Adaptive Data Analysis

Thomas Steinke
tsteinke@seas.harvard.edu

April 21, 2016

Abstract

These lecture notes are based on [BNS+16] and were compiled for a guest lecture in the course CS229r "Information Theory in Computer Science" taught by Madhu Sudan at Harvard University in Spring 2016.

Menu for today's lecture:

- Motivation
- Model
- Overfitting & comparison to non-adaptive data analysis
- What can we do adaptively?
- KL Divergence recap
- Proof
- Differential privacy (time permitting)

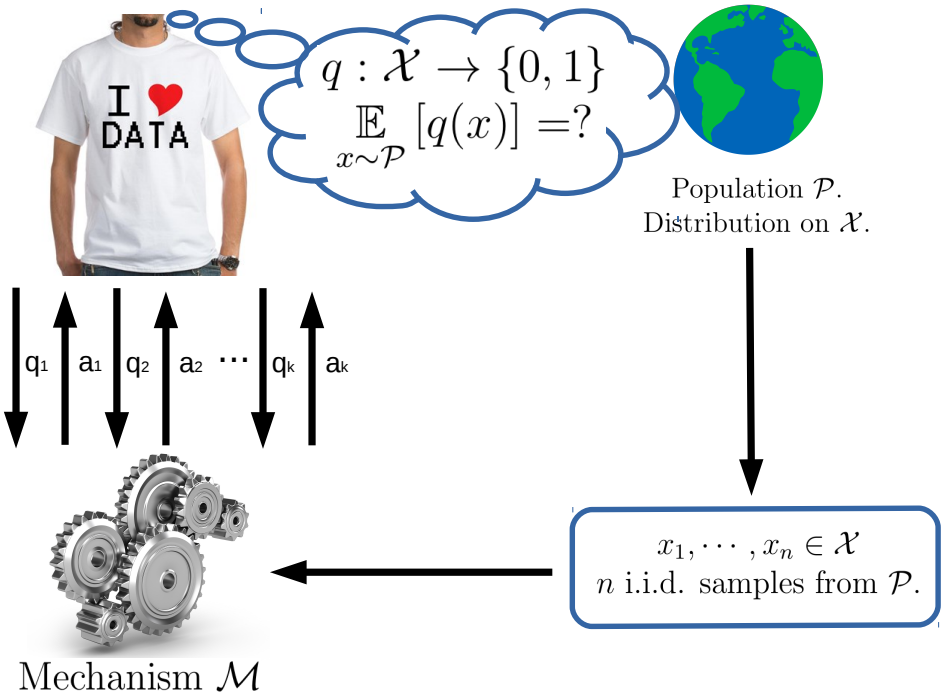
1 Motivation

Ideally, science is non-adaptive — that is, hypotheses are formulated before the data is collected and the hypothesis tested. In particular, a dataset should only be used once. However, in practice, datasets are used repeatedly, with previous analyses informing subsequent analyses. This may lead to erroneous conclusions. [Har15b, DFH+15b]

Adaptive use of data has been identified as a major problem in empirical sciences. One proposed solution is pre-registration, where scientists commit to their methodology before examining the data. In today's lecture we will consider a different approach. Namely, we will consider ways of answering adaptively-chosen queries while preserving validity.

2 Model

First we must define a formal model for this problem.



There is an unknown population $\mathcal{P}$ (modelled as a distribution on a data universe $\mathcal{X}$) and an analyst $\mathcal{A}$ who wants to know about the population. The analyst's queries are of the simple form "what fraction of the population satisfies predicate $q : \mathcal{X} \to \{0,1\}$." To answer his queries, the analyst collects n independent samples from $\mathcal{P}$. For each query q_j she wants an answer $a_j \approx q_j(\mathcal{P}) = \mathbb{E}_{z\sim\mathcal{P}}[q_j(z)]$, and tolerates an additive error α. The samples are given to a mechanism $\mathcal{M}$ which must use them to answer the queries.

Crucially, the analyst's queries are chosen depending on the answers given by $\mathcal{M}$ to previous queries. Thus the queries and the sample are not independent.

3 Cf. Non-adaptive Data Analysis

In the non-adaptive setting the analyst must specify the entire sequence of queries $q_1, \cdots, q_k$ before receiving any of the answers $a_1, \cdots, a_k$. Thus the queries and the samples are independent. We can view the queries as being fixed before the samples are drawn.

By Hoeffding's inequality, for any query $q : \mathcal{X} \to \{0,1\}$,

$$\mathbb{P}_{x_{1\cdots n}\overset{\text{i.i.d.}}{\sim}\mathcal{P}}[|q(x) - q(\mathcal{P})| > \alpha] = \mathbb{P}_{x_{1\cdots n}\overset{\text{i.i.d.}}{\sim}\mathcal{P}}\left[\left|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - \mathbb{E}_{z\sim\mathcal{P}}[q(z)]\right| > \alpha\right] \leq 2 \cdot e^{-2\alpha^2 n}.$$

We refer to $q(x) = \frac{1}{n}\sum_{i=1}^{n} q(x_i)$ as the empirical answer to q and $q_j(\mathcal{P}) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[q_j(z)]$ as the true or population answer to q. By a union bound, for any fixed sequence of queries $q_1, \cdots, q_k : \mathcal{X} \to \{0,1\}$,

$$\mathop{\mathbb{P}}_{x_{1\cdots n} \overset{\text{i.i.d.}}{\sim} \mathcal{P}} \left[\max_{j=1}^{k} |q_j(x) - q_j(\mathcal{P})| > \alpha\right] \leq 2k \cdot e^{-2\alpha^2 n}.$$

Thus, if $n \geq \log(2k/\beta)/2\alpha^2$, then $\mathop{\mathbb{P}}_{x_{1\cdots n} \overset{\text{i.i.d.}}{\sim} \mathcal{P}} \left[\max_{j=1}^{k} |q_j(x) - q_j(\mathcal{P})| > \alpha\right] \leq \beta$. In other words, to answer k non-adaptive queries with accuracy α, we need $n = O(\log(k)/\alpha^2)$ samples.

3.1 What goes wrong in the adaptive setting?

If the queries $q_1, \cdots, q_k$ are adaptive, then we can no longer take a union bound as the queries depend on the samples. We would have to union bound over all queries the analyst *could* have asked, rather than just the queries he did ask.

We can give an easy example where giving empirical answers ($a_j = q_j(x)$) fails: Suppose $\mathcal{X} = \{1, \cdots, k\}$ and $\mathcal{P}$ is the uniform distribution on $\mathcal{X}$. Now let

$$q_j(\ell) = \begin{cases} 1 & j = \ell \\ 0 & j \neq \ell \end{cases}$$

and suppose the analyst is given $a_j = q_j(x)$ for $j = 1 \cdots k$. Then the analyst learns exactly what the samples $x_1, \cdots, x_n$ are. Now the analyst can ask q_{k+1}, where

$$q_{k+1}(\ell) = \begin{cases} 1 & \ell \in \{x_1, \cdots x_n\} \\ 0 & \ell \notin \{x_1, \cdots, x_n\} \end{cases}. \tag{1}$$

Then $q_{k+1}(x) = 1$, but $q_{k+1}(\mathcal{P}) \leq n/k$. So, unless $n = \Omega(k)$, empirical answer fail to be accurate.

What went wrong here? The analyst chose q_{k+1} adaptively and q_{k+1} overfitted her sample. A way to view this example is as follows. Suppose an alien visited earth and met Thomas, Madhu, and Badih. It might wonder wheter all humans are named either Thomas, Madhu, or Badih. Testing this hypothesis on its dataset confirms that this is the case. However, this conclusion is clearly not valid of the population as a whole. This example may seem contrived, but the underlying idea is useful [Har15a].

So we see that adaptive and non-adaptive data analysis are different.

4 What Can We Do Adaptively?

Theorem 1 ([DFH+15a, BNS+16]). *There exists a mechanism $\mathcal{M}$ that takes $n = \tilde{O}(\sqrt{k}/\alpha^2)$ samples from an unknown distribution $\mathcal{P}$ on $\mathcal{X}$ and answers k adaptively-chosen queries $q_1, \cdots, q_k : \mathcal{X} \to \{0,1\}$ with $a_1, \cdots, a_k \in [0,1]$ such that*

$$\mathbb{P}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})| > \alpha\right] \leq \frac{1}{100},$$

where the probability is over both the sample and the randomness of the mechanism.

Requiring $n = \tilde{O}(\sqrt{k}/\alpha^2)$ samples is much worse than $n = O(\log(k)/\alpha^2)$ in the non-adaptive setting. Surprisingly, this is inherent — in the sense that there is an almost-matching lower bound, namely $n = \Omega(\sqrt{k}/\alpha)$ [HU14, SU15b]. (This lower bound holds assuming $|\mathcal{X}| \geq 2^k$. Alternatively, this is a computational lower bound for mechanisms that are not powerful enough to break cryptography with seed length $\log |\mathcal{X}|$. Moreover, it is known that, when $|\mathcal{X}| \leq 2^{\tilde{o}(k)}$ and the mechanism is allowed exponential time, we can do better.) This shows an exponential separation between the adaptive and non-adaptive settings.

5 Proof

The mechanism $\mathcal{M}$ in Theorem 1 is extremely simple: Given a sample $x_1, \cdots, x_n \in \mathcal{X}$, for each query $q_j : \mathcal{X} \to \{0, 1\}$, it returns a random answer

$$a_j \sim \mathcal{N}\left(q_j(x), \sigma^2\right), \tag{2}$$

for an appropriate value of σ^2.

Why? The intuition is that it is hard to overfit noisy data. The addition of noise prevents the analyst from identifying the samples, which means she cannot overfit à la (1).

5.1 Key steps

We want to show

$$\mathbb{P}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})| > \alpha\right] \leq \frac{1}{100},$$

where the probability is taken over the sample $x_1, \cdots, x_n \sim \mathcal{P}$ as well as the randomness of $\mathcal{M}$ and $\mathcal{A}$. By Markov's inequality, it suffices to show that

$$\mathbb{E}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})|\right] \leq \frac{\alpha}{100}.$$

The maximum is annoying to work with. So we simply pick out the worst query. We define a function f which takes the entire transcript and picks out a single worst query that maximizes $|a_{j_*} - q_{j_*}(\mathcal{P})|$. Moreover, f may negate the query to ensure that $a_* - q_*(\mathcal{P}) \geq 0$. Formally, define $f : (Q \times [0, 1])^k \to Q \times [0, 1]$ by

$$f(q_1, a_1, \cdots, q_k, a_k) = \begin{cases} (q_{j_*}, a_{j_*}) & a_{j_*} - q_j(\mathcal{P}) \geq 0 \\ (1 - q_{j_*}, 1 - a_{j_*}) & a_{j_*} - q_{j_*}(\mathcal{P}) < 0 \end{cases}, \qquad \text{where} \qquad j_* = \operatorname*{armax}_{j \in \{1, \cdots, k\}} |a_j - q_j(\mathcal{P})|.$$

This allows us to remove the maximum:

$$\mathbb{E}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})|\right] = \mathbb{E}\left[a_* - q_*(\mathcal{P}) | (q_*, a_*) = f(q_1, a_1, \cdots, q_k, a_k)\right].$$

We use the triangle (in)equality:

$$\mathbb{E}\left[a_* - q_*(\mathcal{P})\right] \leq \mathbb{E}\left[a_* - q_*(x)\right] + \mathbb{E}\left[q_*(x) - q_*(\mathcal{P})\right],$$

which may be paraphrased as

$$\text{true error} \;\leq\; \text{empirical error} \;+\; \text{generalization error}.$$

Now we bound the terms separately. The emprical error is easy to bound:

$$\mathbb{E}\left[a_* - q_*(x)\right] \leq \mathbb{E}\left[\max_{j=1}^{k} |a_j - q_j(x)|\right] = \mathbb{E}\left[\max_{j=1}^{k} |\mathcal{N}(0, \sigma^2)|\right] \leq 2\sigma\sqrt{\log k}.$$

The generalization error is more involved, but we will show that

$$\mathbb{E}\left[q_*(x) - q_*(\mathcal{P})\right] \leq \frac{\sqrt{k}}{2n\sigma}. \tag{3}$$

Thus we have

$$\mathbb{E}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})|\right] \leq 2\sigma\sqrt{\log k} + \frac{\sqrt{k}}{2n\sigma}.$$

To minimize this we set $\sigma^2 = \sqrt{k}/4n\sqrt{\log k}$, giving

$$\mathbb{E}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})|\right] \leq 2\sqrt{\frac{\sqrt{k\log k}}{n}}.$$

So, to prove Theorem 1, we just need

$$n \geq \frac{\sqrt{k\log k}}{(\alpha/200)^2}.$$

It just remains to prove (3).

5.2 Recap: KL Divergence

The Kullback-Leibler divergence between distributions P and Q is defined as

$$\mathrm{D}_{\mathrm{KL}}\left(P\|Q\right) = \mathop{\mathbb{E}}_{x\sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right].$$

It satisfies the following useful properties:

- Non-negativity: $\mathrm{D}_{\mathrm{KL}}\left(P\|Q\right) \geq 0$. (But it is not symmetric i.e. $\mathrm{D}_{\mathrm{KL}}\left(P\|Q\right) \neq \mathrm{D}_{\mathrm{KL}}\left(Q\|P\right)$ in general.)

- Pinsker's inequality: $\Delta(P,Q) \leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}\left(P\|Q\right)}$. (Note that I'm dropping the $\ln 2$ that usually appears in this inequality. This just means I'm defining KL divergence using the natural logarithm, rather than base-2 logarithm. So the units become *nats* rather than *bits*. Today all logarithms will be natural, even though this is an information theory course.)

  Alternatively: Let X and Y be random variables in $[0,1]$. Then

$$\left|\mathbb{E}\left[X\right] - \mathbb{E}\left[Y\right]\right| \leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}\left(P_X\|Q_Y\right)}.$$

- Chain rule: Let $P \times P'$ and $Q \times Q'$ denote product distributions. Then

$$\mathrm{D}_{\mathrm{KL}}\left(P \times P'\|Q \times Q'\right) = \mathrm{D}_{\mathrm{KL}}\left(P\|Q\right) + \mathrm{D}_{\mathrm{KL}}\left(P'\|Q'\right)$$

  More generally, let P be a distribution on two variables, let P' be the marginal distribution on the first variable and let P'_x be the conditional distribution on the second variable given that the first variable is x. Define Q, Q', and Q'_x likewise. Then

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(P\|Q\right) =& \mathop{\mathbb{E}}_{(x,y)\sim P}\left[\log\left(\frac{P(x,y)}{Q(x,y)}\right)\right] \\
=& \mathop{\mathbb{E}}_{x\sim P'}\left[\mathop{\mathbb{E}}_{y\sim P'_x}\left[\log\left(\frac{P'(x)P'_x(y)}{Q'(x)Q'_x(y)}\right)\right]\right] \\
=& \mathop{\mathbb{E}}_{x\sim P'}\left[\log\left(\frac{P'(x)}{Q'(x)}\right) + \mathop{\mathbb{E}}_{y\sim P'_x}\left[\log\left(\frac{P'_x(y)}{Q'_x(y)}\right)\right]\right] \\
=& \mathrm{D}_{\mathrm{KL}}\left(P'\|Q'\right) + \mathop{\mathbb{E}}_{x\sim P'}\left[\mathrm{D}_{\mathrm{KL}}\left(P'_x\|Q'_x\right)\right] \\
\leq& \mathrm{D}_{\mathrm{KL}}\left(P'\|Q'\right) + \max_x \mathrm{D}_{\mathrm{KL}}\left(P'_x\|Q'_x\right).
\end{aligned}$$

- Data processing inequality: Let f be a function. Then

$$\mathrm{D}_{\mathrm{KL}}\left(f(P)\|f(Q)\right) \leq \mathrm{D}_{\mathrm{KL}}\left(P\|Q\right), \tag{4}$$

  where $f(P)$ and $f(Q)$ represent the distribution of the output of f when given a random sample from P or Q as input. This also holds if f is a randomized function.

- For all $a, b, \sigma \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(a,\sigma^2)\middle\|\mathcal{N}(b,\sigma^2)\right) =& \mathop{\mathbb{E}}_{x\sim\mathcal{N}(a,\sigma^2)}\left[\log\left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-a)^2/2\sigma^2}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-b)^2/2\sigma^2}}\right)\right] \\
=& \mathop{\mathbb{E}}_{x\sim\mathcal{N}(a,\sigma^2)}\left[\frac{1}{2\sigma^2}\left(-(x-a)^2 + (x-b)^2\right)\right] \\
=& \mathop{\mathbb{E}}_{x\sim\mathcal{N}(a,\sigma^2)}\left[\frac{1}{2\sigma^2}\left(a-b\right)\left(2x-a-b\right)\right] \\
=& \frac{(a-b)^2}{2\sigma^2}
\end{aligned}$$

5.3 The transcript

Consider a fixed analyst $\mathcal{A}$ and the mechanism $\mathcal{M}$ from (2). Given samples $x_1, \cdots, x_n \in \mathcal{X}$, we can simulate $\mathcal{A}$ and $\mathcal{M}$ interacting conditioned on these samples being drawn. This yields a randomized function mapping the samples to a sequence of queries and answers. Call this the transcript function $T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}} : \mathcal{X}^n \to (Q \times [0,1])^k$.

We now formalise the intuition that $\mathcal{M}$ does not permit $\mathcal{A}$ to identify any samples. Formally, we show that $T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}$ is stable in the sense that changing one sample does not affect the output distribution much.

Lemma 2. *Let $x_1, \cdots, x_n, z \in \mathcal{X}$. Then*

$$\mathrm{D}_{KL}\left(T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}(x_1, \cdots, x_n) \middle\| T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}(x_1, \cdots, x_{i-1}, z, x_{i+1}, \cdots, x_n)\right) \leq \frac{k}{2n^2\sigma^2}.$$

Proof. We show this by induction on k. Denote

$$T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}(x) = T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}(x_1, \cdots, x_n) = (q_1, a_1, q_2, a_2, \cdots, q_k, a_k)$$

and

$$T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}(x_{-i}, z) = T_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{M}}(x_1, \cdots, x_{i-1}, z, x_{i+1}, \cdots, x_n) = (q'_1, a'_1, q'_2, a'_2, \cdots, q'_k, a'_k).$$

We may assume inductively that

$$\mathrm{D}_{\mathrm{KL}}\left((q_1, a_1, \cdots, q_{k-1}, a_{k-1}) \middle\| (q'_1, a'_1, \cdots, q'_{k-1}, a'_{k-1})\right) \leq \frac{k-1}{2n^2\sigma^2}.$$

Firstly q_k is chosen by $\mathcal{A}$ given only $q_1, a_1, \cdots, q_{k-1}, a_{k-1}$ – that is, it does not depend on x other than through the transcript of the first $k-1$ rounds. Thus, by the data processing inequality,

$$\mathrm{D}_{\mathrm{KL}}\left((q_1, a_1, \cdots, q_{k-1}, a_{k-1}, q_k) \middle\| (q'_1, a'_1, \cdots, q'_{k-1}, a'_{k-1}, q'_k)\right) \leq \frac{k-1}{2n^2\sigma^2}.$$

Finally, by the chain rule,

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left((q_1, a_1, \cdots, q_k, a_k) \| (q'_1, a'_1, \cdots, q'_k, a'_k)\right) \\
&\leq \mathrm{D}_{\mathrm{KL}}\left((q_1, a_1, \cdots, q_{k-1}, a_{k-1}, q_k) \middle\| (q'_1, a'_1, \cdots, q'_{k-1}, a'_{k-1}, q'_k)\right) + \max_{q_k} \mathrm{D}_{\mathrm{KL}}\left(a_j \middle\| a'_j\right) \\
&\leq \frac{k-1}{2n^2\sigma^2} + \max_{q_k} \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(q_k(x), \sigma^2) \middle\| \mathcal{N}(q_k(x_{-i}, z), \sigma^2)\right) \\
&\leq \frac{k-1}{2n^2\sigma^2} + \max_{q_k} \frac{(q_k(x) - q_k(x_{-i}, z))^2}{2\sigma^2} \\
&\leq \frac{k-1}{2n^2\sigma^2} + \max_{q_k} \frac{(q_k(x_i) - q_k(z))^2}{2n^2\sigma^2} \\
&\leq \frac{k}{2n^2\sigma^2}.
\end{aligned}$$

□

5.4 Overfitting

Now that we have elucidated the key property of $\mathcal{M}$, we show that this property can be used to bound the generalization error (3).

Lemma 3 (Stability Prevents Overfitting)**.** *Suppose*

$$\mathrm{D}_{KL}\left(T(x) \| T(x_{-i}, z)\right) \leq 2\varepsilon^2$$

for all $x \in \mathcal{X}^n$, $i \in [n]$, *and* $z \in \mathcal{X}$. *Then*

$$\mathbb{E}\left[q_*(x) - q_*(\mathcal{P}) | q_*\right) = f(T(x))] \leq \varepsilon$$

for all f, *where the expectation is taken over the randomness of* T *and* $x_1, \cdots, x_n \sim \mathcal{P}$.

Combining Lemmas 2 and 3 yields (3): $\mathrm{D}_{\mathrm{KL}}\left(T_{\mathcal{A} \overrightarrow{\leftarrow} \mathcal{M}}(x) \middle\| T_{\mathcal{A} \overrightarrow{\leftarrow} \mathcal{M}}(x_{-i}, z)\right) \leq k/2n^2\sigma^2 = 2\varepsilon^2$, whence $\mathbb{E}\left[q_*(x) - q_*(\mathcal{P})\right] \leq \varepsilon = \sqrt{k/4n^2\sigma^2}$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[q_*(x) - q_*(\mathcal{P}) | q_* = f(T(x))\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[q_*(x_i) - q_*(\mathcal{P}) | q_* = f(T(x))\right] && \text{(by linearity of expectation)} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[q_*(x_i) - q_*(\mathcal{P}) | q_* = f(T(x_{-i}, z))\right] \\
&\quad + \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}\left(f(T(x)) \| f(T(x_{-i}, z))\right)} && \text{(by Pinsker's inequality)} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[q_*(x_i) - q_*(\mathcal{P}) | q_* = f(T(x_{-i}, z))\right] + \varepsilon && \text{(by assumption and (4))} \\
&= 0 + \varepsilon,
\end{aligned}$$

where the final equality follows from the fact that x_i and $(q_*, a_*) = f(T(x_{-i}, z))$ are independent.

This yields one side of the lemma and the other side is symmetric. □

6 Differential Privacy

I have presented these results in terms of KL divergence. However, these ideas were originally formulated using *differential privacy*:

Definition 4 (Differential Privacy [DMNS06, DKM+06])**.** *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is said to satisfy (ε, δ)-differential privacy if, for all $x, x' \in \mathcal{X}^n$ differing in one entry,*

$$\forall f : \mathcal{Y} \to \{0,1\} \qquad \mathop{\mathbb{P}}_{M}[f(M(x)) = 1] \leq e^{\varepsilon} \mathop{\mathbb{P}}_{M}[f(M(x')) = 1] + \delta.$$

The original motivation for this definition was data privacy: Imagine that x is a database containing sensitive personal information and $M(x)$ is being released publicly. For example, x may be medical records and $M(x)$ may be the published outcome of medical research. We want to ensure that the publicly released output does not reveal any sensitive personal information, such as an individual's medical condition. Differential privacy provides a mathematical way to formalise this requirement. If one person's data were removed from x or replaced, we would have x' instead and differential privacy guarantees that the outcomes $M(x)$ and $M(x')$ are indistinguishable — that is, that person's information is not revealed. For more discussion about the motivation of differential privacy, as well as various results about differential privacy, read Dwork's survey [Dwo06] or the textbook [DR14] on the subject.

Regardless of the original motivations for differential privacy, the definition can be repurposed for adaptive data analysis. Indeed it can be used to give a sharper analysis of the mechanism $\mathcal{M}$:

Theorem 5. *There exists a mechanism $\mathcal{M}$ that takes $n = O(\sqrt{k}\log(k/\alpha\beta)/\alpha^2)$ samples from an unknown distribution $\mathcal{P}$ on $\mathcal{X}$ and answers k adaptively-chosen queries $q_1, \cdots, q_k : \mathcal{X} \to \{0,1\}$ with $a_1, \cdots, a_k \in [0,1]$ such that*

$$\mathbb{P}\left[\max_{j=1}^{k} |a_j - q_j(\mathcal{P})| > \alpha\right] \leq \beta,$$

where the probability is over both the sample and the randomness of the mechanism.

Note that the logarithmic term in the above theorem can be improved [SU15a]. This theorem follows from the following lemmata.

Lemma 6 (Analog of Lemma 2)**.** *The mechanism $\mathcal{M}$ given by (2)satisfies (ε, δ)-differential privacy for all $\delta > 0$ and*

$$\varepsilon = \frac{k}{2n^2\sigma^2} + \frac{\sqrt{2k\log(1/\delta)}}{n\sigma}.$$

(More precisely, $T_{\mathcal{A} \overset{\to}{\leftarrow} \mathcal{M}}$ satisfies the above differential privacy bound for all $\mathcal{A}$.)

Lemma 7 (Analog of Lemma 3)**.** *Suppose T satisfies (ε, δ)-differential privacy with $0 < \delta < \varepsilon/4 < 1/12$. If $n \geq \log(4\varepsilon/\delta)/\varepsilon^2$, then*

$$\mathbb{P}\left[|q_*(x) - q_*(\mathcal{P})| > 10\varepsilon \mid q_* = f(T(x))\right] \leq \frac{\delta}{\varepsilon}$$

for any function f, where the probability is taken over n i.i.d. samples $x_1, \cdots, x_n \sim \mathcal{P}$ as well as the randomness of T.

Lemma 7 is sharper than Lemma 3 in that it gives high-probability bounds, rather than just a bound on the expectation. Lemma 7 can also be applied to a wide variety of mechanisms from the differential privacy literature. In particular, we can attain sample complexity $n = \tilde{O}(\sqrt{\log |\mathcal{X}|} \cdot \log(k)/\alpha^3)$ [HR10], which is an improvement if $\log |\mathcal{X}| \ll \alpha^2 k$ (at the price of computational efficiency).

References

[BNS+16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *ACM Symposium on the Theory of Computing (STOC)*, 2016. http://arxiv.org/abs/1511.02513.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *ACM Symposium on the Theory of Computing (STOC)*. ACM, June 2015. http://arxiv.org/abs/1411.2664.

[DFH+15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, June 2015. http://science.sciencemag.org/content/349/6248/636.full.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503, 2006. http://www.iacr.org/cryptodb/archive/2006/EUROCRYPT/2319/2319.pdf.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, March 4-7 2006. http://www.iacr.org/cryptodb/archive/2006/TCC/3650/3650.pdf,.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. https://www.cis.upenn.edu/ aaroth/Papers/privacybook.pdf.

[Dwo06] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, pages 1–12, 2006. http://dx.doi.org/10.1007/11787006_1.

[Har15a] Moritz Hardt. Competing in a data science contest without reading the data, March 2015. http://blog.mrtz.org/2015/03/09/competition.html.

[Har15b] Moritz Hardt. The reusable holdout: Preserving validity in adaptive data analysis, August 2015. http://googleresearch.blogspot.com/2015/08/the-reusable-holdout-preserving.html.

[HR10] Moritz Hardt and Guy Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proc. 51st Foundations of Computer Science (FOCS)*, pages 61–70. IEEE, 2010. http://ww.mrtz.org/papers/HR10mult.pdf.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*. IEEE, October 19-21 2014. http://arxiv.org/abs/1408.1655.

[SU15a] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *CoRR*, abs/1501.06095, 2015. http://arxiv.org/abs/1501.06095.

[SU15b] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of The 28th Conference on Learning Theory (COLT'15)*, pages 1588–1628, 2015. http://arxiv.org/abs/1410.1228.