Enjoying Math Series

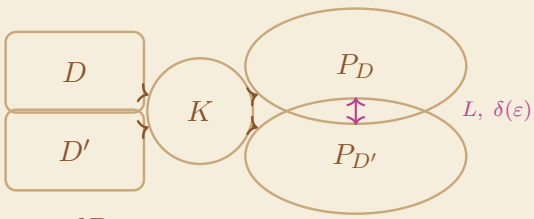
공개된 결과만 보고

한 사람의 흔적은 얼마나 남는가

확률 계산과 정보기하에서 시작하는 차등정보보호

The Shape of Privacy Loss

From Randomized Response to Exact DP and Fisher–Rao Geometry



$$L(y) = \log \frac{dP}{dQ}(y) \qquad \delta(\varepsilon) = \sup_A \{P(A) - e^{\varepsilon} Q(A)\}$$

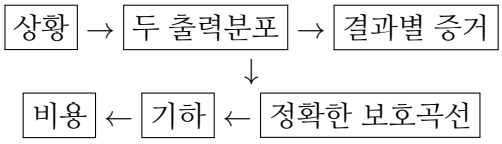
Enjoying Math · 2026

Gunhee Cho

머리말

이 책은 차등정보보호의 공식을 먼저 외우게 하지 않는다. 한 사람의 자료가 있을 때와 없을 때 생기는 두 출력분포를 직접 계산하고, 어느 결과가 두 상황을 구별하는 증거가 되는지 확인한다. 그 계산에서 ε, δ, privacy loss, exact composition, Fisher 정보와 Rao 거리를 차례로 발견한다.

이 책이 반복하는 순서는 다음과 같다.



차등정보보호는 특정한 소프트웨어 이름이나 잡음 공식이 아니다. 공개 결과를 보았을 때 한 사람의 참여 여부에 관한 판단이 얼마나 달라질 수 있는지를 제한하는 확률적 약속이다. 그 약속을 정확히 읽으려면 평균만 아니라 드문 결과의 꼬리까지 보아야 하고, 한 번의 공개만 아니라 반복 공개 전체를 살펴야 한다.

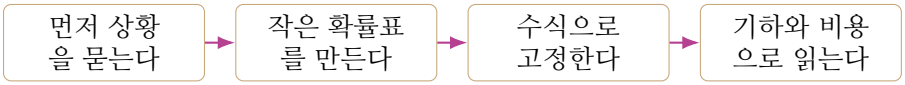
이 책을 사용하는 방법

먼저 물기

확률이나 정보기하를 거의 모르는 독자도 Differential Privacy를 이해할 수 있을까요?

가능합니다. 다만 정의를 먼저 외우지 않습니다. 두 개의 작은 확률표를 만들고, 공개 결과가 어느 상황을 더 지지하는지 손으로 계산한 뒤, 그 계산에 DP라는 이름을 붙입니다.

이 책의 대화법



잠깐, 이렇게 생각해 봅시다

독자: ε이 privacy budget이라는 말은 들었지만 무엇을 재는 값인지 모르겠습니다.
이 책: 예산이라는 비유보다 먼저, 한 사람이 있는 경우와 없는 경우에 같은 결과가 나올 확률의 비를 계산합니다. ε은 그 비가 얼마나 크게 달라질 수 있는지를 로그척도로 제한합니다.

잠깐, 이렇게 생각해 봅시다

독자: δ는 개인정보 보호가 실패할 확률인가요?
이 책: 일반적으로 그렇게만 말하면 부정확합니다. δ는 e^{ε}배의 확률로 덮이지 않는 두 출력분포의 남은 차이입니다. 제3장에서 사건별로 직접 계산합니다.

이 책이 답하려는 질문

1. 한 사람의 기록만 바뀐 두 자료는 무엇인가?
2. 무작위 공개절차는 두 자료를 어떤 출력분포로 바꾸는가?
3. 공개결과 하나가 어느 자료를 더 지지하는가?
4. (ε, δ)는 정확히 어떤 차이를 제한하는가?
5. 여러 번 공개하면 작은 단서들이 어떻게 쌓이는가?
6. 같은 보호수준에서 오차가 가장 작은 메커니즘은 무엇인가?
7. Fisher 정보와 Rao 거리는 DP의 어느 부분을 설명하는가?
8. 국소적 정보수축을 언제 전역 DP로 바꿀 수 있는가?

한 절을 공부하는 방법

첫 번째 식을 건너뛰더라도 두 상황과 공개결과를 말로 설명한다.

두 번째 작은 숫자를 넣어 확률비와 오류율을 손으로 계산한다.

세 번째 정확한 식과 상계, 국소근사를 구분한다.

네 번째 같은 대상을 확률론과 정보기하의 언어로 다시 설명한다.

주장의 등급

주장의 등급

Exact 주어진 분포에서 등식으로 계산한 결과다.

Bound 실제 위험보다 클 수 있지만 안전성을 보장하는 상계다.

Local 가까운 분포에서만 맞는 Taylor 근사다.

Conditional 추가적인 tail 또는 regularity 가정 아래 성립한다.

Open 수치증거나 부분결과는 있으나 아직 일반 증명이 없다.

선수지식

조건부확률, 기댓값, 로그와 미분을 알면 본문을 따라갈 수 있다. 생소한 내용은 부록에서 다시 설명한다. 측도론은 절대연속성과 Radon–Nikodym 미분을 읽는 데 필요한 최소한만 제3장에서 도입한다.

완독의 의미

모든 DP 정의를 암기하는 것이 목표가 아니다. 독자가 새로운 메커니즘 M을 만났을 때 다음 과정을 완성하면 된다.

$$D \sim D' \longrightarrow P = \mathcal{L}(M(D)), \quad Q = \mathcal{L}(M(D')),$$
$$\longrightarrow L = \log \frac{dP}{dQ} \longrightarrow \delta(\varepsilon) \longrightarrow \text{합성} \cdot \text{오차} \cdot \text{한계}.$$

Contents

PART 1

한 사람의 흔적을 확률로 계산하다

chapter 1

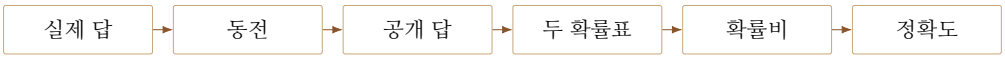
무작위로 대답하면 왜 개인을 보호할 수 있는가

먼저 묻기

사실과 다른 답을 조금 섞는 일이 어떻게 정직한 통계를 가능하게 할까요?

한 사람에게 "민감한 경험이 있습니까?"라고 직접 물으면 두 가지 문제가 생긴다. 사실대로 답하면 개인의 비밀이 드러날 수 있고, 비밀을 지키려고 거짓으로 답하면 집단 통계가 왜곡된다. 무작위응답은 개인의 답과 공개된 답 사이에 우연을 끼워 넣어 두 문제를 함께 다룬다. 이 조사기법의 고전적 출발점은 Warner의 randomized response 논문이다[1].

⟶ 이 장의 대상을 하나의 구별 문제로 옮기는 길



1.1 두 개의 가능한 세상

실제 답을 $X \in \{0,1\}$로 적자. $X = 1$은 "예", $X = 0$은 "아니오"다. 공개되는 답을 $Y \in \{0,1\}$라 한다. 다음 규칙을 사용한다.

1. 확률 p로 실제 답을 공개한다.
2. 확률 $1-p$로 반대 답을 공개한다.

따라서

$$\mathbb{P}(Y = X) = p, \qquad \mathbb{P}(Y \neq X) = 1 - p.$$

예시 1.1 (숫자를 먼저 넣기). $p = 3/4$라고 하자. 실제 답이 1이면 공개 답의 분포는

$$\mathbb{P}(Y = 1 \mid X = 1) = \frac{3}{4}, \qquad \mathbb{P}(Y = 0 \mid X = 1) = \frac{1}{4}.$$

실제 답이 0이면 두 확률은 뒤바뀐다.

$$\mathbb{P}(Y = 1 \mid X = 0) = \frac{1}{4}, \qquad \mathbb{P}(Y = 0 \mid X = 0) = \frac{3}{4}.$$

	$Y = 0$	$Y = 1$
$X = 0$	p	$1 - p$
$X = 1$	$1 - p$	p

말과 그림으로 읽는 직관

공개 답이 1이라고 해서 실제 답이 반드시 1은 아니다. 관찰자는 "사실을 말한 1"과 "뒤집혀 나온 1"을 구별하지 못한다. 보호는 답을 지우는 데서 생기는 것이 아니라 두 원인이 같은 결과를 만들게 하는 데서 생긴다.

1.2 같은 결과가 나올 확률을 비교하기

공개 결과 $Y = 1$이 실제 답 1과 0에서 나올 확률의 비는

$$\frac{\mathbb{P}(Y = 1 \mid X = 1)}{\mathbb{P}(Y = 1 \mid X = 0)} = \frac{p}{1 - p}.$$

$Y = 0$에서는 반대방향의 비가 같은 값을 갖는다. $p \geq 1/2$라면 모든 출력과 두 입력방향을 고려한 최대비는 $p/(1 - p)$다.

정의 1.1 (무작위응답의 개인정보 수준).

$$\varepsilon = \log \frac{p}{1 - p}$$

로 둔다. 그러면

$$\frac{\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y \mid X = x')} \leq e^{\varepsilon}$$

가 모든 $x, x', y \in \{0, 1\}$에 대해 성립한다.

$p = 3/4$에서는

$$\varepsilon = \log 3 \approx 1.099.$$

$p = 1/2$이면 Y는 X와 무관한 동전이므로 $\varepsilon = 0$이다. $p \to 1$이면 사실을 거의 그대로 공개하므로 $\varepsilon \to \infty$다.

여기서 자주 틀립니다

ε은 "유출된 개인정보의 개수"가 아니다. 두 입력에서 같은 결과가 나올 상대적 가능성이 최악의 경우 얼마나 바뀌는지를 로그척도로 표현한 값이다. 단위가 없고, 작을수록 두 상황을 구별하기 어렵다.

1.3 정확도도 함께 계산하기

실제 답을 맞게 공개할 확률은 p이고 틀리게 공개할 확률은 $1-p$다. ε으로 p를 다시 쓰면

$$e^{\varepsilon} = \frac{p}{1-p}$$

이므로

$$p = \frac{e^{\varepsilon}}{1+e^{\varepsilon}}.$$

따라서 공개오류는

$$\mathbb{P}(Y \neq X) = \frac{1}{1+e^{\varepsilon}}.$$

ε	사실 공개확률 p	오류확률 $1-p$
0	0.500	0.500
$\log 2$	0.667	0.333
$\log 3$	0.750	0.250
$\log 9$	0.900	0.100

잠깐, 이렇게 생각해 봅시다

독자: $\varepsilon = 0$이면 가장 좋은 것 아닌가요?
이 책: 보호만 보면 그렇지만 공개 답은 실제 답에 관한 정보를 전혀 주지 않는다. DP 설계는 보호수준만 최소화하는 문제가 아니라 요구한 보호수준