Entropy-minimizing Mechanism for Differential Privacy of Discrete-time Linear Feedback Systems

Yu Wang, Zhenqi Huang, Sayan Mitra and Geir E. Dullerud

September 25, 2014

General Question

Trade-off between "privacy" and "accuracy": a common strategy to protect some data private is to randomize it, but this undermines the accuracy of the data.

Example[1]:

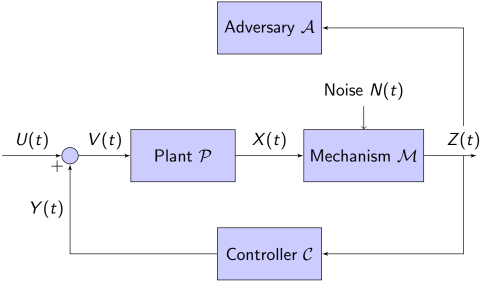


Figure: Block Diagram for ϵ-Differentially Private Discrete-time Linear Feedback System

[1] Huang et al., HiCoNS 14.

Preliminaries

In this work, we use the concept of ϵ-differential privacy as a measure of privacy. It originates from the study of privacy-preserving queries of datasets [2] and later extends to dynamic systems.

Definition

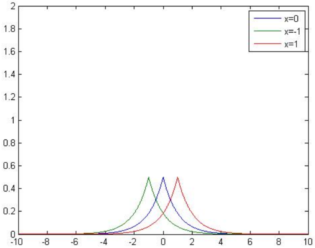
The mechanism $\mathcal{M}$ is ϵ-differentially private if the inequality

$$\mathbb{P}\left[\mathcal{M}(\boldsymbol{x}_1) \subseteq O\right] \leq \exp\left(\epsilon \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_1\right) \mathbb{P}\left[\mathcal{M}(\boldsymbol{x}_2) \subseteq O\right] \tag{1}$$

holds for any inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ and a set of possible outputs O, where $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$.

[2] C. Dwork, 2006.

Preliminaries



Accuracy is measured by Shannon entropy. For a random variable X on $\mathbb{R}^n$ with probability distribution function $f(\boldsymbol{x})$,

$$\mathbf{H}(X) = -\int_{\mathbb{R}^n} f(\boldsymbol{x}) \ln(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{2}$$

One-shot Query

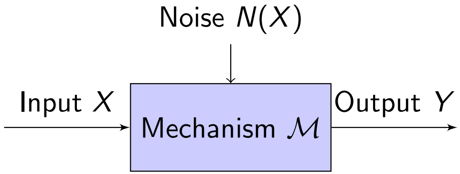


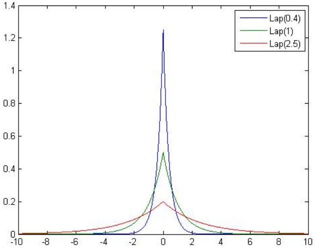
Figure: Block Diagram for a ϵ-Differentially Private Mechanism

Conditions:

- $X, Y \in (\mathbb{R}^n, \|\cdot\|_1)$
- the joint p.d.f. $p(x, y)$ is absolute continuous;
- the noise $N(X)$ is zero-mean;
- the accuracy is measured by $\mathbf{H}(\mathcal{M}) = \sup_X \mathbf{H}(Y)$.

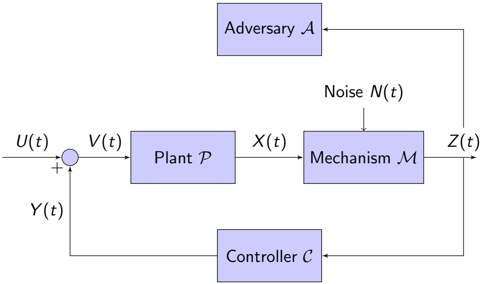
Theorem

For an ϵ-differentially private mechanism $\mathcal{M}$ with input set $(\mathbb{R}^n, \|\cdot\|_1)$, we have $\mathbf{H}(\mathcal{M}) \geq n - n\ln(\epsilon/2)$ and the minimum is achieved by $p(\boldsymbol{x}, \boldsymbol{y}) = (\frac{\epsilon}{2})^n \exp(-\epsilon\|\boldsymbol{y} - \boldsymbol{x}\|_1) = \prod_{i=1}^n \left(\frac{\epsilon}{2} e^{-\epsilon|y_i - x_i|}\right)$.



Trade-off: Privacy$\uparrow \implies \epsilon\downarrow \implies \mathbf{H}(\mathcal{M})\uparrow \implies$ Accuracy $\downarrow$

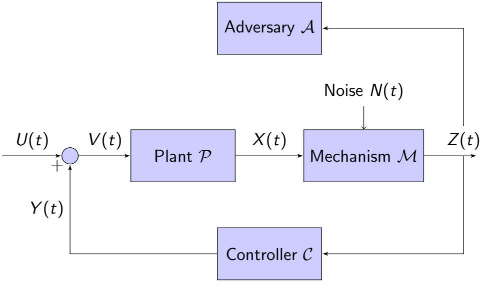
Control Systems



Conditions:

- $X(t), Y(t), Z(t), U(t), V(t) \in (\mathbb{R}^n, \|\cdot\|_1)$
- zero input: $U(t) = 0$
- unit gain feedback: $V(t) = Y(t) = Z(t)$
- dynamics: $X(t+1) = AX(t) + BV(t)$.

Control Systems



The adversary $\mathcal{A}$ only has access to the randomized outputs $\{Z(i) \mid i \in [t]\}$. Since

$$X(t) = A^t X(0) + \sum_{i=0}^{t-1} A^{t-i-1} B Z(i), \tag{3}$$

protecting the ϵ-differential privacy of the initial system state is equivalent to protecting the ϵ-differential privacy of the whole trajectory.

Control Systems

The adversary $\mathcal{A}$ estimates the initial system state from the past history of randomized outputs $\{Z(i) \mid i \in [t]\}$ by

$$\tilde{X}(t) = \mathbb{E}\left[X(0) \mid Z(0), Z(1), \ldots, Z(t)\right], \tag{4}$$

The accuracy of the output of the mechanism $\mathcal{M}$ at time $t \in \mathbb{N}$ is measured by

$$\mathbf{H}\left(\mathcal{M}, t\right) = \mathbf{H}\left(\tilde{X}(t)\right). \tag{5}$$

Control Systems

The mechanism $\mathcal{L}$ is ϵ-differentially private up to time $t \in \mathbb{N}$, if for any pair of initial states $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$, and output history $\{\boldsymbol{z}(i) \mid i \in [t]\}$,

$$\frac{\mathbb{P}\left[Z(1) = \boldsymbol{z}(1), \ldots, Z(t) = \boldsymbol{z}(t) \mid X(0) = \boldsymbol{x}_1\right]}{\mathbb{P}\left[Z(1) = \boldsymbol{z}(1), \ldots, Z(t) = \boldsymbol{z}(t) \mid X(0) = \boldsymbol{x}_2\right]} \le \exp\left(\epsilon \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\right). \tag{6}$$

By Bayes formula, (6) is equivalent to

$$\tilde{h}_t(\boldsymbol{x}_1) \le \exp\left(\epsilon \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\right) \tilde{h}_t(\boldsymbol{x}_2). \tag{7}$$

where $\tilde{h}_t$ is the probability density function of $\tilde{X}(t)$.

Control Systems

Theorem

If a mechanism is ϵ-differentially private up to time $t \geq 0$, then

$$\mathbf{H}(\mathcal{L}, i) \geq n - n\ln(\frac{\epsilon}{2}) \tag{8}$$

for $i \in 1, \ldots, t$. The equality holds when $N(0) \sim Lap(1/\epsilon)$, and for $t \geq 1$, $N(t) = AN(t-1)$. In this case

$$\mathbf{H}(\mathcal{L}, 1) = \mathbf{H}(\mathcal{L}, 2) = \ldots = \mathbf{H}(\mathcal{L}, t) = n - n\ln(\frac{\epsilon}{2}). \tag{9}$$

Proof of Theorem

Assume $X, Y \in \mathbb{R}$.

Problem

$$\begin{aligned} \textit{Minimize: } & \mathbf{H}\left(\mathcal{M}\right) \\ \textit{subject to: } & \mathbb{P}\left[\mathcal{M}(x_1) \subseteq O\right] \leq \exp\left(\epsilon \|x_1 - x_2\|_1\right) \mathbb{P}\left[\mathcal{M}(x_2) \subseteq O\right] \end{aligned}$$

Proof Step 1

Claim 1: for fixed x, $p(x, y-x)$ is even.

$$\mathbf{H}_1^+(\mathcal{M}) = \sup_{x\in\mathbb{R}} \int_{[x,\infty)} -p(x,y)\ln p(x,y)\mathrm{d}y, \tag{10}$$

$$\mathbf{H}_1^-(\mathcal{M}) = \sup_{x\in\mathbb{R}} \int_{(-\infty,x]} -p(x,y)\ln p(x,y)\mathrm{d}y. \tag{11}$$

$$q(x,y) = \begin{cases} p(x,y) & \begin{array}{c}\text{if } y > x, \mathbf{H}_1^+(\mathcal{M}) \le \mathbf{H}_1^-(\mathcal{M}) \\ \text{or } y < x, \mathbf{H}_1^+(\mathcal{M}) > \mathbf{H}_1^-(\mathcal{M}),\end{array} \\ p(x, 2x-y) & \begin{array}{c}\text{if } y > x, \mathbf{H}_1^+(\mathcal{M}) > \mathbf{H}_1^-(\mathcal{M}) \\ \text{or } y < x, \mathbf{H}_1^+(\mathcal{M}) \le \mathbf{H}_1^-(\mathcal{M}).\end{array} \end{cases} \tag{12}$$

$$\mathbf{H}(\mathcal{N}) = 2\min\{\mathbf{H}_1^+(\mathcal{M}), \mathbf{H}_1^-(\mathcal{M})\} \le \mathbf{H}_1^+(\mathcal{M}) + \mathbf{H}_1^-(\mathcal{M}) = \mathbf{H}(\mathcal{M}), \tag{13}$$

Proof Step 1

Claim 2: for any x, $p(x,y) = p(2a-x, 2a-y)$.

$$\mathbf{H}^{+}(\mathcal{M}) = \sup_{x>a} \int_{\mathbb{R}} -p(x,y)\ln p(x,y)\mathrm{d}y, \tag{14}$$

$$\mathbf{H}^{-}(\mathcal{M}) = \sup_{x\leq a} \int_{\mathbb{R}} -p(x,y)\ln p(x,y)\mathrm{d}y. \tag{15}$$

If $\mathbf{H}^{+}(\mathcal{M}) \leq \mathbf{H}^{-}(\mathcal{M})$, then define

$$q(x,y) = \begin{cases} p(x,y), & x > a, \\ p(2a-x, 2a-y), & x \leq a, \end{cases} \tag{16}$$

otherwise, define

$$q(x,y) = \begin{cases} p(2a-x, 2a-y), & x > a, \\ p(x,y), & x \leq a. \end{cases} \tag{17}$$

$$\mathbf{H}(\mathcal{N}) = \min\{\mathbf{H}^{+}(\mathcal{M}), \mathbf{H}^{-}(\mathcal{M})\} \leq \max\{\mathbf{H}^{+}(\mathcal{M}), \mathbf{H}^{-}(\mathcal{M})\} = \mathbf{H}(\mathcal{M}) \tag{18}$$

Proof Step 1

Claim 3: $p(x, y) = f(y - x)$.
Let $q(x, y) = p(x, y - x)$. By Claim 2, $q(x, y) = q(2a - x, -y)$.
By Claim 1, $q(2a - x, -y) = q(2a - x, y)$.
Now the problem becomes,

Problem

$$\begin{aligned} \textit{Minimize: } & H(f) = -\int_{[0,\infty)} f(x) \ln f(x) \mathrm{d}x, \\ \textit{subject to: } & f(x) \textit{ is absolutely continuous}, \\ & f(x) \geq 0, \\ & |f'(x)| \leq \epsilon f(x) \textit{ a.e.}, \\ & \int_{[0,\infty)} f(x) \mathrm{d}x = \frac{1}{2}. \end{aligned}$$

Proof Step 2

Claim 4: $f(x)$ is decreasing.
Let x^* be a local minimum on $(0,1)$. Then there exists $x^* \in [a, b]$ such that $f(a) = f(b) > f(x)$ for $x \in (a, b)$. Let $d = \frac{1}{f(a)} \int_a^b f(x)\mathrm{d}x$ and

$$h(x) = \begin{cases} f(x), & x \in [0, a], \\ f(b), & x \in [a, a+d], \\ f(x+b-a-d), & x \in [a+d, \infty]. \end{cases} \tag{19}$$

Then $H(h) < H(f)$.

Proof Step 2

Let $F(x) = \int_x^\infty f(y)\mathrm{d}y$.

$$\begin{aligned} F(x) &\geq \int_x^\infty \frac{|f'(x)|}{\epsilon}\mathrm{d}y \geq \frac{1}{\epsilon}\left|\int_x^\infty f'(x)\mathrm{d}y\right| \\ &= \frac{1}{\epsilon}|f(\infty) - f(x)| = \frac{f(x)}{\epsilon} \end{aligned} \tag{20}$$

In particular, $f(0) \geq \epsilon F(0) = \frac{\epsilon}{2}$.

Proof Step 2

$$
\begin{aligned}
H(f) &= -\int_0^\infty f(x)\ln f(x)\mathrm{d}x \\
&= -\int_0^\infty f(x)\left(\ln f(0) + \int_0^x \frac{f'(y)}{f(y)}\mathrm{d}y\right)\mathrm{d}x \\
&= -\frac{1}{2}\ln f(0) - \int_0^\infty \frac{f'(y)}{f(y)}\left(\int_x^\infty f(x)\mathrm{d}x\right)\mathrm{d}y \\
&= -\frac{1}{2}\ln f(0) - \int_0^\infty \frac{f'(y)F(y)}{f(y)}\mathrm{d}y \\
&\geq -\frac{1}{2}\ln f(0) - \int_0^\infty \frac{f'(y)}{\epsilon}\mathrm{d}y \\
&= \frac{f(0)}{\epsilon} - \frac{1}{2}\ln f(0) \geq \frac{1}{2} - \ln(\frac{\epsilon}{2}),
\end{aligned}
\tag{21}
$$

The minimum is achieved by

$$
f(x) = \frac{\epsilon}{2}\exp(-\epsilon x). \tag{22}
$$

Thanks!