조합론적 블록 설계에 기반한 통신 효율적인 국소 차등 정보보호 기법에 관한 연구

박현영, 이시현*
한국과학기술원 전기및전자공학부

phy811@kaist.ac.kr, *sihyeon@kaist.ac.kr

A Study on Communication-Efficient Local Differential Privacy Mechanism based on Combinatorial Block Design

Hyun-Young Park, Si-Hyeon Lee*
School of Electrical Engineering, KAIST

요 약

본 논문은 조합론적 블록 설계를 국소 차등 정보보호(LDP) 기법 설계에 활용하는 방법에 대해 제안한다. 이를 통해 기존에 알려진 대표적인 LDP 기법을 하나의 일반적인 이론으로 통합하고, 나아가 기존에 알려진 최적의 추론 정확도를 유지하면서도 보다 적은 통신 자원량을 사용하는 새로운 방법론을 제시한다.

Ⅰ. 서 론

다양한 정보를 수집 및 활용하는 빅데이터 기술에서 데이터 제공자의 개인정보 침해는 사회적으로 중대한 문제이다. 이를 위해 데이터 수집 전 개인정보를 복원하기 어렵도록 데이터를 변형하는 국소 차등 정보보호(Local Differential Privacy, 이하 LDP) 기술이 널리 연구 및 활용되고 있다 [1]. 특히 최근에는 통계적 추론 성능 뿐만 아니라 데이터 수집 과정에서 수반되는 통신 자원량을 함께 고려한 LDP 기술 연구가 활발히 진행되고 있다 [2]-[3]. 본 논문에서는 조합론적 블록 설계에 기반하여 기존에 알려진 LDP 기술들을 하나의 이론으로 정리한다. 나아가 이 이론을 활용하여, 통신 효율을 고려하지 않았을 때 LDP 제약하에서 얻을 수 있는 최적의 추론 정확도를 달성하면서도, 최소한의 통신 자원을 요구하는 새로운 방법을 제시한다.

Ⅱ. 본 론

Ⅱ-1. 시스템 모델

본 논문에서는 LDP 제약이 있는 상황에서의 이산 분포 추정 문제를 다룬다. $\mathcal{X}=\{1,2,\cdots,k\}$에 속하는 값을 가질 수 있는 n 개의 사용자 데이터 $X_1, X_2, \cdots, X_n$ 가 있으며, 이들은 확률질량함수 $P=(P_1, P_2, \cdots, P_k) \in \Delta^{k-1}$ 에 따라 IID 하게 분포되어 있다. 각 사용자는 자신의 데이터 X_i를 서버로 전송할 데이터 Y_i로 변환하며, 이는 조건부 분포 $Q(y|x)=Pr(Y_i=y|X_i=x)$ 를 따른다. 각각의 Y_i가 가질 수 있는 값의 집합은 $\mathcal{Y}$ 로 표시한다. 이 때 서버에 보내는 데이터 Y_i는 원본 데이터 X_i에 대한 정보를 많이 누출시키지 않을 것을 요구해야 한다. 다음의 국소 차등 정보보호 (LDP)는 이러한 정보 보호 정도에 대해 널리 활용되는 지표이다.

정의 1. 국소 차등 정보보호(LDP)[4]

$\epsilon>0$ 에 대해, 조건부 분포 $Q(y|x)$ 가 다음의 조건을 만족할 때 ϵ-LDP 를 만족한다고 한다.

$$Q(y|x) \le e^{\epsilon} Q(y|x'), \forall x, x' \in \mathcal{X}, y \in \mathcal{Y}$$

서버는 전송받은 데이터 $Y_1, Y_2, \cdots, Y_n$ 를 이용해서, 원본 데이터의 분포 P 를 추정하려 한다. 이를 위해, 서버는 분포 P의 추론값 $\hat{P}=(\hat{P}_1, \hat{P}_2, \cdots, \hat{P}_k)$을 만들며, 각각의 $\hat{P}_i$ 는 $Y_1, Y_2, \cdots, Y_n$의 함수이다.

이 모델에서의 주요 목적은 다음의 두가지 목표를 달성하는 ϵ-LDP 를 만족하는 조건부 분포 $Q(y|x)$와 이에 대응되는 추론치 $\hat{P}$를 설계하는 것이다.

1. 통신에 사용되는 자원량 최소화
2. 추론 오차 최소화

여기서, 통신 자원량은 $|\mathcal{Y}|$ 에 해당하며, 추론 오차로는 다음의 평균 제곱 오차를 고려한다.

$$L(P, Q, n, k, \hat{P}) = E\left[\sum_{i=1}^{k} |P_i - \hat{P}_i|^2\right]$$

Ⅱ-2. 기존의 기법 및 알려진 결과

기존에 제시된 대표적인 LDP 기법들로 Subset selection[1], Hadamard response[2], Projective geometry response[3] 등이 있다(이하 SS, HR, PGR). 통신 자원량을 고려하지 않을 경우, 주어진 LDP 제약하에서 수학적으로 가능한 최소의 추론 오차는 알려져 있으며, SS[1]는 이러한 최적의 오차를 달성하는 기법임이 알려져 있다. 그러나 SS 는 통신 자원량이 매우 크며, 구체적으로 $|\mathcal{Y}|$ 의 값이 k에 대해 지수적 이상의 증가를 보인다. 한편, HR[2]과 PGR[3]은 SS 에 비해 통신 자원량이 매우 작으며, 계산적 및 실험적으로 추론 오차가 최적의 값과 근접함이 알려져 있다.

Ⅱ-3. 제안하는 기법 설계 방법 및 주요 결과

본 연구에서는 조합론의 주요 대상인 블록 설계[5]를 이용한 LDP 기법을 제안하고 이의 성능에 대해 분석하였다. 이를 통해 아래와 같은 결론을 얻어냈다.

1. 제안한 방법은 Ⅱ-2 장에서 나열한 기법을 모두 포함할 뿐만 아니라 기존에 알려지지 않은 새로운 기법을 다수 설계할 수 있게 한다.

2. 제안한 방법을 통해 설계한 기법은 최적의 오차를 달성하며, 좋은 블록 설계의 사용을 통해 통신 자원량 역시 크게 감소할 수 있다.

먼저, 블록 설계는 특정한 균일성을 갖춘 집합들의 모임이며, 다음과 같이 정의된다.

정의 2. 블록 설계((Balanced Incomplete) Block Design) 크기가 k인 유한 집합 $\mathcal{X}$와 두 정수 $d \in [1, k-1]$, $c \geq 0$에 대해, 다음의 조건을 만족하는 $\mathcal{X}$의 부분집합들의 모임 $\mathcal{B} \subset 2^{\mathcal{X}}$을 (k, d, c)-블록 설계라 한다.

(a) $|B| = d, \forall B \in \mathcal{B}$

(b) 임의의 서로 다른 두 $x_1, x_2 \in \mathcal{X}$에 대해, $\{x_1, x_2\} \subset B$를 만족하는 $B \in \mathcal{B}$의 개수는 c개이다.

위의 블록 설계를 활용해, 본 논문에서는 아래 정리 1 과 같은 LDP 기법을 제안하고자 한다.

정리 1. $\mathcal{X} = \{1, 2, \cdots, k\}$에서 정의된 (k, d, c)-블록 설계 $\mathcal{B}$와 양수 ϵ에 대해, ϵ-LDP 를 만족하는 조건부 분포 Q를 다음과 같이 만들 수 있다. 먼저, $\mathcal{Y} = \mathcal{B}$으로, $y \in \mathcal{Y}$는 $\mathcal{X}$의 부분집합으로 주어진다. 이 때, $Q(y|x)$는 다음과 같다.

$$x \in y\text{이면, } Q(y|x) = \frac{d(d-1)}{c(k-1)(e^{\epsilon}d+k-d)} e^{\epsilon},$$
$$x \notin y\text{이면, } Q(y|x) = \frac{d(d-1)}{c(k-1)(e^{\epsilon}d+k-d)}.$$

위의 정리를 이용하면, Ⅱ-2 장에서 나열한 SS, HR, PGR 은 특정한 종류의 블록 설계를 사용해 만든 기법으로 표현할 수 있다. 즉 블록 설계를 이용한 LDP 기법은 기존 기법들을 포함하는 보다 일반화된 기법이다. 이 기법의 성능은 다음과 같다. 증명은 분량 상 생략한다.

정리 2. [6] 제안한 기법의 통신 자원량은 다음과 같다.

$$|\mathcal{Y}| = |\mathcal{B}| = \frac{k(k-1)}{d(d-1)} c \geq k$$

정리 3. 제안한 기법 Q에 대해, 분포 P의 비편향 추정치 (Unbiased estimator) $\hat{P}$가 존재하며, 이의 평균 제곱 오차는 다음과 같다.

$$L(P, Q, n, k, \hat{P}) = \frac{1}{n}\left(\frac{(k-1)^2(e^{\epsilon}d+k-d)^2}{kd(k-d)(e^{\epsilon}-1)^2} - \sum_{i=1}^{k}\left(P_i - \frac{1}{k}\right)^2\right)$$

이 때 위의 오차는 c의 값에 무관함을 알 수 있다.

SS 는 가능한 모든 k, d의 조합에 대해 설계가 가능하며, c의 값이 가장 큰 블록 설계로 표현된다. 정리 3 에 의해 추론 오차는 c 의 값에 무관하므로, k, d 의 값을 유지하면서 c 의 값이 매우 작은 블록 설계를 활용해 기법을 만든다면, SS 가 달성하였던 최적의 오차를 유지하면서도 통신 자원량을 크게 줄일 수 있다. 이는 본 논문의 핵심 결과로, 아래의 정리 4 로 정리할 수 있다.

정리 4. $\ln((k/d)-1)$ 근방(neighborhood)의 ϵ에 대해, 모든 ϵ-LDP 기법들 중에서 정리 1 과 정리 3 에서 정의한 $Q, \hat{P}$는 아래에 정의된 asymptotic worst-case 평균 제곱 오차를 최소화한다.

$$\lim_{n\to\infty} \sup_{P \in \Delta^{k-1}} n \times L(P, Q, n, k, \hat{P})$$

한편, HR 과 PGR 은 설계가 가능한 k, d 의 조합이 한정적이지만, 정리 2 에서의 하한 $|\mathcal{B}| = k$를 달성한다. 따라서 이들은 통신 자원량을 최소화하면서, 정리 4 에서의 ϵ의 상황에서는 최적의 오차를 달성할 수 있다. 기존에는 두 기법이 최적의 오차와 근접하다는 결과가 있었는데, 본 논문의 결과에 의해 특정 값 근방의 ϵ에 대해서는 위의 두 기법이 정확히 최적의 오차를 보인다는 강화된 결과를 얻을 수 있다.

나아가, 위의 기법에서 활용되지 않은 다른 블록 설계를 사용해 기법을 만들어도 특정 값 근방의 ϵ에서의 최적의 오차를 달성할 수 있다. 특히, 정리 2 의 하한을 만족하는 블록 설계는 조합론의 중요한 주제이며, 많은 예시가 발견되어 있다. 따라서 이들을 활용하면 기존에 알려지지 않았던, 최적의 오차를 유지하면서도 HR 이나 PGR 이 달성하였던 매우 적은 통신 자원량을 갖는 기법을 다수 설계할 수 있다. 아래는 본 논문에서 제안한 블록 설계 기반 LDP 기법으로 찾은 새로운 결과의 예시이다.

예시 1. k가 소수이고 $k = 4x^2 + 1$인 홀수 x가 존재하면, $d = x^2$ 이면서 $|\mathcal{B}| = k$ 인 블록 설계가 존재한다[6]. 따라서 이를 활용한 LDP 기법은 $\ln(3 + 1/x^2) \simeq \ln 3$ 의 근방에 속한 ϵ에 대해 최적의 오차와 매우 적은 통신 자원량을 달성한다.

Ⅲ. 결 론

본 논문에서는 조합론적 블록 설계를 LDP 연구에 활용하여, 기존 LDP 기법들을 하나의 이론으로 통일할 뿐만 아니라, 최적의 추론 정확도를 보다 적은 통신 자원량을 사용하여 달성하는 새로운 방법에 대해 다루었다. 개인정보 보호와 통계적 추론 성능, 그리고 통신에 소모되는 자원량 모두 중요한 요소인 만큼, 이 세가지 요소를 모두 고려한 본 논문의 기법 및 결과는 다양한 빅데이터 활용 상황에 두루 사용될 수 있다.

ACKNOWLEDGMENT

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1A2C2092151).

참 고 문 헌

[1] Ye, Min, and Alexander Barg. "Optimal locally private estimation under l_p loss for $1 < p < 2$." *Electronic Journal of Statistics* 13.2 (2019): 4102-4120.

[2] Acharya, Jayadev, Ziteng Sun, and Huanyu Zhang. "Hadamard response: Estimating distributions privately, efficiently, and with little communication." *The 22nd International Conference on Artificial Intelligence and Statistics.* PMLR, 2019.

[3] Feldman, Vitaly, et al. "Private frequency estimation via projective geometry." *International Conference on Machine Learning.* PMLR, 2022.

[4] Duchi, John C., Michael I. Jordan, and Martin J. Wainwright. "Local privacy and statistical minimax rates." *2013 IEEE 54th Annual Symposium on Foundations of Computer Science.* IEEE, 2013.

[5] Jukna, Stasys. *Extremal combinatorics: with applications in computer science.* Vol. 571. Berlin: Springer, 2011.

[6] Moore, Emily H., and Harriet Suzanne Katcher Pollatsek. *Difference sets: connecting algebra, combinatorics, and geometry.* Vol. 67. American Mathematical Soc., 2013.