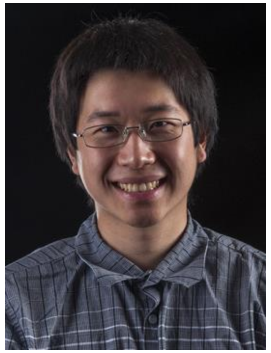
Approximate $\mathbb{F}_2$-Sketching of Valuation Functions

Grigory Yaroslavtsev

Samson Zhou



Carnegie Mellon University

$\mathbb{F}_2$-Sketching

Input $x \in \{0,1\}^n$

Parity = Linear function over $\mathbb{G}F_2$: $\bigoplus_{i \in S} x_i$

Deterministic linear sketch: set of k parities:

$$\ell(x) = \bigoplus_{i_1 \in S_1} x_{i_1}; \quad \bigoplus_{i_2 \in S_2} x_{i_2}; \quad \dots; \quad \bigoplus_{i_k \in S_k} x_{i_k}$$

E.g. $\quad x_4 \oplus x_2 \oplus x_{42}; \quad x_{239} \oplus x_{30}; \quad x_{566};\dots$

Randomized linear sketch: distribution over k parities (random $S_1, S_2, \dots, S_k$):

$$\ell(x) = \bigoplus_{i_1 \in S_1} x_{i_1}; \bigoplus_{i_2 \in S_2} x_{i_2}; \dots; \bigoplus_{i_k \in S_k} x_{i_k}$$

Linear sketching over $\mathbb{F}_2$

Given $f(x): \{0,1\}^n \rightarrow \{0,1\}$

Question:

Can one recover $f(x)$ from a small ($k \ll n$) linear sketch over $\mathbb{F}_2$?

Allow randomized computation (99% success)

- Probability over choice of random sets
- Sets are known at recovery time
- Recovery is deterministic (w.l.o.g)

Application: Distributed Computing

Distributed computation among M machines:

$x = (x_1, x_2, \dots, x_M)$ (more generally $x = \bigoplus_{i=1}^{M} x_i$)

M machines can compute sketches locally:

$$\ell(x_1), \dots, \ell(x_M)$$

Send them to the coordinator who computes:

$$\ell_i(x) = \ell_i(x_1) \oplus \cdots \oplus \ell_i(x_M) \text{ (coordinate-wise XORs)}$$

Coordinator computes $f(x)$ with kM communication

$\boldsymbol{x}$	0	1	1	1	1	1	0	0	0	0
			$\boldsymbol{x_1}$					$\boldsymbol{x_2}$		

Application: Streaming

x generated through a sequence of updates

Updates $i_1, \ldots, i_m$: update i_t flips bit at position i_t

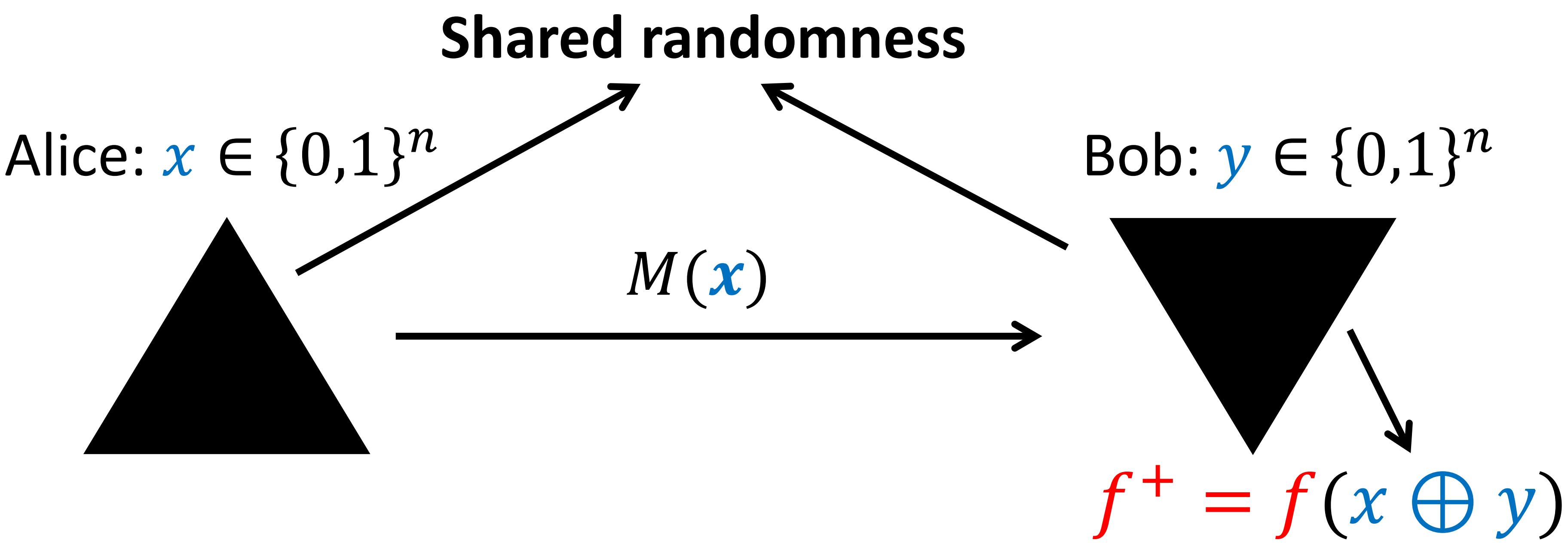
$x^{(0)}$	0	0	0	0	0	0	0	0	0	0

Updates: (1, 3, 8, 3)

$x^{(1)}$	1	0	0	0	0	0	0	0	0	0
$x^{(2)}$	1	0	1	0	0	0	0	0	0	0
$x^{(3)}$	1	0	1	0	0	0	0	1	0	0
x	1	0	0	0	0	0	0	1	0	0

$\ell(x)$ allows to recover $f(x)$ with k bits of space

Puzzle: Open Problem 78 on Sublinear.info



Conjecture: (Almost) shortest message is a randomized $\mathbb{F}_2$-sketch

https://sublinear.info/index.php?title=Open_Problems:78

Deterministic vs. Randomized

Fact: f has a deterministic sketch if and only if

$$f = g(\bigoplus_{i_1 \in S_1} x_{i_1}; \bigoplus_{i_2 \in S_2} x_{i_2}; \dots; \bigoplus_{i_k \in S_k} x_{i_k})$$

Equivalent to "f has Fourier dimension k"

Randomization can help:

OR: $f(x) = x_1 \vee \cdots \vee x_n$

Has "Fourier dimension" $= n$

Pick $t = \log 1/\delta$ random sets $S_1, \dots, S_t$

If there is j such that $\bigoplus_{i \in S_j} x_i = 1$ output 1, otherwise output 0

Error probability δ

Approximate $\mathbb{F}_2$-Sketching

Exact sketching complexity of many functions studied by [Kannan, Mossel, Sanyal, Yaroslavtsev'17].

Recursive majority functions, Fourier sparse functions, etc.

$f(x_1, \ldots, x_n): \{0,1\}^n \to \mathbb{R}$

Normalize: $\|f\|_2$

Question:

Can one compute f': $\mathbb{E}[\left(f(x) - f'(x)\right)^2 \leq \epsilon]$ from a small ($k \ll n$) linear sketch over $\mathbb{F}_2$?

Our Results

Additive ($\sum_{i=1}^{n} w_i x_i$):

- $\Theta\left(\min\left(\frac{||w||_1^2}{\epsilon}, n\right)\right)$ (optimal via Index/Gap Hamming)

Budget-additive ($\min(b, \sum_{i=1}^{n} w_i x_i)$):

- $\Theta\left(\min\left(\frac{||w||_1^2}{\epsilon}, n\right)\right)$

Coverage:

- Optimal $O\left(\frac{1}{\epsilon}\right)$ (via L_1-Sampling)

Matroid rank (various results depending on rank r)

α-Lipschitz submodular functions:

- $\Omega(n)$ communication lower bound for $\alpha = \Omega(1/n)$
- Uses a large family of matroids from [Balcan, Harvey'10]

Technical Theorems

Any $f\colon \{0,1\}^n \to \mathbb{R}$ has a randomized linear sketch of size $O\left(\frac{\|\hat{f}\|_1^2}{\epsilon}\right)$.

(θ, m)-LTF have randomized linear sketches of size $O\left(\frac{\theta}{m}\log\frac{\theta}{m}\right)$.

$\mathrm{HAM}_{\leq d}\left(\bigvee_{i\in S_1} x_i, \bigvee_{i\in S_2} x_i, \ldots\right)$ has a randomized linear sketch of size $O(d^2 \log d)$.

Linear Threshold Functions

$f:\{0,1\}^n \to \{0,1\}$ is a linear threshold function (LTF) if there exist constants $w_1, w_2, \dots, w_n, \theta$ such that $f(x) = 1$ if $\sum_{i=1}^{n} w_i x_i \geq \theta$ and $f(x) = 0$ otherwise.

$f:\{0,1\}^n \to \{0,1\}$ is a (θ, m)-LTF if f is monotone and for all x, $m \leq |\sum_{i=1}^{n} w_i x_i - \theta|$.

- Randomized linear sketches of size $O\left(\frac{\theta}{m}\log n\right)$ [Liu, Zhang'13].
- Question [Montanaro, Osborne'09]: Does there exist a protocol for $f(x \oplus y)$ with communication complexity $O\left(\frac{\theta}{m}\log\frac{\theta}{m}\right)$?

Sketching (θ, m)-LTFs

$\sum_{i=1}^{n} w_i x_i \; ? \; \theta$ and $m \leq |\sum_{i=1}^{n} w_i x_i - \theta|$

Observation 1: Any $w_i \leq \frac{m}{2}$ can be set to 0.

Observation 2: Support of $\{x \mid f(x) = 0\}$ is small $\sim n^{\frac{2\theta}{m}}$

Theorem [Montanaro, Osbourne'09, KMSY'18]: If $\Pr[f(x) = 0] \leq \zeta$, then there exists a sketch of size $O(\log 2^{n+1} \zeta)$.

Already enough to get sketch of size $O\left(\frac{\theta}{m} \log n\right)$

Sketching (θ, m)-LTFs

Observation 3: Any w_i can be rounded down to $w_i' = \frac{m}{2}(1+\xi)^k$.

For $f(x) = 0$, $-m \geq \sum_{i=1}^{n} w_i x_i - \theta \geq \sum_{i=1}^{n} w_i{}' x_i - \theta$, so a margin of m remains.

For $f(x) = 1$, $m + \theta \leq \sum_{i=1}^{n} w_i x_i \leq \sum_{i=1}^{n} (1+\xi) w_i' x_i$, so

$$\sum_{i=1}^{n} w_i' x_i \geq (1-\xi)(m+\theta) \geq \theta + m - 2\xi\theta,$$

since $\theta \geq m$ and a margin of $\frac{4}{5}m$ remains when setting $\xi = \frac{\theta}{10m}$.

Separation Sketch

There is a randomized linear sketch with size $O(1)$ for the function $g(x) = 1$ if $\|x\|_0 \geq 2d$ and $g(x) = 0$ if $\|x\|_0 \leq d$ where $x \in \{0,1\}^n$ and g can answer arbitrarily if one of the above cases doesn't hold. [HuangShiZhangZhu'06]

Recall: at most $\frac{2\theta}{m}$ nonzero coordinates when $f(x) = 0$.

Use above sketch to catch all instances with more than $\frac{2\theta}{m}$ nonzero coordinates.

Sparse recovery when fewer than $\frac{2\theta}{m}$ nonzero coordinates.

Sparse "Recovery"

Recall: all weights $\frac{m}{2}(1+\xi)^k$, $k = O\left(\frac{\theta}{m}\log\frac{\theta}{m}\right)$, interested in $\frac{2\theta}{m}$ nonzeros.

Use $O\left(\frac{\theta}{m}\log\frac{\theta}{m}\right)$ levels and $O\left(\left(\frac{\theta}{m}\right)^2\right)$ buckets to avoid hash collision.

Consider each entry as a separate variable, reduction to $O\left(\left(\frac{\theta}{m}\right)^3 \log\frac{\theta}{m}\right)$ variables.

Sketching $(\theta, O(m))$-LTF on $O\left(\left(\frac{\theta}{m}\right)^3 \log\frac{\theta}{m}\right)$ variables, using $O\left(\frac{\theta}{m}\log\frac{\theta}{m}\right)$ space.

Technical Theorems

Any $f : \{0,1\}^n \to \mathbb{R}$ has a randomized linear sketch of size $O\left(\frac{\|\hat{f}\|_1^2}{\epsilon}\right)$.

(θ, m)-LTF have randomized linear sketches of size $O\left(\frac{\theta}{m} \log \frac{\theta}{m}\right)$.

$\mathrm{HAM}_{\leq d}\left(\bigvee_{i \in S_1} x_i, \bigvee_{i \in S_2} x_i, \ldots\right)$ has a randomized linear sketch of size $O(d^2 \log d)$.

Our Results

Class	Error	Distribution	Complexity	Result
Additive/Budget additive $\min(b, \sum_{i=1}^{n} w_i x_i)$	ϵ	any	$\Theta\left(\frac{\|w\|_1^2}{\epsilon}\right)$	Theorem A.7, D.1 Corollary A.3, A.6
$\min(c\sqrt{n}, \frac{2c}{\sqrt{n}} \sum_{i=1}^{n} x_i)$	constant	uniform	$\Omega(n)$	Theorem D.1
Coverage	ϵ	any	$O\left(\frac{1}{\epsilon}\right)$	Corollary A.4
Matroid Rank 2	exact	any	$\Theta(1)$	Theorem 3.1
Graphic Matroids Rank r	exact	any	$O(r^2 \log r)$	Theorem 3.5
Matroid Rank r	exact	any	$\Omega(r)$	Corollary 3.24
Matroid Rank r	exact	uniform	$O((r \log r + c)^{r+1})$	Corollary E.6
Matroid Rank	$1/\sqrt{n}$	uniform	$\Theta(1)$	Corollary E.8
$\frac{c}{n}$-Lipschitz Submodular	constant	any	$\Theta(n)$	Theorem 3.17

Table 1: Linear sketching complexity of classes of valuation functions

Frequently Asked Questions

Thanks!

Q: Why $\mathbb{F}_2$ updates instead of ± 1?

Often doesn't help if you know the sign

Q: How to store random sets?

Derandomize using Nisan's PRG – extra $O(\log n)$ factor in space

Q: Specific applications?

Essentially all dynamic graph streaming algorithms can be based on L_0-sampling

L_0-sampling can be done optimally using $\mathbb{F}_2$-sketching [Kapralov et al. FOCS'17]