Trace reconstruction for deletion channels

Robin Pemantle

U. Penn. Dept. of Math

Trace reconstruction for deletion channels

Robin Pemantle

U. Penn. Dept. of Math

Based on joint work with

Nina Holden
M.I.T.

Yuval Peres
Microsoft Research

Trace reconstruction for deletion channels

Robin Pemantle

U. Penn. Dept. of Math

Based on joint work with

Nina Holden
M.I.T.

Yuval Peres
Microsoft Research

Thanks for the slides!

February 18, 2018

Problem statement

Deletion channel

- Suppose Alice wants to send to Bob an n-bit string $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in \{0, 1\}^n$.

Deletion channel

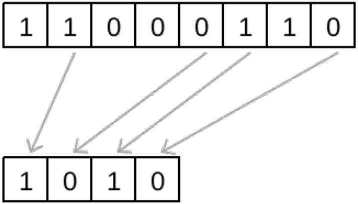
- Suppose Alice wants to send to Bob an n-bit string $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in \{0,1\}^n$.
- Alice transmits the bits one by one, but each bit has some probability q of being deleted.

Deletion channel

- Suppose Alice wants to send to Bob an n-bit string $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in \{0,1\}^n$.
- Alice transmits the bits one by one, but each bit has some probability q of being deleted.
- Bob doesn't know which positions were deleted; all he sees is a shortened string $(y_0, y_1, \ldots, y_{\ell-1})$

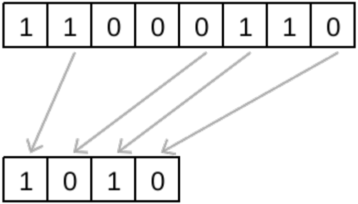
Deletion channel

- Suppose Alice wants to send to Bob an n-bit string $\mathbf{x} = (x_0, \dots, x_{n-1}) \in \{0,1\}^n$.
- Alice transmits the bits one by one, but each bit has some probability q of being deleted.
- Bob doesn't know which positions were deleted; all he sees is a shortened string $(y_0, y_1, \dots, y_{\ell-1})$



Deletion channel

- Suppose Alice wants to send to Bob an n-bit string $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in \{0,1\}^n$.
- Alice transmits the bits one by one, but each bit has some probability q of being deleted.
- Bob doesn't know which positions were deleted; all he sees is a shortened string $(y_0, y_1, \ldots, y_{\ell-1})$



- Erasures/mistakes much easier than deletions, due to synchronicity.

Questions

- Notation: $\mathcal{D}_q(\mathbf{x})$ denotes the distribution over strings that Bob receives after passing through deletion channel.

Questions

- Notation: $\mathcal{D}_q(\mathbf{x})$ denotes the distribution over strings that Bob receives after passing through deletion channel.
- Given T i.i.d. samples ("traces") $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^T$ with each $\mathbf{y}^t \sim \mathcal{D}_q(\mathbf{x})$, can Bob reconstruct $\mathbf{x}$?

Questions

- Notation: $\mathcal{D}_q(\mathbf{x})$ denotes the distribution over strings that Bob receives after passing through deletion channel.
- Given T i.i.d. samples ("traces") $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^T$ with each $\mathbf{y}^t \sim \mathcal{D}_q(\mathbf{x})$, can Bob reconstruct $\mathbf{x}$?
- Can ask for worst case $\mathbf{x}$ or for "average case" $\mathbf{x}$ (where $\mathbf{x}$ is chosen uniformly at random).

Questions

- Notation: $\mathcal{D}_q(\mathbf{x})$ denotes the distribution over strings that Bob receives after passing through deletion channel.
- Given T i.i.d. samples ("traces") $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^T$ with each $\mathbf{y}^t \sim \mathcal{D}_q(\mathbf{x})$, can Bob reconstruct $\mathbf{x}$?
- Can ask for worst case $\mathbf{x}$ or for "average case" $\mathbf{x}$ (where $\mathbf{x}$ is chosen uniformly at random).
- Arises naturally in various contexts: sensor networks, DNA sequencing.

Prior Results

- Problem raised in this form by Batu, Kannan, Khanna and McGregor (2004). Their lower bound: For all $n > 1$ there exist strings x, x' of n bits such that $\Omega(n)$ traces are needed to distinguish whether the input was x or x'.

Prior Results

- Problem raised in this form by Batu, Kannan, Khanna and McGregor (2004). Their lower bound: For all $n > 1$ there exist strings x, x' of n bits such that $\Omega(n)$ traces are needed to distinguish whether the input was x or x'.
- (Holenstein-Mitzenmacher-Panigrahy-Wieder 2008): $e^{O(\sqrt{n})}$ in worst case and $n^{O(1)}$ in random case for $q < 1/100$

Prior Results

- Problem raised in this form by Batu, Kannan, Khanna and McGregor (2004). Their lower bound: For all $n > 1$ there exist strings x, x' of n bits such that $\Omega(n)$ traces are needed to distinguish whether the input was x or x'.
- (Holenstein-Mitzenmacher-Panigrahy-Wieder 2008): $e^{O(\sqrt{n})}$ in worst case and $n^{O(1)}$ in random case for $q < 1/100$
- (Nazarov-Peres and De-O'Donnell-Servedio, both in STOC 2017): For worst case $\mathbf{x}$, we can reconstruct using $e^{O(n^{1/3})}$ traces. Moreover, this is optimal for linear (mean-based) tests.

Prior Results

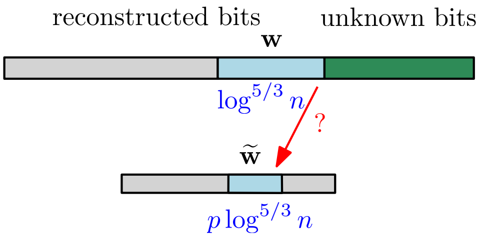
- Problem raised in this form by Batu, Kannan, Khanna and McGregor (2004). Their lower bound: For all $n > 1$ there exist strings x, x' of n bits such that $\Omega(n)$ traces are needed to distinguish whether the input was x or x'.
- (Holenstein-Mitzenmacher-Panigrahy-Wieder 2008): $e^{O(\sqrt{n})}$ in worst case and $n^{O(1)}$ in random case for $q < 1/100$
- (Nazarov-Peres and De-O'Donnell-Servedio, both in STOC 2017): For worst case $\mathbf{x}$, we can reconstruct using $e^{O(n^{1/3})}$ traces. Moreover, this is optimal for linear (mean-based) tests.
- (Peres-Zhai, FOCS 2017): For $q < 1/2$, we can reconstruct a uniform random input $\mathbf{x}$ with probability $1 - o(1)$ using

$$T = e^{C\sqrt{\log n}} = n^{o(1)} \text{ traces.}$$

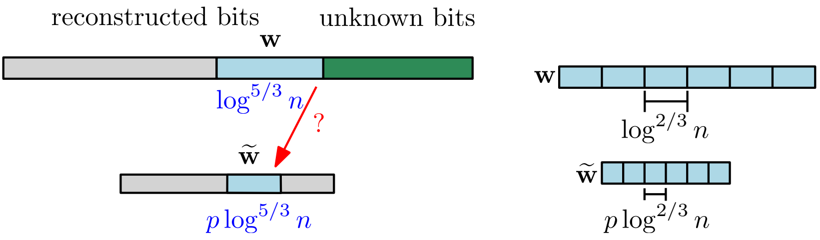
Today

Lower	AVERAGE $q < 1/100$	AVERAGE $q < 1/2$	AVERAGE	Worst case
$\Omega(n)$ [BKKM04]				
	$n^{O(1)}$ [HMPW08]			$e^{\sqrt{n}}$ [HMPW08]
				$e^{\sqrt[3]{n}}$ STOC17
		$e^{\sqrt{\log n}}$ [FOCS17]		
			$e^{\sqrt[3]{\log n}}$ [HPP18]	

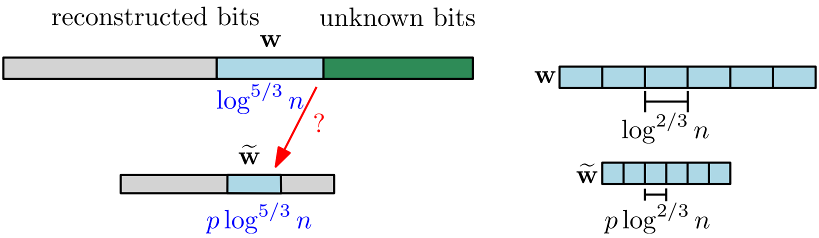
Alignment with error $\sqrt[3]{\log(n)}$



- Was $\widetilde{\mathbf{w}}$ likely obtained by sending $\mathbf{w}$ through the deletion channel?

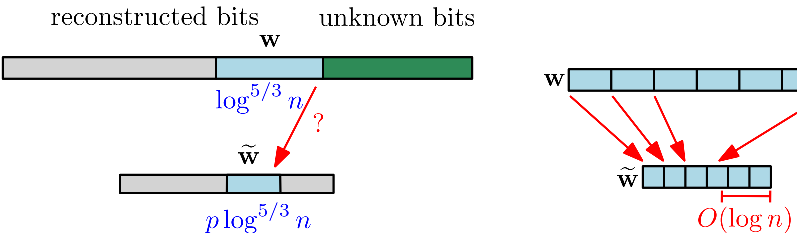


- Was $\widetilde{\mathbf{w}}$ likely obtained by sending $\mathbf{w}$ through the deletion channel?
  - Divide $\widetilde{\mathbf{w}}$ and $\mathbf{w}$ into $\log n$ blocks, see how many blocks have same majority in $\widetilde{\mathbf{w}}$ and $\mathbf{w}$

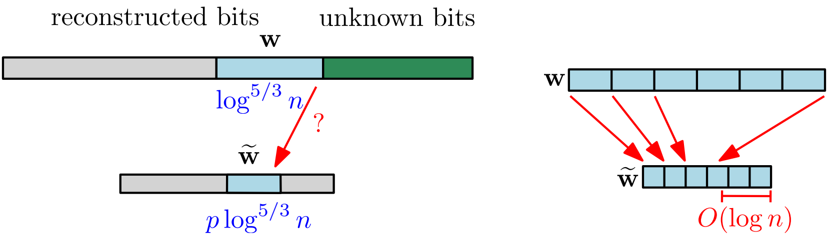


- Was $\widetilde{\mathbf{w}}$ likely obtained by sending $\mathbf{w}$ through the deletion channel?
  - Divide $\widetilde{\mathbf{w}}$ and $\mathbf{w}$ into $\log n$ blocks, see how many blocks have same majority in $\widetilde{\mathbf{w}}$ and $\mathbf{w}$
- Repeat with all strings $\widetilde{\mathbf{w}}$ of appropriate length.

Alignment with error $\sqrt[3]{\log(n)}$

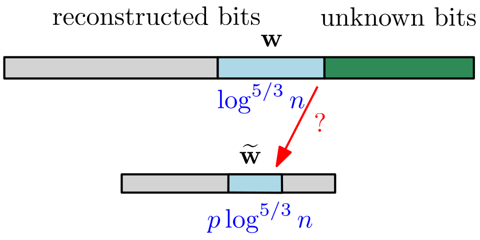


- Was $\widetilde{\mathbf{w}}$ likely obtained by sending $\mathbf{w}$ through the deletion channel?
  - Divide $\widetilde{\mathbf{w}}$ and $\mathbf{w}$ into $\log n$ blocks, see how many blocks have same majority in $\widetilde{\mathbf{w}}$ and $\mathbf{w}$
- Repeat with all strings $\widetilde{\mathbf{w}}$ of appropriate length.
- Alignment error $O(\log n)$ with probability $1 - \exp(-\Omega(\log^{1/3} n))$.



- Was $\widetilde{\mathbf{w}}$ likely obtained by sending $\mathbf{w}$ through the deletion channel?
  - Divide $\widetilde{\mathbf{w}}$ and $\mathbf{w}$ into $\log n$ blocks, see how many blocks have same majority in $\widetilde{\mathbf{w}}$ and $\mathbf{w}$
- Repeat with all strings $\widetilde{\mathbf{w}}$ of appropriate length.
- Alignment error $O(\log n)$ with probability $1 - \exp(-\Omega(\log^{1/3} n))$.
- Error improved to $\sqrt[3]{\log n}$ with second step, window size $\log n$

Alignment with error $\sqrt[3]{\log(n)}$



- Was $\widetilde{\mathbf{w}}$ likely obtained by sending $\mathbf{w}$ through the deletion channel?
  - Divide $\widetilde{\mathbf{w}}$ and $\mathbf{w}$ into $\log n$ blocks, see how many blocks have same majority in $\widetilde{\mathbf{w}}$ and $\mathbf{w}$
- Repeat with all strings $\widetilde{\mathbf{w}}$ of appropriate length.
- Alignment error $O(\log n)$ with probability $1 - \exp(-\Omega(\log^{1/3} n))$.
- Error improved to $\sqrt[3]{\log n}$ with second step, window size $\log n$
- $\sqrt[3]{\log n}$ error implies $e^{\sqrt[3]{\log n}}$ traces by Peres-Nazarov