the mechanism is α-accurate for $\alpha \geq n^{-1/2}$, and with $O(1/\alpha^2)$ reference samples, the attack is able to trace for $d = O(\alpha^2 n^2)$.

2.3.6 The effect of DP on the performance metrics of MI games

In this section, we present some results that explore the consequences of Differential Privacy on the power of the optimal attacker.

Effect on the advantage. First, [YGFJ18] shows that, if $\mathcal{M}$ is ε-pure DP, then

$$\mathrm{Adv}_n(\mathcal{A}) \leq e^{\varepsilon} - 1,$$

for any adversary $\mathcal{A}$, and any n number of samples, and any data-generating Distribution $\mathcal{D}$.

Then, [EMRS19] generalises this bound for (ε, δ)-DP mechanisms, and shows that

$$\mathrm{Adv}_n(\mathcal{A}) \leq 1 - e^{-\varepsilon}(1 - \delta),$$

for any adversary $\mathcal{A}$, and any n number of samples, and any data-generating Distribution $\mathcal{D}$. This bound has the advantage of being always smaller than 1.

Finally, [HOT+23] improve on this bound, and shows that for any (ε, δ)-DP

$$\mathrm{Adv}_n(\mathcal{A}) \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1},$$

for any adversary $\mathcal{A}$, and any n number of samples, and any data-generating Distribution $\mathcal{D}$. This bound is tighter than [EMRS19]'s bound, for any $\varepsilon \geq 0$ and any $\delta \in [0, 1]$.

The proof of the three upper bounds on the advantage is based on the hypothesis formulation of DP of Theorem 2.4. Specifically, Theorem 2.4 implies that, for any adversary $\mathcal{A}$, we have that

$$\alpha_n(\mathcal{A}) + e^{\varepsilon}\beta_n(\mathcal{A}) \geq 1 - \delta, \quad \text{and,} \tag{2.28}$$

$$e^{\varepsilon}\alpha_n(\mathcal{A}) + \beta_n(\mathcal{A}) \geq 1 - \delta. \tag{2.29}$$

Since $\mathrm{Adv}_n(\mathcal{A}) = 1 - \alpha_n(\mathcal{A}) - \beta_n(\mathcal{A})$, playing with the two inequalities above gives the three advantage upper bounds.

Effect on the Trade-off functions. The DP constraint not only upper bounds the advantage of an adversary in the MI game, but restricts the whole trade-off function. Specifically, for any adversary $\mathcal{A}_{s,\tau}$ that thresholds over any score function s, then if the mechanism $\mathcal{M}$ is (ε, δ)-DP, we have that

$$\mathrm{Pow}_n(s, \alpha) \leq 1 - f_{\varepsilon,\delta}(\alpha),$$

where

$$f_{\varepsilon,\delta}(\alpha) \triangleq \max\left\{0, 1 - \delta - e^{\varepsilon}\alpha, e^{-\varepsilon}(1 - \delta - \alpha)\right\}. \tag{2.30}$$

This is a also direct consequence of Inequalities (2.28).

Remark 2.33 (f-DP [DRS19]). *It is possible to define a DP notion, based purely on trade-off functions. Let P and Q be two distributions on the same space. Define the trade-off function $T(P, Q) : [0, 1] \to [0, 1]$ as*

$$T(P, Q)(\alpha) = \inf\{\beta_\phi : \alpha_\phi \leq \alpha\},$$

*where $\phi \in [0, 1]$ is a rejection rule, $\alpha_\phi \triangleq \mathbb{E}_P[\phi]$ is the Type I error, $\beta_\phi \triangleq 1 - \mathbb{E}_Q[\phi]$ is the Type II error, and the infimum is taken over all (measurable) rejection rules. We remark that this trade-off function is defined as the infimum of Type II errors when Type I error is at most α, while the trade-off function definition we consider in Equation (2.25) of MI games metric is the maximum over the power (*i.e. *1 - Type II error) when the Type I is as most α.*

A mechanism $\mathcal{M}$ satisfies f-DP if, for all neighbouring $d \sim d'$

$$T(\mathcal{M}_d, \mathcal{M}_{d'}) \geq f.$$

This definition is parameterised by a function f, compared to real-valued parameters for (ε, δ)-DP or ρ-zCDP. Let P and Q be the distributions such that $f = T(P, Q)$. Then, informally, f-DP implies that distinguishing any two neighbouring datasets based on the released output is at least as difficult as distinguishing between P and Q based on a single draw.

The notion of f-DP can be seen as a generalisation of (ε, δ). In fact, a mechanism is (ε, δ)-DP if and only if it is $f_{\varepsilon,\delta}$-DP, where $f_{\varepsilon,\delta}$ is defined in Equation (2.30).

2.3.7 Privacy auditing

The goal of privacy auditing is to lower bound the privacy budgets of a mechanism.

For simplicity, let us first consider the case of auditing pure DP. A privacy audit procedure $\mathcal{U}$ has query access to the mechanism $\mathcal{M}$, *i.e.* can send as input a dataset d and observe an output $o \sim \mathcal{M}_d$. If the mechanism $\mathcal{M}$ is ε-DP, the privacy audit should output a guess ε_{low} that is a high probability lower bound on the real budget ε.

Definition 2.34 (γ-correct privacy audit for ε-pure DP). *Suppose that the mechanism $\mathcal{M}$ is ε-pure DP. Let $\gamma \in (0, 1)$ be the confidence parameter. Let $\mathcal{U}$ be a (possibly randomised) privacy audit procedure with query access to the mechanism $\mathcal{M}$. The audit procedure $\mathcal{U}$ outputs a guess $\mathcal{U}(\mathcal{M}) \triangleq \varepsilon_{\text{low}} \in \mathbb{R}^+$.*

The audit procedure $\mathcal{U}$ is said to be γ-correct if

$$\Pr(\mathcal{U}(\mathcal{M}) > \varepsilon) \leq \gamma.$$

where the probability is over the randomness of the mechanism $\mathcal{M}$ and the auditing procedure $\mathcal{U}$. For example, the randomness of the auditing procedure could be from sampling randomly input datasets D to send to the mechanism $\mathcal{M}$.

On the other hand, the utility of an audit is measured by the closeness of the lower bound to the real budget, in expectation over the randomness of $\mathcal{U}$ and $\mathcal{M}$ i.e. $\mathbb{E}[|\mathcal{U}(\mathcal{M}) - \varepsilon|]$.

In addition to the query access to the mechanism $\mathcal{M}$, which is a minimal requirement in auditing, and the confidence parameter γ, the audit procedure $\mathcal{U}$ can have additional inputs. For example, the audit $\mathcal{U}$ can be parameterised with the number of query interactions T with the mechanism $\mathcal{M}$, the size of the dataset n, a data-generating distribution $\mathcal{D}$ to generate the input datasets D to query $\mathcal{M}$.

Hypothesis test formulation. The γ-correctness framework in Definition 2.34 could be thought of as statistical estimation of ε in a frequentist way. A privacy audit could also be seen as a hypothesis test. The auditor starts with an initial guess ε_0 and tries to distinguish between

$$H_0\text{: The mechanism } \mathcal{M} \text{ satisfies } \varepsilon_0\text{-DP}$$
$$\text{vs}$$
$$H_1\text{: The mechanism } \mathcal{M} \text{ does not satisfy } \varepsilon_0\text{-DP.}$$

If the auditor rejects H_0 with confidence $1 - \gamma$, then providing as a guess $\varepsilon_{\text{low}} = \varepsilon_0$ is γ-correct. Otherwise, the auditor can increment the guess ε_0, formulate a new hypothesis test and try to reject it again.

The difference between the hypothesis testing and statistical estimation formulations is that: the hypothesis test starts with an initial guess ε_0 and outputs a binary decision to reject or not. On the other, an estimator outputs directly a number ε_{low}.

Standard recipe for privacy auditing. To obtain a lower bound on the privacy budget, a natural approach is to directly use the definition of DP. Typically, an audit based on the definition of DP would first construct a pair of neighbouring input datasets D and D', and an event E on the output space. Then, the audit estimates the probabilities $p_0 \triangleq \mathcal{M}_D(E)$ and $p_1 \triangleq \mathcal{M}_{D'}(E)$. If we have access to the real p_0 and p_1, a lower bound on the privacy budget can be set to $\varepsilon_{\text{low}} \triangleq \max\{\log(p_0/p_1), \log(p_1/p_0)\}$.

However, the auditor has only query access to the mechanism $\mathcal{M}$ and cannot observe the real probabilities p_0 and p_1. Thus, these probabilities should be estimated accurately. A natural approach for estimating p_0 and p_1 is using Monte Carlo estimation. Specifically, the probability $\mathcal{M}_D(E)$ for $D \in \{D, D'\}$ can be seen as the expectation of Bernoulli random variables $\mathbb{1}(o \in E)_{o \sim \mathcal{M}_D}$ over the randomness of the mechanism $\mathcal{M}$. Thus, by querying for T times mechanism $\mathcal{M}$ with the same input dataset D, and observing the outputs $o_1, \ldots, o_T$ from $\mathcal{M}_d$, p_0 can be estimated with the variable

$$\hat{p}_0^T \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(o_t \in E).$$

Similarly, observing the outputs $o'_1, \dots, o'_T$ from $\mathcal{M}_d$ gives the empirical estimate

$$\hat{p}_1^T \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(o'_t \in E\right)$$

for the real probability p_1. Finally, since the goal is to provide a high probability lower bound for the real quantity $\log(p_0/p_1)$, it is important to construct high probability upper and lower bounds on the estimators $\hat{p}_0$ and $\hat{p}_1$. Providing high confidence intervals for Bernoulli distributions is a well-studied problem with several off-the-shelf techniques, such as Clopper-Pearson intervals [CP34] or Wilson intervals [Wil27] among others. The standard confidence interval scales in $1/\sqrt{T}$ for T i.i.d Bernoulli samples. Specifically, there exists $C_0(\gamma)$ and $C_1(\gamma)$ independent from T such that, with probability $1-\gamma$ we get that

$$|\hat{p}_0^T - p_0| \leq \frac{C_0(\gamma)}{\sqrt{T}}$$
$$|\hat{p}_1^T - p_1| \leq \frac{C_1(\gamma)}{\sqrt{T}}.$$

Then, a γ-correct estimate of ε is

$$\varepsilon_{\text{low}} \triangleq \max\left\{0, \log\left(\frac{\hat{p}_0^T - \frac{C_0(\gamma)}{\sqrt{T}}}{\hat{p}_1^T + \frac{C_1(\gamma)}{\sqrt{T}}}\right), \log\left(\frac{\hat{p}_1^T - \frac{C_1(\gamma)}{\sqrt{T}}}{\hat{p}_0^T + \frac{C_0(\gamma)}{\sqrt{T}}}\right)\right\}.$$

The main ingredients left to specify in this recipe are:

(a) how to construct the input datasets D and D',

(b) how to choose the event E.

Using MI games as a proxy for constructing an event E. In order to estimate the privacy budget ε of a mechanism $\mathcal{M}$, it is possible for the auditor to run an MI game. First, the auditor estimates the performance metric of the game, for example, the Type I error α and Type II errors β. Then, by inverting the Inequalities (2.28), the auditor recovers the following lower bound on ε

$$\varepsilon \geq \max\left\{0, \log\left(\frac{1-\beta}{\alpha}\right), \log\left(\frac{1-\alpha}{\beta}\right)\right\}.$$

Since the auditor does not have access directly to the type I and type II errors to compute this lower bound, the auditor estimates these quantities using Monte Carlo estimation. Specifically, the MI game is run for T independent rounds. Then, the empirical estimate $\hat{\alpha}_T$ is defined to be the empirical proportion of rounds when the adversary guesses $\hat{b} = 1$ when $b = 0$. Similarly, $\hat{\beta}_T$ is defined to be the empirical proportion of rounds when the adversary guesses $\hat{b} = 0$ when $b = 1$. Again, to have a high probability lower bound on ε, we need to have high probability

upper and lower bounds on α and β using off-the-shelf techniques for Bernoulli variables, such as Clopper-Pearson, Bernstein, etc. Again, let $C'_0(\gamma)$ and $C'_1(\gamma)$ two constants independent from T such that we get

$$|\hat{\alpha}_T - \alpha| \leq \frac{C'_0(\gamma)}{\sqrt{T}},$$
$$|\hat{\beta}_T - \beta| \leq \frac{C'_1(\gamma)}{\sqrt{T}},$$

with probability $1 - \gamma$.

Then, a γ-correct estimate of ε is

$$\varepsilon_{\text{low}}^{\text{MI}} \triangleq \max\left\{0, \log\left(\frac{1 - \hat{\beta}_T - \frac{C'_1(\gamma)}{\sqrt{T}}}{\hat{\alpha}_T + \frac{C'_0(\gamma)}{\sqrt{T}}}\right), \log\left(\frac{1 - \hat{\alpha}_T - \frac{C'_0(\gamma)}{\sqrt{T}}}{\hat{\beta}_T + \frac{C'_1(\gamma)}{\sqrt{T}}}\right)\right\}.$$

Compared to the standard recipe, the MI game, in a sense, acts as a proxy for designing the event E. Designing an event E is replaced by designing a strong adversary $\mathcal{A}$ in the MI game audit procedure. In this sense, using MI games for audit is simpler, since the audit can directly benefit from stronger adversaries from the MI literature.

However, to audit using MI games, one still needs to construct input datasets D and D'. If the MI game used for audit is the game presented in Algorithm 5, the auditor only needs to design a data-generating distribution, rather than datasets D and D'. Thus, the audit should optimise for the data-generating distribution that minimises the Type I and Type II errors, to get tighter privacy budget estimations.

On the other hand, privacy is a worst-case guarantee. A tight estimate of the privacy budget should estimate the log-likelihood over the worst neighbouring datasets D and D', *i.e.* $\max_{D\sim D',E} \log\left(\frac{\mathcal{M}_D(E)}{\mathcal{M}_{D'}(E)}\right)$. Motivated by this intuition, other versions of the MI game can be proposed. In the game presented in Algorithm 5, there are two additional sources of randomness that could be fixed. First, the initial dataset is generated by sampling n i.i.d points from the data-generating distribution. A first variant of the games could be to have a fixed initial distribution. Then, the choice of the target point $z^\star$ is also stochastically generated in Algorithm 5. A second variant of the game can be played with a fixed target point.

In Chapter 6, we study the variant of Algorithm 5 where the target point is fixed. This means that the metrics of the MI game are target-dependent. For a fixed adversary, it is then possible to optimise for the target point that is easiest to attack, *i.e.* has the lowest Type I/ Type II errors. This target point is called the canary and is a crucial ingredient in auditing. We will characterise the optimal canary selection strategy for auditing the empirical mean in Chapter 6, with an application to auditing machine learning algorithms in Chapter 7.

Auditing (ε, δ)-DP. It is possible to formulate the auditing problem for (ε, δ)-DP in two ways. The first natural way is to ask the auditor to output two estimates ε_{low} and δ_{low}, such that with high probability, these two estimates are lower bounds on the real budgets ε and δ.

Equivalently, we can think of a mechanism $\mathcal{M}$ as achieving an infinite set of condition $(\varepsilon(\delta), \delta)$-DP constraints. Then, the audit procedure can be formulated as follows: for every input arameter δ, the auditor outputs a high probability lower bound on $\varepsilon(\delta)$.

Definition 2.35 (γ-correct (ε, δ)-DP privacy audit procedure)**.** *Suppose that the mechanism $\mathcal{M}$ is $(\varepsilon(\delta), \delta)$-DP, for all $\delta \in (0, 1)$. Let $\gamma \in (0, 1)$ be the confidence parameter. Let $\mathcal{U}$ be a (possibly randomised) privacy audit procedure with query access to the mechanism $\mathcal{M}$. The audit procedure $\mathcal{U}$ outputs a guess $\mathcal{U}(\mathcal{M}, \delta) \triangleq \varepsilon_{\text{low}}(\delta) \in \mathbb{R}^+$.*

The audit procedure $\mathcal{U}$ is said to be γ-correct if, for all $\delta \in (0, 1)$,

$$\Pr(\mathcal{U}(\mathcal{M}, \delta) > \varepsilon(\delta)) \leq \gamma.$$

where the probability is over the randomness of the mechanism $\mathcal{M}$ and the auditing procedure $\mathcal{U}$.

Again, the utility of an audit is measured by the closeness of the lower bound to the real budget, in expectation over the randomness of $\mathcal{U}$ and $\mathcal{M}$ i.e. $\mathbb{E}\left[|\mathcal{U}(\mathcal{M}, \delta) - \varepsilon(\delta)|\right]$*, for each $\delta \in (0, 1)$.*

It is possible to then adapt the equations of the definition-based audit and MI games to correct for δ. The equations become

$$\varepsilon_{\text{low}}(\delta) \triangleq \max\left\{0, \log\left(\frac{\hat{p}_0^T - \delta - \frac{C_0(\gamma)}{\sqrt{T}}}{\hat{p}_1^T + \frac{C_1(\gamma)}{\sqrt{T}}}\right), \log\left(\frac{\hat{p}_1^T - \delta - \frac{C_1(\gamma)}{\sqrt{T}}}{\hat{p}_0^T + \frac{C_0(\gamma)}{\sqrt{T}}}\right)\right\},$$

$$\varepsilon_{\text{low}}^{\text{MI}}(\delta) \triangleq \max\left\{0, \log\left(\frac{1 - \delta - \hat{\beta}_T - \frac{C_1'(\gamma)}{\sqrt{T}}}{\hat{\alpha}_T + \frac{C_0'(\gamma)}{\sqrt{T}}}\right), \log\left(\frac{1 - \delta - \hat{\alpha}_T - \frac{C_0'(\gamma)}{\sqrt{T}}}{\hat{\beta}_T + \frac{C_1'(\gamma)}{\sqrt{T}}}\right)\right\}.$$

Threat models of privacy auditing in supervised machine learning. In supervised machine learning, the mechanism to be audited is a learning algorithm that takes as input a dataset D and outputs a machine learning model $o \triangleq f$. The dataset D is composed of n tuples (x_i, y_i) where x_i is a feature and y_i is a label, *i.e.* $D \triangleq \{(x_1, y_1), \ldots, (x_n, y_n)\}$. The machine learning model f produced can then be queried for an input feature x to get a label $y = f(x)$. The model f is generally found by minimising over a class of models $\mathcal{F}$ some type of error ℓ in the input dataset D, *i.e.* $f \triangleq \arg\min_{g \in \mathcal{F}} \ell(g, D)$.

The class of models $\mathcal{F}$ can be parameterised by $\theta \in \mathbb{R}^d$, *i.e.* $f = f_\theta$. In this case, the threat model for auditing depends on whether the auditor has access to the parameter θ or only query access to the model f_θ. The setting where the auditor can observe the parameter $\theta \in \mathbb{R}^d$ is called

the **white-box setting**. On the other hand, when the auditor can only query the final model f_θ, *i.e.* send input features x to f and observe the outputs $y = f(x)$ is called the **black-box setting**.

In the parameterised setting, the quintessential training algorithms are based on Gradient Descent. The Gradient Descent algorithm start with an initial parameter $\theta_0 \in \mathbb{R}^d$, and then updates sequentially the parameter at each step t by $\theta_t \triangleq \theta_{t-1} - \eta \nabla_{\theta_{t-1}} \ell(\theta_{t-1}, d)$. In white-box auditing, the auditor may have access to only the final parameter θ_T, and we call this setting **white-box final parameter**. The auditor can have access to all (or a subset of) the intermediate parameters sequence $(\theta_0, \ldots, \theta_T)$, and we call this setting **white-box federated learning** setting [MSS22].

In Chapter 7, we discuss how analysing the fixed-target MI game for the empirical mean mechanism can directly provide an adversary and a canary design strategy to audit gradient descent algorithms in the white box federated learning setting.

2.4 Asymptotic Statistics

In this section, we present some classic results from asymptotic statics, used later in Chapter 6 to analyse the asymptotic distribution of the Likelihood Ratio (LR) test in Membership Inference (MI) games.

2.4.1 Stochastic convergence and basic properties

A sequence of random variables X_n is said to converge in distribution to a random variable X, *i.e.* $X_n \rightsquigarrow_n X$ if $\Pr(X_n \leq x) \to \Pr(X \leq x)$, for every x at which the limit distribution $x \to \Pr(X \leq x)$ is continuous.

A sequence of random variables X_n is said to converge in probability to X if for every $\varepsilon > 0$, $\Pr(\|X_n - X\| > \varepsilon) \to 0$, denoted by $X_n \to^{\mathrm{P}} X$.

A sequence of random variable (X_n) is called uniformly tight if: for every ε, $\exists M > 0$, such that $\sup_n \Pr(\|X_n\| > M) < \varepsilon$.

We recall that continuous mappings preserve both convergences.

Theorem 2.36 (Continuous mappings preserve stochastic convergence)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a continuous function at every point of a set C such that $P(X \in C) = 1$.*

(a) If $X_n \rightsquigarrow_n X$, then $g(X_n) \rightsquigarrow_n g(X)$,

(b) If $X_n \to^{\mathrm{P}} X$, then $g(X_n) \to^{\mathrm{P}} g(X)$,

Next, Prohorov's theorem provides a link between convergence in distribution and being uniformly tight.

Theorem 2.37 (Prohorov's theorem). *Let X_n be a random vector in $\mathbb{R}^d$.*

1. *If $X_n \rightsquigarrow X$, for some X, then the sequence (X_n) is uniformly tight;*
2. *If (X_n) is uniformly tight, then there exits a sub-sequence with $X_{n_j} \rightsquigarrow X$ as $j \to \infty$ for some X.*

We also recall the stochastic o_p and O_p notation for random variables.

Definition 2.38 (Stochastic o_p and O_p). *We say that $X_n = o_p(R_n)$ if $X_n = Y_n R_n$ and $Y_n \to^{\mathrm{P}} 0$*

We say that $X_n = O_p(R_n)$ if $X_n = Y_n R_n$ and $Y_n = O_p(1)$ where $O_p(1)$ denotes a sequence that is uniformly tight (also called bounded in probability).

The following lemma is used to get Taylor expansions of random variables.

Lemma 2.39 (Lemma 2.12 in [VdV00]). *Let R be a function on $\mathbb{R}^k$, such that $R(0) = 0$. Let $X_n = o_p(1)$.*

Then, for every $p > 0$,

(a) if $R(h) = o(\|h\|^p)$ as $h \to 0$, then $R(X_n) = o_p(\|X_n\|^p)$;

(b) if $R(h) = O(\|h\|^p)$ as $h \to 0$, then $R(X_n) = O_p(\|X_n\|^p)$.

2.4.2 The Lindeberg-Feller central limit theorem

The Lindeberg-Feller theorem is the simplest extension of the classical central limit theorem (CLT) and is applicable to independent but not necessarily identically distributed random variables with finite variances.

Theorem 2.40 (Lindeberg-Feller CLT). *Let $Y_{n,1}, \ldots Y_{n,d_n}$ be independent random vectors with finite variances such that*

1. *for every $\varepsilon > 0$, $\sum_{j=1}^{d_n} \mathbb{E}\left[\|Y_{n,i}\|^2 \mathbb{1}\left(\|Y_{n,i}\| > \varepsilon\right)\right] \to 0$,*
2. *$\sum_{j=1}^{d_n} \mathbb{E}\left[Y_{n,i}\right] \to \mu$,*
3. *$\sum_{j=1}^{d_n} \mathrm{Cov}\left[Y_{n,i}\right] \to \Sigma$.*

Then

$$\sum_{j=1}^{d_n} Y_{n,j} \rightsquigarrow \mathcal{N}\left(\mu, \Sigma\right).$$

2.4.3 The Edgeworth asymptotic expansions

Finally, the last result from asymptotic statistics is the Edgeworth asymptotic expansion in the CLT.

Theorem 2.41 (Edgeworth expansion, Theorem 15 of Chapter 7 in [Pet12]). *Let $Z_1, \ldots, Z_n$ sampled i.i.d from $\mathcal{D}$, where $\mathcal{D}$ has a finite absolute moment of k-th order,* i.e. $\mathbb{E}[|X_1|^k] < \infty$. *Let d_n be the density of the centred normalised mean $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i$, then*

$$d_n(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \sum_{\nu=1}^{k-2} \frac{q_\nu(x)}{n^{\nu/2}} + o\left(\frac{1}{n^{(k-2)/2}}\right)$$

uniformly in x, where $q_v(x)$ are related to the Chebyshev-Hermite Polynomials H_k. Specifically,

$$q_\nu(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \sum_{(k_1,k_2,\ldots,k_\nu)\in\mathcal{C}} H_{\nu+2s}(x) \prod_{m=1}^{\nu} \frac{1}{k_m!} \left(\frac{\gamma_{m+2}}{(m+2)!\sigma^{m+2}}\right)^{k_m}, \tag{2.31}$$

where the sum is over $(k_1, k_2, \ldots, k_\nu)$ which verify the condition $k_1 + 2k_2 + \ldots \nu k_\nu = \nu$ and $s = k_1 + \cdots + k_\nu$. Here, H_m is the Chebyshev-Hermite polynomial of degree m:

$$\begin{aligned} H_m(x) &\triangleq (-1)^m e^{x^2/2} \frac{d^m}{dx^m} e^{-x^2/2} \\ &= m! \sum_{k=0}^{[m/2]} \frac{(-1)^k x^{m-2k}}{k!(m-2k)!2^k}, \end{aligned}$$

where γ_m is the cumulant of order m of Z_1 and σ^2 its variance.

For example, for $k = 4$, we get:

$$\begin{aligned} q_1(x) &= \frac{\lambda_3}{\sqrt{2\pi}} e^{-x^2/2} \left(x^3 - 3x\right) \\ q_2(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(\frac{\lambda_3^2}{72}\left(x^6 - 15x^4 + 45x^2 - 15\right) + \frac{\lambda_4}{24}\left(x^4 - 6x^2 + 3\right)\right) \end{aligned}$$

where $\lambda_k \triangleq \frac{\gamma_k}{\sigma^k}$.

Part I

The Complexity of Differential Privacy for Multi-Armed Bandits

Chapter 3

Defining Privacy for Bandits

In this chapter, we extend the definition of Differential Privacy (DP) to the bandit setting. First, we present the three main challenges of adapting DP to bandits: the online and sequential nature of the bandit interaction, and partial information. To overcome the first two challenges of the setting, *i.e.* the online and sequential nature, we propose two ways of representing the bandit policy as a mechanism: (a) as a "batch" mechanism with fixed in-advance inputs, *i.e.* the non-interactive setting, or (b) as an interactive mechanism with inputs adaptively and adversarially chosen, *i.e.* the interactive setting. On the other hand, partial information provides two ways to represent the input dataset of the bandit mechanism: (a) the input is the table of all "potential" rewards, *i.e.* Table DP, or (b) the input is only the list of "observed" rewards, *i.e.* View DP. We prove the relation between the different definitions. The interactive formulation is stronger than the non-interactive one, and Table DP is stronger than View DP. Finally, we discuss other threat models for DP in bandits beyond our setting, *i.e.* the local model, user-lever DP, and instantaneous DP, among others.

Contents

3.1 Introduction

Bandits (Section 2.2) are increasingly used in a wide range of sequential decision-making tasks under uncertainty, such as recommender systems [SWS+22], strategic pricing [BV96], clinical trials [Tho33b] to name a few. These applications often involve individuals' sensitive data, such as personal preferences, financial situation, and health conditions, and thus, naturally, invoke data privacy concerns in bandits.

Example 3.1 (DoctorBandit)**.** *Let us revisit the example of clinical trials. A health researcher (i.e. the bandit policy) wants to find the best medicine between K candidates. Thus, they design a sequential clinical trial. On the t-th trial round, a new patient u_t arrives. The researcher recommends medicine $a_t \in [K]$ to the patient. Then, the patient's reaction to the medicine is observed. If the medicine cures the patient, the observed reward $r_t = 1$, otherwise $r_t = 0$. This observed reward can reveal sensitive information about the health condition of patient u_t. On the other hand, to recommend medicine a_t, the policy might also consider the specific medical conditions (or context) of patient u_t as additional information. This corresponds to the contextual bandits setting (i.e. Section 2.2.7), and the context, in this case, also encodes private patient information.* The goal of a privacy-preserving bandit policy is to recommend a sequence of medicines (actions) that cures the maximum number of patients while protecting the privacy of the patients. *We present this interactive process in Algorithm 6.*

Motivated by such data-sensitive scenarios, privacy issues are widely studied for bandits in different settings, such as finite-armed bandits [MT15, TD16, SS19, HH22], adversarial bandits [TS13, TD17] and linear contextual bandits [SS18, NR18, HGFD22]. All these works adhere to Differential Privacy (DP) [DR14a] as the framework to ensure the data privacy of users, which is presently the gold standard of privacy-preserving data analysis. Also, multiple formulations of DP, namely *local* and *global*, are extended to bandits [BDT19]. Here, *we focus on the global DP formulation*, where *users trust the centralised decision-maker*, i.e. the policy, and provide it access to the raw sensitive rewards. The goal of the policy is to reveal the sequence of actions while protecting the privacy of the users and achieving either minimal regret or minimal sample complexity.

The main contribution of this chapter is to define privacy for bandits rigorously. We compare different ways of adopting pure DP and its relaxations for bandits. We observe that, though some of these definitions are equivalent for pure DP, more care is needed for approximate and zero concentrated DP. We illustrate two main distinctions in the definitions. The first deals with bandit feedback when defining the private input dataset. The second deals with the interactive nature of the policy as a mechanism. Formalising and linking these definitions is a crucial step missing in the private bandit's literature. Our first contribution is to fill this gap.

Algorithm 6 Sequential interaction between a policy and users

Input: A policy $\pi = \{\pi_t\}_{t=1}^T$ and Users $\{u_t\}_{t=1}^T$
Output: A sequence of actions $a_1, \ldots, a_T$
for $t = 1, \ldots, T$ **do**
    The user u_t sends the **sensitive** context c_t to π (if available)
    The policy π recommends action $a_t \sim \pi_t(. \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1})$
    The user u_t sends the **sensitive** reward r_t to π
end for

3.2 Challenges in Adapting DP for Bandits

In this section, we discuss three main challenges in extending DP for bandits: the online and sequential nature of the interaction, in addition to partial information, a.k.a the bandit feedback.

All the definitions of DP in bandits considered hereafter share two main ingredients:

(a) The DP constraint is a property of the policy π alone. Recall that DP is a constraint on algorithms, and in bandits, the algorithm is modelled by the policy, formally defined in Defintion 2.16. This means that all the DP definitions in bandits should only depend on the policy π and be independent of any (stochastic) environment ν.

(b) The published output is the sequence of recommended actions, *i.e.* $(a_1, \ldots, a_T)$.

Let us fix a policy $\pi \triangleq (\pi_1, \ldots \pi_T)$. Inspired by the DP promise, the policy π satisfies DP if the distribution over the sequence of actions $(a_1, \ldots, a_T)$ is "essentially" the same when the policy π interacts with two "neighbouring" sets of users, according to the interactive process of Algorithm 6. Compared to the classic threat model of DP presented in Section 2.1, the interaction protocol of Algorithm 6 has three challenging properties:

(a) The interaction is *online*. At each step t, the policy receives a new input r_{t-1} and should produce a new output a_t.

(b) The interaction is *sequential*. The outputed action a_t affects the new observed input r_t, and depends only on the history $(a_1, r_1, \ldots, a_{t-1}, r_{t-1})$.

(c) The feedback to the bandit policy is *partial*. Specifically, each user u_t can be represented by a K-dimentional vector of "potential" rewards $x_t \triangleq (x_{t,1}, \ldots, x_{t,K})$. When the policy recommends action a_t at step t, it only observes the reward r_t corresponding to a_t, *i.e.* $r_t \triangleq x_{t,a_t}$ and does not observe the other $K-1$ coordinates of x_t.

Next, we present four bandit DP definitions that deal with the three challenges in different ways: Table DP (Definition 3.2), View DP (Definition 3.3), Interactive DP (Definition 3.5) and DP in the adaptive continual release model (Definition 3.9). For each definition, we describe

the mechanism considered, its input, outputs, the interaction protocol and the formalisation of the definition. Then, we explain how each definition tackles the three challenges.

3.3 Table DP vs View DP

Here, we adhere to the non-interactive threat model, *i.e.* the inputs of the policy are supposed to be fixed in advance. Specifically, for a policy $\pi = (\pi_1, \dots, \pi_T)$, we induce a "batch" mechanism that takes as input a "reward dataset" and outputs in "one-shot" a sequence of actions.

In the following, we explicit two "batch" mechanisms that can be induced by the same policy π. These two mechanisms only differ in the input "reward dataset" representation. In short, Table DP considers that the input is the table of all potential rewards, while View DP only considers the list of "observed rewards".

Table DP. We represent each user u_t by the vector $x_t \triangleq (x_{t,1}, \dots, x_{t,K}) \in \mathbb{R}^K$ of all its K "potential rewards". We call this the vector of potential rewards since the policy only observes $r_t \triangleq x_{t,a_t}$ when it recommends action a_t. Then, the Table DP defintions represent a set of T users $\{u_t\}_{t=1}^T$ by the dataset $\mathbf{x} \triangleq \{x_t\}_{t=1}^T \in (\mathbb{R}^K)^T$, that we call the *table of rewards*. The hamming distance between two table of rewards $\mathbf{x}, \mathbf{x}' \in (\mathbb{R}^K)^T$ is the number of different *rows* in $\mathbf{x}$ and $\mathbf{x}'$, *i.e.* $d_{\mathrm{Ham}}(\mathbf{x}, \mathbf{x}') \triangleq \sum_{t=1}^T \mathbb{1}\left(x_t \neq x'_t\right) = \sum_{t=1}^T \mathbb{1}\left(\exists i \in [K],\ x_{t,i} \neq x'_{t,i}\right)$. Neighbouring table of rewards, denoted by $\mathbf{x} \sim \mathbf{x}'$, are table of rewards with hamming distance less than equal to one, *i.e.* $d_{\mathrm{Ham}}(\mathbf{x}, \mathbf{x}') \leq 1$.

In Table DP, we induce a "batch" mechanism $\mathcal{M}^\pi$ from the policy π, which takes as input a table of rewards $\mathbf{x} \triangleq \{(x_{t,i})_{i\in[K]}\}_{t\in[T]} \in (\mathbb{R}^K)^T$, and outputs a sequence of actions $(a_1, \dots, a_T) \in [K]^T$. Specifically,

$$\begin{aligned} \mathcal{M}^\pi :\ (\mathbb{R}^K)^T &\to \mathcal{P}([K]^T) \\ \mathbf{x} &\to \mathcal{M}^\pi_{\mathbf{x}}, \end{aligned}$$

where $\mathcal{M}^\pi_{\mathbf{x}}$ is a distribution over the sequence of actions, and

$$\mathcal{M}^\pi_{\mathbf{x}}(a_1, \dots, a_T) \triangleq \prod_{t=1}^T \pi_t\left(a_t | a_1, x_{1,a_1}, \dots a_{t-1}, x_{t-1,a_{t-1}}\right) \tag{3.1}$$

is the probability of observing the sequence $(a_1, \dots, a_T)$ for the input table of rewards $\mathbf{x}$. Notice that indeed $\sum_{(a_1,\dots,a_t)\in[K]^T} \mathcal{M}^\pi_{\mathbf{x}}(a_1, \dots, a_T) = 1$. Now that we rigorously defined the induced mechanism $\mathcal{M}^\pi$, its input and output, the definition of Table DP follows naturally.

Definition 3.2 (Table DP)**.**

- *A policy π satisfies (ε, δ)-Table DP if and only if $\mathcal{M}^\pi$ is (ε, δ)-DP.*

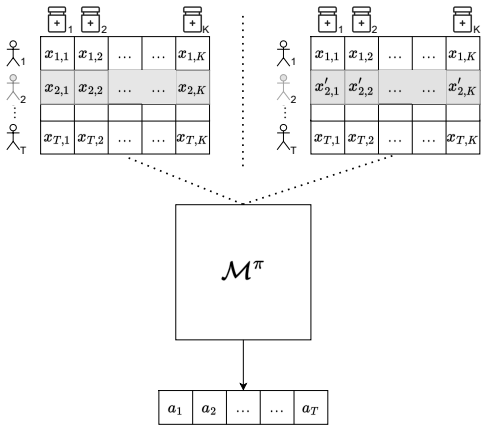


Figure 3.1 – Table DP

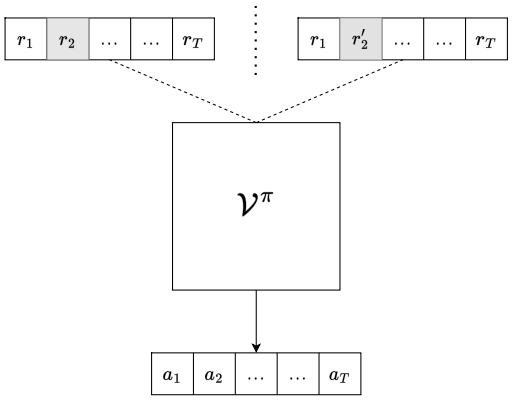


Figure 3.2 – View DP

- *A policy π satisfies ρ-Table zCDP if and only if $\mathcal{M}^\pi$ is ρ-zCDP.*

In simple words, a policy is Table DP if the sequence of output actions is "essentially" the same when the policy interacts with two neighbouring, fixed-in-advance tables of rewards.

Table DP is a formalisation of the privacy definition adopted in [TS13, MT15, NR18]. Table DP deals with the three challenges of defining DP in the following:

(a) To deal with the online nature of the interaction, Table DP uses the "batch" reduction idea, similar to the formulation of DP under continual observation (Section 2.1.5), *i.e.* the input dataset is fixed in advance.

(b) The sequential nature of the interaction is captured by the expression of Equation (3.1). Specifically, in Equation (3.1), the probability of outputting arm a_t depends only on the past $< t$ rewards, *i.e.* $(x_1, \dots, x_{t-1})$, and the past outputed actions, *i.e.* $(a_1, \dots, a_{t-1})$.

(c) To deal with partial information, Table DP considers that the "right" input representation for the "batch" mechanism is the table of "potential" rewards.

View DP. In this definition, we also induce a "batch" mechanism from the policy π which takes as input a fixed in-advance dataset of rewards and outputs a sequence of actions. The difference is in the representation of the input dataset. Since in bandits, the policy only observes the reward corresponding to the action chosen, another natural choice for the input is a list of rewards, *i.e.* $\mathbf{r} \triangleq \{r_1, \dots, r_T\} \in \mathbb{R}^T$. The Hamming distance between two lists of rewards $r, r' \in \mathbb{R}^T$ is the number of different elements in r and r', *i.e.* $d_{\text{Ham}}(r, r') \triangleq \sum_{t=1}^T \mathbb{1}(r_t \neq r'_t)$. Neighbouring list of rewards, denoted by $\mathbf{r} \sim \mathbf{r}'$, is a list of rewards with hamming distance less than equal to one, *i.e.* $d_{\text{Ham}}(\mathbf{r}, \mathbf{r}') \leq 1$.

In View DP, we induce a "batch" mechanism $\mathcal{V}^\pi$ from the policy π, which takes as input a list of rewards $\mathbf{r} \triangleq \{r_1, \dots, r_T\} \in \mathbb{R}^T$, and outputs a sequence of actions $(a_1, \dots, a_T) \in [K]^T$.

Specifically,

$$\begin{aligned} \mathcal{V}^{\pi}: \ \mathbb{R}^T &\to \mathcal{P}([K]^T) \\ \mathbf{r} &\to \mathcal{V}^{\pi}_{\mathbf{r}}, \end{aligned}$$

where $\mathcal{V}^{\pi}_{\mathbf{r}}$ is a distribution over the sequence of actions, and

$$\mathcal{V}^{\pi}_{\mathbf{r}}(a_1, \dots, a_T) = \prod_{t=1}^{T} \pi_t(a_t | a_1, r_1, \dots a_{t-1}, r_{t-1}) \tag{3.2}$$

is the probability of observing the sequence $(a_1, \dots, a_T)$ for the input list of rewards $\mathbf{r}$. Notice that indeed $\sum_{(a_1,\dots,a_t)\in[K]^T} \mathcal{V}^{\pi}_{\mathbf{r}}(a_1, \dots, a_T) = 1$. Again, the definition of View DP follows naturally.

Definition 3.3 (View DP)**.**

- *A policy π satisfies (ε, δ)-View DP if and only if $\mathcal{V}^{\pi}$ is (ε, δ)-DP.*
- *A policy π satisfies ρ-View zCDP if and only if $\mathcal{V}^{\pi}$ is ρ-zCDP.*

In simple words, a policy is View DP if the sequence of output actions is "essentially" the same when the policy interacts with two neighbouring, fixed-in-advance lists of rewards.

View DP is a formalisation of the definition adopted in [SS19, HHM21, HGFD22]. View DP deals with the three challenges of defining DP in the following:

(a) To deal with the online nature of the interaction, View DP also uses the "batch" reduction idea, similar to Table DP.

(b) The sequential nature of the interaction is captured by the expression of Equation (3.2). Similar to Table DP, in Equation (3.2), the probability of outputting arm a_t depends only on the past $< t$ rewards, *i.e.* $(r_1, \dots, r_{t-1})$, and the past outputed actions, *i.e.* $(a_1, \dots, a_{t-1})$. However, there is a subtle difference between Equation (3.2) of View DP and Equation (3.1) of Table DP, which we discuss in Remark 3.4.

(c) To deal with partial information, View DP considers that the "right" input representation for the "batch" mechanism is the list of "observed" rewards.

Remark 3.4. [*Difference between the equations in Table DP and View DP*] *At first glance, Equations (3.1) and (3.2) look very similar. However, the differences arise when $\mathcal{V}^{\pi}_{\mathbf{r}}$ and $\mathcal{M}^{\pi}_{d}$ are applied to a "non-atomic" event $E \in \mathcal{P}([K]^T)$. For example, if we define an event $E \triangleq \{(a_1, \dots, a_T), (b_1, \dots, b_T)\}$, i.e. that the output is either the sequence of actions $(a_1, \dots, a_T)$ or the sequence $(b_1, \dots, b_T)$. Then*

$$\mathcal{V}^{\pi}_{\mathbf{r}}(E) = \prod_{t=1}^{T} \pi_t(a_t | a_1, r_1, \dots a_{t-1}, r_{t-1}) + \prod_{t=1}^{T} \pi_t(b_t | b_1, r_1, \dots b_{t-1}, r_{t-1}),$$

while

$$\mathcal{M}_d^\pi(E) = \prod_{t=1}^{T} \pi_t(a_t|a_1, x_{1,a_1}, \dots a_{t-1}, x_{t-1,a_{t-1}}) + \prod_{t=1}^{T} \pi_t(b_t|b_1, x_{1,b_1}, \dots b_{t-1}, x_{t-1,b_{t-1}}).$$

In the expression of $\mathcal{V}_{\mathbf{r}}^\pi(E)$, the same rewards $(r_1, \dots, r_T)$ appear in the two elements of the sum. In contrast, in the expression of $\mathcal{M}_d^\pi(E)$, each sequence of actions generates different trajectories of reward in the table, i.e. $(x_{1,a_1}, \dots x_{T,a_T})$ *vs* $(x_{1,b_1}, \dots x_{T,b_T})$. *As we show in Section 3.6, this subtle difference is the source of the difference between Table DP and View DP in approximate DP.*

3.4 Interactive Differential Privacy

In this section, we consider the interactive threat model, where an adversary chooses adaptively the reward input to send to the policy at step t, based on previous outputs $a_1, \dots, a_t$. We adhere to the Interactive DP framework of [VZ22] to extend interactive DP to bandits. This framework considers the policy as a party in an interaction protocol, interacting with a possibly adversarial analyst.

The interaction protocol has three components:

(a) The policy $\pi = \{\pi_t\}_{t=1}^T$.

(b) An adversary $B \triangleq \{B_t\}_{t=1}^T$, such that $B_t : (a_1, \dots a_t) \in [K]^t \to q_t \in [K]$

(c) A table of potential rewards $\mathbf{x} \triangleq (x_1, \dots, x_T) \in (\mathbb{R}^K)^T$.

The interaction protocol is the following:

For $t = 1, \dots, T$

1. The bandit algorithm selects an action

$$a_t \sim \pi_t(\cdot \mid q_1, x_{1,q_1}, \dots, q_{t-1}, x_{t-1,q_{t-1}}),\ a_t \in [K]$$

2. The adversary returns a "query" action, which may depend on past recommended actions by the policy

$$q_t = B_t(a_1, a_2, \dots, a_t),\ q_t \in [K]$$

3. The bandit algorithm observes the reward corresponding to query action q_t for user u_t in the table $\mathbf{x}$, *i.e.* x_{t,q_t}.

We represent this interaction by $\pi \leftrightarrow^{\mathbf{x}} B$, and illustrate it in Figure 3.3.

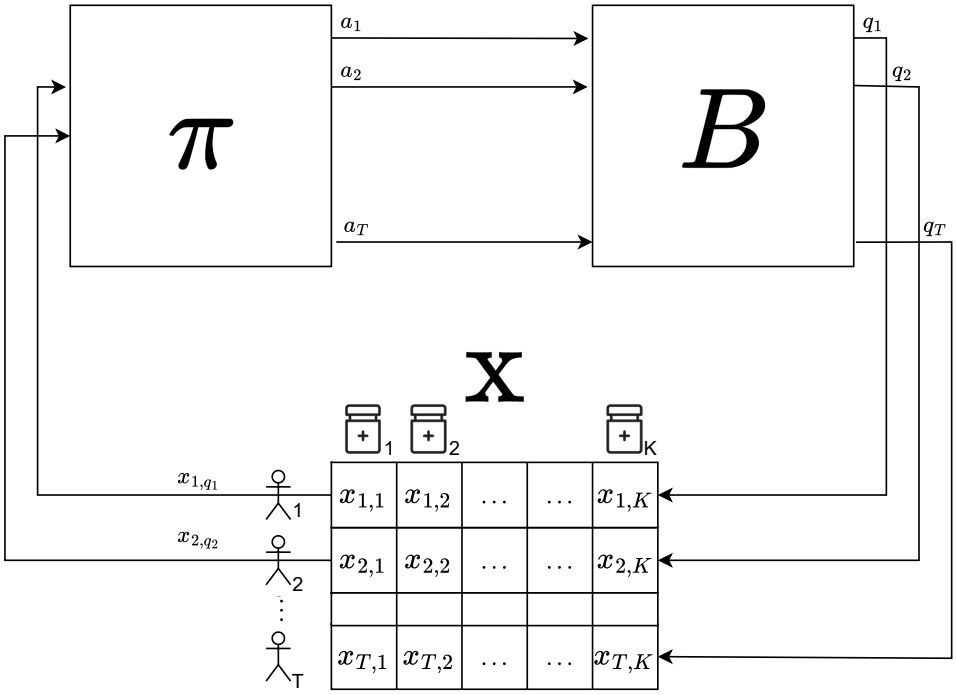


Figure 3.3 – Interaction protocol between the policy, an adversary B, and a table of rewards $\mathbf{x}$.

In this interaction protocol, the reward input x_{t,q_t} that the policy observes at step t is adaptively chosen by the adversary B. Specifically, the adversary chooses a query action q_t that depends on the policy's past actions $(a_1, \ldots, a_t)$. Then, this query action is used to produce reward x_{t,q_t} from the table of rewards $\mathbf{x}$. The policy is then updated using the query action q_t and reward x_{t,q_t} to produce the next action a_{t+1}. Following the Interactive DP framework [VZ22], the policy π satisfies interacive DP if the view of adversary B, *i.e.*

$$\mathrm{View}_{B,\pi,\mathbf{x}} \triangleq \mathrm{View}_B(\pi \leftrightarrow^{\mathbf{x}} B) \triangleq (a_1, \ldots, a_T),$$

is indistinguishable when the interaction is run on two neighbouring tables of rewards $\mathbf{x}$ and $\mathbf{x}'$.

Definition 3.5 (Interactive DP)**.**

- *A policy π satisfies (ε, δ)-Interactive DP for a given $\varepsilon \geq 0$ and $\delta \in [0, 1)$, if for all adversaries B and all subset of views $\mathcal{S} \subseteq [K]^T$,*

$$\sup_{\boldsymbol{x} \sim \boldsymbol{x}'} \mathbb{P}[\mathrm{View}_{B,\pi,\boldsymbol{x}} \in \mathcal{S}] - e^{\varepsilon}\mathbb{P}[\mathrm{View}_{B,\pi,\boldsymbol{x}'} \in \mathcal{S}] \leq \delta.$$

- *A policy π satisfies ρ-Interactive zCDP policy for a given $\rho \geq 0$, if for every $\alpha > 1$, and every adversary B,*

$$\sup_{\boldsymbol{x} \sim \boldsymbol{x}'} D_\alpha(\text{View}_{B,\pi,\boldsymbol{x}} \| \text{View}_{B,\pi,\boldsymbol{x}'}) \leq \rho\alpha.$$

In simple words, a policy is Interactive DP if the view of any adversary is essentially the same when the policy and adversary interact with two neighbouring tables of rewards.

Interactive DP provides stronger privacy guarantees than Table DP and View DP since Interactive DP deals with a stronger adversary that adaptively queries inputs. In contrast, View DP and Table DP only consider fixed-in-advance datasets.

Interactive DP deals with the challenges of adapting DP to bandits in the following:

(a) The online nature of the bandit interaction is captured by the online nature of the adversary B. This adversary provides a new query q_t at each step t, which translates to a new reward input x_{t,q_t}.

(b) The sequential nature of the bandit interaction is captured by the sequential nature of the adversary B. This adversary provides a querry q_t at each step t that only depends on the policy's past outputs $a_1, \ldots, a_t$.

(c) Similar to Table DP, Inteactive DP deals with partial information by considering the input to be the table of "potential" rewards. It is possible to derive a View DP version of Interactive DP, where the adversary directly comes up with a new reward query to the policy.

In addition to modelling for a stronger adaptive adversary, Interactive DP has two additional interesting qualities. (a) Interactive DP protects the privacy of the users even when the users are non-compliant [Kal18, SJ18]. A non-compliant user is a user who decides to ignore the recommendation of the policy and chooses a different arm. Intuitively, in clinical trials, we want to protect the patients' privacy, even if they do not follow the recommended medicine by the doctor. (b) Interactive DP decouples actions and rewards in the privacy definition, yielding stronger group privacy properties. We discuss this further in detail in Section 4.2.

Remark 3.6. [*Expanding the View of the Adversary B*] *For any deterministic adversary B, any policy π, reward table $\boldsymbol{x} \in (\mathbb{R}^K)^T$, and any $(a_1, \ldots, a_T) \in [K]^T$, we have*

$$\begin{aligned}\mathbb{P}[\text{View}_{B,\pi,\boldsymbol{x}} = (a_1, \ldots, a_T)] = \pi_1(a_1)\pi_2(a_2 \mid B_1(a_1), x_{1,B_1(a_1)}) \cdots \times \\ \pi_T(a_T \mid B_1(a_1), x_{1,B_1(a_1)}, \ldots, B_{T-1}(a_1, \ldots, a_{T-1}), x_{T-1,B_{T-1}(a_1,\ldots,a_{T-1})})\end{aligned}$$

Thus, for any $S \subseteq [K]^T$, we get

$$\mathbb{P}[\text{View}_{B,\pi,\boldsymbol{x}} \in S] = \mathcal{M}_{\boldsymbol{x}}^{\pi^B}(S)$$

where $\pi^B \triangleq \{\pi_t^B\}_{t=1}^T$ is the B post-processed policy, defined by

$$\pi_t^B(a \mid a_1, r_1, .., a_{t-1}, r_{t-1}) \triangleq \pi_t\left(a \mid B_1(a_1), r_1, B_2(a_1, a_2), r_2, .., B_{t-1}(a_1, .., a_{t-1}), r_{t-1}\right).$$

Remark 3.7 (Privacy protocol vs Utility protocol)**.** *The interaction protocols to analyse privacy and utility are different. To express the Interacive DP constraint, the protocol between the policy, an adversary, and a table of rewards is defined (Fig 3.3). In this protocol, an Interactive DP policy is constrained to show a similar view to any "privacy" adversary when interacting with two neighbouring reward tables. Given a policy that verifies this Interactive DP constraint, we measure the utility (regret/sample complexity) of this policy when it interacts with an environment using the canonical bandit protocol (Section 2.2.2). There is no adversary in the interaction used to measure the utility of the policy.*

3.5 Adaptive Continual Release Model

In this section, we extend the adaptive continual release model of [JRSS23] to bandits. In this model, similar to the Interactive DP definition of Section 3.4, the policy interacts with an adversary that chooses adaptively rewards based on previous outputs of the policy. The difference between these two models is in the nature of the adversary:

(a) In Interactive DP (Definition 3.5), the adversary $B \triangleq \{B_t\}_{t=1}^T$ is a sequence of functions that map the history of observed actions to query actions. Specifically, at step t, the adversary observes the history $(a_1, \ldots, a_t) \in [K]^t$ of actions recommended, and comes up with a query action $q_t = B_t(a_1, \ldots, a_t) \in [K]$. This query action q_t is then used to generate a reward from a fixed table of rewards $\mathbf{x}$; *i.e.* $r_t = x_{t,q_t}$. The policy updates its recommendations at step $t+1$ based on q_t and r_t.

(b) In the adaptive continual release model of [JRSS23], the adversary directly comes up with a reward. In the following, we formalise this new notion of adversary and call it a "reward-feeding" adversary. This is in contrast to the adversary B of Interactive DP, which is a "query-action" feeding adversary.

Definition 3.8 (A reward-feeding adversary $\mathcal{A}$)**.** *A reward-feeding adversary $\mathcal{A}$ is a sequence of functions $(\mathcal{A}_t)_{t=1}^T$ such that, for $t \in \{1, \ldots, T\}$,*

$$\mathcal{A}_t : a_1, \ldots, a_t \to (r_t^L, r_t^R).$$

A "reward-feeding" adversary $\mathcal{A}$ is a sequence of "reward" functions which take as input the action-history and outputs a pair of rewards (r_t^L, r_t^R). The reward-feeding adversary $\mathcal{A}$ has two channels: a left "standard" channel L and a right channel R. These channels are used to simulate "neighbouring" rewards.

Precisely, to simulate "neighbouring" rewards, the interactive protocol between the policy π and the reward-feeding adversary $\mathcal{A}$ has two hyper-parameters: (a) a specific "challenge" time $t^\star \in \{1, T\}$, and (b) a binary $b \in \{L, R\}$. For steps $t \neq t^\star$, the policy observes a reward coming from the adversary's left "standard" channel, i.e. $r_t = r_t^L$. Otherwise, when $t = t^\star$, the policy observes a reward from the channel corresponding to the secret binary b.

In other words, if $b = L$, the policy π always observes a reward from the left channel. When $b = R$, the policy observes the left channel reward for all steps, except at $t^\star$ where the policy observes a right channel reward. Thus, for any sequence of actions $(a_1, \ldots, a_T)$ chosen by the policy π, and for any $t^\star$, the sequence of rewards observed by π when $b = L$ is neighbouring to the sequence of rewards observed when $b = R$. In addition, these two sequences only differ at the reward observed at the challenge time $t^\star$, and the rewards have been adaptively chosen by the adversary.

Thus, we formalise the adaptive continual release interaction as follows:

Let $b \in \{L, R\}$ and $t^\star \in \{1, \ldots, T\}$.
For $t = 1, \ldots, T$

1. The policy π selects an action

$$a_t \sim \pi_t(\cdot \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1}),\ a_t \in [K]$$

2. The adversary $\mathcal{A}$ selects an adaptively chosen pair of rewards:

$$(r_t^L, r_t^R) = \mathcal{A}_t(a_1, \ldots, a_t)$$

   - If $t \neq t^\star$:

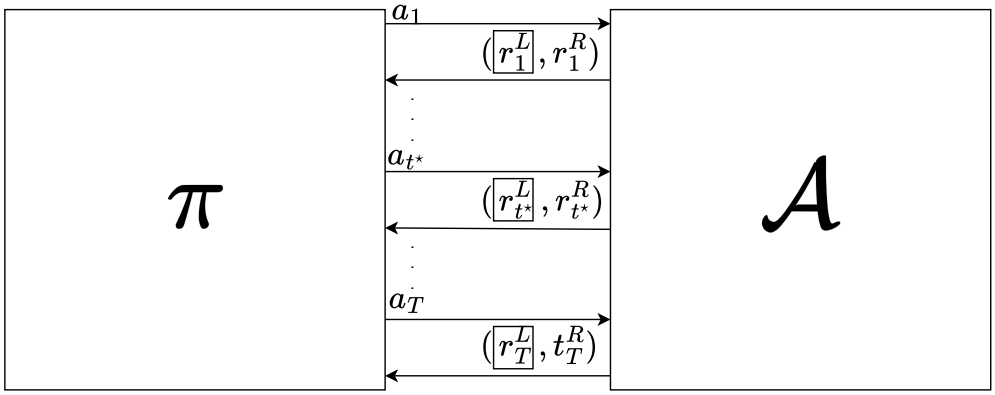
$$r_t = r_t^L$$

   - If $t = t^\star$:

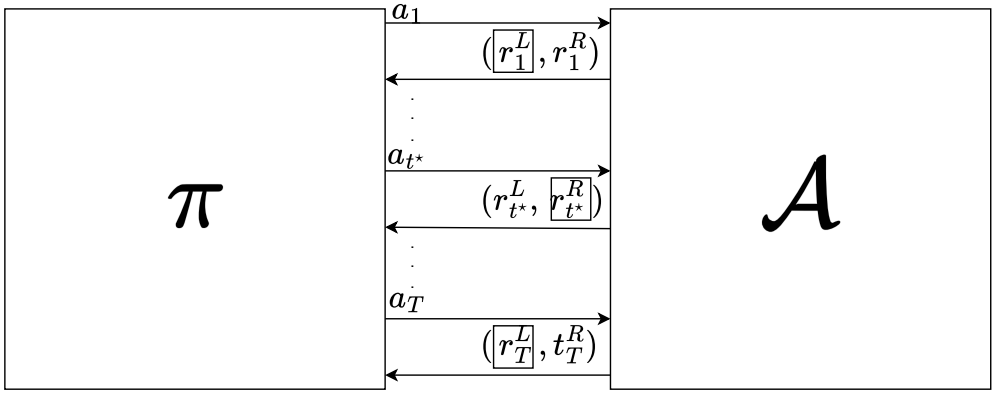
$$r_{t^\star} = r_{t^\star}^b$$

3. The policy π observes the reward r_t

When this interaction is run with parameters $t^\star$ and b, we represent the interaction by $\pi \overset{b,t^\star}{\Leftrightarrow} \mathcal{A}$, and illustrate it in Figure 3.4. The view of the adversary $\mathcal{A}$ in the interaction $\pi \overset{b,t^\star}{\Leftrightarrow} \mathcal{A}$ is the



$(a)\, b = L$



$(b)\, b = R$

Figure 3.4 – Interactive protocol in the adaptive continual release model between a policy π and a reward-feeding adversary $\mathcal{A}$. The protocol in Figure (a) is run with $b = L$, while the protocol in Figure (b) is run with $b = L$. The framed part corresponds to the reward observed by the policy.

sequence of actions chosen by the policy π, *i.e.*

$$\mathrm{View}_{\mathcal{A},\pi}^{b,t^\star} \triangleq \mathrm{View}_{\mathcal{A}}(\pi \overset{b,t^\star}{\Leftrightarrow} \mathcal{A}) \triangleq (a_1, \dots, a_T).$$

A policy is DP in the adaptive continual release model if the view of the adversary is indistinguishable when the interaction is run on $b = L$ and $b = R$ for any challenge step $t^\star$.

Definition 3.9 (DP in the Adaptive Continual Release Model)**.**

- *A policy π is (ε, δ)-DP in the adaptive continual release model for a given $\varepsilon \geq 0$ and $\delta \in [0, 1)$, if for all reward-feeding adversaries $\mathcal{A}$, all subset of views $\mathcal{S} \subseteq [K]^T$,*

$$\sup_{t^\star \in \{1,\dots,T\}} \mathbb{P}[\mathrm{View}_{\mathcal{A},\pi}^{L,t^\star} \in \mathcal{S}] - e^\varepsilon \mathbb{P}[\mathrm{View}_{\mathcal{A},\pi}^{R,t^\star} \in \mathcal{S}] \leq \delta.$$

- *A policy π is ρ-zCDP in the adaptive continual release model for a given $\rho \geq 0$, if for every $\alpha > 1$, and every reward-feeding adversary $\mathcal{A}$,*

$$\sup_{t^\star \in \{1,\dots,T\}} D_\alpha(\mathrm{View}_{\mathcal{A},\pi}^{L,t^\star} \| \mathrm{View}_{\mathcal{A},\pi}^{R,t^\star}) \leq \rho\alpha.$$

The adaptive continual release model deals with the challenges of adapting DP to bandits in the following:

(a) The online nature of the bandit interaction is captured by the online nature of the reward-feeding adversary $\mathcal{A}$. This adversary provides a pair of rewards (r_t^L, r_t^R) at each step t.

(b) The sequential nature of the bandit interaction is captured by the sequential nature of the reward-feeding adversary $\mathcal{A}$. This adversary provides a pair of rewards (r_t^L, r_t^R) at each step t that only depends on the policy's past actions $a_1, \dots, a_t$.

(c) Similar to View DP, the adaptive continual release model deals with partial information by considering the input to be the observed rewards.

Remark 3.10. [*Expanding the View of the Reward-feeding Adversary $\mathcal{A}$*] *For any reward-feeding adversary $\mathcal{A}$, any policy π and any $t^\star \in \{1, \dots, T\}$, and any $(a_1, \dots, a_T) \in [K]^T$, we have for the left view:*

$$\begin{aligned}\mathbb{P}[\mathrm{View}_{\mathcal{A},\pi}^{L,t^\star} = (a_1, \dots, a_T)] = \pi_1(a_1)\pi_2(a_2 \mid a_1, \mathcal{A}_1^L(a_1)) \cdots \times \\ \pi_T(a_T \mid a_1, \mathcal{A}_1^L(a_1), \dots, a_{T-1}, \mathcal{A}_{T-1}^L(a_1, \dots, a_{T-1}))\end{aligned}$$

On the other hand, for the right view:

$$\begin{aligned}\mathbb{P}[\mathrm{View}_{\mathcal{A},\pi}^{R,t^\star} = (a_1, \dots, a_T)] = \pi_1(a_1)\pi_2(a_2 \mid a_1, \mathcal{A}_1^L(a_1)) \cdots \times \\ \pi_{t^\star+1}(a_{t^\star+1} \mid a_1, \mathcal{A}_1^L(a_1), \dots, a_{t^\star}, \mathcal{A}_{t^\star}^R(a_1, \dots, a_{t^\star})) \cdots \times \\ \pi_T(a_T \mid a_1, \mathcal{A}_t^L(a_1), \dots, a_{T-1}, \mathcal{A}_{T-1}^L(a_1, \dots, a_{t-1}))\end{aligned}$$

Let us define

$$\mathcal{A}^{L,t^\star}(a_1, \dots, a_T) \triangleq (\mathcal{A}_1^L(a_1), \mathcal{A}_2^L(a_1, a_2), \dots, \mathcal{A}_T^L(a_1, \dots, a_T))$$

to be the list of rewards that the policy observes when the protocol is run on the left channel. Also,

$$\mathcal{A}^{R,t^\star}(a_1,\dots,a_T) \triangleq (\mathcal{A}_1^L(a_1),\dots,\mathcal{A}_{t^\star}^R(a_1,\dots,a_{t^\star})\dots\mathcal{A}_T^L(a_1,\dots,a_T))$$

is the list of rewards that the policy observes when the protocol is run on the right channel and $t^\star$.

We observe that, for any $(a_1,\dots,a_T)\in[K]^T$,

(a) $\mathbb{P}[\mathrm{View}_{\mathcal{A},\pi}^{L,t^\star} = (a_1,\dots,a_T)] = \mathcal{V}^\pi((a_1,\dots,a_T) \mid \mathcal{A}^{L,t^\star}(a_1,\dots,a_T))$

(b) $\mathbb{P}[\mathrm{View}_{\mathcal{A},\pi}^{R,t^\star} = (a_1,\dots,a_T)] = \mathcal{V}^\pi((a_1,\dots,a_T) \mid \mathcal{A}^{R,t^\star}(a_1,\dots,a_T))$

(c) *$\mathcal{A}^{L,t^\star}(a_1,\dots,a_T)$ and $\mathcal{A}^{R,t^\star}(a_1,\dots,a_T)$ are neighbouring lists of rewards, and only differ at the $t^\star$-th element*

This remark will help connect the adaptive continual release model with View DP later.

Remark 3.11. [*Reward-feeding Adversary as a Tree Reward Input*] *A reward-feeding adversary can be represented by a tree of rewards. Each node in the tree corresponds to a reward input. The tree has a depth of size T. At depth $t\in[T]$ of the tree reside all possible rewards the policy can observe at step t. Going from depth t to depth $t+1$ depends on the action a_{t+1}. Finally, the policy only observes the reward corresponding to its trajectory in the tree. An example of the tree is presented in Figure 3.5c for $T=3$ and $K=2$.*

A policy π is DP in the adaptive continual release model if and only if π is DP when interacting with two neighbouring trees of rewards. Two trees of rewards are neighbouring if they only differ at rewards at one depth $t^\star\in[T]$.

3.6 Relation between DP Definitions

This section proves the relationship between Table DP, View DP, and Interactive DP constraints. First, we summarise the relations between Table DP and View DP in the following proposition.

Proposition 3.12 (Relation between Table DP and View DP)**.** *For any policy π, we have that*

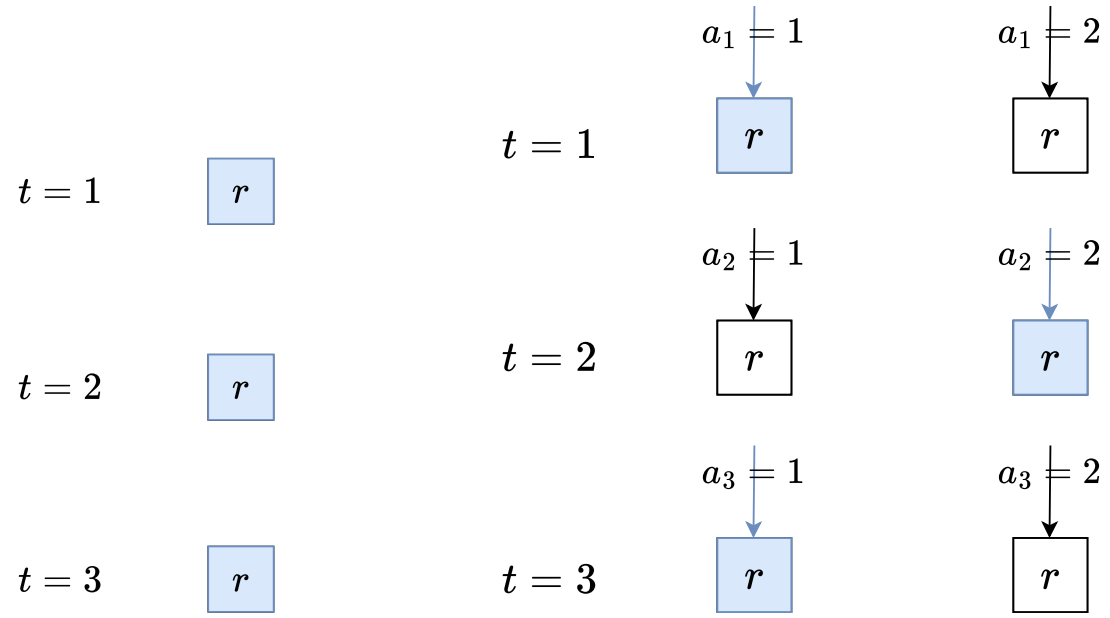
(a) *π is ε-Table DP $\Leftrightarrow$ π is ε-View DP.*

(b) *π is (ε,δ)-Table DP $\Rightarrow$ π is (ε,δ)-View DP.*

(c) *π is ρ-Table zCDP $\Rightarrow$ π ρ-View zCDP.*

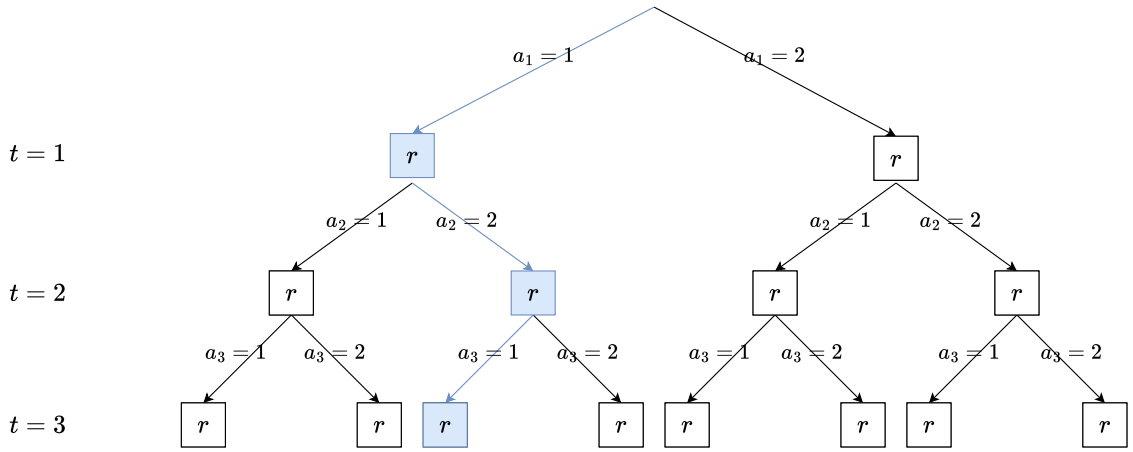
(d) *π is (ε,δ)-View DP $\Rightarrow$ π is $(\varepsilon,K^T\delta)$-Table DP.*

(e) *$\Pi_{\mathit{Table}}^{(\varepsilon,\delta)} \subsetneq \Pi_{\mathit{View}}^{(\varepsilon,\delta)}$, where $\Pi_{\mathit{Table}}^{(\varepsilon,\delta)}$ and $\Pi_{\mathit{View}}^{(\varepsilon,\delta)}$ are the class of all policies verifying (ε,δ)-Table DP and (ε,δ)-View DP, respectively.*



(**a**) List of rewards

(**b**) Table of rewards



(**c**) Tree of rewards

Figure 3.5 – Different reward representations for $T = 3$ and $K = 2$. The highlighted rewards are the rewards observed by the policy for the trajectory $(a_1, a_2, a_3) = (1, 2, 1)$

Consequences of Proposition 3.12. Proposition 3.12 establishes that Table DP is a "stronger" notion of privacy than View DP. Table DP protects all the **potential** responses of an individual rather than just the **observed** ones.

Specifically, Proposition 3.12(a) shows that Table DP and View DP are equivalent for pure DP. For relaxations of pure DP, *i.e.* for (ε, δ)-DP and ρ-zCDP, Proposition 3.12(b) and 3.12(c) show that Table DP always implies View DP with the same privacy budget.

However, the converse from View DP to Table DP happens with a loss in the privacy budget. Proposition 3.12(e) states that the class of policies verifying (ε, δ)-Table DP is *strictly* included in the class of policies verifying (ε, δ)-View DP. To prove this, we build a policy that verifies some $(\varepsilon_1, \delta_1)$-View DP but is shown to be never $(\varepsilon_1, \delta_1)$-Table DP. This validates that going from View DP to Table DP must happen with a *loss* in the privacy budgets. Proposition 3.12(d) yields a simple loss quantification. We leave it an open problem to quantify the best privacy loss conversion from View DP to Table DP.

The proof is presented in Appendix A. The proof uses the following two handy reductions, going from list to table of rewards and vice versa.

Reduction 1 (From lists to table of rewards)**.** *For a list of reward $r \in \mathbb{R}^T$, we define $x(r)$ to be the table of rewards that concatenates r column-wise K times,* i.e. *$x(r)_{t,a} = r_t$ for all $a \in [K]$ and all $t \in [T]$. This transformation has two interesting properties:*

- *For every $r \in \mathbb{R}^T$, we have $\mathcal{V}^\pi_r = \mathcal{M}^\pi_{x(r)}$*
- *If $r \sim r'$ are neighbouring list of rewards, then $x(r) \sim x(r')$ are neighbouring table of rewards*

Reduction 2 (From table of rewards to lists)**.** *For every atomic event $a^T \triangleq (a_1, \dots, a_T)$ and a table of reward $\boldsymbol{x} \in (\mathbb{R}^K)^T$, we define the list of rewards $r(\boldsymbol{x}, a^T) \in \mathbb{R}^T$ such that $r(\boldsymbol{x}, a^T)_t = d_{t,a_t}$. In other words, $r(\boldsymbol{x}, a^T)$ is the list of rewards corresponding to the trajectory of a^T in $\boldsymbol{x}$. This transformation has two interesting properties:*

- *For every table of rewards $\boldsymbol{x}$ and every atomic event a^T, we have $\mathcal{M}^\pi_{\boldsymbol{x}}(a^T) = \mathcal{V}^\pi_{r(\boldsymbol{x},a^T)}(a^T)$*
- *If $\boldsymbol{x} \sim \boldsymbol{x}'$ are neighbouring table of rewards, then for every atomic event a^T, $r(\boldsymbol{x}, a^T) \sim r(\boldsymbol{x}', a^T)$ are neighbouring list of rewards.*

An Intuition. Reduction 1 shows that $\mathcal{V}^\pi_r$ can be represented by $\mathcal{M}^\pi$ on a specific table of rewards (the column-wise concatenation of r). If π is Table DP, then $\mathcal{M}^\pi_{\boldsymbol{x}}$ and $\mathcal{M}^\pi_{\boldsymbol{x}'}$ are indistinguishable for all $\mathbf{x} \sim \mathbf{x}'$. Specifically, for column-wise identical tables, the indistinguishability property is still true, which recovers that π is View DP using Reduction 1. On the other hand, Reduction 2 shows that $\mathcal{M}^\pi_x$ can be represented by $\mathcal{V}^\pi$ on a specific reward list (the trajectory), but only for atomic events. This provides that View DP implies Table DP only for pure DP where the indistinguishability condition is enough to be verified for atomic events.

Now, we relate Interactive DP and Table DP in the following proposition.

Proposition 3.13 (Relation between Interactive DP and Table DP)**.** *For any policy π, we have that*

(*a*) *π is Interactive DP $\Rightarrow$ π is Table DP*

(*b*) *π is Interactive DP if and only if, for every deterministic adversary $B = \{B_t\}_{t=1}^T$, π^B is Table DP. Here, $\pi^B \triangleq \{\pi_t^B\}_{t=1}^T$ is a post-processing of the policy π induced by the adversary B such that*

$$\pi_t^B(a \mid a_1, r_1, .., a_{t-1}, r_{t-1}) \triangleq \pi_t\left(a \mid B_1(a_1), r_1, B_2(a_1, a_2), r_2, .., B_{t-1}(a_1, .., a_{t-1}), r_{t-1}\right).$$

Proposition 3.13 shows that Interactive DP is stronger than Table DP. On the other hand, for policies that are "closed" under interactive post-processing, Interactive DP and ρ-Table are equivalent. A policy is "closed" under interactive post-processing if its privacy property does not depend on the order/the source of the actions, *i.e.* whether they were recommended by the policy itself or queried by an adversarial analyst. In Chapter 5, we show that algorithms in the private bandit literature, which are based on adding the Laplace mechanism to the empirical means or use the binary tree mechanism [DNP+10, CSS11], verify both Table DP and Interactive DP, *i.e.* are "closed" under post-processing.

Remark 3.14 (Deterministic adversaries are enough)**.** *We recall that to check the interactive DP condition, it is enough only to consider deterministic adversaries* (*Lemma 2.2 in* [*VW21*])*.*

Proof. (a) is direct by taking the identity-adversary B^{id} defined by $B_t^{\mathrm{id}}(o_1, \ldots, o_t) = o_t$.

(b) is direct by observing that for every deterministic adversary B, $\mathrm{View}_B(\pi \leftrightarrow^{\mathbf{x}} B)$ reduces to $\mathcal{M}_{\mathbf{x}}^{\pi^B}$, *i.e.* $\mathrm{View}_B(\pi \leftrightarrow^{\mathbf{x}} B) = \mathcal{M}_{\mathbf{x}}^{\pi^B}$. □

Now, we relate DP in the adaptive continual release model with View DP and Table DP.

Proposition 3.15 (Relation between the Adaptive Continual Release Model, View DP and Table DP)**.** *For any policy π, we have that*

(*a*) *π is DP in the adaptive continual release model $\Rightarrow$ π is Table DP*

(*b*) *π is ε-DP in the adaptive continual release model $\Leftrightarrow$ π is ε-Table DP $\Leftrightarrow$ π is ε-View DP*

Proposition 3.15 shows that the adaptive continual release model is stronger than Table DP. For pure ε-DP, the adaptive continual release model, Table DP and View DP are all equivalent.

To prove this proposition, we use the following reduction.

Reduction 3 (From table of rewards to "reward-feeding" adversaries)**.** *For a pair of reward tables $\boldsymbol{x}, \boldsymbol{x'} \in (\mathbb{R}^K)^T$, we define $\mathcal{A}(\boldsymbol{x}, \boldsymbol{x'})$ to be the "reward-feeding" adversary defined by*

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{x'})_t : a_1, \ldots, a_t \to (x_{t,a_t}, x'_{t,a_t}).$$

In other words, at step t, the adversary $\mathcal{A}(\boldsymbol{x}, \boldsymbol{x'})$ only uses the last action a_t and returns the a_t-th column from x_t on the left channel, and the a_t-th column from x'_t on the right channel.

For neighbouring tables $\boldsymbol{x}$ and $\boldsymbol{x'}$ which only differ at some step $t^\star$, it is possible to show that, for every $S \in \mathbb{R}^T$, we have

- $\mathbb{P}[\text{View}^{L,t^\star}_{\mathcal{A}(\boldsymbol{x},\boldsymbol{x'}),\pi} \in \mathcal{S}] = \mathcal{M}^\pi_{\boldsymbol{x}}(S)$
- $\mathbb{P}[\text{View}^{R,t^\star}_{\mathcal{A}(\boldsymbol{x},\boldsymbol{x'}),\pi} \in \mathcal{S}] = \mathcal{M}^\pi_{\boldsymbol{x'}}(S)$

In other words, the batch mechanism $\mathcal{M}^\pi$ combined with neighbouring tables can be "simulated" using a specific type of "reward-feeding" adversaries that only care about the last action from the history.

Proof. (a) Suppose that π is DP in the adaptive continual release model.

Let $t^\star \in [T]$, and $x \sim x'$ be two tables of rewards in $(\mathbb{R}^K)^T$ that only differ at step $t^\star$. Using Reduction 3, we build $\mathcal{A}(x, x')$.

For this construction, we have that $\mathcal{M}^\pi_x = \text{View}^{L,t^\star}_{\mathcal{A}(\mathbf{x},\mathbf{x'}),\pi}$ and $\mathcal{M}^\pi_{x'} = \text{View}^{R,t^\star}_{\mathcal{A}(\mathbf{x},\mathbf{x'}),\pi}$.

Since π is DP in the adaptive continual release model, $\text{View}^{L,t^\star}_{\mathcal{A}(\mathbf{x},\mathbf{x'}),\pi}$ and $\text{View}^{L,t^\star}_{\mathcal{A}(\mathbf{x},\mathbf{x'}),\pi}$ are indistinguishable. Thus, $\mathcal{M}^\pi_x$ and $\mathcal{M}^\pi_{x'}$ are indistinguishable, *i.e.* $\mathcal{M}^\pi$ is DP and π is Table DP.

(b) To prove this part, it is enough to show that ε-View DP implies ε-DP in the adaptive continual release model.

Suppose that π is ε-View DP, *i.e.* $\mathcal{V}^\pi$ is ε-DP. Let $\mathcal{A}$ be a "reward-feeding" adversary, and $(a_1, \dots, a_T) \in [K]^T$ a sequence of arms.

Using Remark 3.10 and the notation defined there, we have:

$$\begin{aligned}
\mathbb{P}[\text{View}^{L,t^\star}_{\mathcal{A},\pi} = (a_1, \dots, a_T)] &= \mathcal{V}^\pi((a_1, \dots, a_T) \mid \mathcal{A}^{L,t^\star}(a_1, \dots, a_T)) \\
&\leq e^\varepsilon \mathcal{V}^\pi((a_1, \dots, a_T) \mid \mathcal{A}^{R,t^\star}(a_1, \dots, a_T)) \\
&= e^\varepsilon \mathbb{P}[\text{View}^{L,t^\star}_{\mathcal{A},\pi} = (a_1, \dots, a_T)]
\end{aligned}$$

where the inequality holds because $\mathcal{V}^\pi$ is DP, and $\mathcal{A}^{L,t^\star}(a_1, \dots, a_T)$ and $\mathcal{A}^{R,t^\star}(a_1, \dots, a_T)$ are neighbouring lists of rewards.

Finally, this means that π is ε-DP in the adaptive continual release model, since for pure DP, it is enough to check the atomic events $(a_1, \dots, a_T)$.

Note that the proof breaks if we consider composite events, which are necessary for approximate DP proofs. □

Summary of the relationship between definitions. We introduced three increasingly stronger input representations and their corresponding DP definitions: list of rewards with View DP, table of rewards with Table DP, and tree of rewards with DP in the adaptive continual release. These representations are summarised in Figure 3.5 for $T = 3$ and $K = 2$.

In general, DP in the adaptive continual release is stronger than Table DP, which is stronger than View DP. For ε-pure DP, these three definitions are equivalent, with the same privacy budget ε. More care is needed for other variants of DP, where going from one definition to another happens with a loss in the privacy budgets.

Besides reward representation, Interactive DP is a definition that deals with another degree of freedom: which actions are "fed" to the policy. In Interactive DP, these actions do not come from the policy itself like previous definitions, but from a "query-action" adversary that chooses the query action depending on the interaction history. This additional freedom helps protect users' privacy even if they do not follow the actions recommended by the policy. Also, the Interactive DP definition decouples the rewards from actions which helps to prove stronger group privacy properties. For policies closed under post-processing, *i.e.* policies that do not care about the source of the actions, Interactive DP is equivalent to its "normal" notion counterpart. In definition 3.5, we used the table of rewards as the reward representation, and thus, the "normal" counterpart is Table DP. It is also possible to use other reward representations for Interactive DP, *i.e.* lists or trees. All policies considered later are closed under post-processing.

3.7 Other DP Threat Models for Bandits

Here, we briefly discuss other threat models for bandits. Specifically, Table DP, View DP, and Interactive DP definitions are

(a) **global DP.** Our definitions adhere to the global DP formulation, where the users trust the centralised decision maker, i.e. the policy, with their private inputs. In the case of a clinical trial, we can imagine that the patients trust the doctors with their true input. In contrast, providing a **local DP** [DN03, EGS03, DMNS06, DJW13] formulation is possible where the users do not trust the centralised server. In local DP, each user uses a local perturbation mechanism to send a "noisy" version of the rewards to the policy. Though local DP provides stronger privacy as the policy has no access to the original rewards, it injects too much noise, leading to higher regret/sample complexity. Also, the fundamental hardness of local DP in bandits regarding regret lower bound and corresponding optimal algorithms are well-understood [ZCH+20]. Lastly, **shuffle DP** [Che21] is a trust model designed to attain the best of both worlds, i.e. removing the need to trust the data curator as in local DP, while providing utility similar to global DP. In the shuffle model, each user feeds their data again to a local perturbation. However, now they trust some entity to apply a uniformly random permutation on all user data. Thanks to privacy amplifications due to permutation, each user needs to add less noise to each data, and thus, the utility is better. On the other hand, shuffle DP replaces the need to trust a centralised server of global DP with the less strong assumption of trusting a randomised shuffler. **Shuffle DP** is studied for bandits in [HGFD22, GCPP22].

(b) **event-level DP.** Our definitions adpot an event-level neighbouring relation, where each user's data is supposed to appear in a single row of the dataset. On the other hand, in **user-level** neighbouring relation, a user's data can appear on multiple rows of the dataset. We refer to [DNPR10b] for an in-depth discussion of these two variants. In essence, any event-level DP algorithm could be transformed to a user-level DP algorithm using group privacy considerations [GKK+24].

(c) **over the entire sequence of actions.** Our defintions provides a privacy guarantee over the full sequence $(a_1, \ldots, a_T)$ of recommended actions. It is possible to define an **instantaneous** definition. Specifically, a policy π is ε-instantaneous DP if for all steps $t \in [T]$ and all neighbouring histories H_t and H'_t, and for all actions $a \in [K]$, we have that $\pi_t(a|H_t) \leq e^{\varepsilon}\pi_t(a|H_t)$. It is easy to show that if a policy is ε-instantaneous DP, then it is $(T\varepsilon)$-View DP.

Joint DP for contextual linear bandits. Until now, our definitions only consider the case when only the rewards are private quantities, and the contexts are either non-available or supposed to be non-private. The definition of Joint DP [SS18] is proposed for the case where contexts are private. First, we define neighbouring context-reward sequences.

Definition 3.16 (t-neighbouring context-reward sequences)**.** *Let $S \triangleq \{(\mathcal{A}_1, r_1), \ldots, (\mathcal{A}_T, r_T)\}$ and $S' \triangleq \{(\mathcal{A}'_1, r'_1), \ldots, (\mathcal{A}'_T, r'_T)\}$ be two context-reward sequences. S and S' are said to be* t-neighbours *if for all $s \neq t$ it holds that $(\mathcal{A}_s, r_s) = (\mathcal{A}'_s, r'_s)$.*

Definition 3.17 (JDP, [SS18])**.** *A randomised policy π for the contextual bandit problem is (ε, δ)-*Jointly Differentially Private *(JDP) if for any t and any pair of t-neighbouring context-reward sequences S and S', and any subset $E_{>t} \subset \mathcal{A}_{t+1} \times \mathcal{A}_{t+2} \times \cdots \times \mathcal{A}_T$ of sequence of actions ranging from step $t+1$ to the end of the sequence, it holds that*

$$\mathbb{P}\{\pi(S) \in E_{>t}\} \leq e^{\varepsilon}\mathbb{P}\{\pi(S') \in E_{>t}\} + \delta. \tag{3.3}$$

where $\pi(S)$ represents the sequence of actions recommended by the policy π when interacting with S, and $\mathbb{P}$ accounts only for randomness due to the policy.

JDP requires that changing the context at step t does not affect the actions chosen *only in the future rounds* $(> t)$, i.e. $(a_{t+1}, \ldots, a_T)$. In contrast, the standard notion of DP would require that the change does not affect the entire sequence of actions $(a_1, \ldots, a_T)$, including the action chosen at step t. Claim 13 of [SS18] shows that the standard notion of DP for linear contextual bandits, where both the reward and contexts are private, and the *entire* sequence of actions is published, always leads to linear regret. In addition, in the reduced model based on decision sets $\mathcal{A}_t$, the standard notion of DP is ill-defined, as it requires the entire sequence of actions to remain unchanged under any change in context-reward. This is true because two t-neighbouring context-reward sequences might yield different sets $\mathcal{A}_t$ and $\mathcal{A}'_t$. Since the action

a_t should be an element of the decision set at step t, i.e. $\mathcal{A}_t$ or $\mathcal{A}'_t$, then it is impossible to expect that a_t is unchanged between the two neighbouring cases.

DP for FC-BAI. Up till now, our definitions are tailored for bandit policies (Definition 2.16) used for the regret minimisation objective. An FC-BAI strategy (Definition 2.17) can be seen as an augmented "regret" policy. In addition to a sampling rule, an FC-BAI must determine when to stop the interaction (stopping rule) and recommend a final guess (recommendation rule). This means that Table DP, View DP and Interactive DP could be adapted to the FC-BAI setting by modifying the output of the mechanism. The output is changed from the recommended sequence of actions $a^T \triangleq (a_1, \dots, a_T)$ to $(a^T, \hat{a}, T)$ the sequence of sampled action a^T, final recommendation $\hat{a}$ and stopping time T. Also, the mechanisms should take a potentially "infinite" dataset as input since the final size of the dataset depends on when the FC-BAI strategy decides to stop.

For completeness, we provide a complete example of adaptations of Table DP and View DP for FC-BAI strategies. Let π^{BAI} be an FC-BAI strategy. We define the batch Table mechanism $\mathcal{M}^{\pi^{\text{BAI}}}$ as

$$\begin{aligned} \mathcal{M}^{\pi^{\text{BAI}}} : \ (\mathbb{R}^K)^\star &\to \mathcal{P}([K]^\star \times \mathbb{N}) \\ \mathbf{x} &\to \mathcal{M}_{\mathbf{x}}^{\pi^{\text{BAI}}}, \end{aligned}$$

where

$$\mathcal{M}_{\mathbf{x}}^{\pi^{\text{BAI}}}(a^T, \widehat{a}, T) \triangleq \text{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T^{\mathbf{x}}\right) \text{S}_{T+1}\left(\top \mid \mathcal{H}_T^{\mathbf{x}}\right) \prod_{t=1}^{T} \text{S}_t\left(a_t \mid \mathcal{H}_{t-1}^{\mathbf{x}}\right) \tag{3.4}$$

is the probability that the FC-BAI strategy π^{BAI} samples the sequence of actions a^T, recommends $\hat{a}$ and stops after T steps of interaction, and $\mathcal{H}_t^{\mathbf{x}} \triangleq (a_1, x_{1,a_1}, \dots, a_t, x_{t,a_t})$.

Similarly, we define the batch View mechanism $\mathcal{M}^{\pi^{\text{BAI}}}$ as

$$\begin{aligned} \mathcal{V}^{\pi^{\text{BAI}}} : \ (\mathbb{R})^\star &\to \mathcal{P}([K]^\star \times \mathbb{N}) \\ \mathbf{r} &\to \mathcal{V}_{\mathbf{x}}^{\pi^{\text{BAI}}}, \end{aligned}$$

where

$$\mathcal{V}_{\mathbf{r}}^{\pi^{\text{BAI}}}(a^T, \widehat{a}, T) \triangleq \text{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T^{\mathbf{r}}\right) \text{S}_{T+1}\left(\top \mid \mathcal{H}_T^{\mathbf{r}}\right) \prod_{t=1}^{T} \text{S}_t\left(a_t \mid \mathcal{H}_{t-1}^{\mathbf{r}}\right) \tag{3.5}$$

is the probability that the FC-BAI strategy π^{BAI} samples the sequence of actions a^T, recommends $\hat{a}$ as the final recommendation and stops after T steps of interaction, and $\mathcal{H}_t^{\mathbf{r}} \triangleq (a_1, r_1, \dots, a_t, r_t)$.

Then, π^{BAI} is a Table DP FC-BAI strategy if and only $\mathcal{M}^{\pi^{\text{BAI}}}$ is DP. Similarly, π^{BAI} is a View DP FC-BAI strategy if and only $\mathcal{V}^{\pi^{\text{BAI}}}$ is DP.

3.8 Conclusion

We define four ways of extending DP to bandits: Table DP, View DP, Interactive DP and DP in the adaptive continual release model. The first two definitions are in a non-interactive DP setting, where the inputs are fixed in advance and differ in the input considered due to the bandit feedback. Inteactive DP adapts the interactive threat model of [VZ22] to bandits, while DP in the adaptive continual release model adapts the threat model of [JRSS23] to bandits. We prove the relationship between our four definitions: Interactive DP is stronger than DP in adaptive continual release model, which is stronger than Table DP, which is in turn stronger than View DP. Finally, we discuss other threat models for privacy in bandits beyond our four definitions.

Chapter 4

Lower Bound Techniques

This chapter provides lower bounds on the regret and sample complexity of any policy satisfying DP. These lower bounds provide valuable insight into the inherent hardness of the problem and establish a target for optimal algorithm design. First, we illustrate the general idea for proving lower bounds in bandits. Then, we introduce coupling ideas that express the additional indistinguishability due to DP as upper bounds on the KL. We plug this KL upper bounds in classic lower bound proofs to generate new regret and sample complexity lower bounds, for bandits under ε-View DP and ρ-Interactive zCDP.

Contents

4.1 Lower Bounds for Bandits: Basic Ideas

The main idea to prove lower bounds in bandits is to reduce the bandit problem to hypothesis testing. Specifically, the goal is to provide two bandit instances that are conflicting:

(a) Actions that are good in one bandit instance are bad for the other bandit instance.

(b) The two bandit instances are hard to distinguish, *i.e.* if a policy interacts with one of the bandit instances, it cannot identify the true bandit instance with high statistical power.

The lower bounds emerge from the tension between these two requirements. To illustrate better this tension, let us revisit the proof of the regret minimax lower bound of Theorem 2.28.

Let π be a policy and consider the Gaussian environment ν_μ with means $\mu = (\Delta, 0, \dots, 0)$ and unit variances, where Δ is a constant we fix later. The interaction between π and ν_μ produces the canonical distribution over histories $\mathbb{P}_{\pi,\nu_\mu}$ and over sequence of actions $\mathbb{M}_{\pi,\nu_\mu}$ (*i.e.* Section 2.2.2), that we denote by $\mathbb{P}_\mu$ and $\mathbb{M}_\mu$ for brievity.

Consider $i = \arg\min_{a>1} \mathbb{E}_\mu[N_a(T)]$ the least played sub-optimal action in expectation, when π interacts with ν_μ. It is easy to show that $\mathbb{E}_\mu[N_i(T)] \leq \frac{T}{K-1}$ since $T = \mathbb{E}_\mu[N_1(T)] + \sum_{a>1} \mathbb{E}_\mu[N_a(T)] \geq (K-1)\mathbb{E}_\mu[N_i(T)]$. Then, to choose the second environment, we make action i optimal by considering the means to be $\mu' = (\Delta, 0, \dots, 0, 2\Delta, 0, \dots, 0)$, *i.e.* $\mu'_i = 2\Delta$ and $\mu'_j = \mu_j$ for all $j \neq i$. Again, the interaction between π and $\nu_{\mu'}$ produces the canonical distribution over histories $\mathbb{P}_{\pi,\nu_{\mu'}}$ and over sequence of actions $\mathbb{M}_{\pi,\nu_{\mu'}}$, that we denote by $\mathbb{P}_{\mu'}$ and $\mathbb{M}_{\mu'}$ for briecity.

The environment ν_μ and $\nu_{\mu'}$ verify the two conflicting constraints:

(a) Action i is optimal in $\nu_{\mu'}$ and suboptimal in ν_μ. Also, action 1 is optimal in ν_μ and suboptimal in $\nu_{\mu'}$.

(b) The environments ν_μ and $\nu_{\mu'}$ only differ at the mean of action i, which is an action rarely chosen by the policy when interacting with ν_μ. The hardness of distinguishing between ν_μ and $\nu_{\mu'}$ depends on the parameter Δ, which we finetune to get the tightest lower bounds.

To choose the Δ parameter, we move from the high-level idea and start doing some calculations. The first step is to use the regret decomposition combined with the Markov inequality to get that

$$\mathrm{Reg}_T(\pi, \nu_\mu) = (T - \mathbb{E}_\mu[N_1(T)])\,\Delta \geq \mathbb{M}_\mu(N_1(T) \leq T/2)\,\frac{T\Delta}{2},$$

and

$$\mathrm{Reg}_T(\pi, \nu_{\mu'}) = \Delta\,\mathbb{E}_{\mu'}[N_1(T)] + \sum_{a \notin \{1,i\}} 2\Delta\,\mathbb{E}_{\mu'}[N_a(T)] \geq \mathbb{M}_{\mu'}(N_1(T) > T/2)\,\frac{T\Delta}{2}.$$

Let us define the event $A \triangleq \{N_1(T) \leq T/2\} = \{(a_1, a_2, \ldots, a_T) : \text{card}(\{j : a_j = 1\}) \leq T/2\}$.

The second important step in the proof is applying the Bretagnolle Huber inequality to get:

$$\begin{aligned}\text{Reg}_T(\pi, \nu_\mu) + \text{Reg}_T(\pi, \nu_{\mu'}) &\geq \frac{T\Delta}{2}(\mathbb{M}_\mu(A) + \mathbb{M}_{\mu'}(A^c)) \\ &\geq \frac{T\Delta}{4}\exp\left(-D_{\text{KL}}\left(\mathbb{M}_\mu \,\|\, \mathbb{M}_{\mu'}\right)\right)\end{aligned}$$

The use of the Bretagnolle Huber inequality can be seen as the main "reduction step" to hypothesis testing. This inequality shows that the testing error gets bigger as the Kullback-Leibler (KL) divergence between the marginals gets smaller. This KL quantifies how distinguishable the two environments are. Providing a tight lower bound on the regret then boils down to providing a tight upper bound on the KL over marginals.

The data processing inequality provides an upper bound on the KL in the classic proofs of bandit lower bounds. Specifically, $D_{\text{KL}}\left(\mathbb{M}_\mu \,\|\, \mathbb{M}_{\mu'}\right) \leq D_{\text{KL}}\left(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}\right)$.

Then, the general KL decomposition lemma (Exercise 15.8, (b) in [LS20]) gives that

$$\begin{aligned}D_{\text{KL}}\left(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}\right) &= \mathbb{E}_\mu\left[\sum_{t=1}^{T} D_{\text{KL}}\left(P_{a_t} \,\|\, P'_{a_t}\right)\right] \\ &\overset{(a)}{=} \sum_{a=1}^{K} \mathbb{E}_\mu[N_a(T)] D_{\text{KL}}\left(P_a \,\|\, P'_a\right) \\ &\overset{(b)}{=} \mathbb{E}_\mu[N_i(T)] D_{\text{KL}}\left(\mathcal{N}(0,1) \,\|\, \mathcal{N}(2\Delta, 1)\right) \\ &\overset{(c)}{=} \mathbb{E}_\mu[N_i(T)] \frac{(2\Delta)^2}{2} \\ &\overset{(d)}{\leq} \frac{2T\Delta^2}{K-1}\end{aligned}$$

where (a) the KL decomposition lemma for finite-armed bandits (Lemma 15.1 in [LS20]), (b) is by the definition of μ and μ', (c) by the definition of the KL between Gaussians and (d) due to the choice of i.

All in all, we get

$$\text{Reg}_T(\pi, \nu_\mu) + \text{Reg}_T(\pi, \nu_{\mu'}) \geq \frac{T\Delta}{4}\exp\left(-\frac{2T\Delta^2}{K-1}\right)$$

We conclude the proof by optimizing for Δ to find that $\Delta = \sqrt{\frac{K-1}{4T}}$ gives the tightest lower bound.

When a policy satisfies a DP constraint, this translates into a stability condition between neighbouring rewards. The main technical challenge explored in this chapter is to express this additional indistinguishability condition as an upper bound on the KL between the marginals, *i.e.* $D_{\mathrm{KL}}\left(\mathbb{M}_{\mu} \parallel \mathbb{M}_{\mu'}\right)$.

4.2 Coupling Techniques for Lower bounds under DP

As portrayed in the previous section, to provide lower bounds on bandits with privacy, it is important to translate the privacy constraint to an upper bound on the KL between the marginals over the outputs, when the inputs are stochastically generated. In the following, we use coupling techniques to generate these upper bounds on the KL between marginals. We first explore the batch setting, where the data-generating distributions are product distributions. Then, we adapt the same techniques to the sequential setting of bandits.

4.2.1 KL Decomposition for product distributions under DP

Let $\mathcal{M}$ be a mechanism defined with the notation introduced in Definition 2.1. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two data-generating distributions over $\mathcal{X}^n$. We define the marginals M_1 and M_2 over the output of the mechanism $\mathcal{M}$ as

$$M_{\nu}(A) \triangleq \int_{D \in \mathcal{X}^n} \mathcal{M}_D\left(A\right) \, \mathrm{d}\mathcal{P}_{\nu}\left(D\right), \tag{4.1}$$

when the inputs are generated from $\mathcal{P}_{\nu}$ for $\nu \in \{1, 2\}$ and $A \in \mathcal{F}$.

The goal in this section is to provide an upper bound on the quantity $D_{\mathrm{KL}}\left(M_1 \parallel M_2\right)$ when the mechanism $\mathcal{M}$ satisfies DP.

Before showing the main result, we recall the definition of an f-divergence and its two main properties.

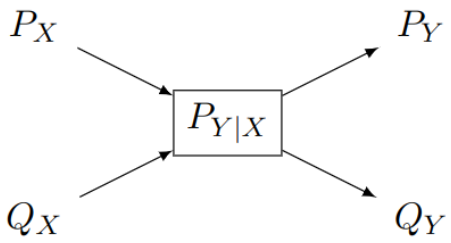
Definition 4.1 (f-divergence)**.** *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let P and Q be two probability distributions on a measurable space $(\mathcal{X}, \mathcal{F})$. If $P \ll Q$,* i.e. *P is absolutely continuous with respect to Q, then the f-divergence is defined as*

$$D_f(P \| Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative and $f(0) \triangleq f(0+)$.

The first property is the data-processing inequality.

Theorem 4.2 (Data processing). *Consider a channel that produces Y given X based on the conditional law $P_{Y|X}$, as shown below.*



Then,

$$D_f\left(P_Y \| Q_Y\right) \leq D_f\left(P_X \| Q_X\right).$$

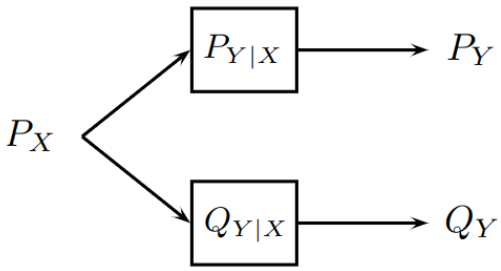
A consequence of the data processing inequality is that

$$D_f\left(M_1 \| M_2\right) \leq D_f\left(\mathcal{P}_1 \| \mathcal{P}_2\right),$$

for any mechanism $\mathcal{M}$ and any f-divergence.

To get a more interesting upper bound on the KL that captures the privacy constraint, we use the second property of f divergence,*i.e.* conditioning increases f-divergence.

Theorem 4.3 (Conditioning Increases f-divergence). *Let $P_X \overset{P_{Y|X}}{\longrightarrow} P_Y$ and $P_X \overset{Q_{Y|X}}{\longrightarrow} Q_Y$.*



Then,

$$D_f\left(P_Y \| Q_Y\right) \leq \mathbb{E}_{X \sim P_X}\left[D_f\left(P_{Y|X} \| Q_{Y|X}\right)\right].$$

To use this property for our goal, we will shift the vision from having two data-generating distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ and one mechanism "channel", into having only one data-generating distribution and two channels.

Define $\mathcal{C}$ as a coupling of $(\mathcal{P}_1, \mathcal{P}_2)$, *i.e.* the marginals of $\mathcal{C}$ are $\mathcal{P}_1$ and $\mathcal{P}_2$. We denote by $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ the set of all the couplings between $\mathcal{P}_1$ and $\mathcal{P}_2$. We consider the data-generating distribution to be the coupling $\mathcal{C}$ between $\mathcal{P}_1$ and $\mathcal{P}_2$. This means that sampling from $\mathcal{C}$ gives a pair of dataset (D, D'), where the marginal distribution of D is $\mathcal{P}_1$, and the marginal of D' is $\mathcal{P}_2$. Then, we consider two channels based on $\mathcal{M}$. The first channel applies the mechanism $\mathcal{M}$ only to D and ignores D', while the second channel applies the mechanism $\mathcal{M}$ only to D' and ignores D. In other words, using the notation of the figure in Theorem 4.3:

- $X = (D, D')$ a pair of datasets in $\mathcal{X}^n$
- the input distribution is $P_X = \mathcal{C}$ the coupling distribution.
- the first channel is the mechanism applied to the first dataset $P_{Y|X} = \mathcal{M}(Y \mid D)$.
- the second channel is the mechanism applied to the second dataset $Q_{Y|X} = \mathcal{M}(Y \mid D')$.
- Y is the output of the mechanism

Using this notation, we have that

- $P_Y = M_1$
- $Q_Y = M_2$
- $D_f\left(P_{Y|X} \| Q_{Y|X}\right) = D_f(\mathcal{M}_D \| \mathcal{M}_{D'})$.

Using Theorem 4.3, we get that

$$D_f(M_1 \| M_2) \leq \mathbb{E}_{(D,D')\sim\mathcal{C}}[D_f(\mathcal{M}_D \| \mathcal{M}_{D'})].$$

which is true for every coupling $\mathcal{C}$. Taking the infimum over the couplings provides the proof of our main theorem, which we summarise here:

Theorem 4.4. *For any mechanism $\mathcal{M}$ and any f-divergence, we have that*

$$D_f(M_1 \| M_2) \leq \inf_{\mathcal{C}\in\Pi(\mathcal{P}_1,\mathcal{P}_2)} \mathbb{E}_{(D,D')\sim\mathcal{C}}[D_f(\mathcal{M}_D \| \mathcal{M}_{D'})]. \tag{4.2}$$

Next, we need to upper-bound the $D_f(\mathcal{M}_D \| \mathcal{M}_{D'})$, when $\mathcal{M}$ satisfies DP. By considering the KL, which is an f-divergence for $f(x) = x\log(x)$, a direct consequence of Group Privacy (Proposition 2.8) gives the following corollary.

Corollary 4.5 (Group privacy and the KL)**.**

- *If $\mathcal{M}$ is ε-pure DP, then for any two datasets,*

$$D_{\mathrm{KL}}\left(\mathcal{M}_D \parallel \mathcal{M}_{D'}\right) \leq \varepsilon d_{\mathrm{Ham}}(D, D').$$

- *If $\mathcal{M}$ is ρ-zCDP, then*
$$D_{\mathrm{KL}}\left(\mathcal{M}_D \parallel \mathcal{M}_{D'}\right) \leq \rho d_{\mathrm{Ham}}(D, D')^2.$$

Combining Corollary with Theorem 4.4 gives the following theorem.

Theorem 4.6 (KL Upper Bound as a Transport Problem)**.**

- *If $\mathcal{M}$ is ε-pure DP, then*
$$D_{\mathrm{KL}}\left(M_1 \parallel M_2\right) \leq \varepsilon \inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(D,D') \sim \mathcal{C}}[d_{\mathrm{Ham}}(D, D')].$$

- *If $\mathcal{M}$ is ρ-zCDP, then*
$$D_{\mathrm{KL}}\left(M_1 \parallel M_2\right) \leq \rho \inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(D,D') \sim \mathcal{C}}[d_{\mathrm{Ham}}(D, D')^2].$$

Deriving the sharpest upper bound for the KL requires solving the transport problem

$$\inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(D,D') \sim \mathcal{C}}[d_{\mathrm{Ham}}(D, D')] \tag{4.3}$$

for ε-pure DP, and the transport problem

$$\inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(D,D') \sim \mathcal{C}}[d_{\mathrm{Ham}}(D, D')^2] \tag{4.4}$$

for ρ-zCDP.

As a proxy, we use maximal couplings.

Proposition 4.7. *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two probability distributions with the same σ-algebra. There exists a coupling $c_\infty(\mathcal{P}_1, \mathcal{P}_2) \in \Pi(\mathcal{P}_1, \mathcal{P}_2)$ called a maximal coupling, such that*

$$\mathbb{E}_{(X_1, X_2) \sim c_\infty(\mathcal{P}_1, \mathcal{P}_2)}\left[\mathbb{1}\left(X_1 \neq X_2\right)\right] = \mathrm{TV}\left(\mathcal{P}_1 \parallel \mathcal{P}_2\right)$$

Suppose that $\mathcal{P}_1$ and $\mathcal{P}_2$ are two product distributions over $\mathcal{X}^n$, *i.e.* $\mathcal{P}_1 = \bigotimes_{i=1}^n p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^n p_{2,i}$, where $p_{\nu,i}$ for $\nu \in \{1, 2\}$ and $i \in [1, n]$ are distributions over $\mathcal{X}$. Let c_∞^i be a maximal coupling between $p_{1,i}$ and $p_{2,i}$ for all $i \in [1, n]$. We define the coupling $\mathcal{C}_\infty \triangleq \bigotimes_{i=1}^n c_\infty^i$. Then $\mathcal{C}_\infty$ is a coupling of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Theorem 4.8 (KL Decomposition for Product Distributions)**.** *Using the $\mathcal{C}_\infty$ coupling between the product distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ as a proxy to solve the transport problems of Equation (4.3) and Equation (4.4), we show that:*

- *If $\mathcal{M}$ is ε-pure DP, then*
$$D_{\mathrm{KL}}\left(M_1 \parallel M_2\right) \leq \varepsilon \left(\sum_{i=1}^n t_i\right) \tag{4.5}$$

- *If $\mathcal{M}$ is ρ-zCDP, then*

$$D_{\mathrm{KL}}\left(M_1 \parallel M_2\right) \leq \rho\left(\sum_{i=1}^{n} t_i\right)^2 + \rho \sum_{i=1}^{n} t_i(1-t_i) \tag{4.6}$$

where $t_i \triangleq \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$.

Proof. Since $d_{\mathrm{Ham}}(D, D') = \sum_{i=1}^{n} \mathbb{1}\left(d_i \neq D'_i\right)$ we get that, for $(D, D') \sim \mathcal{C}_\infty$,

$$d_{\mathrm{Ham}}(D, D') \sim \sum_{i=1}^{n} \mathrm{Bernoulli}(t_i),$$

where $t_i \triangleq \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$, and the terms in the sum are mutually independent.

This further yields that

$$\mathbb{E}_{(D,D')\sim\mathcal{C}_\infty}[d_{\mathrm{Ham}}(D, D')] = \sum_{i=1}^{n} t_i,$$

and

$$\mathbb{E}_{(D,D')\sim\mathcal{C}_\infty}[d_{\mathrm{Ham}}(D, D')^2] = \left(\sum_{i=1}^{n} t_i\right)^2 + \sum_{i=1}^{n} t_i(1-t_i).$$

□

Stochastic generalisation of group privacy. Theorem 4.8 can be seen as a stochastic generalisation of the group privacy property of DP. Specifically, the results from Theorem 4.8 suggest that two random datasets D and D' sampled from $\mathcal{P}_1 = \bigotimes_{i=1}^{n} p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^{n} p_{2,i}$ respectively could be thought of as $(\sum_{i=1}^{n} t_i)$-neighboring datasets, where $t_i = \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$.

Relation to similar results in the literature. Lemma 6.1 in [KV18] shows that, for any event E, $M_1(E) \leq e^{6\varepsilon n \mathrm{TV}(p_1 \parallel p_2)} M_2(E)$, when the mechanism is ε-pure DP, and the data-generating distributions are i.i.d from p_1 or p_2, *i.e.* $\mathcal{P}_\nu = \bigotimes_{i=1}^{n} p_\nu$ for $\nu \in \{1, 2\}$. The Karwa Vadhan is a stronger result than Theorem 4.8 since it controls the multiplicative difference between the marginals at each event. This gives the following direct KL upper bound $D_{\mathrm{KL}}\left(M_1 \parallel M_2\right) \leq 6\varepsilon n \mathrm{TV}\left(p_1 \parallel p_2\right)$ for i.i.d distributions. Also, the Karwa Vadhan lemma builds explicitly the maximal coupling in their proof. Our result generalises this upper bound to product distributions and improves the dependence of factor 6 there. Also, it is worth noting that similar coupling ideas have been developed in [LGG22] to derive DP and zCDP variants of LeCam and Fano inequalities.

4.2.2 Sequential KL decomposition for bandits under DP

Now, we adapt Theorem 4.8 for the bandit marginals. Let $\nu = \{P_a, a \in [K]\}$ and $\nu' = \{P'_a, a \in [K]\}$ be two bandit instances. We recall that, when the policy π interacts with the bandit instance ν, it induces a marginal distribution $\mathbb{M}_{\nu\pi}$ over the sequence of actions, where

$$m_{\nu\pi}(a_1, \dots, a_T) \triangleq \int_{r_1,\dots,r_T} \prod_{t=1}^{T} \pi_t(a_t \mid H_{t-1}) p_{a_t}(r_t) \, \mathrm{d}r_t.$$

and for all $C \in \mathcal{P}([K]^T)$,

$$\mathbb{M}_{\nu\pi}(C) \triangleq \sum_{(a_1,\dots,a_T)\in C} m_{\nu\pi}(a_1, a_2, \dots, a_T).$$

We define $\mathbb{M}_{\nu'\pi}$ similarly.

The goal is to upper bound the quantity $D_{\mathrm{KL}}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right)$. The marginals $\mathbb{M}_{\nu\pi}$ and $\mathbb{M}_{\nu\pi}$ in the sequential setting "look like" marginals generated by "product distributions". However, the hardness of the sequential setting resides in the fact that the data-generating distributions depend on the actions chosen, which are stochastic. Thus, the results of the previous section cannot directly be applied. To adapt the proof ideas of the previous section to the bandit case, we introduce the idea of a coupled bandit instance.

Let $\nu = \{P_a : a \in [K]\}$ and $\nu' = \{P'_a : a \in [K]\}$ be two bandit instances. Define c_a as the maximal coupling between P_a and P'_a. Fix a policy $\pi = \{\pi_t\}_{t=1}^T$.

Here, we build a coupled environment γ of ν and ν'. The policy π interacts with the coupled environment γ up to a given time horizon T to produce an augmented history $\{(a_t, r_t, r'_t)\}_{t=1}^T$. The iterative steps of this interaction process are:

1. The probability of choosing an action $a_t = a$ at time t is dictated only by the policy π_t and $a_1, r_1, a_2, r_2, \dots, a_{t-1}, r_{t-1}$, *i.e.* the policy ignores $\{r'_s\}_{s=1}^{t-1}$.

2. The distribution of rewards (r_t, r'_t) is c_{a_t} and is conditionally independent of the previous observed history $\{(a_s, r_s, r'_s)\}_{t=1}^{t-1}$.

This interaction is similar to the interaction process of policy π with the first bandit instance ν, with the addition of sampling an extra r'_t from the coupling of P_{a_t} and P'_{a_t}. This, in essence, corresponds to the "up" branch in Theorem 4.3.

The distribution of the augmented history induced by the interaction of π and the coupled environment can be defined as

$$p_{\gamma\pi}(a_1, r_1, r'_1 \dots, a_T, r_T, r'_T) \triangleq \prod_{t=1}^{T} \pi_t(a_t \mid a_1, r_1, \dots, a_{t-1}, r_{t-1}) c_{a_t}(r_t, r'_t)$$

To simplify the notation, let $\mathbf{a} \triangleq (a_1, \dots, a_T)$, $\mathbf{r} \triangleq (r_1, \dots, r_T)$ and $\mathbf{r'} \triangleq (r'_1, \dots, r'_T)$. Also, let $c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'}) \triangleq \prod_{t=1}^{T} c_{a_t}(r_t, r'_t)$ and $\pi(\mathbf{a} \mid \mathbf{r}) \triangleq \prod_{t=1}^{T} \pi_t(a_t \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$. We put $\mathbf{h} \triangleq (\mathbf{a}, \mathbf{r}, \mathbf{r'})$.

With the new notation

$$p_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \triangleq \pi(\mathbf{a} \mid \mathbf{r}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})$$

Similarly, we define

$$q_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \triangleq \pi(\mathbf{a} \mid \mathbf{r'}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})$$

which corresponds to the "down" branch in Theorem 4.3, where the policy ignores the rewards $r_1, \dots, r_T$ in the interaction.

It follows that $m_{\nu\pi}$ is the marginal of $p_{\gamma\pi}$ when integrated over $(\mathbf{r}, \mathbf{r'})$, and $m_{\nu'\pi}$ is the marginal of $q_{\gamma\pi}$ when integrated over $(\mathbf{r}, \mathbf{r'})$, *i.e.*

$$m_{\nu\pi}(\mathbf{a}) = \int_{\mathbf{r},\mathbf{r'}} p_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \, \mathrm{d}\mathbf{r} \, \mathrm{d}\mathbf{r'}$$

and

$$m_{\nu'\pi}(\mathbf{a}) = \int_{\mathbf{r},\mathbf{r'}} q_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \, \mathrm{d}\mathbf{r} \, \mathrm{d}\mathbf{r'}.$$

By the data-processing inequality, we get that

$$D_{\mathrm{KL}}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right) \leq D_{\mathrm{KL}}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) \tag{4.7}$$

We have that

$$\begin{aligned} D_{\mathrm{KL}}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) &\overset{(a)}{=} \mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[\log \left(\frac{\pi(\mathbf{a} \mid \mathbf{r}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})}{\pi(\mathbf{a} \mid \mathbf{r'}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})} \right) \right] \\ &\overset{(b)}{=} \mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[\log \left(\frac{\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a})}{\mathcal{V}^{\pi}_{\mathbf{r'}}(\mathbf{a})} \right) \right] \end{aligned}$$

where: (a): by definition of $p_{\gamma\pi}$, $q_{\gamma\pi}$ and the KL divergence

(b): by definition of $\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a}) \triangleq \pi(\mathbf{a} \mid \mathbf{r})$ and $\mathcal{V}^{\pi}_{\mathbf{r'}}(\mathbf{a}) \triangleq \pi(\mathbf{a} \mid \mathbf{r'})$.

View DP. Now, if the policy π is ε-View DP, then by group privacy $\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a}) \leq e^{\varepsilon d_{\mathrm{Ham}}(\mathbf{r},\mathbf{r'})} \mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a})$, for any sequence of actions, and any two sequence of rewards $\mathbf{r}$ and $\mathbf{r'}$. Thus, computing the expectation of $d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r'})$ when $\mathbf{r}$ and $\mathbf{r'}$ are generated through the coupled environment provides the following theorem.

Theorem 4.9 (KL Decomposition for ε-View DP)**.** *If π is ε-View DP, then*

$$D_{\mathrm{KL}}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right) \leq \varepsilon \, \mathbb{E}_{\nu\pi} \left(\sum_{t=1}^{T} t_{a_t} \right),$$

where $t_{a_t} \triangleq \mathrm{TV}\left(P_{a_t} \parallel P'_{a_t}\right)$ *and* $\mathbb{E}_{\nu\pi}$ *is the expectation under* $m_{\nu\pi}$.

Proof. The proof follows by computing

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}}\left[\log\left(\frac{\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a})}{\mathcal{V}_{\mathbf{r'}}(\mathbf{a})}\right)\right] &\overset{(a)}{\leq} \mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}}\left[\varepsilon d_{\mathrm{Ham}}(\mathbf{r},\mathbf{r'})\right] \\
&\overset{(b)}{=} \varepsilon\sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}}\left[\mathbb{1}\left(r_t \neq r'_t\right)\right] \\
&\overset{(c)}{=} \varepsilon\sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}}\left[\mathbb{1}\left(r_t \neq r'_t\right)|a_t\right]\right] \\
&\overset{(d)}{=} \varepsilon\sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}}\left[t_{a_t}\right] \\
&\overset{(e)}{=} \varepsilon\sum_{t=1}^{T}\mathbb{E}_{\mathbf{a}\sim m_{\nu\pi}}\left[t_{a_t}\right]
\end{aligned}$$

where: (a) is by group privacy, (b) is by the definition of the hamming distance, (c) is by the towering property of the expectation, (d) is by the definition of the maximal coupling and (e) is because the sum only depends on the sequence of actions, with marginal distribution $m_{\nu\pi}$. □

Comparison to the product distribution setting: The result of Theorem 4.9 generalises the result of Theorem 4.8 to the sequential setting under pure DP. Since the actions are stochastic, there is an additional expectation over the generation process of the sequence of actions, sampled from $\mathbb{M}_{\nu\pi}$. Also, Theorem 4.9 can be seen as an ε-DP version of the general KL decomposition lemma (Exercice 15.8, (b) in [LS20]), which recall states that $D_{\mathrm{KL}}\left(\mathbb{P}_{\nu\pi} \parallel \mathbb{P}_{\nu'\pi}\right) = \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^{T} D_{\mathrm{KL}}\left(P_{a_t} \parallel P'_{a_t}\right)\right)$.

Remark 4.10 (Improvement of a factor of 6 comapred to [AB22]:)**.** *In Theorem of [AB22], we have showed that*

$$D_{\mathrm{KL}}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right) \leq 6\varepsilon\,\mathbb{E}_{\nu\pi}\left(\sum_{t=1}^{T} t_{a_t}\right).$$

We used a generalisation of Karwa Vadhan lemma to prove this result for product distributions. Using the coupled environment idea in the new proof, we eliminate the extra 6 *factor in the upper bound. This improves all the regret and sample complexity lower bounds in this manuscript by a factor of* 6 *compared to the results in [AB22, AJMB23].*

Remark 4.11 (Stopping time version of the KL decomposition for FC-BAI under View DP)**.** *Let* π^{BAI} *be a DP BAI strategy. Let* ν *and* λ *be two bandit instances. Denote by* $\mathbb{M}_{\boldsymbol{\nu},\pi^{BAI}}$ *the marginal distribution of* $(\underline{A}, \widehat{A}, \tau)$ *when the BAI strategy* π^{BAI} *interacts with* ν*. By adapting the techniques of*

Theorem 4.9 to the canonical bandit setting under FC-BAI, we get that

$$D_{\mathrm{KL}}\left(\mathbb{M}_{\nu,\pi^{BAI}} \,\middle\|\, \mathbb{M}_{\lambda,\pi^{BAI}}\right) \leq \varepsilon\, \mathbb{E}_{\nu,\pi^{BAI}}\left(\sum_{t=1}^{\tau} t_{a_t}\right)$$

where τ is the stopping time.

Interactive zCDP. Now, we suppose that the policy is ρ-Interactive zCDP.

First, we show a *strong* group privacy property.

Theorem 4.12 (Group Privacy for ρ-Interactive DP)**.** *If π is a ρ-Interactive zCDP policy then, for any sequence of actions $(a_1, \dots, a_T)$ and any two sequence of rewards $\boldsymbol{r} \triangleq \{r_1, \dots, r_T\}$ and $\boldsymbol{r'} \triangleq \{r'_1, \dots, r'_T\}$, we have that*

$$\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_t(. \mid H_{t-1}) \,\|\, \pi_t(. \mid H'_{t-1})\right) \leq \rho d_{\mathrm{Ham}}(\boldsymbol{r}, \boldsymbol{r'})^2$$

where $H_t \triangleq (a_1, r_1, \dots, a_t, r_t)$, $H'_t \triangleq (a_1, r'_1, \dots, a_t, r'_t)$ and $d_{\mathrm{Ham}}(\boldsymbol{r}, \boldsymbol{r'}) = \sum_{t=1}^{T} \mathbb{1}\left(r_t \neq r'_t\right)$.

Proof. Let $\mathbf{a} \triangleq (a_1, \dots, a_T)$ be a fixed sequence of actions. Let $\mathbf{r} \triangleq \{r_1, \dots, r_T\}$ and $\mathbf{r'} \triangleq \{r'_1, \dots, r'_T\}$ be two sequences of rewards.

Step 1: The constant adversary. We consider the constant adversary $B_{\mathbf{a}}$ defined as

$$B_{\mathbf{a}}(o_1, \dots, o_t) \triangleq a_t$$

i.e. $B_{\mathbf{a}}$ is the adversary that always queries at step t the action a_t, independently of the actions recommended by the policy. Let $\pi_{\mathbf{a}} \triangleq \pi^{B_{\mathbf{a}}}$ be the policy corresponding to the post-processing B_a.

Since π is ρ-Interactive zCDP, using Proposition 3.13, (b), then $\mathcal{M}^{\pi_{\mathbf{a}}}$ is ρ-zCDP. And Proposition 3.12, (c) gives that $\mathcal{V}^{\pi_{\mathbf{a}}}$ is ρ-zCDP.

Step 2: Group privacy of zCDP. Using the group privacy property of ρ-zCDP *i.e.* Theorem 2.8 with $\alpha = 1$, we get that

$$D_{\mathrm{KL}}\left(\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r}} \,\|\, \mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r'}}\right) \leq \rho\, d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r'})^2. \tag{4.8}$$

Step 3: Decomposing the view of the constant adversary. On the other hand, we have that

$$\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r}}(o_1, \dots, o_T) = \prod_{t=1}^{T} \pi_t(o_t \mid a_1, r_1, \dots, a_{t-1}, r_{t-1}).$$

In other words $\mathcal{V}_{\mathbf{r}}^{\pi_{\mathbf{a}}} = \bigotimes_{t=1}^{T} \pi_t(. \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1})$.

Similarly, $\mathcal{V}_{\mathbf{r}'}^{\pi_{\mathbf{a}}} = \bigotimes_{t=1}^{T} \pi_t(. \mid a_1, r'_1, \ldots, a_{t-1}, r'_{t-1})$.

Hence, we get

$$D_{\mathrm{KL}}\left(\mathcal{V}_{\mathbf{r}}^{\pi_{\mathbf{a}}} \,\|\, \mathcal{V}_{\mathbf{r}'}^{\pi_{\mathbf{a}}}\right) = \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_t(. \mid H_{t-1}) \,\|\, \pi_t(. \mid H'_{t-1})\right) \tag{4.9}$$

Plugging Equaion (4.9) in Inequality (4.8) concludes the proof. □

Remark 4.13 (Decoupling of the adversary provides stronger Group Privacy)**.** *The result of Theorem 4.12 is a strong group privacy property because it shows that the sum of the KL of each policy kernel π_t applied on any two histories H_T and H'_T is upper bounded by ρ times the hamming distance squared between the reward sequence.*

In contrast, if the policy was only View or Table DP, we can only upper bound the sum of the KLs in "expectation over the generation of the actions". Specifically, if π is ρ-View zCDP, then group privacy gives that $D_{\mathrm{KL}}\left(\mathcal{V}_{\boldsymbol{r}}^{\pi} \,\|\, \mathcal{V}_{\boldsymbol{r}'}^{\pi}\right) \leq \rho d_{\mathrm{Ham}}(\boldsymbol{r}, \boldsymbol{r}')^2$, for any two fixed sequences $\boldsymbol{r}$ and $\boldsymbol{r}'$ However,

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathcal{V}_{\boldsymbol{r}}^{\pi} \,\|\, \mathcal{V}_{\boldsymbol{r}'}^{\pi}\right) &= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{V}_{\boldsymbol{r}'}^{\pi}}\left[\log \frac{\mathcal{V}_{\boldsymbol{r}}^{\pi}(\boldsymbol{a})}{\mathcal{V}_{\boldsymbol{r}'}^{\pi}(\boldsymbol{a})}\right] \\
&= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{V}_{\boldsymbol{r}}^{\pi}}\left[\sum_{t=1}^{T} \log \frac{\pi_t(a_t \mid H_{t-1})}{\pi_t(a_t \mid H'_{t-1})}\right] \\
&= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{V}_{\boldsymbol{r}}^{\pi}} \sum_{t=1}^{T}\left[\mathbb{E}_{\boldsymbol{a}\sim\mathcal{V}_{\boldsymbol{r}'}^{\pi}}\left[\log \frac{\pi_t(a_t \mid H_{t-1})}{\pi_t(a_t \mid H'_{t-1})}\right] \mid A_1, \ldots, A_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{V}_{\boldsymbol{r}}^{\pi}}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_t(. \mid H_{t-1}) \,\|\, \pi_t(. \mid H'_{t-1})\right)\right].
\end{aligned}$$

Thus, the decoupling introduced by the adversary in the Interactive DP definition provides a stronger KL upper bound.

Computing the expectation of $d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r}')^2$ when $\mathbf{r}$ and $\mathbf{r}'$ are generated through the coupled argument provides the following theorem.

Theorem 4.14 (KL Decomposition for ρ-Interactive zCDP)**.** *If π is ρ-Interactive zCDP, then*

$$D_{\mathrm{KL}}\left(\mathbb{M}_{\nu\pi} \,\|\, \mathbb{M}_{\nu'\pi}\right) \leq \rho\, \mathbb{E}_{\nu\pi}\left[\left(\sum_{t=1}^{T} t_{a_t}\right)^2\right] + \rho\, \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^{T} t_{a_t}(1 - t_{a_t})\right)$$

where $t_{a_t} \triangleq \mathrm{TV}\left(P_{a_t} \,\|\, P'_{a_t}\right)$ and $\mathbb{E}_{\nu\pi}$ and $\mathbb{V}_{\nu\pi}$ is the expectation under $m_{\nu\pi}$ respectively.

Proof. First, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r}')\sim p_{\gamma\pi}}\left[\log\left(\frac{\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a})}{\mathcal{V}^{\pi}_{\mathbf{r}'}(\mathbf{a})}\right)\right] &\overset{(a)}{=} \sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\log\left(\frac{\pi_t(a_t \mid H_{t-1})}{\pi_t(a_t \mid H'_{t-1})}\right)\right] \\
&\overset{(b)}{=} \sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\log\left(\frac{\pi_t(a_t \mid H_{t-1})}{\pi_t(a_t \mid H'_{t-1})}\right) \mid H_{t-1}\right]\right] \\
&\overset{(c)}{=} \sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\mathbb{E}_{a_t\sim\pi_t(.|H_{t-1})}\left[\log\left(\frac{\pi_t(a_t \mid H_{t-1})}{\pi_t(a_t \mid H'_{t-1})}\right)\right]\right] \\
&\overset{(d)}{=} \sum_{t=1}^{T}\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[D_{\mathrm{KL}}\left(\pi_t(. \mid H_{t-1}) \,\|\, \pi_t(. \mid H'_{t-1})\right)\right] \\
&\overset{(e)}{\leq} \mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\rho d^2_{\mathrm{Ham}}(\mathbf{r},\mathbf{r}')\right]
\end{aligned}$$

where: (a) by definition of $\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a})$ and $\mathcal{V}^{\pi}_{\mathbf{r}'}(\mathbf{a})$.

(b) using the towering property of the expectation.

(c) using that, conditioned on the history H_{t-1}, the distribution of a_t is $\pi_t(. \mid H_{t-1})$.

(d) by definition of the KL divergence.

(e) by the strong group privacy for Interactive DP, *i.e.* Theorem 4.12.

The proof is direct by computing

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[d^2_{\mathrm{Ham}}(\mathbf{r},\mathbf{r}')\right] &\overset{(a)}{=} \mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[d^2_{\mathrm{Ham}}(\mathbf{r},\mathbf{r}') \mid \mathbf{A}\right]\right] \\
&\overset{(b)}{=} \mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[d_{\mathrm{Ham}}(\mathbf{r},\mathbf{r}') \mid \mathbf{A}\right]^2 + \mathbb{V}\left[d_{\mathrm{Ham}}(\mathbf{r},\mathbf{r}') \mid \mathbf{a}\right]\right] \\
&\overset{(c)}{=} \mathbb{E}_{\mathbf{h}\sim p_{\gamma\pi}}\left[\left(\sum_{t=1}^{T} t_{a_t}\right)^2 + \sum_{t=1}^{T} t_{a_t}(1-t_{a_t})\right] \\
&\overset{(d)}{=} \mathbb{E}_{\nu\pi}\left[\left(\sum_{t=1}^{T} t_{a_t}\right)^2\right] + \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^{T} t_{a_t}(1-t_{a_t})\right)
\end{aligned}$$

where we obtain (a) using the towering property of the expectation, (b) by definition of the variance and (c) uses that $d_{\mathrm{Ham}}(\mathbf{r},\mathbf{r}') = \sum_{t=1}^{T}\mathbb{1}\left(r_t \neq r'_t\right)$ where $\mathbb{1}\left(r_t \neq r'_t\right) \mid a_t \sim \text{Bernoulli}(t_{a_t})$ by the definition of the maximal coupling. □

Comparison to the product distribution setting: Again, the result of Theorem 4.14 generalises the result of Theorem 4.8 to the sequential setting under zCDP. Since the actions are stochastic,

there is an additional expectation over the generation process of the sequence of actions, sampled from $\mathbb{M}_{\nu\pi}$.

4.3 Regret Lower Bounds under DP

In this section, we answer the main question:

How much additional regret do we have to endure in bandits with DP compared to bandits without DP?

First, we answer this question for stochastic and linear bandit under ε-View DP by providing minimax and problem-dependent lower bounds. For ρ-Interactive zCDP, we provide only minimax lower bounds on the regret for stochastic and linear bandits. All these lower bounds are retrieved by plugging the upper bounds on the KL from the previous section.

4.3.1 Regret lower bounds under ε-View DP

We summarise the lower bounds presented in this section in Table 4.1.

Stochastic Bandits. First, we consider the stochastic bandit problem with K-arms, as in Section 2.2.2.

Minimax Regret. Under DP, the minimax regret is the lowest achievable regret by a View DP policy under the worst environment among a family of environments $\mathcal{E}^K$. Specifically, we define

$$\mathrm{Reg}^\star_{T,\varepsilon}(\mathcal{E}^K) = \inf_{\pi\in\Pi^\varepsilon} \sup_{\nu\in\mathcal{E}^K} \mathrm{Reg}_T(\pi,\nu)$$

where Π^ε is the class of all policies satisfying ε-View DP.

Theorem 4.15 (Minimax lower bound)**.** *Let $\mathcal{E}^K_{\mathcal{G}}$ be the set of K-armed Gaussian bandits, with unit variance. Then, for $K > 1$ and $T \geq K-1$,*

$$\mathrm{Reg}^\star_{T,\varepsilon}(\mathcal{E}^K_{\mathcal{G}}) \geq \max\left\{ \underbrace{\frac{1}{27}\sqrt{T(K-1)}}_{\text{without DP}}, \quad \underbrace{\frac{1}{22}\frac{K-1}{\varepsilon}}_{\text{with } \varepsilon\text{-View DP}} \right\}. \tag{4.10}$$

Proof. First, since $\Pi^\varepsilon \subset \Pi$, Theorem 2.28 gives that

$$\mathrm{Reg}^\star_{T,\varepsilon}(\mathcal{E}^K_{\mathcal{G}}) \geq \inf_{\pi\in\Pi} \sup_{\nu\in\mathcal{E}^K_{\mathcal{N}}} \mathrm{Reg}_T(\pi,\nu) \geq \frac{1}{27}\sqrt{T(K-1)}$$

To get the private term of the lower bound, we revisit the same proof structure explained in Section 4.1. Specifically, by choosing the same two environments as in Section 4.1 and

combining a Markov inequality then the Bretagnolle Huber inequality gives that

$$\mathrm{Reg}_T(\pi, \nu_\mu) + \mathrm{Reg}_T(\pi, \nu_{\mu'}) \geq \frac{T\Delta}{4} \exp\left(-D_{\mathrm{KL}}\left(\mathbb{M}_\mu \,\|\, \mathbb{M}_{\mu'}\right)\right) .$$

Now, we plug the KL decomposition under View DP of Theorem 4.9, which gives that

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{M}_{\nu\pi} \,\|\, \mathbb{M}_{\nu'\pi}\right) &\leq \varepsilon\, \mathbb{E}_\mu\left(\sum_{t=1}^{T} t_{a_t}\right) \\ &=^{(a)} \varepsilon\, \mathbb{E}_\mu[N_i(T)]\, \mathrm{TV}\left(P_i \,\|\, P_i'\right) \\ &\leq^{(b)} \varepsilon \frac{T}{K-1} \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathcal{N}(0,1) \,\|\, \mathcal{N}(2\Delta, 1)\right)} \\ &= \varepsilon \frac{T\Delta}{K-1} \end{aligned}$$

where: (a) is by definition of the environments ν_μ and $\nu_{\mu'}$ which only differ at the mean of arm i and (b) combines a Pinsker inequality to upper bound the TV, and the upper bound $\mathbb{E}_\mu[N_i(T)] \leq \frac{T}{K-1}$ due to the choice of i.

All in all, we get

$$\mathrm{Reg}_T(\pi, \nu_\mu) + \mathrm{Reg}_T(\pi, \nu_{\mu'}) \geq \frac{T\Delta}{4} \exp\left(-\varepsilon \frac{T\Delta}{K-1}\right)$$

By optimising for Δ, we choose $\Delta = \frac{K-1}{\varepsilon T}$. We conclude the proof by lower bounding $\exp(-1)$ with $\frac{8}{22}$, and using $2\max(a,b) \geq a + b$. □

Consequences of Theorem 4.15. Equation (4.10) shows two regimes of hardness: *high-privacy*, corresponding to lower values of ε, and *low-privacy*, corresponding to higher values of ε. In the high-privacy regime, specifically for $\varepsilon < \frac{22}{27}\sqrt{\frac{(K-1)}{T}}$, the hardness depends only on the number of the arms K and the privacy budget ε and is higher than the lower bound for bandits without privacy. In the low-privacy regime, *i.e.* for $\varepsilon \geq \frac{22}{27}\sqrt{\frac{(K-1)}{T}}$, the lower bound coincides with that of the bandits without privacy. This indicates the phenomena that for higher values of ε, *i.e.* signifying lower privacy, bandits with View DP are not harder than the bandits without privacy. Especially for significantly large T, the threshold between low and high privacy regimes is smaller than most of the practically used privacy budget values. For example, if $T = 10^6$ and $K = 100$, the bandits with and without View DP are equivalently hard for any privacy budget $\varepsilon \geq 0.01$. This shows that for stochastic bandits, we can deploy very low privacy budgets ε without losing anything in performance.

Problem-dependent Regret. Let π be a consistent ε-View DP policy, *i.e.* $\pi \in \Pi_{\mathrm{cons}}(\mathcal{E}) \cap \Pi^\varepsilon$.

Table 4.1 – Regret lower bounds for bandits with ε-View DP

	Minimax	**Problem Dependent**
Stochastic Multi-armed bandit	$\max\left(\frac{1}{27}\sqrt{T(K-1)}, \frac{1}{22}\frac{K-1}{\varepsilon}\right)$	$\sum_{a:\Delta_a>0} \frac{\Delta_a \log(T)}{\min(d_a, \varepsilon t_a)}$
Stoachastic Linear bandit	$\max\left(\frac{\exp(-2)}{8}d\sqrt{T}, \frac{\exp(-1)}{4}\frac{d}{\varepsilon}\right)$	$\inf_{\alpha\in[0,\infty)^{\mathcal{A}}} \sum_{a\in\mathcal{A}} \alpha(a)\Delta_a \log(T)$ s.t. $\|a\|^2_{H_\alpha^{-1}} \leq 0.5\Delta_a \min(\Delta_a, \varepsilon\rho(\mathcal{A}))$

Before deriving the lower bound, we first recall the KL inf quantity, which controls the problem-dependent hardness of the non-private bandit problem, *i.e.*

$$\mathrm{KL}_{\inf}(P, \mu^\star, \mathcal{M}) \triangleq \inf_{P'\in\mathcal{M}} \left\{ D_{\mathrm{KL}}\left(P \,\|\, P'\right) : \mu(P') > \mu^\star \right\}.$$

Similarly, we introduce a total variation version of the KL inf, which we call the TV inf, *i.e.*

$$\mathrm{TV}_{\inf}\left(P, \mu^\star, \mathcal{M}\right) \triangleq \inf_{P'\in\mathcal{M}} \left\{ \mathrm{TV}\left(P \,\|\, P'\right) : \mu\left(P'\right) > \mu^\star \right\}.$$

Theorem 4.16 (Problem-dependent Regret Lower Bound)**.** *Let $\mathcal{E} \triangleq \mathcal{M}_1 \times \cdots \times \mathcal{M}_K$ and $\pi \in \Pi_{cons}(\mathcal{E}) \cap \Pi^\varepsilon$ an ε-View DP consistent policy over $\mathcal{E}$. Then, for any $\nu = (P_a : a \in K) \in \mathcal{E}$,*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \nu)}{\log(T)} \geq \sum_{a:\Delta_a>0} \frac{\Delta_a}{\min\left(\underbrace{KL_{inf}\left(P_a, \mu^*, \mathcal{M}_a\right)}_{\text{without DP}}, \underbrace{\varepsilon\, TV_{inf}\left(P_a, \mu^*, \mathcal{M}_a\right)}_{\text{with } \varepsilon\text{-View DP}}\right)}. \tag{4.11}$$

Consequences of Theorem 4.16. We summarise interesting observations about the lower bound.

1. *Universality:* The lower bound of Theorem 4.16 holds for any environment with K arms and reward distributions with finite means. This is the first general lower bound for bandits with ε-View DP.

2. *For Bernoulli distributions of reward:* TV-distinguishability gap $\mathrm{TV}_{\inf}\left(P_a, \mu^*, \mathcal{M}_a\right) = \Delta_a$, and the KL-distinguishability gap $\mathrm{KL}_{\inf}\left(P_a, \mu^*, \mathcal{M}_a\right) \approx 2\Delta_a^2$. Thus, our problem-dependent lower bound reduces to $\Omega\left(\sum_{a\neq a^*} \frac{\log T}{\min\{\Delta_a, \varepsilon\}}\right)$. For Bernoulli rewards, our lower bound is able to retrieve the $\Omega\left(\frac{K\log T}{\varepsilon}\right)$ private lower bound of [SS18] with explicit constants.

3. *High and Low-privacy Regimes:* Like the minimax regret bound, the problem-dependent regret also indicates two clear regimes in regret due to high and low privacy (resp. small and large privacy budgets ε). *In the low-privacy regime,* i.e. *for $\varepsilon \geq \frac{KL_{inf}(P_a,\mu^*,\mathcal{M}_a)}{TV_{inf}(P_a,\mu^*,\mathcal{M}_a)}$, the regret achievable by bandits with View DP and without View DP are same.* Thus, there is no loss in performance due to privacy in this regime. *In the high-privacy regime,* i.e. *for $\varepsilon < \frac{KL_{inf}(P_a,\mu^*,\mathcal{M}_a)}{TV_{inf}(P_a,\mu^*,\mathcal{M}_a)}$, the regret depends on a coupled effect of privacy and partial information.* This effect is quantified by the inverse of the privacy budget times the inverse of the TV-distinguishability gap. Approximately, one can

think of $\mathrm{KL}_{\inf}(P_a, \mu^*, \mathcal{M}_a) \approx \Delta_a^2$ and $\mathrm{TV}_{\inf}(P_a, \mu^*, \mathcal{M}_a) \approx \Delta_a$, and thus the change of regimes happen at $\varepsilon \approx \Delta_a$.

The proof is similar to the proof of Theorem 4.15 and is deferred to Appendix B.1.

Stochastic Linear Bandits. We also derive new minimax and problem-dependent regret lower bounds for stochastic linear bandits [LS17].

In this section, we consider a linear bandit model with parameter $\theta \in \mathbb{R}^d$ and Gaussian noise. It implies that for an action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ the reward is $R_t = \langle A_t, \theta \rangle + \eta_t$, where $\eta_t \sim \mathcal{N}(0,1)$ is a sequence of independent Gaussian noises. The regret of a policy is

$$\mathrm{Reg}_T(\pi, \mathcal{A}, \theta) \triangleq \mathbb{E}_\theta \left[\sum_{t=1}^{T} \Delta_{A_t} \right],$$

where the suboptimality gap $\Delta_a \triangleq \max_{a' \in \mathcal{A}} \langle a' - a, \theta \rangle$, and $\mathbb{E}_\theta[\cdot]$ is the expectation with respect to the measure of outcomes induced by the interaction of the policy and the linear bandit determined by θ. Given this structure, we state the minimax and problem-dependent regret lower bounds for stochastic linear bandits.

Theorem 4.17 (Minimax regret lower bound)**.** *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = \mathbb{R}^d$. Let π be an ε-View DP policy. Then, there exists a vector $\theta \in \Theta$ such that*

$$\mathrm{Reg}_T(\pi, \mathcal{A}, \theta) \geq \max \left\{ \underbrace{\frac{\exp(-2)}{8} d\sqrt{T}}_{\text{without DP}}, \ \underbrace{\frac{\exp(-1)}{4} \frac{d}{\varepsilon}}_{\text{with } \varepsilon\text{-View DP}} \right\}. \tag{4.12}$$

Theorem 4.18 (Problem-dependent regret lower bound)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set spanning $\mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ be such that there is a unique optimal action. Then, for any consistent and ε-View DP policy π satisfies*

$$\liminf_{T \to \infty} \frac{\mathrm{Reg}_T(\pi, \mathcal{A}, \theta)}{\log(T)} \geq c(\mathcal{A}, \theta), \tag{4.13}$$

where the structural distinguishability gap *is the solution of a constraint optimisation*

$$c(\mathcal{A}, \theta) \triangleq \inf_{\alpha \in [0,\infty)^{\mathcal{A}}} \sum_{a \in \mathcal{A}} \alpha(a) \Delta_a, \text{ such that } \|a\|^2_{H_\alpha^{-1}} \leq \min \left\{ \underbrace{0.5\Delta_a^2}_{\text{without DP}}, \ \underbrace{0.5\varepsilon\rho_a(\mathcal{A})\Delta_a}_{\text{with } \varepsilon\text{-View DP}} \right\}$$

for all $a \in \mathcal{A}$ with $\Delta_a > 0$, $H_\alpha = \sum_{a \in \mathcal{A}} \alpha(a) a a^\top$, and a structure dependent constant $\rho_a(\mathcal{A})$.

Remarks. The minimax regret bound also shows a clear distinction between high and low-privacy regimes for $\varepsilon < 2e/\sqrt{T}$ and $\varepsilon \geq 2e/\sqrt{T}$. For the problem-dependent bound, the difference between high and low-privacy regimes is more subtle, and depends on the structure

of the problem. The proofs also plug the new KL upper bounds in the proofs in [LS17], and are deferred to Appendix B.2 and Appendix B.3.

Remark 4.19 (Lower bounds for ε-Table DP and ε-Interactive DP)**.** *Since* $\Pi^{\text{Interactive}} \subset \Pi^{\text{Table}} \subset \Pi^{\text{View}}$*, a lower bound on* ε*-View DP policies is a lower bound for both* ε*-Table DP and* ε*-Interactive DP policies.*

4.3.2 Regret lower bounds under ρ-Interactive zCDP

Let Π^{ρ}_{Int} be the set of all ρ-Interactive zCDP policies.

Theorem 4.20 (Minimax lower bound for finite-armed bandits)**.** *For any* $K > 1$*,* $T \geq K - 1$*, and* $0 < \rho \leq 1$*,*

$$\begin{aligned}\operatorname{Reg}^{\star}_{T,\rho}(\mathcal{E}^K_{\mathcal{G}}) &\triangleq \inf_{\pi \in \Pi^{\rho}_{Int}} \sup_{\nu \in \mathcal{E}^K_{\mathcal{G}}} \operatorname{Reg}_T(\pi, \nu) \\ &\geq \max\left\{ \underbrace{\frac{1}{27}\sqrt{T(K-1)}}_{\text{without DP}}, \quad \underbrace{\frac{1}{124}\sqrt{\frac{K-1}{\rho}}}_{\text{with } \rho\text{-Interactive zCDP}} \right\}.\end{aligned}$$

Theorem 4.21 (Minimax Lower Bounds for Linear Bandits)**.** *Let* $\mathcal{A} = [-1, 1]^d$ *and* $\Theta = \mathbb{R}^d$*. Then, we have that*

$$\begin{aligned}\operatorname{Reg}^{\star}_{T,\rho}(\mathcal{A}, \Theta) &\triangleq \inf_{\pi \in \Pi^{\rho}_{Int}} \sup_{\theta \in \Theta} \operatorname{Reg}_T(\pi, \mathcal{A}, \theta) \\ &\geq \max\left\{ \underbrace{\frac{e^{-2}}{8} d\sqrt{T}}_{\text{without DP}}, \quad \underbrace{\frac{e^{-2.25}}{4} \frac{d}{\sqrt{\rho}}}_{\text{with } \rho\text{-Interactive zCDP}} \right\}\end{aligned}$$

We summarise the lower bounds presented in this section in Table 4.2. The proofs are also found by plugging the KL upper bound of Theorem 4.14 in the classic proofs. The detailed proofs are deferred to Appendix B.4 and Appendix B.5.

Again, the minimax regret lower bounds suggest the existence of two hardness regimes depending on ρ and T. The change of regimes happens at $\rho \approx \frac{1}{T}$. On the other hand, for ε-View DP, the change of regimes happens at $\varepsilon \approx \frac{1}{\sqrt{T}}$. This is in accordance with the observation that an ε-DP mechanism is $\left(\frac{1}{2}\varepsilon^2\right)$-zCDP of Proposition 2.6.

Remark 4.22 (Lower bounds for View zCDP and Table zCDP)**.** *The lower bounds for* ρ*-zCDP are only provided for* ρ*-Interactive zCDP since we use the stronger group privacy property thanks to the decoupling from the adversary. An interesting open question is to provide lower bounds for the* ρ*-View zCDP and* ρ*-Table zCDP.*

Table 4.2 – Regret lower bounds for bandits with ρ-Interactive DP

	Minimax
Stochastic Multi-armed bandit	$\max\left(\frac{1}{27}\sqrt{T(K-1)}, \frac{1}{124}\sqrt{\frac{K-1}{\rho}}\right)$
Stoachastic Linear bandit	$\max\left(\frac{\exp(-2)}{8}d\sqrt{T}, \frac{\exp(-2.25)}{4}\frac{d}{\sqrt{\rho}}\right)$

4.4 Sample Complexity Lower Bounds under DP

The central question that we address in this section is

How many additional samples a BAI strategy must select for ensuring ε-View DP?

In response, we prove a lower bound on the sample complexity of any δ-correct ε-View DP BAI strategy.

First, we derive an ε-View DP variant of the transportation lemma, *i.e.* Lemma 1 in [KCG16].

Lemma 4.23 (Transportation lemma under ε-View DP)**.** *Let $\delta \in (0,1)$ and $\varepsilon > 0$. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$. For any δ-correct ε-View DP BAI strategy π^{BAI}, we have that*

$$\varepsilon \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi^{BAI}}\left[N_a(\tau)\right] \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \geq \mathrm{kl}(1-\delta,\delta),$$

where $\mathrm{kl}(1-\delta,\delta) \triangleq x \log \frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ *for* $x, y \in (0,1)$.

Proof. Let π^{BAI} be a δ-correct ε-View DP BAI strategy. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$.

Let $\mathbb{M}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}}$ denote the probability distribution of $(\underline{A}, \widehat{A}, \tau)$ when the BAI strategy π^{BAI} interacts with ν. For any alternative instance $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$, the data-processing inequality gives that

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{M}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}} \,\middle\|\, \mathbb{M}_{\boldsymbol{\lambda},\pi^{\mathrm{BAI}}}\right) &\geq \mathrm{kl}\left(\mathbb{M}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right), \mathbb{M}_{\boldsymbol{\lambda},\pi^{\mathrm{BAI}}}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right)\right) \\ &\geq \mathrm{kl}(1-\delta,\delta). \end{aligned} \tag{4.14}$$

where the second inequality is because π is δ-correct *i.e.* $\mathbb{M}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right) \geq 1-\delta$ and $\mathbb{M}_{\boldsymbol{\lambda},\pi}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right) \leq \delta$, and the monotonicity of the kl.

Now, using the stopping time version of the KL decomposition for FC-BAI, we get that

$$D_{\mathrm{KL}}\left(\mathbb{M}_{\nu,\pi^{\mathrm{BAI}}} \,\middle\|\, \mathbb{M}_{\lambda,\pi^{\mathrm{BAI}}}\right) \leq \varepsilon\, \mathbb{E}_{\nu,\pi^{\mathrm{BAI}}}\left(\sum_{t=1}^{\tau} t_{a_t}\right)$$
$$= \varepsilon \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}}\left[N_a(\tau)\right] \mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right). \tag{4.15}$$

Combining Inequality (4.14) and Inequality (4.15) concludes the proof. □

Leveraging Lemma 4.23, we derive a sample complexity lower bound for any ε-DP-FC-BAI strategy.

Theorem 4.24 (Sample complexity lower bound for ε-DP-FC-BAI)**.** *Let $\delta \in (0,1)$ and $\varepsilon > 0$. For any δ-correct ε-View DP BAI strategy π^{BAI}, we have that*

$$\mathbb{E}_{\boldsymbol{\nu},\pi^{BAI}}[\tau] \geq T^{\star}\left(\boldsymbol{\nu};\varepsilon\right)\log(1/3\delta), \tag{4.16}$$

where $\left(T^{\star}\left(\boldsymbol{\nu};\varepsilon\right)\right)^{-1} \triangleq \sup_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\nu})} \min\left(\underbrace{\sum_{a=1}^{K}\omega_a D_{\mathrm{KL}}\left(\nu_a \,\|\, \lambda_a\right)}_{\textit{without DP}}, \underbrace{\varepsilon\sum_{a=1}^{K}\omega_a \mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right)}_{\textit{with } \varepsilon\textit{-View DP}}\right).$

Comments on the lower bound. Similar to the lower bound for the non-private BAI [GK16], the lower bound of Theorem 2 is the value of a two-player zero-sum game between a MIN player and MAX player. MIN plays an alternative instance λ close to ν in order to confuse MAX. The latter plays an allocation $\omega \in \Sigma_K$ to explore the different arms, with the purpose of maximising the divergence between ν and the confusing instance λ that MIN played. On top of the KL divergence present in the non-private lower bound, our bound features the TV distance that appears naturally when incorporating the ε-View DP constraint. The proof is deferred to Appendix B. In order to compare the lower bound of an ε-View BAI strategy with the non-private lower bound of [GK16], we relax Theorem 4.24 to further derive a simpler bound, as in Corollary 4.25.

Proof. Let π be a δ-correct ε-global DP BAI strategy. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$.

Let $\mathbb{E}$ denote the expectation under $\mathbb{M}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}}$, ie $\mathbb{E} \triangleq \mathbb{E}_{\boldsymbol{\nu},\pi^{\mathrm{BAI}}}$.

By Lemma 4.23, we have that $\varepsilon\sum_{a=1}^{K}\mathbb{E}\left[N_a(\tau)\right]\mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right) \geq \mathrm{kl}(1-\delta,\delta)$.

On the other, Lemma 1 of [KCG16] gives that $\sum_{a=1}^{K}\mathbb{E}\left[N_a(\tau)\right] D_{\mathrm{KL}}\left(\nu_a \,\|\, \lambda_a\right) \geq \mathrm{kl}(1-\delta,\delta)$.

Since these two inequalities hold for all $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})$, we get

$$\begin{aligned}
\mathrm{kl}(1-\delta,\delta) &\leq \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})} \min\left(\varepsilon\sum_{a=1}^{K}\mathbb{E}\left[N_a(\tau)\right]\mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right), \sum_{a=1}^{K}\mathbb{E}\left[N_a(\tau)\right] D_{\mathrm{KL}}\left(\nu_a \,\|\, \lambda_a\right)\right) \\
&\overset{(a)}{=} \mathbb{E}[\tau] \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})} \min\left(\varepsilon\sum_{a=1}^{K}\frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]}\mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right), \sum_{a=1}^{K}\frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]} D_{\mathrm{KL}}\left(\nu_a \,\|\, \lambda_a\right)\right) \\
&\overset{(b)}{\leq} \mathbb{E}[\tau]\left(\sup_{\omega\in\Sigma_K}\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})} \min\left(\varepsilon\sum_{a=1}^{K}\omega_a \mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right), \sum_{a=1}^{K}\omega_a D_{\mathrm{KL}}\left(\nu_a \,\|\, \lambda_a\right)\right)\right) .
\end{aligned}$$

(a) is due to the fact that $\mathbb{E}[\tau]$ does not depend on $\boldsymbol{\lambda}$. (b) is obtained by noting that the vector $(\omega_a)_{a\in[K]} \triangleq \left(\frac{\mathbb{E}_{\nu,\pi}[N_a(\tau)]}{\mathbb{E}_{\nu,\pi}[\tau]}\right)_{a\in[K]}$ belongs to the simplex Σ_K.

The theorem follows by noting that for $\delta \in (0,1), \mathrm{kl}(1-\delta,\delta) \geq \log(1/3\delta)$. □

Corollary 4.25. *For any δ-correct ε-View DP BAI strategy π^{BAI}, we have that*

$$\mathbb{E}_{\boldsymbol{\nu},\pi^{BAI}}[\tau] \geq \max\left(\underbrace{T^\star_{\mathrm{KL}}(\boldsymbol{\nu})}_{\textit{without DP}}, \underbrace{\frac{1}{\varepsilon}T^\star_{\mathrm{TV}}(\boldsymbol{\nu})}_{\textit{with } \varepsilon\textit{-View DP}}\right)\log(1/3\delta),$$

where $\left(T^\star_{\boldsymbol{d}}(\boldsymbol{\nu})\right)^{-1} \triangleq \sup_{\omega\in\Sigma_K}\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})}\sum_{a=1}^{K}\omega_a \boldsymbol{d}(\nu_a,\lambda_a)$, and $\boldsymbol{d}$ is either KL *or* TV.

Proof. The proof is direct by observing that $T^\star\left(\boldsymbol{\nu};\varepsilon\right) \geq T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ and $T^\star\left(\boldsymbol{\nu};\varepsilon\right) \geq \frac{1}{\varepsilon}T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$. □

Comparison with the non-private lower bound. $T^\star_{\mathrm{KL}}$ is the characteristic time in the non-private lower bound [GK16], and we refer to Section 2.2 of [GK16] for a detailed discussion on its properties. The sample complexity lower bound suggests the existence of *two hardness regimes depending on ε, $T^\star_{KL}$ and $T^\star_{\mathrm{TV}}$*. (1) *Low-privacy regime*: When $\varepsilon > T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/(T^\star_{\mathrm{KL}}(\boldsymbol{\nu}))$, the lower bound retrieves the non-private lower bound, *i.e.* $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$, and thus, **privacy can be achieved for free**. (2) *High-privacy regime:* When $\varepsilon < T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$, the lower bound becomes $T^\star_{\mathrm{TV}}/\varepsilon$ and ε-View DP δ-BAI requires more samples than non-private ones.

In the following proposition, we characterise $T^\star_{\mathrm{TV}}$ for Bernoulli instances.

Proposition 4.26 (TV characteristic time for Bernoulli instances)**.** *Let ν be a bandit instance,* i.e. *such that $\nu_a = Bernoulli(\mu_a)$ and $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Let $\Delta_a \triangleq \mu_1 - \mu_a$ and $\Delta_{min} \triangleq \min_{a\neq 1}\Delta_a$. We have that*

$$T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) = \frac{1}{\Delta_{min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}, \qquad \textit{and} \qquad \frac{1}{\Delta_{min}} \leq T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) \leq \frac{K}{\Delta_{min}}.$$

Proof sketch. The proof is direct by solving the optimisation problem defining $T^\star_{\mathrm{TV}}$ and using that $\mathrm{TV}\left(\mathrm{Bernoulli}(p) \,\|\, \mathrm{Bernoulli}(q)\right) = |p-q|$. We refer to Appendix B.6 for details.

Comment. The aforementioned bound on TV characteristic time for Bernoulli instances is ε-View DP parallel of the KL-characteristic time bound $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \leq \sum_{a=1}^{K} \Delta_a^{-2}$ [GK16]. Using Pinsker's inequality, one can connect the TV and KL characteristic times by

$$T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) \geq \sqrt{2T^\star_{\mathrm{KL}}(\boldsymbol{\nu})}.$$

4.5 Discussion

We provided regret and sample complexity lower bounds for bandits under DP. All the lower bounds are retrieved by plugging a new upper bound on the KL between marginals, quantifying the extra indistinguishability due to privacy. We use coupling techniques to show that the sharpest upper bound on the KL between the marginal is retrieved by solving an optimal transport problem. Then, all the bandit lower bounds show that the hardness of bandits with privacy has two regimes. A low privacy regime where the hardness of bandits with privacy reduces to the hardness of non-private bandits. And a high privacy regime, where privacy has an extra cost. The change between the regimes happens at $\varepsilon \approx \Delta$, where Δ is the order of the mean gap.

In our proposed lower bounds, the Total Variation is the notion that controls the extra hardness of privacy in the high privacy regime, i.e. $\mathrm{TV}_{\mathrm{inf}}$ for regret and $T^\star_{\mathrm{TV}}$ for FC-BAI. The high-level intuition for the total variation being the quantity that characterises the problem is that Pure DP can be seen as a multiplicative stability constraint of e^ε when one data point changes. With group privacy, if two datasets differ in d_{ham} points, then one incurs a factor $e^{\varepsilon d_{\mathrm{Ham}}}$. Now, by sampling n i.i.d points from a distribution P and n i.i.d points from a distribution Q, these two dataset could be though as being $n\mathrm{TV}\left(P \,\|\, Q\right)$ neighboors in expectation, or $\sum_{t=1}^{n} \mathrm{TV}\left(P_i \,\|\, Q_i\right)$ for product distributions. This is a simple way to interpret the results of Theorem 4.8. In brief, *the total variation naturally appears in lower bounds since it is the quantity that characterises the hardness of the optimal transport problem minimising the hamming distance*, i.e. $TV(P,Q) = \inf_{(X,Y)\sim(P,Q)} E(1_{X\neq Y})$. On the other hand, other f-divergences may also characterize the hardness of the problem. An interesting open problem is to provide regret and sample complexity lower bounds for (ε,δ)-DP, where group privacy and coupling techniques are not tight.

Chapter 5

Algorithm Design

In this chapter, we provide regret and sample complexity upper bounds that match the lower bounds of Chapter 4. First, by considering a warm-up setting of finite-armed bandits with ε-pure View DP, we explain the main intuitions for a generic recipe to make bandit algorithms achieve DP near optimally: the algorithms run in phases, and the private statistics are computed on non-overlapping-sequences to add less Laplace/Gaussian noise. We instantiate this generic wrapper for regret minimisation algorithms under different settings and for best-arm identification algorithms. For each algorithm, we recall the specific setting, make explicit the details of the algorithm, and provide privacy and utility guarantees. We conclude the chapter with an experimental analysis of the different private algorithms, confirming the theoretical findings.

Contents

5.1 Introduction

The following sections introduce the main ingredients for designing near-optimal private bandit algorithms. However, before, we recall in this section the main two algorithms that were studied before in the private bandit literature: DP-UCB [MT15, TD16] and DP-SE [SS19].

DP-UCB [MT15, TD16] was the first DP version of the UCB algorithm. DP-UCB uses the tree-based mechanism [DNPR10b, CSS11] to compute privately the sum of rewards. For each arm, the tree mechanism maintains a binary tree of depth $\log(T)$ over the T streaming reward observations. At each step t, DP-UCB only updates the binary tree of the arm pulled a_t, then yields the private sum of the first $N_{a_t}(t)$ rewards from the root-to-the-leaf path in the arm-dependent binary tree. As a result, the noise added to the sum of rewards is $\mathcal{O}\left(\log(T)^{2.5}/\varepsilon\right)$ for Bernoulli/rewards in $[0,1]$. To apply the optimism in the face of uncertainty, DP-UCB builds a high probability upper bound on the means using the noisy sum of rewards to get the following private UCB index

$$\text{DP-UCB}_a^{\varepsilon}(t-1,\delta) \triangleq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(2/\delta)}{N_a(t-1)}} + \frac{Y_a(t-1)}{N_a(t-1)} + \frac{\sqrt{8}\log(T)^{3/2}\log(2/\delta)}{\varepsilon N_a(t-1)},$$

where $Y_a(t-1)$ is the noise added from the tree based mechanism, which correspnds to the sum of at most $\log(T)$ i.i.d Laplace noise terms, each of scale $\text{Lap}\left(\frac{\log(T)}{\varepsilon}\right)$. DP-UCB achieves ε-View DP for rewards in $[0,1]$ (*i.e.* $1/2$-subgaussian bandits) [1]. Analysing the regret of this index in the UCB meta-algorithm shows that DP-UCB yields a regret upper bound of

$$\text{Reg}_T(\text{DP-UCB}^{\varepsilon},\nu) \leq \frac{8\sqrt{8}K\log^{2.5}(T)}{\varepsilon} + 4\sum_{a=1}^{K}\Delta_a + \sum_{a:\Delta_a>0}\frac{4\log(T)}{\Delta_a}$$

On the other hand, for Bernoulli bandits, the additional regret lower bound due to privacy is $\Omega\left(\frac{K\log(T)}{\varepsilon}\right)$, as discussed in Section 4.3. This means that DP-UCB has an extra multiplicative $\log^{1.5}(T)$ regret compared to the lower bound.

DP-SE [SS19] was the first DP bandit algorithm to eliminate the additional multiplicative factor $\log(T)^{1.5}$ in the regret. DP-SE is a DP version of the Successive Elimination algorithm [EDMM02]. DP-SE runs in episodes: at each episode, the algorithm explores a set of active arms uniformly. At the end of each episode, DP-SE eliminates provably sub-optimal arms. Results from the concentration of the mean are used to determine the sub-optimal arms to eliminate. Also, to use better concentration inequalities, the algorithm runs in *independent* episodes: at the end of an episode, DP-SE only uses the samples collected at the current episode to compute the empirical means which decides the arms to eliminate. If we suppose that the rewards are in $[0,1]$, it is enough to add a noise of $\text{Lap}\left(\frac{1}{\varepsilon}\right)$ to each sum of arm rewards to make

[1] It is possible to show that DP-UCB satistifes the stronger notion of ε-Interactive DP.

the continual release of all the empirical means achieve ε-DP thanks to parallel composition (Lemma 2.10). Due to the addition of the Laplace noise to the sum of rewards, each arm is explored longer. The additional exploration at each phase can be derived from the concentration of Laplace random variables, and is responsible for the additional $\mathcal{O}\left(\log(T)/\varepsilon\right)$ in the regret, which matches the regret lower bound. However, DP-SE has two main drawbacks. The algorithm is not anytime since the DP-SE needs to know the T in advance to decide the length of each episode. Also, in general, Successive Elimination algorithms have the drawback of not matching the problem-dependent regret lower bound exactly and committing to one arm when all the other arms have been eliminated.

A careful analysis of DP-SE suggests that what made the algorithm get rid of the extra $\log(T)^{1.5}$ in the regret was the fact that the algorithm runs in independent episodes: the private means were only computed using the samples collected from that episode. In the next section, *we detect these ingredients and generalise them to propose a general framework to make bandit algorithms achieve privacy near-optimally.*

5.2 A Generic Wrapper: Warm-up Setting

This section focuses on the finite-armed stochastic bandit problem under ε-pure View DP and Bernoulli rewards, *i.e.* $r_t \in \{0,1\}$. This section aims to present the intuitions that lead to a generic blueprint for designing private bandit algorithms. We later generalise this blueprint for bandits under different settings and notions of DP. As discussed in Section 2.2.5, the UCB meta-algorithm (Algorithm 2) is the state-of-the-art optimal regret algorithm for finite-armed bandits. Thus, it is interesting to explore whether designing a private version of UCB that achieves the lower bound for private regret is possible.

The main challenge to making UCB private is the continual private release of the sum of rewards, or equivalently, the empirical means. UCB computes and stores a table of KT empirical means, *i.e.* T means for each arm. Using Simple Composition (Proposition 2.9), a first attempt to make UCB achieve ε-DP is to make each computed mean $\frac{\varepsilon}{KT}$-DP. Since rewards are in $\{0,1\}$, it is then enough to add noise $\mathrm{Lap}\left(\frac{KT}{\varepsilon N_a(t)}\right)$ to the empirical mean at step t, *i.e.* $\tilde{\mu}_a(t-1) \triangleq \hat{\mu}_a(t-1) + \mathrm{Lap}\left(\frac{KT}{\varepsilon N_a(t)}\right)$. Building a high probability upper bound using $\tilde{\mu}_a(t-1)$ gives the following UCB index $\tilde{\mu}_a(t-1) + \sqrt{\frac{2\log(2/\delta)}{N_a(t-1)}} + \frac{KT\log(2/\delta)}{\varepsilon N_a(t-1)}$. Adapting the analysis of UCB for this index shows that this private version of UCB yields linear regret.

A second attempt to design a private version is to consider the counting problem's structure and thus use the binary tree mechanism [DNPR10b, CSS11] to privately estimate the empirical means. This reduces exactly to the DP-UCB algorithm, presented in the Introduction section.

Table 5.1 – A comparison of ε-View DP algorithms for bandits.

Algorithm	Regret	# Private Means	Anytime	Forgetfulness
DP-UCB [MT15, TD16];	$\mathcal{O}\left(\frac{K\log(T)^{2.5}}{\varepsilon}+\sum_{a\neq a^*}\frac{\log(T)}{\Delta_a}\right)$	T	Yes	No
DP-SE [SS19]	$\mathcal{O}\left(\frac{K\log(T)}{\varepsilon}+\sum_{a\neq a^*}\frac{\log(T)}{\Delta_a}\right)$	$\mathcal{O}(\log(T))$	No	Yes
AdaP-UCB	$\mathcal{O}\left(\sum_{a\neq a^*}\frac{\Delta_a\log(T)}{\min(\Delta_a^2,\varepsilon\Delta_a)}\right)$	$\mathcal{O}(\log(T))$	Yes	Yes
AdaP-KLUCB	$\mathcal{O}\left(\sum_{a\neq a^*}\frac{\Delta_a\log(T)}{\min(d(\mu_a,\mu^*),\varepsilon\Delta_a)}\right)$	$\mathcal{O}(\log(T))$	Yes	Yes

As explained before, DP-UCB han an extra multiplicative $\log(T)^{1.5}$ compared to the lower bound. Is it possible to provide a private version of UCB that removes the extra $\log(T)^{1.5}$ term?

Going back to DP-SE, it seems that an important ingredient in designing the algorithm is that the private means were computed on non-overlapping sequences of rewards, *i.e.* DP-SE runs in independent episodes. Thus, using the parallel composition lemma (Lemma 2.10), it is enough for each computed mean to be ε-DP for the whole algorithm to be ε-DP. This is in contrast to the first attempts that need each mean to be $\varepsilon/(KT)$-DP using simple composition, or the sum of $\log(T)$ Laplace noise each of scale $\frac{\log(T)}{\varepsilon}$ for DP-UCB. The main observation to make UCB achieve privacy near optimally is that UCB also does not need to calculate the empirical mean at each step using all the rewards observed till that step. Specifically, it is possible to provide an episodic version of UCB, where the means are computed on non-overlapping sequences of rewards. We present the episodic version of UCB in algorithm 7 and an illustration of its execution in the following example.

Algorithm 7 An episodic version of UCB

Input: K number of arms, optimism parameter β, UCB indexes
Initialisation: Choose each arm once and let $t = K$
for $\ell = 1, 2, \ldots$ **do**
    Let $t_\ell = t + 1$
    Compute $a_\ell = \text{argmax}_{a\in[K]}\, \text{UCB}_a(t_\ell - 1, \beta)$
    Choose arm a_ℓ until round t such that $N_{a_\ell}(t) = 2N_{a_\ell}(t_\ell - 1)$
end for

Example 5.1 (Illustration of Algortihm 7.)**.** *To clarify the schematic, we illustrate a few steps of executing Algorithm 7 in Figure 5.1 for a bandit with only $K = 2$ arms. After playing each arm once, the first episode begins at t_1. We focus on step $t_4 = 7$ to observe the different ingredients. The index of Arm 1 at t_4 uses the private empirical mean $\frac{r_4+r_5}{2} + Lap\left(\frac{1}{2\varepsilon}\right)$ to build a high probability upper bound of the real mean μ_1 with confidence $t_4^{-\beta}$. The index of Arm 2 uses $r_6 + Lap\left(\frac{1}{\varepsilon}\right)$. If we assume that the index of Arm 1 is higher at t_4, Arm 1 is played for an entire episode from t_4 until $t_5 - 1$. The last time Arm 1 was played, the episode's length was 2. Thus, following t_4, the episode's length is doubled to*

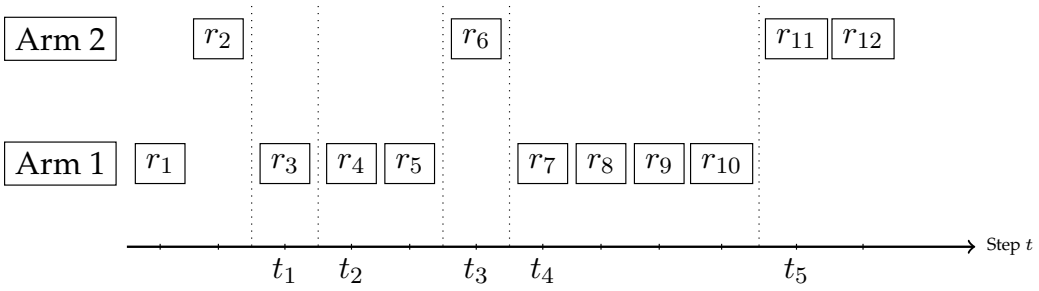


Figure 5.1 – An illustration of adaptive episodes with per-arm doubling.

4. *Again, at t_5, only the rewards r_7, r_8, r_9, r_10 are used to compute the index of Arm 1, and only r_6 is used to compute the index of Arm 2. This ensures that the indexes are computed on non-overlapping sequences of rewards.*

Our private version of UCB relies on three ingredients: *arm-dependent doubling*, *forgetting*, and *adding calibrated noise*. First, the algorithm runs in episodes. The *same arm* is played for a whole episode, and *double* the number of times it was last played. Second, at the beginning of a new episode, the index of arm a is computed only using samples from the last episode where arm a was played, *i.e.* the last *active* episode of arm a, while forgetting all the other samples. In a given episode, the arm with the highest index is played for all the steps. Due to these two ingredients, namely *doubling* and *forgetting*, each empirical mean is computed on non-overlapping sequences of rewards only needs to be ε-DP for the algorithm to be ε-View DP, avoiding the need of composition theorems.

Theorem 5.2. *If Algorithm 7 is instantiated with indexes that only use the private empirical mean of the rewards collected in the last active episode of arm a, then Algorithm 7 satisfies ε-View DP.* [2]

Proof. The main idea is that a change in reward will only affect the empirical mean calculated in one episode, which is made private using the Laplace Mechanism and Lemma 2.10. Since the actions are only calculated using the private empirical means, the algorithm is ε-View DP following the post-processing lemma. We refer to Appendix C.1.2 for a complete proof. □

To concretise an algorithm, we only need to explicitly explain how the indexes are calculated. Let $\hat{\mu}_a^\ell$ be the empirical mean reward of arm a computed using the samples collected between $t_{\psi_a(\ell)}$ and $t_{\psi_a(\ell)+1} - 1$. For an episode ℓ, $\psi_a(\ell) = \ell_a$ is the last active episode of arm a. In Example 5.1, $\psi_1(4) = 2$ and $\psi_2(4) = 3$. Thus, due to the doubling of episode length, the empirical mean corresponds to $\frac{1}{2}N_a(t_\ell - 1)$ samples of arm a. Since the rewards are in $[0, 1]$, the private empirical mean as $\tilde{\mu}_{a,\varepsilon}^\ell = \hat{\mu}_a^\ell + \text{Lap}\left(\frac{2}{\varepsilon N_a(t_\ell - 1)}\right)$ satisfies ε-DP (Theorem 2.13). Now, we want to ensure that $\text{I}_a^\varepsilon(t_\ell - 1, \beta)$, computed using only $\tilde{\mu}_{a,\varepsilon}^\ell$, is a high-probability upper bound

[2] In the following section, we show that this same algorithm achieves even the stronger notion of Interactive DP.

on the true mean. Here, we introduce two specific indexes that satisfy this criterion.

$$\text{For AdaP-UCB:} \qquad \mathrm{UCB}_a^{\varepsilon}(t_\ell - 1, \beta) = \tilde{\mu}_{a,\varepsilon}^{\ell} + \sqrt{\frac{\beta \log(t_\ell)}{2 \times \frac{1}{2} N_a(t_\ell - 1)}} + \frac{\beta \log(t_\ell)}{\varepsilon \times \frac{1}{2} N_a(t_\ell - 1)} \tag{5.1}$$

$$\text{For AdaP-KLUCB:} \qquad \mathrm{UCB}_a^{\varepsilon}(t_\ell - 1, \beta) = \max \left\{ q \in [0,1] : d\left(\breve{\mu}_{a,\varepsilon}^{\ell,\beta}, q\right) \leq \frac{\beta \log(t_\ell)}{\frac{1}{2} N_a(t_\ell - 1)} \right\} \tag{5.2}$$

where $\breve{\mu}_{a,\varepsilon}^{\ell,\beta} \triangleq \mathrm{Clip}_{0,1}\left(\tilde{\mu}_{a,\varepsilon}^{\ell} + \frac{\beta \log(t_\ell)}{\varepsilon \times \frac{1}{2} N_a(t_\ell - 1)}\right) \triangleq \min\{\max\{0, \tilde{\mu}_{a,\varepsilon}^{\ell} + \frac{\beta \log(t_\ell)}{\varepsilon \times \frac{1}{2} N_a(t_\ell - 1)}\}, 1\}$ is the private empirical mean clipped between zero and one.

Theorem 5.3 (Regret Analysis of AdaP-UCB)**.** *For rewards in* $[0,1]$, AdaP-UCB *satisfies* ε*-global DP, and for* $\beta > 3$*, it yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-UCB}, \nu) \leq \sum_{a:\Delta_a > 0} \left(\frac{16\beta}{\min\{\Delta_a, \varepsilon\}} \log(T) + \frac{3\beta}{\beta - 3} \right).$$

Theorem 5.4 (Regret Analysis of AdaP-KLUCB)**.** *When the rewards are sampled from Bernoulli distributions,* AdaP-KLUCB *satisfies* ε*-global DP, and for* $\beta > 3$ *and constants* $C_1(\beta), C_2 > 0$*, it yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-KLUCB}, \nu) \leq \sum_{a:\Delta_a > 0} \left(\frac{C_1(\beta)\Delta_a}{\min\{d_{\inf}(\mu_a, \mu^*), C_2 \varepsilon \Delta_a\}} \log(T) + \frac{3\beta}{\beta - 3} \right).$$

In appendix C.2.4, we also derive problem-independent or minimax regret upper bounds for AdaP-UCB and AdaP-KLUCB, which are of order $\mathcal{O}\left(\sqrt{KT \log(T)} + \frac{K \log(T)}{\varepsilon}\right)$.

Proof Sketch. Our two algorithms AdaP-UCB and AdaP-KLUCB have three main differences compared to the vanilla UCB algorithms of Section 2.3. (a) They run in arm-dependent doubling. We show that the effect of this change is a multiplicative $\times 2$ in the regret compared to Vanilla UCB. The reason is that if the vanilla UCB would have played the sub-optimal arm a for n_a step, the arm-dependent doubling version will play that sub-optimal arm at most $2 \times n_a$ since the algorithm cannot stop in the middle of a phase. (b) They forget samples. At each phase, the algorithm only uses samples from the last active phase. This means that the algorithm "throws" half of the samples. This has an additional multiplicative $\times 2$ effect on the regret. (c) They have new UCB indexes with new exploration bonuses. This is the step where the effect of privacy intervenes. The bonus can be written as the sum of two terms. A first non-private bonus, retrieved by concentration inequalities over the empirical mean. A second private bonus, retrieved by concentration inequalities over the Laplace noise. Using the same techniques as in the proof of Theorem 2.23, we define the "good" event that all the means are well estimated within the noisy empirical means, which happens with probability $1 - \delta$. Then, under this good event, our private versions of UCB stop sampling a sub-optimal

arm a as soon as the confidence width is smaller than the mean gap. Specifically, for example for AdaP-UCB, if $\tilde{n}_a$ is the number of times that AdaP-UCB samples a sub-optimal arm a, then $\tilde{n}_a$ verifies that $2\sqrt{\frac{2\log(2/\delta)}{n_a}} + 2\frac{\log(2/\delta)}{\varepsilon \tilde{n}_a} \leq \Delta_a$. Solving for $\tilde{n}_a$ and replacing $\delta = 1/T$ gives the problem-dependent regret upper bound. Optimising for the worst-case instance retrieves the gap-free upper bound. Another way of proving the gap-free bound is using the fact that the regret is upper bounded by the sum of the confidence width. The sum of the non-private part of the bonus gives the classic $\sqrt{T}$ upper bound in the regret. The sum of the private part can be dealt with similarly to the non-private part. Specifically, for example for AdaP-UCB, the sum of the private bonus is approximately $\sum_t \frac{2\log(t)}{\varepsilon N_{a_t}(t-1)}$. Again, using that $N_{a_t}(t-1) \approx t$, and $\sum_t \frac{1}{t} \approx \int_t \frac{1}{t} dt = \log(T)$ gives an intuition on the additional $\log(T)/\varepsilon$ part in the private upper bond. We provide a generic analysis of Algorithm 7 in Appendix C.2.2, and an instanciation of the proof for AdaP-UCB and AdaP-KLUCB in Appendix C.2.3. We also provide a complete proof for the gap-free regret upper bound in Appendix C.2.4.

The regret upper bound of Theorem 5.4 matches the problem-dependent regret lower bound of Theorem 4.16, for Bernoulli bandits, up to constants. The minimax regret upper bound matches the minimax regret lower bound of Theorem 4.15 up to logarithmic terms in the horizon T. Also, the upper bounds reflect the same two privacy regimes observation from the lower bounds *i.e.* in the low-privacy regime the regrets of AdaP-UCB and AdaP-KLUCB are independent of ε, and in the high-privacy regime, they depend on ε and Δ_a.

Generic Blueprint. Here, we detect the main steps to design a near-optimal private bandit algorithm: (1) Characterise the main private quantity. This corresponds to the empirical mean of rewards $\hat{\mu}$ for finite-armed bandits, or the least-square estimate $\hat{\theta}$ in linear bandits, (2) Design an episodic version of the bandit algorithm that computes the main private quantity on non-overlapping input sequences. This corresponds to the arm-dependent doubling trick for UCB. (3) Add calibrated noise to the quantity of interest using Parallel Composition (Lemma 2.10). This helps to add less noise. (4) Calibrate for the addition of the noise in your algorithm. This corresponds to adapting the exploration bonus in UCB or exploring the arms more in DP-SE. This is the step where the concentration results of the noise are used. (5) Quantify the effect of the noise addition. For AdaP-UCB and AdaP-KLUCB, this corresponds to a multiplicative 4 in the non-private regret due to arm-dependent doubling and forgetting, and an additional $\mathcal{O}\left(K \log(T)/\varepsilon\right)$ private regret due to the additional privacy bonus.

5.3 Private Algorithms for Regret Minimisation

In this section, we instantiate the generic blueprint of Section 5.2 for three bandit settings under ρ-Interactive DP: finite-armed bandits, linear bandits and contextual linear bandits.

5.3.1 Finite-armed bandits

We revisit the finite-armed setting of Section 5.2 under ρ-Interactive zCDP. Let $\nu = (P_a : a \in [K])$ be a bandit instance with K arms and means $(\mu_a)_{a\in[K]}$. The goal is to design a ρ-Interactive zCDP policy π that maximises the cumulative reward, or minimises regret over a horizon T:

$$\mathrm{Reg}_T(\pi, \nu) \triangleq T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[N_a(T)\right] \tag{5.3}$$

Here, $\mu^\star \triangleq \max_{a\in[K]} \mu_a$ is the mean of the optimal arm $a^\star$, $\Delta_a \triangleq \mu^\star - \mu_a$ is the sub-optimality gap of the arm a and $N_a(T) \triangleq \sum_{t=1}^{T} \mathbb{1}\left(a_t = a\right)$ is the number of times the arm a is played till T, where the expectation is taken both on the randomness of the environment ν and the policy π.

We present AdaC-UCB, an extension of the generic algorithmic wrapper proposed in Section 5.2 to ρ-Interactive zCDP. Again, AdaC-UCB relies on the three ingredients : *arm-dependent doubling*, *forgetting*, and *adding calibrated noise*. The only difference is that AdaC-UCB adds calibrated Gaussian noise to achieve zCDP (Theorem 2.14).

For AdaC-UCB, we use the following private index to select the arms (Line 6 of Algorithm 7) as

$$\mathrm{I}_a^\rho(t_\ell - 1, \beta) \triangleq \hat{\mu}_a^\ell + \mathcal{N}\left(0,\, \sigma^2_{a,\ell,\rho}\right) + B_a(t_\ell - 1, \beta, \rho). \tag{5.4}$$

Here, $\hat{\mu}_a^\ell$ is the empirical mean of rewards collected in the last episode in which arm a was played. The variance of the Gaussian noise is

$$\sigma^2_{a,\ell,\rho} \triangleq \frac{1}{2\rho \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}$$

and the exploration bonus $B_a(t_\ell - 1, \beta, \rho)$ is defined as

$$B_a(t_\ell - 1, \beta, \rho) \triangleq \sqrt{\left(\frac{1}{2 \times \frac{1}{2}N_a(t_\ell - 1)} + \frac{1}{\rho \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}\right)\beta \log(t_\ell)}.$$

The term in blue rectifies the non-private confidence bound of UCB for the added Gaussian noise.

Theorem 5.5 (Privacy of AdaC-UCB)**.** *For rewards in* $[0, 1]$, AdaC-UCB *satisfies* ρ*-Interactive zCDP.*

Proof. The main intuition is that a change in one input in the table of rewards only affects the view of the adversary at one episode, which is made ρ-zCDP using the Gaussian mechanism and parallel composition (Lemma 2.10). In appendix C.1.2, we provide a generic proof of the

Interactive DP guarantee for any algorithm which uses the generic blueprint of Section 5.2. The proof of this theorem is a direct consequence, that we specify exactly in Appendix C.1.3. □

Theorem 5.6 (Regret analysis of AdaC-UCB)**.** *For rewards in* $[0,1]$ *and* $\beta > 3$, AdaC-UCB *yields a problem-dependent regret upper bound of*

$$\sum_{a:\Delta_a>0}\left(\frac{8\beta}{\Delta_a}\log(T) + 8\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{2\beta}{\beta-3}\right).$$

and a gap-free regret upper bound of

$$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(K\sqrt{\frac{\log(T)}{\rho}}\right).$$

Proof Sketch. The proof shares the same steps as the proof of the upper bounds in Theorem 5.3 and Theorem 5.4. The only difference is a new "private bonus" part in the UCB index due to the concentration of Gaussian noise, in contrast to the concentration of Laplace noise for AdaP-UCB and AdaP-KLUCB. We present a generic analysis of Algorithm 7 in Appendix C.2.2, and instanciate it for AdaC-UCB in Appendix C.2.6.

AdaC-UCB minimax regret upper bound matches the regret lower bound of Theorem 4.20 up to a $\sqrt{\log(T)}$ term. Also, there is a multiplicative $\sqrt{K}$ gap between the upper and lower bounds.

Remark 5.7 (Extensions to (ε,δ)-Interactive DP and (α,ε)-Interactive RDP)**.** *The difference comes from the different calibrations of the Gaussian Mechanism* (*Thm 2.14*)*. Adapting the analysis from* ρ*-zCDP reduces to changing the* $\frac{1}{2\rho}$ *factor to* $\frac{2}{\varepsilon^2}\log(\frac{1.25}{\delta})$ *for* (ε,δ)*-DP and to* $\frac{\alpha}{2\varepsilon}$ *for* (α,ε)*-RDP,* i.e. *varying the constant* b *in Theorem 2.14.*

5.3.2 Stochastic linear bandits

Here, we study ρ-Interactive zCDP for stochastic linear bandits with a finite number of arms. We consider that a fixed set of actions $\mathcal{A} \subset \mathbb{R}^d$ is available at each round, such that $|\mathcal{A}| = K$. The rewards are generated by a linear structural equation. Specifically, at step t, the observed reward is $r_t \triangleq \langle \theta^\star, a_t\rangle + \eta_t$, where $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is a conditionally 1-subgaussian noise, *i.e.* $\mathbb{E}\left[\exp\left(\lambda\eta_t\right) \mid a_1, \eta_1, \ldots, a_{t-1}\right] \leq \exp\left(\lambda^2/2\right)$ almost surely for all $\lambda \in \mathbb{R}$.

For any horizon $T > 0$, the regret of a policy π is

$$\mathrm{Reg}_T(\pi, \mathcal{A}, \theta^\star) \triangleq \mathbb{E}_{\theta^\star}\left[\sum_{t=1}^{T}\Delta_{A_t}\right], \tag{5.5}$$

where suboptimality gap $\Delta_a \triangleq \max_{a' \in \mathcal{A}} \langle a' - a, \theta^\star \rangle$. $\mathbb{E}_{\theta^\star}[\cdot]$ is the expectation with respect to the measure of outcomes induced by the interaction of π and the linear bandit environment $(\mathcal{A}, \theta^\star)$.

We propose AdaC-GOPE (Algorithm 8), which is a ρ-Interactive zCDP extension of the G-Optimal design-based Phased Elimination (GOPE) algorithm, *i.e.* Algorithm 12 in [LS20]. AdaC-GOPE is a phased elimination algorithm. At the end of each episode ℓ, AdaC-GOPE eliminates the arms that are likely to be sub-optimal, *i.e.* the ones with an empirical gap exceeding the current threshold ($\beta_\ell = 2^{-\ell}$). The elimination criterion only depends on the samples collected in the current episode. In addition, the actions to be played during an episode are chosen based on the solution of an optimal design problem (Equation (5.7)) that helps to exploit the structure of arms and to minimise the number of samples needed to eliminate a sub-optimal arm.

In particular, if π_ℓ is the G-optimal solution (Definition 5.8) for $\mathcal{A}_\ell$ at phase ℓ, then each action $a \in \mathcal{A}_\ell$ is played $T_\ell(a) \triangleq \lceil c_\ell \pi_\ell(a) \rceil$ times, where for $\delta_{K,\ell} \triangleq \frac{\delta}{K\ell(\ell+1)}$ and $f(d,\delta) \triangleq d + 2\sqrt{d \log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)$,

$$c_\ell \triangleq \frac{8d}{\beta_\ell^2} \log\left(\frac{4}{\delta_{K,\ell}}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2}{\rho} f\left(d, \delta_{K,\ell}\right)} \tag{5.6}$$

The term in blue is the additional length of the episode to compensate for the noisy statistics used to ensure privacy. The samples collected in the current episode do not influence which actions are played in it. This decoupling allows (a) the use of the tighter confidence bounds available in the fixed design setting (Appendix C.3.1) and (b) avoiding privacy composition theorems and using, therefore, Lemma 2.10 to make the algorithm private. Note that AdaC-GOPE can be seen as a generalisation of DP-SE [SS19] to the linear bandit setting.

Here, we present the definitions of optimal design and a classic equivalence result required to state Algorithm 8.

Definition 5.8 (Optimal design [LF23])**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ and $\pi : \mathcal{A} \to [0,1]$ be a distribution on $\mathcal{A}$ so that $\sum_{a \in \mathcal{A}} \pi(a) = 1$. Let $V(\pi) \in \mathbb{R}^{d \times d}$ and $f(\pi), g(\pi) \in \mathbb{R}$ be given by*

$$V(\pi) \triangleq \sum_{a \in \mathcal{A}} \pi(a) a a^T, \quad f(\pi) \triangleq \log\det V(\pi), \quad g(\pi) \triangleq \max_{a \in \mathcal{A}} \|a\|_{V(\pi)^{-1}}.$$

- *π is called a **design**.*
- *The set $\mathrm{Supp}(\pi) \triangleq \{a \in \pi : \pi(a) \neq 0\}$ is called the **core set** of $\mathcal{A}$.*
- *A design that maximises f is called a **D-optimal design**.*
- *A design that minimises g is called a **G-optimal design**.*

Theorem 5.9 (Kiefer–Wolfowitz theorem [KW60]). *Assume that $\mathcal{A}$ is compact and $span(\mathcal{A}) = \mathbb{R}^d$. The following are equivalent*

- *$\pi^\star$ is a minimiser of g,*
- *$\pi^\star$ is a maximiser of f, and*
- *$g(\pi^\star) = d$.*

Also, there exists a minimiser $\pi^\star$ of g such that $|\mathrm{Supp}\,(\pi^\star)| \leq \frac{d(d+1)}{2}$.

Algorithm 8 AdaC-GOPE

Input: Privacy budget ρ, $\mathcal{A} \subset \mathbb{R}^d$ and δ
Output: Actions satisfying ρ-Interactive zCDP
Initialisation: Set $\ell = 1$, $t_1 = 1$ and $\mathcal{A}_1 = \mathcal{A}$
for $\ell = 1, 2, \ldots$ **do**
  $\beta_\ell \leftarrow 2^{-\ell}$
  Step 1: Find the G-optimal design π_ℓ for $\mathcal{A}_\ell$:

$$\max_{\substack{\pi \in \mathcal{P}(\mathcal{A}_\ell) \\ |\mathrm{Supp}(\pi)| \leq d(d+1)/2}} \log\det V(\pi). \tag{5.7}$$

  Step 2: $\mathcal{S}_\ell \leftarrow \mathrm{Supp}\,(\pi_\ell)$
  Choose each action $a \in \mathcal{S}_\ell$ for $T_\ell(a) \triangleq \lceil c_\ell \pi_\ell(a) \rceil$ times where c_ℓ is defined by Eq (5.6).
  Observe rewards $\{r_t\}_{t=t_\ell}^{t_\ell + \sum_a T_\ell(a)}$
  $T_\ell \leftarrow \sum_{a \in \mathcal{S}_\ell} T_\ell(a)$ and $t_{\ell+1} \leftarrow t_\ell + T_\ell + 1$
  Step 3: Estimate the parameter as

$$\hat{\theta}_\ell = V_\ell^{-1} \sum_{t=t_\ell}^{t_{\ell+1}-1} a_t r_t \quad \text{with} \quad V_\ell = \sum_{a \in \mathcal{S}_\ell} T_\ell(a) a a^\top$$

  Step 4: Make the parameter estimate private

$$\tilde{\theta}_\ell = \hat{\theta}_\ell + V_\ell^{-\frac{1}{2}} N_\ell,$$

  where $N_\ell \sim \mathcal{N}\left(0, \frac{2d}{\rho c_\ell} I_d\right)$.
  Step 5: Eliminate low rewarding arms:

$$\mathcal{A}_{\ell+1} = \left\{ a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \left\langle \tilde{\theta}_\ell, b - a \right\rangle \leq 2\beta_\ell \right\}.$$

end for

Now, we state some classic assumptions that bound the quantities of interest.

Assumption 5.10 (Boundedness). *We assume that:*
(1) actions are bounded: $\forall a \in \mathcal{A}$, $\|a\|_2 \leq 1$ in linear bandits, and $\forall t \in [1, T], \forall a \in \mathcal{A}_t$, $\|a\|_2 \leq 1$ in

contextual bandits
(2) *rewards are bounded:* $|r_t| \leq 1$, *and*
(3) *the unknown parameter is bounded:* $\|\theta^\star\|_2 \leq 1$.

Theorem 5.11 (Privacy of AdaC-GOPE)**.** *Under Assumption 5.10,* AdaC-GOPE *satisfies* ρ*-Interactive zCDP.*

Proof Sketch. AdaC-GOPE follows the blueprint of Section 5.2: the algorithm runs in independent episodes and each θ^ℓ is computed on non-overlapping sequence of rewards. The generic privacy proof is presented in Appendix C.1.2, and is instantiated for AdaC-GOPE in Appendix C.1.3.

Theorem 5.12 (Regret Analysis of AdaC-GOPE)**.** *Under Assumption 5.10 and for* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, the regret* R_T *of* AdaC-GOPE *is upper-bounded by*

$$A\sqrt{dT\log\left(\frac{K\log(T)}{\delta}\right)} + \frac{Bd}{\sqrt{\rho}}\sqrt{\log\left(\frac{K\log(T)}{\delta}\right)\log(T)},$$

where A *and* B *are universal constants. If* $\delta = \frac{1}{T}$*, then*

$$\mathbb{E}(R_T) \leq \mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}}(\log(KT))^{\frac{3}{2}}\right).$$

Proof Sketch. Under the "good event" that all the private parameters $\tilde{\theta}_\ell$ are well estimated, we show that the optimal action never gets eliminated. But the sub-optimal arms get eliminated as soon as the elimination threshold β_ℓ is smaller than their sub-optimality gaps. The regret upper bound follows directly. We refer to Appendix C.3.2 for complete proof.

We discuss the implications of our regret upper bound:

1. Achieving ρ-Interactive zCDP 'almost for free': Theorem 5.12 shows that the price of ρ-Interactive zCDP is the additive term $\tilde{\mathcal{O}}\left(\frac{d}{\sqrt{\rho}}\right)$[3]. For a fixed RDP budget ρ and as $T \to \infty$, the regret due to privacy becomes negligible in comparison with the privacy-oblivious term in regret, *i.e.* $\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$.

2. Optimality of AdaC-GOPE. In Theorem 4.21, we prove a $\Omega(\frac{d}{\sqrt{\rho}})$ minimax private regret lower bound that matches the regret upper bound of AdaC-GOPE up to an extra $(\log KT)^{\frac{3}{2}}$ factor. If K is exponential in d, then there is a mismatch between the regret upper and lower bounds, in their dependence on the dimension d. This gap could be improved with a better mechanism to make $\hat{\theta}$ private (Step 4 in Algorithm 2). In Appendix C.3.3, we discuss in detail how different ways of adding noise at Step 4 impact the dependence of the regret upper bound on d.

[3] $\tilde{\mathcal{O}}$ hides poly-logarithmic factors in the horizon T.

Related Algorithms and Bounds. Concurrently to our work, both [HGFD22] and [LZJ22] study private variants of the GOPE algorithm for pure ε-View DP and (ε, δ)-View DP, respectively. However, both algorithms differ in how they make private the estimated parameter $\hat{\theta}$ compared to AdaC-GOPE. Both [HGFD22] and [LZJ22] add noise to each sum of rewards $\sum_{t=t_\ell}^{t_{\ell+1}-1} r_t$ (Line 11, Alg. 8), whereas AdaC-GOPE add noise in $\hat{\theta}_l$ (Line 12, Alg. 8). As a result, though AdaC-GOPE achieves linear dependence on the dimension d as suggested by the lower bound, others do not (d^2 for [HGFD22] and $d^{3/2}$ for [LZJ22]).

In Appendix C.3.3, we analyse in detail the impact of adding noise at different steps of GOPE, both theoretically and experimentally.

5.3.3 Contextual linear bandits

Now, we consider an even more general setting of bandits, where the feasible arms at each step may vary and depend on some contextual information.

Contextual bandits generalise the finite-armed bandits by allowing the learner to use side information. At each step t, the policy observes a context $c_t \in \mathcal{C}$, which might be random or not. Having observed the context, the policy chooses an action $a_t \in [K]$ and observes a reward r_t. For the linear contextual bandits, the reward r_t depends on both the arm a_t and the context c_t in terms of a linear structural equation:

$$r_t = \langle \theta^\star, \psi(a_t, c_t) \rangle + \eta_t. \tag{5.8}$$

Here, $\psi : [K] \times \mathcal{C} \to \mathbb{R}^d$ is the feature map, $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is the noise, which we assume to be conditionally 1-subgaussian.

Under Equation (5.8), all that matters is the feature vector that results in choosing a given action rather than the identity of the action itself. This justifies studying a reduced model: in round t, the policy is served with the decision set $\mathcal{A}_t \subset \mathbb{R}^d$, from which it chooses an action $a_t \in \mathcal{A}_t$ and receives a reward

$$r_t = \langle \theta^\star, a_t \rangle + \eta_t,$$

where η_t is 1-subgaussian given $\mathcal{A}_1, a_1, R_1, \ldots, \mathcal{A}_{t-1}, a_{t-1}, R_{t-1}, \mathcal{A}_t$, and A_t.

Different choices of $\mathcal{A}_t$ lead to different settings. If $\mathcal{A}_t = \{\psi(c_t, a) : a \in [K]\}$, then we have a contextual linear bandit. On the other hand, if $\mathcal{A}_t = \{e_1, \ldots, e_d\}$, where $(e_i)_i$ are the unit vectors of $\mathbb{R}^d$ then the resulting bandit problem reduces to the stochastic finite-armed bandit.

The goal is to design a ρ-Interactive zCDP policy that minimises the regret, which is defined as

$$\hat{R}_T \triangleq \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle \theta^\star, a - a_t \rangle, \quad R_T \triangleq \mathbb{E}[\hat{R}_T].$$

Remark 5.13. *We suppose that c_t is **public** information, and thus $\mathcal{A}_t$ is public too. Rewards are the only private statistics to protect. The main difference compared to Section 5.3.2 is that the set of actions $\mathcal{A}_t$ is allowed to change at each time-step t. Thus, the action-elimination-based strategies, as used in Section 5.3.2, are not useful.*

We propose AdaC-OFUL, a ρ-Interactive zCDP extension of the Rarely Switching OFUL algorithm [AYPS11]. The OFUL algorithm applies the "optimism in the face of uncertainty principle" to the contextual linear bandit setting, which is to act in each round as if the environment is as nice as plausibly possible. The Rarely Switching OFUL Algorithm (RS-OFUL) can be seen as an "adaptively" phased version of the OFUL algorithm. RS-OFUL runs in episodes. At the beginning of each episode, the least square estimate and the confidence ellipsoid are updated. For the whole episode, the same estimate and confidence ellipsoid are used to choose the optimistic action. The condition to update the estimates (Line 6 of Algorithm 9) is to accumulate enough "useful information" in terms of the design matrix, which makes an update worth enough. RS-OFUL only updates the estimates $\log(T)$ times, while OFUL updates the estimates at each time step. RS-OFUL achieves similar regret as OFUL, up to a $\sqrt{1+C}$ multiplicative constant.

AdaC-OFUL (Algorithm 9) extends RS-OFUL by privately estimating the least-square estimate (Line 8 of Algorithm 9) while adapting the confidence ellipsoid accordingly. Specifically, we set $\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$, where $\beta_t = \mathcal{O}\left(\sqrt{d\log(t)}\right)$ and $\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho}d\log(t)}\right)$. Further details are in Appendix C.4.1.

Algorithm 9 AdaC-OFUL

1: **Input:** Privacy budget ρ, Horizon T, Regulariser λ, Dimension d, Doubling Schedule C
2: **Output:** A sequence of T-actions satisfying ρ-Interactive zCDP
3: **Initialisation:** $V_0 = \lambda I_d$, $\tilde{\theta} = 0_d$, $\tau = 0$, $\ell = 1$
4: **for** $t = 1, 2, \ldots$ **do**
5: Observe $\mathcal{A}_t$
6: **if** $\det(V_{t-1}) > (1+C)\det(V_\tau)$ **then**
7: Sample $Y_\ell \sim \mathcal{N}(0, \frac{2}{\rho} I_d)$
8: Compute $\tilde{\theta}_{t-1} = (V_{t-1})^{-1}(\sum_{s=1}^{t-1} a_s r_s + \sum_{m=1}^{\ell} Y_m)$
9: $\ell \leftarrow \ell + 1$ and $\tau \leftarrow t-1$
10: **end if**
11: Compute $a_t = \operatorname{argmax}_{a \in \mathcal{A}_t} \left\langle \tilde{\theta}_\tau, a \right\rangle + \tilde{\beta}_\tau \|a\|_{(V_\tau)^{-1}}$
12: Play arm a_t, Observe reward r_t
13: $V_t \leftarrow V_{t-1} + a_t a_t^T$
14: **end for**

Theorem 5.14 (Privacy of AdaC-OFUL)**.** *Under Assumption 5.10,* AdaC-OFUL *satisfies ρ-Interactive zCDP.*

Proof Sketch. AdaC-OFUL follows the blueprint of Section 5.2: the algorithm runs in independent adaptive episodes. Also, each $\tilde{\theta}_{t-1}$ can be retrieved by summing over quantities computed on non-overlapping sequences of rewards. The generic privacy proof is presented in Appendix C.1.2, and is instantiated for AdaC-OFUL in Appendix C.1.3.

To analyse the regret of AdaC-OFUL, we impose a stochastic assumption on the context generation. Specifically, we adopt the same assumption that is often used in on-policy [GLZ14, LRJ+22] and off-policy [ZDLB21, JLB22] linear contextual bandits.

Assumption 5.15 (Stochastic Contexts)**.** *At each step t, the context set $\mathcal{A}_t \triangleq \{a_1^t, \dots, a_{k_t}^t\}$ is generated i.i.d conditionally on k_t and the history $H_t \triangleq \{\mathcal{A}_1, a_1, X_1, \dots, \mathcal{A}_{t-1}, a_{t-1}, X_{t-1}, \mathcal{A}_t, a_t\})$ from a random process A such that 1. $\|A\|_2 = 1$*

2. *$\mathbb{E}[AA^T]$ is full rank, with minimum eigenvalue $\lambda_0 > 0$*
3. *$\forall z \in \mathbb{R}^d$, $\|z\|_2 = 1$, the random variable $(z^T A)^2$ is conditionally subgaussian, with variance*

$$\nu_t^2 \triangleq \mathbb{V}\left[(z^T A)^2 \mid k_t, H_t\right] \leq \frac{\lambda_0^2}{8\log(4k_t)}$$

This additional assumption helps control the minimum eigenvalue of the design matrix $V_t \triangleq \sum_{s=1}^t a_s a_s^T$.

Lemma 5.16 (Lemma 2, Equation (6) of [GLZ14])**.** *Let, at each round, $\mathcal{A}_t = \{a_1^t, \dots, a_{k_t}^t\}$ be generated i.i.d (conditioned on k_t and the history H_t) from a random process A such that*

- *$\|A\| = 1$*
- *$\mathbb{E}[AA^T]$ is full rank, with minimum eigenvalue $\lambda_0 > 0$*
- *$\forall z \in \mathbb{R}^d$, $\|z\| = 1$, the random variable $(z^T A)^2$ is conditionally subgaussian, with variance*

$$\nu_t^2 = \mathbb{V}\left[(z^T A)^2 \mid k_t, H_t\right] \leq \frac{\lambda_0^2}{8\log(4k_t)}$$

Then

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \lambda_{min}\left(\sum_{s=1}^t A_s A_s^T\right) \leq \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}\right) \leq \delta$$

Using Lemma 5.16 on the minimum eigenvalue, we quantify more precisely the effect of the added noise due to ρ-Interactive zCDP and derive tighter confidence bounds.

Theorem 5.17 (Regret Analysis of AdaC-OFUL)**.** *Under Assumptions 5.10 and 5.15, and for* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$*, the regret* R_T *of* AdaC-OFUL *is upper bounded by*

$$R_T \leq \mathcal{O}\left(d \log(T)\sqrt{T}\right) + \mathcal{O}\left(\frac{d^2}{\sqrt{\rho}} \log(T)^2\right)$$

Proof Sketch. The main challenge in the regret analysis is to design tight ellipsoid confidence sets around the private estimate $\tilde{\theta}_t$, since the regret can be shown to be the sum of the confidence widths. To design the non-private part of the ellipsoid confidence sets, we rely on the self-normalised bound for vector-valued martingales theorem of [AYPS11]. For the private part, we rely on the assumption of stochastic contexts controlling $\lambda_{\min}(G_t)$ and the concentration of χ^2 distribution to control the introduced Gaussian noise. The rest of the proof is adapted from the analysis of RS-OFUL [AYPS11]. We also show that the number of episodes, *i.e.* updates of the estimated parameters, is in $\mathcal{O}(\log(T))$. We refer to Appendix C.4.2 for the complete proof.

We discuss the implications of our regret upper bound:

1. Achieving ρ-Interactive zCDP 'almost for free': The upper bound of Theorem 5.17 shows that the price of ρ-Interactive zCDP for linear contextual bandits is the additive term $\tilde{\mathcal{O}}\left(\frac{d^2}{\sqrt{\rho}}\right)$. For a fixed budget ρ and as $T \to \infty$, the regret due to zCDP turns negligible in comparison with the privacy-oblivious regret term of $\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$.

2. Adapting AdaC-OFUL *for private contexts:* To make AdaC-OFUL achieve Joint-DP [SS18], the estimate $\tilde{\theta}$ at line 8 should be made private with respect to both rewards and context. A straightforward way to do so is by estimating the design matrix V_t privately, e.g. as it is done in [SS18]. A first regret analysis of this adaptation shows that the price of privacy in the regret will become not negligible, *i.e.* the regret is $O\left(\sqrt{T} + \sqrt{T/\rho}\right)$. This shows that the bottleneck in the problem is the private estimation of the design matrix.

3. Connecting Related Settings. [NR18] proposes LinPriv, which is an ε-global DP extension of OFUL. The context is assumed to be public but *adversely chosen*. Theorem 5 in [NR18] states that the regret of LinPriv is $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\varepsilon}Kd \log T\right)$. We revisit their regret analysis and show that the bound should be $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\varepsilon}Kd\sqrt{T}\right)$ instead. Refer to Appendix C.4.3 for details. Also, [SS18] proposes an (ε, δ)-Joint DP algorithm for *private and adversarial contexts*. The algorithm is based on OFUL and privately estimates $\hat{\theta}_t$ at each step using the tree-based mechanism [DNPR10b, CSS11]. However, this algorithm has an additional regret of $\frac{1}{\varepsilon}\sqrt{T}$ due to privacy.

Open Problem. It is still an open problem *whether it is possible* to design a private algorithm for linear contextual bandits with *private and/or adversarially chosen contexts*, such that the additional regret due to privacy in $\mathcal{O}(\log(T))$. We discuss this further in Chapter 9.

Table 5.2 – Regret bounds for bandits with ρ-Interactive zCDP.

Bandit Setting	Regret Upper Bound	Regret Lower Bound
Finite-armed bandits	$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(\frac{K}{\sqrt{\rho}}\sqrt{\log(T)}\right)$ (Thm 5.6)	$\Omega\left(\max\left(\sqrt{KT}, \sqrt{\frac{K}{\rho}}\right)\right)$ (Thm 4.20)
Linear bandits	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}}\log^{\frac{3}{2}}(KT)\right)$ (Thm 5.12)	$\Omega\left(\max\left(d\sqrt{T}, \frac{d}{\sqrt{\rho}}\right)\right)$ (Thm 4.21)
Linear Contextual bandits	$\mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\frac{d^2}{\sqrt{\rho}}\log(T)^2\right)$ (Thm 5.17)	

5.4 Private Algorithms for Best-Arm Identification

Following the ε-DP sample complexity lower bounds of Section 4.4, we aim to design an efficient ε-Interactive DP FC-BAI policy that simultaneously achieves the lower bound near-optimally and is computationally efficient. Due to the superior empirical performance and computational efficiency of Top Two algorithms (Algorithm 3), we design a DP variant of Top Two algorithms, which follows the generic blueprint of Section 5.2. First, we present AdaP-TT, which directly applies the generic recipe of Section 5.2 to the Top Two Algorithm. Then, we improve AdaP-TT by adapting the transportation costs to the private lower bounds. We call this algorithm AdaP-TT*. We show that AdaP-TT* achieves better sample complexity and matches the lower bound up to constants.

5.4.1 Adapting the generic wrapper for the Top Two algorithm

We instantiate the generic wrapper presented in Section 5.2 to the Top Two Algorithm (Algorithm 3). We refer to Section 2.2.9 for a general presentation of the Top Two Algorithm.

In (meta-) Algorithm 10, we present in-depth an adaptation of our private generic wrapper to the Top Two. Algorithm 10 is an FC-BAI strategy with

(a) **The Empirical Best (EB) recommendation rule:** Algorithm 10 recommends the arm with the highest (private) empirical mean, *i.e.* Line 14 in Algorithm 10.

(b) **The Generalised Likelihood Ratio (GLR) stopping rule:** Algorithm 10 decides to stop when the GLR stopping rule at Line 12 is met. This GLR stopping rule is decided using the transportation costs $(W_{a,b})_{(a,b)\in[K]^2}$ and thresholds $(c^{\varepsilon}_{a,b})_{(a,b)\in[K]^2}$ specified later.

(c) **The Top Two sampling rule:** Algorithm 10 is an instance of TTUCB [JDB+22], since it uses the following ingredients

1. A UCB leader, Line 16 in Algorithm 10
2. A Transportation Cost (TC) challenger, Line 17 in Algorithm 10
3. A β tracking procedure, Line 18 in Algorithm 10

Algorithm 10 Private Top Two Meta Algorithm.

1: **Input:** Privacy budget ε, risk $\delta \in (0,1)$, target allocation $\beta \in (0,1)$, transportation costs $W_{a,b} : \mathbb{R}^K \times \mathbb{N}^K \to \mathbb{R}^+$, thresholds $c^{\varepsilon}_{a,b} : \mathbb{N}^K \times (0,1) \to \mathbb{R}^+$,
2: **Output:** Recommendation $\hat{a}$, Stopping time τ and Sequence $(a_1, \dots, a_\tau)$
3: **Initialization:** $\forall a \in [K]$, pull arm a, set $k_a = 1$, $T_1(a) = K+1$, $L_{n,a} = 0$, $N_{n,a} = 1$, $n = K+1$.
4: **for** $n > K$ **do**
5: **if** there exists $a \in [K]$ such that $N_{n,a} \geq 2N_{T_{k_a}(a),a}$ **then** ▷ Per-arm doubling
6: Change phase $k_a \leftarrow k_a + 1$ for this arm a
7: Set $T_{k_a}(a) = n$ and $\tilde{N}_{k_a,a} = N_{T_{k_a}(a),a} - N_{T_{k_a-1}(a),a}$ ▷ Pulls of a in its last phase
8: Set $\hat{\mu}_{k_a,a} = \tilde{N}^{-1}_{k_a,a} \sum_{s=T_{k_a-1}(a)}^{T_{k_a}(a)-1} X_s \mathbb{1}(I_s = a)$ ▷ Empirical mean of a in its last phase
9: Set $\tilde{\mu}_{k_a,a} = \hat{\mu}_{k_a,a} + Y_{k_a,a}$ where $Y_{k_a,a} \sim \mathrm{Lap}((\varepsilon \tilde{N}_{k_a,a})^{-1})$ ▷ Make it private
10: **end if**
11: Set $\hat{a}_n = \arg\max_{b \in [K]} \tilde{\mu}_{k_b,b}$ ▷ Arm with highest private mean
12: **if** $W_{\hat{a}_n,b}\left((\tilde{\mu}_{k_a,a})_{a\in[K]}, \left(\tilde{N}_{k_a,a}\right)_{a\in[K]}\right) \geq c^{\varepsilon}_{\hat{a}_n,k_b}\left(\left(\tilde{N}_{k_a,a}\right)_{a\in[K]}, \delta\right)$ for all $b \neq \hat{a}_n$ **then**
13: Set $a_n = \top$
14: **return** $(\hat{a}_n, n)$ ▷ If GLR condition is met, recommend the private empirical best
15: **end if**
16: Set $b_n = \arg\max_{a\in[K]} \{\tilde{\mu}_{k_a,a} + \sqrt{k_a/\tilde{N}_{k_a,a}} + k_a/(\varepsilon \tilde{N}_{k_a,a})\}$ ▷ Private UCB leader
17: Set $c_n = \arg\min_{a \neq B_n} W_{b_n,a}\left((\tilde{\mu}_{k_a,a})_{a\in[K]}, (N_{k_a,a})_{a\in[K]}\right)$ ▷ Private TC challenger
18: Set $a_n = b_n$ if $N^{b_n}_{n,b_n} \leq \beta L_{n+1,b_n}$, else $a_n = c_n$ ▷ Tracking
19: Pull a_n and observe $r_n \sim \nu_{a_n}$
20: Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $N^{b_n}_{n+1,a_n} \leftarrow N^{b_n}_{n,a_n} + 1$ and $L_{n+1,b_n} \leftarrow L_{n,b_n} + 1$. Set $n \leftarrow n+1$
21: **end for**

Algorithm 10 incorporates the ingredients from the generic wrapper since

(a) The main private quantity is the empirical mean of rewards.

(b) An arm-dependent doubling combined with forgetting (Lines 5-10) is incorporated, so that the means $(\hat{\mu}_{k_a,a})_{a\in[K]}$ are computed on non-overlapping sequences of rewards. Thus, adding a Laplace noise of scale $1/(\varepsilon \tilde{N}_{k_a,a})$ at Line 9 of Algorithm 10 is enough to make the whole sequence of all computed $(\tilde{\mu}_{k_a,a})_{a\in[K]}$ satisfy ε-DP.

(c) The sampling rule, recommendation rule and stopping rule are all solely based on the private $(\tilde{\mu}_{k_a,a})_{a\in[K]}$

(d) The algorithm calibrates for the noise addition by adapting the thresholds $(c^{\varepsilon}_{a,b})_{(a,b)\in[K]^2}$ and the UCB bonus at Line 16. As we will show later, AdaP-TT* even adapts the transportation costs $(W_{a,b})_{(a,b)\in[K]^2}$ for privacy.

Following the generic wrapper and its generic privacy proof, Algorithm 10 is ε-Interactive DP. To finalise the algorithm design, Algorithm 10 needs the specification of

(a) Transportation costs $(W_{a,b})_{(a,b)\in[K]^2}$ used in the GLR stopping rule (Line 12 in Algorithm 10) and to choose the challenger (Line 17 in Algorithm 10). Each transportation cost $W_{a,b}$ is a function that takes as argument two vectors in $\mathbb{R}^K \times \mathbb{N}^K$, and returns a positive real number in $\mathbb{R}^+$.

(b) Thresholds $(c^{\varepsilon}_{a,b})_{(a,b)\in[K]^2}$ used for the GLR stopping rule (Line 12 in Algorithm 10). Each threshold $c^{\varepsilon}_{a,b}$ is a function that takes as argument a count vector in $\mathbb{N}^K$ and a risk parameter δ, and returns a positive real number in $\mathbb{R}^+$.

5.4.2 A plug-in approach: the AdaP-TT algorithm

AdaP-TT is an instance of Algorithm 10 which uses Gaussian transportation costs

$$W^G_{a,b}(\tilde{\mu}, \omega) = \frac{(\tilde{\mu}_a - \tilde{\mu}_b)^2_+}{2\sigma^2(1/\omega_a + 1/\omega_b)} \tag{5.9}$$

where $\tilde{\mu} \in \mathbb{R}^K$, $\omega \in \mathbb{N}^K$, $a, b \in [K]$ and G stands for Gaussian.

AdaP-TT uses Gaussian thresholds adapted for the private empirical mean estimators. For Gaussian distributions, the non-private thresholds are defined

$$c^G_{a,b}(\omega, \delta) = 2\mathcal{C}_G(\log\left((K-1)/\delta\right)/2) + 2\log(4 + \ln\omega_a) + 2\log(4 + \ln\omega_b)\,, \tag{5.10}$$

where the function $\mathcal{C}_G$ is defined in (C.30). It satisfies $\mathcal{C}_G(x) \approx x + \ln(x)$. For bounded distributions on $[0, 1]$ such as Bernoulli, we take $\sigma = 1/2$.

Using the concentration of Laplace noise, the private thresholds are then chosen to be

$$c^{G,\varepsilon}_{a,b}(\omega, \delta) = 2c^G_{a,b}(\omega, \delta(2\zeta(s)^2 k(\omega_a)^s k(\omega_b)^s)^{-1}) + \frac{1}{\varepsilon^2\sigma^2}\sum_{c\in\{a,b\}}\frac{1}{\omega_c}\left(\log\frac{2K\zeta(s)k(\omega_c)^s}{\delta}\right)^2 \tag{5.11}$$

where $s > 1$, ζ is the Riemann function and $k(x) = \log_2 x + 2$.

Lemma 5.18 (AdaP-TT is δ-correct)**.** *Given any sampling rule, the GLR stopping rule with $W^G_{a,b}$ as in (5.9) and the stopping threshold $c^{G,\varepsilon}_{a,b}$ as in (5.11) yields a δ-correct algorithm for σ-sub-Gaussian distributions.*

Proof. Proving δ-correctness of a GLR stopping rule is done by leveraging concentration results. Specifically, we start by decomposing the failure probability $\mathbb{P}_\mu\left(\tau_\delta < +\infty, \hat{a} \neq a^\star\right)$ into a non-private and a private part using the basic property of $\mathbb{P}(X+Y \geq a+b) \leq \mathbb{P}(X \geq a) + \mathbb{P}(Y \geq b)$. The two-factor in front of $c^G_{a,b}$ originates from the looseness of this decomposition, and we improve on it in Section 5.4.3. We conclude using concentration results from σ-sub-Gaussian and Laplace random variables. The proof is detailed in Appendix D of [AJMB24]. □

Remark 5.19 (Interpretable asymptotic shape for the thresholds of AdaP-TT)**.** *Asymptotically, our threshold is* $c_{a,b}^{G,\varepsilon}(\omega,\delta) \approx_{\delta\to 0} 2\log(1/\delta) + (1/\omega_a + 1/\omega_b)\log(1/\delta)^2/(\varepsilon^2\sigma^2)$.

Theorem 5.20 (Sample complexity of AdaP-TT)**.** *Let* $(\delta,\beta) \in (0,1)^2$ *and* $\varepsilon > 0$*. The* AdaP-TT *algorithm is* ε*-Interactive DP,* δ*-correct and satisfies that, for all* $\mu \in \mathbb{R}^K$ *such that* $\min_{a\neq b} |\mu_a - \mu_b| > 0$*,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \left(1 + \sqrt{1 + \frac{\Delta_{\max}^2}{2\sigma^4\varepsilon^2}}\right) \quad \textit{with} \quad \sigma = 1/2\,.$$

We adopt the asymptotic proof of the TTUCB algorithm, which is based on the unified analysis of the Top Two algorithms from [JDB+22]. We sketch high-level ideas for the proof and specify the effect of the generic wrapper on the expected sample complexity.

Proof. (1) The *non-private TTUCB algorithm* [JD24] achieves a sample complexity of $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})$ for sub-Gaussian random variables. The proof relies on showing that the empirical pulling counts are converging towards the β-optimal allocation $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})$. (2) The *effect of doubling and forgetting* is a multiplicative four-factor, *i.e.* $4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})$. The first multiplicative two-factor is due to forgetting since we throw away half of the samples. The second multiplicative two-factor is due to doubling since we have to wait for the end of an episode to evaluate the stopping condition. (3) The *Laplace noise* only affects the empirical estimate of the mean. Since the Laplace noise has no bias and a sub-exponential tail, the private means will still converge towards their true values. Therefore, the empirical counts will also converge to $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})$ asymptotically. (4) While the *Laplace noise has little effect on the sampling rule* itself, it *changes the dependency in* $\log(1/\delta)$ *of the threshold* used in the GLR stopping rule. The private threshold $c_{a,b}^{G,\varepsilon}$ has an extra factor $\mathcal{O}(\log^2(1/\delta))$ compared to the non-private one $c_{a,b}^{G}$. Using the convergence towards $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})$, the stopping condition is met as soon as $\frac{n}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})} \lesssim 2\log(1/\delta) + \frac{\Delta_{\max}^2}{2\sigma^4\varepsilon^2}\frac{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})}{n}\log^2(1/\delta)$. Solving the inequality for n concludes the proof while adding a multiplicative four-factor. The proof is detailed in Appendix E of [AJMB24]. □

Discussion. In the non-private regime where $\varepsilon \to +\infty$, our upper bound recovers the non-private lower bound for Gaussian distributions $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ up to a multiplicative factor 16, for $\beta = 1/2$ since $T^\star_{\mathrm{KL},1/2}(\boldsymbol{\nu}) \leq 2T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$. For Bernoulli distributions (or bounded distributions in $[0,1]$), there is still a mismatch between the upper and lower bounds due to the mismatch between the KL divergence of Bernoulli distributions and that of Gaussian (e.g. large ratio when the means are close to 0 or 1). This is, in essence, similar to the mismatch between UCB and KL-UCB in the regret-minimisation literature (e.g. Chapter 10 in [LS20]). To overcome this mismatch, it is necessary to adapt the transportation costs to the family of distributions considered. While the Top Two algorithms for Bernoulli distributions (or bounded distributions in $[0,1]$) have been studied in [JDB+22], the analysis is more involved. Therefore, it would obfuscate where and how privacy is impacting the expected sample complexity.

In the asymptotic high privacy regime i.e $\varepsilon \to 0$, our upper bound gives $\mathcal{O}(T^\star_{\mathrm{KL}}(\boldsymbol{\nu})\Delta_{\max}/\varepsilon)$ while the lower bound is $\Omega(T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\varepsilon)$. Therefore, our upper bound is only asymptotically tight for instances such that $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) = \mathcal{O}(T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\Delta_{\max})$, e.g. instances where the mean gaps have the same order of magnitude.

Specifically, for $\beta = 1/2$, it is well known that $T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}) \le 2T^\star_{\mathrm{KL}}(\boldsymbol{\mu}) \le 8\sum_{a\neq a^\star}\Delta_a^{-2}$. We consider Bernoulli instances ($0 < \Delta_{\min} \le \Delta_{\max} < 1$), where the gaps have the same order of magnitude, *i.e. Condition 1*: there exists a constant $C \ge 1$ such that $\Delta_{\max}/\Delta_{\min} \le C$. For such instances, there exists a universal constant c, such that

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \;\le\; c\,\max\left\{T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}), C\varepsilon^{-1}\sum\Delta_a^{-1}\right\}.$$

To show this, first, we upper bound

$$\begin{aligned}\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} &\le 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})\left(1+\sqrt{1+\frac{\Delta_{\max}^2}{2\varepsilon^2}}\right)\\ &\underset{(a)}{\le} 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})\left(2+\frac{\Delta_{\max}}{\sqrt{2}\varepsilon}\right)\end{aligned}$$

where $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\mu})$ is the β-characteristic time for Gaussian bandits, and (a) is due to the sub-additivity of the square root.

For $\beta = 1/2$, [Rus16] showed that $T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}) \le 2T^\star_{\mathrm{KL}}(\boldsymbol{\mu})$. On the other hand, [GK16] showed that $H(\boldsymbol{\mu}) \le T^\star_{\mathrm{KL}}(\boldsymbol{\mu}) \le 2H(\boldsymbol{\mu})$, where $H(\boldsymbol{\mu}) \triangleq \sum_{a\in[K]} 2\Delta_a^{-2}$ with $\Delta_{a^\star} = \Delta_{\min}$.

Plugging these two inequalities in the upper bound with $\beta = 1/2$ gives that

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \le 8T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}) + 16H(\boldsymbol{\mu})\frac{\Delta_{\max}}{\sqrt{2}\varepsilon}$$

Since we consider Bernoulli distributions, we know that $0 < \Delta_{\min} \le \Delta_{\max} < 1$. If we restrict ourselves to instances such that all the gaps have the same order of magnitude (Condition 1): there exists a constant $C \ge 1$ such that $\Delta_{\max} \le C\Delta_{\min}$.

For such instances, we obtain

$$\begin{aligned}\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} &\le 8T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}) + 16H(\boldsymbol{\mu})\frac{C\Delta_{\min}}{\sqrt{2}\varepsilon}\\ &\le 8T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}) + 16\sqrt{2}\frac{C}{\varepsilon}\left(\frac{1}{\Delta_{\min}}+\sum_{a=2}^{K}\frac{1}{\Delta_a}\right)\end{aligned}$$

where the last inequality is due to $H(\boldsymbol{\mu})\Delta_{\min} \le \frac{2}{\Delta_{\min}} + \sum_{a=2}^K \frac{2}{\Delta_a}$.

Finally using that $a+b \leq 2\max(a,b)$, we get that

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq c \max\left\{ T^\star_{\mathrm{KL},1/2}(\boldsymbol{\mu}), \frac{C}{\varepsilon}\left(\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K} \frac{1}{\Delta_a}\right)\right\}$$

for the universal constant $c = 45.26$.

For Bernoulli instances, Corollary 4.25 gives that the lower bound of the expected sample complexity of any δ-correct ε-global DP BAI strategy is

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \geq \max\left\{ T^\star_{\mathrm{KL}}(\boldsymbol{\nu}), \frac{1}{\varepsilon}\left(\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K} \frac{1}{\Delta_a}\right)\right\}.$$

where we use Proposition 4.26 to replace $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$ and $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ is the characteristic time for Bernoulli bandits.

However, in all the other cases where the gaps are not of the same order, the plug-in approach of AdaP-TT is sub-optimal due to a problem-dependent gap.

5.4.3 A lower bound based ppproach: the AdaP-TT* algorithm

To overcome the limitation of AdaP-TT, we adapt the transportation costs to reflect the private lower bound of Theorem 4.24, instead of "ignoring" the privacy constraint by using the non-private transportation costs $W^G_{a,b}$ as in (5.9) which are tailored for non-private FC-BAI.

Therefore, we propose the AdaP-TT* algorithm. AdaP-TT* is an instance of Algorithm 10 with the transportation costs

$$W^{G,\varepsilon}_{a,b}(\tilde{\mu},\omega) = \frac{(\tilde{\mu}_a - \tilde{\mu}_b)_+ \min\{3\varepsilon, (\tilde{\mu}_a - \tilde{\mu}_b)_+\}}{2\sigma^2(1/\omega_a + 1/\omega_b)} \quad \text{with} \quad \sigma = 1/2\,. \tag{5.12}$$

The transportation cost $W^{G,\varepsilon}_{a,b}$ is inspired by the relaxed private β-characteristic time

$$T^\star_{\varepsilon,\beta}(\boldsymbol{\nu})^{-1} \triangleq \max_{\omega\in\Sigma_K, \omega_{a^\star}=\beta} \min_{a\neq a^\star} \frac{(\mu_{a^\star} - \mu_a)\min\{3\varepsilon, \mu_{a^\star} - \mu_a\}}{2\sigma^2(1/\beta + 1/\omega_a)} \quad \text{with} \quad \sigma = 1/2\,. \tag{5.13}$$

AdaP-TT* uses the thresholds

$$\hat{c}^{G,\varepsilon}_{a,b}(\tilde{\mu},\omega,\delta) \triangleq \begin{cases} \frac{1}{2}c^{G,\varepsilon}_{a,b}(\omega, 2\delta/3) + \frac{\sqrt{2}}{\varepsilon\sigma}\sum_{c\in\{a,b\}}\sqrt{\frac{h(\omega_c,\delta)}{\omega_c}}\log\left(\frac{3K\zeta(s)k(\omega_c)^s}{\delta}\right), \text{if } (\tilde{\mu}_a - \tilde{\mu}_b)_+ < 3\varepsilon \\ \frac{3}{\sigma^2}\log\left(3K\zeta(s)\max_{c\in\{a,b\}} k(w_c)/\delta\right) + \frac{3\varepsilon}{\sqrt{2}\sigma}\sum_{c\in\{a,b\}}\sqrt{\omega_c h(\omega_c,\delta)} \quad \text{else} \end{cases} \tag{5.14}$$

where $s>1$, ζ is the Riemann function and $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for all $x \geq 1$, where W_{-1} is the negative branch of the Lambert W function. It satisfies $\overline{W}_{-1}(x) \approx x + \log x$. Finally, $c^{G,\varepsilon}_{a,b}$

is as in (5.11), $k(x) = \log_2 x + 2$ and

$$h(x,\delta) = \overline{W}_{-1}\left(2\log(3K\zeta(s)k(x)^s/\delta) + 4\log(4+\log x) + 1/2\right)/2\,.$$

Compared to thresholds of AdaP-TT, $\tilde{c}_{a,b}^{G,\varepsilon}$ depends also on the mean estimator $\tilde{\mu}$.

Lemma 5.21 (AdaP-TT* is δ-correct)**.** *Given any sampling rule, the GLR stopping rule with $W_{a,b}^{G,\varepsilon}$ as in (5.12) and the stopping threshold $\tilde{c}_{a,b}^{G,\varepsilon}$ as in (5.14) yields a δ-correct algorithm for σ-sub-Gaussian distributions.*

Proof. The proof is similar to the one of Lemma 5.18 with tighter manipulations allowing to divide $c_{a,b}^{G,\varepsilon}$ by 2. The proof is detailed in Appendix D of [AJMB24]. □

Remark 5.22 (Interpretable asymptotic shape for the thresholds of AdaP-TT*)**.** *Our threshold is*

$$\frac{3}{\sigma^2}\log(1/\delta) + \frac{3\varepsilon}{\sqrt{2}\sigma}(\sqrt{\omega_b}+\sqrt{\omega_a})\sqrt{\log(1/\delta)}$$

when $\tilde{\mu}_a - \tilde{\mu}_b \geq 3\varepsilon$, and

$$\log(1/\delta) + \frac{1}{2\varepsilon^2\sigma^2}(1/\omega_a + 1/\omega_b)\log(1/\delta)^2 + \frac{\sqrt{2}}{\varepsilon\sigma}(\sqrt{1/\omega_a}+\sqrt{1/\omega_b})\log(1/\delta)^{3/2}$$

otherwise.

Theorem 5.23 (Sample complexity of AdaP-TT*)**.** *Let $(\delta,\beta)\in(0,1)^2$ and $\varepsilon>0$. The* AdaP-TT* *algorithm is ε-Interactive DP, δ-correct and satisfies that, for all $\mu\in\mathbb{R}^K$ such that $\min_{a\neq b}|\mu_a-\mu_b|>0$,*

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq \begin{cases} 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})g_1\left(\Delta_{\max}/(\sigma^2\varepsilon)\right) & \text{if } \Delta_{\max}<3\varepsilon \\ 12T^\star_{\varepsilon,\beta}(\boldsymbol{\nu})g_2(3\varepsilon^2 T^\star_{\varepsilon,\beta}(\boldsymbol{\nu})\max\{\beta,1-\beta\}/2)/\sigma^2 & \text{otherwise}\end{cases},$$

where $T^\star_{\varepsilon,\beta}(\boldsymbol{\nu})$ as in (5.13) with $\sigma=1/2$. The function $g_1(y) = \sup\left\{x \mid x^2 < x + y\sqrt{2x} + \frac{y^2}{4}\right\}$ is increasing on $[0,12]$ and satisfies that $g_1(0)=1$ and $g_1(12)\leq 10$. The function $g_2(y) = 1 + 2(\sqrt{1+1/y}-1)^{-1}$ is increasing on $\mathbb{R}_+^\star$ and satisfies that $\lim_{y\to 0} g_2(y) = 1$.

Proof. The proof is similar to the one of Theorem 5.20 with tighter manipulations. The complete proof is detailed in Appendix E of [AJMB24]. □

Discussion. When $\Delta_{\max}<3\varepsilon$, our upper bound recovers the non-private lower bound for Gaussian distributions $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ up to a multiplicative factor $8g_1(4\Delta_{\max}/\varepsilon)\in[8,80]$, whose limit is 8 in non-private regime where $\varepsilon\to+\infty$. When $\Delta_{\min}\geq 3\varepsilon$, we have $12T^\star_{\varepsilon,\beta}(\boldsymbol{\nu}) \leq 8T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\varepsilon$. In the asymptotic highly privacy regime where $\varepsilon\to 0$, our upper bound matches the lower bound up to a multiplicative factor 48. Therefore, we close the gap left open by the algorithm

in Section 5.4.2. While the regime $\Delta_{\max} \geq 3\varepsilon > \Delta_{\min}$ is relevant for practical application, it is harder to understand how the different quantities interact in the upper/lower bounds in transitional phases. Thus, it is harder to claim optimality in those phases.

Comparison to DP-SE. DP-SE [SS19] is a DP version of the Successive Elimination algorithm introduced for the regret minimisation setting. The algorithm samples active arms uniformly during phases of geometrically increasing length. Based on the private confidence bounds, DP-SE eliminates provably sub-optimal arms at the end of each phase. Due to its phased-elimination structure, DP-SE can be easily converted into a DP FC-BAI algorithm, where we stop once there is *only one active arm left*. In particular, the proof of Theorem 4.3 of [SS19] shows that with high probability any sub-optimal arm $a \neq a^\star$ is sampled no more than $\mathcal{O}(\Delta_a^2 + (\varepsilon\Delta_a)^{-1})$. From this result, it is straightforward to extract a sample complexity upper bound for DP-SE, *i.e.* $\mathcal{O}(\sum_{a\neq a^\star} \Delta_a^{-2} + \sum_{a\neq a^\star}(\varepsilon\Delta_a)^{-1})$. This shows that DP-SE, too, achieves (ignoring constants) the high-privacy lower bound $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\varepsilon$ for Bernoulli instances. However, due to its uniform sampling within the phases, DP-SE is less adaptive than the Top Two sampling rule. Inside a phase, DP-SE continues to sample arms that might already be known to be bad, while Top Two algorithms adapt their sampling based on the transportation costs that reflect the amount of evidence collected in favour of the hypothesis that the leader is the best arm. Finally, our top two algorithms have the advantage of being anytime, *i.e.* their sampling strategy does not depend on the risk δ.

Another adaptation of DP-SE, namely DP-SEQ, is proposed in [KNSS21] for the problem of privately finding the arm with the highest quantile at a fixed level. Hence, it is different from BAI. For multiple agents, [RBCS23] studies privacy for BAI under fixed confidence. They propose and analyse the sample complexity of DP-MASE, a multi-agent version of DP-SE. They show that multi-agent collaboration leads to better sample complexity than independent agents, even under privacy constraints. While the multi-agent setting with federated learning allows tackling large-scale clinical trials taking place at several locations simultaneously, we study the single-agent setting, which is relevant for many small-scale clinical trials.

Remark 5.24 (On the number of rounds of adaptivity)**.** *Using our arm-dependent phases technique, it is possible to compute, at the end of the episode of arm a, the sequence of all the arms to be pulled before the end of the next episode (for another arm), without taking the collected observations into account. In contrast to the classical batched setting, where the batch size is fixed, the size of the resulting batches is adaptive and data-dependent. In the non-private setting ($\varepsilon = +\infty$), we recover Batched Best-Arm Identification (BBAI) in the fixed-confidence setting.* AdaP-TT *and* AdaP-TT* *are asymptotically optimal up to a multiplicative factor 4 with solely $\mathcal{O}\left(K \log_2(T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \log(1/\delta))\right)$ rounds of adaptivity. We refer the reader to Appendix F of [AJMB24] for more details on this remark, including a comparison to existing works.*

5.5 Experimental Analysis

In this section, we test experimentally the performance of the private bandit algorithms, which all are instantiations of the generic wrapper presented in Section 5.2.

5.5.1 Finite-armed bandits under Pure DP

We perform empirical evaluations to test two hypotheses:

(a) AdaP-KLUCB is the most optimal algorithm among the existing bandit algorithms with ε-pure DP

(b) The transition between high and low-privacy regimes is reflected in the empirical performance.

First, we compare the performances of AdaP-UCB and AdaP-KLUCB with those of DP-SE and DP-UCB. We set $\alpha = 3.1$ to comply with the regret upper bounds of AdaP-UCB and AdaP-KLUCB. We assign $\gamma = 0.1$ for DP-UCB and $\beta = 1/T$ for DP-SE. We implement all the algorithms in Python (version 3.8) and on an 8-core 64-bit Intel i5@1.6 GHz CPU. We test the algorithms for Bernoulli bandits with 5-arms and means $\{0.75, 0.625, 0.5, 0.375, 0.25\}$ (as in [SS19]). We run each algorithm 20 times for a horizon $T = 10^7$, and plot corresponding average and standard deviations of regrets in Figure 5.2. AdaP-KLUCB achieves the lowest regret followed by AdaP-UCB. Both of them achieve 10 times lower regret than the competing algorithms.

In Figure 5.3, we plot regret of AdaP-KLUCB at $T = 10^7$ for a Bernoulli bandit with mean rewards $\{0.8, 0.1, 0.1, 0.1, 0.1\}$. We plot the average regret over 20 runs as a function of the privacy budget $\varepsilon \in [0.05, 10]$. As indicated by the theoretical regret lower bounds and upper bounds, the experimental performance of AdaP-KLUCB demonstrates two regimes: a high-privacy regime (for $\varepsilon < 0.3$), where the regret of AdaP-KLUCB depends on the privacy budget ε, and a low privacy regime (for $\varepsilon > 0.3$), where the regret of AdaP-KLUCB does not depend on ε.

In brief, *our experimental results validate that* AdaP-KLUCB *is the most optimal algorithm for stochastic bandits that satisfies ε-global DP, and performance of* AdaP-KLUCB *transits from high- to low-privacy regimes, where its performance turns independent of the privacy budget ε.*

5.5.2 Regret bandits under ρ-Interactive zCDP

For finite-armed bandits, we test AdaC-UCB with $\beta = 1$ and compare it to its non-private counterpart, *i.e.* a UCB algorithm with adaptive episodes and forgetting. We test the algorithms for Bernoulli bandits with 5-arms and means $\{0.75, 0.625, 0.5, 0.375, 0.25\}$ (as in [SS19]).

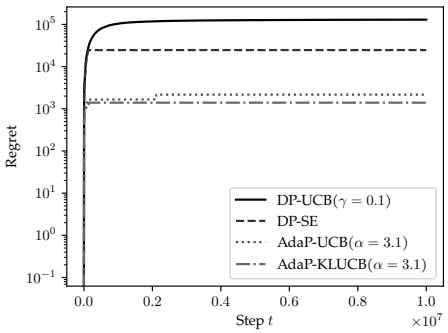


Figure 5.2 – Evolution of regret over time for DP-UCB, DP-SE, AdaP-UCB, and AdaP-KLUCB with $\varepsilon = 1$. Each algorithm is run 20 times with $T = 10^7$, and Bernoulli distributions with means $\{0.75, 0.625, 0.5, 0.375, 0.25\}$. AdaP-KLUCB achieves the lowest regret.

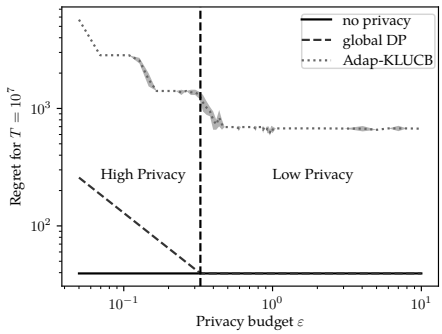


Figure 5.3 – Dependence of lower bounds and regret of AdaP-KLUCB with respect to the privacy budget ε. We run AdaP-KLUCB for 20 runs with $T = 10^7$. Echoing the theoretical analysis, the regret of AdaP-KLUCB transits between privacy regimes and is independent of ε for low-privacy.

For linear bandits with finitely many arms, we implement AdaC-GOPE and compare it to GOPE. We set the failure probability to $\delta = 0.001$ and the noise to be $\rho_t = \mathcal{N}(0, 1)$. We use the Frank-Wolfe algorithm to solve the G-optimal design problem [LS20]. We chose $K = 10$ actions randomly on the unit tri-dimensional sphere ($d = 3$). The true parameter $\theta^\star$ is also chosen randomly on the tri-dimensional sphere.

For linear contextual bandits, we implement AdaC-OFUL and compare it to RS-OFUL. We set $C = 1$, the regularisation constant $\lambda = 0.1$, the failure probability to $\delta = 0.001$ and the noise $\rho_t = \mathcal{N}(0, 1)$. We set $K = 10$ and $d = 3$. To generate the contexts, at each time step, we sample from a new set of actions $\mathcal{A}_t$ which is 10 dimensional multivariate Gaussian $\mathcal{N}\left(\left(\frac{1}{\sqrt{d}}, \ldots, \frac{1}{\sqrt{d}}\right), \frac{1}{10}\mathbf{I}_d\right)$. This way, we sample the contexts near the unit sphere, while having a sub-Gaussian generation process corresponding to the context-generation Assumption 5.15. The true parameter $\theta^\star$ is chosen randomly on the tri-dimensional sphere.

For the three settings, we run the private and non-private algorithms 100 times for a horizon $T = 10^7$, and compare their average regrets (Figure 6.2).

From the experimental results illustrated in Figure 6.2, we reach two conclusions for all three settings.

1. *Free-privacy in low-privacy regime.* For a fixed horizon T, the difference between the private and non-private regret, $Reg_{priv} - Reg_{non-priv}$, converges to zero as the privacy budget $\rho \to \infty$. Thus, our algorithms achieve the same regret as their non-private counterparts in the low-privacy regime.

2. *Asymptotic no price of privacy.* For a fixed privacy budget ρ, the Price of Privacy (PoP), i.e. PoP $\triangleq \frac{\mathrm{Reg}_{priv} - \mathrm{Reg}_{non-priv}}{\mathrm{Reg}_{non-priv}}$ converges to zero as the horizon T increases. This observation

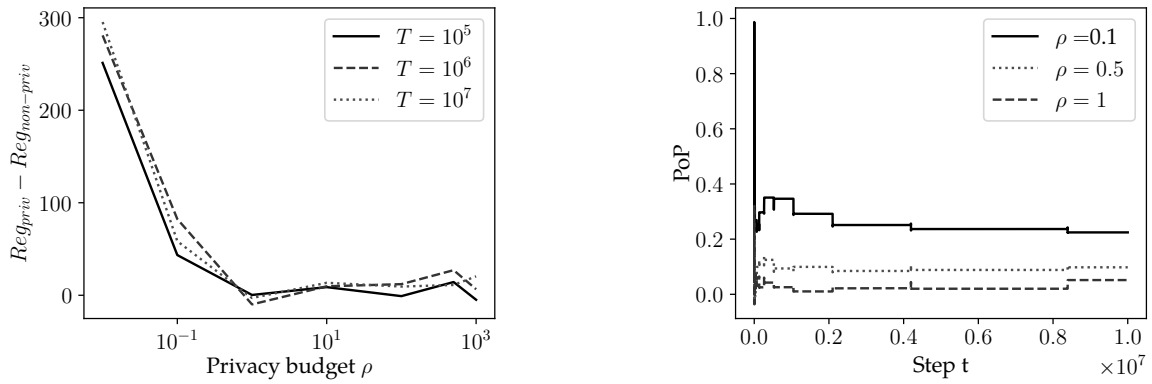


Figure 5.4 – Finite-armed Bandits under ρ-Interactive DP

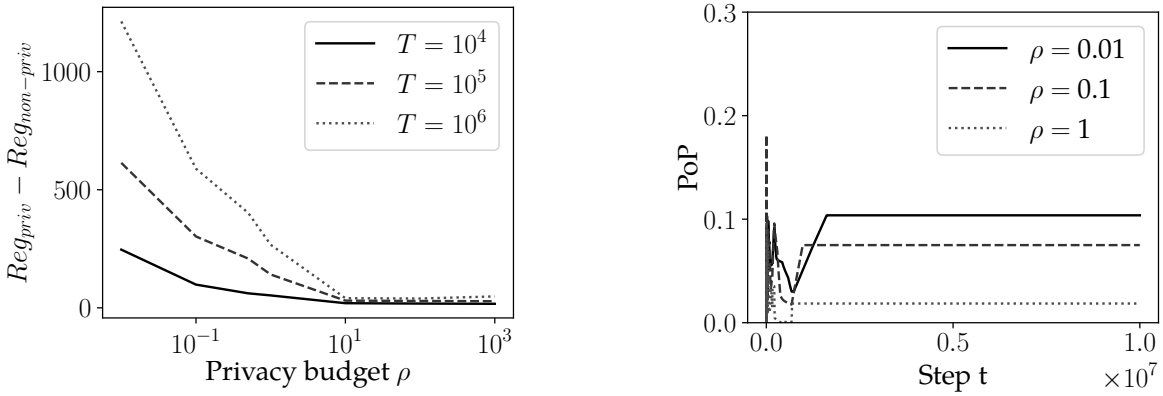


Figure 5.5 – Linear Bandits under ρ-Interactive DP

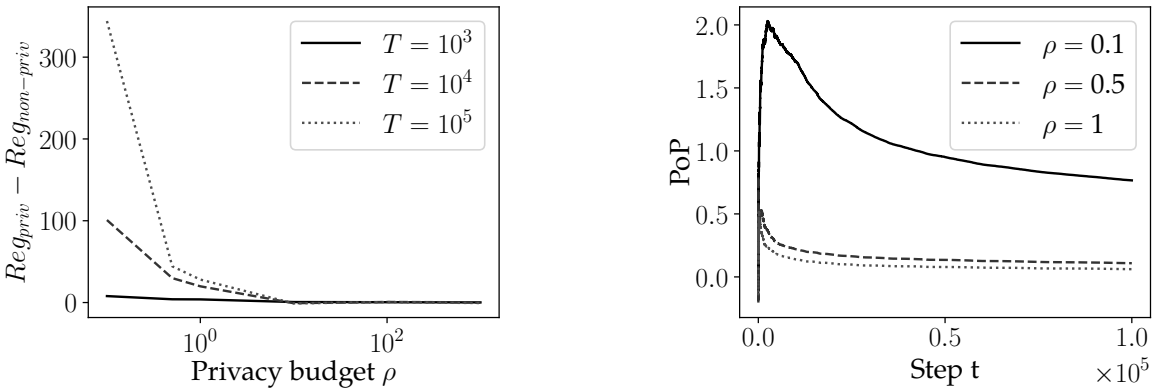


Figure 5.6 – Contextual Linear Bandits under ρ-Interactive DP

resonates with both the theoretical regret upper bounds of the algorithms and the hardness suggested by the lower bounds, where cost due to privacy appears as lower-order terms.

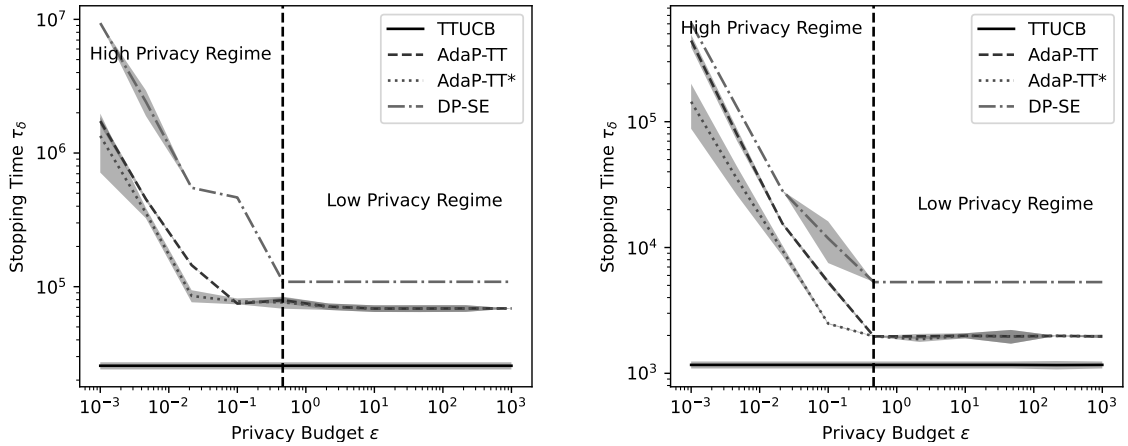


Figure 5.7 – Empirical stopping time τ_δ (mean $\pm$ std. over 1000 runs) with respect to the privacy budget ε for ε-global DP on Bernoulli instance μ_1 (left) and μ_2 (right). The shaded vertical line separates the two privacy regimes.

5.5.3 FC-BAI setting under Pure DP

We compare the performances of AdaP-TT, AdaP-TT* and DP-SE for FC-BAI in different Bernoulli instances as in [SS19]. The first instance has means $\mu_1 = (0.95, 0.9, 0.9, 0.9, 0.5)$ and the second instance has means $\mu_2 = (0.75, 0.7, 0.7, 0.7, 0.7)$. As a benchmark, we also compare to the non-private TTUCB. We set the risk $\delta = 10^{-2}$ and implement all the algorithms in Python (version 3.8). We run each algorithm 1000 times, and plot corresponding average and standard deviations of the empirical stopping times in Figure 5.7.

Figure 5.7 shows that: (a) AdaP-TT and AdaP-TT* require fewer samples than DP-SE to provide a δ-correct answer, for different values of ε and in all the instances tested. AdaP-TT and AdaP-TT* have the same performance in the low privacy regimes, while AdaP-TT* improves the sample complexity in the high privacy regime, as predicted theoretically. (b) The experimental performance of AdaP-TT and AdaP-TT* demonstrate two regimes. A high-privacy regime (for $\varepsilon < 0.1$ for μ_1 and $\varepsilon < 0.4$ for μ_2), where the stopping time depends on the privacy budget ε, and a low privacy regime (for $\varepsilon > 0.1$ for μ_1 and $\varepsilon > 0.4$ for μ_2), where the performance of AdaP-TT and AdaP-TT* does not depend on ε, and is four times the samples required by TTUCB in the worst case, as shown theoretically.

5.6 Conclusion

We propose a generic wrapper to design near-optimal private bandit algorithms. The main ingredient of this wrapper builds on the fact that the main private quantities of interest (*i.e.* the em-

pirical means of rewards $\hat{\mu}$ and the least-square estimate $\hat{\theta}$) are computed over non-overlapping sequences of inputs. This helps to add less noise in the bandit algorithm thanks to the Parallel Composition lemma (Lemma 2.10), thus achieving better utility. The noise addition step is then calibrated inside each algorithm differently: the exploration bonus is adapted for UCB and LinUCB, the exploration of each phase is augmented in elimination-based algorithms, and the thresholds are adapted for the FC-BAI algorithm. For different settings of interest, we instantiate the wrapper by expliciting the algorithm's details, inherit the wrapper's privacy guarantee, and then prove utility guarantees, *i.e.* regret and sample complexity upper bounds. For bandits under pure DP, we propose AdaP-UCB and AdaP-KLUCB, the first DP versions of UCB, which achieve the regret lower bounds. For ρ-Interactive zCDP, we propose AdaC-UCB, AdaC-GOPE and AdaC-OFUL, for finite-armed, linear and contextual bandits, and show that the additional cost in the regret due to ρ-Interactive zCDP is negligible in comparison to the regret incurred oblivious to privacy. Finally, for FC-BAI, we propose AdaP-TT and AdaP-TT*, DP versions of the Top Two algorithm. If AdaP-TT only matches the lower bound in instances with similar gaps, the AdaP-TT* algorithm overcomes this limitation by adapting the transportation costs. Our experimental analysis of the different algorithms validates our theoretical results.

Part II

Privacy Auditing and the Hardness of Membership Inference Games

Chapter 6

The Hardness of Target-dependent Membership Inference Games

We study *fixed-target* tracing attacks, where an attacker aims to infer whether a *fixed* target point was included or not in the input dataset of an algorithm. First, we define the *target-dependent leakage* of a point $z^\star$ as the advantage of the *optimal* adversary inferring its membership, and express it as a Total Variation distance. Then, we quantify both the target-dependent leakage and the trade-off functions for *the empirical mean* in terms of the Mahalanobis distance between the target point and the data-generating distribution. We further assess the impacts of two privacy defences, *i.e.* adding Gaussian noise and sub-sampling, and that of target misspecification by deriving their *target-dependent leakages* and trade-off functions. Our asymptotic analysis builds on a novel proof technique that combines an Edgeworth expansion of the Likelihood Ratio (LR) test and a Lindeberg-Feller central limit theorem. Our analysis yields that the LR attack for the empirical mean is a scalar product attack corrected by the inverse of the covariance matrix. This connects the LR and scalar product scores in the tracing attacks literature. Also, our Mahalanobis leakage score justifies the empirical success of orthogonality and other canary selection strategies used for privacy auditing. Finally, our experiments demonstrate the impacts of the leakage score, the sub-sampling ratio, and the noise scale on the target-dependent leakage, as indicated by the theory. Finally, our experiments validate that the Mahalanobis leakage score explains the hardness of fixed-target tracing attacks.

Contents

6.1 Introduction

In this section, we answer the main two questions:

Why are some points statistically harder to trace than others, and how can we quantify this hardness?
Can we quantify the target-dependent effect of privacy-preserving mechanisms?

Answering these two questions leads to the following contributions.

1. *Defining the target-dependent leakage.* We instantiate a *fixed-target MI game* (Algorithm 12, [YMM+22]). We define the leakage of a target point as the advantage of the optimal attacker, *i.e.* the LR attacker, trying to identify this fixed target point. We also characterise the target-dependent leakage in terms of a Total Variation distance (Equation (6.1)).

2. *Explaining the target-dependent leakage using the Mahalanobis distance.* We investigate the fixed-target MI game for the empirical mean. First, we find the asymptotic distributions of the LR scores if the target datum is included in the empirical mean and also if not. Then, we recover the optimal advantage (Equation (6.2)) and trade-off functions (Equation (6.3)). This shows that the target-dependent hardness of MI games depends on the Mahalanobis distance between the target point $z^\star$ and the true data-generating distribution (Table 6.1).

3. *A new covariance attack.* We analyse the LR score for the empirical mean asymptotically. Our novel proof technique that combines an Edgeworth expansion with Lindeberg-Feller central limit theorem shows that the *LR score is asymptotically a scalar product attack, corrected by the inverse of the covariance matrix* (Equation (6.4)). This enables us with a novel score for attacks, and improves on the scalar product by correcting it for the geometry of the data.

4. *Tight quantification of the effects of noise addition, sub-sampling, and misspecified targets on leakage.* We further study the impact of privacy-preserving mechanisms, such as the Gaussian mechanism [DR14b] and sub-sampling, on the target-dependent leakage. As shown in Table 6.1, both of them reduce the leakage scores and, thus, the powers of the optimal attacks. We numerically validate them. Finally, we quantify how target misspecification affects the leakage, and how it depends on the similarity between the real and misspecified targets.

6.2 Fixed-target Membership Game

First, we introduce the fixed-target Membership Inference (MI) game. Then, we discuss different performance metrics to assess the power of the adversary in the fixed-target game. Finally, we connect the fixed-target MI game (Algorithm 12) to the "average-target" MI game (Algorithm 5) presented in Section 2.3.

6.2.1 The target-dependent threat model

Let $\mathcal{M}$ be a randomised mechanism that takes as input a dataset D of n points belonging to $\mathcal{Z}$ and outputs $o \in \mathcal{O}$. In a Membership Inference (MI) game, an adversary attempts to infer whether a given target point $z^\star \in \mathcal{Z}$ was included in the input dataset of $\mathcal{M}$. Given access to an output $o \sim \mathcal{M}(D)$, the adversary tries to infer whether $z^\star \in D$ where D is the input dataset that generated the output o.

A fixed-target MI game (Algorithm 12) is a game between two entities: the fixed-target Crafter (Algorithm 11) and the adversary $\mathcal{A}_{z^\star}$. The MI game runs in multiple rounds. At each round t, the crafter samples a pair (o_t, b_t), where o_t is an output of the mechanism and b_t is the secret binary membership of $z^\star$. The adversary $\mathcal{A}_{z^\star}$ takes as input only o_t and outputs $\hat{b}_t$ trying to reconstruct b_t.

The specificity of the *fixed-target* MI game is that the target $z^\star$ is fixed throughout the game. Thus, the performance metrics of the attacker, *i.e.* the advantage and trade-off functions, are target-dependent. In contrast, in the MI game of Algorithm 5, the target $z^\star$ is sampled randomly at each step of the game, either sampled from the data-generating distribution (Line 6 in Algorithm 4), or uniformly sampled from the input dataset (Line 8 in Algorithm 4). Thus, in the MI game of Algorithm 5, the performance metrics of the attacker are averaged over the sampling of the target points. This averaging obfuscates the dependence of the leakage on each target point. To study the effect of target points on the hardness of MI games, we use this fixed-target formulation of MI games, which has also been proposed in Definition 3.3 of [YMM+22].

A fixed-target MI game can also be seen as a hypothesis test. Here, the adversary tries to test the hypothesis "H_0: *The output o observed was generated from a dataset sampled i.i.d. from $\mathcal{D}$*", *i.e.* $b = 0$, versus "H_1: *The target point $z^\star$ was included in the input dataset producing the output o*", *i.e.* $b = 1$. We denote by $p_{\text{out}}(o \mid z^\star)$ and $p_{\text{in}}(o \mid z^\star)$ the distributions of the output o under H_0 and H_1 respectively.

6.2.2 Performance metrics for the fixed-target adversary

An adversary $\mathcal{A}_{z^\star}$ is a possibly randomised function that takes as input o the output of the mechanism $\mathcal{M}$, and generates a guess $\hat{b} \sim \mathcal{A}_{z^\star}(o)$ trying to infer b. The performance of $\mathcal{A}_{z^\star}$ can be assessed either with aggregated metrics like the accuracy and the advantage, or with test-based metrics like Type-I/Type-II errors, and trade-off functions.

The *accuracy* of $\mathcal{A}_{z^\star}$ is defined as

$$\text{Acc}_n(\mathcal{A}_{z^\star}) \triangleq \Pr[\mathcal{A}_{z^\star}(o) = b],$$

Algorithm 11 The Fixed-target Crafter

Input: Mechanism $\mathcal{M}$, Data distribution $\mathcal{D}$, Number of samples n, Target $z^\star$
Output: (o, b), where $o \in \mathcal{O}$ and $b \in \{0, 1\}$
Build a dataset $D \sim \bigotimes_{i=1}^{n} \mathcal{D}$
Sample $b \sim \text{Bernoulli}\left(\frac{1}{2}\right)$
if $b = 1$ **then**
    Sample $i \sim \mathcal{U}[n]$
    $D \leftarrow \text{Replace}(D, i, z^\star)$ ▷ Put $z^\star$ at position i in D
end if
Let $o \sim \mathcal{M}(D)$
Return (o, b)

where the probability is over the generation of (o, b) using Algorithm 11 with input $(\mathcal{M}, \mathcal{D}, n, z^\star)$, and any randomness in the adversary.

The *advantage* of an adversary is the centred accuracy

$$\text{Adv}_n(\mathcal{A}_{z^\star}) \triangleq 2\text{Acc}_n(\mathcal{A}_{z^\star}) - 1.$$

We can also define two errors from the hypothesis testing formulation. *The Type-I error*, aka False Positive Rate, is

$$\alpha_n(\mathcal{A}_{z^\star}) \triangleq \Pr\left[\mathcal{A}_{z^\star}(o) = 1 \mid b = 0\right].$$

The Type-II error, aka the False Negative Rate, is

$$\beta_n(\mathcal{A}_{z^\star}) \triangleq \Pr\left[\mathcal{A}_{z^\star}(o) = 0 \mid b = 1\right].$$

The *power* of the test is $1 - \beta_n(\mathcal{A}_{z^\star})$.

In MI games, an adversary can threshold over a score function s to conduct the MI games, *i.e.* for $\mathcal{A}_{s,\tau,z^\star}(o) \triangleq \mathbb{1}\left(s(o; z^\star) > \tau\right)$ where s is a score function and τ is a threshold. We want to design score functions that maximise the power under a fixed significance level α, *i.e.*

$$\text{Pow}_n(s, \alpha, z^\star) \triangleq \max_{\tau \in T_\alpha} 1 - \beta_n(\mathcal{A}_{s,\tau,z^\star})$$

where $T_\alpha \triangleq \{\tau \in \mathbb{R} : \alpha_n(\mathcal{A}_{s,\tau,z^\star}) \leq \alpha\}$. $\text{Pow}_n(s, \alpha, z^\star)$ is also called a *trade-off function*.

6.2.3 Connection between the fixed-target and average-target MI games

An adversary $\mathcal{A}$ in an average-target MI game (Algorithm 5) can be regarded as an infinite collection of target-dependent adversaries $(\mathcal{A}_{z^\star})_{z^\star}$, where $\mathcal{A}(z^\star, o) = \mathcal{A}_{z^\star}(o)$.

Algorithm 12 Fixed-target MI Game

Input: Mechanism $\mathcal{M}$, Data distribution $\mathcal{D}$, Number of samples n, Target $z^\star$, Adversary $\mathcal{A}_{z^\star}$, Rounds T
Output: A list $L \in \{0,1\}^T$, where $L_t = 1$ if the adversary succeeds at step t.
Initialise a empty list L of length T
for t = 1, . . . , T **do**
    Sample $(o_t, b_t) \sim$ Fixed-target Crafter (Algorithm 11), with inputs ($\mathcal{M}$, $\mathcal{D}$, n, $z^\star$)
    Sample $\hat{b}_t \sim \mathcal{A}_{z^\star}(o_t)$
    Set $L_t \leftarrow \mathbb{1}\left(b_t = \hat{b}_t\right)$
end for
Return L

The advantage (and accuracy) of $\mathcal{A}$ is the expected advantage (and accuracy) of $\mathcal{A}_{z^\star}$, when $z^\star \sim \mathcal{D}$, *i.e.*

$$\mathrm{Adv}_n(\mathcal{A}) = \mathbb{E}_{z^\star \sim \mathcal{D}}\left[\mathrm{Adv}_n(\mathcal{A}_{z^\star})\right].$$

Studying the performance metrics of an adversary under average-target MI game hides the dependence on the target, by averaging out the performance on different target points. Consequently, using the average-target MI games to audit privacy can hurt performance. A gain could directly be observed by running the same attack on a fixed "easy to attack" fixed-target MI game.

Also, we observe that the optimal LR test for the average-target MI game is the same LR test for the fixed-target MI game. Specifically,

$$\begin{aligned}
\ell_n(o; z^\star) &\triangleq \log\left(\frac{p_n^{\mathrm{in}}(z^\star, o)}{p_n^{\mathrm{out}}(z^\star, o)}\right) \\
&= \log\left(\frac{p_n^{\mathrm{in}}(o \mid z^\star) p_n^{\mathrm{in}}(z^\star)}{p_n^{\mathrm{out}}(o \mid z^\star) p_n^{\mathrm{out}}(z^\star)}\right) \\
&= \log\left(\frac{p_n^{\mathrm{in}}(o \mid z^\star)}{p_n^{\mathrm{out}}(o \mid z^\star)}\right)
\end{aligned}$$

since $p_n^{\mathrm{in}}(z^\star) = p_n^{\mathrm{out}}(z^\star) = \mathcal{D}(z^\star)$.

Thus, the same LR attack optimally solves both fixed-target and average-target MI games. The only difference is in the resulting performance metric, *i.e.* whether we average out the effect of $z^\star \sim \mathcal{D}$ in average-target MI games or we keep the dependence on the target $z^\star$, by fixing $z^\star$ in the fixed-target MI games.

6.3 Optimal Adversary and Definition of Membership Leakage

It is a fundamental result of statistics that given two data generating distributions p_0 and p_1 under hypotheses H_0 and H_1 respectively, no test can achieve better power than the Likelihood Ratio (LR) test [NP33].

Now, we recall the hypothesis testing formulation of the fixed-target MI games, where $p_n^{\text{out}}(o \mid z^\star)$ is the distribution of the output o under H_0 and $p_n^{\text{in}}(o \mid z^\star)$ is the distribution of the output o under H_1. Then, the *log-Likelihood Ratio* (LR) test (or score) for fixed-target MI game is

$$\ell_n(o; z^\star) \triangleq \log\left(\frac{p_n^{\text{in}}(o \mid z^\star)}{p_n^{\text{out}}(o \mid z^\star)}\right).$$

The LR attacker uses a threshold τ on the log-likelihood score, *i.e.* $\mathcal{A}_{\ell,\tau,z^\star}(o) \triangleq \mathbb{1}\left(\ell_n(o; z^\star) > \tau\right)$. We denote by $\mathcal{A}_{\text{Bayes},z^\star} \triangleq \mathcal{A}_{\ell,0,z^\star}$ the LR attacker with threshold $\tau = 0$. We provide Theorem 6.1 to characterise optimal adversaries under both aggregated and test-based metrics.

Theorem 6.1 (Characterising Optimal Adversaries)**.**

(*a*) *For every $\alpha \in [0, 1]$, the log-likelihood test ℓ_n is the test that maximises the power under significance α,* i.e. *for any α and any test function s,*

$$\text{Pow}_n(\ell_n, \alpha, z^\star) \geq \text{Pow}_n(s, \alpha, z^\star).$$

(*b*) *$\mathcal{A}_{\text{Bayes},z^\star}$ is the adversary that maximises the advantage (and accuracy),* i.e. *for any adversary $\mathcal{A}_{z^\star}$, we have that*

$$\text{Adv}_n(\mathcal{A}_{\text{Bayes},z^\star}) \geq \text{Adv}_n(\mathcal{A}_{z^\star}).$$

(*c*) *Let* TV *denote the total variation distance. The advantage of the optimal Bayes adversary is*

$$\text{Adv}_n(\mathcal{A}_{\text{Bayes},z^\star}) = \text{TV}\left(p_n^{\text{out}}(. \mid z^\star) \,\middle\|\, p_n^{\text{in}}(. \mid z^\star)\right). \tag{6.1}$$

Proof. First, (a) is a direct consequence of the Neyman-Pearson lemma. To prove (b), we observe that the log-likelihood adversary with threshold $\tau = 0$ is exactly the Bayes optimal classifier. Specifically, since $\Pr(b = 0) = \Pr(b = 1) = 1/2$, we can rewrite the log-likelihood as the

$$\ell_n(o; z^\star) = \log\left(\frac{\Pr(b = 1 \mid o, z^\star)}{\Pr(b = 0 \mid o, z^\star)}\right).$$

Thus, thresholding with 0 gives the Bayes optimal classifier exactly, which has the highest accuracy among all classifiers.

For (c), we observe that

$$\text{Adv}_n(\mathcal{A}_{\text{Bayes},z^\star}) = \Pr(\ell_n(o; z^\star) \leq 0 \mid b = 0) - \Pr(\ell_n(o; z^\star) \leq 0 \mid b = 1)$$

$$= p_n^{\text{out}}(O \mid z^\star) - p_n^{\text{in}}(O \mid z^\star)$$

where $O \triangleq \{o \in \mathcal{O} : p_n^{\text{out}}(o \mid z^\star) \geq p_n^{\text{in}}(o \mid z^\star)\}$.

The last equation is exactly the definition of the TV $(p_n^{\text{out}}(. \mid z^\star) \,\|\, p_n^{\text{in}}(. \mid z^\star))$.

□

As a consequence of Theorem 6.1, we *define* the target-dependent leakage of $z^\star$.

Definition 6.2 (Targent-dependent leakage)**.** *The target-dependent leakage of $z^\star$, for mechanism $\mathcal{M}$ and data-generating distribution $\mathcal{D}$, is the advantage of the optimal Bayes attacker on $z^\star$,* i.e.

$$\xi_n(z^\star, \mathcal{M}, \mathcal{D}) \triangleq \text{Adv}_n(\mathcal{A}_{\text{Bayes},z^\star}) = \text{TV}\left(p_n^{\text{out}}(. \mid z^\star) \,\middle\|\, p_n^{\text{in}}(. \mid z^\star)\right).$$

Our main goal is to *quantify the target-dependent leakage $\xi_n(z^\star, \mathcal{M}, \mathcal{D})$ and trade-off functions for different mechanisms*, namely the empirical mean and its variations. These two quantities may be intractable to characterise for any general data-generating distribution. To overcome this limitation, we use the asymptotic properties of the empirical mean as the main tool.

6.4 Target-dependent Leakage of the Empirical Mean

We instantiate the fixed-target MI game with the empirical mean mechanism. First, we characterise the asymptotic distribution of the LR scores under H_0 and H_1. Then, we quantify the target-dependent leakage of a target $z^\star$, and show that it depends on the Mahalanobis distance $z^\star$ and the data generating distribution $\mathcal{D}$. Finally, we connect our results to tracing attacks [SOJH09, DSS+15], and propose a new canary selection strategy and white-box attack on gradient descents.

Notations and the asymptotic regime. We denote by $\mathcal{M}_n^{\text{emp}}$ the empirical mean mechanism. $\mathcal{M}_n^{\text{emp}}$ takes as input a dataset of size n of d-dimensional points, *i.e.* $D = \{Z_1, \ldots, Z_n\} \in (\mathbb{R}^d)^n$, and outputs the exact empirical mean $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n Z_i \in \mathbb{R}^d$. We denote by $\rightsquigarrow$ convergence in distribution. Let Φ represent the Cumulative Distribution Function (CDF) of the standard normal distribution, *i.e.* $\Phi(\alpha) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha} e^{-t^2/2}\,\mathrm{d}t$ for $\alpha \in \mathbb{R}$. For a matrix M and a vector x, we write $\|x\|_M^2 \triangleq x^T M x$. Since the LR test can be non-tractable in general cases, we study the *asymptotic behaviour of the LR test*, when both the sample size n and the dimension d tend to infinity such that $d/n = \tau > 0$.

Assumptions on the data generating distribution. We suppose that the data-generating distribution is column-wise independent, *i.e.* $\mathcal{D} \triangleq \bigotimes_{j=1}^d \mathcal{D}_j$ and has a finite $(4+\delta)$-th moment for some small δ, *i.e.* there exists $\delta > 0$, such that $\mathbb{E}[Z^{4+\delta}] < \infty$. We denote by $\mu \triangleq (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$

the mean of $\mathcal{D}$, and by $C_\sigma \triangleq \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2) \in \mathbb{R}^{d\times d}$ the covariance matrix. We recall that the Mahalanobis distance [Mah36] of $z^\star$ with respect to $\mathcal{D}$ is $\|z^\star - \mu\|_{C_\sigma^{-1}}$.

Asymptotic distribution of the LR score. For the empirical mean and a column-wise independent distribution $\mathcal{D}$, the LR score is

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) = \sum_{j=1}^{d} \log\left(\frac{p_{n,j}^{\mathrm{in}}(\hat{\mu}_{n,j}; z_j^\star, \mu_j, \sigma_j)}{p_{n,j}^{\mathrm{out}}(\hat{\mu}_{n,j}; z_j^\star, \mu_j, \sigma_j)}\right),$$

where $\hat{\mu}_n = (\hat{\mu}_{n,j})_{j=1}^d$, $p_{n,j}^{\mathrm{out}}$ is the distribution of $\hat{\mu}_{n,j}$ for $b = 0$, and $p_{n,j}^{\mathrm{in}}$ is the distribution of $\hat{\mu}_{n,j}$ for $b = 1$. In Theorem 6.3, we characterise the asymptotic distribution of the LR test, under H_0 and H_1 in the target-dependent MI game.

Theorem 6.3 (Asymptotic distribution of the LR score)**.** *Using an Edgeworth asymptotic expansion of the likelihood ratio score and a Lindeberg-Feller central limit theorem, we show that*

(*a*) *Under* H_0,

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{1}{2}m^\star, m^\star\right)$$

(*b*) *Under* H_1,

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{1}{2}m^\star, m^\star\right)$$

The convergence is a convergence in distribution, such that $d, n \to \infty$*, while* $d/n = \tau$*. We call*

$$m^\star \triangleq \lim_{n,d} \frac{1}{n}\|z^\star - \mu\|_{C_\sigma^{-1}}^2 = \lim_{n,d} \sum_{j=1}^{d} \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2}$$

the leakage score *of target* $z^\star$*.*

Proof Sketch. The proof has three main steps. First, we rewrite the LR score with respect to $d_{n,j}$ the density of the centred normalised mean $\sqrt{n}\left(\frac{\hat{\mu}_{n,j}-\mu_j}{\sigma}\right)$. Then, we use the Edgeworth asymptotic expansion (Theorem 2.41) of $d_{n,j}$ to get an expansion of the LR score. Finally, we conclude the asymptotic distribution of the LR test using the Lindeberg-Feller theorem (Theorem 2.40). The detailed proof of Theorem 6.3 is presented in Appendix D.1.

Using testing results between Gaussians, we retrieve the leakage and trade-off functions.

Corollary 6.4 (Target-dependent leakage of the empirical mean)**.** *The asymptotic target-dependent leakage of* $z^\star$ *in the empirical mean is*

$$\lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^{\mathrm{emp}}, \mathcal{D}) = \Phi\left(\frac{\sqrt{m^\star}}{2}\right) - \Phi\left(-\frac{\sqrt{m^\star}}{2}\right). \tag{6.2}$$

The asymptotic trade-off function, achievable with threshold $\tau_\alpha = -\frac{m^\star}{2} + \sqrt{m^\star}\Phi^{-1}(1-\alpha)$*, is*

$$\lim_{n,d} \mathrm{Pow}_n(\ell_n, \alpha, z^\star) = \Phi\left(\Phi^{-1}(\alpha) + \sqrt{m^\star}\right). \tag{6.3}$$

Proof. From the asymptotic distribution of the LR score, we get directly that

$$\begin{aligned}
\lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^{\mathrm{emp}}, \mathcal{D}) &= \Pr\left(\mathcal{N}\left(-\frac{m^\star}{2}, m^\star\right) < 0\right) - \Pr\left(\mathcal{N}\left(\frac{m^\star}{2}, m^\star\right) < 0\right) \\
&= \Phi\left(\frac{m^\star/2}{\sqrt{m^\star}}\right) - \Phi\left(-\frac{m^\star/2}{\sqrt{m^\star}}\right) \\
&= \Phi\left(\frac{\sqrt{m^\star}}{2}\right) - \Phi\left(-\frac{\sqrt{m^\star}}{2}\right)
\end{aligned}$$

The threshold τ_α for which the asymptotic LR attack achieves significance α verifies:

$$\Pr\left(\mathcal{N}\left(-\frac{m^\star}{2}, m^\star\right) \geq \tau_\alpha)\right) = \alpha$$

Thus $\tau_\alpha = -\frac{m^\star}{2} + \sqrt{m^\star}\Phi^{-1}(1-\alpha)$.

Finally, we find the power of the test by

$$\begin{aligned}
\lim_{n,d} \mathrm{Pow}_n(\ell_n, \alpha, z^\star) &= \Pr\left(\mathcal{N}\left(\frac{m^\star}{2}, m^\star\right) \geq \tau_\alpha\right) \\
&= \Pr\left(\frac{m^\star}{2} + \sqrt{m^\star}\mathcal{N}(0,1) \geq -\frac{m^\star}{2} + \sqrt{m^\star}\Phi^{-1}(1-\alpha)\right) \\
&= \Pr\left(\sqrt{m^\star}\mathcal{N}(0,1) \geq -m^\star - \sqrt{m^\star}\Phi^{-1}(\alpha)\right) \\
&= \Pr\left(\mathcal{N}(0,1) \leq \sqrt{m^\star} + \Phi^{-1}(\alpha)\right) \\
&= \Phi\left(\Phi^{-1}(\alpha) + \sqrt{m^\star}\right)
\end{aligned}$$

□

Empirical LR attack. Following the proof of Theorem 6.3, we show in Remark D.1 that

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \sim (z^\star - \mu)^T C_\sigma^{-1}(\hat{\mu}_n - \mu) - \frac{1}{2n}\|z^\star - \mu\|^2_{C_\sigma^{-1}} \tag{6.4}$$

asymptotically in n and d. Equation (6.4) shows that the LR score is a scalar product between $z^\star - \mu$ and $\hat{\mu}_n - \mu$, corrected by the precision matrix C_σ^{-1}. The optimal LR score is computed using the true mean μ and covariance matrix C_σ. A straightforward way to convert it into a realistic attack is by replacing the true mean μ and covariance C_σ in Equation (6.4) with empirical estimates. We can use a set of reference points $D_{n_0}^{\mathrm{ref}} \triangleq \{Z_1^{\mathrm{ref}}, \ldots, Z_{n_0}^{\mathrm{ref}}\}$ sampled independently from $Z_1, \ldots, Z_n, z^\star$ to get $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} Z_i^{\mathrm{ref}}$ and $\hat{C}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} Z_i^{\mathrm{ref}}(Z_i^{\mathrm{ref}})^T$. This

leads to an empirical LR score

$$\ell_n^{\text{emp}}(\hat{\mu}_n; z^\star, D_{n_0}^{\text{ref}}) = (z^\star - \hat{\mu}_0)^T \hat{C}_0^{-1}(\hat{\mu}_n - \hat{\mu}_0) - \frac{1}{2n}\|z^\star - \hat{\mu}_0\|^2_{\hat{C}_0^{-1}} \,. \tag{6.5}$$

Since the attack estimates the mean and covariance, the empirical LR attack is no longer optimal. The drop in the empirical LR attack's power depends on the estimation's accuracy, and thus on n_0 the number of reference points.

Connection to [SOJH09]. For Bernoulli distributions, [SOJH09] shows that the LR test in the "average-target" MI game is asymptotically distributed as $\mathcal{N}\left(-\frac{1}{2}\tau, \tau\right)$ under H_0 and $\mathcal{N}\left(\frac{1}{2}\tau, \tau\right)$ under H_1. Since $\mathbb{E}_{z^\star\sim\mathcal{D}}\left[\|z^\star - \mu\|^2_{C_\sigma^{-1}}\right] = d$, we have $\mathbb{E}_{z^\star\sim\mathcal{D}}[m^\star] = \lim_{n,d} \frac{d}{n} = \tau$. Thus, our results retrieve the "averaged" results of [SOJH09, Section T8.1.1], which we presented in Section 2.3.4. To prove their result, [SOJH09] uses an analysis tailored only for Bernoulli distributions. (a) The starting point of the proof is an exact characterisation of the LR score, only true for Bernoulli distribution, *i.e.* $\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) = \sum_{j=1}^d z_j^\star \log\left(\frac{\hat{\mu}_{n,j}}{\mu_j}\right) + (1 - z_j^\star)\log\left(\frac{1-\hat{\mu}_{n,j}}{1-\mu_j}\right)$ for every n [SOJH09, Sec. T1.2]. (b) The proof also uses specific concentration results of Bernoulli distributions [SOJH09, Sec. T8.1.2]. Our analysis generalises their results without assuming the knowledge specific to Bernoulli distributions.

Connection to the scalar product attack. [DSS+15] proposes a scalar product attack for tracing the empirical mean that thresholds over the score $s^{\text{scal}}(\hat{\mu}_n; z^\star, z^{\text{ref}}) \triangleq (z^\star - z^{\text{ref}})^T \hat{\mu}_n$. The intuition behind this attack is to compare the target-output correlation $(z^\star)^T\hat{\mu}_n$ with a reference-output correlation $(z^{\text{ref}})^T\hat{\mu}_n$. The analysis of [DSS+15] shows that with only one reference point $z^{\text{ref}} \sim \mathcal{D}$, and even for noisy estimates of the mean, the attack is able to trace the data of some individuals in the regime $d \sim n^2$. *Our asymptotic analysis shows that the LR test is also a scalar-product attack* (Equation (6.4)), but *corrected for the geometry of the data using the inverse covariance matrix*. If the data-generating distribution has the same variance over the columns, *i.e.* $C_\sigma = \sigma I_d$, the LR test and the scalar product test are equivalent up to a multiplicative constant, and thus, have the same power.

Explaining the privacy onion effect. Removing a layer of outlier points is equivalent to sampling from a new data-generating distribution, with a smaller variance. In this new data-generating distribution with smaller variance, the points which are not removed will naturally have an increased Mahalanobis distance. Thus, removing a layer of outlier points yields a layer of newly exposed target points. Hence, the Mahalanobis leakage score explains the privacy onion effect [CJZ+22].

Inherent privacy of the empirical mean. The hypothesis testing interpretation of Differential Privacy (DP) implies a trade-off between the Type-I and the Type-II errors of any adversary trying to infer the presence of any target point, under any data-generating distribution. Our results show that, under the specific threat model of the MI games, the empirical mean already

imposes a trade-off between the Type-I and Type-II errors of any MI adversary. Another way to interpret the result is that, if an auditor uses the fixed-target MI game to audit the privacy of the empirical mean, the auditor would conclude that the empirical mean is $\sqrt{m^\star}$-Gaussian DP [DRS19], or equivalently (ε, δ)-DP where for all $\varepsilon \geq 0$, $\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon}{\sqrt{m^\star}} + \frac{\sqrt{m^\star}}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon}{\sqrt{m^\star}} - \frac{\sqrt{m^\star}}{2}\right)$. The result is a direct consequence of Equation (6.3) and [DRS19, Corollary 2.13].

The column-wise independence asymption. Our analysis assumes that the data-generating distribution $\mathcal{D}$ is a product distribution, i.e. the columns of the input are independent. This assumption is standard and has been used in different related works in the tracing literature [HSR+08, SOJH09, DSS+15]. Our proof could be adapted to the dependent case using a *multivariate* Edgeworth expansion in the likelihood ratio test. The same conclusions of our analysis will follow, with the only difference being that the covariance matrix will no longer be diagonal but a full matrix. To rigorously use a high-dimensional *multivariate* Edgeworth expansion, additional technical assumptions must be added, making the analysis very technical without yielding additional insights. We leave it as a future direction to adapt the proof to the dependent case.

6.5 Impact of Privacy Defences and Misspecification on the Leakage

We quantify the effect of adding noise and sub-sampling on the leakage of the empirical mean. Both defences act like contractions of the leakage score. We also study the effect of target misspecification. The detailed proofs for this section are presented in Appendix D.

6.5.1 Effect of adding noise

We denote by $\mathcal{M}_n^\gamma$ the mechanism releasing the noisy empirical mean of a dataset using the Gaussian mechanism [DR14b]. Specifically, $\mathcal{M}_n^\gamma$ takes as input a dataset of size n of d-dimensional points, *i.e.* $D = \{Z_1, \dots, Z_n\} \in (\mathbb{R}^d)^n$, and outputs the noisy mean $\tilde{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n Z_i + \frac{1}{\sqrt{n}} N_d \in \mathbb{R}^d$, where $N_d \sim \mathcal{N}(0, C_\gamma)$ such that $\gamma = (\gamma_1, \dots, \gamma_d) \in \mathbb{R}^d$ and $C_\gamma = \operatorname{diag}(\gamma_1^2, \dots, \gamma_d^2) \in \mathbb{R}^{d\times d}$. Similar to Section 6.4, we assume that the data-generating distribution $\mathcal{D}$ is column-wise independent, has a mean $\mu \triangleq (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$, a covariance matrix $C_\sigma \triangleq \operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2) \in \mathbb{R}^{d\times d}$, and a finite $(4+\delta)$-th moment.

The LR score for $\mathcal{M}_n^\gamma$ is

$$\tilde{\ell}_n^\gamma(\tilde{\mu}_n; z^\star, \mu, C_\sigma) = \sum_{j=1}^d \log\left(\frac{\tilde{p}_{n,j}^{\text{in}}(\tilde{\mu}_{n,j}; z_j^\star, \mu_j, \sigma_j)}{\tilde{p}_{n,j}^{\text{out}}(\tilde{\mu}_{n,j}; z_j^\star, \mu_j, \sigma_j)}\right)$$

where $\tilde{\mu}_n = (\tilde{\mu}_{n,j})_{j=1}^d$, $\tilde{p}_{n,j}^{\text{out}}$ is the distribution of $\tilde{\mu}_{n,j}$ under $b=0$, and $\tilde{p}_{n,j}^{\text{in}}$ is the distribution of $\tilde{\mu}_{n,j}$ under $b=1$.

The output of $\mathcal{M}_n^\gamma$ could be re-written as $\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^n (Z_i + N_i)$, where $(N_i) \sim^{\text{i.i.d}} \mathcal{N}(0, C_\gamma)$. This means that $\mathcal{M}_n^\gamma$ could be seen as the exact empirical mean of n i.i.d samples from a new data-generating distribution $\tilde{\mathcal{D}} \triangleq \mathcal{D} \bigotimes \mathcal{N}(0, C_\gamma)$. *i.e.* $\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^n \tilde{Z}_i$, where $\tilde{Z}_i \sim \tilde{\mathcal{D}}$. This means that the results of Section 6.4 directly apply to $\mathcal{M}_n^\gamma$, by replacing $\mathcal{D}$ by $\tilde{\mathcal{D}}$. The Mahalanobis distance of $z^\star$ with respect to $\tilde{\mathcal{D}}$ is $\|z^\star - \mu\|_{(C_\sigma + C_\gamma)^{-1}}$, where $\|z^\star - \mu\|^2_{(C_\sigma+C_\gamma)^{-1}} = \sum_{j=1}^d \frac{(z_j^\star - \mu_j)^2}{\sigma_j^2 + \gamma_j^2}$. We call

$$\tilde{m}_\gamma^\star \triangleq \lim_{n,d} \frac{1}{n}\|z^\star - \mu\|^2_{(C_\sigma+C_\gamma)^{-1}}$$

the noisy leakage score.

Theorem 6.5 (Target-dependent leakage of the noisy empirical mean)**.** *As $d, n \to \infty$ s.t. $d/n = \tau$, $\tilde{\ell}_n^\gamma(\tilde{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{\tilde{m}_\gamma^\star}{2}, \tilde{m}_\gamma^\star\right)$ under H_0, and $\tilde{\ell}_n^\gamma(\tilde{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{\tilde{m}_\gamma^\star}{2}, \tilde{m}_\gamma^\star\right)$ under H_1.*

The asymptotic target-depend leakage of $z^\star$ in the noisy empirical mean is

$$\lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^\gamma, \mathcal{D}) = \Phi\left(\frac{\sqrt{\tilde{m}_\gamma^\star}}{2}\right) - \Phi\left(-\frac{\sqrt{\tilde{m}_\gamma^\star}}{2}\right).$$

The optimal trade-off function, achievable with the threshold $\tau_\alpha = -\frac{\tilde{m}_\gamma^\star}{2} + \sqrt{\tilde{m}_\gamma^\star}\Phi^{-1}(1-\alpha)$, is

$$\lim_{n,d} \text{Pow}_n(\tilde{\ell}_n, \alpha, z^\star) = \Phi\left(\Phi^{-1}(\alpha) + \sqrt{\tilde{m}_\gamma^\star}\right).$$

Theorem 6.5 shows that the Gaussian Mechanism acts by increasing the variance of the data-generating distribution, thus decreasing the Mahalanobis distance of target points and their leakage.

6.5.2 Effect of sub-sampling

We consider the *empirical mean with sub-sampling* mechanism [BBG18] $\mathcal{M}_n^{\text{sub},\rho}$ that uniformly sub-samples k_n rows without replacement from the original dataset, and then computes the exact empirical mean of the sub-sampled rows. $\mathcal{M}_n^{\text{sub},\rho}$ takes as input a dataset $D = \{Z_1, \ldots, Z_n\} \in (\mathbb{R}^d)^n$ and outputs $\hat{\mu}_{k_n}^{\text{sub}} \triangleq \frac{1}{k_n}\sum_{i=1}^n Z_i \mathbb{1}(\varsigma(i) \leq k_n)$. Here, $k_n \triangleq \rho n$, $0 < \rho < 1$ and $\varsigma \sim^{\text{unif}} S_n$ is a permutation sampled uniformly from the set of permutations of $\{1 \ldots, n\}$, *i.e.* S_n, and independently from $(Z_1, \ldots, Z_n)$. The LR score for $\mathcal{M}_n^{\text{sub},\rho}$ is

$$\ell_n^{\text{sub},\rho}(\hat{\mu}_n^{\text{sub}}; z^\star, \mu, C_\sigma) = \sum_{j=1}^d \log\left(\frac{p_{n,j}^{\text{in,sub}}(\hat{\mu}_{n,j}^{\text{sub}}; z_j^\star, \mu_j, \sigma_j)}{p_{n,j}^{\text{out,sub}}(\hat{\mu}_{n,j}^{\text{sub}}; z_j^\star, \mu_j, \sigma_j)}\right).$$

Here, $\hat{\mu}_n^{\text{sub}} = (\hat{\mu}_{n,j}^{\text{sub}})_{j=1}^d$, $p_{n,j}^{\text{out,sub}}$ is the distribution of $\hat{\mu}_{n,j}^{\text{sub}}$ under $b = 0$, and $p_{n,j}^{\text{in,sub}}$ is the distribution of $\hat{\mu}_{n,j}^{\text{sub}}$ under $b = 1$.

Theorem 6.6 (Target-dependent leakage of the sub-sampling empirical mean)**.** *As $d, n \to \infty$ s.t. $d/n = \tau$, $\ell_n^{\text{sub},\rho}(\hat{\mu}_n^{\text{sub}}; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{\rho m^\star}{2}, \rho m^\star\right)$ under H_0, $\ell_n^{\text{sub},\rho}(\hat{\mu}_n^{\text{sub}}; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{\rho m^\star}{2}, \rho m^\star\right)$ under H_1.*

The asymptotic target-dependent leakage of $z^\star$ in $\mathcal{M}_n^{\text{sub},\rho}$ is

$$\lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^{\text{sub},\rho}, \mathcal{D}) = \Phi\left(\frac{\sqrt{\rho m^\star}}{2}\right) - \Phi\left(-\frac{\sqrt{\rho m^\star}}{2}\right).$$

The optimal trade-off function obtained with $\tau_\alpha = -\frac{\rho m^\star}{2} + \sqrt{\rho m^\star}\Phi^{-1}(1-\alpha)$, is

$$\lim_{n,d} \text{Pow}_n(\ell_n^{\text{sub},\rho}, \alpha, z^\star) = \Phi\left(z_\alpha + \sqrt{\rho m^\star}\right).$$

Proof sketch. The proof uses the same three steps of the proof of Theorem 6.3. The additional technical hardness of this proof comes from the "mixture" nature of the "in" distribution $p_{n,j}^{\text{in,sub}}$, due to the sub-sampling. The detailed proof is presented in Appendix D.3.

Theorem 6.6 shows that the sub-sampling mechanism acts by increasing the number of "effective samples" from n to n/ρ, thus decreasing the leakage score.

6.5.3 Attacking with a misspecified target

Now, we suppose that the adversary has a misspecified target z^{targ}, *i.e.* different from the real $z^\star$ in the fixed-target MI game (Algorithm 12). The adversary then builds the LR test tailored for z^{targ}, *i.e.*

$$\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma),$$

where ℓ_n is defined in Section 6.4. The misspecified adversary will never be optimal, but it can still leak enough information depending on the amount of misspecification. In the following, we quantify the sub-optimality of the misspecified adversary, which we define as a measure of leakage similarity between z^{targ} and $z^\star$, *i.e.* we quantify how much z^{targ} leaks information about the presence of $z^\star$.

Theorem 6.7 (Leakage of a misspecified adversary)**.** *Let $\mathcal{A}_{\text{miss}}$ the adversary that uses the misspecified LR score $\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma)$. Then,*

$$\lim_{n,d} \text{Adv}_n(\mathcal{A}_{\text{miss}}) = \Phi\left(\frac{|m^{\text{scal}}|}{2\sqrt{m^{\text{targ}}}}\right) - \Phi\left(-\frac{|m^{\text{scal}}|}{2\sqrt{m^{\text{targ}}}}\right).$$

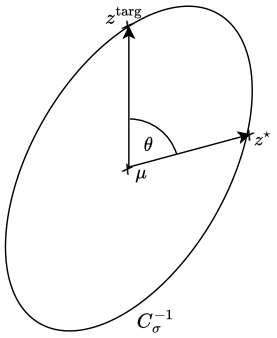


Figure 6.1 – The effect of misspecifying the target datum depends on the relative angle θ, between $z^\star - \mu$ and $z^{\text{targ}} - \mu$, corrected by C_σ^{-1}.

Here, $m^{\text{scal}} \triangleq \lim_{n,d} \frac{1}{n} \left(z^{\text{targ}} - \mu\right)^T C_\sigma^{-1} \left(z^\star - \mu\right)$ *and* $m^{\text{targ}} \triangleq \lim_{n,d} \frac{1}{n} \|z^{\text{targ}} - \mu\|^2_{C_\sigma^{-1}}$.

If the MI adversary had well specified the target datum, *i.e.* used $z^\star$ rather than z^{targ}, then they achieve the optimal asymptotic advantage

$$\lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^{\text{emp}}, \mathcal{D}) = \Phi\left(\frac{\sqrt{m^\star}}{2}\right) - \Phi\left(-\frac{\sqrt{m^\star}}{2}\right).$$

Theorem 6.7 *quantifies the sub-optimality of the misspecified adversary*, which is

$$\Delta(z^{\text{targ}}, z^\star) = \lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^{\text{emp}}, \mathcal{D}) - \text{Adv}_n(\mathcal{A}_{\text{miss}}).$$

Indeed, $\Delta(z^{\text{targ}}, z^\star) \geq 0$, since by the Cauchy Schwartz inequality, $|m^{\text{scal}}| \leq \sqrt{m^{\text{targ}} m^\star}$. The misspecified attack is still strong as long as $\sqrt{m^{\text{targ}} m^\star} - |m^{\text{scal}}| = \sqrt{m^{\text{targ}} m^\star}(1 - |\cos(\theta)|)$ stays small. We geometrically illustrate θ in Figure 6.1.

Proof sketch. The proof uses the same three steps of the proof of Theorem 6.3. The difference in the proof happens at the step of computing the expectations and variances before concluding using the Lindeberg-Feller theorem. The detailed proof is presented in Appendix D.4.

6.6 Experimental Analysis

We validate the theoretical analysis empirically on synthetic data.

We test: *Are the powers of the LR tests tightly determined by Theorem 6.3, Theorem 6.5, and Theorem 6.6 for the empirical mean, noisy empirical mean, and sub-sampled empirical mean mechanisms, respectively?*

Table 6.1 – Target-dependent leakage score in different settings

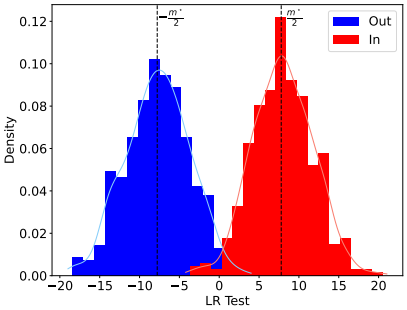
Setting	Leakage Score
Empirical mean	$\frac{1}{n}\|z^\star - \mu\|^2_{C_\sigma^{-1}}$
Gaussian Noise ($\gamma > 0$)	$\frac{1}{n}\|z^\star - \mu\|^2_{(C_\sigma + C_\gamma)^{-1}}$
Sub-sampling ($\rho < 1$)	$\frac{\rho}{n}\|z^\star - \mu\|^2_{C_\sigma^{-1}}$
Similar point	$\frac{1}{n}\left(z^\star_{\text{targ}} - \mu\right)^T C_\sigma^{-1}\left(z^\star_{\text{true}} - \mu\right)$

Experimental setup. We take $n = 1000$, $\tau = 5$, and thus, $d = 5000$. The data-generating distribution $\mathcal{D}$ is a d dimensional Bernoulli, with parameter $p \in [0,1]^d$. The three mechanisms considered are $\mathcal{M}_n^{\text{emp}}$, $\mathcal{M}_n^{\gamma}$ and $\mathcal{M}_n^{\text{sub},\rho}$. The adversaries chosen for each mechanism are the thresholding adversaries based on the asymptotic approximations of LR tests, as in our analysis. Finally, we choose three target data points in $\{0,1\}^d$. (a) The *easiest point to attack* $z^\star_{\text{easy}}$ is the point with the highest Mahalanobis distance with respect to p. It is the point with binary coordinates furthest away from those of p. $z^\star_{\text{easy}} = (\mathbb{1}(p_i \leq 1/2))_{i=1}^d$. (b) The *hardest point to attack* is $z^\star_{\text{hard}} = (\mathbb{1}(p_i > 1/2))_{i=1}^d$, that has the coordinates closest to p. (c) A *medium point to attack* $z^\star_{\text{med}}$ is randomly sampled from the data-generating distribution $\text{Bern}(p)$, for which the Mahalanobis distance and the leakage score are of orders d and $\tau = d/n$.

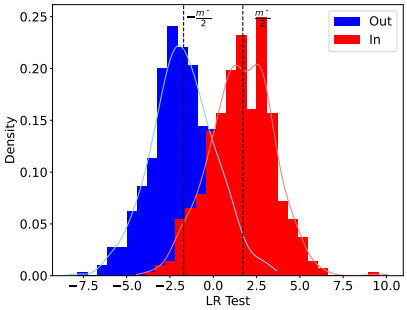
Results and discussions. We illustrate the results in Figure 6.2. (a) *Impact of $m^\star$, noise, and sub-sampling ratio:* Figure 6.2 (a) shows that the power of LR test uniformly increases with an increase in $m^\star$. Figure 6.2 (b) shows that the power of the LR test uniformly decreases with an increase in the noise variance γ^2 of the Gaussian mechanism. Figure 6.2 (c) shows that the power of the LR test uniformly decreases with a decrease in the sub-sampling ratio ρ. (b) *Tightness of the power of test analysis:* Figure 6.2 validates that our theoretical analysis tightly captures the impacts of the target-dependent hardness of leakage and privacy-preserving mechanisms on the experimental ROC curves.

6.7 Conclusion

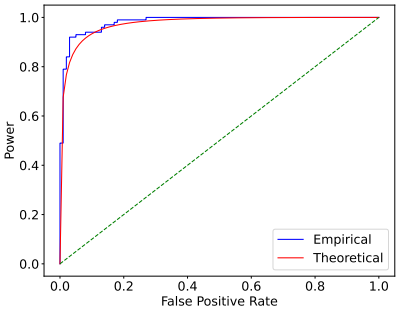
We study fixed-target MI games, and characterise the target-dependent leakage and trade-off functions of the empirical mean and its variations. We summarise the results in Table 6.1. We show that the leakage due to a target point depends on its Mahalanobis distance from the mean of the data generating distribution, and captures precisely the hardness of fixed-target MI games. Our generic analysis captures the impact of different DP mechanisms, like Gaussian noise addition, sub-sampling, and using a misspecified target datum on the leakage. Finally, we numerically validate our theoretical results.



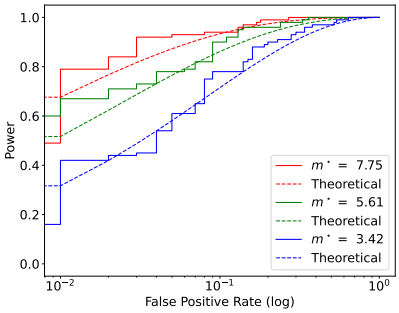
(**a**) LR distribution on easy target $z^{\star}_{\text{easy}}$



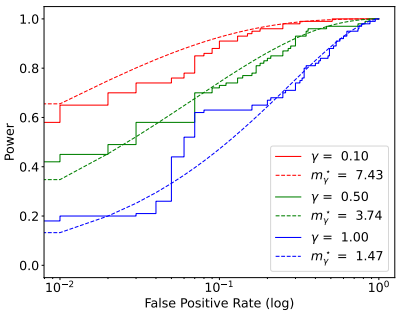
(**b**) LR distribution on hard target $z^{\star}_{\text{hard}}$



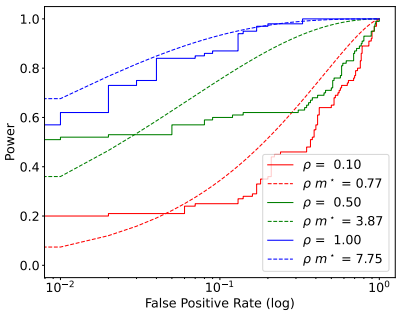
(**c**) Theoretical and empirical trade-offs



(**d**) Effect of $m^{\star}$



(**e**) Effect of γ



(**f**) Effect of ρ

Figure 6.2 – Experimental demonstration of the theoretical results and impacts of $m^{\star}$, noise, and subsampling ratio on leakage. Dotted lines represent theoretical bounds and solid lines represent the empirical results.

Chapter 7

White-Box Membership Inference Games for Gradient Descents

In this chapter, we focus on auditing supervised learning gradient descent algorithms. The main observation is that gradient descent algorithms operate by sequentially updating a parameter estimate θ_t *in the direction of the empirical mean of gradients*. Thus, if an auditor has access to all the intermediates parameters $\{\theta_t\}_t$, *i.e.* the white-box federated learning setting, auditing gradient descent algorithms reduces to auditing the empirical mean mechanism. Using this observation, we use the results of target-dependent MI games for the empirical mean from Chapter 6 to propose (a) an optimal covariance attack for gradient descents and (b) an optimal canary selection strategy based on the Mahalanobis leakage score. We test the two methods for a logistic regression algorithm trained on the FMNIST dataset and a CNN model trained on the CIFAR10 dataset. Our results show that the covariance attack improves over the scalar product attack for gradients. Also, the Mahalanobis score predicts the hardness of the MI game well and thus provides good candidates for canaries. We also connect our attack and canary strategy to the heuristics proposed in the white-box federated learning auditing literature.

Contents

7.1 The White-Box Federated Learning Setting

In this section, we present the white-box federated learning setting for privacy auditing. Then, we discuss techniques from the literature in this setting, both attack strategies and canary selection strategies.

7.1.1 Presentation of the threat model

First, we recall the learning problem's setting. We adhere to supervised learning, where the (private) input dataset contains n examples of features and label pairs, *i.e.* $D \triangleq \{(x_i, y_i)\}_{i=1}^n$. The goal in supervised learning is to learn a model f that explains well the dataset D. We suppose that the model f is parameterised, *i.e.* $f = f_\theta$. This reduces the learning problem to finding the best parameter $\theta \in \mathbb{R}^d$ which explains the dataset D with respect to a loss function ℓ, *i.e.* to find $\theta^\star \triangleq \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(f_\theta(x_i), y_i)$. Throughout this chapter, we only focus on gradient descent algorithms. Gradient Descent algorithms start with an initial parameter $\theta_0 \in \mathbb{R}^d$, and then update sequentially the parameter at each step t by

$$\theta_t \triangleq \theta_{t-1} - \eta_t \nabla_{\theta_{t-1}} Q(\theta_{t-1}),$$

where η_t is the learning rate at step t, and $Q(\theta_{t-1})$ is a quantity that depends on the loss on "some input samples". For example,

(a) in batch gradient descent

$$\nabla_{\theta_{t-1}} Q(\theta_{t-1}) \triangleq \frac{1}{n}\sum_{i=1}^n \nabla_{\theta_{t-1}} \ell(f_{\theta_{t-1}}(x_i), y_i)$$

is the gradient with respect to the whole dataset.

(b) in mini-batch gradient descent, the dataset is divided into a set of mini-batches $D = \cup B_k$. At each step t, a mini-batch B is sampled uniformly and

$$\nabla_{\theta_{t-1}} Q(\theta_{t-1}) \triangleq \frac{1}{|B|}\sum_{i\in B} \nabla_{\theta_{t-1}} \ell(f_{\theta_{t-1}}(x_i), y_i).$$

We call $|B|$ the batch size.

(c) in stochastic gradient descent,

$$\nabla_{\theta_{t-1}} Q(\theta_{t-1}) \triangleq \nabla_{\theta_{t-1}} \ell(f_{\theta_{t-1}}(x_i), y_i),$$

where $i \sim \mathcal{U}([1, n])$ is sampled randomly from $\{1, n\}$.

(d) in DP-SGD [ACG+16],

$$\nabla_{\theta_{t-1}} Q(\theta_{t-1}) \triangleq \left(\frac{1}{|B|} \sum_{i \in B} \text{ClipNorm}_C \left[\nabla_{\theta_{t-1}} \ell(f_{\theta_{t-1}}(x_i), y_i) \right] \right) + \mathcal{N}\left(0, \gamma^2 C^2 I_d\right),$$

where B is again a batch uniformly sampled, $\text{ClipNorm}_C(x) \triangleq \min\{1, C/\|x\|\}x$ is the clipping norm function, *i.e.* $\text{ClipNorm}_C(x) = x$ if $\|x\| \leq C$ otherwise $\text{ClipNorm}_C(x) = C\frac{x}{\|x\|}$. Here, $C > 0$ is a gradient norm bound and $\gamma > 0$ is the noise magnitude. DP-SGD can be shown to verify a DP constraint, where the privacy budget depends on the clipping gradient norm C, the noise γ, the batch size $|B|$ and the number of gradient iterations T.

The goal is to audit the privacy guarantee of gradient descent algorithms. Specifically, the mechanism to be audited is the "training" algorithm that takes as input the private dataset $D \triangleq \{(x_i, y_i)\}_{i=1}^n$, and produces sequentially the parameter estimates $\{\theta_t\}_{t=1}^T$. Depending on which part the auditor can observe, it is possible to define different threat models. In the white-box federated learning setting, the auditor can observe the full sequence of iterates $\{\theta_t\}_{t=1}^T$. On the other hand, in the white-box final parameter setting, the auditor only observes the last iterate θ_T. For both these two settings, the auditor observes the value of the parameter $\theta \in \mathbb{R}^d$, either the last one or the full intermediate ones. In contrast, the auditor has only query access to the final model θ_t in the black box model, *i.e.* the auditor can choose an input x to send to the model, and observes the output $f_{\theta_T}(x)$ of the final model.

The white-box federated learning setting is a fundamental setting for understanding privacy auditing. In addition, this setting has practical uses too. For example, a potential use of this setting is for "debugging" implementations of private gradient descent algorithms. In this case, the auditor is the programmer trying to release an implementation of their favourite private descent algorithm. To verify the guarantees of their algorithm, the programmer runs a white-box federated learning audit on their implementation and compares the empirical privacy guarantees retrieved with the theoretical analysis. The white-box federated learning setting captures well this use case since the auditor *is* the programmer themselves, and thus have access to all the intermediate iterates.

On the other hand, as the name of the setting suggests, another natural application of this threat model is auditing the private Federated Learning (FL) protocol. In a standard FL protocol (Figure 7.1), a server computes a global model θ_t at each step t by taking the (noisy) average of client updates, where each client computes a gradient estimate using their local datasets. We suppose that the auditor is also a client in an "honest-but-curious" FL protocol, *i.e.* the auditor plays the role of the "adversary" but follows the rules of the protocol: At each step t of the protocol, the server sends a global model θ_t to each client. Then, each client i computes their local gradient update $g_{t,i}$ and sends it to the server. The auditor, being also a client in the protocol, computes a gradient update $g_t^\star \triangleq \nabla_{\theta_t} \ell(f_{\theta_t(x^\star)}, y^\star)$ on their local "canary"

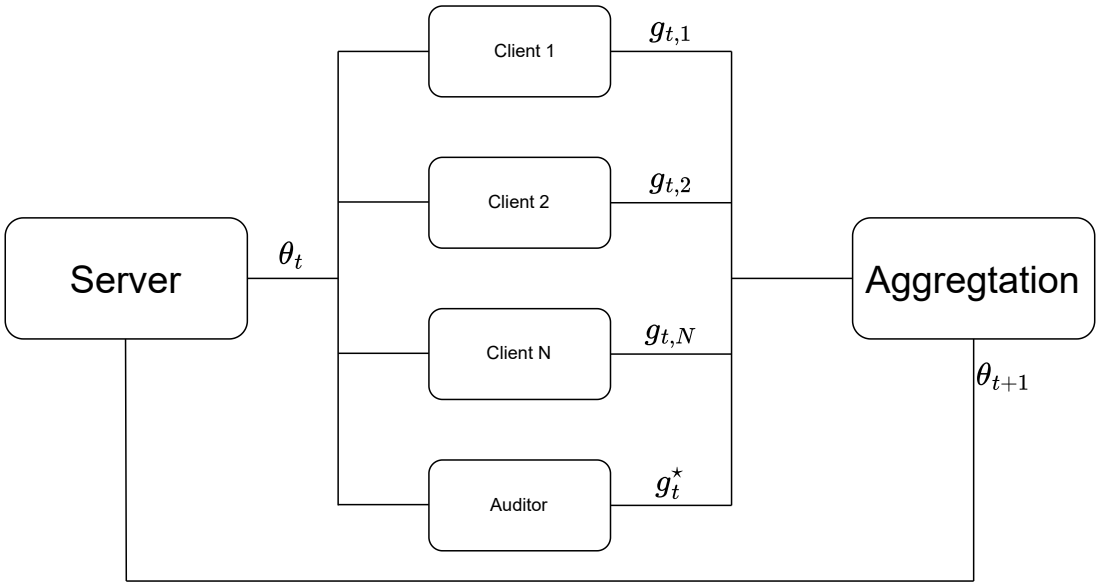


Figure 7.1 – The White-box Federated learning threat model. At each step, the server sends a global model θ_t to each client. The auditor is a client, too. Each client i computes and sends the local update $g_{i,t}$ to the server. The auditor also computes the local update $g^\star$ on a canary $z^\star$, and sends it to the global server with probability $1/2$. The server aggregates the updates received to compute θ_{t+1}.

datapoint $z^\star \triangleq (x^\star, y^\star)$, *i.e.* the most-leaking sample. Then, the auditor sends the gradient update to the server with probability $1/2$, or otherwise sends nothing to the server. In the next step of the interaction, the auditor observes the new updated global model θ_{t+1}. The goal of the auditor is to decide based solely on θ_{t+1}, θ_t and $g_t^\star$ whether the canary update was indeed sent to the server or not. This is exactly a tracing problem of the aggregation mechanism, which is generally a variant of the empirical mean. To design an audit in this setting, the auditor has to design two algorithms. First, an attack score to determine based on θ_{t+1}, θ_t and $g_t^\star$ whether $g_t^\star$ was included in θ_{t+1} or not, *i.e.* a score function that takes as input θ_{t+1}, θ_t and $g_t^\star$ and outputs a score that would be high if the canary was included, otherwise the score is low. The auditor also needs to choose a good canary $z_t^\star$ at each step t. A good canary should be an easy point to trace, *i.e.* , one for which the target-dependent leakage is high.

To summarise, auditing a gradient descent algorithm (batch, minibatch and DP-SGD) in the white-box federated learning setting reduces to auditing variants of the empirical mean mechanism, applied to loss gradient data $\{\nabla_{\theta_{t-1}} \ell(f_{\theta_{t-1}}(x_i), y_i)\}_{i=1}^n$. In turn, as explained above and also in more detail in Section 2.3.7, designing an audit algorithm reduces to designing an MI attack (or score) and a canary-selection strategy which determines how to choose the target point for the MI attack. In the following, we present the main techniques used in the white-box federated learning auditing literature. Then, we show how our results from analysing the

target-dependent hardness of MI games can be used to design a new score and a new canary selection strategy for white-box gradient descent algorithms.

7.1.2 Related works

We present the MI scores and canary selection strategies used in the white-box federated learning literature.

Attack scores. The scalar product [DSS+15] is the most popular score used in white-box attacks in the literature [MSS22, NHS+23, SNJ23, AKO+23]. The scalar product score takes as input θ_{t+1}, θ_t and $g_t^\star$ and outputs the scalar product $\langle \theta_{t+1} - \theta_t, g_t^\star \rangle$. This score is a direct application of the tracing attack against of [DSS+15] to the white-box federated learning setting, since $\theta_{t+1} - \theta_t$ is an empirical-mean like quantity.

On the other hand, [LPK23] proposes a different score attack, called Gradient Likelihood Ratio (GLiR) Attack, *i.e.* Algorithm 1 in [LPK23]. This attack is based on an analysis of the LR test of the empirical mean. However, the analysis of [LPK23] arrives at a different score function compared to our results. Specifically, let $g_{\text{batch}}^t = \frac{\theta_{t+1}-\theta_t}{\eta_t}$ be the batch gradient. Then, the GLiR attack needs to estimate an empirical mean $\hat{\mu}_0$ and covariance $\hat{C}_0$ of reference data's gradient. Then, the attack computes the statistics $\hat{S} = (|B|-1)(g_{\text{batch}}^t - g_t^\star)^T \hat{C}_0^{-1}(g_{\text{batch}}^t - g_t^\star)$ and $\hat{K} = \|\hat{C}_0^{-1/2}(\hat{\mu}_0 - g_t^\star)\|$. The GLiR score is

$$s^{\text{GLiR}}(g_{\text{batch}}^t, g_t^\star) = \log\left(F^{-1}_{\chi^2_d(|B|\hat{K})}\left(\hat{S}\right)\right)$$

where $F^{-1}_{\chi^2_d(\gamma)}$ is the inverse of the CDF of the non-central chi-squared distribution with d degrees of freedom and non-centrality parameter γ and $|B|$ is the batch size. For some threshold τ, if $s^{\text{GLiR}}(g_{\text{batch}}^t, g_t^\star) < \tau$, the GLiR attack suggests that $g^\star$ was included, otherwise it was not.

We connect the GLiR score to our covariance attack in Section 7.2, and provide some comments on the LR analysis of [LPK23].

Canary selection strategies. The main intuition for canary selection strategies in the literature is to propose heuristics to generate out-of-distribution data [JUO20, MSS22, NHS+23, SNJ23, AKO+23]. For example, [NHS+23] proposes the Dirac canary strategy which suggests to use as canaries gradient updates with all the values zero except at a single index. The intuition behind this choice is that the sparse nature of the Dirac gradient makes it an out-of-distribution sample in natural datasets. For CIFAR10, [NHS+23] shows the effectiveness of such a canary choice for auditing neural nets in the white-box federated learning setting. The Dirac canary is a type of *gradient canaries*, because it directly suggests what gradient $g^\star$ the auditor should include or not in the training. The other type of canaries is *input canaries*. As the name suggests, input canaries are pairs of features and label $z^\star = (x^\star, y^\star)$ that the auditor chooses to include

or not in the training. There are different heuristics in the literature to generate input canaries, *i.e.* mislabeled examples or blank examples [NHS+23], or adversarial examples [JUO20]. In Section 7.3, we explain the experimental success of Dirac canaries using our Mahalanobis leakage score, and propose a new canary selection strategy based on Mahalanobis leakage score that can be used to generate both gradient and input canaries.

On the other hand, [MSS22] proposes CANIFE, an algorithm that learns to craft canaries by back-propagating to the input level the following loss function

$$\ell^{\text{CANIFE}}(z^\star) \triangleq \sum_i^{n_r} \langle u_i, g_t^\star \rangle^2 + \max(C - \|g_t^\star\|, 0)^2. \tag{7.1}$$

Here, $g_t^\star \triangleq \nabla_{\theta_t} \ell(f_{\theta_t}(x^\star), y^\star)$ is the gradient of the loss of the canary $z^\star$ at step t, u_i is the gradient of the loss with respect to a reference sample and n_r is the number of reference samples. CANIFE can be used to craft an input canary $z^\star$ by directly minimising the loss ℓ^{CANIFE}. We connect the CANIFE loss to our Mahalanobis score in Section 7.3.

7.2 The Covariance Score for Gradient Descents

We present our covariance attack in Algorithm 13. Given a target gradient of $g_t^\star$ and the batch gradient $g_{\text{batch}}^t \triangleq \frac{\theta_{t+1}-\theta_t}{\eta_t}$, the covariance attack at step t uses the empirical LR score of Equation (6.5) to provide the covariance score

$$s_t^{\text{cov}} = (g_t^\star - \hat{\mu}_0^t)^T (\hat{C}_0^t)^{-1} \left(g_{\text{batch}}^t - \hat{\mu}_0^t\right) - \frac{1}{2|B|} \|g_t^\star - \hat{\mu}_0^t\|^2_{(\hat{C}_0^t)^{-1}}. \tag{7.2}$$

Here, $\hat{\mu}_0^t$ and $\hat{C}_0^t$ are the estimated empirical mean and empirical variance of the loss computed on reference samples $\{(x_1, y_1), \ldots, (x_{n_r}, y_{n_r})\}$ at step t.

Let us suppose that the auditor wants to run the attack over *only one step* t of the gradient descent iterations. Then, an attack over a "one-step update" is exactly an attack on the "empirical mean" g_{batch}^t. In this case, $\hat{\mu}_0^t$ and $\hat{C}_0^t$ in Equation (7.2) are estimated using the gradients at step t of the reference samples, *i.e.* $\hat{\mu}_0^t = \frac{1}{n_r}\sum_i^{n_r} u_i^t$ and $\hat{C}_0^t = \frac{1}{n_r}\sum_i^{n_r} u_i^t (u_i^t)^T$, where $u_i^t \triangleq \nabla_{\theta_t} \ell(f_{\theta_t}(x_i), y_i)$.

The covariance attack can also be used to trace gradient descent algorithms over *multiple steps of gradient iterations*. At each step t, the attack computes the score s_t^{cov}. Then, the final score over T iterations is just the sum of scores $\sum_{t=1}^T s_t^{\text{cov}}$. This sum is big if at least the target datapoint $z^\star$ is detected at one iteration. Otherwise, if the target is not detected at any step, the sum of scores is low. For the covariance attack of Algorithm 13 to be able to trace the presence of a target point $z^\star = (x^\star, y^\star)$

Algorithm 13 The covariance attack

Input: Estimated $(\hat{\mu}_0, \hat{C}_0)$, canary $z^\star = (x^\star, y^\star)$, learning rates (η_t), batch size $|B|$, number of gradient steps T.
for $t = 1, \cdots, T$ **do**
  Set $g_t^\star = \nabla_{\theta_t} \ell(f_{\theta_t}(x^\star), y^\star)$
  Set $g_{\text{batch}}^t = (\theta_{t+1} - \theta_t)/\eta_t$
  Compute $s_t^{\text{cov}} = (g_t^\star - \hat{\mu}_0)^T \hat{C}_0^{-1} (g_{\text{batch}}^t - \hat{\mu}_0) - \frac{1}{2|B|} \|g_t^\star - \hat{\mu}_0\|^2_{\hat{C}_0^{-1}}$
end for
Return $\sum_{t=1}^T s_t^{\text{cov}}$

Ideally, each score s_t^{cov} in Equation (7.2) should be computed with newly estimated mean $\hat{\mu}_0^t$ and covariance $\hat{C}_0^t$ of the gradients at step t. However, this may be computationally expensive. To deal with this, in Algorithm 13, we only estimate $\hat{\mu}_0$ and covariance $\hat{C}_0$ once at the beginning of an epoch, *i.e.* using θ_0. The covariance attack at each step t is computed using the same estimated $\hat{\mu}_0^t$ and $\hat{C}_0^t$, and then summed over the iterations. This means that at each step $t > 1$, the covariance attack of Algorithm 13 uses a misspecified reference empirical mean and covariance. On the other hand, if the number of iterations is small enough, for example by running the attack on only one epoch, then we can argue that the model parameters did not change much, and the effect of misspecification is negligible.

The covariance attack of Algorithm 13 is provably better than the scalar product attack, at the expense of estimating the inverse of the covariance matrix well. The shape of the covariance matrix is $d \times d$, where d is the number of parameters of the model θ. This means that storing and inverting this covariance matrix is computationally expensive for models with many parameters. A simple trick to deal with this problem is only running the attack on a subset of the parameters. For example, we can run the covariance attack over the last layer of a neural net. If the last layer has d_ℓ parameters, the covariance matrix becomes $d_\ell \times d_\ell$ with $d_\ell \ll d$.

As Section 7.1.2 explains, [LPK23] also provides a score based on analysing the LR test. We provide the following remarks to connect our analysis to that of [LPK23].

(a) In Step 1 of the proof, in [LPK23, Section E.1.] declares that "We suppose that the number of averaged samples is sufficiently large such that we can apply the Central Limit Theorem", and thus, considers[1] under H_0,

$$\hat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{1}{n} C_\sigma\right) \tag{7.3}$$

and under H_1,

$$\hat{\mu}_n \sim \mathcal{N}\left(\frac{1}{n} z^\star + \frac{n-1}{n}\mu, \frac{n-1}{n^2} C_\sigma\right) . \tag{7.4}$$

[1] Following equations are restatements of Equations 45 and 46 of [LPK23] using our notations.

However, the Central Limit Theorem is a "limit in distribution" of the empirical means. Thus, *the limit distribution of $\hat{\mu}_n$ is just the constant μ under both H_0 and H_1.* The effect of $z^\star$ disappears in this statement, as $n \to \infty$. For their claim to be "rigorously" correct, one should assume that the data-generating distribution is exactly a Gaussian distribution. This gives the exact distribution of $\hat{\mu}_n$ under H_0 and H_1 as expressed by the two equations above. Supposing that the data-generating distributions are Gaussian distributions simplifies the analysis, since now there is no need to go for asymptotics in n and d, and thus, there is no need for Edgeworth expansions and Lindeberg CLT. In contrast, our results provide a way to rigorously justify under which conditions this holistic view of "equivalence to testing between Gaussians" is correct, *i.e.* finite 4-th moment of the data distribution.

(b) As a score function, [LPK23] chooses to analyse the distribution of the "norm squared" of a re-centred and normalised version of the mean *i.e.* $S_n \approx \|C^{-1/2}(\hat{\mu}_n - \mu)\|^2$ while hiding some constants specific to their analysis. They characterise the distribution of S_n and show that it is a (scaled) non-central chi-squared distribution with d degrees of freedom, with different parameters under H_0 and H_1. In our case, we provide the asymptotic distribution of the LR score directly under H_0 and H_1, which provides a simpler covariance score.

7.3 Choosing Canaries Using the Mahalanobis Distance

We present our Mahalanobis-based canary selection strategy in Algorithm 14.

Algorithm 14 can either be run to generate a gradient canary or an input canary. The algorithm takes as input candidate canaries, either gradients or inputs, and outputs the easiest point to attack between the proposed candidates. To do so, Algorithm 14 starts by estimating the reference empirical mean $\hat{\mu}_0$ and covariance matrix $\hat{C}_0$ of gradients over reference points. Then, for each canditate canary k, Algorithm 14 computes its estimated Mahalanobis score $m_k^\star$ with respect to the estimated reference means $\hat{\mu}_0$ and $\hat{C}_0$. Finally, the algorithm returns the candidate with the highest estimated Mahalanobis score, *i.e.* the easiest point to attack according to the results of Chapter 5.

In addition to proposing a new gradient and a new input canary strategy, our Mahalanobis score also explains the success of the heuristics presented in Section 7.1.2. Specifically, Dirac canaries, black examples, or mislabeled examples are all points with high Mahalanobis distance, thus making them great canary candidates. Our Mahalanobis score can also be run over "in-distribution" canary candidates to find the most leaking one over them. This could come in handy when the auditor, while trying to participate in the white-box audit protocol (e.g. Figure 7.1), does not want to hurt the accuracy of the final model. Thus, the auditor wants to send gradient updates that are helpful for accuracy, *i.e.* "in-distribution", but with a high enough Mahalanobis score to be distinguishable. The Mahalanobis score solves the tradeoff

Algorithm 14 Canary selection strategy using the Mahalanobis score.

Input: $\{(x_1, y_1), \ldots, (x_{n_r}, y_{n_r})\}$ reference points, $\{(x_1^\star, y_1^\star), \ldots, (x_{n_c}^\star, y_{n_c}^\star)\}$ candidate input canaries, $\{g_1^\star, \ldots, g_{n_c}^\star\}$ candidate gradient canaries.
Step1: Estimate the empirical mean and covariance of the gradients
Initialise weights and biases of the model *i.e.* set θ^0.
for $i = 1, \cdots, n_r$ **do**
    Compute $u_i = \nabla_{\theta_0} \ell(f_{\theta_0}(x_i), y_i)$
end for
Compute $\hat{\mu}_0 = \frac{1}{n_r} \sum_i^{n_r} u_i$ and $\hat{C}_0 = \frac{1}{n_r} \sum_i^{n_r} u_i u_i^T$
Return $(\hat{\mu}_0, \hat{C}_0)$
Step2: Compute the Mahalanobis score for the candidates, using the estimated mean and covariance
if "Input Canaries" **then**
    for $k = 1, \cdots, n_c$ **do**
        Compute $g_k^\star = \nabla_{\theta_0} \ell(f_{\theta_0}(x_k), y_k)$
        Compute the Mahalanobis score $m_k^\star = \|g_k^\star - \hat{\mu}_0\|^2_{\hat{C}_0^{-1}}$
    end for
    Return $(x_{k^\star}^\star, y_{k^\star}^\star)$ where $k^\star \triangleq \arg\max_{k=1}^{n_c} m_k^\star$
else "Gradient canaries"
    for $k = 1, \cdots, n_c$ **do**
        Compute the Mahalanobis score $m_k^\star = \|g_k^\star - \hat{\mu}_0\|^2_{\hat{C}_0^{-1}}$
    end for
    Return $g_{k^\star}^\star$ where $k^\star \triangleq \arg\max_{k=1}^{n_c} m_k^\star$
end if

between the "accuracy of the model" and the "success of the MI attack" by choosing points with a moderate Mahalanobis score.

Finally, our Mahalanobis score also explains the CANIFE loss ℓ^{CANIFE} of Equation (7.1). In [MSS22, Appendix A], the intuition to explain the CANIFE loss starts by expressing the LR score between two Gaussian distributions. Then, [MSS22] concludes that to make the two Guassians distinguishable (separable) enough, one should maximise $(g_t^\star)^T C^{-1} g_t^\star$, which is precisely the Mahalanobis score. Finally, they claim that maximising the score is "equivalent" to minimising $(g_t^\star)^T C g_t^\star$, which yields exactly the CANIFE loss ℓ^{CANIFE} when substituting $C = \sum_i u_i u_i^\top$. Thus, the Mahalanobis score also explains the CANIFE loss, and our results rigorously justify the success of this approach beyond Gaussian distributions.

7.4 Experimental Analysis

We test: *Does the Mahalanobis leakage score explain the target-dependent hardness of MI games on real datasets? Does the covariance-corrected LR attack improve the scalar product attack?*

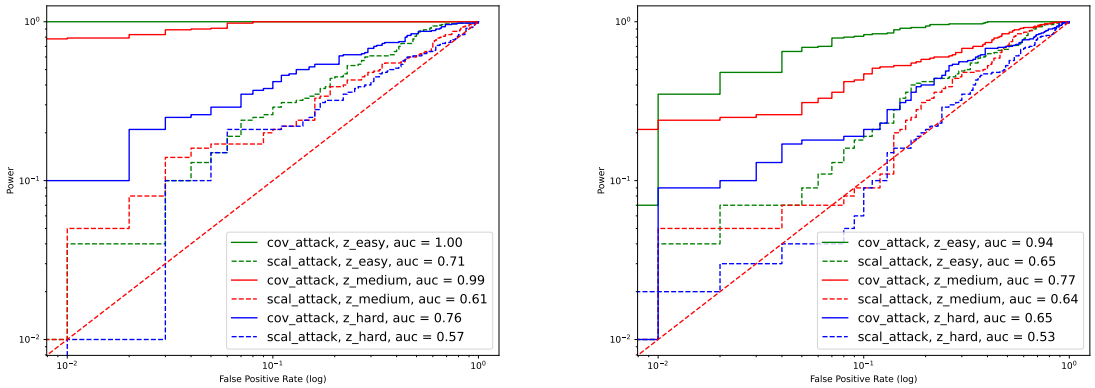


Figure 7.2 – Covariance and scalar product attacks.

Experimental Setup. We attack two models. We train a single linear layer $(\text{Linear}(28 \times 28, 10))$, *i.e.* a Logistic Regressor, on Fashion MNIST (FMNIST) [XRV17]. Thus, the number of weights and biases in the model is $d_F = 7850$. We train a Convolutional Neural Net (CNN) [LBH15] with three convolutional layers and a final linear layer on CIFAR10 [KH+09].

The number of weights and biases in CNN is $d_C = 18786$, while the last linear layer has $d_L = 4080$ parameters. We use a mini-batch SGD with a batch size 64, learning rate 10^{-3}, and cross-entropy loss for training. We attack the models in a white-box FL setting [MSS22]. First, we run Algorithm 14 to estimate the empirical mean and covariance, using $n_r = 20$k reference points from training data. Then, we estimate the Mahalanobis score for every point in the training data. Finally, we run the covariance attack (Algo. 13) and a scalar product attack on the points in the training data with the highest and lowest Mahalanobis scores, *i.e.* z_{easy} and z_{hard}, respectively. The scalar product attack replaces the score s_t in Algo. 13 with the scalar product score $s_t^{\text{scal}} = (g_t^\star)^T g_{\text{batch}}^t$. Both attacks are run *only* on one epoch of SGD, *i.e.* one loop over the training data. For FMNIST, the attack is implemented with the *full* gradient of the loss. For CIFAR10, we only attack the last linear layer of the CNN, leading to $d_L \times d_L$ covariance matrix rather than $d_C \times d_C$. This improves our attack's time and space complexity by storing and inverting a smaller matrix. It still maintains the strength of the tracing attack since d_L is still significantly larger than the batch size. We show the ROC curves of the two attacks against easy and hard targets of FMNIST and CIFAR10 in Fig. 7.2.

Results and discussion. Figure 7.2 shows: (a) The point with the highest Mahalanobis score is easier to attack than the point with the lowest Mahalanobis score for both the datasets and models. (b) The covariance attack improves on the scalar product attack. (c) Also, in practice, we can run the covariance attack over one epoch of training and use the same covariance matrix computed with only the last layer at each step.

7.5 Conclusion

This chapter introduces the covariance attack (Algorithm 13) and Mahalanobis canary selection strategy (Algorithm 14) for white-box federated learning auditing of gradient descent algorithms. These two algorithms are a direct consequence of analysing the target-dependent hardness of MI games on variants of the empirical mean mechanisms. Also, these two algorithms can be directly plugged into state-of-the-art auditing procedures, where our new covariance attack can replace the scalar product, and our Mahalanobis gradient can replace heuristics used in the literature (*i.e.* Dirac canaries and black example canaries). Our two algorithms have the advantage of being provably optimal. On the other hand, the price of optimality is the computational burden of estimating, storing and inverting a covariance matrix. We propose only running the attack on the last layer for big neural nets to deal with this extra computational expense.

Part III

Conclusion and Perspectives

Chapter 8

Conclusion

In this thesis, we have contributed to the field of privacy-preserving data analysis. The main framework to define privacy throughout the thesis is Differential Privacy (DP). In the first part, the main setting for studying utility is the stochastic multi-armed bandit problem. For the second part, the main framework for analysing the leakage of mechanisms is Membership Inference (MI) games. From both attack and defence points of view, the thesis aims to quantify the utility-privacy tradeoffs. For bandits, we express the privacy-utility tradeoffs as lower bounds and matching upper bounds of bandit algorithms' regret and sample complexity. For MI games, we express the tradeoffs as bounds on the power of any adversary trying to infer the presence of a target data point.

In Part I, we study the complexity of bandits under privacy constraints. First, in Chapter 3, we propose four extensions of DP to the bandit setting: View DP, Table DP, Interactive DP and DP in the adaptive continual release model. Each definition deals differently with the challenges of extending DP, i.e. the online and sequential nature of the bandit interaction with partial feedback. We also provide different relations between the definitions. Then, in Chapter 4, we provide lower bounds on the minimal regret and sample complexity that any DP policy should satisfy. We provide a generic proof of the lower bounds, all generated from a "stochastic generalisation" of group privacy using coupling techniques. The lower bounds show the existence of two regimes of hardness: a low privacy regime where the hardness of private bandits reduces to the hardness of non-private bandits, and a high privacy regime where the extra price of privacy is characterised by new information-theoretic quantities based on the total variation (i.e. the $\mathrm{TV}_{\mathrm{Inf}}$ and $T^{\star}_{\mathrm{TV}}$). Approximatively, the change of regmine from low to high privacy happens at $\varepsilon \approx \Delta$, where Δ is the order of the mean gaps. The lower bounds of Chapter 4 provide a target for optimal algorithm design. In Chapter 5, we provide a generic wrapper to generate near-optimal private bandits. The main intuition of the wrapper is to compute the main private statistics of the algorithm on non-overlapping sequences, to add less noise thanks to parallel composition. The challenge is to find ways to run state-of-the-art bandit algorithms

in a non-overlapping way, for example, by running in phases without forgetting. We instantiate the wrapper for finite-armed bandits under ε-Interactive DP (AdaP-UCB and AdaP-KLUCB) and under ρ-Interactive zCDP (AdaC-UCB), for linear (AdaC-GOPE) and contextual bandits (AdaC-OFUL) under ρ-Interactive DP, and for Best-arm Identification under ε-Interactive DP (AdaP-TT and AdaP-TT*). For each algorithm, we provide a privacy guarantee resulting from the generic wrapper and a utility guarantee in terms of upper bounds on the regret or sample complexity. The upper bounds of each algorithm are compared to the lower bounds and match up to constants or logarithmic terms, depending on the setting. The theoretical upper bounds also reflect the two hardness regime observations from the lower bounds, which in turn is validated by the experimental analysis.

In Part II, we study fixed-target Membership Inference (MI) games. In MI games, the goal of an adversary is to determine whether a target point was included or not in the input dataset of a mechanism by only looking at its output. The fixed-target MI game is a threat model for MI games that models the MI problem as a game between a crafter and an adversary. The specificity of this game is that the target point is fixed throughout the game. This means that the metrics of the game, i.e. the advantage and power of the adversary, are target-dependent. Our main goal is to quantify the target-dependent advantage and power of the optimal adversaries (the LR test), which we call the target-dependent leakage. For the empirical mean and variants of interest, we use asymptotic properties from generalisations of the Central Limit Theorem (CLT) to quantify the optimal advantage and power of the optimal LR attack. The main result of the analysis is that the hardness of the fixed-target MI game depends on the Mahalanobis distance between the target $z^\star$ and the data-generating distribution. Specifically, target points with high Mahalanobis distance are easy to attack, while points with low Mahalanobis distance are hard to trace. Another by-product of our asymptotic analysis of the LR score is showing that the LR score is a scalar product attack, corrected by the inverse of the covariance of the data-generating distribution. This observation connects the two primary attacks in the tracing literature, i.e. the LR score and the scalar product attack. It also provides a new covariance attack that dominates the scalar product attack. Using these two by-products of the asymptotic analysis, we provide a new MI attack (Algorithm 13) and a new canary-chosing strategy (Algorithm 14) for auditing gradient descent algorithms, in the white-box federated learning setting. The main observation is that one step of gradient descent computes the empirical mean of gradient quantities, and thus, auditing gradient descents reduces to auditing the empirical mean of gradients. We implement our two algorithms for a logistic regression model trained with FMNIST and a CNN model trained with CIFAR10. Our results show that the covariance attack improves on the scalar product attack and that the Mahalanobis leakage predicts well the hardness of the MI games. Our two algorithms can then be integrated into any auditing scheme to improve the tightness of the privacy budget estimates.

Chapter 9

Perspectives

Throughout this thesis's chapters, we have encountered many open questions worth exploring in future research.

Privacy Defintions. In Chapter 3, Proposition 3.12 shows that any (ε, δ)-View DP policy is $(\varepsilon, K^T\delta)$-Table DP. This means that the conversion from View to Table DP happens at a loss in the δ parameter. An interesting open problem is to provide an "optimal" conversion from View DP to Table DP, especially at low δ regimes.

Lower bounds for (ε, δ)-DP. In Chapter 4, all the lower bounds presented are either for ε-pure DP or ρ-zCDP. The reason behind this is that the coupling techniques of Section 4.2 are based on the group privacy property. However, the group privacy property for (ε, δ)-DP (Equation (2.5)) has extra terms corresponding to δ making hard to express this group privacy property as a non-vacuous upper bound on the $D_{\mathrm{KL}}\left(\mathcal{M}_d \parallel \mathcal{M}_{d'}\right)$. Thus, other techniques should be developed for this setting, e.g. adapting fingerprinting proofs [BUV14] to the bandit setting.

Contextual bandits under Joint DP. In Chapter 5, we only provide a private contextual linear bandit algorithm where only the rewards are considered private, while the context is supposed to be public. In the setting where both rewards and context are private, i.e. Joint DP, it is still an open problem to design a near-optimal Joint DP bandit algorithm. In particular, the best regret upper bound is known to be $O\left(d\sqrt{T}\log(T) + d^{3/4}\sqrt{T\log(1/\delta)}/\sqrt{\varepsilon}\right)$ [SS18], while the lower bound is $\Omega\left(\sqrt{dT\log(K)} + d/(\varepsilon+\delta)\right)$ [HZZ22]. To solve linear contextual bandits under JDP, [SS18] propose a variant of LinUCB [AYPS11], where the regression parameter θ is estimated privately using the tree-based mechanism. Specifically, let us write the least squares estimator at step t as $\hat{\theta}_t = V_t^{-1}u_t$, where $V_t \triangleq \lambda I_d + \sum_{s=1}^{t-1} a_s a_s^T$ and $u_t \triangleq \sum_{s=1}^{t-1} a_s r_s$. Then, since these two quantities are written as sums, [SS18] estimates the quantities V_t and u_t privately using the *tree-based mechanism*. Let us refer to the private estimations of V_t and u_t as $\tilde{V}_t$ and $\tilde{u}_t$. Thus, $\tilde{\theta}_t = \tilde{V}_t^{-1}\tilde{u}_t$ and $a_t = \arg\max_{a\in\mathcal{A}_t} \left\langle \tilde{\theta}_t, a \right\rangle + \sqrt{\tilde{\beta}_t}\|a\|_{\tilde{V}_t^{-1}}$, where $\tilde{\beta}_t$ accounts for the noise addition. [SS18] analyse the corresponding algorithm for adversarial contexts, and show that

it yields a regret upper bound of $O\left(d\sqrt{T}\log(T) + d^{3/4}\sqrt{T\log(1/\delta)}/\sqrt{\varepsilon}\right)$, where the price of JDP is non-negligible even asymptotically in T. Given the advancements in other settings, we wonder: *Is it possible to propose a JDP variant of LinUCB, such that the price of JDP is negligible for adversarial contexts?*. We postulate that the main bottleneck in the JDP variant of LinUCB proposed by [SS18] is the "sufficient statistic" method used to make the least square estimator achieve DP. For example, even for the offline batch setting of regression, [BHH+24] provides several drawbacks of using the sufficient statistics method for least squares, e.g. requiring $d^{3/2}$ samples or having errors growing with the condition number of the design matrix. They also propose a new private version of least squares called ISSP, which is near-optimal and overcomes the pitfalls of the "sufficient statistic" method. Thus, we ask: *Is it possible to propose a JDP variant of LinUCB based on the ISSP estimator with a negligible price of privacy in the regret?*. We discuss the recent progress on this problem, both from the algorithm design and lower bound techniques, and posit the open problem in [AB24c].

MI games/auditing on Z-estimators and relation to influence functions. In Chapter 6, the main technical tool used to provide an asymptotic expansion of the LR test is the "asymptotic normality" of the empirical mean, i.e. the Edgeworth expansion in Theorem 2.41. Conversely, the empirical mean estimator is only one instance of a more general class of estimators verifying the asymptotic normality property, called Z-estimators. Suppose we are interested in estimating a parameter θ that is attached (a "functional") of the distribution of observations $X_1, \dots, X_n$. A popular method to construct an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ is to maximise a criterion function of the type

$$\theta \to M_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i). \tag{9.1}$$

Here, $m_\theta : \mathcal{X} \to \bar{\mathbb{R}}$ are known functions. An estimator maximising M_n is called an Z-estimator. Often, maximising a function is found by setting a derivative to zero. Therefore, the name Z-estimator is also used for estimators satisfying a set of equations of the type

$$\Psi_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = 0. \tag{9.2}$$

The class of Z-estimators retrieves the empirical mean with $\psi_\theta(X_i) \triangleq X_i - \theta$ and the median with $\psi_\theta(X_i) \triangleq \mathrm{sign}(X_i - \theta)$. The class of Z-estimators also recovers many other estimators of interest, such as Maximum likelihood estimators, least square estimators and Empirical risk minimisation.

Under some technical condition on the data-generating distribution and the "regularity" of the function ψ_θ, it is possible to show that θ_n converges in probability to a parameter θ_0 a zero of the function $\Psi(\theta) \triangleq \mathbb{E}_X(\psi_\theta(X))$. Also, for any Z-estimator, Theorem 5.21 in [VdV00] shows

that

$$\hat{\theta}_n - \theta_0 = -V_{\theta_0}^{-1} \frac{1}{n} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{9.3}$$

where V_θ is a non-singular derivative matrix of the map $\theta \to \Psi(\theta)$ at θ_0. Generlly, $I_{\theta_0}(X_i) \triangleq V_{\theta_0}^{-1}\psi_{\theta_0}(X_i)$ is called the influence function. Thus, Equation (9.3) shows that, asymptotically, any Z-estimator can be thought of as the empirical mean of its influence functions. Using the results of the target-dependent MI game for the empirical mean mechanism, it is direct to provide a new covariance and a new canary selection strategy for Z estimators. For the score, the covariance attack becomes

$$(I_{\theta_0}(X^\star) - \theta_0)^T V_{\theta_0}^{-1}(\hat{\theta}_n - \theta_0) - \frac{1}{2n}\|I_{\theta_0}(X^\star) - \theta_0\|^2_{V_{\theta_0}^{-1}}.$$

Similarly, to choose canaries, find samples for which the estimated Mahalanobis distance of the influence functions at $X^\star$ is high, i.e. $\|I_{\theta_0}(X^\star) - \theta_0\|_{V_{\theta_0}^{-1}}$. However, we leave it for future work to provide a rigorous statement of when these statements are correct, i.e. rigorous conditions on the data-generating distribution and regularity of the ψ_θ functions.

Tight black-box auditing. In Chapter 7, we propose an attack score and a canary strategy for auditing gradient descent algorithms in the white-box federated learning setting. These two ingredients can be plugged into any state-of-the-art auditing procedure [NHS+23, SNJ23, MSS22] to improve the privacy lower bound estimates. For the white-box federated learning setting, the audit procedure applied to gradient descent algorithms trained on datasets like FMINST and CIFAR10 [NHS+23, SNJ23, MSS22] already seems to provide tight estimates of the privacy budgets. However, the auditing procedures are still not tight in the black-box setting where the auditor can only query the final trained model. The main score for MI attacks in the black-box setting are thresholds over the loss computed over the canary [CCN+22]. Specifically, if the canary is $z^\star = (x^\star, y^\star)$, then the auditor queries the final model at $z^\star$ (and augmented versions of $z^\star$ [CCN+22]). Then, the auditor computes the loss of the final model at $z^\star$, i.e. $\ell(f_{\theta_T}(x^\star), y^\star)$. The main intuition is that the loss of the model over a point that is included in the training is small. Thus the attack concludes that z^star is in when the loss $\ell(f_{\theta_T}(x^\star), y^\star)$ is smaller than some fine-tuned threshold τ. An interesting open problem is to prove optimal scores for the MI problem in the black-box setting. Are the adversaries thresholding over the loss used in the literature already optimal? If not, can we design better scores? Also, how do we design optimal canary strategies in the black-box setting?

MI games/auditing of online/sequential algorithms. In part II, all the mechanisms analysed were offline batch algorithms, which produce a one-shot output computed on a full dataset. An interesting future direction is to analyse the leakage of online and sequential algorithms by instantiating MI game threat models for these algorithms. Analysing the sequential LR test for

this setting may have multiple applications, such as auditing algorithms that are continuously updated [JWO+23] or auditing bandit algorithms and recommendation systems.

Appendix A

Supplementary for Chapter 3

Proof of Proposition 3.12

Proposition 3.12 (Relation between Table DP and View DP)**.** *For any policy π, we have that*

(*a*) *π is ε-Table DP $\Leftrightarrow$ π is ε-View DP.*

(*b*) *π is (ε, δ)-Table DP $\Rightarrow$ π is (ε, δ)-View DP.*

(*c*) *π is ρ-Table zCDP $\Rightarrow$ π ρ-View zCDP.*

(*d*) *π is (ε, δ)-View DP $\Rightarrow$ π is $(\varepsilon, K^T\delta)$-Table DP.*

(*e*) *$\Pi^{(\varepsilon,\delta)}_{Table} \subsetneq \Pi^{(\varepsilon,\delta)}_{View}$, where $\Pi^{(\varepsilon,\delta)}_{Table}$ and $\Pi^{(\varepsilon,\delta)}_{View}$ are the class of all policies verifying (ε, δ)-Table DP and (ε, δ)-View DP, respectively.*

Proof. (b): Suppose that $\mathcal{M}^\pi$ is (ε, δ)-DP.

Let $r \sim r'$ two neighbouring lists of rewards. For every event $E \in \mathcal{P}([K]^T)$, we have that

$$\mathcal{V}^\pi_{\mathbf{r}}(E) - e^\varepsilon \mathcal{V}^\pi_{r'}(E) = \mathcal{M}^\pi_{d(r)}(E) - e^\varepsilon \mathcal{M}^\pi_{d(r')}(E) \leq \delta$$

where the last inequality is because $\mathcal{M}^\pi$ is (ε, δ)-DP and $d(r) \sim d(r')$.

We conclude that $\mathcal{V}^\pi$ is (ε, δ)-DP.

(c): Suppose that $\mathcal{M}^\pi$ is ρ-zCDP.

Let $r \sim r'$ two neighbouring lists of rewards. For every $\alpha > 1$, we have that

$$D_\alpha(\mathcal{V}^\pi_r \| \mathcal{V}^\pi_{r'}) = D_\alpha(\mathcal{M}^\pi_{d(r)} \| \mathcal{M}^\pi_{d(r')}) \leq \rho\alpha$$

where the last inequality is because $\mathcal{M}^\pi$ is ρ-zCDP and $d(r) \sim d(r')$.

We conclude that $\mathcal{V}^\pi$ is ρ-zCDP.

(a) $\Rightarrow$) Is a direct consequence of (b) for $\delta = 0$.

$\Leftarrow$) Suppose that $\mathcal{V}^\pi$ is ε-DP.

Let $d \sim d'$ be two tables of rewards in $(\mathbb{R}^K)^T$.

For ε-DP, it is enough to consider atomic events $a^T \triangleq (a_1, \dots, a_T)$.

For any atomic event a^T, we have that

$$\mathcal{M}_d^\pi(a^T) = \mathcal{V}_{r(d,a^T)}^\pi(a^T) \leq e^\varepsilon \mathcal{V}_{r(d',a^T)}^\pi(a^T) = e^\varepsilon \mathcal{M}_{d'}^\pi(a^T)$$

where the first inequality is because $\mathcal{V}^\pi$ is ε-DP and $r(d, a^T) \sim r(d', a^T)$.

We conclude that $\mathcal{M}^\pi$ is ε-DP.

(d) Suppose that $\mathcal{V}^\pi$ is (ε, δ)-DP.

Let $d \sim d'$ be two tables of rewards in $(\mathbb{R}^K)^T$.

Let $E \in \mathcal{P}([K]^T)$ be an event, i.e. a set of sequences. We have that

$$\begin{aligned}
\mathcal{M}_d^\pi(E) = \sum_{a^T \in E} \mathcal{M}_d^\pi(a^T) &= \sum_{a^T \in E} \mathcal{V}_{r(d,a^T)}^\pi(a^T) \\
&\underset{(1)}{\leq} \sum_{a^T \in E} (e^\varepsilon \mathcal{V}_{r(d',a^T)}^\pi(a^T) + \delta) \\
&\underset{(2)}{\leq} e^\varepsilon \mathcal{M}_{d'}^\pi(E) + K^T \delta,
\end{aligned}$$

where (1) holds true because $\mathcal{V}^\pi$ is (ε, δ)-DP, and (2) is true because $\mathrm{card}(E) \leq K^T$.

We conclude that $\mathcal{M}^\pi$ is $(\varepsilon, K^T\delta)$-DP.

(e) To prove the strict inclusion, we build a policy π for $T = 3$, $K = 2$ with action 0 and action 1, and rewards in $\{0, 1\}$.

A policy here is a sequence of three decision rules

$$\pi = \{\pi_1, \pi_2, \pi_3\},$$

where each decision rule is a function from the history. Since the possible histories at each step are finite, specifying a decision rule is just specifying the probability weights of choosing action 0 and action 1 for every possible history.

We consider the following decision rules

$$\pi_1 = \begin{bmatrix} 2/3 & 1/3 \end{bmatrix}$$

$$\pi_2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \\ 1/4 & 3/4 \\ 1/3 & 2/3 \end{bmatrix}$$

$$\pi_3 = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \\ 1/4 & 3/4 \\ 1/5 & 4/5 \\ 1/2 & 1/2 \\ 2/3 & 1/3 \\ 1/4 & 3/4 \\ 0 & 1 \\ 1/3 & 2/3 \\ 1/7 & 6/7 \\ 3/4 & 1/4 \\ 2/5 & 3/5 \\ 1/2 & 1/2 \\ 1 & 0 \\ 1/4 & 3/4 \\ 2/3 & 1/3 \end{bmatrix}$$

The history is first represented as a binary string, and then converted to decimals. Finally, the index in the decision rule corresponding to this decimal value is chosen. We elaborate this procedure in the two examples below.

Example 1. If the policy observed the history $\{1, 0\}$, i.e. action 1 was played in the first round and the reward 0 was observed, this leads to index 2 in π_2, so the policy plays arm 0 with probability $1/4$ and arm 1 with probability $3/4$.

Example 2. If the policy observed the history $\{0, 1, 1, 1\}$, i.e. action 0 was played in the first round, the reward 1 was observed, then action 1 was played in the second round and the reward 1 was observed. This corresponds to index 7 in π_3. Thus, the policy plays arm 0 with probability 0 and arm 1 with probability 1.

Since the events and the neighbouring datasets are finite (and have a small number), it is easy to build the following two sets:

$$A = \left\{ \left(\mathcal{V}^{\pi}_{\mathbf{r}}(E), \mathcal{V}^{\pi}_{\mathbf{r}'}(E) \right), \forall E \in \mathcal{P}([2]^3), \text{ and } \forall \mathbf{r} \sim \mathbf{r}' \right\}$$

$$B = \{ (\mathcal{M}^{\pi}_{d}(E), \mathcal{M}^{\pi}_{d'}(E)), \forall E \in \mathcal{P}([2]^3), \text{ and } \forall d \sim d' \}$$

A and B represent all the probability tuples (p, q) computed on all neighbouring lists and tables of rewards, respectively, for all possible events on the sequence of actions.

Then, by checking over all the elements of A and B, it is possible to show that π is $(\varepsilon_1, \delta_1)$-View DP but never $(\varepsilon_1, \delta_1)$-Table DP for $\varepsilon_1 = 0.95$ and $\delta_1 = 0.17$. Specifically, we mean that for $\varepsilon_1 = 0.95$ and $\delta_1 = 0.17$, we obtain that $\forall (p, q) \in A,\ p \leq e^{\varepsilon_1} q + \delta_1$, while $\exists (p', q') \in B,\ p' > e^{\varepsilon_1} q' + \delta_1$. In fact, we can show that the smallest ε_0, for which π is $(\varepsilon_0, \delta_1)$-Table DP, is $\varepsilon_0 = 0.98$.

Thus, we conclude our proof with this construction. □

Appendix B

Supplementary for Chapter 4

Contents

B.1 Proof of Theorem 4.16

Theorem 4.16 (Problem-dependent Regret Lower Bound). *Let $\mathcal{E} \triangleq \mathcal{M}_1 \times \cdots \times \mathcal{M}_K$ and $\pi \in \Pi_{cons}(\mathcal{E}) \cap \Pi^{\varepsilon}$ an ε-View DP consistent policy over $\mathcal{E}$. Then, for any $\nu = (P_a : a \in K) \in \mathcal{E}$,*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\pi,\nu)}{\log(T)} \geq \sum_{a:\Delta_a>0} \frac{\Delta_a}{\min\Big(\underbrace{KL_{inf}\left(P_a,\mu^*,\mathcal{M}_a\right)}_{\text{without DP}}, \underbrace{\varepsilon\, TV_{inf}\left(P_a,\mu^*,\mathcal{M}_a\right)}_{\text{with }\varepsilon\text{-View DP}}\Big)}. \tag{B.1}$$

Proof. Let μ_a be the mean of the a-th arm in ν. We denote $d_a = \mathrm{KL}_{\inf}\left(P_a,\mu^*,\mathcal{M}_a\right)$ and $t_a = \mathrm{TV}_{\inf}\left(P_a,\mu^*,\mathcal{M}_a\right)$ for brievity. Let $\pi \in \Pi_{\text{cons}}\left(\mathcal{E}\right) \cap \Pi^{\varepsilon}$ be a consistent ε-View DP policy. Recall that $\mathrm{Reg}_T(\pi,\nu) = \sum_{a\neq a^\star} \Delta_a\, \mathbb{E}_{\nu\pi}[N_a(T)]$.

Since π is consistent, by Theorem 2.31, it holds that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_a(T)\right]}{\log(T)} \geq \frac{1}{d_a}.$$

The theorem will follow by showing, for every suboptimal arm a, that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_a(T)\right]}{\log(T)} \geq \frac{1}{\varepsilon\, t_a}$$

Fix a suboptimal arm a, and let $\alpha > 0$ be an arbitrary constant.

Step 1: Choosing the 'Hard-to-distinguish' Environment. Let $\nu' \triangleq \left(P'_j\right)_{j=1}^{K} \in \mathcal{E}$ be a bandit with $P'_j = P_j$ for $j \neq a$ and $P'_a \in \mathcal{M}_a$ be such that $\mathrm{TV}\left(P_a \parallel P'_a\right) \leq t_a + \alpha$ and $\mu\left(P'_a\right) > \mu^*$, which exists by the definition of t_a. Let $\mu' \in \mathbb{R}^K$ be the vector of means of distributions of ν'.

Step 2: From Lower Bounding Regret to Upper Bounding KL-divergence. For simplicity of notations, we use $\mathrm{Reg}_T = \mathrm{Reg}_T(\pi,\nu)$, $\mathrm{Reg}'_T = \mathrm{Reg}_T(\pi,\nu)$, and $A = \{(a_1,a_2,\ldots,a_T) : \mathrm{card}(\{j : a_j = 1\}) \leq T/2\}$.

Then, by regret decomposition and Markov Inequality, we obtain

$$\begin{aligned}
\mathrm{Reg}_T + \mathrm{Reg}'_T &\geq \frac{T}{2}\left(M_{\nu\pi}(A)\Delta_a + M_{\nu'\pi}\left(A^c\right)\left(\mu'_a - \mu^*\right)\right) \\
&\geq \frac{T}{2}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\left(M_{\nu\pi}(A) + M_{\nu'\pi}\left(A^c\right)\right) \\
&\geq \frac{T}{4}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\exp(-D_{\mathrm{KL}}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right))
\end{aligned} \tag{B.2}$$

Step 3: KL-divergence Decomposition with ε-View DP. By Theorem 4.9 and the construction of the 'hard-to-distinguish' environments, we obtain

$$\begin{aligned} D_{\mathrm{KL}}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right) &\leq \varepsilon \mathbb{E}_{\nu\pi}\left[N_a(T)\right] \mathrm{TV}\left(P_a \parallel P_a'\right) \\ &\leq \varepsilon \mathbb{E}_{\nu\pi}\left[N_a(T)\right]\left(t_a + \alpha\right) \end{aligned}$$

Step 4: Rearranging and taking the limit inferior. Thus, we get

$$\mathrm{Reg}_T + \mathrm{Reg}_T' \geq \frac{T}{4} \min\left\{\Delta_a, \mu_a' - \mu^*\right\} \exp\left(-\varepsilon \mathbb{E}_{\nu\pi}\left[N_a(T)\right]\left(t_a + \alpha\right)\right)$$

Now, taking the limit inferior on both sides leads to

$$\begin{aligned} \liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_a(T)\right]}{\log(T)} &\geq \frac{1}{\varepsilon\left(t_a + \alpha\right)} \liminf_{T\to\infty} \frac{\log\left(\frac{T \min\{\Delta_a, \mu_a' - \mu^*\}}{4\left(\mathrm{Reg}_T + \mathrm{Reg}_T'\right)}\right)}{\log(T)} \\ &= \frac{1}{\varepsilon\left(t_a + \alpha\right)} \left(1 - \limsup_{T\to\infty} \frac{\log\left(\mathrm{Reg}_T + \mathrm{Reg}_T'\right)}{\log(T)}\right) = \frac{1}{\varepsilon\left(t_a + \alpha\right)}. \end{aligned}$$

The last equality follows from the definition of consistency, which says that for any $p > 0$, there exists a constant C_p such that for sufficiently large T, $\mathrm{Reg}_T + \mathrm{Reg}_T' \leq C_p T^p$. This property implies that

$$\limsup_{T\to\infty} \frac{\log\left(\mathrm{Reg}_T + \mathrm{Reg}_T'\right)}{\log(T)} \leq \limsup_{T\to\infty} \frac{p\log(T) + \log\left(C_p\right)}{\log(T)} = p,$$

which gives the result since $p > 0$ was an arbitrary constant.

We arrive at the claimed result by taking the limit as α tends to zero.

□

B.2 Proof of Theorem 4.17

Theorem 4.17 (Minimax regret lower bound). *Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = \mathbb{R}^d$. Let π be an ε-View DP policy. Then, there exists a vector $\theta \in \Theta$ such that*

$$\mathrm{Reg}_T(\pi, \mathcal{A}, \theta) \geq \max\left\{\underbrace{\frac{\exp(-2)}{8} d\sqrt{T}}_{\textit{without DP}}, \underbrace{\frac{\exp(-1)}{4} \frac{d}{\varepsilon}}_{\textit{with } \varepsilon\textit{-View DP}}\right\}. \tag{B.3}$$

Proof. Due to Theorem 24.1 in [LS20], it holds that,

$$\mathrm{Reg}_T^{\text{minimax}}(\mathcal{A}, \Theta) \geq \exp(-2)\frac{d}{8}\sqrt{T}.$$

Now, we focus on proving the ε-View DP part of the lower bound.

Let $\Theta = \left\{-\frac{1}{\varepsilon T}, \frac{1}{\varepsilon T}\right\}^d$. For $\theta, \theta' \in \Theta$, let ν and ν' be the bandit instances corresponding resp. to θ and θ'. We denote $\mathbb{M}_\theta = \mathbb{M}_{\nu,\pi}$ and $\mathbb{M}_{\theta'} = \mathbb{M}_{\nu',\pi}$. Let $\mathbb{E}_\theta$ and $\mathbb{E}_{\theta'}$ the expectations under $\mathbb{M}_\theta$ and $\mathbb{M}_{\theta'}$ respectively.

Step 1: From Lower Bounding Regret to Upper Bounding KL-divergence We begin with

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{A}, \theta) &= \mathbb{E}_\theta\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\mathrm{sign}\left(\theta_i\right) - A_{ti}\right)\theta_i\right] \\
&\geq \frac{1}{\varepsilon T}\sum_{i=1}^{d}\mathbb{E}_\theta\left[\sum_{t=1}^{T}\mathbb{I}\left\{\mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right)\right\}\right] \\
&\geq \frac{1}{\varepsilon}\sum_{i=1}^{d}\mathbb{M}_\theta\left(\sum_{t=1}^{T}\mathbb{I}\left\{\mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right)\right\} \geq T/2\right)
\end{aligned}$$

In this derivation, the first equality holds because the optimal action satisfies $a_i^* = \mathrm{sign}\left(\theta_i\right)$ for $i \in [d]$. The first inequality follows from an observation that

$$\left(\mathrm{sign}\left(\theta_i\right) - A_{ti}\right)\theta_i \geq \left|\theta_i\right|\mathbb{I}\left\{\mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right)\right\}$$

. The last inequality is a direct application of Markov's inequality.

For $i \in [d]$ and $\theta \in \Theta$, we define

$$p_{\theta,i} \triangleq \mathbb{M}_\theta\left(\sum_{t=1}^{T}\mathbb{I}\left\{\mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right)\right\} \geq T/2\right).$$

Now, let $i \in [d]$ and $\theta \in \Theta$ be fixed. Also, let $\theta'_j = \theta_j$ for $j \neq i$ and $\theta'_i = -\theta_i$. Then, by the Bretagnolle-Huber inequality,

$$p_{\theta,i} + p_{\theta',i} \geq \frac{1}{2}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right)\right).$$

Step 2: KL-divergence Decomposition with ε-View DP. From Theorem 4.9, we obtain that

$$D_{\mathrm{KL}}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right) \leq \varepsilon\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\mathrm{TV}\left(\mathcal{N}\left(\left\langle A_t, \theta\right\rangle, 1\right) \parallel \mathcal{N}\left(\left\langle A_t, \theta'\right\rangle, 1\right)\right)\right]$$

$$
\begin{aligned}
&\leq \varepsilon \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathcal{N}\left(\langle A_t,\theta\rangle,1\right)\,\|\,\mathcal{N}\left(\langle A_t,\theta'\rangle,1\right)\right)}\right]\\
&= \varepsilon \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\sqrt{\frac{1}{4}\left[\langle A_t,\theta-\theta'\rangle^2\right]}\right]\\
&= \frac{1}{2}\varepsilon \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\left|\langle A_t,\theta-\theta'\rangle\right|\right] \qquad \text{(B.4)}\\
&= \frac{1}{2}\varepsilon \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\left|A_{t,i}\right|\left(2\left|\theta_i\right|\right)\right]\\
&\leq \frac{1}{2}\varepsilon \mathbb{E}_{\nu\pi}\left[T\times 2\frac{1}{\varepsilon T}\right]=1 \qquad \text{(B.5)}
\end{aligned}
$$

Here, the second inequality is a consequence of Pinsker's inequality. The last inequality holds true because $A_t \in [-1,1]^d$ and $\theta,\theta' \in \left\{-\frac{1}{\varepsilon T},\frac{1}{\varepsilon T}\right\}^d$

Step 3: Choosing the 'Hard-to-distinguish' θ. We already have that

$$p_{\theta,i}+p_{\theta',i}\geq\frac{1}{2}\exp\left(-1\right)$$

Now, we apply an 'averaging hammer' over all $\theta\in\Theta$, such that $|\Theta|=2^d$, to obtain

$$\sum_{\theta\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^{d}p_{\theta,i}=\frac{1}{|\Theta|}\sum_{i=1}^{d}\sum_{\theta\in\Theta}p_{\theta,i}\geq\frac{d}{4}\exp(-1).$$

This implies that there exists a $\theta\in\Theta$ such that $\sum_{i=1}^{d}p_{\theta,i}\geq d\exp(-1)/4$.

Step 4: Plugging Back θ in the Regret Decomposition. With this choice of θ, we conclude

$$
\begin{aligned}
\mathrm{Reg}_T(\mathcal{A},\theta)&\geq\frac{1}{\varepsilon}\sum_{i=1}^{d}p_{\theta,i}\\
&\geq\frac{\exp(-1)}{4}\frac{d}{\varepsilon}
\end{aligned}
$$

□

B.3 Proof of Theorem 4.18

Theorem 4.18 (Problem-dependent regret lower bound)**.** *Let $\mathcal{A}\subset\mathbb{R}^d$ be a finite set spanning $\mathbb{R}^d$ and $\theta\in\mathbb{R}^d$ be such that there is a unique optimal action. Then, for any consistent and ε-View DP policy*

π satisfies

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \mathcal{A}, \theta)}{\log(T)} \geq c(\mathcal{A}, \theta), \tag{B.6}$$

where the structural distinguishability gap *is the solution of a constraint optimisation*

$$c(\mathcal{A}, \theta) \triangleq \inf_{\alpha \in [0,\infty)^{\mathcal{A}}} \sum_{a\in\mathcal{A}} \alpha(a)\Delta_a, \text{ such that } \|a\|^2_{H_\alpha^{-1}} \leq \min\left\{ \underbrace{0.5\Delta_a^2}_{\text{without DP}} , \underbrace{0.5\varepsilon\rho_a(\mathcal{A})\Delta_a}_{\text{with } \varepsilon\text{-View DP}} \right\}$$

for all $a \in \mathcal{A}$ with $\Delta_a > 0$, $H_\alpha = \sum_{a\in\mathcal{A}} \alpha(a)aa^\top$, and a structure dependent constant $\rho_a(\mathcal{A})$.

Proof. Let $a^* = \mathrm{argmax}_{a\in\mathcal{A}}\langle a, \theta\rangle$ be the optimal action, which we assumed to be unique.

By Theorem 25.1, [LS20],

$$\limsup_{T\to\infty} \log(T)\, \|a - a^*\|^2_{\bar{G}_T^{-1}} \leq \frac{1}{2}\Delta_a^2. \tag{B.7}$$

Let $\mathbb{M}$ and $\mathbb{M}'$ be the measures on the sequence of outcomes $A_1, \ldots, A_T$ induced by θ and θ' respectively. Let $\mathbb{E}[\cdot]$ and $\mathbb{E}'[\cdot]$ be the expectation operators of $\mathbb{M}$ and $\mathbb{M}'$, respectively.

Step 1: Choosing the 'Hard to distinguish' θ'. Let $\theta' \in \mathbb{R}^d$ be an alternative parameter to be chosen subsequently. We follow the usual plan of choosing θ' to be close to θ, but also ensuring that the optimal action in the bandit determined by θ' is not a^*. Let $\Delta_{\min} = \min\{\Delta_a : a \in \mathcal{A}, \Delta_a > 0\}$, $\alpha \in (0, \Delta_{\min})$ and H be a positive definite matrix (to be chosen later) such that $\|a - a^*\|^2_H > 0$.

Given this setting, we define

$$\theta' \triangleq \theta + \frac{\Delta_a + \alpha}{\|a - a^*\|^2_H} H\,(a - a^*)\,,$$

which is chosen such that $\langle a - a^*, \theta'\rangle = \langle a - a^*, \theta\rangle + \Delta_a + \alpha = \alpha$.

This means that a^* is α-suboptimal for the environment corresponding to θ'.

Step 2: From Lower Bounding Regret to Upper Bounding KL-divergence. For simiplicity, we abbreviate $\mathrm{Reg}_T = \mathrm{Reg}_T(\mathcal{A}, \theta)$ and $\mathrm{Reg}'_T = \mathrm{Reg}_T(\mathcal{A}, \theta')$.

Then, by applying the classic regret decomposition and Markov's inequality, we obtain

$$\mathrm{Reg}_T = \mathbb{E}\left[\sum_{a\in\mathcal{A}} N_a(T)\Delta_a\right] \geq \frac{T\Delta_{\min}}{2}\mathbb{M}\left(N_{a^*}(T) < T/2\right) \geq \frac{T\alpha}{2}\mathbb{M}\left(N_{a^*}(T) < T/2\right),$$

Since a^* is α-suboptimal in bandit θ', it implies that

$$\mathrm{Reg}'_T \geq \frac{T\alpha}{2}\mathbb{M}'\left(N_{a^*}(T) \geq T/2\right).$$

Now, Bretagnolle Huber inequality implies that

$$\begin{aligned}\mathrm{Reg}_T + \mathrm{Reg}'_T &\geq \frac{T\alpha}{2}\left(\mathbb{M}\left(N_{a^*}(T) < T/2\right) + \mathbb{M}'\left(N_{a^*}(T) \geq T/2\right)\right)\\ &\geq \frac{T\alpha}{4}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{M} \parallel \mathbb{M}'\right)\right)\end{aligned}$$

Step 3: KL-divergence Decomposition with ε-View DP. By Equation B.4, we have that

$$\begin{aligned}D_{\mathrm{KL}}\left(\mathbb{M} \parallel \mathbb{M}\right) &\leq \frac{1}{2}\varepsilon\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\left|\langle A_t, \theta - \theta'\rangle\right|\right]\\ &= \frac{1}{2}\varepsilon\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\left|\left\langle A_t, \frac{\Delta_a + \alpha}{\|a - a^*\|_H^2}H\left(a - a^*\right)\right\rangle\right|\right]\\ &= \frac{1}{2}\varepsilon\frac{\Delta_a + \alpha}{\|a - a^*\|_{\bar{G}_T^{-1}}^2}\rho_T(H),\end{aligned}$$

where we define

$$\rho_T(H) \triangleq \frac{\|a - a^*\|_{\bar{G}_T^{-1}}^2}{\|a - a^*\|_H^2}\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T}\left|\langle A_t, H\left(a - a^*\right)\rangle\right|\right]$$

Thus, after re-arrangement, we get

$$\frac{\varepsilon\left(\Delta_a + \alpha\right)}{2\log(T)\|a - a^*\|_{\bar{G}_T^{-1}}^2}\rho_T(H) \geq 1 - \frac{\log\left(\left(4R_T + 4R'_T\right)/\alpha\right)}{\log(T)}. \tag{B.8}$$

Step 4: Choosing H and Taking the Limit. The definition of consistency means that Reg_T and Reg'_T are both sub-linear in T. This implies that the second term in Equation (B.8) tends to zero for large T. Thus, by tending T to ∞ and α to zero, we obtain

$$\liminf_{T\to\infty}\frac{\rho_T(H)}{\log(T)\|a - a^*\|_{\bar{G}_T^{-1}}^2} \geq \frac{2}{\varepsilon\Delta_a}.$$

We now choose H to be a cluster point of the sequence $\left(\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|\right)_{T\in S}$ where $\left\|\bar{G}_T^{-1}\right\|$ is the spectral norm of the matrix $\bar{G}_T^{-1}$.

Lemma B.1: For this choice of H,

$$\liminf_{T\to\infty}\rho_T(H) \leq \rho_a(\mathcal{A}),$$

where

$$\rho_a(\mathcal{A}) \triangleq \sum_{j=1,\|a_j\|\neq 0}^{K} \frac{\left|a_j^T(a-a^*)\right|}{\|a_j\|^2}.$$

Finally,

$$\limsup_{T\to\infty} \log(T)\, \|a-a^*\|^2_{\bar{G}_T^{-1}} \leq \frac{1}{2}\varepsilon\Delta_a\rho_a(\mathcal{A}).$$

Combined with Equation B.7, we get that

$$\limsup_{T\to\infty} \log(T)\, \|a-a^*\|^2_{\bar{G}_T^{-1}} \leq \min\left(\frac{1}{2}\Delta_a^2, \frac{1}{2}\varepsilon\Delta_a\rho_a(\mathcal{A})\right).$$

Using that

$$\lim_{T\to\infty} \frac{\|a-a^*\|_{\bar{G}_T^{-1}}}{\|a\|_{\bar{G}_T^{-1}}} = 1$$

from Theorem 25.1, [LS20], we get that

$$\limsup_{T\to\infty} \log(T)\, \|a\|^2_{\bar{G}_T^{-1}} \leq \min\left(\frac{1}{2}\Delta_a^2, 3\varepsilon\Delta_a\rho_a(\mathcal{A})\right).$$

Step 5: Getting Back to the Regret. We conclude using the same steps as in the Corollary 2 [LS17]. □

Now, we prove Lemma B.1.

Lemma B.1. *If H is a cluster point of the sequence $\left(\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|\right)_{T\in S}$ and $\left\|\bar{G}_T^{-1}\right\|$ is the spectral norm of the matrix $\bar{G}_T^{-1}$, then the following inequality holds true:*

$$\liminf_{T\to\infty} \rho_T(H) \leq \rho_a(\mathcal{A}),$$

where

$$\rho_a(\mathcal{A}) \triangleq \sum_{j=1,\|a_j\|\neq 0}^{K} \frac{\left|a_j^T(a-a^*)\right|}{\|a_j\|^2}.$$

Proof. We let S be a subset so that $\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|$ converges to H on $T\in S$. Then,

$$\begin{aligned}
\liminf_{T\to\infty} \rho_T(H) &\leq \liminf_{T\in S} \rho_T(\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|) \\
&= \liminf_{T\in S} \mathbb{E}_\theta\left[\sum_{t=1}^{T}\left|\left\langle A_t, \bar{G}_T^{-1}(a-a^*)\right\rangle\right|\right] \\
&= \liminf_{T\in S} \sum_{j=1}^{K} \mathbb{E}_\theta(N_j(T))\left|a_j^T\bar{G}_T^{-1}(a-a^*)\right|
\end{aligned}$$

$$= \liminf_{T \in S} \sum_{j=1, \|a_j\| \neq 0}^{K} \mathbb{E}_\theta(N_j(T)) \left| a_j^T \bar{G}_T^{-1} (a - a^*) \right|$$

Let j be such that $\|a_j\| \neq 0$. Now, we aim to upper bound the term $\left| a_j^T \bar{G}_T^{-1} (a - a^*) \right|$

First, we decompose $a - a^*$ into two orthogonal components, which are aligned and orthogonal to a_j respectively.

$$a - a^* = \alpha_j a_j + b_j,$$

where $a_j^\top b_j = 0$ and $\alpha_j = \frac{a_j^T (a - a^*)}{\|a_j\|^2}$.

On the other hand, we have that

$$\bar{G}_T = \mathbb{E}_\theta \left[\sum_{t=1}^{T} A_t A_t^\top \right] = \sum_{j=1}^{K} \mathbb{E}_\theta(N_j(T)) a_j a_j^\top \succeq \mathbb{E}_\theta(N_j(T)) a_j a_j^\top$$

Since

$$\left(\mathbb{E}_\theta(N_j(T)) a_j a_j^\top \right)^\dagger = \frac{1}{\mathbb{E}_\theta(N_j(T))(a_j^\top a_j)^2} a_j a_j^\top,$$

and

$$\left(\mathbb{E}_\theta(N_j(T)) a_j a_j^\top \right)^\dagger b_j = 0,$$

only the component of $a - a^*$ in the direction of a_j matters in the dot product $a_j^T \bar{G}_T^{-1} (a - a^*)$. Thus,

$$\begin{aligned} \left| a_j^T \bar{G}_T^{-1} (a - a^*) \right| &\leq \frac{|\alpha_j|}{\mathbb{E}_\theta(N_j(T))(a_j^\top a_j)^2} a_j^T a_j a_j^T a_j \\ &= \frac{|\alpha_j|}{\mathbb{E}_\theta(N_j(T))} \end{aligned}$$

Consequently,

$$\liminf_{T \to \infty} \rho_T(H) \leq \sum_{j=1, \|a_j\| \neq 0}^{K} \frac{\left| a_j^T (a - a^*) \right|}{\|a_j\|^2} \triangleq \rho_a(\mathcal{A})$$

□

Example B.2 ($\rho_a(\mathcal{A})$ for an orthogonal set of arms)**.** *If the action space is the orthogonal basis, then* $\rho_a(\mathcal{A}) = 2$*, because:*

$$\bar{G}_T = \begin{bmatrix} \mathbb{E}(N_1(T)) & & \\ & \ddots & \\ & & \mathbb{E}(N_d(T)) \end{bmatrix}$$

and:

$$\left|\left\langle A_t, \bar{G}_T^{-1}\left(a - a^*\right)\right\rangle\right| = \frac{1}{\mathbb{E}(N_a(T))}\mathbb{I}_{A_t=a} + \frac{1}{\mathbb{E}(N_{a^\star}(T))}\mathbb{I}_{A_t=a^\star}$$

so:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left|\left\langle A_t, \bar{G}_T^{-1}\left(a - a^*\right)\right\rangle\right|\right] = 2$$

B.4 Proof of Theorem 4.20

Theorem 4.20 (Minimax lower bound for finite-armed bandits)**.** *For any* $K > 1$*,* $T \geq K - 1$*, and* $0 < \rho \leq 1$*,*

$$\begin{aligned}\mathrm{Reg}^{\star}_{T,\rho}(\mathcal{E}^K_{\mathcal{G}}) &\triangleq \inf_{\pi \in \Pi^{\rho}_{Int}} \sup_{\nu \in \mathcal{E}^K_{\mathcal{G}}} \mathrm{Reg}_T(\pi, \nu) \\ &\geq \max\left\{\underbrace{\frac{1}{27}\sqrt{T(K-1)}}_{\text{without DP}}, \quad \underbrace{\frac{1}{124}\sqrt{\frac{K-1}{\rho}}}_{\text{with } \rho\text{-Interactive zCDP}}\right\}.\end{aligned}$$

Proof. The non-private part of the lower bound is due to Theorem 15.2 in [LS20]. To prove the private part of the lower bound, we plug our KL decomposition theorem into the proofs of regret lower bounds for bandits.

Step 1: Choosing the 'hard-to-distinguish' environments. First, we fix a ρ-zCDP policy π . Let Δ be a constant (to be specified later), and ν be a Gaussian bandit instance with unit variance and mean vector $\mu = (\Delta, 0, 0, ..., 0)$.

To choose the second bandit instance, let $a \triangleq \arg\min_{i \in [2,K]} \mathbb{E}_{\nu,\pi}[N_i(T)]$ be the least played arm in expectation other than the optimal arm 1. The second environment ν' is then chosen to be a Gaussian bandit instance with unit variance and mean vector $\mu' = (\Delta, 0, 0, \ldots 0, 2\Delta, 0 \ldots, 0)$, where $\mu'_j = \mu_j$ for every j except for $\mu'_a = 2\Delta$.

The first arm is optimal in ν and the arm i is optimal in ν'.

Since $T = \mathbb{E}_{\nu\pi}[N_1(T)] + \sum_{i>1}\mathbb{E}_{\nu\pi}[N_i(T)] \geq (K-1)\mathbb{E}_{\nu\pi}[N_a(T)]$, we observe that

$$n_a \triangleq \mathbb{E}_{\nu\pi}[N_a(T)] \leq \frac{T}{K-1}$$

Step 2: From lower bounding regret to upper bounding KL-divergence. Now by the classic regret decomposition and Markov inequality, we get

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu) &= (T - \mathbb{E}_{\nu\pi}[N_1(T)])\,\Delta \\ &\geq \mathbb{M}_{\nu\pi}(N_1(T) \leq T/2)\,\frac{T\Delta}{2},\end{aligned}$$

and

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu') &= \Delta\mathbb{E}_{\nu'\pi}[N_1(T)] + \sum_{a\notin\{1,i\}} 2\Delta\mathbb{E}_{\nu'\pi}[N_a(T)] \\ &\geq \mathbb{M}_{\nu'\pi}(N_1(T) > T/2)\,\frac{T\Delta}{2}.\end{aligned}$$

Let us define the event $A \triangleq \{N_1(T) \leq T/2\} = \{(a_1, a_2, \ldots, a_T) : \mathrm{card}(\{j : a_j = 1\}) \leq T/2\}$.

By applying the Bretagnolle–Huber inequality, we have:

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu) + \mathrm{Reg}_T(\pi,\nu') &\geq \frac{T\Delta}{2}(M_{\nu\pi}(A) + M_{\nu'\pi}(A^c)) \\ &\geq \frac{T\Delta}{4}\exp(-D_{\mathrm{KL}}(M_{\nu\pi} \parallel M_{\nu'\pi}))\end{aligned}$$

Step 3: KL-divergence decomposition with ρ-Interactive zCDP. Since ν and ν' only differ in arm a, we get that $\sum t_{a_t} = t_a \sum \mathbb{1}(a_t = a)$, where $t_a \triangleq \mathrm{TV}(\nu_a \parallel \nu'_a)$.

Now, applying Theorem 4.14 gives

$$\begin{aligned}D_{\mathrm{KL}}(M_{\nu\pi} \parallel M_{\nu'\pi}) &\leq \rho(n_a^2 t_a^2 + n_a t_a(1-t_a) + t_a^2\,\mathbb{V}_{\nu\pi}(N_a(T))) \\ &\leq \rho(n_a^2 t_a^2 + n_a t_a + t_a^2\,\mathbb{V}_{\nu\pi}(N_a(T)))\,.\end{aligned}$$

where the last inequality is due to the fact that $1 - t_a \leq 1$.

On the other hand, we have the following upper bounds,

$$n_a \leq \frac{T}{K-1}$$

and

$$\mathbb{V}_{\nu\pi}(N_a(T)) \leq \mathbb{E}_{\nu\pi}[N_a(T)]\,(T - \mathbb{E}_{\nu\pi}[N_a(T)]) \leq \frac{T^2}{K-1}$$

and finally, using Pinsker's Inequality

$$t_a = \mathrm{TV}\left(\nu_a \,\|\, \nu'_a\right) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathcal{N}(0,1) \,\|\, \mathcal{N}(2\Delta, 1)\right)} = \Delta$$

Step 4: Choosing the worst Δ. Plugging back in the regret expression, we find

$$\mathrm{Reg}_T(\pi, \nu) + \mathrm{Reg}_T(\pi, \nu') \geq \frac{T\Delta}{4} \exp\left(-\rho\left[\frac{T^2}{K-1}\left(1 + \frac{1}{K-1}\right)\Delta^2 + \frac{T}{K-1}\Delta\right]\right)$$

Let $\alpha \triangleq \frac{T}{4}$, $\beta \triangleq \frac{\rho T^2}{K-1}(1 + \frac{1}{K-1})$ and $\gamma \triangleq \frac{\rho T}{K-1}$.

We have then

$$\begin{aligned}\mathrm{Reg}_T(\pi, \nu) + \mathrm{Reg}_T(\pi, \nu') &\geq \alpha\Delta \exp\left(-\beta\Delta^2 - \gamma\Delta\right) \\ &\geq \alpha\Delta \exp\left(-\beta\left(\Delta + \frac{\gamma}{2\beta}\right)^2\right)\end{aligned}$$

By optimising for Δ, we choose $\Delta = \frac{1}{\sqrt{\beta}} - \frac{\gamma}{2\beta}$.

Putting back in Δ we have

$$\begin{aligned}\Delta &= \frac{1}{\sqrt{\beta}} - \frac{\gamma}{2\beta} \\ &= \sqrt{\frac{K-1}{\rho T^2\left(1 + \frac{1}{K-1}\right)}} - \frac{1}{2T\left(1 + \frac{1}{K-1}\right)} \\ &\geq \sqrt{\frac{K-1}{2\rho T^2}} - \frac{1}{2T} = \frac{\sqrt{K-1}}{T}\left(\frac{1}{\sqrt{2\rho}} - \frac{1}{2\sqrt{K-1}}\right) \\ &\geq \frac{\sqrt{K-1}}{T}\left(\frac{1}{\sqrt{2\rho}} - \frac{1}{2}\right) \\ &\geq \frac{\sqrt{K-1}}{T}\left(\frac{1}{4\sqrt{2\rho}}\right)\end{aligned}$$

where all the inequalities use that $K \geq 2$ and $\rho \leq 1$.

This gives that

$$\mathrm{Reg}_T(\pi, \nu) + \mathrm{Reg}_T(\pi, \nu') \geq \frac{\sqrt{K-1}}{4}\left(\frac{1}{4\sqrt{2\rho}}\right)\exp\left(-1\right)$$

We conclude the proof by using $\frac{1}{16\sqrt{2}}\exp(-1) \geq \frac{1}{62}$, and using $2\max(a,b) \geq a + b$. □

B.5 Proof of Theorem 4.21

Theorem 4.21 (Minimax Lower Bounds for Linear Bandits). *Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = \mathbb{R}^d$. Then, we have that*

$$\begin{aligned}\mathrm{Reg}^{\star}_{T,\rho}(\mathcal{A}, \Theta) &\triangleq \inf_{\pi \in \Pi^{\rho}_{Int}} \sup_{\theta \in \Theta} \mathrm{Reg}_T(\pi, \mathcal{A}, \theta) \\ &\geq \max\left\{ \underbrace{\frac{e^{-2}}{8} d\sqrt{T}}_{\text{without DP}}, \quad \underbrace{\frac{e^{-2.25}}{4} \frac{d}{\sqrt{\rho}}}_{\text{with } \rho\text{-Interactive zCDP}} \right\}\end{aligned}$$

Proof. For the non-private lower bound, Theorem 24.1 of [LS20] gives that,

$$\mathrm{Reg}_T^{\text{minimax}}(\mathcal{A}, \Theta) \geq \exp(-2) \frac{d}{8} \sqrt{T}.$$

Now, we focus on proving the ρ-zCDP part of the lower bound.

Let $\Theta = \left\{ -\frac{1}{T\sqrt{\rho}}, \frac{1}{T\sqrt{\rho}} \right\}^d$. For $\theta, \theta' \in \Theta$, let ν and ν' be the bandit instances corresponding resp. to θ and θ'. We denote $\mathbb{M}_\theta = \mathbb{M}_{\nu,\pi}$ and $\mathbb{M}_{\theta'} = \mathbb{M}_{\nu',\pi}$. Let $\mathbb{E}_\theta$ and $\mathbb{E}_{\theta'}$ the expectations under $\mathbb{M}_\theta$ and $\mathbb{M}_{\theta'}$ respectively.

Step 1: From lower bounding regret to upper bounding KL-divergence. We begin with

$$\begin{aligned}\mathrm{Reg}_T(\mathcal{A}, \theta) &= \mathbb{E}_\theta \left[\sum_{t=1}^{T} \sum_{i=1}^{d} \left(\mathrm{sign}\left(\theta_i\right) - A_{ti} \right) \theta_i \right] \\ &\geq \frac{1}{T\sqrt{\rho}} \sum_{i=1}^{d} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\} \right] \\ &\geq \frac{1}{\sqrt{\rho}} \sum_{i=1}^{d} \mathbb{M}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\} \geq T/2 \right)\end{aligned}$$

In this derivation, the first equality holds because the optimal action satisfies $a_i^* = \mathrm{sign}\left(\theta_i\right)$ for $i \in [d]$. The first inequality follows from an observation that

$$\left(\mathrm{sign}\left(\theta_i\right) - A_{ti} \right) \theta_i \geq \left| \theta_i \right| \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\}.$$

The last inequality is a direct application of Markov's inequality.

For $i \in [d]$ and $\theta \in \Theta$, we define

$$p_{\theta,i} \triangleq \mathbb{M}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\} \geq T/2 \right).$$

Now, let $i \in [d]$ and $\theta \in \Theta$ be fixed. Also, let $\theta'_j = \theta_j$ for $j \neq i$ and $\theta'_i = -\theta_i$. Then, by the Bretagnolle-Huber inequality,

$$p_{\theta,i} + p_{\theta',i} \geq \frac{1}{2} \exp\left(-D_{\mathrm{KL}}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right)\right).$$

Step 2: KL-divergence decomposition with ρ-Interactive zCDP.

Define $p_t \triangleq \mathrm{TV}\left(\mathcal{N}\left(\langle A_t, \theta\rangle, 1\right) \parallel \mathcal{N}\left(\langle A_t, \theta'\rangle, 1\right)\right)$.

From Lemma 4.14, we obtain that

$$D_{\mathrm{KL}}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right) \leq \rho\left(\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right]\right)^2 + \rho\left(\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right]\right) + \rho\,\mathbb{V}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right]$$

On the other hand, using Pinsker's inequality, we have that

$$\begin{aligned}
\sum_{t=1}^{T} p_t &\leq \sum_{t=1}^{T} \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathcal{N}\left(\langle A_t, \theta\rangle, 1\right) \parallel \mathcal{N}\left(\langle A_t, \theta'\rangle, 1\right)\right)} \\
&\leq \sum_{t=1}^{T} \sqrt{\frac{1}{4}\left[\langle A_t, \theta - \theta'\rangle^2\right]} \\
&\leq \frac{1}{2}\left[\sum_{t=1}^{T} \left|\langle A_t, \theta - \theta'\rangle\right|\right] \\
&\leq \frac{1}{2}\left[\sum_{t=1}^{T} \left|A_{t,i}\right| \left(2\left|\theta_i\right|\right)\right] \\
&\leq \frac{1}{2}\left[T \times 2\frac{1}{T\sqrt{\rho}}\right] = \frac{1}{\sqrt{\rho}}.
\end{aligned}$$

The last inequality holds true because $A_t \in [-1, 1]^d$ and $\theta, \theta' \in \left\{-\frac{1}{T\sqrt{\rho}}, \frac{1}{T\sqrt{\rho}}\right\}^d$.

This gives that

$$\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right] \leq \frac{1}{\sqrt{\rho}} \quad \text{and} \quad \mathbb{V}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right] \leq \frac{1}{4\rho}$$

Plugging back in the KL decomposition, we get that,

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right) &\leq \rho\left(\frac{1}{\sqrt{\rho}}\right)^2 + \rho\left(\frac{1}{\sqrt{\rho}}\right) + \rho\left(\frac{1}{4\rho}\right) \\
&= 1 + \sqrt{\rho} + \frac{1}{4} \leq \frac{9}{4}
\end{aligned}$$

where the last inequality is due to $\rho \leq 1$.

Step 3: Choosing the 'hard-to-distinguish' θ. Now, we have that

$$p_{\theta,i} + p_{\theta',i} \geq \frac{1}{2}\exp\left(-9/4\right)$$

Now, we apply an 'averaging hammer' over all $\theta \in \Theta$, such that $|\Theta| = 2^d$, to obtain

$$\sum_{\theta\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^{d} p_{\theta,i} = \frac{1}{|\Theta|}\sum_{i=1}^{d}\sum_{\theta\in\Theta} p_{\theta,i} \geq \frac{d}{4}\exp(-\frac{9}{4}).$$

This implies that there exists a $\theta \in \Theta$ such that $\sum_{i=1}^{d} p_{\theta,i} \geq d\exp(-\frac{9}{4})/4$.

Step 4: Plugging back θ in the regret decomposition. With this choice of θ, we conclude that

$$\begin{aligned}\mathrm{Reg}_T(\mathcal{A},\theta) &\geq \frac{1}{\sqrt{\rho}}\sum_{i=1}^{d} p_{\theta,i}\\ &\geq \frac{\exp(-\frac{9}{4})}{4}\frac{d}{\sqrt{\rho}}\end{aligned}$$

□

B.6 Proof of Proposition 4.26

Proposition 4.26 (TV characteristic time for Bernoulli instances)**.** *Let ν be a bandit instance, i.e. such that $\nu_a = Bernoulli(\mu_a)$ and $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Let $\Delta_a \triangleq \mu_1 - \mu_a$ and $\Delta_{min} \triangleq \min_{a\neq 1}\Delta_a$. We have that*

$$T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) = \frac{1}{\Delta_{min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}, \qquad \textit{and} \qquad \frac{1}{\Delta_{min}} \leq T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) \leq \frac{K}{\Delta_{min}}.$$

Proof. **Step 1:** Let ν be a bandit instance, i.e. such that $\nu_a \triangleq$ Bernoulli(μ_a) and $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$.

For the alternative bandit instance $\boldsymbol{\lambda}$, we refer to the mean of arm a as ρ_a, i.e. $\lambda_a \triangleq$ Bernoulli(ρ_a).

By the definition of $T^\star_{\mathbf{TV}}$, we have that

$$\begin{aligned}(T^\star_{\mathbf{TV}}(\boldsymbol{\nu}))^{-1} &= \sup_{\omega\in\Sigma_K}\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})}\sum_{a=1}^{K}\omega_a \mathrm{TV}\left(\nu_a \,\|\, \lambda_a\right)\\ &\overset{(a)}{=} \sup_{\omega\in\Sigma_K}\min_{a\neq 1}\inf_{\boldsymbol{\lambda}:\rho_a>\rho_1}\omega_1\left|\mu_1-\rho_1\right| + \omega_a\left|\mu_a-\rho_a\right|\end{aligned}$$

$$\overset{(b)}{=} \sup_{\omega\in\Sigma_K} \min_{a\neq 1} \min(\omega_1,\omega_a)\Delta_a$$
$$\overset{(c)}{=} \sup_{\omega\in\Sigma_K} \omega_1 \min_{a\neq 1} \min(1,\frac{\omega_a}{\omega_1})\Delta_a$$
$$\overset{(d)}{=} \sup_{(x_2,\dots,x_K)\in(\mathbb{R}^+)^{K-1}} \frac{\min_{a\neq 1} g_a(x_a)}{1+x_2+\cdots+x_K},$$

where $g_a(x_a) \triangleq \min(1,x_a)\Delta_a$.

Equality (a) is obtained due to the fact that $\mathrm{Alt}(\boldsymbol{\nu}) = \bigcup_{a\neq 1}\{\boldsymbol{\lambda} : \rho_a > \rho_1\}$, and for Bernoullis, $\mathrm{TV}(\nu_a \parallel \lambda_a) = |\mu_a - \rho_a|$.

Equality (b) is true, since $\inf_{\boldsymbol{\lambda}:\rho_a>\rho_1} \omega_1 |\mu_1-\rho_1| + \omega_a|\mu_a-\rho_a| = \min(\omega_1,\omega_a)\Delta_a$.

Equality (c) holds true, since $\omega_1 \neq 1$ (if $\omega_1 = 0$, the value of the objective is 0).

Equality (d) is obtained by the change of variable $x_a \triangleq \frac{\omega_a}{\omega_1}$

Step 2: Let $(x_2,\dots,x_K)\in(\mathbb{R}^+)^{K-1}$. By the definition of g_a, we have that

$$g_a(x_a) \le x_a\Delta_a \quad \text{and} \quad g_a(x_a) \le \Delta_a.$$

This leads to the inequalities

$$\min_{a\neq 1} g_a(x_a) \le g_a(x_a) \le x_a\Delta_a \quad \text{and} \quad \min_{a\neq 1} g_a(x_a) \le \Delta_{\min}.$$

Thus,

$$\left(\min_{a\neq 1} g_a(x_a)\right)\left(\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}\right) = \frac{\min_{a\neq 1} g_a(x_a)}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{\min_{a\neq 1} g_a(x_a)}{\Delta_a}$$
$$\le 1 + \sum_{a=2}^{K} x_a\,.$$

This means that for every $(x_2,\dots,x_K)\in(\mathbb{R}^+)^{K-1}$,

$$\frac{\min_{a\neq 1} g_a(x_a)}{1+x_2+\cdots+x_K} \le \frac{1}{\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}}.$$

Here, the upper bound is achievable for $x_a^\star = \frac{\Delta_{\min}}{\Delta_a}$, since $g_a(x_a^\star) = \Delta_{\min}$ for all $a\neq 1$.

This concludes that

$$(T^\star_{\mathbf{TV}}(\boldsymbol{\nu}))^{-1} = \frac{1}{\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}} \qquad \Longrightarrow\ (T^\star_{\mathbf{TV}}(\boldsymbol{\nu})) = \frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}\,.$$

Step 3: The lower and upper bounds on $(T^{\star}_{\mathbf{TV}}(\boldsymbol{\nu}))$ follow from the fact that $\frac{1}{\Delta_a} \geq 0$ for all a, and $\frac{1}{\Delta_a} \leq \frac{1}{\Delta_{\min}}$ for all $a \neq 1$.

Hence, we conclude the proof.

□

Appendix C

Supplementary for Chapter 5

Contents

C.1 Privacy Proof of the Generic Wrapper

In this section, we give complete proof of the privacy of the generic wrapper, introduced in Section 5.2. First, the intuition behind the blueprint is formalised in Lemma 2.10, that we first recall in more detail. Then a generic proof of privacy is proposed, followed by a specification for each algorithm given after.

C.1.1 The parallel composition lemma

The Parallel Composition lemma shows that when the mechanism $\mathcal{M}$ is applied to non-overlapping subsets of the input dataset, there is no need to use the composition theorems. Plus, there is no additional cost in the privacy budget.

Let $\mathcal{M}$ be a mechanism that takes a set as input. Let $\ell < T$ and $t_1, \dots t_\ell, t_{\ell+1}$ be in $[1, T]$ such that $1 = t_1 < \cdots < t_\ell < t_{\ell+1} - 1 = T$.
Let's define the following mechanism

$$\mathcal{G} : \{x_1, \dots, x_T\} \to \bigotimes_{i=1}^{\ell} \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}}$$

$\mathcal{G}$ is the mechanism we get by applying $\mathcal{M}$ to the partition of the input dataset $\{x_1, \dots, x_T\}$ according to $t_1 < \cdots < t_\ell < t_{\ell+1}$, i.e.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix} \xrightarrow{\mathcal{G}} \begin{pmatrix} o_1 \\ \vdots \\ o_\ell \end{pmatrix} \tag{C.1}$$

where $o_i \sim \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}}$.

We have that

(a) If $\mathcal{M}$ is (ε, δ)-DP then $\mathcal{G}$ is (ε, δ)-DP

(b) If $\mathcal{M}$ is ρ-zCDP then $\mathcal{G}$ is ρ-zCDP

Proof. Let $x \triangleq \{x_1, \dots, x_T\}$ and $x' \triangleq \{x'_1, \dots, x'_T\}$ be two neighboring datasets. This implies that $\exists j \in [1, T]$ such that $x_j \neq x'_j$ and $\forall t \neq j$, $x_t = x'_t$.

Let ℓ' be such that $t_{\ell'} \leq j \leq t_{\ell'+1} - 1$.

We denote $\{x\}_{t_i}^{t_{i+1}} \triangleq \{x_{t_i}, \dots, x_{t_{i+1}-1}\}$ the records in x corresponding to the episode from t_i until $t_{i+1} - 1$.

(**a**) Suppose that $\mathcal{M}$ is (ε, δ)-DP.

For every output event $E = E_1 \times \cdots \times E_\ell$, we have that

$$\begin{aligned}
\mathcal{G}_x(E) &= \prod_{i=1}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&= \mathcal{M}_{\{x\}_{t_{\ell'}}^{t_{\ell'+1}}}(E_{\ell'}) \prod_{i=1, i\neq\ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&\leq \left(e^{\varepsilon} \mathcal{M}_{\{x'\}_{t_{\ell'}}^{t_{\ell'+1}}}(E_{\ell'}) + \delta \right) \prod_{i=1, i\neq\ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&= e^{\varepsilon} \mathcal{G}_{x'}(E) + \delta \times \prod_{i=1, i\neq\ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&\leq e^{\varepsilon} \mathcal{G}_{x'}(E) + \delta
\end{aligned}$$

since $\prod_{i=1, i\neq\ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \leq 1$

Which gives that $\mathcal{G}$ is (ε, δ)-DP.

(**b**) Suppose that $\mathcal{M}$ is ρ-zCDP. Let denote $o^\ell \triangleq (o_1, \ldots, o_\ell)$ We have that

$$D_\alpha(\mathcal{G}_x \| \mathcal{G}_{x'}) = \frac{1}{\alpha - 1} \log \left(\int_{o^\ell} \mathcal{G}_{x'}(o) \left(\frac{\mathcal{G}_x(o)}{\mathcal{G}_{x'}(o)} \right)^{\alpha} \right)$$

Since

$$\mathcal{G}_x(o) = \prod_{i=1}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(o_i)$$

and

$$\mathcal{G}_{x'}(o) = \prod_{i=1}^{\ell} \mathcal{M}_{\{x'\}_{t_i}^{t_{i+1}}}(o_i)$$

we get

$$\frac{\mathcal{G}_x(o)}{\mathcal{G}_{x'}(o)} = \frac{\mathcal{M}_{\{x\}_{t_{\ell'}}^{t_{\ell'+1}}}(o_i)}{\mathcal{M}_{\{x'\}_{t_{\ell'}}^{t_{\ell'+1}}}(o_i)}$$

Thus,

$$D_\alpha(\mathcal{G}_x \| \mathcal{G}_{x'}) = D_\alpha(\mathcal{M}_{\{x\}_{t_{\ell'}}^{t_{\ell'+1}}} \| \mathcal{M}_{\{x'\}_{t_{\ell'}}^{t_{\ell'+1}}}) \leq \alpha\rho$$

Which gives that $\mathcal{G}$ is ρ-zCDP.

□

For each of the algorithms instanciated using the generic wrapper of Section 5.2, the final actions can be seen as a post-processing of some private quantity of interest (empirical means for finite-armed bandits or the parameter $\hat{\theta}$ for linear and contextual bandits). However, we cannot directly conclude the privacy of the proposed algorithms using just a post-processing argument and Lemma 2.10. This is because the steps corresponding to the start of an episode in the algorithms $t_1 < \cdots < t_\ell < t_{\ell+1}$ are adaptive and depend on the dataset itself, while for Lemma 2.10, those have been fixed before.

To deal with the adaptive episode, we propose a generic privacy proof.

C.1.2 Generic privacy proof

In this section, we give one generic proof that works for the all the proposed algorithms.

First, we give a summary of the intuition of the proof for dealing with adaptive episodes. By fixing two neighbouring tables of rewards d and d' that only differ at some user u_j, and a deterministic adversary B, we have that

- the view of the adversary B from the beginning of the interaction until step j will be the same
- the adaptive episodes generated by the policy in the first j steps will be the same, which means that step j will fall in the same episode in the view of B when interacting with $\pi(d)$ or $\pi(d')$
- for these fixed similar episodes, we use the privacy Lemma 2.10
- the view of B from step $j+1$ until T will be private by post-processing

Let $d = \{x_1, \dots, x_T\}$ and $d' = \{x'_1, \dots, x'_T\}$ two neighbouring reward tables in $(\mathbb{R}^K)^T$. Let $j \in [1, T]$ such that, for all $t \neq j$, $x_t = x'_t$. Let B be a deterministic adversary. We want to show that $D_\alpha(\text{View}(B \leftrightarrow \pi(d)) \| \text{View}(B \leftrightarrow \pi(d'))) \leq \alpha\rho$.

Step 1. Sequential decomposition of the view of the adversary B

We observe that due to the sequential nature of the interaction, the view of B can be decomposed to a part that depends on $d_{<j} \triangleq \{x_1, \dots, x_{j-1}\}$, which is identical for both d and d' and a second conditional part on the history.

First, let us denote $\mathcal{P}_d^{B,\pi} \triangleq \text{View}(B \leftrightarrow^d \pi)$, $\mathbf{o}_{\leq j} \triangleq (o_1, \dots, o_j)$ and $\mathbf{o}_{>j} \triangleq (o_{j+1}, \dots, o_T)$.

We have that, for every sequence of actions $\mathbf{o} \triangleq (o_1, \dots, o_T) \in [K]^T$

$$\begin{aligned}\mathcal{P}_d^{B,\pi}(\mathbf{o}) &= \prod_{t=1}^{T} \pi_t \left(o_t \mid B(o_1), x_{1,B(o_1)}, \dots, B(\mathbf{o}_{\leq t-1}), x_{t-1,B(\mathbf{o}_{\leq t-1})}\right) \\ &\triangleq \mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})\end{aligned}$$

where

$$\mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \triangleq \prod_{t=1}^{j} \pi_t \left(o_t \mid B(o_1), x_{1,B(o_1)}, \ldots, B(\mathbf{o}_{\leq t-1}), x_{t-1,B(\mathbf{o}_{\leq t-1})}\right)$$

and

$$\mathcal{P}_{d}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j}) \triangleq \prod_{t=j+1}^{T} \pi_t \left(o_t \mid B(o_1), x_{1,B(o_1)}, \ldots, B(\mathbf{o}_{\leq t-1}), x_{t-1,B(\mathbf{o}_{\leq t-1})}\right)$$

Similarly

$$\mathcal{P}_{d'}^{B,\pi}(\mathbf{o}) = \mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j})\mathcal{P}_{d'}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})$$

since $d'_{<j} = d_{<j}$.

Step 2. Decomposing the Rényi divergence.

We have that

$$\begin{aligned}
e^{(\alpha-1)D_\alpha(\mathcal{P}_d^{B,\pi} \| \mathcal{P}_{d'}^{B,\pi})} &= \sum_{\mathbf{o}\in[K]^T} \mathcal{P}_{d'}^{B,\pi}(\mathbf{o}) \left(\frac{\mathcal{P}_d^{B,\pi}(\mathbf{o})}{\mathcal{P}_{d'}^{B,\pi}(\mathbf{o})}\right)^\alpha \\
&= \sum_{\mathbf{o}\in[K]^T} \mathcal{P}_{d'}^{B,\pi}(\mathbf{o}) \left(\frac{\mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})}{\mathcal{P}_{d'}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})}\right)^\alpha \\
&= \sum_{\mathbf{o}_{\leq j}\in[K]^j} \mathcal{P}_{d'_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \sum_{\mathbf{o}_{>j}\in[K]^{T-j}} \mathcal{P}_{d'}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j}) \left(\frac{\mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})}{\mathcal{P}_{d'}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})}\right)^\alpha \\
&= \sum_{\mathbf{o}_{\leq j}\in[K]^j} \mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) e^{(\alpha-1)D_\alpha(\mathcal{P}_d^{B,\pi}(.|\mathbf{o}_{\leq j}) \| \mathcal{P}_{d'}^{B,\pi}(.|\mathbf{o}_{\leq j}))} \\
&= \mathbb{E}_{\mathbf{o}_{\leq j}\sim\mathcal{P}_{d_{<j}}^{B,\pi}} \left[e^{(\alpha-1)D_\alpha(\mathcal{P}_d^{B,\pi}(.|\mathbf{o}_{\leq j}) \| \mathcal{P}_{d'}^{B,\pi}(.|\mathbf{o}_{\leq j}))}\right]
\end{aligned}$$

Step 3. The adaptive episodes are the same, before step j.

Let ℓ such that $t_\ell \leq j < t_{\ell+1}$ in the view of B when interacting with d. Let us call it $\psi_d^\pi(j) \triangleq \ell$. Similarly, let ℓ' such that $t_{\ell'} \leq j < t_{\ell'+1}$ in the view of B when interacting with d. Let us call it $\psi_{d'}^\pi(j) \triangleq \ell'$.

Since $\psi_d^\pi(j)$ only depends on $d_{<j}$, which is identical for d and d', we have that $\psi_d^\pi(j) = \psi_{d'}^\pi(j)$ with probability 1.

We call ξ_j the last **time-step** of the episode $\psi_d^\pi(j)$, i.e $\xi_j \triangleq t_{\psi_d^\pi(j)+1} - 1$.

Step 4. Private sufficient statistics.

Fix $\mathbf{o}_{\leq j}$.

Let $r_s \triangleq x_{s,B(o_1,\dots,o_s)}$, for $s \in [1,j]$, be the reward corresponding to the action chosen by B in the table d. Similarly, $r'_s \triangleq x'_{s,B(o_1,\dots,o_s)}$ for d'.

Let us define $L_j \triangleq \mathcal{G}_{\{r_1,\dots,r_{\xi_j}\}}$ and $L'_j \triangleq \mathcal{G}_{\{r'_1,\dots,r'_{\xi_j}\}}$, where $\mathcal{G}$ is defined as in Eq. C.1, using the same episodes for d and d'. The underlying mechanism $\mathcal{M}$, used to define $\mathcal{G}$, will be specified for each algorithm in Section C.1.3.

In addition, the specified mechanism $\mathcal{M}$ will verify ρ-zCDP with respect to its set input.

Using the structure of the policy π, there exists a randomised mapping $f_{x_{\xi_j+1},\dots,x_T}$ such that $\mathcal{P}_d^{B,\pi}(. \mid \mathbf{o}_{\le j}) = f_{x_{\xi_j+1},\dots,x_T}(L_j)$ and $\mathcal{P}_{d'}^{B,\pi}(. \mid \mathbf{o}_{\le j}) = f_{x_{\xi_j+1},\dots,x_T}(L'_j)$.

In other words, the view of the adversary B from step $\xi_j + 1$ until T only depends on the sufficient statistics L_j and the new inputs $x_{\xi_j+1},\dots,x_T$, which are the same for d and d'.

For example, the sufficient statistics are the private mean estimate of the active arm in each episode for AdaC-UCB and the noisy parameter estimate $\hat{\theta}$ for AdaC-GOPE.

Step 5. Concluding with Lemma 2.10 and post-processing.

Using Lemma 2.10, we have that

$$D_\alpha(L_j, L'_j) \le \alpha\rho$$

Using the post-processing property of D_α, we get that

$$\begin{aligned} D_\alpha(\mathcal{P}_d^{B,\pi}(. \mid \mathbf{o}_{\le j}) \| \mathcal{P}_{d'}^{B,\pi}(. \mid \mathbf{o}_{\le j})) &= D_\alpha(f_{x_{\xi_j+1},\dots,x_T}(L_j) \| f_{x_{\xi_j+1},\dots,x_T}(L'_j)) \\ &\le D_\alpha(L_j, L'_j) \le \alpha\rho \end{aligned}$$

Finally, we conclude by taking the expectation with respect to $\mathbf{o}_{\le j} \sim \mathcal{P}_{d_{<j}}^{B,\pi}$

$$\begin{aligned} e^{(\alpha-1)D_\alpha(\mathcal{P}_d^{B,\pi} \| \mathcal{P}_{d'}^{B,\pi})} &= \mathbb{E}_{\mathbf{o}_{\le j} \sim \mathcal{P}_{d_{<j}}^{B,\pi}} \left[e^{(\alpha-1)D_\alpha(\mathcal{P}_d^{B,\pi}(.|\mathbf{o}_{\le j}) \| \mathcal{P}_{d'}^{B,\pi}(.|\mathbf{o}_{\le j}))} \right] \\ &\le e^{(\alpha-1)\alpha\rho} \end{aligned}$$

Thus, we conclude

$$D_\alpha(\mathcal{P}_d^{B,\pi} \| \mathcal{P}_{d'}^{B,\pi}) \le \alpha\rho$$

Remark C.1 (Beyond zCDP)**.** *The same proof could be adapted to Pure DP and other relaxations of Pure DP under Interactive DP.*

C.1.3 Instantiating the specifics of the privacy proof for each algorithm

In this section, we instantiate Step 4 of the generic proof for each algorithm, by specifying the mechanism $\mathcal{M}$ in the proof and showing that they are ρ-zCDP.

• **For** AdaC-UCB, the mechanism $\mathcal{M}$ is the private empirical mean statistic, i.e $\mathcal{M}_{\{r_1,\dots,r_t\}} \triangleq \frac{1}{t}\sum_{s=1}^{t} r_s + \mathcal{N}\left(0, \frac{1}{2\rho t^2}\right)$. Since rewards are in $[0,1]$, by the Gaussian Mechanism (i.e. Theorem 2.14) $\mathcal{M}$ is ρ-zCDP.

• **For** AdaC-GOPE, the mechanism $\mathcal{M}$ is a private estimate of the linear parameter θ, i.e $\mathcal{M}_{\{r_{t_\ell},\dots,r_{t_{\ell+1}-1}\}} \triangleq V_\ell^{-1}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right) + V_\ell^{-\frac{1}{2}} N_\ell$ where $V_\ell = \sum_{a\in\mathcal{S}_\ell} T_\ell(a) aa^\top$, $N_\ell \sim \mathcal{N}\left(0, \frac{2}{\rho} g_\ell^2 I_d\right)$ and $g_\ell = \max_{b\in\mathcal{A}_\ell} \|b\|_{V_\ell^{-1}}$.

To show that $\mathcal{M}$ is ρ-zCDP, we rewrite $\hat{\theta}_\ell = V_\ell^{-1}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right) = V_\ell^{-\frac{1}{2}}\phi_\ell$ where $\phi_\ell \triangleq V_\ell^{-\frac{1}{2}}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right)$.

Let $\{r_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ and $\{r'_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ two neighbouring sequence of rewards that differ at only step $j \in [t_\ell, t_{\ell+1}-1]$. We have that

$$\begin{aligned}\|\phi_\ell - \phi'_\ell\|_2 &= \|V_\ell^{-\frac{1}{2}}\left[a_j(r_s - r'_s)\right]\|_2 \\ &\leq 2\|V_\ell^{-\frac{1}{2}} a_j\|_2 \leq 2g_\ell\end{aligned}$$

since $r_j, r'_j \in [-1,1]$.

Using the Gaussian Mechanism (i.e. Theorem 2.14), this means that $\phi_\ell + N_\ell$ is ρ-zCDP and $\mathcal{M}$ is too by post-processing.

• **For** AdaC-OFUL, the mechanism $\mathcal{M}$ is the private estimate of the sum $\sum_{s=t_\ell}^{t_{\ell+1}-1} a_s r_s$, i.e $\mathcal{M}_{\{r_{t_\ell},\dots,r_{t_{\ell+1}-1}\}} \triangleq \sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s + \mathcal{N}(0, \frac{2}{\rho} I_d)$.

Since rewards are in $[-1,1]$ and $\|a\|_2 \leq 1$, the L2 sensitivity of $\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s$ is 2. By Theorem 2.14, $\mathcal{M}$ is ρ-zCDP.

We need an extra step of cumulatively summing the outputs of $\mathcal{G}$, which is still private by post-processing, i.e

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix} \xrightarrow{\mathcal{G}} \begin{pmatrix} o_1 \\ \vdots \\ o_\ell \end{pmatrix} \to \begin{pmatrix} o_1 \\ o_1 + o_2 \\ \vdots \\ o_1 + o_2 + \cdots + o_\ell \end{pmatrix}$$

Then, we have that $\left(\sum_{t=1}^{t_j} a_s r_s + \sum_{m=1}^{j} Y_m\right)_{j\in[1,\ell]}$ is ρ-zCDP, where $Y_m \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \frac{2}{\rho} I_d\right)$

This shows that the price of not forgetting is, for each estimate at the end of an episode j, to have to sum all the previous independent noises i.e. $\sum_{m=1}^{j} Y_j$, compared to just Y_j when forgetting.

• **For** AdaP-UCB, AdaP-KLUCB, AdaP-TT **and** AdaP-TT*, the mechanism $\mathcal{M}$ is the private empirical mean statistic, i.e $\mathcal{M}_{\{r_1,\dots,r_t\}} \triangleq \frac{1}{t}\sum_{s=1}^{t} r_s + \text{Lap}\left(\frac{1}{\varepsilon t}\right)$. Since rewards are in $[0,1]$, by the Laplace Mechanism (i.e. Theorem 2.13) $\mathcal{M}$ is ε-DP. As observed in Remark C.1, the generic proof of Section C.1.2 is still valid for ε-Interactive DP. Also, for FC-BAI strategies, the same proof is adapted with the minor change of havaing the final recommendation and stopping time as additional outputs of the mechanism. Since this two additional outputs are also solely computed using the sequence of private empirical mean, the same conclusions are valid.

C.2 Finite-armed Bandits with Pure DP and zCDP

C.2.1 Concentration inequalities for pure DP

Lemma C.2. *Assume that $(X_i)_{1\leq i\leq n}$ are iid random variables in $[0,1]$, with $\mathbb{E}(X_i)=\mu$. Then, for any $\delta \geq 0$,*

$$\mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) - \frac{\log\left(\frac{1}{\delta}\right)}{n\varepsilon} - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \geq \mu\right) \leq \frac{3}{2}\delta, \tag{C.2}$$

and

$$\mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log\left(\frac{1}{\delta}\right)}{n\varepsilon} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \leq \mu\right) \leq \frac{3}{2}\delta, \tag{C.3}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$

Proof. We have that

$$\begin{aligned}
p_1 &\triangleq \mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) - \frac{\log\left(\frac{1}{\delta}\right)}{n\varepsilon} - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \geq \mu\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_n - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \geq \mu\right) + \mathbb{P}\left(Lap\left(\frac{1}{n\varepsilon}\right) - \frac{\log\left(\frac{1}{\delta}\right)}{n\varepsilon} \geq 0\right) \\
&\leq \delta + \frac{\delta}{2} = \frac{3}{2}\delta,
\end{aligned}$$

where the last inequality is due to Lemma C.23 and Lemma C.22.

Similarly,

$$\begin{aligned}
p_2 &\triangleq \mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log\left(\frac{1}{\delta}\right)}{n\varepsilon} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \leq \mu\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_n + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \leq \mu\right) + \mathbb{P}\left(Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log\left(\frac{1}{\delta}\right)}{n\varepsilon} \leq 0\right) \\
&\leq \delta + \frac{\delta}{2} = \frac{3}{2}\delta,
\end{aligned}$$

where the last inequality is due to Lemma C.23 and Lemma C.22. □

Lemma C.3. *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables sampled from a Bernoulli distribution with mean μ, and let $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ be the sample mean. Let*

$$\breve{\mu}_n(\delta) \triangleq \mathrm{Clip}_{0,1}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\varepsilon}\right) \tag{C.4}$$

for $\delta > 0$ be the clipped and private empirical mean.

Claim 1. *For any $\delta > 0$ and $\alpha \in [0, \mu]$, the following inequality holds:*

$$\mathbb{P}(\mu \geq \breve{\mu}_n(\delta) + \alpha) \leq \exp(-nd(\mu - \alpha, \mu)) + \frac{1}{2}\delta \tag{C.5}$$

Claim 2. *Furthermore for $\delta \geq 0$, we define*

$$U_n(\delta) \triangleq \max\left\{q \in [0,1] : d\left(\breve{\mu}_n\left(\delta\right), q\right) \leq \frac{\log\left(\frac{1}{\delta}\right)}{n}\right\} \tag{C.6}$$

Then,

$$\mathbb{P}(\mu \geq U_n(\delta)) \leq \frac{3}{2}\delta \tag{C.7}$$

Proof. Here, we prove Claim 1 followed by Claim 2.

Claim 1. Since $\breve{\mu}_n(\delta) = \min\left\{\max\left\{0, \hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\varepsilon}\right\}, 1\right\}$, we have that

$$\begin{aligned}
\mu - \alpha \geq \breve{\mu}_n(\delta) &\Rightarrow \mu - \alpha \geq 1 \quad \text{or} \quad \mu - \alpha \geq \max\left\{0, \left(\hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\varepsilon}\right)\right\} \\
&\Rightarrow \mu - \alpha \geq \hat{\mu}_n + Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\varepsilon} \quad (\text{since } \mu \leq 1) \\
&\Rightarrow \mu - \alpha \geq \hat{\mu}_n \quad \text{or} \quad Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\varepsilon} \leq 0.
\end{aligned}$$

It implies that

$$\begin{aligned}\mathbb{P}(\mu \geq \breve{\mu}_n(\delta) + \alpha) &\leq \mathbb{P}\Big(\mu \geq \hat{\mu}_n + \alpha\Big) + \mathbb{P}\left(Lap\left(\frac{1}{n\varepsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\varepsilon} \leq 0\right) \\ &\leq \exp(-nd(\mu - \alpha, \mu)) + \frac{1}{2}\delta.\end{aligned}$$

The last inequality is due to Equation C.29 of Lemma C.26 and Lemma C.22.

Claim 2.

We have that the sets

$$\begin{aligned}\{\mu \geq U_n(\delta)\} &\underset{(a)}{=} \{\mu \geq U_n(\delta) \geq \breve{\mu}_n(\delta)\} \\ &\underset{(b)}{=} \{d(\breve{\mu}_n(\delta), \mu) \geq d(\breve{\mu}_n(\delta), U_n(\delta)), \mu \geq \breve{\mu}_n(\delta)\} \\ &\underset{(c)}{=} \{d(\breve{\mu}_n(\delta), \mu) \geq \frac{\log(\frac{1}{\delta})}{n}, \mu \geq \breve{\mu}_n(\delta)\} \\ &\underset{(d)}{=} \{\breve{\mu}_n(\delta) \leq \mu - \alpha\}\end{aligned}$$

Here, we chose an $\alpha > 0$ such that $d(\mu - \alpha, \mu) = \frac{\log(\frac{1}{\delta})}{n}$.

Step (a) holds because $U_n(\delta) \geq \breve{\mu}_n(\delta)$ by the definition of $U_n(\delta)$. Step (b) also holds true since $d(\breve{\mu}_n(\delta), \cdot)$ is increasing on $[\breve{\mu}_n(\delta), 1]$. Since $d(\breve{\mu}_n(\delta), U_n(\delta)) = \frac{\log(\frac{1}{\delta})}{n}$ by the definition of $U_n(\delta)$, we obtain the equality in Step (c). Finally, Step (d) is obtained by inverting the relative entropy.

We conclude the proof by

$$\begin{aligned}\mathbb{P}\{\mu \geq U_n(\delta)\} &= \mathbb{P}\left\{\breve{\mu}_n(\delta) \leq \mu - \alpha\right\} \\ &\leq \exp(-nd(\mu - \alpha, \mu)) + \frac{1}{2}\delta \quad \text{(by } \textbf{Claim 1}\text{)} \\ &= \delta + \frac{\delta}{2} = \frac{3}{2}\delta \quad \text{(by substituting } \alpha\text{)}\end{aligned}$$

□

C.2.2 Generic regret analysis for Algorithm 7

Algorithm 7 is a generic framework to construct an extension of any optimistic index-based bandit algorithm, which would satisfy ε-Interactive DP. The algorithm is based on the index I_a^ε of each arm. I_a^ε is computed using the private empirical mean of the last active episode of arm a and is a high probability upper bound of the real mean μ_a.

To explicate the two conditions on arm indexes, we introduce the notation $I_a^\varepsilon(t-1,\beta,s)$, which is the index of arm a, at time-step t and computed using s reward samples from arm a.

Thus, we can express the index computed using just the last active episode as

$$I_a^\varepsilon(t-1,\beta) = I_a^\varepsilon(t-1,\beta,\frac{1}{2}N_a(t-1)). \tag{C.8}$$

Because $I_a^\varepsilon(t-1,\beta)$ only uses samples collected from the last active episode, and due to the doubling, the last active episode's size is exactly half the number of times arm a was pulled since the beginning.

The optimism of the index is ensured by the fact that

$$\mathbb{P}\left(I_a^\varepsilon(t-1,\beta,s) \le \mu_a\right) \le \frac{3}{2}\frac{1}{t^\beta} \tag{C.9}$$

for every arm a, every sample size s and every time-step t, where β is the confidence level.

Theorem C.4. *Let a be a suboptimal arm and $\ell \in \mathbb{N}$ such that $2^\ell < T$. Then, Algorithm 7 using an index I_a^ε satisfying Equations C.8 and C.9, also satisfies that for any $\beta > 3$,*

$$\mathbb{E}[N_a(T)] \le 2^{\ell+1} + \mathbb{P}\left(G^c_{a,\ell,T}\right)T + \frac{\beta}{\beta-3},$$

where $G_{a,\ell,T} = \{\mathrm{I}_a^\varepsilon(T-1,\beta,2^\ell) < \mu^*\}$ and $G^c_{a,\ell,T}$ is the complement of $G_{a,\ell,T}$

Proof. Without loss of generality, we assume the first arm is the optimal one ($\mu^* = \mu_1$) and denote a suboptimal arm by a ($1 < a \le K$).

We leverage the standard idea of UCB-type proofs: if arm a is chosen at the beginning of an episode ℓ, then either its index at t_ℓ is larger than the true mean of the first arm, or the true mean of the first arm is larger than the first arm's index at t_ℓ.

Since decisions, i.e. playing the arm with the highest index, are only taken at the beginning of an episode, we introduce ϕ which takes as input a time step and outputs the time step corresponding to the beginning of an episode. Formally, for each $t \in [K+1, T]$, let $\phi(t) = t_\ell$ such that $t_\ell \le t \le t_{\ell+1} - 1$. In Example 5.1, $\phi(5) = 4$ and $\phi(9) = 7$.

Formally, $\phi(t)$ is a random variable such that

$$\forall t : \phi(t) \le t \le 2\phi(t) \tag{C.10}$$

Step 1: Decomposition of $N_a(T)$. We observe that

$$N_a(T) = 1 + \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a\}$$

$$= 1 + \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a \text{ and } \mathrm{I}_1^{\varepsilon}(\phi(t)-1,\beta) > \mu_1\} + \mathbb{I}\{A_t = a \text{ and } \mathrm{I}_1^{\varepsilon}(\phi(t)-1,\beta) \le \mu_1\}$$

$$\le 1 + \underbrace{N'_a(T)}_{Term1} + \underbrace{\sum_{t=K+1}^{T} \mathbb{I}\{\mathrm{I}_1^{\varepsilon}(\phi(t)-1,\beta) \le \mu_1\}}_{Term2}$$

We define $N'_a(T) \triangleq \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a \text{ and } \mathrm{I}_1^{\varepsilon}(t_{\ell'}-1,\beta) > \mu_1\}$

Step 2: Decomposition of Term 1: $N'_a(T)$**.** Let $G_{a,\ell,T}$ be the 'good' event defined by

$$G_{a,\ell,T} = \{\mathrm{I}_a^{\varepsilon}(T-1,\beta,2^{\ell}) < \mu_1\}.$$

The main part of the proof is decomposing $N'_a(T)$ among the 'good' and the 'bad' events, i.e.

$$\mathbb{E}[N'_a(T)] = \mathbb{E}[\mathbb{I}\{G_{a,\ell,T}\}N'_a(T)] + \mathbb{E}[\mathbb{I}\{G^c_{a,\ell,T}\}N'_a(T)] \le 2^{\ell+1} + \mathbb{P}(G^c_{a,\ell,T})T.$$

$G^c_{a,\ell,T}$ denotes the complement of $G_{a,\ell,T}$.

To prove the last inequality, we only need to prove that when $G_{a,\ell,T}$ happens, $N'_a(T) \le 2^{\ell+1}$. We prove it by contradiction.

Hence, let us assume that $G_{a,\ell,T}$ holds but $N'_a(T) > 2^{\ell+1}$.

This assumption implies that the arm a is played more than $2^{\ell+1}$ times. Thus, there must exist a round $t_{\ell'}$, where $N_a(t_{\ell'}-1) = 2^{\ell+1}$, $A_{t_{\ell'}} = i$ and $\mathrm{I}_1^{\varepsilon}(t_{\ell'}-1,\beta) \ge \mu_1$. Since indices are computed only using the samples from the last active episode, $\mathrm{I}_a^{\varepsilon}(t_{\ell'}-1,\beta)$ is computed using exactly 2^{ℓ} reward samples from arm a.

Thus, we obtain

$$\begin{aligned}
\mathrm{I}_a^{\varepsilon}(t_{\ell'}-1,\beta) &= \mathrm{I}_a^{\varepsilon}(t_{\ell'}-1,\beta,2^{\ell}) \\
&\le \mathrm{I}_a^{\varepsilon}(T-1,\beta,2^{\ell}) \quad \text{(because } t_{\ell'} \le T \text{ and } \mathrm{I}_a^{\varepsilon}(\cdot,\beta,2^{\ell}) \text{ is increasing)} \\
&< \mu_1 \quad \text{(definition of } G_{a,\ell,T}) \\
&\le \mathrm{I}_1^{\varepsilon}(t_{\ell'}-1,\beta)
\end{aligned}$$

The last inequality contradicts the fact that $A_{t_{\ell'}} = i$ and thus, establishes the claim that $N'_a(T) \le 2^{\ell+1}$ under the 'good' event.

Step 3: Upper-bounding Term 2. To conclude,

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{\mathrm{I}_1^{\varepsilon}(\phi(t)-1,\beta) \le \mu_1\}\right] = \sum_{t=K+1}^{T} \mathbb{P}\{\mathrm{I}_1^{\varepsilon}(\phi(t)-1,\beta) \le \mu_1\}$$

$$
\begin{aligned}
&\leq \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \mathbb{P}\{\mathrm{I}_1^{\varepsilon}(\phi-1,\beta) \leq \mu_1\} \\
&\leq \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \sum_{s=1}^{\phi} \mathbb{P}\{\mathrm{I}_1^{\varepsilon}(\phi-1,\beta,s) \leq \mu_1\} \\
&\leq \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \sum_{s=1}^{\phi} \frac{3}{2}\frac{1}{\phi^{\beta}} \qquad \text{(Equation C.9)} \\
&= \frac{3}{2} \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \frac{1}{\phi^{\beta-1}} \\
&\leq \frac{3}{2} \sum_{t=K+1}^{T} \frac{2^{\beta-2}}{t^{\beta-2}} \quad \text{(because } \phi \geq \frac{t}{2}\text{)} \\
&\leq \frac{3}{2} 2^{\beta-2} \int_{K}^{T} \frac{1}{x^{\beta-2}} dx \quad \text{(sum-integral inequality)} \\
&\leq \frac{3}{2} 2^{\beta-2} \frac{1}{\beta-3} \frac{1}{K^{\beta-3}} = \frac{3}{2}\frac{2}{\beta-3}\left(\frac{2}{K}\right)^{\beta-3} \\
&\leq \frac{3}{\beta-3}
\end{aligned}
$$

for $\beta > 3$ and $K \geq 2$.

Here, the first inequality is due to an union bound on $\phi(t) \in [t/2, t]$ (Equation C.10), and the second inequality is due to a union bound on $N_1(\phi - 1)$.

Step 4: Combining the Bounds on Terms 1 and 2.

$$
\begin{aligned}
\mathbb{E}[N_a(T)] &\leq 1 + 2^{\ell+1} + \mathbb{P}\left(G^c_{a,\ell,T}\right) T + \frac{3}{\beta-3} \\
&= 2^{\ell+1} + \mathbb{P}\left(G^c_{a,\ell,T}\right) T + \frac{\beta}{\beta-3}
\end{aligned}
$$

□

Now we design indexes that satisfy the conditions of Theorem C.4, namely AdaP-UCB and AdaP-KLUCB.

To obtain the final regret bounds, we only have to choose ℓ big enough such that

$$\mathbb{P}\left(\mathrm{I}_a(T, 2^{\ell}) \geq \mu_1\right) T$$

is negligible. This corresponds to the leading term in the regret upper-bounds, and this is where the regrets of AdaP-UCB and AdaP-KLUCB differ.

We explicate the issues of designing the indexes and choosing corresponding ℓ in the following section, which leads to the regret upper bounds of AdaP-UCB and AdaP-KLUCB.

C.2.3 Regret analysis for AdaP-UCB and AdaP-KLUCB

Theorem 5.3 (Regret Analysis of AdaP-UCB)**.** *For rewards in* $[0,1]$, AdaP-UCB *satisfies* ε*-Interactive DP, and for* $\beta > 3$*, it yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-UCB}, \nu) \leq \sum_{a:\Delta_a>0} \left(\frac{16\beta}{\min\{\Delta_a, \varepsilon\}} \log(T) + \frac{3\beta}{\beta - 3} \right).$$

Proof. The proof is constituted of three steps.

Step 1: Designing an Index satisfying Equation (C.8), Equation (C.9), and ε-Interactive DP. For AdaP-UCB, the index is defined as

$$\mathrm{I}_a^{\varepsilon}(t_\ell - 1, \beta) = \tilde{\mu}_{a,\varepsilon}^{\ell} + \sqrt{\frac{\beta \log(t_\ell)}{2 \times \frac{1}{2} N_a(t_\ell - 1)}} + \frac{\beta \log(t_\ell)}{\varepsilon \times \frac{1}{2} N_a(t_\ell - 1)},$$

where

$$\tilde{\mu}_{a,\varepsilon}^{\ell} = \hat{\mu}_{a, \frac{1}{2} N_a(t_\ell - 1)} + Lap\left(\frac{1}{\varepsilon \times \frac{1}{2} N_a(t_\ell - 1)} \right) \tag{C.11}$$

is the private empirical mean of arm a computed using only samples from the last active episode, and $\hat{\mu}_{a,s}$ is the empirical mean of arm a calculated using s samples of reward from arm a.

This index verifies the first condition (Equation C.8) of Theorem C.4.

The second condition (Equation C.9) of Theorem C.4 follows directly from Equation C.3 of Lemma C.2

By Section C.1.3, AdaP-UCB is ε-Interactive DP.

By Theorem C.4, for every suboptimal arm a, we have that

$$\mathbb{E}[N_a(T)] \leq 2^{\ell+1} + \mathbb{P}\left(G_{a,\ell,T}^c \right) T + \frac{\beta}{\beta - 3},$$

where

$$G_{a,\ell,T} = \left\{ \hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell \varepsilon} \right) + \sqrt{\frac{\beta \log(T)}{2 \times 2^\ell}} + \frac{\beta \log(T)}{\varepsilon 2^\ell} < \mu_1 \right\}.$$

Step 2: Choosing an ℓ. Now, we observe that

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell\varepsilon}\right) + \sqrt{\frac{\beta\log(T)}{2\times 2^\ell}} + \frac{\beta\log(T)}{\varepsilon 2^\ell} \geq \mu_1\right)\\ &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell\varepsilon}\right) - \sqrt{\frac{\beta\log(T)}{2\times 2^\ell}} - \frac{\beta\log(T)}{\varepsilon 2^\ell} \geq \mu_a + \gamma\right)\end{aligned}$$

for $\gamma = \left(\Delta_a - 2\sqrt{\frac{\beta\log(T)}{2\times 2^\ell}} - 2\frac{\beta\log(T)}{\varepsilon 2^\ell}\right)$.

The idea is to choose ℓ big enough so that $\gamma \geq 0$.

Let us consider the contrary, i.e.

$$\begin{aligned}\gamma < 0 \Rightarrow \sqrt{2^\ell} &< \sqrt{\frac{\beta\log(T)}{2\Delta_a^2}}\left(1+\sqrt{1+\frac{4\Delta_a}{\varepsilon}}\right)\\ \Rightarrow 2^\ell &< \frac{\beta\log(T)}{2\Delta_a^2}\left(4+\frac{8\Delta_a}{\varepsilon}\right)\\ \Rightarrow 2^\ell &< \frac{4\beta\log(T)}{\Delta_a\min\{\varepsilon, 2\Delta_a\}}.\end{aligned} \tag{C.12}$$

Thus, by choosing

$$\ell = \left\lceil \frac{1}{\log(2)}\log\left(\frac{4\beta\log(T)}{\Delta_a\min\{\varepsilon,2\Delta_a\}}\right)\right\rceil$$

we ensure $\gamma > 0$. This also implies that

$$\mathbb{P}(G^c_{a,\ell,T}) \leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell\varepsilon}\right) - \sqrt{\frac{\beta\log(T)}{2\times 2^\ell}} - \frac{\beta\log(T)}{\varepsilon 2^\ell} \geq \mu_a\right) \leq \frac{3}{2T^\beta}$$

The last inequality is due to Equation C.2 of Lemma C.2.

Step 3: The Regret Bound. Combining Steps 1 and 2, we get that

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \frac{\beta}{\beta-3} + 2^{\ell+1} + T\times\frac{3}{2T^\beta}\\ &\leq \frac{16\beta\log(T)}{\Delta_a\min\{\varepsilon,2\Delta_a\}} + \frac{3\beta}{\beta-3}.\end{aligned} \tag{C.13}$$

Plugging this upper bound back in the definition of problem-dependent regret concludes the proof. □

Remark C.5. *The leading term of the regret is $\frac{16\beta\log(T)}{\Delta_a\min\{\varepsilon,2\Delta_a\}}$, which is 4 times more than what we got from Equation C.12. A multiplicative factor of 2 is introduced due to the doubling and another*

multiplicative factor of 2 is due to the forgetting. Thus, the combined price of doubling and forgetting is a multiplicative constant 4 in the leading term of regret.

Theorem 5.4 (Regret Analysis of AdaP-KLUCB)**.** *When the rewards are sampled from Bernoulli distributions,* AdaP-KLUCB *satisfies ε-global DP, and for $\beta > 3$ and constants $C_1(\beta), C_2 > 0$, it yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-KLUCB}, \nu) \leq \sum_{a:\Delta_a>0} \left(\frac{C_1(\beta)\Delta_a}{\min\{d_{\inf}(\mu_a, \mu^*), C_2\varepsilon\Delta_a\}} \log(T) + \frac{3\beta}{\beta - 3} \right).$$

Proof. The proof is constituted of three steps.

Step 1: Designing an Index satisfying Equation (C.8), Equation (C.9), and ε-Interactive DP. For AdaP-KLUCB, the index is defined as

$$\mathrm{I}_a^{\varepsilon}(t_\ell - 1, \beta) = \max\left\{ q \in [0,1] : d\left(\breve{\mu}_{a,\varepsilon}^{\ell,\beta}, q\right) \leq \frac{\beta \log(t_\ell)}{\frac{1}{2}N_a(t_\ell - 1)} \right\} \triangleq U_{a,\frac{1}{2}N_a(t_\ell-1)}\left(\frac{1}{t_\ell^\beta}\right),$$

where $\breve{\mu}_{a,\varepsilon}^{\ell,\beta} = \mathrm{Clip}_{0,1}\left(\tilde{\mu}_{a,\varepsilon}^{\ell} + \frac{\beta\log(t_\ell)}{\varepsilon\frac{1}{2}N_a(t_\ell-1)}\right) = \breve{\mu}_{a,\frac{1}{2}N_a(t_\ell-1)}\left(\frac{1}{t_\ell^\beta}\right)$ as defined in Equation C.4,

$\tilde{\mu}_{a,\varepsilon}^{\ell}$ is the private empirical computed only using the samples from the last active episode (as defined for AdaP-UCB, and $U_{a,s}(\delta) = \max\left\{ q \in [0,1] : d\left(\breve{\mu}_{a,s}\left(\delta\right), q\right) \leq \frac{\log\left(\frac{1}{\delta}\right)}{s} \right\}$ as defined in Equation C.6

This index verifies the first condition (Equation C.8) of Theorem C.4.

The second condition (Equation C.9) of Theorem C.4 follows directly from Equation C.3 of Lemma C.2

By Section C.1.3, AdaP-KLUCB also satisfies ε-Interactive DP.

By Theorem C.4, for every suboptimal arm a, we have that

$$\mathbb{E}[N_a(T)] \leq 2^{\ell+1} + \mathbb{P}\left(G_{a,\ell,T}^c\right) T + \frac{\beta}{\beta - 3},$$

where

$$G_{a,\ell,T} = \left\{ U_{a,2^\ell}\left(\frac{1}{T^\beta}\right) < \mu_1 \right\}.$$

Step 2: Choosing an ℓ. We observe that

$$\begin{aligned}
\mathbb{P}(G_{a,\ell,T}^c) &= \mathbb{P}\left(U_{a,2^\ell}\left(\frac{1}{T^\beta}\right) \geq \mu_1 \right) \\
&\leq \mathbb{P}\left(d^+\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\beta}\right), \mu_1\right) \leq \frac{\beta\log(T)}{2^\ell} \right) \quad \text{(by definition of } U_{a,2^\ell}\text{)}
\end{aligned}$$

where $d^+(p,q) \triangleq d(p,q)\mathbb{I}_{p<q}$and $d(p,q)$ is the relative entropy between Bernoulli distributions as stated in Definition C.24.

Let $\upsilon > 0$, and $c(\upsilon) \in [0,1]$ such that: $d(\mu_a + c(\upsilon)\Delta_a, \mu_1) = \frac{d(\mu_a,\mu_1)}{1+\upsilon}$.

Since $d(\cdot,\mu_1)$ is a bijective function from $[\mu_a,\mu_1]$ to $[0,d(\mu_a,\mu_1)]$, we get that $c(\upsilon)$ always exists and is unique.

In addition, $c(\upsilon)$ verifies: $\lim_{\upsilon\to 0} c(\upsilon) = 0$, $\lim_{\upsilon\to+\infty} c(\upsilon) = 1$ and $c(\upsilon)$ is an increasing function of υ.

First, we choose ℓ such that

$$2^\ell \geq \frac{(1+\upsilon)\beta\log(T)}{d(\mu_a,\mu_1)}. \tag{C.14}$$

This leads to

$$\begin{aligned}
\mathbb{P}(G^c_{a,\ell,T}) &\leq \mathbb{P}\left(d^+\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\beta}\right),\mu_1\right) \leq \frac{d(\mu_a,\mu_1)}{1+\upsilon}\right)\\
&= \mathbb{P}\left(d^+\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\beta}\right),\mu_1\right) \leq d(\mu_a + c(\upsilon)\Delta_a,\mu_1)\right) \quad \text{(definition of } c(\upsilon))\\
&\leq \mathbb{P}\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\beta}\right) \geq \mu_a + c(\upsilon)\Delta_a\right) \quad (d(\cdot,\mu_1) \text{ is decreasing on } [0,\mu_1])\\
&\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell\varepsilon}\right) + \frac{\beta\log(T)}{\varepsilon 2^\ell} \geq \mu_a + c(\upsilon)\Delta_a\right) \quad \text{(definition of } \breve{\mu})
\end{aligned}$$

Let us consider $\gamma_{\ell,T}$ such that $d(\mu_a + \gamma_{\ell,T}\Delta_a,\mu_a) = \frac{\log(T)}{2^\ell}$. We prove its existence and upper bound it later in Fact C.6. Thus, we obtain

$$\begin{aligned}
\mathbb{P}(G^c_{a,\ell,T}) &\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a + Lap\left(\frac{1}{2^\ell\varepsilon}\right) - \frac{\log(T)}{\varepsilon 2^\ell} \geq \mu_a + (c(\upsilon) - \gamma_{\ell,T})\Delta_a - \frac{(1+\beta)\log(T)}{\varepsilon 2^\ell}\right)\\
&= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a + Lap\left(\frac{1}{2^\ell\varepsilon}\right) - \frac{\log(T)}{\varepsilon 2^\ell} \geq \mu_a + \theta\right)
\end{aligned}$$

Here, $\theta \triangleq (c(\upsilon) - \gamma_{\ell,T})\Delta_a - \frac{(1+\beta)\log(T)}{\varepsilon 2^\ell}$.

By choosing

$$2^\ell \geq \frac{(1+\beta)\log(T)}{(c(\upsilon) - \gamma_{\ell,T})\varepsilon\Delta_a}, \tag{C.15}$$

we ensure that $\theta \geq 0$. Thus, we get

$$\begin{aligned}
\mathbb{P}(G^c_{a,\ell,T}) &\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a + Lap\left(\frac{1}{2^\ell\varepsilon}\right) - \frac{\log(T)}{\varepsilon 2^\ell} \geq \mu_a\right)\\
&\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a \geq \mu_a\right) + \mathbb{P}\left(Lap\left(\frac{1}{2^\ell\varepsilon}\right) - \frac{\log(T)}{\varepsilon 2^\ell} \geq 0\right)\\
&\leq \exp\left(-2^\ell d(\mu_a + \gamma_{\ell,T}\Delta_a,\mu_a)\right) + \frac{1}{2T} = \frac{3}{2T}.
\end{aligned}$$

The last inequality is due to Equation C.28 of Lemma C.26 and Lemma C.22.

Fact C.6. $\mathbf{B} \triangleq \{\upsilon > 0 : c(\upsilon) > \gamma_{\ell,T}\} \neq \varnothing$.

Combining both conditions $C.14$ and $C.14$, we choose ℓ to be the smallest integer such that

$$2^{\ell} \geq \inf_{\upsilon \in \mathbf{B}} \max \left\{ \frac{(1+\upsilon)\beta}{d(\mu_a, \mu_1)}, \frac{(1+\beta)}{(c(\upsilon) - \gamma_{\ell,T})\varepsilon\Delta_a} \right\} \log(T) \triangleq \frac{\frac{1}{4}C_1(\beta)}{\min\{d(\mu_a, \mu_1), C_2\varepsilon\Delta_a\}} \log(T)$$

Step 3: The Regret Bound. Combining Steps 1 and 2, we get that

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq 2^{\ell+1} + T \times \frac{3}{2T} + \frac{\beta}{\beta - 3} \\ &\leq \frac{C_1(\beta)}{\min\{d(\mu_a, \mu_1), C_2\varepsilon\Delta_a\}} \log(T) + \frac{3\beta}{\beta - 3}\end{aligned}$$

Plugging this upper bound back in the definition of problem-dependent regret concludes the proof. □

To conclude, we prove Lemma C.6.

Lemma C.6. $\mathbf{B} \triangleq \{\upsilon > 0 : c(\upsilon) > \gamma_{\ell,T}\} \neq \varnothing$.

Proof. **Step 1: Going from $d(\cdot, \mu_a)$ to $d(\cdot, \mu_1)$.** The difficulty of the proof lies in the fact that $\gamma_{\ell,T}$ is defined by inverting $d(\cdot, \mu_a)$ while $c(\upsilon)$ is defined by inverting $d(\cdot, \mu_1)$.

To handle that, we investigate the function $g(x) \triangleq d(x, \mu_a) - d(x, \mu_1)$.

g satisfies the following properties:

- g is continuous and increasing in the interval $[\mu_a, \mu_1]$,
- $g(\mu_a) = -d(\mu_a, \mu_1) < 0$, and
- $g(\mu_1) = d(\mu_1, \mu_a) > 0$.

This implies that there exists a unique root of $g(x)$, where it changes sign. Specifically, there exists a unique $z \in [\mu_a, \mu_1]$ such that:

- $g(z) = 0$
- $\forall x \in [\mu_a, z[: g(x) < 0$
- $\forall x \in]z, \mu_1] : g(x) > 0$

and consequently z verifies $d(z, \mu_a) = d(z, \mu_1)$

Step 2: Choosing υ. We choose υ such that $\frac{d(\mu_a,\mu_1)}{1+\upsilon} = d(z, \mu_a) = d(z, \mu_1)$.

Step 3: Consequence of the choice of υ on $c(\upsilon)$. Thus,

$$d(\mu_a + c(\upsilon)\Delta_a, \mu_1) = d(z, \mu_1),$$

which yields

$$z = \mu_a + c(\upsilon)\Delta_a$$

by uniqueness of z.

Step 4: Consequence of the choice of υ on $\gamma_{\ell,T}$. On the other hand,

$$\begin{aligned} d(\mu_a + \gamma_{\ell,T}\Delta_a, \mu_a) &= \frac{\log(T)}{2^\ell} && \text{(by definition of } \gamma_{\ell,T}) \\ &\leq \frac{d(\mu_a, \mu_1)}{\beta(\upsilon+1)} && \text{(by Equation C.14)} \\ &< d(z, \mu_a) && \text{(since } \beta > 3) \\ &= d(\mu_a + c(\upsilon)\Delta_a, \mu_a) && \text{(C.16)} \end{aligned}$$

As a consequence, we conclude that $\gamma_{\ell,T}$ exists and $\gamma_{\ell,T} < c(\upsilon)$ as $d(\cdot, \mu_a)$ is an increasing function in the interval $[\mu_a, 1]$ □

C.2.4 Gap-free regret bound for AdaP-UCB and AdaP-KLUCB

In this section, we provide problem-independent (or minimax) regret upper bounds for AdaP-UCB.

Theorem C.7. *For rewards in* $[0,1]$, AdaP-UCB *yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-UCB}, \nu) \leq \frac{3\beta}{\beta - 3}\sum_a \Delta_a + 8\sqrt{\beta K T \log(T)} + \frac{16\beta K \log(T)}{\varepsilon}$$

which achieves the minimax lower bound of Thm 4.15 up to $\log(T)$ *factors.*

Proof. Let Δ be a value to be tuned later.
We have that

$$\begin{aligned} \mathrm{Reg}_T(\text{AdaP-UCB}, \nu) = \sum_a \Delta_a \mathbb{E}[N_a(T)] &= \sum_{a:\Delta_a \leq \Delta} \Delta_a \mathbb{E}[N_a(T) + \sum_{a:\Delta_a > \Delta} \Delta_a \mathbb{E}[N_a(T)] \\ &\leq T\Delta + \sum_{a:\Delta_a > \Delta} \Delta_a \left(\frac{16\beta \log(T)}{\Delta_a \min\{\varepsilon, \Delta_a\}} + \frac{3\beta}{\beta - 3} \right) \quad \text{(eq. C.13)} \\ &\leq T\Delta + \frac{16\beta K \log(T)}{\Delta} + \frac{16\beta K \log(T)}{\varepsilon} + \frac{3\beta}{\beta - 3}\sum_a \Delta_a \\ &\leq 8\sqrt{\beta K T \log(T)} + \frac{16\beta K \log(T)}{\varepsilon} + \frac{3\beta}{\beta - 3}\sum_a \Delta_a \end{aligned}$$

where the last step is by taking $\Delta = 4\sqrt{\frac{\beta K \log(T)}{T}}$.

□

Remark C.8. *The same bound is achieved by* AdaP-KLUCB *(up to multiplicative constants) by using that* $d(\mu_a, \mu^*) \geq 2\Delta_a^2$ *and using the same steps in Thm C.7.*

C.2.5 Concentration inequalities for zCDP

Lemma C.9. *Assume that* $(X_i)_{1 \leq i \leq n}$ *are iid random variables in* $[0,1]$*, with* $\mathbb{E}(X_i) = \mu$*. Then, for any* $\delta \geq 0$*,*

$$\mathbb{P}\left(\hat{\mu}_n + Z_n - \sqrt{\left(\frac{1}{2n} + \frac{1}{\rho n^2}\right)\log\left(\frac{1}{\delta}\right)} \geq \mu\right) \leq \delta, \tag{C.17}$$

and

$$\mathbb{P}\left(\hat{\mu}_n + Z_n + \sqrt{\left(\frac{1}{2n} + \frac{1}{\rho n^2}\right)\log\left(\frac{1}{\delta}\right)} \leq \mu\right) \leq \delta, \tag{C.18}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^n X_t$ *and* $Z_n \sim \mathcal{N}\left(0, \frac{1}{2\rho n^2}\right)$.

Proof. Let $Y = (\hat{\mu}_n + Z_n - \mu)$.
Using Properties 2 and 3 of Lemma 2.22, we get that Y is $\sqrt{\frac{1}{4n} + \frac{1}{2\rho n^2}}$-subgaussian.
We conclude using the concentration on subgaussian random variables, i.e. Lemma 2.21. □

C.2.6 Regret analysis for AdaC-UCB

Theorem 5.6 (Regret analysis of AdaC-UCB)**.** *For rewards in* $[0,1]$ *and* $\beta > 3$*,* AdaC-UCB *yields a problem-dependent regret upper bound of*

$$\sum_{a:\Delta_a>0}\left(\frac{8\beta}{\Delta_a}\log(T) + 8\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{2\beta}{\beta-3}\right).$$

and a gap-free regret upper bound of

$$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(K\sqrt{\frac{\log(T)}{\rho}}\right).$$

Proof. By the generic regret decomposition of Theorem C.4, for every sub-optimal arm a, we have that

$$\mathbb{E}[N_a(T)] \leq 2^{\ell+1} + \mathbb{P}\left(G^c_{a,\ell,T}\right)T + \frac{\beta}{\beta-3}, \tag{C.19}$$

where

$$G_{a,\ell,T} = \left\{\hat{\mu}_{a,2^\ell} + Z_\ell + b_{\ell,T} < \mu_1\right\}.$$

such that $b_{\ell,T} \triangleq \sqrt{\left(\frac{1}{2\times 2^\ell} + \frac{1}{\rho\times(2^\ell)^2}\right)\beta\log(T)}$ and $Z_\ell \sim \mathcal{N}\left(0,\, 1/\left(2\rho \times \left(2^\ell\right)^2\right)\right)$.

Step 1: Choosing an ℓ. Now, we observe that

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell + b_{\ell,T} \geq \mu_1\right)\\ &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell - b_{\ell,T} \geq \mu_a + \varepsilon\right)\end{aligned}$$

for $\varepsilon = \Delta_a - 2b_{\ell,T}$.

The idea is to choose ℓ big enough so that $\varepsilon \geq 0$.

Let us consider the contrary, i.e.

$$\begin{aligned}\varepsilon < 0 \Rightarrow 2^\ell &< \frac{2\beta\log(T)}{\Delta_a^2}\left(1 + \Delta_a\sqrt{\frac{1}{\rho\beta\log(T)}}\right)\\ \Rightarrow 2^\ell &< \frac{2\beta}{\Delta_a^2}\log(T) + 2\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)}\end{aligned} \tag{C.20}$$

Thus, by choosing

$$\ell = \left\lceil \frac{1}{\log(2)}\log\left(\frac{2\beta}{\Delta_a^2}\log(T) + 2\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)}\right)\right\rceil$$

we ensure $\varepsilon > 0$. This also implies that

$$\mathbb{P}(G^c_{a,\ell,T}) \leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell - b_{\ell,T} \geq \mu_a\right) \leq \frac{1}{T^\beta}$$

The last inequality is due to Equation C.17 of Lemma C.9, with $n = 2^\ell$ and $\delta = T^{-\beta}$.

Step 2: The regret bound. Plugging the choice of ℓ and the upper bound on $\mathbb{P}(G^c_{a,\ell,T})$ in Inequality C.19 gives

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \frac{\beta}{\beta-3} + 2^{\ell+1} + T\times\frac{1}{T^\beta}\\ &\leq \frac{8\beta}{\Delta_a^2}\log(T) + 8\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)} + \frac{2\beta}{\beta-3}.\end{aligned} \tag{C.21}$$

Plugging this upper bound back in the definition of problem-dependent regret, we get that the regret $\mathrm{Reg}_T(\mathsf{AdaC\text{-}UCB}, \nu)$ is upped bounded by

$$\sum_{a:\Delta_a>0} \left(\frac{8\beta}{\Delta_a} \log(T) + 8\sqrt{\frac{\beta}{\rho}}\sqrt{\log(T)} + \frac{2\beta}{\beta-3} \right).$$

Step 3: The gap-free regret bound. Let Δ be a value to be tuned later.

We observe that

$$\begin{aligned}
&\mathrm{Reg}_T(\mathsf{AdaP\text{-}UCB}, \nu) = \sum_a \Delta_a \mathbb{E}[N_a(T)] \\
&= \sum_{a:\Delta_a\le\Delta} \Delta_a \mathbb{E}[N_a(T) + \sum_{a:\Delta_a>\Delta} \Delta_a \mathbb{E}[N_a(T)] \\
&\le T\Delta + \sum_{a:\Delta_a>\Delta} \Delta_a \left(\frac{8\beta}{\Delta_a^2} \log(T) + 8\sqrt{\frac{\beta\log(T)}{\rho\Delta_a^2}} + \frac{2\beta}{\beta-3} \right) \\
&\le T\Delta + \frac{8\beta K \log(T)}{\Delta} + 8K\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{3\beta}{\beta-3}\sum_a \Delta_a \\
&\le 4\sqrt{2\beta K T \log(T)} + 8K\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{3\beta}{\beta-3}\sum_a \Delta_a
\end{aligned}$$

Here, the last step is tuning $\Delta = \sqrt{\frac{8\beta K \log(T)}{T}}$. □

C.3 Linear Bandits with zCDP

C.3.1 Concentration inequalities

Let $a_1, \dots, a_t$ be deterministically chosen without the knowledge of $r_1, \dots, r_t$. Let π be an optimal design for $\mathcal{A}$.

Let $V_t \triangleq \sum_{s=1}^t a_s a_s^T = \sum_{a\in\mathcal{A}} N_a(t) a a^T$ be the design matrix, $\hat{\theta}_t = V_t^{-1} \sum_{s=1}^t a_s r_s$ be the least square estimate and $\tilde{\theta}_t = \hat{\theta}_t + V_t^{-\frac{1}{2}} N_t$ where $N_t \sim \mathcal{N}\left(0, \frac{2}{\rho} g_t^2 I_d\right)$, where $g_t \triangleq \max_{b\in\mathcal{A}} \|b\|_{V_t^{-1}}$.

Theorem C.10. *Let $\delta \in [0,1]$ and $\beta_t \triangleq g_t\sqrt{2\log\left(\frac{4}{\delta}\right)} + g_t^2\sqrt{\frac{2}{\rho} f(d,\delta)}$, where*

$$f(d,\delta) \triangleq d + 2\sqrt{d\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right).$$

For every $a \in \mathcal{A}$, we have that

$$\mathbb{P}\left(\left| \left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle \right| \ge \beta_t \right) \le \delta.$$

Proof. For every $a \in \mathcal{A}$

$$\begin{aligned}\left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle &= \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle + a^T V_t^{-\frac{1}{2}} N_t \\ &= \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle + Z_t\end{aligned}$$

where $Z_t \triangleq a^T V_t^{-\frac{1}{2}} N_t$.

Step 1: Concentration of the least square estimate. Using Equation (20.2) from Chapter 20 of [LS20], we have that

$$\mathbb{P}\left(\left| \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle \right| \geq g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} \right) \leq \frac{\delta}{2}$$

Step 2: Concentration of the injected Gaussian noise. On the other hand, using Cauchy-Schwartz, we have that

$$|Z_t| = \left| a^T V_t^{-\frac{1}{2}} N_t \right| \leq \|V_t^{-\frac{1}{2}} a\| . \|N_t\| \leq g_t \|N_t\|$$

using that $\|V_t^{-\frac{1}{2}} a\| = \|a\|_{V_t^{-1}} \leq g_t$.

Here, $N_t = \sqrt{\frac{2}{\rho}} g_t \mathcal{N}(0, I_d)$. Thus, using Lemma C.28, we get

$$\mathbb{P}\left(|Z_t| \geq g_t^2 \sqrt{\frac{2}{\rho} f(d, \delta)} \right) \leq \frac{\delta}{2}$$

Steps 1 and 2 together conclude the proof. □

Corollary C.11. *Let β be a confidence level. If each action $a \in \mathcal{A}$ is chosen for $N_a(t) \triangleq \lceil c_t \pi(a) \rceil$ where*

$$c_t \triangleq \frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right) + \frac{2d}{\beta} \sqrt{\frac{2}{\rho} f(d, \delta)}$$

and $f(d, \delta) \triangleq d + 2\sqrt{d \log\left(\frac{2}{\delta}\right)} + 2 \log\left(\frac{2}{\delta}\right)$.

then, for $t = \sum_{a \in \mathrm{Supp}(\pi)} N_a(t)$, we get that

$$\mathbb{P}\left(\left| \left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle \right| \geq \beta \right) \leq \delta \,.$$

Proof. We have that

$$V_t = \sum_{a \in \text{Supp}(\pi)} N_a(t) a a^T \geq c_t V(\pi)$$

This means that

$$g_t^2 = \max_{b \in \mathcal{A}} \|b\|^2_{V_t^{-1}} \leq \frac{1}{c_t} \max_{b \in \mathcal{A}} \|b\|^2_{V(\pi)^{-1}} = \frac{g(\pi)}{c_t} = \frac{d}{c_t},$$

where the last equality is because π is an optimal design for $\mathcal{A}$.

Recall that

$$\beta_t \triangleq g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} + g_t^2 \sqrt{\frac{2}{\rho} f(d, \delta)}$$

Thus,

$$\begin{aligned}
\beta_t &\leq \sqrt{\frac{d}{c_t}} \sqrt{2 \log\left(\frac{4}{\delta}\right)} + \frac{d}{c_t} \sqrt{\frac{2}{\rho} f(d, \delta)} \\
&\leq \frac{\sqrt{2d \log\left(\frac{4}{\delta}\right)}}{\sqrt{\frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right)}} + \frac{d \sqrt{\frac{2}{\rho} f(d, \delta)}}{\frac{2d}{\beta} \sqrt{\frac{2}{\rho} f(d, \delta)}} \\
&= \frac{\beta}{2} + \frac{\beta}{2} = \beta
\end{aligned}$$

The final inequality is due to $c_t \geq \frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right)$, and $c_t \geq \frac{2d}{\beta} \sqrt{\frac{2}{\rho} f(d, \delta)}$.

We conclude the proof using Theorem C.10. □

C.3.2 Regret analysis of AdaC-GOPE

Theorem 5.12 (Regret Analysis of AdaC-GOPE)**.** *Under Assumption 5.10 and for $\delta \in (0, 1)$, with probability at least $1 - \delta$, the regret R_T of* AdaC-GOPE *is upper-bounded by*

$$A \sqrt{dT \log\left(\frac{K \log(T)}{\delta}\right)} + \frac{Bd}{\sqrt{\rho}} \sqrt{\log\left(\frac{K \log(T)}{\delta}\right)} \log(T),$$

where A and B are universal constants. If $\delta = \frac{1}{T}$, then

$$\mathbb{E}(R_T) \leq \mathcal{O}\left(\sqrt{dT \log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}} (\log(KT))^{\frac{3}{2}}\right).$$

Proof. **Step 1: Defining the good event E.** Let

$$E \triangleq \bigcap_{\ell=1}^{\infty} \bigcap_{a \in \mathcal{A}_\ell} \left\{ \left| \left\langle \tilde{\theta}_\ell - \theta_*, a \right\rangle \right| \leq \beta_\ell \right\}.$$

Using Corollary C.11, we get that

$$\begin{aligned} \mathbb{P}(\neg E) &\leq \sum_{\ell=1}^{\infty} \sum_{a \in \mathcal{A}_\ell} \mathbb{P}\left(\left| \left\langle \tilde{\theta}_\ell - \theta_*, a \right\rangle \right| > \beta_\ell \right) \\ &\leq \sum_{\ell=1}^{\infty} \sum_{a \in \mathcal{A}_\ell} \frac{\delta}{K\ell(\ell+1)} \leq \delta \end{aligned}$$

Step 2: Good properties under E. We have that under E

- The optimal arm $a^\star \in \arg\max_{a \in \mathcal{A}} \langle \theta^*, a \rangle$ is never eliminated.

  Proof. for every episode ℓ and $b \in \mathcal{A}_\ell$, we have that under the good event E,

  $$\begin{aligned} \left\langle \tilde{\theta}_\ell, b - a^\star \right\rangle &= \left\langle \tilde{\theta}_\ell - \theta^\star, b - a^\star \right\rangle + \langle \theta^\star, b - a^\star \rangle \\ &\leq \left\langle \tilde{\theta}_\ell - \theta^\star, b - a^\star \right\rangle \\ &\leq \left| \left\langle \tilde{\theta}_\ell - \theta_*, a^\star \right\rangle \right| + \left| \left\langle \tilde{\theta}_\ell - \theta_*, b \right\rangle \right| \leq 2\beta_\ell \end{aligned}$$

  where the first inequality is because $\langle \theta^\star, b - a^\star \rangle \leq 0$ by definition of the optimal arm $a^\star$. This means that $a^\star$ is never eliminated. □

- Each sub-optimal arm a will be removed after ℓ_a rounds where $\ell_a \triangleq \min\{\ell : 4\beta_\ell < \Delta_a\}$.

  Proof. We have that under E,

  $$\begin{aligned} \left\langle \tilde{\theta}_{\ell_a}, a^\star - a \right\rangle &\geq \langle \theta^\star, a^\star \rangle - \beta_{\ell_a} - \langle \theta^\star, a \rangle - \beta_{\ell_a} \\ &= \Delta_a - 2\beta_{\ell_a} > 2\beta_{\ell_a} \end{aligned}$$

  which means that a get eliminated at the round ℓ_a. □

- for $a \in \mathcal{A}_{\ell+1}$, we have that $\Delta_a \leq 4\beta_\ell$.

  Proof. If $\Delta_a > 4\beta_\ell$, then by the definition of ℓ_a, $\ell \geq \ell_a$ and arm a is already eliminated, i.e. $a \notin \mathcal{A}_{\ell+1}$ □

Step 3: Regret decomposition under E.

Fix Δ to be optimised later.

Under E, each sub-optimal action a such that $\Delta_a > \Delta$ will only be played for the first ℓ_Δ rounds where

$$\ell_\Delta \triangleq \min\{\ell : 4\beta_\ell < \Delta\} = \left\lceil \log_2\left(\frac{4}{\Delta}\right)\right\rceil$$

We have that

$$\begin{aligned}
R_T &= \sum_{a\in\mathcal{A}} \Delta_a N_a(T) \\
&= \sum_{a:\Delta_a>\Delta} \Delta_a N_a(T) + \sum_{a:\Delta_a\le\Delta} \Delta_a N_a(T) \\
&= \sum_{\ell=1}^{\ell_\Delta\wedge\ell(T)} \sum_{a\in\mathcal{A}_\ell} \Delta_a T_\ell(a) + T\Delta \\
&\le \sum_{\ell=1}^{\ell_\Delta\wedge\ell(T)} 4\beta_{\ell-1} T_\ell + T\Delta
\end{aligned}$$

where the last inequality is thanks to the third bullet point in **Step 2**, i.e. $\Delta_a \le 4\beta_{\ell-1}$ for $a \in \mathcal{A}_\ell$.

Also $\ell(T)$ is the total number of episodes played until timestep T.

Step 4: Upper-bounding T_ℓ and $\ell(T)$ under E. Let $\delta_{K,\ell} \triangleq \frac{\delta}{K\ell(\ell+1)}$. We recall that $f(d,\delta) \triangleq d + 2\sqrt{d\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)$.

We have that

$$\begin{aligned}
T_\ell &= \sum_{a\in S_\ell} T_\ell(a) \\
&= \sum_{a\in S_\ell} \left\lceil \frac{8d\pi_\ell(a)}{\beta_\ell^2}\log\left(\frac{4}{\delta_{K,\ell}}\right) + \frac{2d\pi_\ell(a)}{\beta_\ell}\sqrt{\frac{2}{\rho}f(d,\delta_{K,\ell})}\right\rceil \\
&\le \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2}\log\left(\frac{4}{\delta_{K,\ell}}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2}{\rho}f(d,\delta_{K,\ell})}.
\end{aligned}$$

since $\beta_{\ell+1} = \frac{1}{2}\beta_\ell$ and $\sum_{\ell=1}^{\ell(T)} T_\ell = T$, there exists a constant C such that $\ell(T) \le C\log(T)$. In other words, the length of the episodes is at least doubling so their number is logarithmic.

Which means that, for $\ell \le \ell(T)$, there exists a constant C' such that

$$\log\left(\frac{4}{\delta_{K,\ell}}\right) = \log\left(\frac{4K\ell(\ell+1)}{\delta}\right) \le C'\log\left(\frac{K\log(T)}{\delta}\right).$$

Define $\alpha_T \triangleq \log\left(\frac{K\log(T)}{\delta}\right)$

$$T_\ell \leq \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2}C'\alpha_T + \frac{4d}{\beta_\ell}\sqrt{\frac{1}{\rho}C'\alpha_T}$$

Step 5: Upper-bounding regret under E.

Under E

$$\begin{aligned}
&\sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 4\beta_{\ell-1}T_\ell \\
&\leq \sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 8\beta_\ell\left(\frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2}C'\alpha_T + \frac{4d}{\beta_\ell}\sqrt{\frac{1}{\rho}C'\alpha_T}\right) \\
&\leq 4d(d+1) + 64dC'\alpha_T\left(\sum_{\ell=1}^{\ell_\Delta} 2^\ell\right) + 32d\sqrt{\frac{1}{\rho}C'\alpha_T}\ell(T) \\
&\leq 4d(d+1) + 16dC'\alpha_T\left(\frac{16}{\Delta}\right) + 32d\sqrt{\frac{1}{\rho}C'\alpha_T}\ell(T) \\
&\leq 4d(d+1) + C_1 d\alpha_T\frac{1}{\Delta} + C_2 d\sqrt{\frac{1}{\rho}\alpha_T}\log(T)
\end{aligned}$$

All in all, we have that

$$R_T \leq 4d(d+1) + C_2 d\sqrt{\frac{1}{\rho}\alpha_T}\log(T) + C_1 d\alpha_T\frac{1}{\Delta} + T\Delta$$

Step 6: Optimizing for Δ. We take

$$\Delta = \sqrt{\frac{C_1 d}{T}\alpha_T}.$$

We get an upper bound on R_T of

$$A\sqrt{dT\log\left(\frac{k\log(T)}{\delta}\right)} + Bd\sqrt{\frac{1}{\rho}\log\left(\frac{k\log(T)}{\delta}\right)}\log(T)$$

Step 7: Upper-bounding the expected regret. For $\delta = \frac{1}{T}$, we get that

$$\mathbb{E}(R_T) \leq (1-\delta)R_T(\delta) + \delta T$$

$$\leq R_T(\delta) + 1$$
$$\leq C_1' \sqrt{dT \log(kT)} + C_2' \sqrt{\frac{1}{\rho}} d \log(kT)^{\frac{3}{2}}$$

□

C.3.3 Adding noise at different steps of GOPE

In order to make the GOPE algorithm differentially private, the main task is to derive a private estimate of the linear parameter θ at each phase ℓ, i.e. $\hat{\theta}_\ell$. If the estimate is private with respect to the samples used to compute it, i.e. $\hat{\theta}_\ell = V_\ell^{-1} \left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s \right)$ w.r.t $\{r_s\}_{s=t_\ell}^{t_{\ell+1}-1}$, then due to forgetting and post-processing, the algorithm turns private too.

We discuss three different ways to make the empirical estimate $\hat{\theta}_\ell$ private.

Adding noise in the end

A first attempt would be to analyse the L_2 sensitivity of $\hat{\theta}_\ell$ directly, and adding Gaussian noise calibrated by the L_2 sensitivity of $\hat{\theta}_\ell$.

Let $\{r_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ and $\{r'_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ two neighbouring sequence of rewards that differ at only step $j \in [t_\ell, t_{\ell+1} - 1]$. Then, we have that

$$\begin{aligned} \|\hat{\theta}_\ell - \hat{\theta'}_\ell\|_2 &= \|V_\ell^{-1} \left[a_j (r_s - r'_s) \right] \|_2 \\ &\leq 2 \|V_\ell^{-1} a_j\|_2 \end{aligned}$$

since $r_j, r'_j \in [-1, 1]$.

However, it is hard to control the quantity $\|V_\ell^{-1} a_j\|_2$ without additional assumptions. The G-optimal design permits only to control another related quantity, i.e. $\|a_j\|_{V_\ell^{-1}} = \|V_\ell^{-\frac{1}{2}} a_j\|_2$. Thus, it is better to add noise at a step before if one does not want to add further assumption.

Adding noise in the beginning

Since $\hat{\theta}_\ell = V_\ell^{-1} \left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s \right)$, another way to make $\hat{\theta}_\ell$ private is by adding noise directly to the sum of observed rewards.

Specifically, one can rewrite the sum

$$\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s = \sum_{a \in S_\ell} a \sum_{a_t = a, t \in [t_\ell, t_{\ell+1}-1]} r_t \, .$$

Since rewards are in $[-1, 1]$, the L_2 sensitivity of $\sum_{a_t=a, t\in[t_\ell, t_{\ell+1}-1]} r_t$ is 2.

Thus, by Theorem 2.14, this means that the noisy sum of rewards $\sum_{a_t=a, t\in[t_\ell, t_{\ell+1}-1]} r_t + \mathcal{N}\left(0, \frac{2}{\rho}\right)$ is ρ-zCDP. Hence, by post-processing lemma, the corresponding noisy estimate $\hat{\theta}_\ell + V_\ell^{-1}\left(\sum_{a\in S_\ell} a\mathcal{N}\left(0, \frac{2}{\rho}\right)\right)$ is a ρ-zCDP estimate of $\hat{\theta}_\ell$.

This is exactly how both [HGFD22] and [LZJ22] derive a private version of GOPE for different privacy definitions, i.e. pure ε-DP for [HGFD22] and (ε, δ)-DP for [LZJ22], respectively. The drawback of this approach is that the variance of the noise depends on the size of the support S_ℓ of the G-optimal design.

To deal with this, both [HGFD22] and [LZJ22] solve a variant of the G-optimal design to get a solution where $|S_\ell| \leq 4d \log\log d + 16$ rather than the full $d(d+1)/2$ support of AdaC-GOPE's optimal design. And still, the dependence on d in the private part of the regret achieved by both these algorithms are d^2 in [HGFD22, Eq (18)], and $d^{\frac{3}{2}}$ in [LZJ22, Eq (56)], respectively. Thus, both of these existing algorithms do not achieve to the linear dependence on d in the regret term due to privacy, as suggested by the minimax lower bound.

Adding noise at an intermediate level

In contrast, AdaC-GOPE adds noise to the statistic

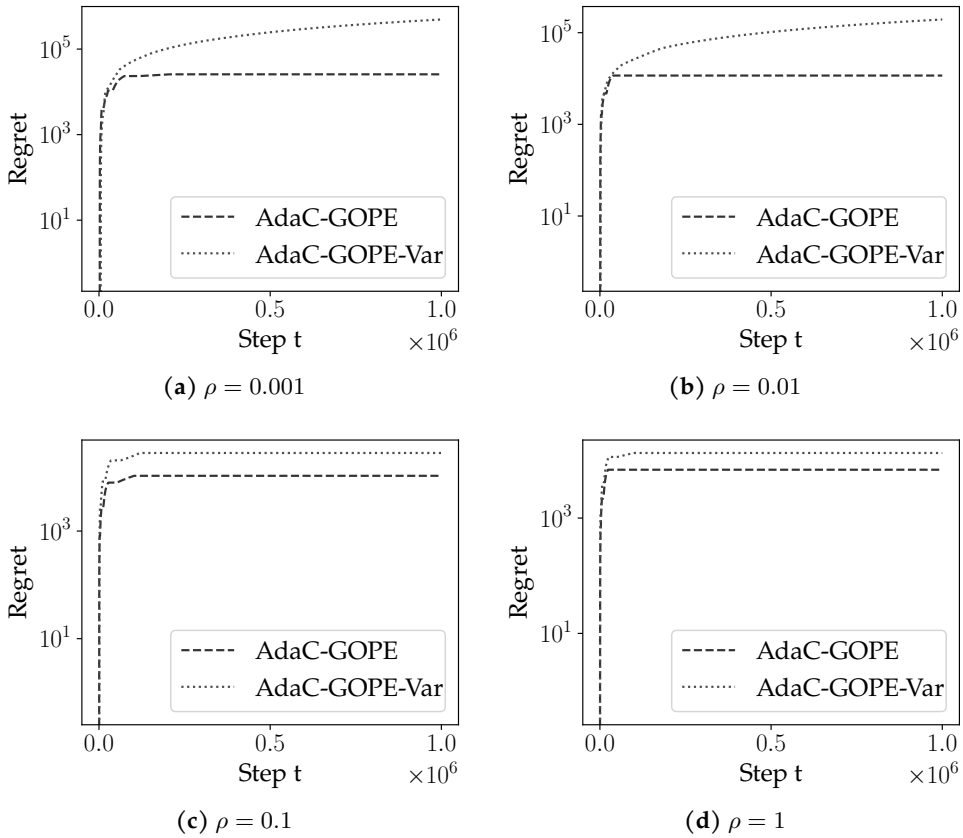
$$\phi_\ell = V_\ell^{-\frac{1}{2}} \left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s \right).$$

ϕ_ℓ is an intermediate quantity between the sum of rewards $\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s$, and the parameter $\hat{\theta}_\ell$, whose L_2 sensitivity can be controlled directly using the G-optimal Design. Due to this subtle observation, the private estimation $\tilde{\theta}_\ell$ of AdaC-GOPE is independent of the size of the support S_ℓ. Hence, the regret term of AdaC-GOPE due to privacy enjoys a linear dependence on d, as suggested by the minimax lower bound.

Conclusion

In brief, to achieve the same DP guarantee with the same budget, one may arrive at it by adding noise at different steps, and the resulting algorithms may have different utilities. In general, adding noise at an intermediate level of computation (not directly to the input, i.e. local and not output perturbation) generally gives the best results.

Remark C.12 (AdaC-GOPE VS variants)**.** *We also compare the empirical performance of* AdaC-GOPE *with a variant where the noise is added to the sum statistic i.e.* $\tilde{\theta}_\ell \triangleq \hat{\theta}_\ell + V_\ell^{-1}\left(\sum_{a\in S_\ell} a\mathcal{N}\left(0, \frac{2}{\rho}\right)\right)$*. we add an experimental comparison between* AdaC-GOPE *and a variant of* AdaC-GOPE *where the way of*



(a) $\rho = 0.001$ **(b)** $\rho = 0.01$

(c) $\rho = 0.1$ **(d)** $\rho = 1$

Figure C.1 – Evolution of the regret over time for AdaC-GOPE and Adar-GOPE-Var for different values of the privacy budget ρ.

making the estimate $\hat{\theta}_\ell$ private is different (Section C.3.3). In AdaR-GOPE-Var, Step 4 changes to

$$\tilde{\theta}_\ell^{\textit{AdaR-GOPE-Var}} = \hat{\theta}_\ell + V_\ell^{-1} \left(\sum_{a \in S_\ell} a\mathcal{N}\left(0, \frac{2}{\rho}\right) \right) .$$

We compare AdaC-GOPE *and AdaR-GOPE-Var in the same experimental setup and instances as in Section 5.5, for different privacy budgets ρ and report the results in Figure C.1.*

As suggested by the regret analysis, AdaC-GOPE *achieves less regret, especially in the high privacy regime where the private part of the regret has more impact.*

C.4 Linear Contextual Bandits with zCDP

C.4.1 Confidence bound for the private least-square estimator

Theorem C.13. *Let $\delta \in (0,1)$. Then, with probability $1-\mathcal{O}(\delta)$, it holds that, for all $t \in [1,T]$,*

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \tilde{\beta}_t$$

where

$$\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$$

such that

$$\beta_t = \mathcal{O}\left(\sqrt{d\log(t)}\right) \text{ and } \gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho}d\log(t)}\right)$$

and β_t and γ_t are increasing in t.

Proof. **Step 1: Decomposing $\tilde{\theta}_t - \theta^\star$.** We have that

$$\begin{aligned}
\tilde{\theta}_t - \theta^\star &= V_t^{-1}\left(\sum_{s=1}^{t} A_s R_s + \sum_{m=1}^{\ell(t)} Y_m\right) - \theta^\star \\
&= V_t^{-1}\left(\sum_{s=1}^{t} A_s(A_s^T\theta^\star + \eta_s) + \sum_{m=1}^{\ell(t)} Y_m\right) - \theta^\star \\
&= V_t^{-1}\left((V_t - \lambda I_d)\theta^\star + \sum_{s=1}^{t} A_s\eta_s + \sum_{m=1}^{\ell(t)} Y_m\right) - \theta^\star \\
&= V_t^{-1}\left(S_t + N_t - \lambda\theta^\star\right)
\end{aligned}$$

where $S_t \triangleq \sum_{s=1}^{t} A_s\eta_s$, $N_t = \sum_{m=1}^{\ell(t)} Y_m \sim \mathcal{N}\left(0, \frac{2\ell(t)}{\rho} I_d\right)$ and $\ell(t)$ is the number of episodes until time-step t number of updates of $\tilde{\theta}$.

Which gives that

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} = \|S_t + N_t - \lambda\theta^\star\|_{V_t^{-1}}$$

Step 2: Defining the Good Event E. We call E_1, E_2 and E_3 respectively the events

$$\left\{\forall t \in [T] : \|S_t\|_{V_t^{-1}} \leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)}\right\},$$

$$\{\forall t \in [T] : \lambda_{\min}(G_t) \geq g(t, \lambda_0, \delta, d)\},$$
$$\left\{\forall t \in [T] : \|N_t\| \leq \sqrt{\frac{2\ell(t)}{\rho} f\left(d, \frac{\delta}{T}\right)}\right\}$$

where $G_t \triangleq \sum_{s=1}^{t} A_s A_s^T$, $g(t, \lambda_0, \delta, d) \triangleq \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}$ and $f(d, \delta) \triangleq d + 2\sqrt{d\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)$.

Let

$$E = E_1 \cap E_2 \cap E_3 \tag{C.22}$$

Step 3: Showing that E Happens with High Probability.

For event E_1:

By a direct application of Lemma C.30, we get that

$$\mathbb{P}(\neg E_1) \leq \delta.$$

For event E_2:

By a direct application of Lemma 5.16, we get that

$$\mathbb{P}(\neg E_2) \leq \delta.$$

For event E_3:

Since $N_t \sim \mathcal{N}\left(0, \frac{2\ell(t)}{\rho} I_d\right)$, a direct application of Lemma C.28 gives that

$$\mathbb{P}(\neg E_3) \leq \delta.$$

All in all, we get that $\mathbb{P}(E) \geq 1 - 3\delta$.

Step 4: Upper-bounding $\|\tilde{\theta}_t - \theta^\star\|_{V_t}$ under E. We have that,

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \|N_t\|_{V_t^{-1}} + \|\lambda\theta^\star\|_{V_t^{-1}}$$

Under E, $V_t \geq (\lambda + \lambda_{\min}(G_t)) I_d \geq \lambda I_d$.

Which gives that, under E,

$$\|N_t\|_{V_t^{-1}} \leq \frac{1}{\sqrt{\lambda + \lambda_{\min}(G_t)}} \|N_t\|$$

$$\leq \sqrt{\frac{\frac{2\ell(t)}{\rho}\left(d + 2\sqrt{d\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{T}{\delta}\right)\right)}{\lambda + \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}}}$$
$$\triangleq \frac{\gamma_t}{\sqrt{t}}$$

and

$$\begin{aligned}
&\|S_t\|_{V_t^{-1}} + \|\lambda\theta^\star\|_{V_t^{-1}} \\
&\leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)} + \frac{\lambda}{\sqrt{\lambda}}\|\theta^\star\| \\
&= \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)} + \sqrt{\lambda}\|\theta^\star\| \triangleq \beta_t
\end{aligned}$$

So, under E, we have that

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \tilde{\beta}_t$$

where

$$\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$$

Step 5: Upper-bounding $\det(V_t)$ and $\ell(t)$.

Under E, using the determinant trace inequality, we have that

$$\det(V_t) \leq \left(\frac{1}{d}\text{trace}(V_t)\right)^d \leq \left(\frac{d\lambda + t}{d}\right)^d$$

which gives that

$$\beta_t = \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t}{\lambda d}\right)} + \sqrt{\lambda}\|\theta^\star\|$$

We can say that $\beta_t = \mathcal{O}(\sqrt{d\log(t)})$.

On the other hand, after each episode, the $\det(V_t)$ is, at least, increased multiplicatively by $(1 + C)$, which means that under E, we have that

$$(1 + C)^{\ell(t)}\det(V_0) \leq \det(V_t) \leq \left(\lambda + \frac{t}{d}\right)^d$$

which gives that

$$\ell(t) \leq \frac{d}{\log(1+C)} \log\left(1 + \frac{t}{\lambda d}\right)$$

so $\ell(t) = \mathcal{O}(d\log(t))$ and $\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho} d\log(t)}\right)$

Step 6: Final Touch.

Under event E, we have that $\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \tilde{\beta}_t$ where $\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$, $\beta_t = \mathcal{O}(\sqrt{d\log(t)})$ and $\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho} d\log(t)}\right)$ such that β_t and γ_t are increasing.

□

C.4.2 Regret analysis of AdaC-OFUL

Theorem 5.17 (Regret Analysis of AdaC-OFUL). *Under Assumptions 5.10 and 5.15, and for $\delta \in (0,1]$, with probability at least $1-\delta$, the regret R_T of* AdaC-OFUL *is upper bounded by*

$$R_T \leq \mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\frac{d^2}{\sqrt{\rho}}\log(T)^2\right)$$

Proof. Let E be the event defined in equation C.22.

Step 1: Regret Decomposition.

Let $A_t^\star = \arg\max_{a\in\mathcal{A}_t} \langle \theta^\star, a\rangle$.

We have that

$$R_T = \sum_{t=1}^{T} r_t, \quad \text{where} \quad r_t = \langle \theta^\star, A_t^\star - A_t\rangle$$

Step 2: Upper-bounding Instantaneous Regret under E.

At step t, let τ_t be the last step where $\tilde{\theta}$ was updated.

Let $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \tilde{\theta}_{t-1}\|_{V_{t-1}} \leq \tilde{\beta}_{t-1}\}$ and $\mathrm{UCB}_t(a) = \max_{\theta\in\mathcal{C}_t} \langle \theta, a\rangle$.

Also, define $\breve{\theta}_{\tau_t} = \arg\max_{\theta\in\mathcal{C}_{\tau_t}} \langle \theta, A_t\rangle$ so that $\mathrm{UCB}_{\tau_t}(A_t) = \left\langle \breve{\theta}_{\tau_t}, A_t\right\rangle$.

Finally, Line 11 of Algorithm 9 could be re-written as $A_t = \arg\max_{a\in\mathcal{A}_t} \mathrm{UCB}_{\tau_t}(a)$.

Under E, we have that

$$\begin{aligned} r_t &= \langle \theta^\star, A_t^\star - A_t\rangle \\ &\overset{(a)}{\leq} \left\langle \breve{\theta}_{\tau_t} - \theta^\star, A_t\right\rangle \\ &\overset{(b}{\leq} \|\breve{\theta}_{\tau_t} - \theta^\star\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}} \end{aligned}$$

$$\overset{(c)}{\leq} \sqrt{\frac{\det(V_{t-1})}{\det(V_{\tau_t})}} \|\breve{\theta}_{\tau_t} - \theta^\star\|_{V_{\tau_t}} \|A_t\|_{V_{t-1}^{-1}}$$
$$\overset{(d)}{\leq} \sqrt{1+C}(2\tilde{\beta}_{\tau_t}) \|A_t\|_{V_{t-1}^{-1}}$$

where:

(a) Under E, $\theta^\star \in \mathcal{C}_{\tau_t}$ and $\langle \theta^\star, A_t^\star \rangle \leq \max_{\theta \in \mathcal{C}_{\tau_t}} \langle \theta, A_t^\star \rangle = \mathrm{UCB}_{\tau_t}(A_t^\star) \leq \mathrm{UCB}_{\tau_t}(A_t) = \left\langle \breve{\theta}_{\tau_t}, A_t \right\rangle$.

(b) By the Cauchy-Schwartz inequality.

(c) By Lemma C.31.

(d) By definition of τ_t and Line 6 of Algorithm 9 , we have that $\det(V_{t-1}) \leq (1+C)\det(V_{\tau_t})$ and under E, $\theta^\star \in \mathcal{C}_{\tau_t}$, so $\|\breve{\theta}_{\tau_t} - \theta^\star\|_{V_{\tau_t}} \leq 2\tilde{\beta}_{\tau_t}$.

We also have that $r_t \leq 2$ and $\tilde{\beta}_{\tau_t} \leq \beta_T + \frac{\gamma_T}{\sqrt{\tau_t}}$, which gives

$$r_t \leq 2\sqrt{1+C}\beta_T \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) + 2\sqrt{1+C} \frac{\gamma_T}{\sqrt{\tau_t}} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right)$$

Step 3: Upper-bounding Regret under E.
Under E, we have that

$$\begin{aligned} R_T &= \sum_{t=1}^T r_t \\ &\leq 2\sqrt{1+C}\beta_T \sum_{t=1}^T \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) \\ &\quad + 2\sqrt{1+C}\gamma_T \sum_{t=1}^T \frac{1}{\sqrt{\tau_t}} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) \\ &\leq 2\sqrt{1+C}\beta_T \sqrt{T \sum_{t=1}^T 1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}} \\ &\quad + 2\sqrt{1+C}\gamma_T \sqrt{\left(\sum_{t=1}^T \frac{1}{\tau_t}\right)\left(\sum_{t=1}^T 1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}\right)} \end{aligned} \tag{C.23}$$

where the last inequality is due to the Cauchy-Schwartz inequality.

Step 4: The Elliptical Potential Lemma.

We use that $1 \wedge x \leq \log(1+x)$ and $\det(V_t) = \det(V_{t-1})\left(1 + \|A_t\|^2_{G_{t-1}(\lambda)^{-1}}\right)$ to have that

$$\begin{aligned}\sum_{t=1}^{T}\left(1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}\right) &\leq 2\sum_{t=1}^{T}\log\left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right) \\ &= 2\log\left(\frac{\det(V_T)}{\det(V_0)}\right) \\ &\leq 2d\log\left(1 + \frac{T}{\lambda d}\right) \end{aligned} \tag{C.24}$$

often known as the elliptical potential lemma (Lemma 19.4, [LS20].

Step 5: Upper-bounding the Length of Every Episode .

Episode ℓ starts at t_ℓ and ends at $t_{\ell+1} - 1$, so we have that

$$\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} \leq 1 + C \tag{C.25}$$

On the other hand,

$$\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} = \prod_{t=t_\ell+1}^{t_{\ell+1}-1}\left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right) \tag{C.26}$$

Under E, we use that

$$V_{t-1} \leq (\lambda + \lambda_{\max}(G_{t-1}))\, I_d \leq (\lambda + t - 1)\, I_d$$

since $\lambda_{\max}(G_{t-1}) \leq \mathrm{trace}(G_{t-1}) \leq t-1$.

which gives that

$$\|A_t\|^2_{V_{t-1}^{-1}} \geq \frac{1}{\lambda + t - 1}$$

Plugging in Equation C.26, we get that

$$\begin{aligned}\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} &\geq \prod_{t=t_\ell+1}^{t_{\ell+1}-1}\left(1 + \frac{1}{\lambda + t - 1}\right) \\ &= \prod_{t=t_\ell+1}^{t_{\ell+1}-1}\left(\frac{\lambda + t}{\lambda + t - 1}\right) = \frac{\lambda + t_{\ell+1} - 1}{\lambda + t_\ell} \\ &\geq \frac{1}{\lambda + 1}\frac{t_{\ell+1}}{t_\ell}\end{aligned}$$

where the last inequality uses that $t_\ell \geq 1$ and $\lambda \geq 1$.

Finally using the upper bound of Equation C.25, we get that

$$\frac{t_{\ell+1}}{t_\ell} \leq (1+C)(1+\lambda)$$

Which gives that

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{\tau_t} &= \sum_{\ell=1}^{\ell(T)} \sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{1}{t_\ell} \\ &= \sum_{\ell=1}^{\ell(T)} \frac{t_{\ell+1}-t_\ell}{t_\ell} \leq (1+C)(1+\lambda)\ell(T) \end{aligned} \tag{C.27}$$

Step 6: Final Touch.

Plugging the upper bounds of Equation C.24 and C.27 in the regret upper bound of Equation C.23, we get that

$$\begin{aligned}R_T \leq 2\sqrt{1+C}\sqrt{2d\log\left(1+\frac{T}{\lambda d}\right)}\Bigg(\beta_T\sqrt{T} \\ +\gamma_T\sqrt{(1+C)(1+\lambda)\ell(T)}\Bigg)\end{aligned}$$

We finalise by using that

$$\beta_T = \mathcal{O}\left(\sqrt{d\log(T)}\right), \gamma_T = \mathcal{O}\left(\sqrt{\frac{1}{\rho}}d\log(T)\right)$$
$$\text{and } \ell(T) = \mathcal{O}\left(d\log(T)\right)$$

We get that

$$R_T \leq \mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{1}{\rho}}d^2\log(T)^2\right)$$

□

C.4.3 Rectifying LinPriv regret analysis

[NR18] propose "LinPriv: Reward-Private Linear UCB", an ε-global DP linear contextual bandit algorithm. The context is assumed to be public but adversely chosen. The algorithm is an

ε-global DP extension of OFUL, where the reward statistics are estimated, at each time-step and for every arm, using a tree-based mechanism [DNPR10b, CSS11].

Theorem 5 in [NR18] claims that the regret of LinPriv is of order

$$\tilde{\mathcal{O}}\left(d\sqrt{T}+\frac{1}{\varepsilon}Kd\log T\right).$$

We believe there is a mistake in their regret analysis. In the proof of Theorem 5, page 25, they say that

"The crux of their analysis is actually the bound $\sum_{t=1}^{n}\|x_{i,t}\|_{V_{i,t}^{-1}} \leq 2d\log\left(1+\frac{n}{\lambda d}\right)$. "

However, we believe that the result they are citing from [AYPS11] is erroneous. The correct one is

$$\sum_{t=1}^{n}\|x_{i,t}\|^2_{V_{i,t}^{-1}} \leq 2d\log\left(1+\frac{n}{\lambda d}\right),$$

which is known as the elliptical potential lemma (Eq. (C.24)).

To get the sum, a Cauchy-Schwartz inequality is generally used which leads to

$$\sum_{t=1}^{n}\|x_{i,t}\|_{V_{i,t}^{-1}} \leq \sqrt{n\sum_{t=1}^{n}\|x_{i,t}\|^2_{V_{i,t}^{-1}}} \leq \sqrt{2nd\log\left(1+\frac{n}{\lambda d}\right)}$$

After n is replaced by $\frac{T}{K}$, an additional multiplicative $\sqrt{T}$ should appear in the private regret.

Thus, the rectified regret should be $\tilde{\mathcal{O}}\left(d\sqrt{T}+\frac{1}{\varepsilon}Kd\sqrt{T}\right)$.

Remark C.14. *In the proof of Theorem 5 [NR18], to bound the sum*

$$\sum w_{i,t} \leq \mathcal{O}(\sqrt{\log T})\sum_{t=1}^{n}\|x_{i,t}\|_{V_{i,t}^{-1}},$$

the correct bound has been used on the sum

$$\sum_{t=1}^{n}\|x_{i,t}\|_{V_{i,t}^{-1}}$$

with the $\sqrt{T}$ appearing. However, it is misused for the private part.

C.5 Existing Technical Results and Definitions

In this section, we summarise the existing technical results and definitions required to establish our proofs.

Lemma C.15 (Post-processing Lemma (Proposition 2.1, [DR14a])). *If $\mathcal{M}$ is a mechanism and f is an arbitrary randomised mapping defined on $\mathcal{M}$'s output, then*

- *If $\mathcal{M}$ is (ε, δ)-DP, then $f \circ \mathcal{M}$ is (ε, δ)-DP.*
- *If $\mathcal{M}$ is ρ-zCDP, then $f \circ \mathcal{M}$ is ρ-zCDP.*

Lemma C.16 (Post-processing property of Renyi Divergence, Lemma 2.2 [BS16]). *Let P and Q be distributions on Ω and let $f : \Omega \to \Theta$ be a function. Let $f(P)$ and $f(Q)$ denote the distributions on Θ induced by applying f to P and Q respectively. Then $\mathrm{D}_\alpha(f(P)\|f(Q)) \leq \mathrm{D}_\alpha(P\|Q)$.*

Lemma C.17 (Markov's Inequality). *For any random variable X and $\varepsilon > 0$,*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}.$$

Definition C.18 (Consistent Policies). *A policy π is called consistent over a class of bandits $\mathcal{E}$ if for all $\nu \in \mathcal{E}$ and $p > 0$, it holds that*

$$\lim_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \nu)}{T^p} = 0.$$

The class of consistent policies over $\mathcal{E}$ is denoted by $\Pi_{cons}(\mathcal{E})$.

Lemma C.19 (Divergence decomposition). *Let $\nu = (P_1, \ldots, P_K)$ and $\nu' = (P'_1, \ldots, P'_K)$ be two bandit instances. Fix some policy π and let $\mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model. Then,*

$$\mathrm{KL}\mathbb{P}_{\nu\pi}\mathbb{P}_{\nu'\pi} = \sum_{a=1}^{K} \mathbb{E}_\nu\left[N_a(T)\right] \mathrm{D}\left(P_a, P'_a\right).$$

Lemma C.20 (Bretagnolle-Huber inequality). *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$\mathbb{P}(A) + \mathbb{Q}\left(A^c\right) \geq \frac{1}{2}\exp(-\mathrm{D}(\mathbb{P}, \mathbb{Q})),$$

where $A^c = \Omega \backslash A$ is the complement of A.

Lemma C.21 (Pinsker's Inequality). *For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same probability space $(\Omega, \mathcal{F})$, we have*

$$\mathrm{KL}\mathbb{P}\mathbb{Q} \geq 2(\mathrm{TV}\left(\mathbb{P} \parallel \mathbb{Q}\right))^2.$$

Lemma C.22 (Tail Bounds for Laplacian Random Variables). *For any $a, b > 0$, we have*

$$\mathbb{P}(Lap(b) > a) = \frac{1}{2}\exp\left(-\frac{a}{b}\right) \quad \textit{and} \quad \mathbb{P}(Lap(b) < -a) = \frac{1}{2}\exp\left(-\frac{a}{b}\right).$$

Lemma C.23 (Hoeffding's Bound). *Assume that $(X_i)_{1\leq i\leq n}$ are iid random variables in $[0,1]$, with $\mathbb{E}(X_i)=\mu$. For any $\delta,\beta\geq 0$ and, we have:*

$$\mathbb{P}\left(\hat{\mu}_n \geq \mu+\beta\right) \leq \exp\left(-2n\beta^2\right) \quad \textit{and} \quad \mathbb{P}\left(\hat{\mu}_n \leq \mu-\beta\right) \leq \exp\left(-2n\beta^2\right),$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Definition C.24 (Relative entropy between Bernoulli distributions). *The relative entropy between Bernoulli distributions with parameters $p,q\in[0,1]$ is*

$$d(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q)),$$

where singularities are defined by taking limits: $d(0,q)=\log(1/(1-q))$ and $d(1,q)=\log(1/q)$ for $q\in[0,1]$ and $d(p,0)=0$ if $p=0$ and ∞ otherwise and $d(p,1)=0$ if $p=1$ and ∞ otherwise.

Lemma C.25 (Properties of the relative entropy between Bernoulli distributions (Lemma 10.2, [LS20])). *Let $p,q,\varepsilon\in[0,1]$.*

1. *The functions $d(\cdot,q)$ and $d(p,\cdot)$ are convex and have unique minimisers at q and p, respectively.*
2. *$d(p,\cdot)$ and $d(\cdot,p)$ are increasing in the interval $[p,1]$ and decreasing in the interval $[0,p]$.*

Lemma C.26 (Chernoff's Bound). *Let $X_1,X_2,\ldots,X_n$ be a sequence of independent random variables that are Bernoulli distributed with mean μ, and let $\hat{\mu}_n=\frac{1}{n}\sum_{t=1}^{n}X_t$ be the sample mean. Then, for $\beta\in[0,1-\mu]$, it holds that:*

$$\mathbb{P}(\hat{\mu}_n \geq \mu+\beta) \leq \exp(-nd(\mu+\beta,\mu)), \tag{C.28}$$

and for $\beta\in[0,\mu]$,

$$\mathbb{P}(\hat{\mu}_n \leq \mu-\beta) \leq \exp(-nd(\mu-\beta,\mu)). \tag{C.29}$$

Lemma C.27 (Theorem 7.8 of [Zha11]). *If $A\geq B\geq 0$, then*

- $\det(A)\geq\det(B)$
- *$A^{-1}\leq B^{-1}$ if A and B are non-singular.*

Lemma C.28 (Concentration of the χ^2-Distribution, Claim 17 of [SS18]). *If $X\sim\mathcal{N}(0,I_d)$ and $\delta\in(0,1)$, then*

$$\mathbb{P}\left(\|X\|^2 \geq d+2\sqrt{d\log\left(\frac{1}{\delta}\right)}+2\log\left(\frac{1}{\delta}\right)\right)\leq\delta$$

Lemma C.29 (Concentration of the Largest Singular Value, Section 4.2 of [SS18]). *If $M\in\mathbb{R}^{d\times d}$ such that $M_{i,j}\overset{iid}{\sim}\mathcal{N}(0,1)$, $\|M\|\triangleq$ the largest singular value of M and $\delta\in(0,1)$, then*

$$\mathbb{P}\left(\|M\| > 4\sqrt{d+1}+2\log\left(\frac{1}{\delta}\right)\right)\leq\delta$$

Lemma C.30 (Theorem 20.4 of [LS20]). *Let the noise ρ_t be conditionally 1-subgaussian (conditioned on $A_1, X_1, \ldots, A_{t-1}, X_{t-1}, A_t$), $S_t = \sum_{s=1}^t A_s \rho_s$ and $V_t(\lambda) = \lambda I_d + \sum_{s=1}^t A_s A_s^T$. Then, for all $\lambda > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \|S_t\|^2_{V_t(\lambda)^{-1}} \geq 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(\lambda))}{\lambda^d}\right)\right) \leq \delta$$

Lemma C.31 (Lemma 12 in [AYPS11]). *Let A, B and C be positive semi-definite matrices such that $A = B + C$. Then, we have that*

$$\sup_{x \neq 0} \frac{x^T A x}{x^T B x} \leq \frac{\det(A)}{\det(B)}$$

Lemma C.32 (Theorem 9 in [KK21]). *Let $\boldsymbol{\nu}$ be a sub-Gaussian bandit with means $\boldsymbol{\mu} \in \mathbb{R}^K$ and variance proxy σ. Let $S \subseteq [K]$ and $x > 0$.*

$$\mathbb{P}_{\boldsymbol{\nu}}\left(\exists n \in \mathbb{N}, \sum_{a \in S} \frac{N_{n,a}}{2\sigma^2}(\mu_{n,a} - \mu_a)^2 > \sum_{a \in S} 2\log\left(4 + \log\left(N_{n,a}\right)\right) + |S|\mathcal{C}_G\left(\frac{x}{|S|}\right)\right) \leq e^{-x},$$

where $\mathcal{C}_G$ is defined in [KK21] as

$$\mathcal{C}_G(x) \triangleq \min_{\lambda \in]1/2,1]} \frac{g_G(\lambda) + x}{\lambda} \text{ and } g_G(\lambda) \triangleq 2\lambda - 2\lambda\log(4\lambda) + \log\zeta(2\lambda) - \frac{1}{2}\log(1-\lambda). \quad \text{(C.30)}$$

Here, ζ is the Riemann ζ function and $\mathcal{C}_G(x) \approx x + \log(x)$.

Appendix D

Supplementary for Chapter 6

Contents

D.1 Proof of Theorem 6.3

Theorem 6.3 (Asymptotic distribution of the LR score)**.** *Using an Edgeworth asymptotic expansion of the likelihood ratio score and a Lindeberg-Feller central limit theorem, we show that*

(*a*) *Under* H_0,

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{1}{2}m^\star, m^\star\right)$$

(*b*) *Under* H_1,

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{1}{2}m^\star, m^\star\right)$$

The convergence is a convergence in distribution, such that $d, n \to \infty$*, while* $d/n = \tau$*. We call*

$$m^\star \triangleq \lim_{n,d} \frac{1}{n}\|z^\star - \mu\|^2_{C_\sigma^{-1}} = \lim_{n,d} \sum_{j=1}^{d} \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2}$$

the leakage score *of target* $z^\star$*.*

Proof. We have that $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, where $Z_i = (Z_{i,j})_{j=1}^{d_n} \in \mathbb{R}^{d_n}$ and $Z_i \sim^{\text{i.i.d}} \mathcal{D} = \bigotimes_{j=1}^{d_n} \mathcal{D}_j$.

Each distribution $\mathcal{D}_j$ has mean μ_j and variance σ_j^2.

We denote $\hat{\mu}_n = (\hat{\mu}_{n,j})_{j=1}^{d_n}$, where $\hat{\mu}_{n,j} = \frac{1}{n}\sum_{i=1}^n Z_{i,j}$.

Step 1: Rewriting the LR score

Let $j \in [1, d_n]$.

Under H_0, we can re-write

$$\hat{\mu}_{n,j} = \mu_j + \frac{\sigma_j}{\sqrt{n}}\hat{Z}_{n,j},$$

where

$$\begin{aligned}\hat{Z}_{n,j} &\triangleq \sqrt{n}\left(\frac{\hat{\mu}_{n,j} - \mu_j}{\sigma}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Z_{i,j} - \mu_j}{\sigma_j}.\end{aligned}$$

Since $(Z_{i,j})_{i=1}^n$ are i.i.d from $\mathcal{D}_j$, using the CLT, $\hat{Z}_{n,j} \rightsquigarrow_{n\to\infty} \mathcal{N}(0,1)$.

Let $d_{n,j}$ be the density function of $\hat{Z}_{n,j}$.

The density $p_{n,j}^{\text{out}}$ of $\hat{\mu}_{n,j}$ under H_0 can be written as

$$p_{n,j}^{\text{out}}(x; z_j^\star, \mu_j, \sigma_j) = \frac{\sqrt{n}}{\sigma_j} d_{n,j}\left[\frac{\sqrt{n}}{\sigma_j}(x - \mu_j)\right]$$

Under H_1, we can re-write

$$\begin{aligned}\hat{\mu}_{n,j} &= \frac{1}{n} z_j^\star + \frac{n-1}{n}\left(\mu_j + \frac{\sigma_j}{\sqrt{n-1}}\hat{Z}_{n-1,j}\right) \\ &= \mu_j + \frac{1}{n}\left(z_j^\star - \mu_j\right) + \frac{\sigma_j\sqrt{n-1}}{n}\hat{Z}_{n-1,j}\end{aligned}$$

The density $p_{n,j}^{\text{in}}$ of $\hat{\mu}_{n,j}$ under H_1 can be written as

$$p_{n,j}^{\text{in}}(x; z_j^\star, \mu_j, \sigma_j) = \frac{n}{\sigma_j\sqrt{n-1}} d_{n-1,j}\left[\frac{n}{\sigma_j\sqrt{n-1}}\left(x - \mu_j - \frac{1}{n}\left(z_j^\star - \mu_j\right)\right)\right]$$

The LR score is

$$\begin{aligned}\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) &= \sum_{j=1}^{d_n} \log\left(\frac{p_{n,j}^{\text{in}}(\hat{\mu}_{n,j}; z_j^\star, \mu_j, \sigma_j)}{p_{n,j}^{\text{out}}(\hat{\mu}_{n,j}; z_j^\star, \mu_j, \sigma_j)}\right) \\ &= \sum_{j=1}^{d_n} -\frac{1}{2}\log\left(1 - \frac{1}{n}\right) + \log\left(\frac{d_{n-1,j}\left(\delta_{n,j}^{\text{in}}\right)}{d_{n,j}\left(\delta_{n,j}^{\text{out}}\right)}\right)\end{aligned}$$

where

$$\begin{aligned}\delta_{n,j}^{\text{out}} &\triangleq \frac{\sqrt{n}}{\sigma_j}\left(\hat{\mu}_{n,j} - \mu_j\right) \\ \delta_{n,j}^{\text{in}} &\triangleq \frac{n}{\sqrt{n-1}\sigma_j}\left(\hat{\mu}_{n,j} - \mu_j + \frac{1}{n}\left(\mu_j - z_j^\star\right)\right)\end{aligned}$$

Step 2: Asymptotic expansion of the LR score

Using Lemma D.2, we have

$$\log\left(\frac{d_{n-1,j}\left(\delta_{n,j}^{\text{in}}\right)}{d_{n,j}\left(\delta_{n,j}^{\text{out}}\right)}\right) = \frac{1}{2}\left(\left(\delta_{n,j}^{\text{out}}\right)^2 - \left(\delta_{n,j}^{\text{in}}\right)^2\right) + \frac{\lambda_{3,j}\left(\mu_j - z_j^\star\right)}{n\sigma_j} R_{n,j} + o_p\left(\frac{1}{n}\right)$$

where $\lambda_{k,j} \triangleq \frac{\gamma_{j,k}}{\sigma_j^k}$ s.t. $\gamma_{j,k}$ is the k-order cumulant of distribution $\mathcal{D}_j$.

Let $Y_{n,j} \triangleq \frac{1}{2}\left(\left(\delta_{n,j}^{\text{out}}\right)^2 - \left(\delta_{n,j}^{\text{in}}\right)^2\right) + \frac{\lambda_{3,j}\left(\mu_j - z_j^\star\right)}{n\sigma_j} R_{n,j}$.

We remark that we need an expansion up to $o_p\left(\frac{1}{n}\right)$, since $d_n/n = \tau + o(1)$.

Thus

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) = \sum_{j=1}^{d_n} -\frac{1}{2}\log\left(1 - \frac{1}{n}\right) + \log\left(\frac{d_{n-1,j}\left(\delta_{n,j}^{\text{in}}\right)}{d_{n,j}\left(\delta_{n,j}^{\text{out}}\right)}\right)$$

$$= \sum_{j=1}^{d_n} \left(\frac{1}{2n} + Y_{n,j} + o_p\left(\frac{1}{n}\right) \right)$$
$$= \frac{\tau}{2} + o_p(1) + \sum_{j=1}^{d_n} Y_{n,j} \tag{D.1}$$

because $\frac{d_n}{n} = \tau + o(1)$.

Step3: Concluding using the Lindeberg-Feller CLT

Under H_0:

Using Lemma D.3, $\mathbb{E}_0[Y_n, j] = -\frac{1}{2n} - \frac{(z_j^\star - \mu_j)^2}{2n\sigma_j^2} + o\left(\frac{1}{n}\right)$ and $V_0[Y_n, j] = \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$.

Since $\sum_{j=1}^{d_n} \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} = \frac{\|z^\star - \mu\|^2_{C_\sigma^{-1}}}{n}$, we get:

- $\sum_{j=1}^{d_n} \mathbb{E}_0[Y_n, j] \to -\frac{\tau}{2} - \frac{m^\star}{2}$
- $\sum_{j=1}^{d_n} V_0[Y_n, j] \to m^\star$

Using Lemma D.4, we have that $Y_{n,j}$ verify the Lindeberg-Feller condition, i.e.

$$\sum_{j=1}^{d_n} \mathbb{E}_0 \left[Y_{n,j}^2 \mathbb{1}\left(|Y_{n,j}| > \varepsilon\right) \right] \to 0$$

for every $\varepsilon > 0$.

We conclude using the Lindeberg-Feller CLT that $\sum_{j=1}^{d_n} Y_{n,j} \rightsquigarrow \mathcal{N}\left(-\frac{\tau}{2} - \frac{m^\star}{2}, m^\star\right)$, and thus

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{m^\star}{2}, m^\star\right)$$

Similarly, *Under H_1:*

Using Lemma D.3, $\mathbb{E}_1[Y_n, j] = -\frac{1}{2n} + \frac{(z_j^\star - \mu_j)^2}{2n\sigma_j^2} + o\left(\frac{1}{n}\right)$ and $V_1[Y_n, j] = \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$.

We get:

- $\sum_{j=1}^{d_n} \mathbb{E}_1[Y_{n,j}] \to -\frac{\tau}{2} + \frac{m^\star}{2}$
- $\sum_{j=1}^{d_n} V_1[Y_{n,j}] \to m^\star$

Using Lemma D.4, we have that $Y_{n,j}$ verify the Lindeberg-Feller condition, i.e.

$$\sum_{j=1}^{d_n} \mathbb{E}_1 \left[Y_{n,j}^2 \mathbb{1}\left(|Y_{n,j}| > \varepsilon\right) \right] \to 0$$

for every $\varepsilon > 0$.

We conclude using the Lindeberg-Feller CLT that $\sum_{j=1}^{d_n} Y_{n,j} \rightsquigarrow \mathcal{N}\left(-\frac{\tau}{2} + \frac{m^\star}{2}, m^\star\right)$, and thus

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{m^\star}{2}, m^\star\right)$$

□

Remark D.1. *Expanding* $\frac{1}{2}\left(\left(\delta_{n,j}^{\text{out}}\right)^2 - \left(\delta_{n,j}^{\text{in}}\right)^2\right)$*, taking the sum from* $j = 1$ *until* d_n*, we get that*

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) \sim (z^\star - \mu)^T C_\sigma^{-1}(\hat{\mu}_n - \mu) - \frac{1}{2n}\|z^\star - \mu\|^2_{C_\sigma^{-1}}$$

Let $X_n \triangleq (z^\star - \mu)^T C_\sigma^{-1}(\hat{\mu}_n - \mu) - \frac{1}{2n}\|z^\star - \mu\|^2_{C_\sigma^{-1}}$.

This asymptotic representation of the LR test is useful to get directly the means and variances of the limit distribution of the LR test. Specifically, since $\mathbb{E}_0(\hat{\mu}_n) = \mu$, $\mathbb{E}_1(\hat{\mu}_n) = \frac{n-1}{n}\mu + \frac{1}{n}z^\star$ *and* $\mathbb{V}_0(\hat{\mu}_n) = \mathbb{V}_1(\hat{\mu}_n) = C_\sigma$*, we get that*

$$\begin{aligned}
\mathbb{E}_0[X_n] &= -\frac{1}{2n}\|z^\star - \mu\|^2_{C_\sigma^{-1}} \\
\mathbb{E}_1[X_n] &= \frac{1}{2n}\|z^\star - \mu\|^2_{C_\sigma^{-1}} \\
\mathbb{V}_0[X_n] &= \mathbb{V}_1[X_n] = \frac{1}{n}\|z^\star - \mu\|^2_{C_\sigma^{-1}}
\end{aligned}$$

Taking the limit as $n \to \infty$ *retrieves the results of Theorem 6.3.*

D.2 The Three Technical Lemmas Used in the Proof of Theorem 6.3

Lemma D.2. *Asymptotic expansion of the LR score*

We show that

$$\log\left(\frac{d_{n-1,j}\left(\delta_{n,j}^{\text{in}}\right)}{d_{n,j}\left(\delta_{n,j}^{\text{out}}\right)}\right) = \frac{1}{2}\left(\left(\delta_{n,j}^{\text{out}}\right)^2 - \left(\delta_{n,j}^{\text{in}}\right)^2\right) + \frac{\lambda_{3,j}\left(\mu_j - z_j^\star\right)}{n\sigma_j}R_{n,j} + o_p\left(\frac{1}{n}\right)$$

where $\delta_{n,j}^{\text{out}} \triangleq \frac{\sqrt{n}}{\sigma_j}(\hat{\mu}_{n,j} - \mu_j)$, $\delta_{n,j}^{\text{in}} \triangleq \frac{n}{\sqrt{n-1}\sigma_j}\left(\hat{\mu}_{n,j} - \mu_j + \frac{1}{n}\left(\mu_j - z_j^\star\right)\right)$, $\lambda_{k,j} \triangleq \frac{\gamma_{j,k}}{\sigma_j^k}$ *where* $\gamma_{j,k}$ *is the* k*-order cumulant of distribution* $\mathcal{D}_j$ *and* $R_{n,j} \triangleq \left(\delta_{n,j}^{\text{out}}\right)^2 + \delta_{n,j}^{\text{out}}\delta_{n,j}^{\text{in}} + \left(\delta_{n,j}^{\text{in}}\right)^2 - 3$.

Proof Sketch. The proof starts by using the Edgeworth expansion of $d_{n,j}$ up to the order $k = 4$. Then, using Taylor expansions of the logarithm, exponential and polynomial function to the 2nd order, the final LR expansion can be found. We present the exact derivations in Appendix C.3 of [AB24b].

Lemma D.3. *Expectation and variance computations*

$$\mathbb{E}_0[Y_{n,j}] = -\frac{1}{2n} - \frac{(z_j^\star - \mu_j)^2}{2n\sigma_j^2} + o\left(\frac{1}{n}\right) \qquad V_0[Y_{n,j}] = \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$$
$$\mathbb{E}_1[Y_{n,j}] = -\frac{1}{2n} + \frac{(z_j^\star - \mu_j)^2}{2n\sigma_j^2} + o\left(\frac{1}{n}\right) \qquad V_1[Y_{n,j}] = \frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$$

Proof Sketch. The proof is direct from expectation and variance of the mean under H_0 and H_1. Specifically, under H_0 we have that $\mathbb{E}_0(\hat{\mu}_{n,j}) = \mu_j$ and $\mathbb{V}_0(\hat{\mu}_{n,j}) = \frac{1}{n}\sigma_j^2$. On the other hand, under H_1, we have that $\mathbb{E}_1(\hat{\mu}_{n,j}) = \mu_j + \frac{1}{n}\left(z_j^\star - \mu_j\right)$ and $\mathbb{V}_1(\hat{\mu}_{n,j}) = \frac{n-1}{n^2}\sigma_j^2$. We present the exact derivations in Appendix C.3 of [AB24b].

Lemma D.4. *The Lindeberg-Feller condition*

The random variables $(Y_{n,j})_{j=1}^{d_n}$ verify the Lindeberg-Feller condition.

Proof. Let $\varepsilon > 0$, $h \in \{0, 1\}$ and $\delta > 0$. We have that

$$\begin{aligned}\mathbb{E}_h\left[Y_{n,j}^2 \mathbb{1}\left(|Y_{n,j}| > \varepsilon\right)\right] &= \mathbb{E}_h\left[\frac{Y_{n,j}^{2+\delta}}{Y_{n,j}^{\delta}} \mathbb{1}\left(|Y_{n,j}| > \varepsilon\right)\right] \\ &\leq \frac{1}{\varepsilon^\delta}\,\mathbb{E}_h\left[Y_{n,j}^{2+\delta}\right]\end{aligned}$$

On the other hand, we have that $Y_{n,j} = \frac{1}{2}\left(\left(\delta_{n,j}^{\text{out}}\right)^2 - \left(\delta_{n,j}^{\text{in}}\right)^2\right) + \frac{\lambda_{3,j}\left(\mu_j - z_j^\star\right)}{n\sigma_j} R_{n,j}$, where

$$\left(\left(\delta_{n,j}^{\text{out}}\right)^2 - \left(\delta_{n,j}^{\text{in}}\right)^2\right) = O_p\left(\frac{1}{\sqrt{n}}\right) \text{ and } R_{n,j} = O_p(1)$$

Thus $Y_{n,j} = O_p\left(\frac{1}{\sqrt{n}}\right)$, and $\mathbb{E}_h\left[Y_{n,j}^{2+\delta}\right] = o\left(\frac{1}{n}\right)$.

Which means that $\mathbb{E}_h\left[Y_{n,j}^2 \mathbb{1}\left(|Y_{n,j}| > \varepsilon\right)\right] = o\left(\frac{1}{n}\right)$ and

$$\sum_{j=1}^{d_n} \mathbb{E}_h\left[Y_{n,j}^2 \mathbb{1}\left(|Y_{n,j}| > \varepsilon\right)\right] = o\left(\frac{d_n}{n}\right) = o(1) \to 0$$

□

D.3 Effect of Sub-sampling, Proof of Theorem 6.6

Theorem 6.6 (Target-dependent leakage of the sub-sampling empirical mean)**.** *As* $d, n \to \infty$ *s.t.* $d/n = \tau$, $\ell_n^{\mathrm{sub},\rho}(\hat{\mu}_n^{\mathrm{sub}}; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{\rho m^\star}{2}, \rho m^\star\right)$ *under* H_0, $\ell_n^{\mathrm{sub},\rho}(\hat{\mu}_n^{\mathrm{sub}}; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{\rho m^\star}{2}, \rho m^\star\right)$ *under* H_1.

The asymptotic target-dependent leakage of $z^\star$ *in* $\mathcal{M}_n^{\mathrm{sub},\rho}$ *is*

$$\lim_{n,d} \xi_n(z^\star, \mathcal{M}_n^{\mathrm{sub},\rho}, \mathcal{D}) = \Phi\left(\frac{\sqrt{\rho m^\star}}{2}\right) - \Phi\left(-\frac{\sqrt{\rho m^\star}}{2}\right).$$

The optimal trade-off function obtained with $\tau_\alpha = -\frac{\rho m^\star}{2} + \sqrt{\rho m^\star}\Phi^{-1}(1-\alpha)$, *is*

$$\lim_{n,d} \mathrm{Pow}_n(\ell_n^{\mathrm{sub},\rho}, \alpha, z^\star) = \Phi\left(z_\alpha + \sqrt{\rho m^\star}\right).$$

Proof. We have that $\hat{\mu}_n^{\mathrm{sub}} = \frac{1}{k_n}\sum_{i=1}^n Z_i \mathbb{1}\left(\varsigma(i) \leq k_n\right)$, where $k_n \triangleq \rho n$, Z_i are i.i.d and $\varsigma \sim^{\mathrm{unif}} S_n$ is a permutation sampled uniformly from the set of permutations of $\{1 \dots, n\}$ i.e. S_n and independently from (Z_i).

We denote $\hat{\mu}_n^{\mathrm{sub}} = (\hat{\mu}_{n,j}^{\mathrm{sub}})_{j=1}^{d_n}$.

Step 1: Rewriting the LR score

Let $j \in [1, d_n]$.

Under H_0, we can re-write

$$\hat{\mu}_{n,j}^{\mathrm{sub}} = \mu_j + \frac{\sigma_j}{\sqrt{k_n}}\hat{Z}_{k_n,j},$$

where

$$\begin{aligned}\hat{Z}_{d_n,j} &\triangleq \sqrt{k_n}\left(\frac{\hat{\mu}_{n,j}^{\mathrm{sub}} - \mu_j}{\sigma}\right) \\ &= \frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n}\frac{Z_{\varsigma^{-1}(i),j} - \mu_j}{\sigma_j}.\end{aligned}$$

Since $(Z_{i,j})_{i=1}^n$ are i.i.d from $\mathcal{D}_j$, and $\varsigma \sim^{\mathrm{unif}} S_n$ and ind. from (Z_i), then $(Z_{\varsigma^{-1}(i),j})_{i=1}^{k_n}$.

Using the CLT, $\hat{Z}_{d_n,j} \rightsquigarrow_{n\to\infty} \mathcal{N}(0,1)$.

Let $d_{n,j}$ be the density function of $\hat{Z}_{n,j}$.

The density $p_{n,j}^{\mathrm{out,sub}}$ of $\hat{\mu}_{n,j}^{\mathrm{sub}}$ under H_0 can be written as

$$p_{n,j}^{\mathrm{out,sub}}(x; z_j^\star, \mu_j, \sigma_j) = \frac{\sqrt{k_n}}{\sigma_j} d_{k_n,j}\left[\frac{\sqrt{k_n}}{\sigma_j}(x - \mu_j)\right]$$

Under H_1, we can re-write

$$\hat{\mu}_{n,j}^{\text{sub}} = \frac{1}{k_n} z_j^\star \mathbb{1}\left(\varsigma(n) \leq k_n\right) + \frac{1}{k_n} \sum_{i=1}^{n-1} Z_i \mathbb{1}\left(\varsigma(i) \leq k_n\right)$$

Let $A = \{\mathbb{1}\left(\varsigma(n) \leq k_n\right)\}$ the event that $z^\star$ was sub-sampled. We have that $\Pr(A) = \rho$. The density $p_{n,j}^{\text{in,sub}}$ of $\hat{\mu}_{n,j}^{\text{sub}}$ under H_1 and given A is

$$\frac{k_n}{\sigma_j \sqrt{k_n - 1}} d_{k_n - 1, j} \left[\frac{k_n}{\sigma_j \sqrt{k_n - 1}} \left(x - \mu_j - \frac{1}{k_n} \left(z_j^\star - \mu_j \right) \right) \right]$$

The density $p_{n,j}^{\text{in,sub}}$ of $\hat{\mu}_{n,j}^{\text{sub}}$ under H_1 and given A^c is

$$\frac{\sqrt{k_n}}{\sigma_j} d_{k_n, j} \left[\frac{\sqrt{k_n}}{\sigma_j} (x - \mu_j) \right]$$

Thus, the density $p_{n,j}^{\text{in,sub}}$ of $\hat{\mu}_{n,j}^{\text{sub}}$ under H_1 can be written as

$$\begin{aligned} p_{n,j}^{\text{in,sub}}(x; z_j^\star, \mu_j, \sigma_j) = (1-\rho) \frac{\sqrt{k_n}}{\sigma_j} d_{k_n, j} \left[\frac{\sqrt{k_n}}{\sigma_j} (x - \mu_j) \right] \\ + \rho \frac{k_n}{\sigma_j \sqrt{k_n - 1}} d_{k_n - 1, j} \left[\frac{k_n}{\sigma_j \sqrt{k_n - 1}} \left(x - \mu_j - \frac{1}{k_n} \left(z_j^\star - \mu_j \right) \right) \right] \end{aligned}$$

The additional technical hardness of this proof comes from the 'mixture' nature of the 'in' distribution.

The LR score is

$$\begin{aligned} \ell_n^{\text{sub},\rho}(\hat{\mu}_n^{\text{sub}}; z^\star, \mu, C_\sigma) &= \sum_{j=1}^{d_n} \log \left(\frac{p_{n,j}^{\text{in,sub}}(\hat{\mu}_{n,j}^{\text{sub}}; z_j^\star, \mu_j, \sigma_j)}{p_{n,j}^{\text{out,sub}}(\hat{\mu}_{n,j}^{\text{sub}}; z_j^\star, \mu_j, \sigma_j)} \right) \\ &= \sum_{j=1}^{d_n} \log \left(\frac{(1-\rho) \frac{\sqrt{k_n}}{\sigma_j} d_{k_n, j} \left(\delta_{k_n, j}^{\text{out,sub}} \right) + \rho \frac{k_n}{\sigma_j \sqrt{k_n - 1}} d_{k_n - 1, j} \left(\delta_{k_n, j}^{\text{in,sub}} \right)}{\frac{\sqrt{k_n}}{\sigma_j} d_{k_n, j} \left(\delta_{k_n, j}^{\text{out,sub}} \right)} \right) \\ &= \sum_{j=1}^{d_n} \log \left((1-\rho) + \rho \sqrt{\frac{k_n}{k_n - 1}} \frac{d_{k_n - 1, j} \left(\delta_{k_n, j}^{\text{in,sub}} \right)}{d_{k_n, j} \left(\delta_{k_n, j}^{\text{out,sub}} \right)} \right) \end{aligned}$$

where

$$\delta_{k_n, j}^{\text{out,sub}} \triangleq \frac{\sqrt{k_n}}{\sigma_j} \left(\hat{\mu}_{n,j}^{\text{sub}} - \mu_j \right)$$

$$\delta_{k_n,j}^{\text{in,sub}} \triangleq \frac{k_n}{\sqrt{k_n-1}\sigma_j}\left(\hat{\mu}_{n,j}^{\text{sub}} - \mu_j + \frac{1}{k_n}\left(\mu_j - z_j^\star\right)\right)$$

Step 2: Asymptotic expansion of the LR score

Using Lemma D.5, we have

$$\log\left((1-\rho) + \rho\sqrt{\frac{k_n}{k_n-1}}\frac{d_{k_n-1,j}\left(\delta_{k_n,j}^{\text{in,sub}}\right)}{d_{k_n,j}\left(\delta_{k_n,j}^{\text{out,sub}}\right)}\right) = W_{n,j} + o_p\left(\frac{1}{n}\right)$$

where

$$W_{n,j} \triangleq \frac{\rho}{2}\left((\delta_{k_n,j}^{\text{out,sub}})^2 - (\delta_{k_n,j}^{\text{in,sub}})^2\right) + \frac{\rho}{2k_n} + \frac{\rho(1-\rho)}{8}\left((\delta_{k_n,j}^{\text{out,sub}})^2 - (\delta_{k_n,j}^{\text{in,sub}})^2\right)^2 + \rho\frac{\lambda_{3,j}\left(\mu_j - z_j^\star\right)}{k_n\sigma_j}R_{k_n,j}$$

The extra hardness of this proof comes from the fact that the density under H_1 is now a mixture of two Gaussians, rather than just one Gaussian in the case of the exact empirical mean.

Thus

$$\ell_n^{\text{sub},\rho}(\hat{\mu}_n^{\text{sub}}; z^\star, \mu, C_\sigma) = o_p(1) + \sum_{j=1}^{d_n} W_{n,j}$$

because $\frac{d_n}{n} = \tau + o(1)$.

Step3: Concluding using the Lindeberg-Feller CLT

Under H_0:

Using Lemma D.6, $\mathbb{E}_0[W_n, j] = -\frac{\rho}{2}\frac{\left(\mu_j - z_j^\star\right)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$ and $\mathbb{V}_0[W_{n,j}] = \rho\frac{(\mu_j - z_j^\star)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$.

Since $\sum_{j=1}^{d_n}\frac{(z_j^\star-\mu_j)^2}{n\sigma_j^2} = \frac{\|z^\star-\mu\|^2_{C_\sigma^{-1}}}{n}$, we get:

- $\sum_{j=1}^{d_n}\mathbb{E}_0[W_n, j] \to -\frac{\rho m^\star}{2}$
- $\sum_{j=1}^{d_n} V_0[W_n, j] \to \rho m^\star$

Similarly to Lemma D.4, we can show that $W_{n,j}$ verify the Lindeberg-Feller condition, i.e.

$$\sum_{j=1}^{d_n}\mathbb{E}_0\left[W_{n,j}^2\mathbb{1}\left(|W_{n,j}| > \varepsilon\right)\right] \to 0$$

for every $\varepsilon > 0$.

We conclude using the Lindeberg-Feller CLT that

$$\ell_n^{\mathrm{sub},\rho}(\hat{\mu}_n^{\mathrm{sub}}; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\rho\frac{m^\star}{2}, \rho m^\star\right)$$

Similarly, *Under H_1:*

Using Lemma D.6, $\mathbb{E}_1[W_n, j] = \frac{\rho}{2}\frac{\left(\mu_j - z_j^\star\right)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$ and $\mathbb{V}_1[W_{n,j}] = \rho\frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$.

We get:

- $\sum_{j=1}^{d_n} \mathbb{E}_1[W_n, j] \to \frac{\rho m^\star}{2}$
- $\sum_{j=1}^{d_n} V_1[W_n, j] \to \rho m^\star$

Similarly to Lemma D.4, we can show that $W_{n,j}$ verify the Lindeberg-Feller condition, i.e.

$$\sum_{j=1}^{d_n} \mathbb{E}_1\left[W_{n,j}^2 \mathbb{1}\left(|W_{n,j}| > \varepsilon\right)\right] \to 0$$

for every $\varepsilon > 0$.

We conclude using the Lindeberg-Feller CLT that

$$\ell_n^{\mathrm{sub},\rho}(\hat{\mu}_n^{\mathrm{sub}}; z^\star, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\rho\frac{m^\star}{2}, \rho m^\star\right)$$

Step4: Characterising the advantage and the power function

Using the same step as in the proof of Corollary 6.4, we conclude. □

Now we present the helpful technical lemmas. Similarly to Section D.2 the proofs are only computational, and the details can be found at the end of Appendix C.5 in [AB24b].

Lemma D.5 (Asymptotic expansion of the LR score for sub-sampling)**.** *We show that*

$$\log\left((1-\rho) + \rho\sqrt{\frac{k_n}{k_n - 1}}\frac{d_{k_n-1,j}\left(\delta_{k_n,j}^{\mathrm{in,sub}}\right)}{d_{k_n,j}\left(\delta_{k_n,j}^{\mathrm{out,sub}}\right)}\right) = W_{n,j} + o_p\left(\frac{1}{n}\right)$$

where

$$W_{n,j} \triangleq \frac{\rho}{2}\left((\delta_{k_n,j}^{\mathrm{out,sub}})^2 - (\delta_{k_n,j}^{\mathrm{in,sub}})^2\right) + \frac{\rho}{2k_n} + \frac{\rho(1-\rho)}{8}\left((\delta_{k_n,j}^{\mathrm{out,sub}})^2 - (\delta_{k_n,j}^{\mathrm{in,sub}})^2\right)^2 + \rho\frac{\lambda_{3,j}\left(\mu_j - z_j^\star\right)}{k_n\sigma_j}R_{k_n,j}$$

Lemma D.6. *Expectation and variance computations for sub-sampling*

$$\mathbb{E}_0[W_{n,j}] = -\frac{\rho}{2}\frac{\left(\mu_j - z_j^\star\right)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right) \qquad \mathbb{V}_0[W_{n,j}] = \rho\frac{(\mu_j - z_j^\star)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$$

$$\mathbb{E}_1[W_{n,j}] = \frac{\rho}{2}\frac{\left(\mu_j - z_j^\star\right)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right) \qquad \mathbb{V}_1[W_{n,j}] = \rho\frac{(z_j^\star - \mu_j)^2}{n\sigma_j^2} + o\left(\frac{1}{n}\right)$$

D.4 Effect of Misspecifiaction, Proof of Theorem 6.7

Theorem 6.7 (Leakage of a misspecified adversary)**.** *We show that as $d, n \leftarrow \infty$ while $d/n = \tau$,*

(*a*) *Under H_0,*

$$\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{m^{\text{targ}}}{2}, m^\star\right)$$

(*b*) *Under H_1,*

$$\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{m^\star - m^{\text{diff}}}{2}, m^\star\right)$$

where $m^{\text{diff}} \triangleq \lim_{n,d} \frac{1}{n}\|z^\star - z^{\text{targ}}\|^2_{C_\sigma^{-1}} = \lim_{n,d} \sum_{j=1}^d \frac{(z_j^\star - z_j^{\text{targ}})^2}{n\sigma_j^2}$.

Let $\mathcal{A}_{\text{miss}}$ the adversary that uses the misspecified LR score $\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma)$. Then,

$$\lim_{n,d} \text{Adv}_n(\mathcal{A}_{\text{miss}}) = \Phi\left(\frac{|m^{\text{scal}}|}{2\sqrt{m^{\text{targ}}}}\right) - \Phi\left(-\frac{|m^{\text{scal}}|}{2\sqrt{m^{\text{targ}}}}\right).$$

Here, $m^{\text{scal}} \triangleq \lim_{n,d} \frac{1}{n}\left(z^{\text{targ}} - \mu\right)^T C_\sigma^{-1}\left(z^\star - \mu\right)$ and $m^{\text{targ}} \triangleq \lim_{n,d} \frac{1}{n}\|z^{\text{targ}} - \mu\|^2_{C_\sigma^{-1}}$.

Proof. **Step 1: Asymptotic expansion of the LR score**

Directly using Equation (D.1) from the proof in Section D.1, by only replacing $z^\star$ by z^{targ}, we get

$$\ell_n(\hat{\mu}_n; z^\star, \mu, C_\sigma) = \frac{\tau}{2} + o_p(1) + \sum_{j=1}^{d_n} Y_{n,j}^{\text{targ}}$$

where

$$Y_{n,j}^{\text{targ}} \triangleq \frac{1}{2}\left(\left(\delta_{n,j}^{\text{out,targ}}\right)^2 - \left(\delta_{n,j}^{\text{in,targ}}\right)^2\right) + \frac{\lambda_{3,j}\left(\mu_j - z_j^{\text{targ}}\right)}{n\sigma_j} R_{n,j}^{\text{targ}},$$

and

$$\begin{aligned}
\delta_{n,j}^{\text{out,targ}} &\triangleq \frac{\sqrt{n}}{\sigma_j}\left(\hat{\mu}_{n,j} - \mu_j\right) \\
\delta_{n,j}^{\text{in,targ}} &\triangleq \frac{n}{\sqrt{n-1}\sigma_j}\left(\hat{\mu}_{n,j} - \mu_j + \frac{1}{n}\left(\mu_j - z_j^{\text{targ}}\right)\right) \\
R_{n,j}^{\text{targ}} &\triangleq \left(\delta_{n,j}^{\text{out,targ}}\right)^2 + \delta_{n,j}^{\text{out,targ}}\delta_{n,j}^{\text{in,targ}} + \left(\delta_{n,j}^{\text{in,targ}}\right)^2 - 3
\end{aligned}$$

Step 2: Computing expectations and variances

This is the step where the effect of misspecification appears.

Computing the expectations and variances under H_0 and H_1 gives that

$$\begin{aligned}
\mathbb{E}_0[Y_{n,j}^{\text{targ}}] &= -\frac{1}{2n} - \frac{(z_j^{\text{targ}} - \mu_j)^2}{2n\sigma_j^2} + o\left(\frac{1}{n}\right) \\
\mathbb{V}_0[Y_{n,j}^{\text{targ}}] &= \frac{1}{n}\frac{(\mu_j - z_j^{\text{targ}})^2}{\sigma_j^2} + o\left(\frac{1}{n}\right) \\
\mathbb{E}_1[Y_{n,j}^{\text{targ}}] &= -\frac{1}{2n} + \frac{(z_j^\star - \mu_j)^2 - \left(z_j^\star - z_j^{\text{targ}}\right)^2}{2n\sigma_j^2} + o\left(\frac{1}{n}\right) \\
\mathbb{V}_1[Y_{n,j}^{\text{targ}}] &= \frac{1}{n}\frac{(\mu_j - z_j^{\text{targ}})^2}{\sigma_j^2} + o\left(\frac{1}{n}\right)
\end{aligned}$$

Step3: Concluding using the Lindeberg-Feller CLT

Under H_0:

Using the results of Step2, we have

- $\sum_{j=1}^{d_n} \mathbb{E}_0[Y_n^{\text{targ}}, j] \to -\frac{\tau}{2} - \frac{m^{\text{targ}}}{2}$
- $\sum_{j=1}^{d_n} V_0[Y_n, j] \to m^{\text{targ}}$

Similarly to Lemma D.4, we can show that $Y_{n,j}^{\text{targ}}$ verify the Lindeberg-Feller condition, i.e.

$$\sum_{j=1}^{d_n} \mathbb{E}_0\left[(Y_{n,j}^{\text{targ}})^2 \mathbb{1}\left(|Y_{n,j}^{\text{targ}}| > \varepsilon\right)\right] \to 0$$

for every $\varepsilon > 0$.

We conclude using the Lindeberg-Feller CLT that $\sum_{j=1}^{d_n} Y_{n,j}^{\text{targ}} \rightsquigarrow \mathcal{N}\left(-\frac{\tau}{2} - \frac{m^{\text{targ}}}{2}, m^{\text{targ}}\right)$, and thus

$$\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(-\frac{m^{\text{targ}}}{2}, m^{\text{targ}}\right)$$

Similarly, *Under H_1:*

Use the results of Step2, we get

- $\sum_{j=1}^{d_n} \mathbb{E}_1[Y_{n,j}^{\text{targ}}] \to -\frac{\tau}{2} + \frac{m^\star - m^{\text{diff}}}{2}$
- $\sum_{j=1}^{d_n} V_1[Y_{n,j}^{\text{targ}}] \to m^{\text{targ}}$

Similarly to Lemma D.4, we can show that $Y_{n,j}^{\text{targ}}$ verify the Lindeberg-Feller condition, i.e.

$$\sum_{j=1}^{d_n} \mathbb{E}_1\left[(Y_{n,j}^{\text{targ}})^2 \mathbb{1}\left(|Y_{n,j}^{\text{targ}}| > \varepsilon\right)\right] \to 0$$

for every $\varepsilon > 0$.

We conclude using the Lindeberg-Feller CLT that $\sum_{j=1}^{d_n} Y_{n,j}^{\text{targ}} \rightsquigarrow \mathcal{N}\left(-\frac{\tau}{2} + \frac{m^\star - m^{\text{diff}}}{2}, m^{\text{targ}}\right)$, and thus

$$\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma) \rightsquigarrow \mathcal{N}\left(\frac{m^\star - m^{\text{diff}}}{2}, m^{\text{targ}}\right)$$

Step4: Getting the advantage of the misspecified attack.

We use that $\text{TV}\left(\mathcal{N}\left(\mu_0, \sigma_0^2\right) \parallel \mathcal{N}\left(\mu_1, \sigma_0^2\right)\right) = \Phi\left(\frac{|\mu_0 - \mu_1|}{2\sigma_0}\right) - \Phi\left(-\frac{|\mu_0 - \mu_1|}{2\sigma_0}\right)$, so that

$$\begin{aligned}
\lim_{n,d} \text{Adv}_n(\mathcal{A}_{\text{miss}}) &= \text{TV}\left(\mathcal{N}\left(-\frac{m^{\text{targ}}}{2}, m^{\text{targ}}\right) \middle\| \mathcal{N}\left(\frac{m^\star - m^{\text{diff}}}{2}, m^{\text{targ}}\right)\right) \\
&= \Phi\left(\frac{|m^\star + m^{\text{targ}} - m^{\text{diff}}|}{4\sqrt{m^\star}}\right) - \Phi\left(-\frac{|m^\star + m^{\text{targ}} - m^{\text{diff}}|}{4\sqrt{m^\star}}\right) \\
&= \Phi\left(\frac{|m^{\text{scal}}|}{2\sqrt{m^{\text{targ}}}}\right) - \Phi\left(-\frac{|m^{\text{scal}}|}{2\sqrt{m^{\text{targ}}}}\right)
\end{aligned}$$

because $m^{\text{diff}} = m^\star + m^{\text{targ}} - 2m^{\text{scal}}$. □

Remark D.7 (Simple way to get the expectations computation)**.** *Thanks to Remark D.1, we recall that*

$$\ell_n(\hat{\mu}_n; z^{\text{targ}}, \mu, C_\sigma) \approx (z^{\text{targ}} - \mu)^T C_\sigma^{-1}(\hat{\mu}_n - \mu) - \frac{1}{2n}\|^{\text{targ}} - \mu\|_{C_\sigma^{-1}}^2$$

And thus taking the expectation under H_0 and H_1, using that $\mathbb{E}_0(\hat{\mu}_n) = \mu$, $\mathbb{E}_1(\hat{\mu}_n) = \frac{n-1}{n}\mu + \frac{1}{n}z^\star$ and $\mathbb{V}_0(\hat{\mu}_n) = \mathbb{V}_1(\hat{\mu}_n) = C_\sigma$, we get back the same expectations and variances values as the result of Theorem 6.7.

List of Figures

List of Algorithms

List of Tables

Bibliography

[AB22] Achraf Azize and Debabrota Basu. When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35:32199–32210, 2022.

[AB24a] Achraf Azize and Debabrota Basu. Concentrated differential privacy for bandits. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, 2024.

[AB24b] Achraf Azize and Debabrota Basu. How much does each datapoint leak your privacy? quantifying the per-datum membership leakage. *arXiv preprint arXiv:2402.10065*, 2024.

[AB24c] Achraf Azize and Debabrota Basu. Open problem: What is the complexity of joint differential privacy in linear contextual bandits? In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5306–5311. PMLR, 2024.

[ABM10] J-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-armed Bandits. In *Conference on Learning Theory*, 2010.

[Abo18] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2867–2867, 2018.

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[AJMB23] Achraf Azize, Marc Jourdan, Aymen Al Marjani, and Debabrota Basu. On the complexity of differentially private best-arm identification with fixed confidence. *arXiv preprint arXiv:2309.02202*, 2023.

[AJMB24] Achraf Azize, Marc Jourdan, Aymen Al Marjani, and Debabrota Basu. Differentially private best-arm identification. *arXiv preprint arXiv:2406.06408*, 2024.

[AKO+23] Galen Andrew, Peter Kairouz, Sewoong Oh, Alina Oprea, H Brendan McMahan, and Vinith Suriyakumar. One-shot empirical privacy estimation for federated learning. *arXiv preprint arXiv:2302.03098*, 2023.

[Ann03] George J Annas. Hipaa regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348:1486, 2003.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[BBG18] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

[BDT19] Debabrota Basu, Christos Dimitrakakis, and Aristide Tossou. Differential privacy for multi-armed bandits: What is it and what is its cost? *arXiv preprint arXiv:1905.12298*, 2019.

[BHH+24] Gavin Brown, Jonathan Hayase, Samuel Hopkins, Weihao Kong, Xiyang Liu, Sewoong Oh, Juan C Perdomo, and Adam Smith. Insufficient statistics perturbation: Stable estimators for private least squares extended abstract. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 750–751. PMLR, 2024.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[BSW11] Dan Boneh, Amit Sahai, and Brent Waters. Functional encryption: Definitions and challenges. In *Theory of Cryptography: 8th Theory of Cryptography Conference, TCC 2011, Providence, RI, USA, March 28-30, 2011. Proceedings 8*, pages 253–273. Springer, 2011.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10, 2014.

[BV96] Dirk Bergemann and Juuso Välimäki. Learning and strategic pricing. *Econometrica: Journal of the Econometric Society*, pages 1125–1149, 1996.

[CCN+22] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[Che21] Albert Cheu. Differential privacy in the shuffle model: A survey of separations. *arXiv preprint arXiv:2107.11839*, 2021.

[CJZ+22] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022.

[CLK+14] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

[CP34] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[CSS11] T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3), nov 2011.

[DJW13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

[DKM19] Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[DNP+10] Cynthia Dwork, Moni Naor, Toniann Pitassi, Guy N Rothblum, and Sergey Yekhanin. Pan-private streaming algorithms. In *Innovations in Computer Science*, pages 66–80, 2010.

[DNPR10a] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *ACM symposium on Theory of computing*, pages 715–724. ACM, 2010.

[DNPR10b] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *ACM Symposium on Theory of Computing*, STOC '10, page 715–724, New York, NY, USA, 2010. Association for Computing Machinery.

[DNR+09] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390, 2009.

[DR14a] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DR14b] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DRS19] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[DSSU17] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[EDMM02] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Conference on Computational Learning Theory*, COLT '02, page 255–270, Berlin, Heidelberg, 2002. Springer-Verlag.

[EDMMM06] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[EGS03] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, 2003.

[EMRS19] Úlfar Erlingsson, Ilya Mironov, Ananth Raghunathan, and Shuang Song. That which we call private. *arXiv preprint arXiv:1908.03566*, 2019.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[GCPP22] Evrard Garcelon, Kamalika Chaudhuri, Vianney Perchet, and Matteo Pirotta. Privacy amplification via shuffling for linear contextual bandits. In *International Conference on Algorithmic Learning Theory*, pages 381–407. PMLR, 2022.

[GGL12] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

[GK16] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[GKK+24] Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Raghu Meka, and Chiyuan Zhang. User-level differential privacy with few examples per user. *Advances in Neural Information Processing Systems*, 36, 2024.

[GLZ14] Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014.

[HGFD22] Osama A Hanna, Antonious M Girgis, Christina Fragouli, and Suhas Diggavi. Differentially private stochastic linear bandits:(almost) for free. *arXiv preprint arXiv:2207.03445*, 2022.

[HH22] Bingshan Hu and Nidhi Hegde. Near-optimal thompson sampling-based algorithms for differentially private stochastic bandits. In *Uncertainty in Artificial Intelligence*, pages 844–852. PMLR, 2022.

[HHM21] Bingshan Hu, Zhiming Huang, and Nishant A. Mehta. Optimal algorithms for private online learning in a stochastic environment, 2021.

[HOT+23] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Investigating membership inference attacks under data dependencies. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pages 473–488. IEEE, 2023.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[HZZ22] Jiahao He, Jiheng Zhang, and Rachel Zhang. A reduction from linear contextual bandit lower bounds to estimation lower bounds. In *International Conference on Machine Learning*, pages 8660–8677. PMLR, 2022.

[JD24] Marc Jourdan and Rémy Degenne. Non-asymptotic analysis of a ucb-based top two algorithm. *Advances in Neural Information Processing Systems*, 36, 2024.

[JDB+22] Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 35:26791–26803, 2022.

[JDK24] Marc Jourdan, Rémy Degenne, and Emilie Kaufmann. An ε-best-arm identification algorithm for fixed-confidence and beyond. *Advances in Neural Information Processing Systems*, 36, 2024.

[JLB22] Matthew Jörke, Jonathan Lee, and Emma Brunskill. Simple regret minimization for contextual bandits using bayesian optimal experimental design. In *ICML2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2022.

[JMNB14] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[JN14] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

[Jou24] Marc Jourdan. *Solving Pure Exploration Problems with the Top Two Approach*. PhD thesis, Université de Lille, 2024.

[JRSS23] Palak Jain, Sofya Raskhodnikova, Satchit Sivakumar, and Adam Smith. The price of differential privacy under continual observation. In *International Conference on Machine Learning*, pages 14654–14678. PMLR, 2023.

[JUO20] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

[JWO+23] Matthew Jagielski, Stanley Wu, Alina Oprea, Jonathan Ullman, and Roxana Geambasu. How to combine membership-inference attacks on multiple updated machine learning models. *Proceedings on Privacy Enhancing Technologies*, 2023.

[Kal18] Nathan Kallus. Instrument-armed bandits. In *Algorithmic Learning Theory*, pages 529–546. PMLR, 2018.

[Kat10] Jonathan Katz. *Digital signatures*, volume 1. Springer, 2010.

[KCG16] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[KK21] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[KNSS21] Dionysios S Kalogerias, Kontantinos E Nikolakakis, Anand D Sarwate, and Or Sheffet. Quantile multi-armed bandits: Optimal best-arm identification and a differentially private scheme. *IEEE Journal on Selected Areas in Information Theory*, 2(2):534–548, 2021.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[KTAS12] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, volume 12, pages 655–662, 2012.

[KV18] Vishesh Karwa and Salil Vadhan. Finite Sample Differentially Private Confidence Intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.

[KW60] Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LeC73] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1(1):38–53, 01 1973.

[LEHT22] David E Losada, David Elsweiler, Morgan Harvey, and Christoph Trattner. A day at the races: using best arm identification algorithms to reduce the cost of information retrieval user studies. *Applied Intelligence*, 52(5):5617–5632, 2022.

[LF23] Jesús López-Fidalgo. *Optimal Experimental Design: A Concise Introduction for Researchers*, volume 226. Springer Nature, 2023.

[LGG22] Clément Lalanne, Aurélien Garivier, and Rémi Gribonval. On the statistical complexity of estimation and testing under privacy constraints. *arXiv preprint arXiv:2210.02215*, 2022.

[LJD+17] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[LPJ22] Simon Lindståhl, Alexandre Proutiere, and Andreas Johnsson. Measurement-based admission control in sliced networks: A best arm identification approach. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 1484–1490. IEEE, 2022.

[LPK23] Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Gaussian membership inference privacy. *arXiv preprint arXiv:2306.07273*, 2023.

[LRJ+22] Zhaoqi Li, Lillian Ratliff, Kevin G Jamieson, Lalit Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

[LS17] 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.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LVR+19] Pieter JK Libin, Timothy Verstraeten, Diederik M Roijers, Jelena Grujic, Kristof Theys, Philippe Lemey, and Ann Nowé. Bayesian best-arm identification for selecting influenza mitigation strategies. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018*, 2019.

[LZJ22] Fengjiao Li, Xingyu Zhou, and Bo Ji. Differentially private linear bandits with partial distributed feedback. In *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 41–48. IEEE, 2022.

[Mah36] Prasanta Chandra Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*, volume 12, pages 49–55, 1936.

[MSS22] Samuel Maddock, Alexandre Sablayrolles, and Pierre Stock. Canife: Crafting canaries for empirical privacy measurement in federated learning. *arXiv preprint arXiv:2210.02912*, 2022.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[MT15] Nikita Mishra and Abhradeep Thakurta. (Nearly) optimal differentially private stochastic multi-arm bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

[NHS+23] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. *arXiv preprint arXiv:2302.07956*, 2023.

[NP33] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[NR18] Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. In *International Conference on Machine Learning*, pages 3720–3729. PMLR, 2018.

[Pet12] Valentin V Petrov. *Sums of independent random variables*, volume 82. Springer Science & Business Media, 2012.

[QKR17] Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

[RBCS23] Alexandre Rio, Merwan Barlier, Igor Colin, and Marta Soare. Multi-agent best arm identification with private communications. In *International Conference on Machine Learning*, 2023.

[Rus16] Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.

[SHM+20] Xuedong Shang, Rianne Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.

[SJ18] Andrew Stirn and Tony Jebara. Thompson sampling for noncompliant bandits. *arXiv preprint arXiv:1812.00856*, 2018.

[SNJ23] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *arXiv preprint arXiv:2305.08846*, 2023.

[SOJH09] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[SS18] Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306, 2018.

[SS19] Touqir Sajed and Or Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pages 5579–5588. PMLR, 2019.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[SU20] Thomas Steinke and Jonathan Ullman. The pitfalls of averagecase differential privacy, 2020.

[SWS+22] Nícollas Silva, Heitor Werneck, Thiago Silva, Adriano CM Pereira, and Leonardo Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022.

[TBD+16] Katherine Tucker, Janice Branson, Maria Dilleen, Sally Hollis, Paul Loughlin, Mark J Nixon, and Zoë Williams. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC medical research methodology*, 16(1):5–14, 2016.

[TD16] Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[TD17] Aristide CY Tossou and Christos Dimitrakakis. Achieving privacy in the adversarial multi-armed bandit. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[Tho33a] W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25, 1933.

[Tho33b] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[TS13] Abhradeep Guha Thakurta and Adam Smith. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. *Advances in Neural Information Processing Systems*, 26, 2013.

[TVV+17] Abhradeep Guha Thakurta, Andrew H Vyrros, Umesh S Vaishampayan, Gaurav Kapoor, Julien Freudiger, Vivek Rangarajan Sridhar, and Doug Davidson. Learning new words. *Granted US Patents*, 9594741:2, 2017.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VVdB17] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[VW21] Salil Vadhan and Tianhao Wang. Concurrent composition of differential privacy. In *Theory of Cryptography: 19th International Conference, TCC 2021, Raleigh, NC, USA, November 8–11, 2021, Proceedings, Part II 19*, pages 582–604. Springer, 2021.

[VZ22] Salil Vadhan and Wanrong Zhang. Concurrent composition theorems for all standard variants of differential privacy. *arXiv preprint arXiv:2207.08335*, 2022.

[War65] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, pages 63–69, 1965.

[Wil27] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[WTP21] Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[YGFJ18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[YMM+22] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.

[YQWY23] Wei You, Chao Qin, Zihao Wang, and Shuoguang Yang. Information-directed selection for top-two algorithms. In *Conference on Learning Theory*, pages 2850–2851. PMLR, 2023.

[ZCH+20] Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning, 2020.

[ZCL14] Yuan Zhou, Xi Chen, and Jian Li. Optimal pac multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning*, pages 217–225. PMLR, 2014.

[ZDLB21] Andrea Zanette, Kefan Dong, Jonathan N Lee, and Emma Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34:22720–22731, 2021.

[Zha11] Fuzhen Zhang. *Matrix theory: basic results and techniques*. Springer, 2011.

[ZYS+20] Bo Zhang, Ruotong Yu, Haipei Sun, Yanying Li, Jun Xu, and Hui Wang. Privacy for all: Demystify vulnerability disparity of differential privacy against membership inference attack. *arXiv preprint arXiv:2001.08855*, 2020.