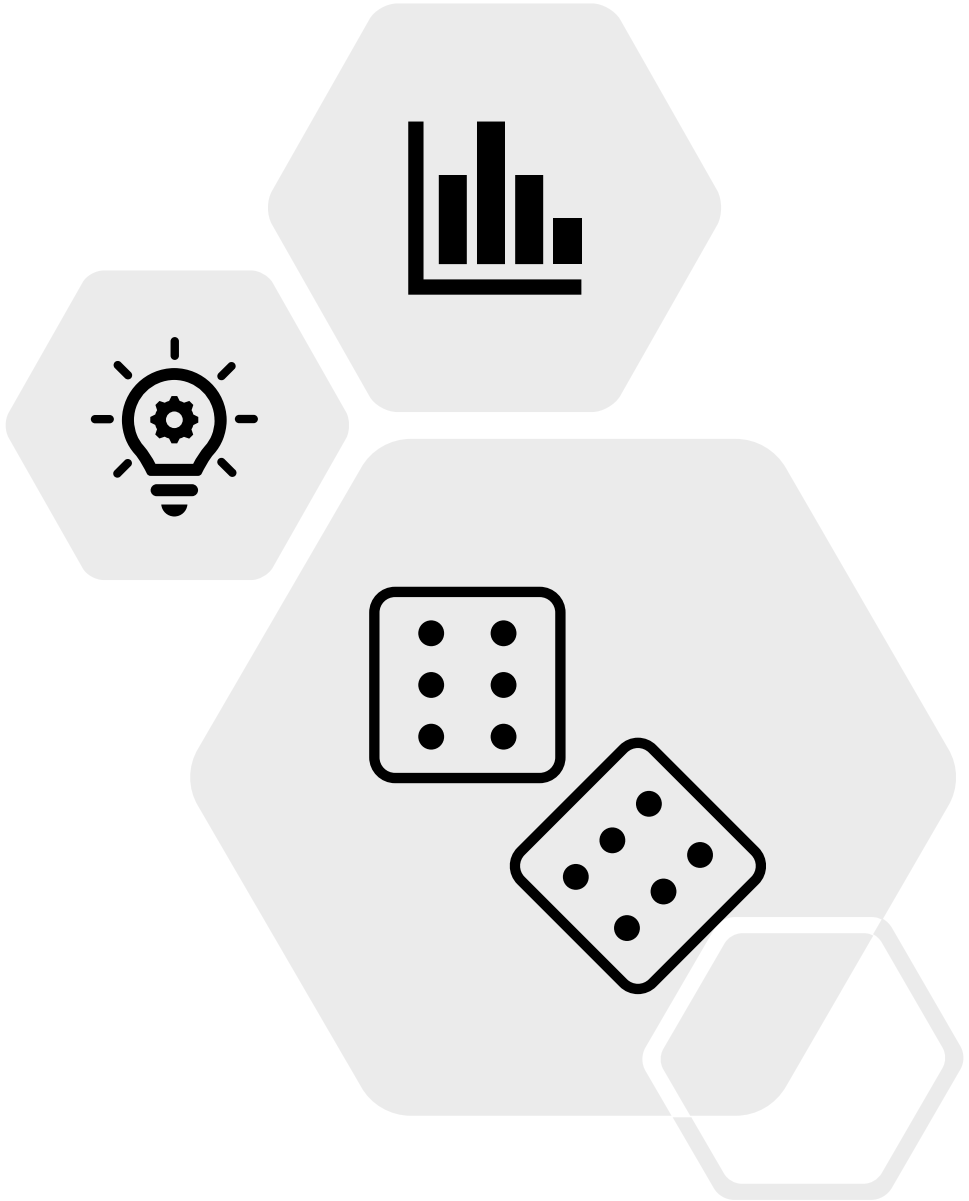


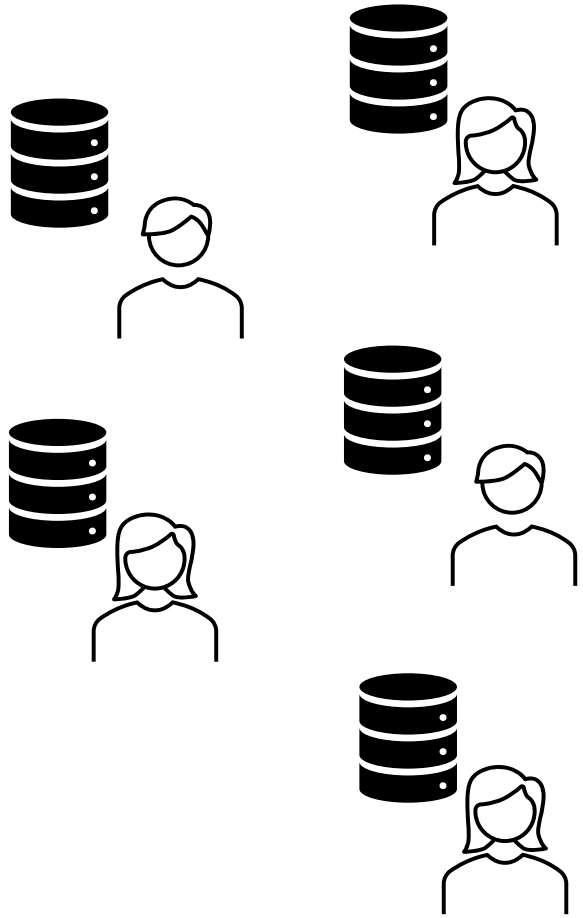
Domain Compression and its Applications to Distribution Testing

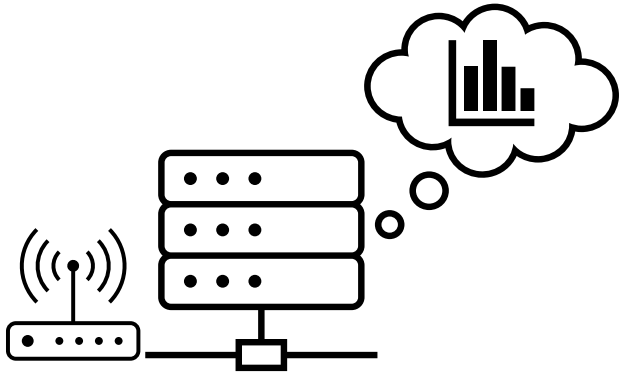
Clément Canonne, University of Sydney

CIS Seminar (U Melbourne), 10/05/2022

The (distributed) setting

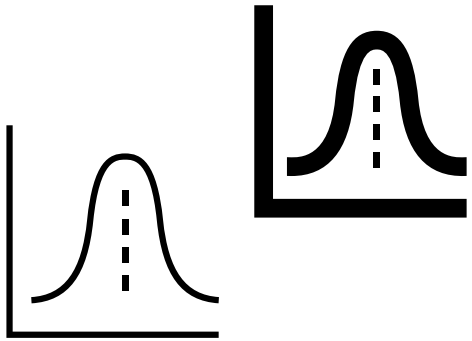
- Users have data (observations)
- Center wants to perform hypothesis testing





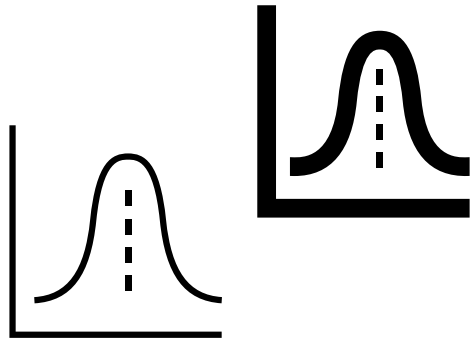
The (distributed) setting: example

- Users have data (observations) $X_1,\ldots,X_n$ i.i.d. from some p
- Does the data distribution p fit the model q?
  - H_0: p=q
  - H_1: TV(p,q) > ε



The (distributed) setting: example

- Users have data (observations) $X_1,\ldots,X_n$ i.i.d. from some p
- Does the data distribution p fit the model q?
  - H_0: p=q
  - H_1: TV(p,q) > ε



One-sample testing [Stats]
(One-sample) goodness-of-fit [Stats/ML]
Identity testing [Computer Science]

The (distributed) setting: example

- Users have data (observations) $X_1,\ldots,X_n$ i.i.d. from some p
- Does the data distribution p fit the model q?
  - H_0: p=q
  - H_1: TV(p,q) > ε

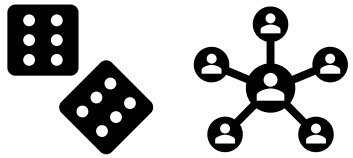
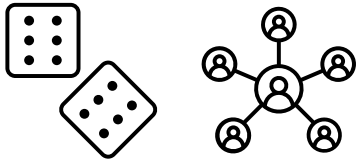
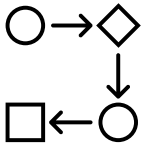
$$\mathrm{TV}(p,q) = \sup_S (p(S) - q(S))$$

One-sample testing [Stats]
(One-sample) goodness-of-fit [Stats/ML]
Identity testing [Computer Science]

The (distributed) setting

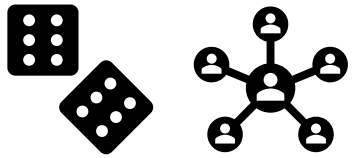
- Users have data (observations) $X_1,\ldots,X_n$ i.i.d. from some p
- Users have constraints: privacy, bandwidth, memory, …
- Center has goals: e.g., minimise number of users n (maximise utility)

The (distributed) setting: more!

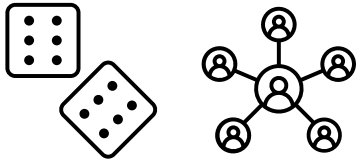
- Private-coin protocols 
- Public-coin protocols 
- Sequentially interactive protocols 
- Blackboard protocols 

The (distributed) setting: more!

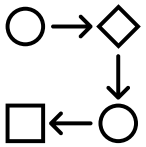
- Private-coin protocols



- Public-coin protocols



- Sequentially interactive protocols



- Blackboard protocols

Back to our example: identity testing

n i.i.d. samples from unknown p over [k] = {1,2,…,k}

known reference q over [k]

Distance parameter ε in (0,1]

$$\mathcal{H}_0 : p = q \quad \text{vs.} \quad \mathcal{H}_1 : \mathrm{TV}(p, q) > \varepsilon$$

How large must n be to distinguish with probability at least 2/3?

Back to our example: identity testing

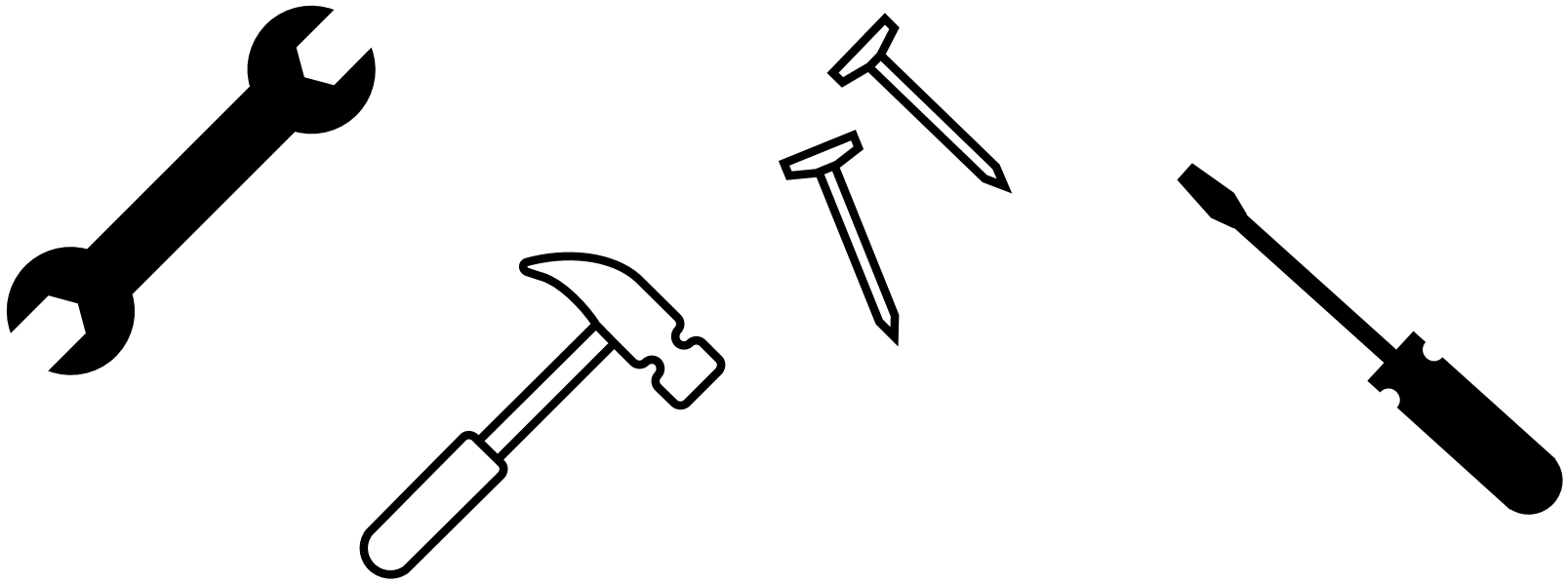
	Private-Coin	Public-Coin
Centralised	$k^{1/2}/\varepsilon^2$	
Local privacy	$k^{3/2}/(\varepsilon^2\rho^2)$	$k/(\varepsilon^2\rho^2)$
Shuffle privacy*	$k^{3/4}/(\varepsilon\rho) + k^{1/2}/\varepsilon^2$	$k^{2/3}/(\varepsilon^{4/3}\rho^{2/3}) + k^{1/2}/\varepsilon^2 + k^{1/2}/(\varepsilon\rho)$
Bandwidth	$k^{3/2}/(\varepsilon^2 2^{\ell}) + k^{1/2}/\varepsilon^2$	$k/(\varepsilon^2 2^{\ell/2}) + k^{1/2}/\varepsilon^2$

Back to our example: identity testing

	Private-Coin	Public-Coin
Centralised	$k^{1/2}/\varepsilon^2$	
Local privacy	$k^{3/2}/(\varepsilon^2\rho^2)$	?
Shuffle privacy*	$k^{3/4}/(\varepsilon\rho) + k^{1/2}/\varepsilon^2$	?
Bandwidth	$k^{3/2}/(\varepsilon^2 2^{\ell}) + k^{1/2}/\varepsilon^2$	?

Claim: all the public-coin bounds are “immediate”

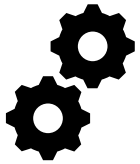
(For every hard-earned upper bound, the second is free!)

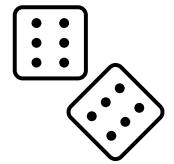


Domain Compression

[Acharya–Canonne–Han–Sun–Tyagi'20], [Amin–Joseph–Mao'20]

- trade domain size for statistical distance using shared randomness
- develop a "private-coin" protocol, get a "public-coin" one for free!

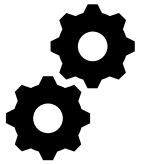


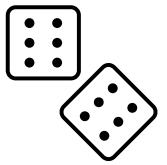


Domain Compression

[Acharya–Canonne–Han–Sun–Tyagi'20], [Amin–Joseph–Mao'20]

- trade domain size for statistical distance using shared randomness
- develop a "private-coin" protocol, get a "public-coin" one for free!





Theorem 2.12 (Domain Compression Lemma). There exist absolute constants $c_1, c_2 > 0$ such that the following holds. For any $2 \le \ell \le k$ and any $\mathbf{p}, \mathbf{q} \in \Delta_k$,

$$\Pr_{\Pi}\left[d_{\mathrm{TV}}(\mathbf{p}_\Pi, \mathbf{q}_\Pi) \ge c_1 \sqrt{\frac{\ell}{k}} d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \right] \ge c_2 \,,$$

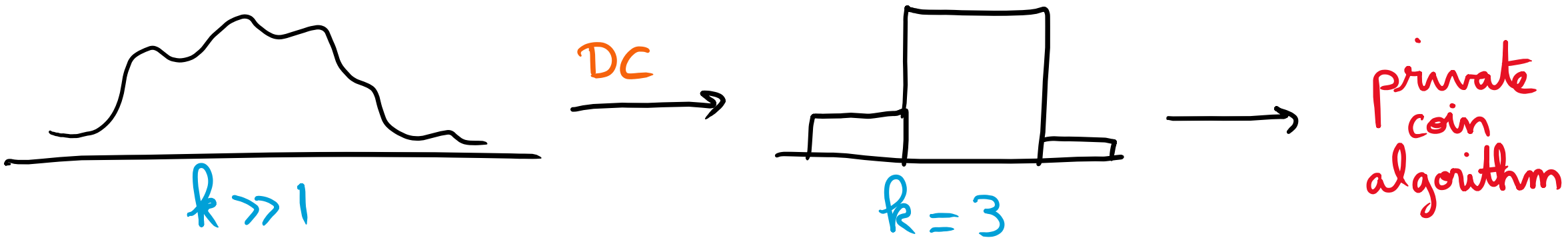
where $\Pi = (\Pi_1, \ldots \Pi_\ell)$ is a uniformly random partition of $[k]$ in ℓ subsets, and $\mathbf{p}_\Pi \in \Delta_\ell$ denotes the probability distribution on $[\ell]$ induced by $\mathbf{p}$ and Π via $\mathbf{p}_\Pi(i) = \mathbf{p}(\Pi_i)$.

Domain Compression

Use **public randomness** to convert a testing instance (k, ε) to a testing instance (L, $\varepsilon\sqrt{L/k}$) , for your choice of $2 \le L \le k$.

(Fine print: need to repeat O(1) times to amplify success probability.)

Now, what could that be good for?



Let’s see: identity testing

	Private-Coin	Public-Coin
Centralised	$k^{1/2}/\varepsilon^2$	
Local privacy	$k^{3/2}/(\varepsilon^2\rho^2)$	?
Shuffle privacy*	$k^{3/4}/(\varepsilon\rho) + k^{1/2}/\varepsilon^2$	?
Bandwidth	$k^{3/2}/(\varepsilon^2 2^{\ell}) + k^{1/2}/\varepsilon^2$	?

Local privacy

$$\frac{k^{3/2}}{\varepsilon^2 \rho^2}$$

DC $\longrightarrow$

Local privacy

$$\frac{k^{3/2}}{\varepsilon^2 \rho^2} \xrightarrow{DC} \frac{L^{3/2}}{\left(\varepsilon\sqrt{\frac{L}{k}}\right)^2 \rho^2} = \frac{L^{1/2} k}{\varepsilon^2 \rho^2}$$

Local privacy

$$\frac{k^{3/2}}{\varepsilon^2 \rho^2} \xrightarrow{DC} \frac{L^{3/2}}{\left(\varepsilon\sqrt{\frac{L}{k}}\right)^2 \rho^2} = \frac{L^{1/2} k}{\varepsilon^2 \rho^2}$$

Minimised for $L^* = 2$

Local privacy

$$\frac{k^{3/2}}{\varepsilon^2 \rho^2} \xrightarrow{DC} \frac{L^{3/2}}{\left(\varepsilon\sqrt{\frac{L}{k}}\right)^2 \rho^2} = \frac{L^{1/2} k}{\varepsilon^2 \rho^2}$$

Minimised for $L^* = 2$

$$\boxed{\frac{k}{\varepsilon^2 \rho^2}}$$

Local privacy

$$\frac{k^{3/2}}{\varepsilon^2 \rho^2} \xrightarrow{DC} \frac{L^{3/2}}{\left(\varepsilon\sqrt{\frac{L}{k}}\right)^2 \rho^2} = \frac{L^{1/2} k}{\varepsilon^2 \rho^2}$$

Minimised for $L^* = 2$

$$\boxed{\frac{k}{\varepsilon^2 \rho^2}}$$

Just a fluke?

Shuffle privacy

$$\frac{k^{3/4}}{\varepsilon\rho} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{DC}$$

Shuffle privacy

$$\frac{k^{3/4}}{\varepsilon \ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{\text{DC}} \frac{L^{3/4}}{\varepsilon \sqrt{\frac{L}{k}}\, \ell} + \frac{L^{1/2}}{\left(\varepsilon \sqrt{\frac{L}{k}}\right)^2}$$

Shuffle privacy

$$\frac{k^{3/4}}{\varepsilon\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{\text{DC}} \frac{L^{1/4}k^{1/2}}{\varepsilon\ell} + \frac{k}{\varepsilon^2 L^{1/2}}$$

Shuffle privacy

$$\frac{k^{3/4}}{\varepsilon\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{DC} \frac{L^{1/4}k^{1/2}}{\varepsilon\ell} + \frac{k}{\varepsilon^2 L^{1/2}}$$

Minimised for $L^* \simeq k^{2/3}\frac{\ell^{4/3}}{\varepsilon^{4/3}}$

Shuffle privacy

$$\frac{k^{3/4}}{\varepsilon\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{\text{DC}} \frac{L^{1/4}k^{1/2}}{\varepsilon\ell} + \frac{k}{\varepsilon^2 L^{1/2}}$$

Minimised for $L^* \simeq k^{2/3}\frac{\ell^{4/3}}{\varepsilon^{4/3}}$

(but careful: we need $2 \leq L \leq k$)

Shuffle privacy

Case 1 $L^* \le 2$ Take $L = 2$.

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \simeq \frac{k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2}$$

Shuffle privacy

$k \lesssim \frac{\varepsilon^2}{\ell^2}$

Case 1 $L^* \le 2$ Take $L = 2$.

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \lesssim \frac{k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2} \lesssim \frac{k^{1/2}}{\varepsilon \ell} + \frac{k^{2/3}}{\varepsilon^{4/3} \ell^{2/3}}$$

$$= \frac{k^{2/3}}{\varepsilon^2} \cdot k^{1/3}$$

Shuffle privacy

$k \lesssim \frac{\varepsilon^2}{\ell^2}$

Case 1 $L^* \le 2$ Take $L = 2$.

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \lesssim \frac{k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2} \lesssim \frac{k^{1/2}}{\varepsilon \ell} + \frac{k^{2/3}}{\varepsilon^{4/3} \ell^{2/3}}$$

$$= \frac{k^{2/3}}{\varepsilon^2} \cdot k^{1/3}$$

Good!

Shuffle privacy

Case 2 $L^* \geq k$ Take $L = k$.

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \simeq \frac{k^{3/4}}{\varepsilon \ell} + \frac{\sqrt{k}}{\varepsilon^2}$$

Shuffle privacy

Case 2 $L^* \geq k$ Take $L = k$.

$k \leq \frac{\ell^4}{\varepsilon^4}$

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \simeq \frac{k^{3/4}}{\varepsilon \ell} + \frac{\sqrt{k}}{\varepsilon^2} \leq 2 \frac{\sqrt{k}}{\varepsilon^2}$$

$$= \frac{\sqrt{k}}{\varepsilon \ell} \cdot k^{1/4}$$

Shuffle privacy

Case 2 $L^* \geq k$ Take $L = k$.

$k \leq \frac{\ell^4}{\varepsilon^4}$

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \asymp \frac{k^{3/4}}{\varepsilon \ell} + \frac{\sqrt{k}}{\varepsilon^2} \leq 2 \frac{\sqrt{k}}{\varepsilon^2}$$

$$= \frac{\sqrt{k}}{\varepsilon \ell} \cdot k^{1/4}$$

Good!

Shuffle privacy

Case 3 $2 \le L^* \le k$. Take $L = L^* \simeq k^{2/3} \frac{\ell^{4/3}}{\varepsilon^{4/3}}$

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \simeq \frac{k^{2/3}}{\varepsilon^{4/3} \ell^{2/3}}$$

Shuffle privacy

Case 3 $2 \le L^* \le k$. Take $L = L^* \simeq k^{2/3} \frac{\ell^{4/3}}{\varepsilon^{4/3}}$

$$\frac{L^{1/4} k^{1/2}}{\varepsilon \ell} + \frac{k}{\varepsilon^2 L^{1/2}} \simeq \frac{k^{2/3}}{\varepsilon^{4/3} \ell^{2/3}}$$

Good!

Shuffle privacy

Summary

$$\frac{\sqrt{k}}{\varepsilon^2} + \frac{\sqrt{k}}{\varepsilon\ell} + \frac{k^{2/3}}{\varepsilon^{4/3}\ell^{2/3}}$$

suffices.

Shuffle privacy

Summary

$$\frac{\sqrt{k}}{\varepsilon^2} + \frac{\sqrt{k}}{\varepsilon\ell} + \frac{k^{2/3}}{\varepsilon^{4/3}\ell^{2/3}}$$

suffices.

One more time?

Bandwidth (each user sends ℓ bits)

$$\frac{k^{3/2}}{\varepsilon^2 2^\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{DC}$$

Bandwidth (each user sends ℓ bits)

$$\frac{k^{3/2}}{\varepsilon^2 2^\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{DC} \frac{L^{3/2}}{\left(\varepsilon\sqrt{\frac{L}{k}}\right)^2 2^\ell} + \frac{L^{1/2}}{\left(\varepsilon\sqrt{\frac{L}{k}}\right)^2}$$

Bandwidth (each user sends ℓ bits)

$$\frac{k^{3/2}}{\varepsilon^2 2^\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{\text{DC}} \frac{L^{1/2} k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2} \varepsilon^2}$$

Minimised for $L^* = 2^\ell$

Bandwidth (each user sends ℓ bits)

$$\frac{k^{3/2}}{\varepsilon^2 2^\ell} + \frac{k^{1/2}}{\varepsilon^2} \xrightarrow{\text{DC}} \frac{L^{1/2}k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2}\varepsilon^2}$$

Minimised for $L^* = 2^\ell$

(again, we need $2 \le L \le k$)

Bandwidth (each user sends ℓ bits)

Case 1: $L^* \geq k$. Take $L = k$

$$\frac{L^{1/2} k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2} \varepsilon^2} = \frac{k}{\varepsilon^2 2^\ell} + \frac{\sqrt{k}}{\varepsilon^2} \simeq \frac{\sqrt{k}}{\varepsilon^2}$$

$2^\ell = L^* \geq k$

Bandwidth (each user sends ℓ bits)

Case 1: $L^* \geq k$. Take $L = k$

$$\frac{L^{1/2} k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2} \varepsilon^2} = \frac{k}{\varepsilon^2 2^\ell} + \frac{\sqrt{k}}{\varepsilon^2} \simeq \frac{\sqrt{k}}{\varepsilon^2}$$

Case 2 $L^* \leq k$. Take $L = L^* = 2^\ell$.

$$\frac{L^{1/2} k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2} \varepsilon^2} \simeq \frac{k}{\varepsilon^2 2^{\ell/2}}$$

Bandwidth (each user sends ℓ bits)

Case 1: $L^* \geq k$. Take $L = k$

$$\frac{L^{1/2} k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2} \varepsilon^2} = \frac{k}{\varepsilon^2 2^\ell} + \frac{\sqrt{k}}{\varepsilon^2} \simeq \boxed{\frac{\sqrt{k}}{\varepsilon^2}}$$

Case 2 $L^* \leq k$. Take $L = L^* = 2^\ell$.

$$\frac{L^{1/2} k}{\varepsilon^2 2^\ell} + \frac{k}{L^{1/2} \varepsilon^2} \simeq \boxed{\frac{k}{\varepsilon^2 2^{\ell/2}}}$$

Great!

We've seen: identity testing

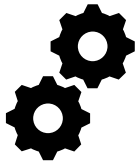
	Private-Coin	Public-Coin
Centralised	$k^{1/2}/\varepsilon^2$	
Local privacy	$k^{3/2}/(\varepsilon^2\rho^2)$	$k/(\varepsilon^2\rho^2)$
Shuffle privacy*	$k^{3/4}/(\varepsilon\rho) + k^{1/2}/\varepsilon^2$	$k^{2/3}/(\varepsilon^{4/3}\rho^{2/3}) + k^{1/2}/\varepsilon^2 + k^{1/2}/(\varepsilon\rho)$
Bandwidth	$k^{3/2}/(\varepsilon^2 2^\ell) + k^{1/2}/\varepsilon^2$	$k/(\varepsilon^2 2^{\ell/2}) + k^{1/2}/\varepsilon^2$

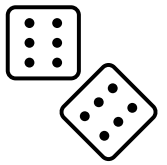
Domain Compression: let's prove it

Theorem 2.12 (Domain Compression Lemma). There exist absolute constants $c_1, c_2 > 0$ such that the following holds. For any $2 \le \ell \le k$ and any $\mathbf{p}, \mathbf{q} \in \Delta_k$,

$$\Pr_{\Pi}\left[d_{\mathrm{TV}}(\mathbf{p}_\Pi, \mathbf{q}_\Pi) \ge c_1 \sqrt{\frac{\ell}{k}} d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \right] \ge c_2 \,,$$

where $\Pi = (\Pi_1, \ldots \Pi_\ell)$ is a uniformly random partition of $[k]$ in ℓ subsets, and $\mathbf{p}_\Pi \in \Delta_\ell$ denotes the probability distribution on $[\ell]$ induced by $\mathbf{p}$ and Π via $\mathbf{p}_\Pi(i) = \mathbf{p}(\Pi_i)$.





Domain Compression: The Plan

- Prove the case L=2 (“one-bit isometry”) for the ℓ_2 norm (*not* TV)
- For general L:
  - partition the domain in L parts
  - apply the one-bit isometry separately to each of them
  - invoke a “simple” probability lemma about additivity of tails to conclude

Part I: One-bit isometry

Theorem. There exist $c_1, c_2 > 0$ s.t. for all $p, q \in [0,1]^k$,

$$\mathbb{P}\{|p(S) - q(S)| \geq c_1 \|p - q\|_2\} \geq c_2,$$

where $S \subseteq [k]$ is uniformly random.

not necessarily proba. distr.!

Part I: One-bit isometry

Theorem. There exist $c_1, c_2 > 0$ s.t. for all $p, q \in [0,1]^k$,

$$\mathbb{P}\{ |p(S) - q(S)| \geq c_1 \|p - q\|_2 \} \geq c_2,$$

where $S \subseteq [k]$ is uniformly random.

Part I: One-bit isometry

Theorem. There exist $c_1, c_2 > 0$ s.t. for all $p, q \in [0,1]^k$,

$$\mathbb{P}\left\{ |p(S) - q(S)| \geq c_1 \|p - q\|_2 \right\} \geq c_2,$$

where $S \subseteq [k]$ is uniformly random.

Proof. Set $\delta := p - q$, and $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

$$Z^2 := (p(S) - q(S))^2 = \left(\sum_{i=1}^{k} \delta_i X_i\right)^2 = \sum_i \delta_i^2 X_i^2 + \sum_{i \neq j} \delta_i \delta_j X_i X_j$$

$$\text{so } \mathbb{E}[Z^2] = \frac{1}{2}\|\delta\|_2^2 + \frac{1}{4}\sum_{i \neq j} \delta_i \delta_j$$

Proof. Set $\delta := p - q$, and $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

$$Z^2 := (p(S) - q(S))^2 = \left(\sum_{i=1}^{k} \delta_i X_i\right)^2 = \sum_i \delta_i^2 X_i + \sum_{i \neq j} \delta_i \delta_j X_i X_j$$

$$\text{so } \mathbb{E}[Z^2] = \frac{1}{2}\|\delta\|_2^2 + \frac{1}{4}\sum_{i \neq j} \delta_i \delta_j = \frac{1}{2}\|\delta\|_2^2 + \frac{1}{4}\left(\left(\sum_i \delta_i\right)^2 - \sum_i \delta_i^2\right)$$

Proof. Set $\delta := p - q$, and $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

$$Z^2 := (p(S) - q(S))^2 = \left(\sum_{i=1}^{k} \delta_i X_i\right)^2 = \sum_i \delta_i^2 X_i + \sum_{i \neq j} \delta_i \delta_j X_i X_j$$

$$\text{so } \mathbb{E}[Z^2] = \frac{1}{2}\|\delta\|_2^2 + \frac{1}{4}\sum_{i \neq j} \delta_i \delta_j = \frac{1}{2}\|\delta\|_2^2 + \frac{1}{4}\left(\left(\sum_i \delta_i\right)^2 - \|\delta\|_2^2\right)$$

$$= \frac{1}{4}\|\delta\|_2^2 + \frac{1}{4}\underbrace{\left(\sum_i \delta_i\right)^2}_{(\mathbb{E}Z)^2} \geq \frac{1}{4}\|\delta\|_2^2$$

Proof. Set $\delta := p - q$, and $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

$$Z^2 := (p(S) - q(S))^2 = \left(\sum_{i=1}^{k} \delta_i X_i\right)^2$$

so $\mathbb{E}[Z^2] \geq \frac{1}{4}\|\delta\|_2^2 + \frac{1}{4}(\mathbb{E}Z)^2 \geq \frac{1}{4}\|\delta\|_2^2$

Suffices now to prove

$$\mathbb{P}\left\{ Z^2 \geq \frac{1}{2}\mathbb{E}[Z^2] \right\} \geq \Omega(1)$$

(anticoncentration)

Proof. $Z^2 := \left(\sum_{i=1}^{k} \delta_i X_i\right)^2$ where $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

By **Paley-Zygmund**,

$$\mathbb{P}\left\{ Z^2 \geq \frac{1}{2}\mathbb{E}[Z^2] \right\} \geq \frac{1}{4}\frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]}$$

Proof. $Z^2 := \left(\sum_{i=1}^{k} \delta_i X_i\right)^2$ where $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

By Paley-Zygmund,

$$\mathbb{P}\left\{ Z^2 \geq \frac{1}{2}\mathbb{E}[Z^2] \right\} \geq \frac{1}{4} \frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]}$$

we got this

this is scary

Proof. $Z^2 := \left(\sum_{i=1}^{k} \delta_i X_i\right)^2$ where $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

By Paley-Zygmund,

$$\mathbb{P}\left\{ Z^2 \geq \frac{1}{2}\mathbb{E}[Z^2] \right\} \geq \frac{1}{4} \frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]}$$

we got this

this is scary

Nice trick: MGF.

Proof. $Z^2 := \left(\sum_{i=1}^{k} \delta_i X_i\right)^2$ where $X_1, \dots, X_k$ iid $\mathrm{Bern}(\frac{1}{2})$.

By Paley-Zygmund,

$$\mathbb{P}\left\{ Z^2 \geq \frac{1}{2}\mathbb{E}[Z^2] \right\} \geq \frac{1}{4} \frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]}$$

we got this

this is scary

Nice trick: MGF. $\forall \lambda$,

$$\mathbb{E}\, e^{\lambda Z} = \prod_{i=1}^{k} \mathbb{E}\, e^{\lambda \delta_i X_i} \overset{\text{HL}}{\leq} \prod_{i=1}^{k} e^{\lambda \frac{\delta_i}{2} + \frac{\lambda^2 \delta_i^2}{8}} = e^{\frac{\lambda}{2}\mathbb{E}[Z] + \frac{\lambda^2}{8}\|\delta\|_2^2}$$

Proof. So $\mathbb{E}\, e^{\lambda \underbrace{(Z-\mathbb{E}Z)}_{\text{symmetric}}} \le e^{\frac{1}{8}\lambda^2 \|\delta\|_2^2}$, and "so"

$$\frac{\lambda^4}{24}\mathbb{E}\left[(Z-\mathbb{E}Z)^4\right] \le e^{\frac{1}{8}\lambda^2\|\delta\|_2^2}$$

Choosing $\lambda = \frac{1}{\|\delta\|_2}$, we get

$$\mathbb{E}\left[(Z-\mathbb{E}Z)^4\right] \le 28\,\|\delta\|_2^4$$

Almost there!

Proof. So $\mathbb{E}\, e^{\lambda \underbrace{(Z-\mathbb{E}Z)}_{\text{symmetric}}} \le e^{\frac{1}{8}\lambda^2 \|\delta\|_2^2}$, and "so"

$$\frac{\lambda^4}{24}\mathbb{E}\left[(Z-\mathbb{E}Z)^4\right] \le e^{\frac{1}{8}\lambda^2\|\delta\|_2^2}$$

Choosing $\lambda = \frac{1}{\|\delta\|_2}$, we get

$$\mathbb{E}\left[(Z-\mathbb{E}Z)^4\right] \le 28\,\|\delta\|_2^4$$

Almost there!

$$\mathbb{E}[Z^4] \le \mathbb{E}\left[8(Z-\mathbb{E}Z)^4 + 8\mathbb{E}[Z]^4\right] \le 224\,\|\delta\|_2^4 + 8\mathbb{E}[Z]^4$$

which is... something.

Proof.

By Paley-Zygmund,

$$\mathbb{P}\left\{ Z^2 \geq \frac{1}{2}\mathbb{E}[Z^2] \right\} \geq \frac{1}{4}\frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]} = \Omega(1)$$

$\geq \frac{1}{4}\|\delta\|_2^2$

$\gtrsim \|\delta\|_2^4 + \mathbb{E}[Z]^4$

$\lesssim \|\delta\|_2^4 + \mathbb{E}[Z]^4$

and we are done! □

Part II: General L

Divide [k] in L parts of equal size, and apply the 1-bit isometry to the L subvectors separately:

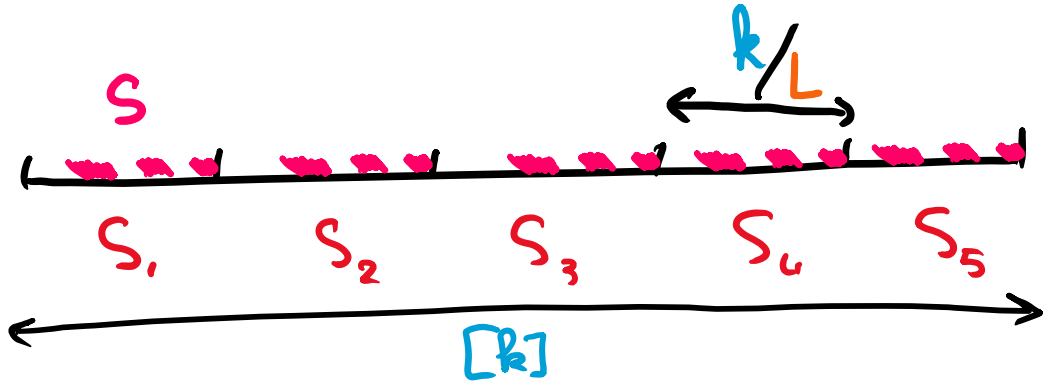
Part II: General L

$S_1, \dots, S_L$

Divide $[k]$ in L parts of equal size, and apply the 1-bit isometry to the L subvectors separately:

$$\forall j \in [L], \quad \mathbb{P}_S\left\{ |p^j(S) - q^j(S)| \geq \frac{c_1}{\sqrt{k/L}} \|p^j - q^j\|_1, \right\} \geqslant c_2$$

$$\|p^j - q^j\|_1 \overset{\leqslant}{c\text{-}s} \|p^j - q^j\|_2$$



Pick $S \subseteq [k/L]$ u.a.r, translate it in all L parts.

Part II: General L

Divide $[k]$ in L parts of equal size:

$$\forall j \in [L], \quad \mathop{\mathbb{P}}_{S}\left\{ |p^j(S) - q^j(S)| \geq \frac{c_1}{\sqrt{k/L}} \|p^j - q^j\|_1 \right\} \geq c_2$$

"Then"

$$\mathop{\mathbb{P}}_{S}\left\{ \sum_{j=1}^{L} |p^j(S) - q^j(S)| \geq \frac{c_1'}{\sqrt{k/L}} \sum_{j=1}^{L} \|p^j - q^j\|_1 \right\} \geq c_2'$$

and we are done. □

Part II: General L

What was that “Then”?

Part II: General L

What was that "Then"?

Lemma 3.4 (Additivity of tails). *Let* $a_1, \ldots, a_m \geq 0$, *and* $Y_1, \ldots, Y_m$ *be non-negative random variables such that for some* $c \in (0,1)$, $\Pr[\, Y_i \geq a_i \,] \geq c$ *for every* $1 \leq i \leq m$. *Then,*

$$\Pr\left[\, Y_1 + \ldots + Y_m \geq c \cdot \frac{a_1 + \ldots + a_m}{2} \,\right] \geq \frac{c}{2-c} \,.$$

Part II: General L

What was that "Then"?

Lemma 3.4 (Additivity of tails). *Let* $a_1, \ldots, a_m \geq 0$, *and* $Y_1, \ldots, Y_m$ *be non-negative random variables such that for some* $c \in (0,1)$, $\Pr[\, Y_i \geq a_i \,] \geq c$ *for every* $1 \leq i \leq m$. *Then,*

$$\Pr\left[\, Y_1 + \ldots + Y_m \geq c \cdot \frac{a_1 + \ldots + a_m}{2} \,\right] \geq \frac{c}{2-c}\,.$$

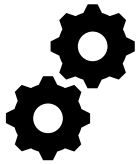
No independence of $Y_1, \ldots, Y_m$ required!

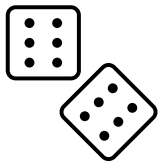
Domain Compression: we proved it

Theorem 2.12 (Domain Compression Lemma). There exist absolute constants $c_1, c_2 > 0$ such that the following holds. For any $2 \leq \ell \leq k$ and any $\mathbf{p}, \mathbf{q} \in \Delta_k$,

$$\Pr_{\Pi}\left[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_\Pi, \mathbf{q}_\Pi) \geq c_1 \sqrt{\frac{\ell}{k}} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \right] \geq c_2 \,,$$

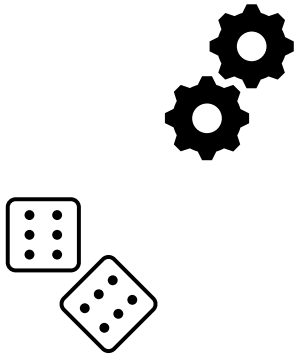
where $\Pi = (\Pi_1, \ldots \Pi_\ell)$ is a uniformly random partition of $[k]$ in ℓ subsets, and $\mathbf{p}_\Pi \in \Delta_\ell$ denotes the probability distribution on $[\ell]$ induced by $\mathbf{p}$ and Π via $\mathbf{p}_\Pi(i) = \mathbf{p}(\Pi_i)$.





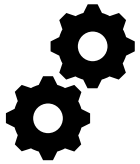
Domain Compression: comments

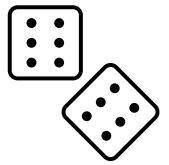
- Optimal, up to the constants
- Can generalise to other than uniformly random partition
  - E.g., equipartition
- Proof via Paley-Zygmund: only requires 4^{th} moments
  - This means 4-wise independent Bernoulli suffice!
  - This means O(log k) bits of shared randomness suffice!



Domain Compression: comments

- Optimal, up to the constants
- Can generalise to other than uniformly random partition
  - E.g., equipartition
- Proof via Paley-Zygmund: only requires 4^{th} moments
  - This means 4-wise independent Bernoulli suffice!
  - This means O(log k) bits of shared randomness suffice!
- But we can do better. General statement with s bits of randomness
  - Proof of the one-bit isometry much more involved
  - Cute derandomisation ideas involved
  - Check [Acharya–Canonne–Han–Sun–Tyagi'20]





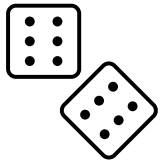
https://arxiv.org/abs/1907.08743

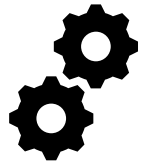
Once more: identity testing

	Private-Coin	Public-Coin
Centralised	$k^{1/2}/\varepsilon^2$	
Local privacy	$k^{3/2}/(\varepsilon^2\rho^2)$	$k/(\varepsilon^2\rho^2)$
Shuffle privacy*	$k^{3/4}/(\varepsilon\rho) + k^{1/2}/\varepsilon^2$	$k^{2/3}/(\varepsilon^{4/3}\rho^{2/3}) + k^{1/2}/\varepsilon^2 + k^{1/2}/(\varepsilon\rho)$
Bandwidth	$k^{3/2}/(\varepsilon^2 2^\ell) + k^{1/2}/\varepsilon^2$	$k/(\varepsilon^2 2^{\ell/2}) + k^{1/2}/\varepsilon^2$

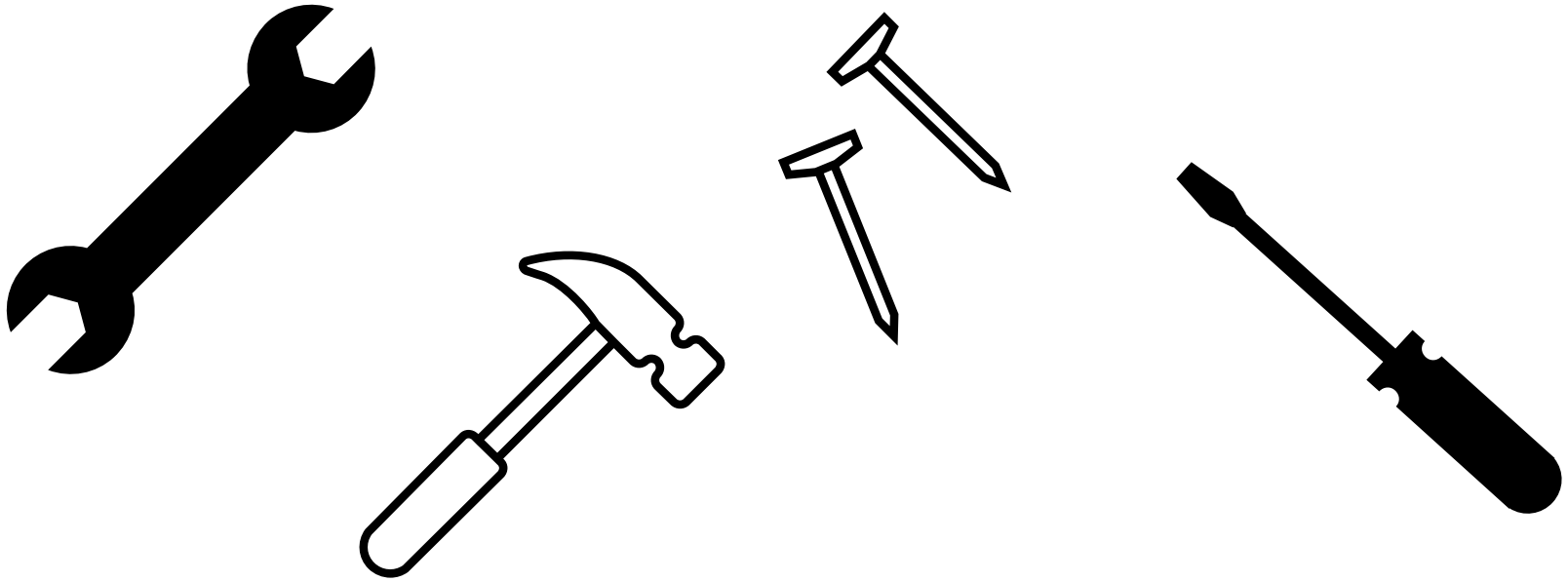
↳ Domain compression with s public coins: $\frac{\sqrt{k}}{\varepsilon^2}$ $\sqrt{\frac{k}{2^\ell \vee 1}}$ $\sqrt{\frac{k}{2^{s+\ell} \vee 1}}$

Conclusion



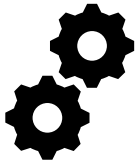


Domain compression is a versatile primitive! More applications?

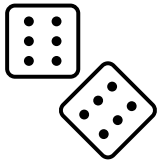


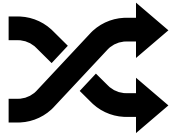
Thank you.

An example



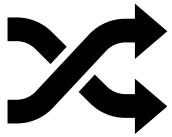
optimal private-coin
locally private tester

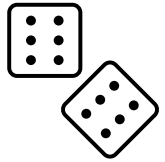




optimal public-coin
locally private tester

optimal private-coin
shuffle private tester





optimal public-coin
shuffle private tester

[Canonne–Lyu'21]