Two-party differential privacy and deterministic extraction from Santha-Vazirani sources

Grigory Yaroslavtsev

Pennsylvania State University
http://www.cse.psu.edu/~gyy5026

December 24, 2010

Plan

1. Introduction
2. Two party DP and Santha-Vazirani sources
3. Deterministic extraction from Santha-Vazirani sources
4. Lower bound for inner product
5. Limitations of the extractor technique

Differential privacy in client-server setting

For strings $x, y \in \{0,1\}^n$, let $|x - y|_H$ denote Hamming distance.
A mechanism M on $\{0,1\}^n$ is a family of probability distributions $\{\mu_x \colon x \in \{0,1\}^n\}$ on $\mathbb{R}$.

Differential privacy in client-server setting

For strings $x, y \in \{0,1\}^n$, let $|x-y|_H$ denote Hamming distance.
A mechanism M on $\{0,1\}^n$ is a family of probability distributions $\{\mu_x \colon x \in \{0,1\}^n\}$ on $\mathbb{R}$.

Definition (Differential privacy)

The mechanism is ϵ-differentially private if for any x and y such that $|x-y|_H = 1$ and any measurable subset $S \subset \mathbb{R}$ we have

$$\mu_x(S) \leq e^{\epsilon}\mu_y(S)$$

Differential privacy in client-server setting

For strings $x, y \in \{0,1\}^n$, let $|x-y|_H$ denote Hamming distance.
A mechanism M on $\{0,1\}^n$ is a family of probability distributions $\{\mu_x \colon x \in \{0,1\}^n\}$ on $\mathbb{R}$.

Definition (Differential privacy)

The mechanism is ϵ-differentially private if for any x and y such that $|x-y|_H = 1$ and any measurable subset $S \subset \mathbb{R}$ we have

$$\mu_x(S) \leq e^{\epsilon} \mu_y(S)$$

Two-party differential privacy

- $VIEW_P^A(x, y) = (T, R_{AB}, R_A)$ random variable, where the probability space is public and private randomness of both parties.
- For each x, $VIEW_P^A(x, \cdot)$ is a mechanism over the y's.
- $VIEW_P^B(\cdot, y)$ is defined similarly.

Two-party differential privacy

- $VIEW_P^A(x, y) = (T, R_{AB}, R_A)$ random variable, where the probability space is public and private randomness of both parties.
- For each x, $VIEW_P^A(x, \cdot)$ is a mechanism over the y's.
- $VIEW_P^B(\cdot, y)$ is defined similarly.

Definition (Differential privacy for two-party protocols)

Protocol $P(x, y)$ has ϵ-differential privacy if the mechanism $VIEW_P^A(x, \cdot)$ is ϵ-differentially private for all values of x and same holds for $VIEW_P^B(\cdot, y)$ and all values of y.

Hamming distance (independent setting)

Question

Suppose Bob knows that Alice's x comes from uniform distribution X, independent of Bob's distribution Y. How can he approximate $|x - y|_H$ up to an expected additive error $O(\sqrt{n})$ without any communication?

Hamming distance (independent setting)

Question

Suppose Bob knows that Alice's x comes from uniform distribution X, independent of Bob's distribution Y. How can he approximate $|x-y|_H$ up to an expected additive error $O(\sqrt{n})$ without any communication?

Answer

- *Just say $n/2$.*
- *W.l.o.g. assume that $y = 0^n$, then correct answer is $|x|_H$.*
- *$|x|_H$ is distributed by $B(n, 1/2)$.*
- *Using Hoeffding's inequality:* $\Pr\left[\left||x|_H - n/2\right| > c\sqrt{n}\right] < 2e^{-c^2}$.

Hamming distance (independent setting)

Question

Suppose Bob knows that Alice's x comes from uniform distribution X, independent of Bob's distribution Y. How can he approximate $|x - y|_H$ up to an expected additive error $O(\sqrt{n})$ without any communication?

Answer

- *Just say $n/2$.*
- *W.l.o.g. assume that $y = 0^n$, then correct answer is $|x|_H$.*
- *$|x|_H$ is distributed by $B(n, 1/2)$.*
- *Using Hoeffding's inequality:* $\Pr\left[||x|_H - n/2| > c\sqrt{n}\right] < 2e^{-c^2}$.

Exercise: How to do better than this?

Hamming distance (independent setting)

Question

Suppose Bob knows that Alice's x comes from uniform distribution X, independent of Bob's distribution Y. How can he approximate $|x-y|_H$ up to an expected additive error $O(\sqrt{n})$ without any communication?

Answer

- *Just say $n/2$.*
- *W.l.o.g. assume that $y = 0^n$, then correct answer is $|x|_H$.*
- *$|x|_H$ is distributed by $B(n, 1/2)$.*
- *Using Hoeffding's inequality:* $\Pr\left[||x|_H - n/2| > c\sqrt{n}\right] < 2e^{-c^2}$.

Exercise: How to do better than this?
Hint: Use randomized response.

Santha-Vazirani sources

Definition (α-unpredictable bit source)

For $\alpha \in [0, 1]$, random variable $X = (X_1, \ldots, X_n)$ taking values in $\{0, 1\}^n$ is an α-unpredictable bit source if for every $i \in [n]$, and every $x_1, \ldots, x_{i-1} \in \{0, 1\}^{i-1}$, we have

$$\alpha \leq \frac{Pr[X_i = 0 | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}]}{Pr[X_i = 1 | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}]} \leq 1/\alpha$$

Santha-Vazirani sources

Definition (α-unpredictable bit source)

For $\alpha \in [0,1]$, random variable $X = (X_1, \dots, X_n)$ taking values in $\{0,1\}^n$ is an α-unpredictable bit source if for every $i \in [n]$, and every $x_1, \dots, x_{i-1} \in \{0,1\}^{i-1}$, we have

$$\alpha \leq \frac{Pr[X_i = 0 | X_1 = x_1, \dots, X_{i-1} = x_{i-1}]}{Pr[X_i = 1 | X_1 = x_1, \dots, X_{i-1} = x_{i-1}]} \leq 1/\alpha$$

Properties

- *No string has probability mass greater than $1/(1+\alpha)^n$*
- *Min-entropy ($\min_x \log_2(1/Pr[X = x])$), is at least βn, where $\beta = \log_2(1+\alpha) \geq \alpha$.*

Santha-Vazirani sources

Definition (Strongly α-unpredictable bit source)

For $\alpha \in [0,1]$, a random variable $X = (X_1, \dots, X_n)$ taking values in $\{0,1\}^n$ is a strongly α-unpredictable bit source if for every $i \in [n]$, and every $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n \in \{0,1\}^{n-1}$, we have

$$\alpha \leq \frac{Pr[X_i = 0 | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \dots, X_n = x_n]}{Pr[X_i = 1 | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \dots, X_n = x_n]} \leq 1/\alpha$$

Two-party DP and Santha-Vazirani sources

Lemma (Two-party DP and Santha-Vazirani sources)

- *Let $P(x, y)$ be a ϵ-differentially private randomized protocol with inputs $x, y \in \{0,1\}^n$.*
- *Let X and Y be independent random variables uniformly distributed in $\{0,1\}^n$.*
- *Let random variable $T(X, Y)$ denote the transcript on input (X, Y).*

Then for every $t \in Supp(T)$, the random variables $X|_{T=t}$ and $Y|_{T=t}$ are independent strongly $e^{-\epsilon}$-unpredictable bit sources.

Two-party DP and Santha-Vazirani sources

Lemma (Two-party DP and Santha-Vazirani sources)

- *Let $P(x, y)$ be a ϵ-differentially private randomized protocol with inputs $x, y \in \{0,1\}^n$.*
- *Let X and Y be independent random variables uniformly distributed in $\{0,1\}^n$.*
- *Let random variable $T(X, Y)$ denote the transcript on input (X, Y).*

Then for every $t \in Supp(T)$, the random variables $X|_{T=t}$ and $Y|_{T=t}$ are independent strongly $e^{-\epsilon}$-unpredictable bit sources.

Proof

1. *Independence*
   Proof by induction on the number of rounds.
2. *Strong $e^{-\epsilon}$unpredictability (next slide)*

Proof of strong $e^{-\epsilon}$ unpredictability

Using Bayes' Rule and the uniformity of X:

$$\frac{Pr[X_i = 0 | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n, T = t]}{Pr[X_i = 1 | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n, T = t]} =$$

$$= \frac{Pr[T = t | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = 0, X_{i+1} = x_{i+1}, \ldots, X_n = x_n]}{Pr[T = t | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = 1, X_{i+1} = x_{i+1}, \ldots, X_n = x_n]} =$$

$$= \frac{Pr[T(x_1 \cdots x_{i-1} 0 x_{i+1} \cdots x_n, Y) = t]}{Pr[T(x_1 \cdots x_{i-1} 1 x_{i+1} \cdots x_n, Y) = t]}$$

By ϵ-differential privacy the latter ratio is between $e^{-\epsilon}$ and e^{ϵ}.

Deterministic extraction from Santha-Vazirani sources

- Vazirani [Vaz87]: Inner product modulo 2 extracts an almost-uniform bit from two independent unpredictable sources
- Not possible for one source (no function can be more than α-unpredictable ([SV86])
  Exercise: Prove this.
- Generalization[MMP+10]: Inner product modulo m extracts an almost-uniform element of $\mathbb{Z}_m$, if n is at least roughly m^2

δ-closeness

Definition (Statistical distance and δ-closeness)

For random variables X and X' taking values in Ω, we say that X and X' are δ-close if the statistical distance between their distributions is at most δ, i.e.,

$$||X - X'||_{SD} := \frac{1}{2} \sum_{x \in \Omega} |Pr[X = x] - Pr[X' = x]| \leq \delta$$

Deterministic extraction from Santha-Vazirani sources

Theorem (Randomness extraction)

There is a constant c such that if:

1. *X is an α-unpredictable bit source on $\{0,1\}^n$,*
2. *Y is a source on $\{0,1\}^n$ with min-entropy at least βn,*
3. *Y is independent from X,*
4. *$Z = \langle X, Y \rangle$ mod m for some $m \in \mathbb{N}$,*

then for every $\delta \in [0,1]$, such that

$$n \geq c \cdot \frac{m^2}{\alpha\beta} \cdot \log\left(\frac{m}{\beta}\right) \cdot \log\left(\frac{m}{\delta}\right) \approx m^2 \log^2 m$$

the random variable (Y, Z) is δ-close to (Y, U) where U is uniform on $\mathbb{Z}_m$ and independent of Y.

Bounding the magnitude of Fourier coefficients of Z

Lemma (Bounding Fourier coefficients)

Let Z be a random variable taking values in $\mathbb{Z}_m$.
Then the statistical distance between Z and the uniform distribution on $\mathbb{Z}_m$ is at most

$$\frac{1}{2}\sqrt{\sum_{\omega \neq 1} |E[\omega^Z]|^2} = \frac{1}{2}\sqrt{\sum_{k=1}^{m-1}\left|\sum_{\ell=0}^{m-1} \Pr[Z=n] e^{-\frac{2\pi i}{m}k\ell}\right|^2},$$

where the sum is over all complex m'th roots of unity ω other than 1.

Bounding the magnitude of Fourier coefficients of Z

Proof

Let $p_Z(\cdot), p_U(\cdot)$ *be probability masses of* Z *and* U*.*

$$||Z - U||_{SD} = \frac{1}{2}||p_Z - p_U||_1 \leq \frac{\sqrt{m}}{2}||p_Z - p_U||_2$$

Bounding the magnitude of Fourier coefficients of Z

Proof

Let $p_Z(\cdot), p_U(\cdot)$ be probability masses of Z and U.

$$||Z - U||_{SD} = \frac{1}{2}||p_Z - p_U||_1 \leq \frac{\sqrt{m}}{2}||p_Z - p_U||_2$$

By Parseval's theorem ($\hat{p}_X(k)$ is k-th Fourier coefficient of DFT of X):

$$\frac{\sqrt{m}}{2}||p_Z - p_U||_2 = \frac{1}{2}\sqrt{\sum_{k=0}^{m-1} |\hat{p}_Z(k) - \hat{p}_U(k)|^2}$$

Bounding the magnitude of Fourier coefficients of Z

Proof

Let $p_Z(\cdot), p_U(\cdot)$ be probability masses of Z and U.

$$||Z-U||_{SD} = \frac{1}{2}||p_Z - p_U||_1 \leq \frac{\sqrt{m}}{2}||p_Z - p_U||_2$$

By Parseval's theorem ($\hat{p}_X(k)$ is k-th Fourier coefficient of DFT of X):

$$\frac{\sqrt{m}}{2}||p_Z - p_U||_2 = \frac{1}{2}\sqrt{\sum_{k=0}^{m-1}|\hat{p}_Z(k) - \hat{p}_U(k)|^2}$$

Substituting Fourier coefficients $\hat{p}_Z(\cdot)$ and $\hat{p}_U(\cdot)$, the claim follows, i.e.

$$||Z-U||_{SD} \leq \frac{1}{2}\sqrt{\sum_{\omega\neq 1}|E[\omega^Z]|^2} = \frac{1}{2}\sqrt{\sum_{k=1}^{m-1}\left|\sum_{\ell=0}^{m-1}\Pr[Z=n]e^{-\frac{2\pi i}{m}k\ell}\right|^2}.$$

Proof of randomness extraction theorem

Proof (Randomness extraction theorem)

- *X is α-unpredictable source on $\{0,1\}^n$, Y is βn-source on $\{0,1\}^n$.*
- *For every $\omega \neq 1$, let $BAD = \bigcup_\omega BAD_\omega$, where*

$$BAD_\omega = \left\{ y \in \{0,1\}^n \colon \left| E\left[\omega^{\langle X,y\rangle}\right]\right| > \frac{\delta}{\sqrt{m}} \right\}.$$

Proof of randomness extraction theorem

Proof (Randomness extraction theorem)

- *X is α-unpredictable source on $\{0,1\}^n$, Y is βn-source on $\{0,1\}^n$.*
- *For every $\omega \neq 1$, let $BAD = \bigcup_\omega BAD_\omega$, where*

$$BAD_\omega = \left\{ y \in \{0,1\}^n \colon \left| E\left[\omega^{\langle X,y \rangle}\right] \right| > \frac{\delta}{\sqrt{m}} \right\}.$$

-

$$\begin{aligned} ||Z - U||_{SD} = (||Z - U||_{SD} | y \notin BAD) Pr(y \notin BAD) \\ + (||Z - U||_{SD} | y \in BAD) Pr(y \in BAD) \leq \\ \leq (||Z - U||_{SD} | y \notin BAD) + Pr(y \in BAD). \end{aligned}$$

Proof of randomness extraction theorem

Proof (Randomness extraction theorem)

- *X is α-unpredictable source on $\{0,1\}^n$, Y is βn-source on $\{0,1\}^n$.*
- *For every $\omega \neq 1$, let $BAD = \bigcup_\omega BAD_\omega$, where*

$$BAD_\omega = \left\{ y \in \{0,1\}^n \colon \left| E\left[\omega^{\langle X,y\rangle}\right]\right| > \frac{\delta}{\sqrt{m}} \right\}.$$

- $$||Z-U||_{SD} = (||Z-U||_{SD} | y \notin BAD) Pr(y \notin BAD) + (||Z-U||_{SD} | y \in BAD) Pr(y \in BAD) \leq \leq (||Z-U||_{SD} | y \notin BAD) + Pr(y \in BAD).$$
- *Using Bounding Fourier coefficients Lemma for every $y \notin BAD$, the statistical distance between $Z|_{Y=y} = \langle X,y\rangle \bmod m$ and the uniform distribution on $\mathbb{Z}_m$ is at most $(1/2)\sqrt{(m-1)\cdot(\delta/\sqrt{m})^2} \leq \delta/2$.*

Estimating $Pr[Y \in BAD_\omega]$

$Pr[Y \in BAD] \leq \delta/2$ follows if $Pr[Y \in BAD_\omega] \leq \delta/2m$ for each $\omega \neq 1$.

Estimating $Pr[Y \in BAD_\omega]$

$Pr[Y \in BAD] \leq \delta/2$ follows if $Pr[Y \in BAD_\omega] \leq \delta/2m$ for each $\omega \neq 1$. Suppose $\omega \neq 1$ is a primitive root of unity. By Estimating $2t$'th moment of Fourier coefficients Lemma, we have:

$$|BAD_\omega| \leq \frac{\sum_{y \in \mathbb{Z}_2^n} \left| E\left[\omega^{\langle X, y \rangle}\right] \right|^{2t}}{(\delta/\sqrt{m})^{2t}} \leq \frac{\sum_{y \in \mathbb{Z}_m^n} \left| E\left[\omega^{\langle X, y \rangle}\right] \right|^{2t}}{(\delta/\sqrt{m})^{2t}} \leq$$
$$\leq \frac{[1 + m \cdot \exp(-\Omega(\alpha t/m^2))]^n}{(\delta^2/m)^t} \leq \frac{2^{\beta n/2}}{(\delta^2/m)^t}$$

for $t = \lceil c_0 \cdot (m^2/\alpha) \cdot \log(m/\beta) \rceil$ for a sufficiently large constant c_0.

Estimating $Pr[Y \in BAD_\omega]$

$Pr[Y \in BAD] \leq \delta/2$ follows if $Pr[Y \in BAD_\omega] \leq \delta/2m$ for each $\omega \neq 1$. Suppose $\omega \neq 1$ is a primitive root of unity. By Estimating $2t$'th moment of Fourier coefficients Lemma, we have:

$$|BAD_\omega| \leq \frac{\sum_{y \in \mathbb{Z}_2^n} \left|E\left[\omega^{\langle X,y \rangle}\right]\right|^{2t}}{(\delta/\sqrt{m})^{2t}} \leq \frac{\sum_{y \in \mathbb{Z}_m^n} \left|E\left[\omega^{\langle X,y \rangle}\right]\right|^{2t}}{(\delta/\sqrt{m})^{2t}} \leq$$

$$\leq \frac{[1 + m \cdot \exp(-\Omega(\alpha t/m^2))]^n}{(\delta^2/m)^t} \leq \frac{2^{\beta n/2}}{(\delta^2/m)^t}$$

for $t = \lceil c_0 \cdot (m^2/\alpha) \cdot \log(m/\beta) \rceil$ for a sufficiently large constant c_0. So if $n \geq (2/\beta) \cdot (t \cdot \log(m/\delta^2) + \log(2m/\delta))$ (holds by hypothesis):

$$Pr[Y \in BAD_\omega] \leq 2^{-\beta n} \cdot |BAD_\omega| \leq \frac{2^{-\beta n/2}}{(\delta^2/m)^t} \leq \frac{\delta}{2m}.$$

Lower bound for inner product (theorem)

No differentially-private protocol can estimate inner product to within error $o(\sqrt{n}/\log n)$:

Lower bound for inner product (theorem)

No differentially-private protocol can estimate inner product to within error $o(\sqrt{n}/\log n)$:

Theorem (Lower bound for inner product)

Let $P(x, y)$ be a randomized protocol with ϵ-differential privacy for inputs $x, y \in \{0, 1\}^n$, and let $\delta > 0$. Then with probability at least $1 - \delta$ over $x, y \leftarrow \{0, 1\}^n$ and the coin tosses of P, party B's output differs from $\langle x, y\rangle$ by at least

$$\Delta = \Omega\left(\frac{\sqrt{n}}{\log n} \cdot \frac{\delta}{e^{\epsilon}}\right).$$

Lower bound for inner product (theorem)

No differentially-private protocol can estimate inner product to within error $o(\sqrt{n}/\log n)$:

Theorem (Lower bound for inner product)

Let $P(x, y)$ be a randomized protocol with ϵ-differential privacy for inputs $x, y \in \{0,1\}^n$, and let $\delta > 0$. Then with probability at least $1 - \delta$ over $x, y \leftarrow \{0,1\}^n$ and the coin tosses of P, party B's output differs from $\langle x, y \rangle$ by at least

$$\Delta = \Omega\left(\frac{\sqrt{n}}{\log n} \cdot \frac{\delta}{e^{\epsilon}}\right).$$

Similar result for Hamming distance is implied, because

$$\langle x, y \rangle = |x|_H + |y|_H - |x - y|_H$$

.

Limitations of the extractor technique

Definition (Sensitivity)

Given $f : \{0,1\}^n \to \mathbb{R}$ *let sensitivity* $= \max_{|x-y|_H=1} |f(x) - f(y)|$.

Limitations of the extractor technique

Definition (Sensitivity)

Given $f \colon \{0,1\}^n \to \mathbb{R}$ *let sensitivity* $= \max_{|x-y|_H=1} |f(x) - f(y)|$.

Theorem (Limitation of extractor technique)

Let $f \colon \{0,1\}^n \times \{0,1\}^n \to \mathbb{R}$ *be a sensitivity-1 function. Then for any distribution* μ *such that for any input* y*, the conditional distribution* $\mu(X|Y=y)$ *is a product distribution* $\prod_{i=1}^n \mu_i(X_i|Y=y)$*, there is a function* $g(y)$ *such that* $Pr_{(x,y)\sim\mu}[|g(y) - f(x,y)| > t] \le 2\exp\left(-t^2/2n\right)$.

Limitations of the extractor technique

Definition (Sensitivity)

Given $f\colon \{0,1\}^n \to \mathbb{R}$ *let sensitivity* $= \max_{|x-y|_H=1} |f(x) - f(y)|$.

Theorem (Limitation of extractor technique)

Let $f\colon \{0,1\}^n \times \{0,1\}^n \to \mathbb{R}$ *be a sensitivity-1 function. Then for any distribution* μ *such that for any input* y*, the conditional distribution* $\mu(X|Y=y)$ *is a product distribution* $\prod_{i=1}^n \mu_i(X_i|Y=y)$*, there is a function* $g(y)$ *such that* $Pr_{(x,y)\sim\mu}[|g(y) - f(x,y)| > t] \le 2\exp\left(-t^2/2n\right)$.

Proof

For any $h\colon \{0,1\}^n \leftarrow \mathbb{R}$ *of sensitivity 1, and any product distribution* ν *on* X,

$$Pr[|h(x) - E_{x\sim\nu}[h(x)]| > t] \le 2\exp\left(-t^2/2n\right).$$

Applying to $f(X,y)$ *and* $g(y) = E_{x\in\mu(X|Y=y)}[f(x,y)]$*, the result follows.*

References

- [MMP+10] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, Salil Vadhan, "The Limits of Two-Party Differential Privacy" (http://research.microsoft.com/pubs/137029/2dplimits.pdf)
- [Vaz87] U. V. Vazirni, "Strong communication complexity or generating quasirandom sequences from two communicating semirandom sources"
- [SV86] Miklos Santha, Umesh V. Vazirani, "Generating quasi-random sequences from semi-random sources"
- Ryan O'Donnell, "Analysis of Boolean functions" (http://www.cs.cmu.edu/~odonnell/boolean-analysis/).
- Devdatt P. Dubhashi, Alessandro Pancones, "Concentration of Measure for the Analysis of Randomized Algorithms"