

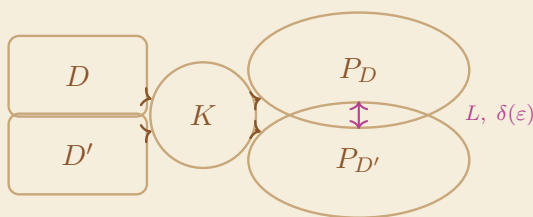
Enjoying Math Series

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

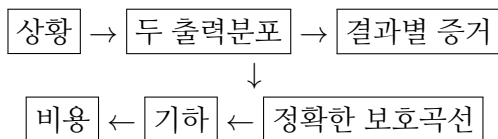


$$L(y) = \log \frac{dP}{dQ}(y) \quad \delta(\epsilon) = \sup_A \{P(A) - e^\epsilon Q(A)\}$$

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.

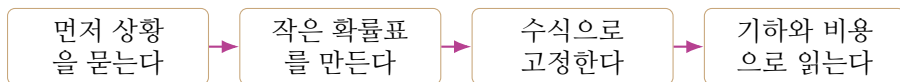
이 책을 사용하는 방법

먼저 묻기

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: ϵ 이 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. ϵ 은 그 비가 얼마나 크게 달라질 수 있는지를 로그척도로 제한합니다.

잠깐, 이렇게 생각해 봅시다

독자: δ 는 개인정보 보호가 실패할 확률인가요?

이 책: 일반적으로 그렇게만 말하면 부정확합니다. δ 는 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 읽는데 필요한 최소한만 제3장에서 도입한다.

완독의 의미

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

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

Contents

머리말	iii
이 책을 사용하는 방법	iv
I 한 사람의 흔적을 확률로 계산하다	1
1 무작위로 대답하면 왜 개인을 보호할 수 있는가	2
1.1 두 개의 가능한 세상	2
1.2 같은 결과가 나올 확률을 비교하기	3
1.3 정확도도 함께 계산하기	4
1.4 한 사람에서 데이터베이스로	5
1.5 일반적인 DP 정의	5
1.6 공개 뒤의 가공은 무엇을 바꾸는가	6
1.7 첫 번째 Python 계산	6
1.8 이 장의 다섯 질문	7
1.9 연습문제	7
2 답에 잡음을 더하면 왜 안전해지는가	9
2.1 민감도: 한 사람이 답을 얼마나 바꾸는가	9
2.2 Laplace mechanism	10
2.3 Laplace 잡음의 비용	11
2.4 여러 좌표와 ℓ_1 기하	11
2.5 Gaussian mechanism은 왜 순수 DP가 아닌가	12
2.6 Approximate DP	12

2.7	고전적 Gaussian 보정과 정확한 보정	13
2.8	Clipping은 무엇을 하는가	13
2.9	후처리와 denoising	14
2.10	Python 검산	14
2.11	이 장의 다섯 질문	14
2.12	연습문제	15
II 개인정보가 새는 양을 정확히 계산하다		16
3	결과 하나가 어느 상황을 더 지지하는가	17
3.1	두 출력분포를 고정하기	17
3.2	Privacy loss random variable	18
3.3	평균 privacy loss는 KL 발산이다	18
3.4	사건 하나에서 남는 차이	19
3.5	최적사건을 직접 찾기	19
3.6	Privacy profile	20
3.7	δ 는 왜 실패확률 하나가 아닌가	20
3.8	유한 출력에서 손으로 계산하기	21
3.9	Profile의 기본 모양	21
3.10	Profile과 가설검정	22
3.11	수치 계산	22
3.12	이 장의 다섯 질문	23
3.13	연습문제	23
4	두 상황을 맞히려는 사람과 Gaussian exact DP	24
4.1	두 shifted Gaussian	24
4.2	로그밀도비를 한 줄씩 계산하기	25

4.3	Exact $\delta(\varepsilon)$ 유도	25
4.4	다변수 Gaussian	26
4.5	Gaussian mechanism의 민감도와 r	26
4.6	가설검정으로 같은 식 읽기	27
4.7	EER와 DP를 혼동하지 않기	28
4.8	Renyi divergence로 모멘트를 보기	28
4.9	상계와 exact 계산 비교	28
4.10	Python으로 exact profile 계산하기	29
4.11	정보기하적 예고	30
4.12	이 장의 다섯 질문	30
4.13	연습문제	30
III 여러 번 공개하고 가장 좋은 방법을 선택하다		32
5	작은 단서를 여러 번 공개하면 어떻게 되는가	33
5.1	두 번의 독립 공개	33
5.2	기본 합성	34
5.3	적응적 공개의 chain rule	35
5.4	Privacy-loss distribution과 convolution	35
5.5	Renyi accountant	36
5.6	고급 합성의 직관	36
5.7	Hypothesis testing과 exact composition	37
5.8	Subsampling은 왜 보호를 증폭하는가	37
5.9	Privacy odometer와 filter	38
5.10	DP와 순차인증을 구분하기	38
5.11	연구실: binary extremality	39
5.12	계산 실습의 기록 형식	39

5.13	이 장의 다섯 질문	40
5.14	연습문제	40
6	같은 보호수준에서 어떤 방법이 가장 정확한가	41
6.1	먼저 비용의 단위를 정하기	41
6.2	가장 작은 최적화 문제	42
6.3	Gaussian 공분산을 설계변수로	42
6.4	왜 비등방성 잡음이 필요한가	43
6.5	Private direction과 utility direction	43
6.6	일반화 고유벡터의 역할	44
6.7	Bregman divergence로 utility를 재기	45
6.8	Matrix mechanism의 핵심	45
6.9	DP-SGD에서의 세 비용	45
6.10	최적설계의 실험 절차	46
6.11	VR gaze 자료에서 DP가 적용되는 위치	46
6.12	이 장의 다섯 질문	47
6.13	연습문제	47
IV	차등정보보호를 정보기하로 다시 읽다	48
7	KL의 굽어짐에서 Fisher-Rao 기하를 발견하다	49
7.1	분포를 점으로 놓기	49
7.2	Score와 Fisher 정보	50
7.3	정규분포: Fisher 정의에서 직접 계산하기	50
7.4	같은 결과를 KL Hessian에서 얻기	51
7.5	선요소와 쌍곡평면	52
7.6	베르누이분포에서 다시 확인하기	53

7.7	제공근 사상과 Hellinger affinity	53
7.8	표본 수와 계량	54
7.9	메커니즘 뒤의 Fisher 정보	54
7.10	KL data processing과 국소화	54
7.11	DP와 만나는 지점	55
7.12	이 장의 다섯 질문	55
7.13	연습문제	56
8	Fisher 정보가 작으면 DP도 좋은가	57
8.1	Gaussian에서는 왜 exact한가	57
8.2	일반 경로의 Fisher energy	58
8.3	왜 이차모멘트만으로 tail을 정할 수 없는가	59
8.4	Tail bound에서 DP로 가는 기본 보조정리	59
8.5	Fisher energy에서 κ, v 를 얻으려면	60
8.6	목표로 삼을 local-to-global 정리	60
8.7	Markov kernel과 contraction coefficient	61
8.8	Ollivier – Ricci 곡률	61
8.9	Ricci 기반 model comparison	62
8.10	Path-space와 SDE	63
8.11	DP·인증·정보기하의 경계	63
8.12	연구 주장 체크리스트	63
8.13	이 장의 다섯 질문	64
8.14	연습문제	64
A	확률론과 미적분 최소 도구상자	65
A.1	확률변수와 분포는 무엇이 다른가	65
A.2	밀도, 절대연속성, Radon – Nikodym 미분	65

A.3	조건부기대와 정보의 압축	66
A.4	적률생성함수와 sub-Gaussian 꼬리	66
A.5	Taylor 전개와 Hessian	67
A.6	행렬과 Mahalanobis 길이	67
A.7	수치 계산에서 지켜야 할 세 가지	67
B	핵심 정리 증명 노트	68
B.1	post-processing 정리	68
B.2	Laplace mechanism	68
B.3	hockey-stick divergence의 사건 표현	69
B.4	Gaussian exact profile	69
B.5	Fisher information contraction	69
B.6	Renyi DP에서 (ϵ, δ) -DP로	70
C	재현 실험실	71
C.1	실험 전에 고정할 계약	71
C.2	실험 1: randomized response	71
C.3	실험 2: Laplace와 Gaussian 비교	72
C.4	실험 3: privacy-loss distribution 합성	72
C.5	실험 4: DP-SGD accountant	72
C.6	실험 5: 비등방성 Gaussian mechanism	72
C.7	실험 6: gaze 인 증은 DP 실험과 무엇이 다른가	73
D	연구문제 지도	74
D.1	기하적 mechanism 설계	74
D.2	Rao 길이에서 global DP로	74
D.3	Ollivier – Ricci contraction	75
D.4	exact composition과 binary extremality	75

D.5	무엇을 “sharp”라고 부를 수 있는가	75
E	연습문제 해설	77
E.1	1장	77
E.2	2장	77
E.3	3장	77
E.4	4장	78
E.5	5장	78
E.6	6장	78
E.7	7장	78
E.8	8장	79
F	기호와 참고문헌 길잡이	80
F.1	주요 기호	80
F.2	읽는 순서별 핵심 참고문헌	80
F.3	마지막 점검표	83

PART 1

한 사람의 흔적을 확률로 계산하다

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

따라서

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

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

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

실제 답이 0이면 두 확률은 뒤바뀐다.

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

	$Y = 0$	$Y = 1$
$X = 0$	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 Y 는 X 와 무관한 동전이므로 $\varepsilon = 0$ 이다. $p \rightarrow 1$ 이면 사실을 거의 그대로 공개하므로 $\varepsilon \rightarrow \infty$ 다.

여기서 자주 틀립니다

ε 은 “유출된 개인정보의 개수”가 아니다. 두 입력에서 같은 결과가 나올 상대적 가능성이 최악의 경우 얼마나 바뀌는지를 로그적으로 표현한 값이다. 단위가 없고, 작을수록 두 상황을 구별하기 어렵다.

1.3 정확도도 함께 계산하기

실제 답을 맞게 공개할 확률은 p 이고 틀리게 공개할 확률은 $1 - p$ 다. ε 으로 p 를 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