Dual Query for DP Query Release

Xi Gong

12/12/2021

1.Introduction

> Background and motivation

A central problem in differential privacy is to accurately answer a large number of statistical queries. One approach is to release a differentially private synthetic dataset and perform all data analytics on it. This is differentially private by post-processing and eliminates the need for designing different algorithms tailored to specific data analytics. One of the classical algorithms for realeasing synthetic data is the Private Multiplicative Weights (PMW) algorithm [Hardt 12], whose max error has a near-optimal dependency on the size of the query and data universe. However, the algorithm's computation depends exponentially on the data dimension, making it infeasible for higher dimensional data that are common in practice.

While exponential runtime is proven to be necessary in the worst case [Ullman 2016], there is still a need for algorithms that work efficiently in practice. One such algorithm is called Dual Query [Gaboardi 2014]. It borrows a novel game-theoretic view of the query release problem that is introduced in [Hsu 2012]. Under this view, the query release problem is reframed into a zero-sum game between a data and a query player. Algorithms such as PMW then belongs to the primal view of the problem (the data player goes first and the query player finds the best responses), while Dual Query belongs to the dual (the query player go first and then lets the data player finds the best responses). The computation gain of Dual Query comes from the fact that the multiplicative weights is now kept over the set of queries (tractable) rather the space of all possible records (intractable).

2. Dual Query for Query Release

> High level summary of the DualQuery algorithm [Gaboardi 2014].

This section presents a brief summary of the algorithmic and proof ideas of the DualQuery algorithm in [Gaboardi 2014]. The algorithm description is contained in the attached picture.

The key idea of this algorithm is to reframe the *synthetic data generation* problem as finding the **approximate mixed Nash equilibrium** to a zero sum game between the data and query player. The game is defined as the following: in each round, the query player plays a *distribution* over the set of queries, and the data player responds to it by choosing a record x. The payoff function of this zero sum game is then defined as the expected difference between the "response to query on x" and "response to query on actual dataset D ". The query player aims to maximize the payoff, while the data player aims to do the opposite.

If the query player maintains the distribution using multiplicative weights algorithm, and the data player finds approximate best responses according to the payoff function, then after $O(\log(|Q|))$ iterations, the "average distribution" and the "average dataset" formed over the rounds forms an approximate Nash-equilibrium. It is then demonstrated that an alpha-approximate Nash-equilibrium translates to a synthetic dataset that is alpha-accurate. Since the distribution is now kept over the queries instead of the set of all possible records, the runtime for the multiplicative weights update is now tractable. This isolates the hardness of the problem to the optimization step the data player plays, and in practice one can either expect the optimization to be easy most of the time, or terminate the optimization program when it exceeds a certain time constraint and pick the suboptimal solution instead (without any compromise to privacy).

The only thing that is left undiscussed is privacy; since the query player updates his distribution based on a function of the dataset, the query distribution itself cannot be revealed directly. Therefore, to access the distribution privately, the authors used sampling, which can be seen as a type of exponential algorithm. The sample size is chosen so that it is large enough to be representative of the "weighted query", and small enough to ensure small privacy loss.

2.1 Engine of Dual Query

> Proof of Von-Neumann's minimax theorem using tools from online learning **[Freund and Schapire]**; implicitly hints a method for finding α-approximate equilibrium; a truly beautiful connection and proof

While the minimax theorem (two player, finite decision setting) is a simple consequence of **strong duality for LP**, it admits a beautiful proof using ideas from **online learning** under the assumption that the payoff matrix has entries bounded between positive and minus one. i.e., $A \in [-1,1]^{n \times m}$. Importantly, this proof also contains in itself a method for finding α-approximate equilibrium. For notational convenience, we assume implicitly that the maximization and minimization are restricted to the probability simplex of the two players' decision set.

One direction of von-Neumann's minimax theorem is trivial

$$\max_v \min_u A(u,v) \leq \min_u \max_v A(u,v).$$

This aligns with our intuition that playing second-hand can only be advantageous.

The reverse inequality, however, seems quite puzzling. One may be tempted to first prove $A(u,v) \leq \min_v A(u,v)$ then gradually adds on to that, but this inequality is not even true. The ordering of the minimization and maximization works against us.

A common trick in this scenario is to first relax the desired inequality by an epsilon term and then show that we can allow epsilon to be arbitrarily small. More precisely, we want to prove

$$\min_u \max_v A(u,v) \leq \max_v \min_u A(u,v) + \epsilon,$$

for arbitrary $\epsilon > 0$.

This allows us to first set the modest goal of proving $A(u,v) \leq \min_v A(u,v) + \epsilon$. If the person who stares at this inequality happens to relish the notion of regret minimization, there is some chance that he will play with the notations and write the following tautology

$$A(u,v) - \min_v A(u,v) = \langle u,v\rangle_A - \min_u \langle u,v\rangle_A = \frac{1}{T}\sum_{i=1}^{T}\langle u,v\rangle_A - \min_u \frac{1}{T}\sum_{t=1}^{T}\langle u,v\rangle_A$$

where $\langle\cdot,\cdot\rangle_A$ is the inner product induced by the payoff matrix A. This is for resemblance of notation to literation in online learning.

Note that the last term looks remarkably similar to the notion of **regret** in online learning, but we are yet to take advantage of the power of online algorithms. To do so, we replace the fixed vector v above by an arbitrary sequence $\{v_t\}$. This sequence corresponds to the loss vectors chosen by the adversary in the expert game framework. Then we let u_t be distribution maintained by the **multiplicative weights algorithm** (MW) at time t. By classical regret bound for the expert game, we have

$$\frac{1}{T}\sum_{i=1}^{T}\langle u_t,v_t\rangle_A - \min_u \frac{1}{T}\sum_{t=1}^{T}\langle u,v_t\rangle_A \leq \epsilon_T$$

where ϵ_T is a term that is sublinear in T.

To simplify notation, we let $\tilde{v} := \frac{1}{T}\sum_{t=1}^{T} v_t$. Rewrite the above inequality with this notation, we get

$$\frac{1}{T}\sum_{t=1}^{T}\langle u_t,v_t\rangle_A \leq \min_u \langle u,\tilde{v}\rangle_A + \epsilon_T \leq \max_v \min_u \langle u,v\rangle_A + \epsilon_T.$$

This completes half of the work. To finish the proof, we dream an inequality of the form

$$\frac{1}{T}\sum_{t=1}^{T}\langle u_t,v_t\rangle_A \overset{(?)}{\geq} \max_v \langle\tilde{u},v\rangle_A \geq \min_u \max_v \langle u,v\rangle_A,$$

where $\tilde{u} := \frac{1}{T}\sum_{t=1}^{T} u_t$. However, the first inequality is obviously false.

With a moment of thought, one can realize that we have not taken full advantage of the regret bound. While we indeed allowed our player to play MW, we have not grant the adversary his full power. In particular, one can easily check that if v_t is **chosen adaptively** to maximize $\langle u_t, v_t\rangle$ at each round t, then the first equality t true while everything else proven so far continue to hold (thanks to the generality of the regret bound).

Combine, we get

$$\min_u \max_v \langle u,v\rangle_A \leq \max_v \min_u \langle u,v\rangle_A + \epsilon_T.$$

The inequality is proved by taking $T \to 0$.

Within this proof, we have two byproducts:

1. $\min_u \max_v \langle u,v\rangle_A \leq \min_u \langle u,\tilde{v}\rangle_A + \epsilon_T$
2. $\max_v \langle\tilde{u},v\rangle_A \leq \max_u \min_v \langle u,v\rangle_A + \epsilon_T$.

Pick T such that $\epsilon_T = \alpha$, then these two inequalities tell us that $(\tilde{u},\tilde{v})$ forms an alpha-approximate equilibrium after T iterations. Recall that $\tilde{u}$ are average responses by the MW player, and $\tilde{u}$ are the average of best responses by the adversary.

3. Improved Privacy Analysis

> A slightly different privacy analysis leads to constant factor improvements under large T regime.
>
> Due to a misread of the deadline I was unable to put an autoDP plot here.

The privacy analysis of DualQuery involves composition of T mechanism, each pure-DP with $\epsilon_t = \frac{2s\eta(t-1)}{n}$. The authors applied advanced composition for *homogeneous pure-DP mechanisms* by upper bounding ϵ_t by $\frac{2s\eta(T-1)}{n}$ for all $t \leq T$, which ignores their individual dif

A slightly different analysis through zCDP can improve this privacy analysis under regime where T is moderately large. At round t, the mechanism is $\frac{1}{2}s(2\eta(t-1)/n)^2$-zCDP. Compositing over T rounds, we get that the algorithm is $\rho - \mathrm{zCDP}$ with

$$\rho = \sum_{t=1}^{T}\frac{1}{2}s\epsilon_t^2 = \frac{2s\eta^2(T-1)(T)(2(T-1)+1)}{6n^2}.$$

Converting from zCDP back to approximate DP, we get that DualQuery is (ϵ, δ)-DP with

$$\begin{aligned}\epsilon &= \frac{s\eta^2(T-1)(T)(2(T-1)+1)}{3n^2} + 2\sqrt{\frac{s\eta^2(T-1)(T)(2(T-1)+1)}{3n^2}\log\left(\frac{1}{\delta}\right)} \\ &= \frac{s\eta^2(T-1)(T-1)(2(T-1))}{3n^2} + \frac{s\eta^2(T-1)(T-1)}{3n^2} + \frac{s\eta^2(T-1)(2(T-1)+1)}{3n^2} + 2\sqrt{\frac{s\eta^2(T-1)(T)(2(T-1)+1)}{3n^2}\log\left(\frac{1}{\delta}\right)} \\ &= \frac{s\alpha^2(T-1)^3}{24n^2} + \ldots\end{aligned}$$

approximate DP. Note that terms that are dominated by $(T-1)^3$ are omitted as we will consider regime where T is relatively large.

On the other hand, the privacy guarantee in the original DualQuery paper guarantee is

$$\begin{aligned}\varepsilon &= \frac{2\eta(T-1)}{n} \cdot \left[\sqrt{2s(T-1)\log(1/\delta)} + s(T-1)\left(\exp\left(\frac{2\eta(T-1)}{n}\right) - 1\right)\right] \\ &\geq \frac{2\eta(T-1)}{n} \cdot \left[\sqrt{2s(T-1)\log(1/\delta)} + s(T-1)\left(\frac{2\eta(T-1)}{n}\right)\right] \\ &= \frac{s\alpha^2(T-1)^3}{4n^2} + \ldots\end{aligned}$$

Under regime where T is large, analysis through zCDP improves the privacy parameter by factor of 6.

4. Subsampled DualQuery

> Subsampling DualQuery yields different dependencies.

Algorithm. Subsampled-Dual Query

Parameters: Target accuracy $a \in (0,1)$, target failure probability $\beta \in (0,1)$

Input: Database D and linear queries Q.

1. Draw $\frac{2\log\left(\frac{2|Q|}{\beta}\right)}{\alpha^2}$ samples without replacement from the database uniformly at random.
2. Initialize $T' = \frac{64\log|Q|}{\alpha^2}, \eta' = \frac{\alpha}{8}, s' = \frac{192\log(2|X|T/\beta)}{\alpha^2}$.
3. Run DualQuery on the new set of parameters.

The only modification of the above algorithm to DualQuery is that we first draw $\frac{2\log\left(\frac{2|Q|}{\beta}\right)}{\alpha^2}$ samples with replacement from the database uniformly at random. This sample is sufficiently large to be representative of the original database on any query of size $|Q|$, but can contain significantly fewer number of records under reasonable assumptions on $|Q|$ and n. The hope is that this will 1) reduce the runtime of the non-private optimization step by reducing the number of records in consideration 2) achieve similar or better privacy guarantee.

Claim 1. (The subsampled database is representative of the original database on Q).

With probability $\beta/2$, we have

$$\max_{q\in Q} |\frac{1}{n}q^T x - \frac{1}{m}q^T\tilde{x}| \leq \frac{\alpha}{2},$$

where $\tilde{x}$ denote the dataset obtained by sampling $\frac{2\log\left(\frac{2|Q|}{\beta}\right)}{\alpha^2}$ records with replacement.

Proof.

We assume that n is sufficient large so that the samples can be viewed as nearly independently.

Note that $E(\frac{1}{m}q^T\tilde{x}) = E(\sum_{i=1}^m \frac{f_q(\phi_i)}{m}) = \frac{1}{n}q^T x$, and that $0 \leq f_q(\phi_i) \leq 1$. By Hoeffding's lemma and union bound, we obtain the claim. This type of argument is presented in lecture when analyzing the smallDB algorithm.

Claim 2. (Output of subsampled-DualQuery is α-accurate with respect to original database)

Let $\hat{x}$ denote the output of the subsampled DualQuery. With probability β, we have

$$|q(\hat{x}) - q(D)| \leq \alpha.$$

Proof. Note that $|q(\hat{x}) - q(D)| \leq |q(\hat{x}) - q(\hat{D})| + |q(\hat{D}) - q(D)|$. The first term is bounded by $\frac{\alpha}{2}$ due to guarantee of the multiplicative weights algorithm a, the second term is bounded by $\frac{\alpha}{2}$ by claim 1.

Claim 3. (Privacy amplification)

Fix failure probability β, accuracy parameter α, and δ parameter in approximate DP.

The approximate-DP guarantee for original DualQuery is

$$(\epsilon, \delta) = (\frac{256 \log^{3/2} |\mathcal{Q}| \sqrt{6 \log(1/\delta) \log(2|\mathcal{X}|T/\beta)}}{\alpha^3 n}, \delta)$$

The approximate-DP guarantee for subsampled DualQuery is

$$\epsilon_{sub} = [\frac{4 \log(2|Q|/\beta)}{\alpha^2 n}] \cdot \frac{8 \cdot 256 \log^{3/2} |Q| \sqrt{6 \log(1/\delta) \log(8|X|T/\beta)}}{a^3 n}.$$

$$\delta_{sub} = [\frac{4 \log(2|Q|/\beta)}{\alpha^2 n}]\delta$$

Proof. This follows from direct application the subsampling lemma (without replacement).

4. References.

Paper

1. [Hardt 2012] https://proceedings.neurips.cc/paper/2012/file/208e43f0e45c4c78cafadb83d2888cb6-Paper.pdf
2. [Gaboardi 2014] https://arxiv.org/abs/1402.1526

Lecture note

1. [Nishant Mehta] http://web.uvic.ca/~nmehta/ml_theory_spring2019/lecture19.pdf