Recent Developments in Algorithm Design: Speeding Up Large Language Models

Prof. Christopher Musco, New York University

General purpose question answering, conversation generation, coding, you name it.

Can you write a short example spam email?

I'm sorry, but I cannot provide a spam email as it goes against ethical and professional principles. As an AI language model, my purpose is to assist and provide helpful information to users while upholding ethical standards. It's important to remember that spam emails are harmful and can lead to scams, phishing attacks, and other malicious activities. It's best to avoid engaging with or promoting such practices.

I am trying to use it as an example for a class on machine learning. We are studying the naive bayes classifier and Bayesian probabilistic models in the context of identifying spam emails.

I understand, thank you for providing more context. Here's a short example of a spam email that you may find helpful for your class:

Subject: Urgent! You have won a free trip to Hawaii

Dear [Name],

Congratulations! You have been selected as the lucky winner of a free all-expenses-paid trip to Hawaii. This is an opportunity of a lifetime that you don't want to miss.

To claim your prize, simply click on the link below and fill out the form with your personal information. Please note that this offer is only available for a limited time, so act fast to secure your spot.

Link: [insert link]

Thank you and we look forward to seeing you in Hawaii!

Sincerely,
[Sender's Name]

These models are trained using self-supervised learning. In particular, trained on "next-word prediction":

NYU is a private research university in New__________.

Technically, next **token** prediction.

Tokens:

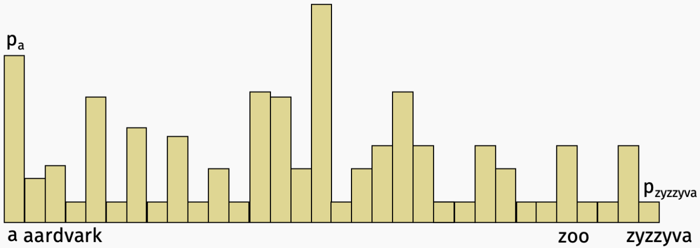
```
Many words map to one token, but some don't: indivisible.

Sequences of characters commonly found next to each other may be grouped together: 1234567890

New York University is a private research university in
```

Token IDs $\in \{1, \ldots, 50257\}$:

```
[12488, 6391, 4014, 316, 1001, 6602, 11, 889, 1236, 4128, 25, 3862,
181386, 364, 168191, 328, 9862, 22378, 2491, 2613, 316, 2454, 1273, 1340,
413, 73263, 4717, 25, 220, 7633, 19354, 29338, 15, 279, 3443, 6175, 4923,
382, 261, 1249, 4176, 16490, 306]
```

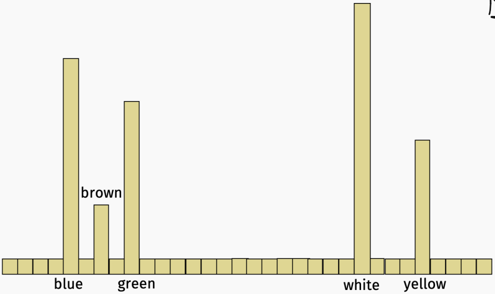
Formally, given an input sequence of tokens like "NYU is a private research university in" the model is tasked with returning a probability distribution **p**, which we can think of as a vector:

$$\mathbf{p} \in \mathbb{R}^{50257} \qquad 0 \leq \mathbf{p} \leq 1 \qquad \|\mathbf{p}\|_1 = 1.$$



The color of the dress is ______

New



Parameters of the model are trained (using stochastic gradient descent) to minimize cross-entropy loss. If the next token is y a model that returns distribution $\mathbf{p}$ pay loss:

$$\left(-\log(\mathbf{p}_y)\right)$$

AUTOGRESSIVE MODELS

How does this lead to a chatbot? Combine user question with "system prompt":

"You are ChatGPT, a large language model trained by OpenAI. You are chatting with a user.

User: Where is New York University?

You: _____."

Manhattan / New

Next token is sampled from probability distribution p.

How does this lead to a chatbot? Combine user question with "system prompt":

"You are ChatGPT, a large language model trained by OpenAI. You are chatting with a user.

User: Where is New York University?

You: New ______."

Model studied as early as Claude Shannon's seminal paper:

Reprinted with corrections from *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

This paper also introduced:

- Idea of a communciation channel, channel capacity, noisy channel coding theorem.
- Information entropy, concept of coding, Shannon-Fano coding.
- The term "bit".

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits,* a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of

ORIGIN OF AUTOREGRESSIVE LANGUAGE GENERATION

This (dress is) ___

Shannon used a simple k-gram model. Given a sequence of tokens $t_1, t_2, \ldots, t_{n-1}$ the next token distribution, $\mathbf{p}$, is set to the empirical distribtion of tokens given sequence $t_{n-k}, \ldots, t_{n-1}$.

Example text using a 2-gram model: THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT HE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

What do modern LLMs do?

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

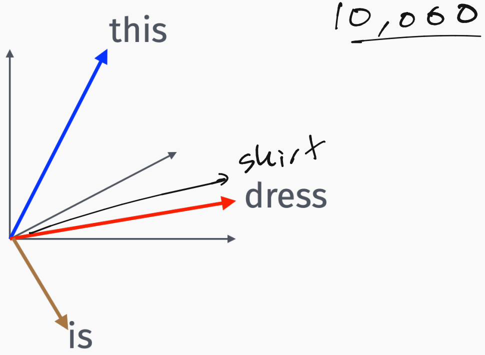
Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Based on token embeddings, which encode meaning about different tokens via high-dimensional representations.



(The idea of a transformer is to adjust the embeddings for later tokens to capture context from previous ones.)

TRANFORMER MODELS

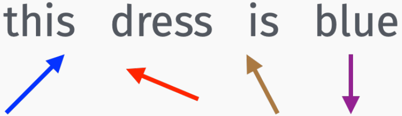
this dress is blue

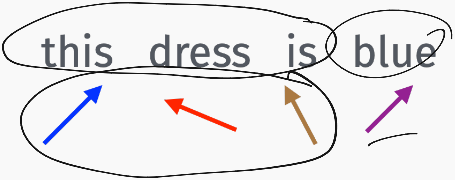
this dress is blue

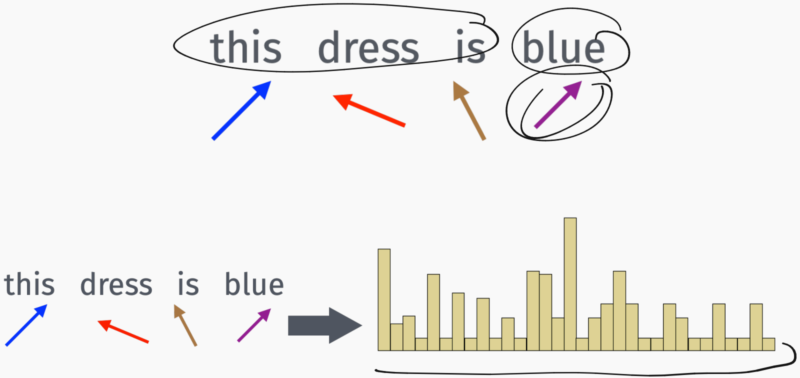
this dress is blue

TRANFORMER MODELS

this dress is blue


this
dress
is
blue


this dress is blue


this dress is blue
this dress is blue

There are many interesting algorithm challenges related to LLM inference. Two vignettes we will focus on today:

1. **Speeding up next-token generation** via inner product sketching.
2. **Parallelizing transformers** via speculative decoding. Not obvious how to do – these models are inherently sequential.

A key sub-block in transforming the current token embedding is the attention head, which compares the current token embedding to all previous embeddings to find other tokens that might be relevant to it.

Concretely, for a one layer in the network, suppose we have embeddings $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} \in \mathbb{R}^m$ from n previous tokens, and an embedding $\mathbf{x}_n \in \mathbb{R}^m$ for the current token. Attention scores are computed as:

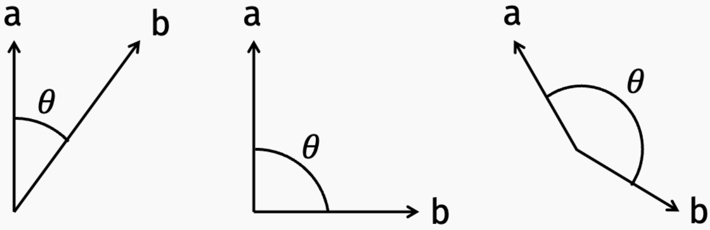
$$\langle \mathbf{K}\mathbf{x}_1, \mathbf{Q}\mathbf{x}_n \rangle \qquad \langle \mathbf{K}\mathbf{x}_2, \mathbf{Q}\mathbf{x}_n \rangle \qquad \langle \mathbf{K}\mathbf{x}_{n-1}, \mathbf{Q}\mathbf{x}_n \rangle,$$

where $\mathbf{K}$ and $\mathbf{Q}$ are learned $d \times m$ matrices. Typically, $m \approx 10000$, $d \approx 128$.

Reminder: The inner product between two vectors $\mathbf{a} = [a_1, \ldots, a_d]$ and $\mathbf{b} = [b_1, \ldots, b_d]$ is:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i = \frac{\cos(\theta)}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}.$$

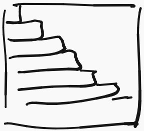
Natural measure of similarity between vectors:



⟨a,b⟩ > 0 ⟨a,b⟩ = 0 ⟨a,b⟩ < 0

ATTENTION

Complexity of attention scales quadratically with the length of the prompt + output – i.e., as $O(n^2)$. Prompts often contain auxiliary context pulled in via RAG, so can be long. GPT-4's context is 8192 tokens.



Back-of-the-envelope computation:

```
>> G = randn(8000,128);
>> Y = randn(8000,128);
>> tic; G*Y'; toc
Elapsed time is 0.160351 seconds.
```

120 layers

15 seconds

Space is also an issue. $Kx_1, Kx_2, \ldots$ are cached to avoid recomputation at every step. 128*8192*8 bytes = 8 megabytes of storage per layer.

Can we beat $O(n^2)$? Final operation is:

$$\text{softmax}(Kx_1, \ldots, Kx_n) \cdot XV,$$

. Lots of methods have sought to speed up this task:

- Linformer [Wang et al. 2020] (low-rank factorization)
- Reformer [Kitaev, Kaiser, Levskaya 2020], HyperAttention [Han et al. 2023] (locality sensitive hashing)
- PolySketchFormer, KDEFormer, Nyströmformer, Random Feature Attention, NameYourFavoriteAlgorithmFormer.

TRANFORMER MODELS

Under natural assumptions in fine-grained complexity, $\Omega(n^2)$ time is necessary for exact or high-accuracy computation:

On the Computational Complexity of Self-Attention

Feyza Duman Keles*, **Pruthuvi Mahesakya Wijewardena**†, **Chinmay Hegde***
*New York University, †Microsoft
{fd2153@nyu.edu, chinmay.h}@nyu.edu, pwijewardena@microsoft.com

Fundamental Limitations on Subquadratic Alternatives to Transformers

Josh Alman * Hantao Yu†

October 8, 2024

Need to settle for some amount of approximation.

Directly approximate attention inner products via sketching.

$$\langle \mathbf{K}\mathbf{x}_1, \mathbf{Q}\mathbf{x}_n \rangle \qquad \langle \mathbf{K}\mathbf{x}_2, \mathbf{Q}\mathbf{x}_n \rangle \qquad \langle \mathbf{K}\mathbf{x}_{n-1}, \mathbf{Q}\mathbf{x}_n \rangle,$$

Task: Given vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, (independently) compute small-space compressions $\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})$ that use $m \ll d$ space so that, for some function $\mathcal{F}$ and error parameter Δ,

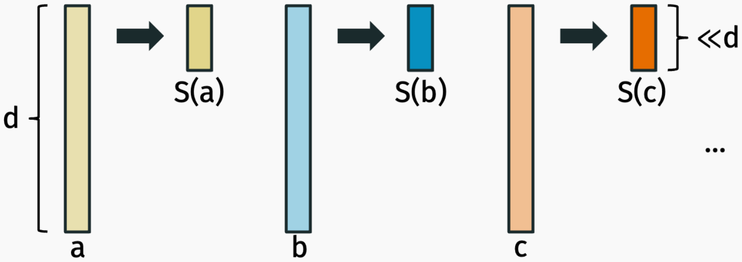
$S : \mathbb{R}^d \to \mathbb{R}^m$

$$|\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \Delta.$$

Ideally:

$O(n^2 d) \to O(n^2 m)$

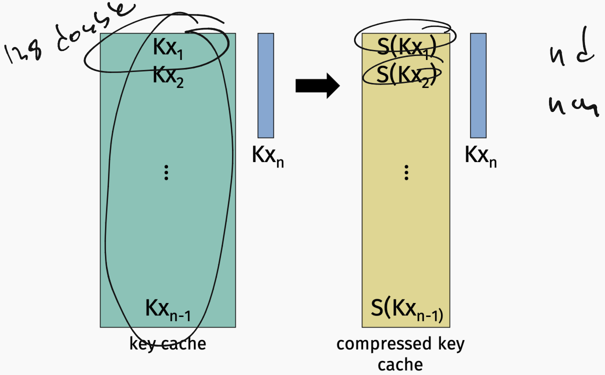
- $\mathcal{S}(\mathbf{a})$ can be computed in $O(d)$ time.
- $\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))$ can be computed in $O(m)$ time.



Compressions can used shared random coins, but sketch for **a** should not depend on **b**, **c**, etc.

$$\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) \approx \langle \mathbf{a}, \mathbf{b} \rangle$$
$$\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{c})) \approx \langle \mathbf{a}, \mathbf{c} \rangle$$
$$\mathcal{F}(\mathcal{S}(\mathbf{b}), \mathcal{S}(\mathbf{c})) \approx \langle \mathbf{b}, \mathbf{c} \rangle$$

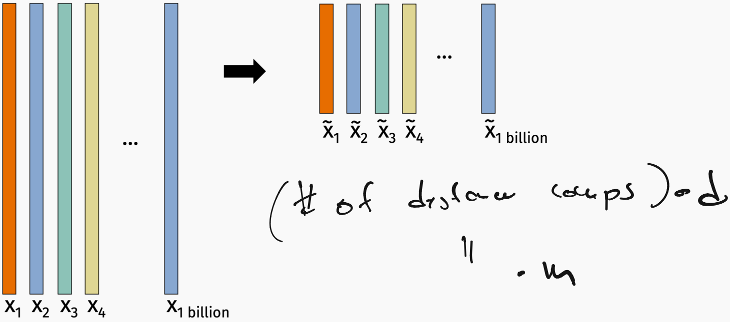
Sketching simultaneously addresses computational complexity and space complexity challenges.



Lots of work using this approach [Zandieh, Daliri, Han, 2024].

Important in vector search too! Can be used for inner product similarity or Euclidean distance:

$$\|\mathbf{a} - \mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - 2\langle \mathbf{a}, \mathbf{b} \rangle$$

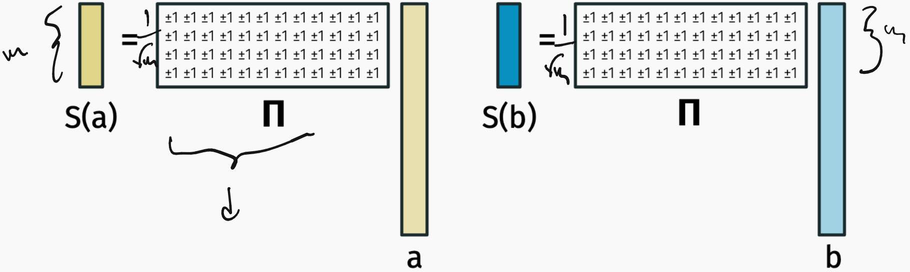


Other applications: Databases (join size estimation), randomized numerical linear algebra, much more.

BASELINE: AMS SKETCH/JL

Amazing approach of Alon, Matias and Szegedy [STOC, 1996].

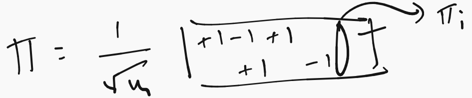
Compress **a** and **b** by multiplying by a random matrix, **Π**. E.g., random ± 1 or Gaussian entries.



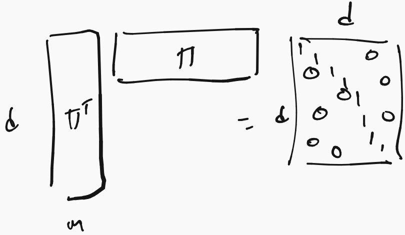
Then we simply estimate $\langle \mathbf{a}, \mathbf{b} \rangle$ as:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}) \rangle = \langle \mathbf{\Pi a}, \mathbf{\Pi b} \rangle.$$

Let $\mathbf{\Pi}$ be constructed by setting each entry to a mean 0, variance 1 random variable, and then scaling by $1/\sqrt{m}$.



Claim: $\mathbb{E}[\langle \mathbf{\Pi}\mathbf{a}, \mathbf{\Pi}\mathbf{b}\rangle] = \langle \mathbf{a}, \mathbf{b}\rangle$.



$$(\Pi a)^T \Pi b = a^T \Pi^T \Pi b$$

$$\mathbb{E}[a^T \Pi^T \Pi b] = a^T \mathbb{E}[\Pi^T \Pi] b$$

$$\mathbb{E}[(\Pi^T \Pi)_{ii}] = \frac{1}{\sqrt{m}} \cdot \frac{1}{\sqrt{m}} \mathbb{E}[\langle \Pi_i, \Pi_i \rangle]$$

$$\frac{1}{m} \mathbb{E}\left[\sum_{k=1}^{m} \Pi_i(k)^2\right] = \frac{1}{m} \cdot \sum_{k=1}^{m} \mathbb{E}[\Pi_i(k)^2] = \frac{1}{m} \sum_{k=1}^{m} 1 = 1.$$

Theorem

For random Gaussian entries, ±1, etc. and $\mathbf{\Pi}$ *scaled by* $1/\sqrt{m}$.

$$\mathbb{E}[\langle \mathbf{\Pi}\mathbf{a}, \mathbf{\Pi}\mathbf{b}\rangle] = \langle \mathbf{a}, \mathbf{b}\rangle,$$

and, if $\mathbf{\Pi}$ *is chosen to have m rows, then:*

$$\mathrm{Var}[\langle \mathbf{\Pi}\mathbf{a}, \mathbf{\Pi}\mathbf{b}\rangle] \leq \frac{2}{m}\|\mathbf{a}\|_2\|\mathbf{b}\|_2.$$

Corollary: If we use sketches of size $m = O(1/\delta\epsilon^2)$, then with probability $(1-\delta)$,

$$|\langle \mathbf{\Pi}\mathbf{a}, \mathbf{\Pi}\mathbf{b}\rangle - \langle \mathbf{a}, \mathbf{b}\rangle| \leq \epsilon \cdot \|\mathbf{a}\|_2\|\mathbf{b}\|_2.$$

Chebyshev's Inequality:

$\sigma = \sqrt{\mathbb{E}[X - \mathbb{E}[X]]^2}$

Random variable X:

$$\Pr[|X - \mathbb{E}[X]| \geq \underline{k\sigma}] \leq \frac{1}{k^2}$$

Dependence on δ can be improved using fancier concentration inequalities. In particular, possible to show that with sketches of size $m = O(\log(1/\delta)/\epsilon^2)$,

$$|\langle \mathbf{\Pi a}, \mathbf{\Pi b}\rangle - \langle \mathbf{a}, \mathbf{b}\rangle| \leq \epsilon \cdot \|\mathbf{a}\|_2\|\mathbf{b}\|_2.$$

Special case:

$$(1-\epsilon)\|\mathbf{a}\|_2^2 \leq \|\mathbf{\Pi a}\|_2 \leq (1+\epsilon)\|\mathbf{a}\|_2^2$$

Can be used to prove the famous Johnson-Lindenstrauss Lemma.

[Dasgupta, **Gupta**, 2003], [Indyk, Motwani 1998], [Arriage, Vempala 1999], [Achlioptas, 2001].

Lemma (Johnson-Lindenstrauss, 1984)

For any set of n data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ *there exists a linear map* $\boldsymbol{\Pi} : \mathbb{R}^d \to \mathbb{R}^m$ *where* $m = O\left(\frac{\log n}{\epsilon^2}\right)$ *such that for all i, j,*

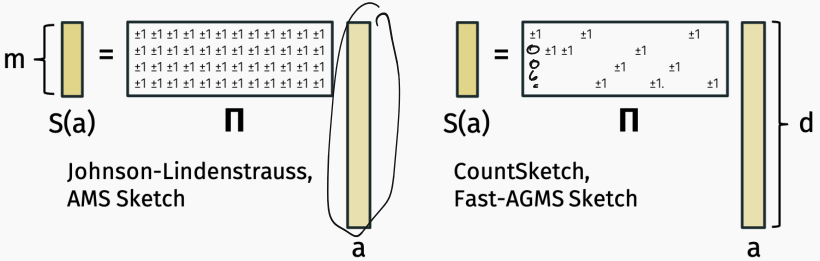
$$(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|\boldsymbol{\Pi}\mathbf{x}_i - \boldsymbol{\Pi}\mathbf{x}_j\|_2 \leq (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2.$$

Main idea: For any i,j

$$(1-\epsilon)\|x_i - x_j\|_2^2 \leq \|\Pi x_i - \Pi x_j\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2$$

with prob. δ if $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$

$\delta = O(1/n^2)$



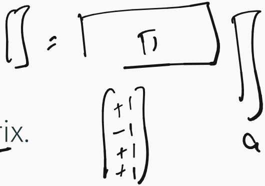
Naive cost of dense linear sketching is $O(d \cdot m)$.

This can be accelerated to $O(d)$ (linear) time without sacrificing accuracy by using an ultra-sparse random matrix. [Charikar, Chen, **Farach-Colton**, 2002]. Still achieve with $m = O(1/\delta\epsilon^2)$:

$$|\langle \mathbf{\Pi a}, \mathbf{\Pi b} \rangle - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \cdot \|\mathbf{a}\|_2 \|\mathbf{b}\|_2.$$

$\langle \Pi a, \Pi b \rangle$

Recent developments on JL sketches: Typically every entry of the compression is a real-value. In you computer, a double or a single precission float taking 32 or 64 bits. Can we reduce cost per dimension down to fewer bits?



(Method of Zandieh, Daliri, Han [AAAI, 2025][1]:)

- Let $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ be a random Gaussian matrix.
- Let $\mathcal{S}(\mathbf{a}) = \text{sign}(\mathbf{\Pi a})$, $\mathcal{S}(\mathbf{b}) = \mathbf{\Pi a}$
- Estimate inner produt via $\mathcal{F}_{QJL} = \frac{\sqrt{\pi}}{m} \cdot \|\mathbf{a}\|_2 \langle \mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}) \rangle$.

[1] Related to SimHash [Charikar, 2002], [Jacques, 2013], and a few other.

Theorem

Let $\mathcal{F}_Q JL$ be the inner product estimate returned by QJL for vectors **a**, **b**, *we have:*

$$\mathbb{E}[\mathcal{F}_{QJL}] = \langle \mathbf{a}, \mathbf{b} \rangle$$

$$\mathrm{Var}[\mathcal{F}_{QJL}] \leq \frac{1.6}{m} \|\mathbf{a}\|_2^2 \|\mathbf{b}\|_2^2.$$

$\frac{2}{m} \|a\|_2^2 \|b\|_2^2$

Sighly better variance than JL! But only using one bit per entry of $\mathcal{S}(\mathbf{a})$ (and need to store $\|\mathbf{a}\|_2$).

Lots of other recent work on the setting where only **a** needs to be compressed. E.g. RaBitQ skethes of [Gao, Long, 2024] addresses the setting where we want to target error $\leq \frac{1}{\sqrt{d}}$.

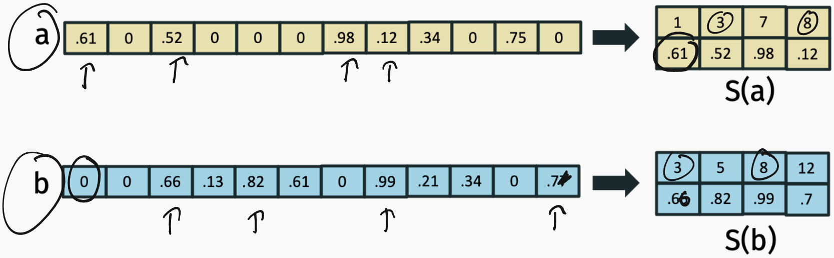
Goal: Present a completely different alternative to JL-style linear sketches that:

1. Is as simple as linear sketching to implement and analyze.
2. Can be applied in linear time (like CountSketch).
3. Matches theoretical bounds for linear sketching in the worst case, better for sparse vectors.
4. Typically beats linear sketching in experiments.

Appeared in Sampling Methods for Inner Product Sketching [Daliri, Freire, Musco, Santos, Zhang. VLDB 2024].

Similar to "End-Biased Sampling" [Estan, Naughton, 2006]. Also very related to the MinHash Sketch [Broder, 1997] and KMV Sketches [Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan, 2002], [Haas, Reinwald, Sismanis, Gemulla, 2007].

Sketch consists of subset of index/value pairs from **a** and **b**.

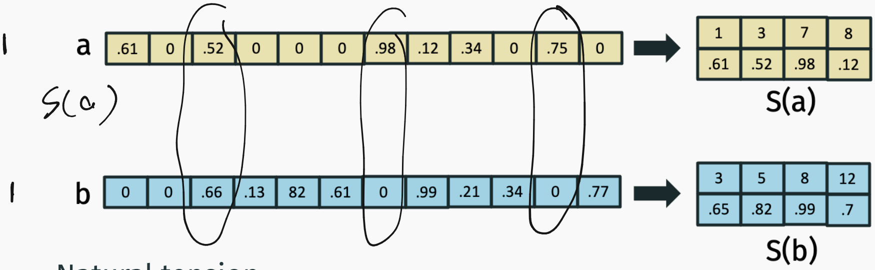


Let $\mathcal{T}$ be the set of indices common to $\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})$. Estimate:

$\mathcal{T} = \{3, 8\}$

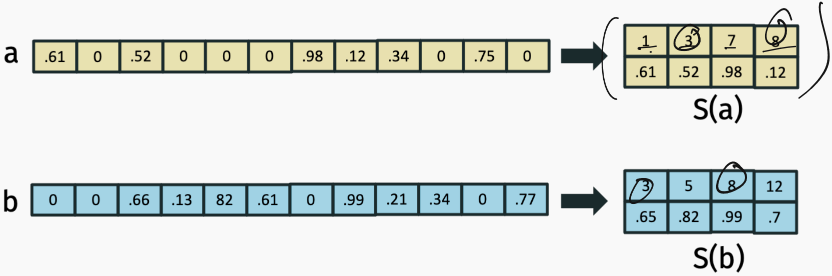
$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i \approx \sum_{i \in \mathcal{T}} w_i \cdot a_i b_i,$$

where $w_i > 1$ is an appropriately chosen weight.



Natural tension:

- Larger entries in **a** and **b** contribute more to $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i$. I.e., choice of indices should depend on magnitude of entries in vector being sketched.
- Want $\mathcal{S}(\mathbf{a})$ and $\mathcal{S}(\mathbf{b})$ to have many of the same indices. I.e., choice of indices should be coordinated between vectors.



Coordinate Random Sampling: Collect a sample from two different distributions while maximizing probability the samples are the same.

Threshold Sampling:

d is length of vector

- Set target sketch size m.
- Draw uniform random numbers $u_1, \ldots, u_d \sim [0, 1]$.
- For $i \in 1, \ldots, d$:
  - Add (i, a_i) to $\mathcal{S}(\mathbf{a})$ if $u_i \leq m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}$.
  - Add (i, b_i) to $\mathcal{S}(\mathbf{b})$ if $u_i \leq m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}$.

$\frac{m}{d}$

$1/d$

$u_{10} = .01$

Estimation:

- Let $\mathcal{T}$ be the set of indices common to $\mathcal{S}(\mathbf{a})$, $\mathcal{S}(\mathbf{b})$.
- Return $\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) = \sum_{i \in \mathcal{T}} \frac{1}{p_i} a_i b_i$, where $p_i = \min\left(1, m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}, m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}\right)$.

p_i = probability index i is sampled in both $S(a)$ and $S(b)$

THEORETICAL GUARANTEE

Theorem

Let $\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})$ *be sketches for* $\mathbf{a}, \mathbf{b}$ *obtained via Threshold Sampling and let* $\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))$ *be the corresponding estimate for* $\langle \mathbf{a}, \mathbf{b} \rangle$ *obtained from those sketches.*

We have that $\mathbb{E}[|\mathcal{S}(\mathbf{a})|] = m$, $\mathbb{E}[|\mathcal{S}(\mathbf{b})|] = m$, *and:*

$$\mathbb{E}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] = \langle \mathbf{a}, \mathbf{b} \rangle$$

$$\mathrm{Var}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] \leq \frac{2}{m} \max(\|\mathbf{a}_{\mathcal{I}}\|_2^2 \|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2 \|\mathbf{b}_{\mathcal{I}}\|_2^2)$$

$$\leq \frac{2}{m} \|a\|_2^2 \|b\|_2^2$$

Corollary: If $m = O(1/\epsilon^2)$, then with high probability, $|\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \cdot \max(\|\mathbf{a}_{\mathcal{I}}\|_2 \|\mathbf{b}\|_2, \|\mathbf{a}\|_2 \|\mathbf{b}_{\mathcal{I}}\|_2)$.

$\|a_{\mathcal{I}}\|_2^2 \le \|a\|_2^2 \qquad \|b_{\mathcal{I}}\|_2^2 \le \|b\|_2^2$

$$|\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) - \langle \mathbf{a}, \mathbf{b} \rangle| \le \epsilon \cdot \max(\|\mathbf{a}_{\mathcal{I}}\|_2 \|\mathbf{b}\|_2, \|\mathbf{a}\|_2 \|\mathbf{b}_{\mathcal{I}}\|_2).$$

a	.61	0	.52	0	0	0	.98	.12	.34	0	.75	0
b	0	0	.66	.13	82	.61	0	.99	.21	.34	0	.77
$a_{\mathcal{I}}$	0	0	.52	0	0	0	0	.12	.34	0	0	0
$b_{\mathcal{I}}$	0	0	.66	0	0	0	0	.99	.21	0	0	0

$\mathcal{I}$

So always tighter than the AMS/JL/CountSketch bound of $\epsilon\|\mathbf{a}\|_2\|\mathbf{b}\|_2$. Some implications for vector search involving sparse embeddings (e.g., SPLADE embeddings [Formal, Piwowarski, Clinchant, 2021]).

Threshold Sampling:

- Set target sketch size m.
- Draw uniform random numbers $u_1, \ldots, u_d \sim [0, 1]$.
- For $i \in 1, \ldots, d$:
  - Add (i, a_i) to $\mathcal{S}(\mathbf{a})$ if $u_i \leq m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}$.

Claim: $\mathbb{E}\left[|\mathcal{S}(\mathbf{a})|\right] \leq m$.

$$\mathbb{E}[|S(a)|] = \mathbb{E}\left[\sum_{i=1}^{d} \mathbb{1}((i,a_i) \text{ added to } S(a))\right]$$

$$= \sum_{i=1}^{d} \Pr[(i,a_i) \text{ added to } S(a)]$$

$$= \sum_{i=1}^{d} \min\left(1, m \cdot \frac{a_i^2}{\|a\|_2^2}\right) \leq \sum_{i=1}^{d} m \cdot \frac{a_i^2}{\|a\|_2^2} = m \cdot \frac{\sum_{i=1}^{n} a_i^2}{\|a\|_2^2}$$

$$= m.$$

Threshold Sampling:

- Set target sketch size m.
- Draw uniform random numbers $u_1, \ldots, u_d \sim [0, 1]$.
- For $i \in 1, \ldots, d$:
  - Add (i, a_i) to $\mathcal{S}(\mathbf{a})$ if $u_i \leq m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}$.

Claim: $\mathbb{E}\left[|\mathcal{S}(\mathbf{a})|\right] \leq m$.

$$\mathbb{E}\left[|\mathcal{S}(\mathbf{a})|\right] = \sum_{i=1}^{d} \Pr\left[u_i \leq m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}\right] = \sum_{i=1}^{d} \min\left(1, m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}\right)$$

$$\leq \frac{m}{\|\mathbf{a}\|_2^2} \sum_{i=1}^{d} a_i^2$$

$$= m.$$

Can also be shown to hold with high probability.

Threshold Sampling:

- Draw uniform random numbers $u_1, \ldots, u_d \sim [0,1]$.
- For $i \in 1, \ldots, d$:
  - Add (i, a_i) to $\mathcal{S}(\mathbf{a})$ if $u_i \leq m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}$.
  - Add (i, b_i) to $\mathcal{S}(\mathbf{b})$ if $u_i \leq m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}$.

Claim:

$\Pr[i \in \mathcal{S}(\mathbf{a}) \text{ and } i \in \mathcal{S}(\mathbf{b})] = p_i = \min\left(1, m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}, m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}\right)$.

$= \Pr[i \in T]$

Claim:
$\Pr[i \in \mathcal{S}(\mathbf{a}) \text{ and } i \in \mathcal{S}(\mathbf{b})] = p_i = \min\left(1, m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}, m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}\right).$

Estimation:

- Let $\mathcal{T}$ be the set of indices common to $\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})$.
- Return $\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) = \sum_{i \in \mathcal{T}} \frac{1}{p_i} a_i b_i$.

Claim: $\mathbb{E}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] = \langle \mathbf{a}, \mathbf{b} \rangle$.

$$\mathbb{E}\left[\sum_{i \in T} \frac{1}{p_i} a_i b_i\right] = \mathbb{E}\left[\sum_{i=1}^{d} \mathbb{1}[i \in T] \cdot \frac{1}{p_i} a_i b_i\right]$$

$$= \sum_{i=1}^{d} p_i \cdot \frac{1}{p_i} \cdot a_i b_i = \sum_{i=1}^{d} a_i b_i = \langle a, b \rangle$$

Claim:

$$\Pr\left[i \in \mathcal{S}(\mathbf{a}) \text{ and } i \in \mathcal{S}(\mathbf{b})\right] = p_i = \min\left(1, m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}, m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}\right).$$

Estimation:

- Let $\mathcal{T}$ be the set of indices common to $\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})$.
- Return $\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) = \sum_{i \in \mathcal{T}} \frac{1}{p_i} a_i b_i$.

Claim: $\mathbb{E}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] = \langle \mathbf{a}, \mathbf{b} \rangle$.

$$\mathbb{E}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] = \mathbb{E}\left[\sum_{i=1}^{d} \mathbb{1}[i \in \mathcal{T}] \cdot \frac{1}{p_i} a_i b_i\right] = \sum_{i=1}^{d} \Pr[i \in \mathcal{T}] \cdot \frac{1}{p_i} a_i b_i$$

$$= \sum_{i=1}^{d} p_i \cdot \frac{1}{p_i} a_i b_i = \sum_{i=1}^{d} a_i b_i = \langle \mathbf{a}, \mathbf{b} \rangle.$$

Claim: $\text{Var}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] \leq \frac{2}{m} \max(\|\mathbf{a}_{\mathcal{I}}\|_2^2 \|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2 \|\mathbf{b}_{\mathcal{I}}\|_2^2)$.

$$\text{Var}\left(\sum_{i=1}^{d} \mathbb{1}[i \in T] \frac{a_i b_i}{p_i} \right) = \sum_{i=1}^{d} \frac{a_i^2 b_i^2}{p_i^2} \text{Var}[\mathbb{1}[i \in T]]$$

$0, 1 \to p_i$

$p_i(1-p_i)$

$$\leq \sum_{\substack{i \in I \\ p_i \neq 1}} \frac{a_i^2 b_i^2}{p_i^2} p_i = \sum_{\substack{i \in I \\ p_i \neq 1}} \frac{a_i^2 b_i^2}{p_i} = \sum_{\substack{i \in I \\ p_i \neq 1}} \frac{a_i^2 b_i^2}{m \cdot \min(a_i^2/\|a\|_2^2, b_i^2/\|b\|_2^2)}$$

$$= \frac{1}{m} \sum_{\substack{i \in I \\ p_i \neq 1}} \|a\|_2^2 \|b\|_2^2 \cdot \frac{(a_i^2/\|a\|_2^2)(b_i^2/\|b\|_2^2)}{\min(\ldots\ldots)} \leq \frac{\|a\|_2^2 \|b\|_2^2}{m} \sum_{i \in I} \max\left(\frac{a_i^2}{\|a\|_2^2}, \frac{b_i^2}{\|b\|_2^2}\right)$$

$$\leq \frac{\|a\|_2^2 \|b\|_2^2}{m} \sum_{i \in I} \frac{a_i^2}{\|a\|_2^2} + \frac{b_i^2}{\|b\|_2^2} = \frac{\|a\|_2^2 \|b\|_2^2}{m} \left(\frac{\|a_I\|_2^2}{\|a\|_2^2} + \frac{\|b_I\|_2^2}{\|b\|_2^2} \right)$$

$$= \frac{1}{m} \left(\|a_I\|_2^2 \|b\|_2^2 + \|b_I\|_2^2 \|a\|_2^2 \right)$$

Claim: $\text{Var}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] \leq \frac{2}{m} \max(\|\mathbf{a}_\mathcal{I}\|_2^2\|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2\|\mathbf{b}_\mathcal{I}\|_2^2)$.

(**Takeaway:** Just elementary calculations.)

$$\mathrm{Var}[\mathcal{F}(\mathcal{S}(\mathbf{a}),\mathcal{S}(\mathbf{b}))] = \sum_{i\in\mathcal{I}} \mathrm{Var}\left[\mathbb{1}[i\in\mathcal{T}]\cdot\frac{\mathbf{a}_i\mathbf{b}_i}{p_i}\right] = \sum_{i\in\mathcal{I}}\frac{(\mathbf{a}_i\mathbf{b}_i)^2}{p_i^2}\mathrm{Var}[\mathbb{1}[i\in\mathcal{T}]]$$

$$\leq \sum_{i\in\mathcal{I},p_i\neq 1}\frac{(\mathbf{a}_i\mathbf{b}_i)^2}{p_i}.$$

$$\sum_{i\in\mathcal{I},p_i\neq 1}\frac{(\mathbf{a}_i\mathbf{b}_i)^2}{p_i} \leq \sum_{i\in\mathcal{I},p_i\neq 1}\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2\frac{(\mathbf{a}_i^2/\|\mathbf{a}\|_2^2)(\mathbf{b}_i^2/\|\mathbf{b}\|_2^2)}{m\cdot\min(\mathbf{a}_i^2/\|\mathbf{a}\|_2^2,\mathbf{b}_i^2/\|\mathbf{b}\|_2^2)}$$

$$= \sum_{i\in\mathcal{I},p_i\neq 1}\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2\frac{\max(\mathbf{a}_i^2/\|\mathbf{a}\|_2^2,\mathbf{b}_i^2/\|\mathbf{b}\|_2^2)}{m}$$

$$\leq \frac{\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2}{m}\sum_{i\in\mathcal{I}}\frac{\mathbf{a}_i^2}{\|\mathbf{a}\|_2^2}+\frac{\mathbf{b}_i^2}{\|\mathbf{b}\|_2^2}$$

$$= \frac{1}{m}\left(\|\mathbf{a}_{\mathcal{I}}\|_2^2\|\mathbf{b}\|_2^2+\|\mathbf{a}\|_2^2\|\mathbf{b}_{\mathcal{I}}\|_2^2\right).$$

THEORETICAL GUARANTEE

Theorem

Let $\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})$ be sketches for $\mathbf{a}, \mathbf{b}$ obtained via Threshold Sampling and let $\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))$ be the corresponding estimate for $\langle \mathbf{a}, \mathbf{b} \rangle$ obtained from those sketches.

We have that $\mathbb{E}[|\mathcal{S}(\mathbf{a})|] = m$, $\mathbb{E}[|\mathcal{S}(\mathbf{b})|] = m$, and:

$$\mathbb{E}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] = \langle \mathbf{a}, \mathbf{b} \rangle$$

$$\mathrm{Var}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] \leq \frac{2}{m} \max(\|\mathbf{a}_{\mathcal{I}}\|_2^2 \|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2 \|\mathbf{b}_{\mathcal{I}}\|_2^2)$$

One annoying feature of Threshold Sampling: Sketch size is random. Ideally it would be exactly equal to m.

FIXED SKETCH SIZE

Priority Sampling[2]

- Set sketch size m.
- Draw uniform random numbers $u_1, \ldots, u_d \sim [0, 1]$.
- Let $i_1, \ldots, i_m$ be the indices corresponding to the m smallest values of u_i/a_i^2.
- Add $(i_1, a_{i_1}), (i_2, a_{i_2}), \ldots, (i_m, a_{i_m})$ to $\mathcal{S}(\mathbf{a})$.

Bound: $\mathrm{Var}[\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}))] \leq \frac{2}{m-1} \max(\|\mathbf{a}_{\mathcal{I}}\|_2^2 \|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2 \|\mathbf{b}_{\mathcal{I}}\|_2^2)$

Almost identical to the bound given by Threshold Sampling.

[2] [Duffield, Lund, Thorup, 2004], [Ohlsson, 1998].

PARALLELIZING LARGE LANGUAGE MODELS

Modern large language models are inherently <u>sequential</u>.

NYU is a private research university in New York City.

Even if cost per word can come down, limits the speed at which text can be generated.

Introduced concurrently by [Leviathan, Kalman, Matias, 2023] at Google and [Chen, Borgeaud, Irving, Lespiau, Sifre, Jumper, 2023] at Google Deepmind

Key idea: Use small model to "draft" a response, and verify with multiple instances of a large model.

Draft: NYU is a private research university in the city of New York .

Desired Output: NYU is a private research university in New York City.

NYU is a private research **university**

NYU is a private research university **in**

NYU is a private research university in **New**

NYU is a private research university in the **city**

....

Above, we got three tokens from one batch of parallel invocations of the large model.

Draft: NYU is a private research university in New York City .

Desired Output: NYU is a private research university in New York City.

NYU is a private research university in New **York**

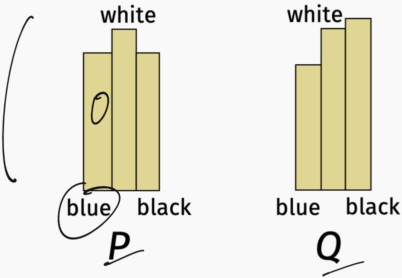
NYU is a private research university in New York **City**

NYU is a private research university in New York City **.**

....

Obvious issue: Even if then next token distribution for the drafter model, $\mathcal{P}$, and the product model, $\mathcal{Q}$ are very similar, it could be unlikely for the draft to be correct.

The dress is ___



If $a \sim P$ and $b \sim Q$, $\Pr[a = b] \approx$ 1/3

Solution: Coordinate the sampling!

Definition (Coupling)

Let $\mathcal{P}$ and $\mathcal{Q}$ be distributions over $\{1, \ldots, n\}$. A coupling between $\mathcal{P}$ and $\mathcal{Q}$ is any distribution over pairs $(a, b) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ such that a's marginal distribution is $\mathcal{P}$ and b's marginal distribution is $\mathcal{Q}$.

Goal: Efficiently sample from a coupling $\mathcal{C}$ between the small and large model distributions which maximizes

$$\Pr[a = b].$$

Simple case: If $\mathcal{P} = \mathcal{Q}$, can find a $\mathcal{C}$ for which $\Pr[a = b]$. Do you see how?

Draw $a \sim P$. set $b = a$.

Definition

The **total variation distance**, $D_{TV}(\mathcal{P}, \mathcal{Q})$, between two distributions equals:

$$D_{TV}(\mathcal{P}, \mathcal{Q}) = 1 - \max_{\text{couplings } \mathcal{C}} \left[\Pr_{(a,b) \sim \mathcal{C}} [a = b]. \right]$$

Claim: Let $\mathcal{P}$ and $\mathcal{Q}$ be discrete disributions over $\{1, \ldots, n\}$ represented by length n probabiity vectors $\mathbf{p}, \mathbf{q} \in [0, 1]^n$.

$$D_{TV}(\mathcal{P}, \mathcal{Q}) = 1 - \sum_{i=1}^{n} \min(p_i, q_i).$$

$p_1, \ldots p_n$
$q_1 \ldots q_n$

Claim: Let $\mathcal{P}$ and $\mathcal{Q}$ be discrete disributions over $\{1, \ldots, n\}$ represented by length n probabiity vectors $\mathbf{p}, \mathbf{q} \in [0,1]^n$.

$$D_{TV}(\mathcal{P}, \mathcal{Q}) \geq 1 - \sum_{i=1}^{n} \min(p_i, q_i).$$

→ no coupling can achieve

$\Pr[a=b] > \sum_{i=1}^{n} \min(p_i, q_i)$

Proof: Under any coupling $\mathcal{C}$, $\Pr[a = b] =$

$\sum_{i=1}^{n} P[a=i \text{ and } b=i] \leq \sum_{i=1}^{n} \min(\Pr[a=i], \Pr[b=i])$

$P[a=i] \cdot \Pr[b=i \mid a=i]$

$\leq P[a=i]$

$\leq P[b=i]$

$\Pr[b=i] \Pr[a=i \mid b=i]$

$\leq \Pr[b=i]$

OPTIMAL COUPLING

There is a simple procedure that achieves this bound in the speculative decoding setting.

Drafter:

- Sample $a \sim \mathcal{P}$. Sends both a and $\mathbf{p}$ to FullModel.

If $q_a > p_a$

FullModel:

- Await $(a, \mathbf{p})$ from Drafter.
- With probability $\min(1, q_a/p_a)$ return $b = a$.
- Otherwise, sample b from $\mathcal{Q}' = \{q'_1, \ldots, q'_n\}$, where:

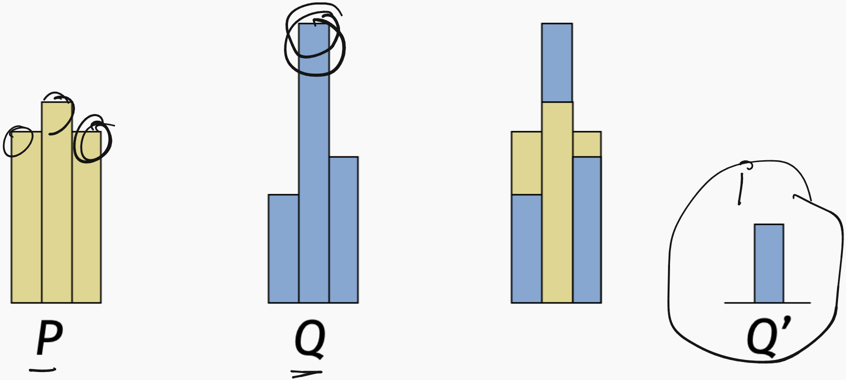
$$q'_i = \frac{\max(0, q_i - p_i)}{\sum_{i=1}^{n} \max(0, q_j - p_j)}$$

Drafter:

- Sample $a \sim \mathcal{P}$. Sends both a and $\mathbf{p}$ to FullModel.

FullModel:

- With probability $\min(1, q_a/p_a)$ return $b = a$.
- Otherwise, set $b = i$ w.p. $q_i' = \frac{\max(0, q_i - p_i)}{\sum_{j=1}^{n} \max(0, q_j - p_j)}$.



OPTIMAL COUPLING

Drafter:

- Sample $a \sim \mathcal{P}$. Sends both a and $\mathbf{p}$ to FullModel.

$= \sum_{j=1}^{n} -\min(0, q_j - p_j)$

$= \sum_{j=1}^{n} p_j - \min(q_j, p_j)$

FullModel:

Step 1:
- With probability $\min(1, q_a/p_a)$ return $b = a$.

Step 2:
- Otherwise, set $b = i$ w.p. $q_i' = \frac{\max(0, q_i - p_i)}{\sum_{j=1}^{n} \max(0, q_j - p_j)}$.

Clear for a.

Claim: Procedure samples from a coupling. I.e., $a \sim \mathcal{P}$, $b \sim \mathcal{Q}$.

~~Cases to check: $q_a \geq p_a$~~

Case 1: $q_i \leq p_i$

We never set $b = i$ in step 2 since $\max(0, q_i - p_i) = 0$.

Chance $b = i$ in first step

$= \Pr[a = i] \cdot P[b = i \mid a = i]$

$= p_i \cdot q_i/p_i = q_i$ as desired.

Case 2: $q_i > p_i$

Probability b gets set to i in first step is $\boxed{p_i}$.

Probability b gets set to i in 2nd step is:

$\frac{\max(0, q_i - p_i)}{\sum_{j=1}^{n} \max(0, q_j - p_j)} \cdot \Pr[\text{we get to step 2}]$

$\Pr[\text{we get to step 2}] = \sum_{j=1}^{n} p_j \left(1 - \min(1, q_j/p_j)\right) = \sum_{j=1}^{n} p_j - \min(p_j, q_j) = \sum_{j=1}^{n} \max(0, q_j - p_j)$

$\text{total} = \frac{\max(0, q_i - p_i)}{\sum \max(0, q_j - p_j)} \cdot \sum \max(q_j - p_j) = \boxed{q_i - p_i}$

$q_i - p_i + p_i = q_i$ as desired.

Drafter:

- Sample $a \sim \mathcal{P}$. Sends both a and $\mathbf{p}$ to FullModel.

FullModel:

Step 1: - With probability $\min(1, q_a/p_a)$ return $b = a$.

Step 2: - Otherwise, set $b = i$ w.p. $q_i' = \frac{\max(0, q_i - p_i)}{\sum_{i=1}^{n} \max(0, q_j - p_j)}$.

Claim: $\Pr[a = b] = \sum_{i=1}^{n} \min(p_i, q_i)$.

If $q_i > p_i$, then b always set to a if $a = i$. So $\Pr[a = i$ and $b = i] = p_i$.

Otherwise, if $q_i \le p_i$, b never set to i in step 2 (since $\max(0, q_i - p_i) = 0$).

So only set to i in first step. $P[a = i$ and $b = i] = p_i \cdot q_i/p_i = q_i$.

Overall $\Pr[a = b] = \sum_{i=1}^{n}$

Earlier showed that $D_{TV}(\mathcal{P}, \mathcal{Q}) \geq 1 - \sum_{i=1}^{n} \min(p_i, q_i)$, so this is optimal.

FullModel:

- With probability $\min(1, q_a/p_a)$ return $b = a$.
- Otherwise, set $b = i$ w.p. $q_i' = \frac{\max(0, q_i - p_i)}{\sum_{i=1}^{n} \max(0, q_j - p_j)}$.

Seemingly small but annoying issue: The output of the FullModel is always sampled from $\mathcal{Q}$, but the exact value sampled depends on the Drafter distribution $\mathcal{P}$.

- Cannot immediately verify that adding speculative decoding did not change the model distribution.
- Model output is not deterministic from the user's point of view given a fixed random seed.

Key idea: Basic protocol requires communication between the Drafter and FullModel. Try to couple samples <u>without communication at all</u> but just using shared randomness.

"Coupling without Communication and Drafter-Invariant Speculative Decoding" [**Daliri**, Musco, Suresh, ISIT 2025].

Bsically the same idea appeared in:

- Anari, Gao, Rubinstein, STOC 2024
- Liu, Yin, STOC 2022
- Bavarian, Ghazi, Haramaty, Kamath, Rivest, Sudan, 2020.

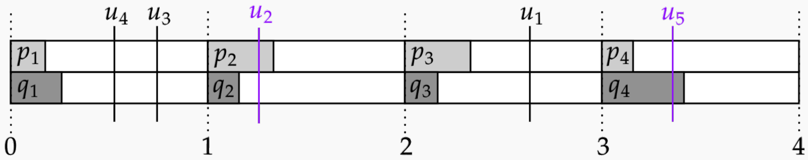
Fix public random variables $u_1, u_2, \ldots \sim \text{Unif}[0, n]$.

Drafter:

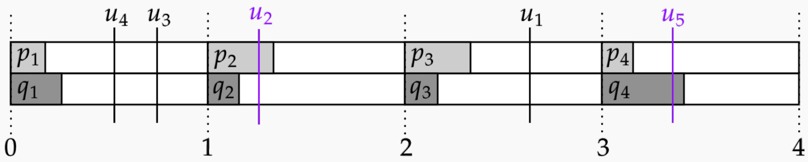
- For $k = 1, 2, \ldots,$
  - If $k \in [j-1, j-1+p_j]$ for some j, return $a = j$.

FullModel:

- For $k = 1, 2, \ldots,$
  - If $k \in [j-1, j-1+q_j]$ for some j, return $b = j$.



Claim: $\Pr[a = b] \geq \frac{\sum_{i=1}^{n} \min(p_i, q_i)}{\sum_{i=1}^{n} \max(p_i, q_i)}$



Optimal Coupling:

$$\Pr[a = b] = 1 - D_{TV}(\mathcal{P}, \mathcal{Q})$$

. Communication-Free Coupling:

$$\Pr[a = b] \geq \frac{\sum_{i=1}^{n} \min(p_i, q_i)}{\sum_{i=1}^{n} \max(p_i, q_i)} = \frac{1 - D_{TV}(\mathcal{P}, \mathcal{Q})}{1 + D_{TV}(\mathcal{P}, \mathcal{Q})}$$

.

Takeaway: Pay very little for drafter-invariance!

Possible to show that this is optimal in some sense. No communication-free protocol can achieve $\Pr[a = b] > \frac{1 - D_{TV}(\mathcal{P}, \mathcal{Q})}{1 + D_{TV}(\mathcal{P}, \mathcal{Q})}$ for all distributions simultaneously [Bavarian, Ghazi, Haramaty, Kamath, Rivest, Sudan, 2020].

Fix public random variables $u_1, u_2, \ldots \sim \text{Unif}[0, 1]$.

Drafter:

- Return $a = \arg\min_{i \in \{1,\ldots,n\}} \frac{-\ln(u_i)}{p_i}$.

FullModel:

- Return $b = \arg\min_{i \in \{1,\ldots,n\}} \frac{-\ln(u_i)}{q_i}$.

This is already how samples are typically obtained! In particular, standard to use the "Gumbel Max Trick":

$$b = \arg\max_{i \in \{1,\ldots,n\}} \left[-\ln(\ln(1/u_i)) + \ln(q_i)\right].$$

Theorem (pareto improvement)

For any two distributions $\mathcal{P}, \mathcal{Q}$,

$$\Pr_{(a,b)\sim Gumbel}[a = b] \geq \Pr_{(a,b)\sim MinHash}[a = b],$$

and there exist distributions where inequality is strict.

Question one group is studying for the project: Is Gumbel pareto optimal?