Positional LSH: Binary Block Matrix Approximation for Attention with Linear Biases

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Positional encoding in transformers is commonly implemented through positional embeddings, attention masks, or bias terms, but formal connections between these mechanisms remain limited. We study attention with positional bias through the lens of locality-sensitive hashing (LSH), focusing on Attention with Linear Biases (ALiBi). We show that the ALiBi bias matrix is the expectation of contiguous block-diagonal binary masks induced by a "positional LSH" scheme. The empirical mean of masks sampled from this scheme yields spectral norm and max-norm approximation guarantees with bounded block sizes with high probability. This structural theorem implies a uniform approximation theorem for ALiBi-biased attention: with high probability over the sampled masks, the approximate attention output is accurate simultaneously for all query-key-value inputs and can be computed in near-linear time in the context length, reducing long-context ALiBi to a collection of randomized short-context regular (positionally unbiased) attention operations. Conceptually, this connects positional bias, masks, and positional embeddings in a single formal framework and suggests an approach to efficient ALiBi-biased attention. Experiments on large language models validate our theoretical findings.

1. Introduction

Although the Transformer architecture has remained remarkably intact since its introduction in [36], even through widespread adoption in large language models, encoding token positions stands out as a still evolving aspect. The challenge is to encode how the structural positions of tokens in the input inform the way they attend to each other. A closely intertwined challenge is long-context attention: attending to all tokens in a long context becomes computationally difficult or infeasible, whereas limiting attention to structurally local windows often degrades long-range reasoning performance.

Three leading paradigms in positional encoding, often used in conjunction, are as follows:

- *Positional embedding*, where each position in the input is endowed with an embedding vector designed to encode its location – e.g., sinusoidal [36] or rotary position embeddings (RoPE) [33];
- *Attention with masks*, which allows tokens to attend to some tokens but not to others through a fixed binary mask – e.g., sliding-window attention [4, 20];
- *Attention with bias*, which modifies attention weights a posteriori according to token positions – e.g., Attention with Linear Biases (ALiBi) [28].

In this work, we undertake a theoretical study of attention with positional bias, focusing on ALiBi, a prominent method that has emerged from empirical research [28] and has been widely

implemented and adopted in popular models [10, 38]. We study it through the lens of *locality sensitive hashing* (LSH), a key paradigm in the theory of efficient algorithms. This leads us to uncover fundamental connections between ALiBi and LSH with implications to long-context attention.

Our main structural result, Theorem 1, shows that the ALiBi bias matrix can be approximated arbitrarily well by a linear combination of binary contiguous block-diagonal matrices. Conceptually, this draws a formal connection between the three main paradigms in positional encoding: bias, masks and embeddings. Namely, it suggests that attention with bias can be approximated by attention with masks, where the masks are induced by locality-sensitive hash collisions of positional embeddings. In the case of ALiBi, this connection holds in a strong formal sense.

Our main algorithmic result, Theorem 2, is an efficient approximation theorem for ALiBi-biased attention over long contexts, which is based on our structural result. It suggests an approach to long-context ALiBi through a reduction to a collection of randomized short-context regular (positionally unbiased) attention operations, which correspond to the contiguous binary blocks in Theorem 1. This yields a near-linear time algorithm for ALiBi with a provable approximation guarantee, addressing the computational bottleneck in long-context ALiBi, a challenge called out in [28].

Conceptually, our contribution goes beyond applying classical LSH to positions in attention: it is in reinterpreting positional bias through hashing on positions, drawing a formal connection between positional embeddings, binary attention masks, and bias terms. Technically, while kernel approximation via LSH is standard, we introduce novel technical tools in order to prove matrix convergence in the spectral norm and high-probability block size control. These results are specialized to attention with ALiBi and do not follow from generic LSH machinery alone. Our proofs draw on matrix concentration theory, Fourier analysis of Toeplitz matrices, and sub-gamma tail bounds.

We validate our theoretical findings through experiments on public large language models.

2. Our Results in Detail

We fix notation to be used throughout. We use "$\odot$" for element-wise matrix product. We denote the following norms for a matrix M: $\|M\|$ for the spectral norm; $\|M\|_{\max} = \max_{ij} |M_{ij}|$; and $\|M\|_{2,\infty} = \max_i \sqrt{\sum_j M_{ij}^2}$. We let $D^{[M]}$ denote the diagonal row-sum matrix $D^{[M]}_{ii} = \sum_{j=1}^{n} M_{ij}$. Finally, we will use $J \in \mathbb{R}^{n\times n}$ for the fixed lower-triangular all-1 matrix, $J_{ij} = \mathbf{1}\{j \le i\}$.

Attention. Let n be the attention context length. Let $Q, K \in \mathbb{R}^{n\times d}$ be the key and query matrices with rows $\{q_i\}, \{k_j\}$ respectively. The unnormalized dot-product attention matrix $A \in \mathbb{R}^{n\times n}$ has entries $A_{ij} = \exp(q_i^T k_j/\sqrt{d})$. The row-wise normalized non-causal attention matrix is $P = (D^{[A]})^{-1}A$. In the causal attention case we instead have $P = (D^{[A\odot J]})^{-1}(A \odot J)$. Given a value matrix $V \in \mathbb{R}^{n\times d'}$, the output of the attention operation without positional bias is $T = PV$.

ALiBi. The ALiBi matrix $L^\star$ is defined in [28] as $L^\star_{ij} = e^{-|i-j|/\sigma}$, where $\sigma > 0$ is a bandwidth hyperparameter which is fixed per head and varies between heads. The unnormalized ALiBi-biased attention weights are $A^\star = A \odot L^\star$ in the non-causal case and $A^\star = A \odot J \odot L^\star$ in the causal case. In either case, the normalized ALiBi-biased attention matrix is $P^\star = (D^{[A^\star]})^{-1}A^\star$, and the goal is to compute the output $T^\star = P^\star V$.

The following theorem is our main structural result: an approximation of the ALiBi matrix by random contiguous block-diagonal binary matrices.

Theorem 1 *There exists a distribution $\mathcal{M}$ over contiguous block-diagonal binary $n \times n$ matrices, where each diagonal block is a square all-1 matrix, such that $\mathbb{E}_{M\sim\mathcal{M}}[M] = L^\star$. Let $M_1, \ldots, M_s \sim$*

$\mathcal{M}$ be i.i.d. samples and let $\widetilde{M} = \frac{1}{s}\sum_{i=1}^{s} M_i$ be their empirical mean. Then we have

$$\textit{Spectral norm expectation:} \quad \mathbb{E}\|L^\star - \widetilde{M}\| \le C\Psi_\sigma \left(\sqrt{\frac{\log n}{s}} + \frac{\log n}{s}\right) \tag{1}$$

$$\textit{Spectral norm concentration:} \quad \forall \varepsilon > 0, \;\; \Pr[\|L^\star - \widetilde{M}\| \ge \varepsilon] \le n \cdot e^{-Cs\min\left\{\frac{\varepsilon^2}{\Psi_\sigma^2}, \frac{\varepsilon}{\Psi_\sigma}\right\}} \tag{2}$$

$$\textit{Max-norm concentration:} \quad \forall \varepsilon > 0, \quad \Pr[\|L^\star - \widetilde{M}\|_{\max} \ge \varepsilon] \le 2n^2 e^{-2s\varepsilon^2} \tag{3}$$

where $\Psi_\sigma = 1 + \sigma + \frac{e^{1/\sigma}+1}{e^{1/\sigma}-1}$ and $C > 0$ is a universal constant. Furthermore, let $b_{\max}$ be the maximum size of a diagonal block in any of the samples $M_1, \ldots, M_s$, then, for all $\delta \in (0, 1/e)$,

$$\Pr[b_{\max} > 1 + \sigma(3\ln(s/\delta) + 2)] < \delta. \tag{4}$$

Structurally, the theorem states that the ALiBi matrix is a convex combination of contiguous non-overlapping local-window attention patterns, each positionally unbiased. Approximation-wise, it implies that with high probability, $s \approx \varepsilon^{-2}\sigma^2 \log n$ samples from $\mathcal{M}$ suffice to approximate ALiBi using diagonal blocks of size at most $b_{\max} \approx \sigma \log\log n$ (suppressing logarithmic factors in $\varepsilon^{-1}, \delta^{-1}, \sigma$ for simplicity). Regarding the values of σ and Ψ_σ, we recall that Press et al. [28] set σ as a geometric series from $2^{8/n_h}$ to 2^8 across n_h heads. Under this setting, Ψ_σ satisfies $2 < \Psi_\sigma < 3\sigma + 2$.

Algorithmically, Theorem 1 suggests that ALiBi-biased attention can be approximated efficiently over long contexts by a reduction to a collection of small local attention windows. Our next theorem formalizes this intuition and yields an approximation guarantee for the output attention weights.

Theorem 2 (uniform ALiBi-biased attention approximation) *Let $\delta \in (0, 1/e)$ and $s \ge \log(n/\delta)$. Let $\widetilde{M} = \frac{1}{s}\sum_{i=1}^{s} M_i$ be the empirical mean of s samples from $\mathcal{M}$ as in Theorem 1. Given an attention instance (K, Q, V), let $A, A^\star, P$ be the corresponding attention weights as above (in either the causal or non-causal variant), and denote:*

$$\Delta_P = C\min\{\Psi_\sigma \|P\|_{2,\infty}, 1\}\sqrt{\frac{\log(n/\delta)}{s}} \quad \textit{and} \quad \beta_P^\star = \min_i \sum_{j=1}^{n} P_{ij} L_{ij}^\star$$

where $C > 0$ is a universal constant. Then, with probability $1-\delta$ over the draw of $\widetilde{M}$, the following hold simultaneously for all input attention instances:

1. *The approximate ALiBi-biased attention output $\widetilde{T}^\star$, obtained by replacing $L^\star$ with $\widetilde{M}$, can be computed in time $O(nd(\sigma s \log(s/\delta) + 1))$.*
2. *Whenever $\Delta_P < \beta_P^\star$, it holds that $\|T^\star - \widetilde{T}^\star\|_{\max} \le \frac{2\Delta_P}{\beta_P^\star - \Delta_P}\|V\|_{\max}$.*

To unpack the statement, we note that the critical quantity $\beta_P^\star$ is the ratio between the minimum row denominator in $A^\star$ and A, that is, in the attention matrix before and after the ALiBi bias. It governs how sensitive the instance is to perturbations in the bias matrix. The term Δ_P which governs the error decays like $1/\sqrt{s}$ with the number of samples s, and may further improve with smaller $\|P\|_{2,\infty}$, which occurs in instances where attention weights are more balanced. Importantly, the convergence in matrix norms in Theorem 1 (eqs. (2) and (3)) is what allows the approximation in Theorem 2 to hold uniformly — that is, simultaneously for all attention inputs with high probability.

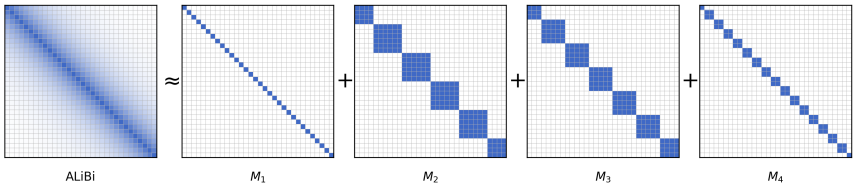

Figure 1: ALiBi approximation with binary block-diagonal matrices by positional LSH.

3. Positional LSH

We give a high-level overview of our positional LSH framework. Full proofs are in Appendix B.

Let U be a space of objects and $\ker : U \times U \to [0, 1]$ a similarity map (kernel) over U. An LSH scheme for $\ker$ is a distribution $\mathcal{H}$ over hash functions operating on U such that for all $u, u' \in U$

$$\ker(u, u') = \Pr_{h \sim \mathcal{H}}[h(u) = h(u')]. \tag{5}$$

This notion was defined in classical works [7, 18] and has been widely influential, including recent applications to efficient attention computation over long contexts (see Section A). While prior work has applied LSH to the token embeddings of the keys and queries in the attention operation, here we will apply it exclusively to their positions in the sequence. Hence, we call this *positional LSH*.

Let L be a positional bias matrix that biases the original attention weights A as $A \odot L$. Let $\{u_i\}_{i=1}^n$ be positional embeddings for the token positions in a context of length n. That is, u_i is an object that encodes the location i within the context. Suppose that L has the form $L_{ij} = \ker(u_i, u_j)$. Furthermore, suppose that $\ker$ admits an LSH scheme $\mathcal{H}$. We may define a distribution $\mathcal{M}$ over $n \times n$ matrices where a sampled hash function $h \sim \mathcal{H}$ induces the random binary matrix $M^{(h)}$ given by

$$M^{(h)}_{ij} = \begin{cases} 1 & \text{if } h(u_i) = h(u_j) \\ 0 & \text{if } h(u_i) \neq h(u_j). \end{cases}$$

The LSH property (5) immediately implies $L = \mathbb{E}_{h \sim \mathcal{H}}[M^{(h)}]$. This naturally suggests that L can be approximated by the empirical average of i.i.d. samples of $M^{(h)}$. See Figure 1 for illustration. This raises the question of the convergence rate to the mean. Entry-wise scalar convergence follows immediately from standard Chernoff-Hoeffding concentration, and this already yields the max-norm bound eq. (3) in Theorem 1. Matrix convergence in the spectral norm (eqs. (1) and (2) in Theorem 1) is considerably more intricate and may not hold in general (see Section C). In Section B we prove it for ALiBi with a specific positional LSH scheme.

4. Experimental Validation

Convergence in Matrix Norms. We validate our theoretical results on convergence in matrix norms in Theorems 1 and 2 directly on attention operations on the pre-trained Mistral-7B model [21] with 30 text segments from the Wikitext-103 dataset [27] as inputs, with context lengths of 4k-4.5k. We retrieve the Q,K,V matrices for each model per input, layer and attention head. We compute $\|L^\star - \widetilde{M}\|$, $\|L^\star - \widetilde{M}\|_{\max}$ and $\|T^\star - \widetilde{T}^\star\|_{\max}$, for varying sample sizes.

The results are reported in Figure 2. Results for additional σ values is displayed in Appendix D. The results show that norm errors decay towards zero as the sample size increases.

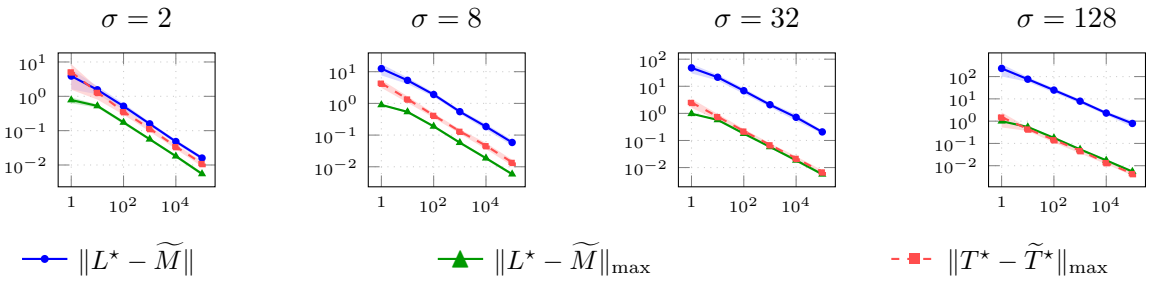

Figure 2: $\|L^\star - \widetilde{M}\|$, $\|L^\star - \widetilde{M}\|_{\max}$ and $\|T^\star - \widetilde{T}^\star\|_{\max}$ as functions of the sample size s, for four attention heads (with σ values: $\sigma \in \{2, 8, 32, 128\}$), for Mistral-7B. Axes are in log scale.

Table 1: Length extrapolation results on Wikitext-103 for Qwen3-0.6B (trained from scratch)

Context length:	8k (training)		16k (extrapolation)	
Method	PPL	Acc	PPL	Acc
Original model	20.213±0.088	44.796±0.182	21.316±0.058	43.652±0.126
ALiBi	18.920±0.009	45.249±0.035	18.753±0.021	45.351±0.020
Positional LSH, $s = 1$	19.217±0.059	45.051±0.044	19.278±0.027	44.985±0.029
Positional LSH, $s = 10$	19.073±0.054	45.125±0.025	19.043±0.070	45.179±0.021
Positional LSH, $s = 20$	19.050±0.033	45.146±0.032	18.961±0.033	45.192±0.046
Fixed blocks	19.788±0.054	44.725±0.058	19.784±0.041	44.658±0.067

Table 2: Length extrapolation results on Wikitext-103 for Mistral-7B (fine-tuned)

Context length:	4k (training)		8k (extrapolation)		16k (extrapolation)	
Method	PPL	Acc	PPL	Acc	PPL	Acc
Original model	6.347±0.001	58.371±0.016	6.273±0.002	58.523±0.016	6.270±0.004	58.501±0.039
ALiBi	6.498±0.008	57.924±0.142	6.455±0.022	58.086±0.134	6.461±0.021	57.960±0.207
Positional LSH, $s = 1$	6.957±0.009	56.807±0.031	6.941±0.015	56.926±0.025	6.923±0.011	56.975±0.027
Positional LSH, $s = 5$	6.725±0.013	57.186±0.051	6.711±0.019	57.325±0.038	6.721±0.012	57.235±0.058
Positional LSH, $s = 10$	6.698±0.010	57.280±0.026	6.683±0.011	57.389±0.028	6.705±0.034	57.284±0.021
Fixed blocks	7.352±0.014	54.919±0.022	7.327±0.017	54.944±0.019	7.378±0.015	54.930±0.013

Model Training Experiments. We evaluate the effects of our approximation method on downstream model performance. We implemented a prototype of our positional LSH method for ALiBi in FlashAttention-2 [12]. Our implementation is not optimized from an engineering perspective (see limitations in Section E); the goal is to gain empirical insight into the approximation quality of our method rather than show efficiency gains. We train on Wikitext-103 two models: Qwen3-0.6B [34] for pre-training from scratch, and Mistral-7B [21] for fine-tuning from its public pre-trained checkpoint. We evaluate perplexity (PPL) and next-token prediction accuracy (Acc), both on the original context length used in training and under context length extrapolation (where the model is evaluated on longer context lengths than it was trained on). As a baseline, we also include a "fixed blocks" baseline with a single block-diagonal mask matrix with a fixed block size of 512.

The results are displayed in Tables 1 and 2. For both models, as the number of samples in Positional LSH increases, performance improves, approaching that of ALiBi.

References

[1] Alex Andoni and Piotr Indyk. Dimension reduction in kernel spaces from locality-sensitive hashing. 2009.

[2] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *Advances in neural information processing systems*, 28, 2015.

[3] Arturs Backurs, Piotr Indyk, and Tal Wagner. Space and time efficient kernel density estimation in high dimensions. *Advances in neural information processing systems*, 32, 2019.

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[5] Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043. IEEE, 2017.

[6] Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 172–183. IEEE, 2020.

[7] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, 2002.

[8] Zhuoming Chen, Ranajoy Sadhukhan, Zihao Ye, Yang Zhou, Jianyu Zhang, Niklas Nolte, Yuandong Tian, Matthijs Douze, Leon Bottou, Zhihao Jia, et al. Magicpig: Lsh sampling for efficient llm generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022.

[10] Sanghun Cho, Ofir Press, and Tri Dao. ALiBi-FlashAttention: Speeding up ALiBi by 3–5x with a hardware-efficient implementation. Princeton Language and Intelligence Blog, Jan 2024. URL https://pli.princeton.edu/blog/2024/alibi-flashattention-speeding-alibi-3-5x-hardware-efficient-implementation. Accessed: 2026-04-27.

[11] Benjamin Coleman and Anshumali Shrivastava. Sub-linear race sketches for approximate kernel density estimation on streaming data. In *Proceedings of The Web Conference 2020*, pages 1739–1749, 2020.

[12] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Aditya Desai, Shuo Yang, Alejandro Cuadron, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. Hashattention: Semantic sparsity for faster inference. In *Forty-second International Conference on Machine Learning*, 2025.

[14] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999.

[15] Robert M Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

[16] Insu Han, R Jayaram, A Karbasi, V Mirrokno, D Woodruff, and A Zandieh. Hyperattention: Long-context attention in near-linear time. In *International Conference on Learning Representations*. International Conference on Learning Representations, 2024.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[18] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[19] Piotr Indyk, Michael Kapralov, Kshiteej Sheth, and Tal Wagner. Improved algorithms for kernel matrix-vector multiplication under sparsity assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

[21] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

[22] Sahil Joshi, Agniva Chowdhury, Amar Kanakamedala, Ekam Singh, Evan Tu, and Anshumali Shrivastava. Race attention: A strictly linear-time attention for long-sequence training. In *The Fourteenth International Conference on Learning Representations*, 2026.

[23] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

[24] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[25] Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[28] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[30] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[31] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

[32] Paris Siminelakis, Kexin Rong, Peter Bailis, Moses Charikar, and Philip Levis. Rehashing kernel evaluation in high dimensions. In *International Conference on Machine Learning*, pages 5789–5798. PMLR, 2019.

[33] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[34] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[35] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Bo Waggoner. Sub-gamma variables and concentration, 2019. URL https://www.bowaggoner.com/blog/2019/02-03-sub-gamma-concentration/. Available at https://www.bowaggoner.com/blog/2019/02-03-sub-gamma-concentration/.

[38] Haixu Wu, Minghao Guo, Yuezhou Ma, Yuanxu Sun, Jianmin Wang, Wojciech Matusik, and Mingsheng Long. Flashbias: Fast computation of attention with bias. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Appendix A. Related Work

Positional encodings and relative-position-aware attention in transformers are widely studied. Apart from RoPE, a central line of work injects positional information directly into attention through relative-position terms. Shaw et al. [31] introduced relative position representations inside self-attention, and T5 [29] used a simplified scalar relative bias added to attention logits. Fixed and kernelized bias schemes for length extrapolation include ALiBi and KERPLE [9], while FIRE [25] provides a unified functional view. Jelassi et al. [20] also studied hard-ALiBi, a binary sliding-window variant with head-dependent window sizes, while Kazemnejad et al. [23] studied NoPE, i.e., using no positional encoding at all. Our work is complementary to these methods: rather than proposing a new positional bias family, we propose a structural and algorithmic analysis of ALiBi.

From an efficiency perspective, exact attention with bias has recently been optimized at the systems level, including ALiBi-FlashAttention [10] for exact computation and FlashBias [38] via a low-rank decomposition. Finally, LSH has a long history in similarity search [2, 7, 14, 18] and kernel methods [5, 6, 11, 32], and has more recently been used for efficient attention and KV-cache compression [8, 13, 16, 19, 22, 24]. Prior LSH-based attention work hashes token representations or semantic content; in contrast, we use LSH in a different way by hashing positions.

Appendix B. Positional LSH for ALiBi: Full Proofs

We note that $M^{(h)}$ has a distinct structure: its 1-entries are organized in non-overlapping principal combinatorial rectangles. Specifically, for each hash bin ℓ in the range of h, denote by $I_\ell = \{i : h(u_i) = \ell\}$ the indices that fall into that bin. $M^{(h)}$ has 1's in its $I_\ell \times I_\ell$ principal submatrix for every ℓ, and 0's elsewhere. Equivalently, the rows and columns of A can be permuted to make $M^{(h)}$ a block-diagonal matrix in which all blocks are all-1 submatrices. Note that without a permutation, the blocks I_ℓ need not necessarily be contiguous; it could be, for example, that $h(u_1) = h(u_3) \neq h(u_2)$, in which case indices 1 and 3 are hashed into the same bin while 2 is hashed into a different bin.

B.1. Positional LSH for ALiBi via Random Binning Features

ALiBi biases attention weights according to the similarity measure $L^\star_{ij} = \ker(i,j) = e^{-|i-j|/\sigma}$ between token positions i, j. This is the familiar one-dimensional Laplacian kernel. Note that the positional embedding object of each index i is simply its integer value, $u_i = i$.

LSH schemes for the Laplacian kernel are well-known [1, 3, 30]. We will use the one due to Rahimi and Recht [30], which they termed Random Binning Features.

Definition 3 ([30]) *The Random Binning Features (RBFs) LSH scheme over $\mathbb{R}$ is defined as follows. To sample $h \sim \mathcal{H}$, first draw $b \sim \Gamma(2, \sigma)$ from the Gamma distribution with shape 2 and scale σ. It is supported on $[0, \infty)$ with density $p(b) = \sigma^{-2} b e^{-b/\sigma}$. Then sample c uniformly from $[0, b]$. The hash function $h = h_{b,c}$ partitions $\mathbb{R}$ into length-b segments $\{[b\ell + c, b\ell + b + c) : \ell \in \mathbb{Z}\}$ and hashes each number into the segment that contains it. Formally, $h(u) = \lfloor \frac{u-c}{b} \rfloor$ for every $u \in \mathbb{R}$.*

Lemma 4 ([30]) *RBFs satisfy eq. (5) for the Laplacian kernel $\ker(u, u') = e^{-|u-u'|/\sigma}$ over $\mathbb{R}$.*

Therefore, RBFs form a positional LSH scheme for ALiBi. Although RBFs for the Laplacian kernel are classical, our use of them will not be a black-box application: the nontrivial step will be to

analyze their interaction particularly with the space of token indices and achieve matrix approximation for ALiBi-biased attention. To this end, we will require specific properties of their underlying Gamma distribution, summarized in the following fact.

Fact 1 *A Gamma random variable $Z \sim \Gamma(2, \sigma)$ has moments $\mathbb{E}[Z^k] = \sigma^k(k+1)!$ for all $k \in \mathbb{N}$ and moment generating function $\mathbb{E}[e^{tZ}] = (1-\sigma t)^{-2}$ for all $t < 1/\sigma$.*

We observe some additional useful properties of RBFs as a positional LSH scheme over sequence indices $\{1, \ldots, n\}$. First, the hash bins it forms are contiguous – the indices in each bin are consecutive in the context. Therefore, the diagonal all-1 blocks are contiguous in $M^{(h)}$ as defined in the previous section. Second, the hash bins have nearly uniform sizes: each bin contains exactly $\lceil b \rceil$ consecutive indices, except for the first and last bins, which may contain fewer indices due to being on the boundary of the context.

B.2. Approximation in Matrix Norms

We turn to proving eqs. (1) to (3) in Theorem 1. As mentioned earlier, eq. (3) follows from standard scalar concentration for Bernoullis (see Theorem 11). For the spectral norm guarantees, eqs. (1) and (2), we will use the following subexponential matrix Bernstein inequality with unbounded moments.

Theorem 5 ([35], Theorem 6.2 and Remark 6.5) *Let $X_1, \ldots, X_s$ be i.i.d. samples from a distribution X over symmetric $n \times n$ matrices with $\mathbb{E}[X] = L$. Suppose there are $r, \alpha > 0$ such that*

$$\forall\, k \in \mathbb{N}, \quad \|\mathbb{E}[(X-L)^k]\| \leq \frac{k!}{2} \cdot r^{k-2} \cdot \alpha^2. \tag{6}$$

Then, the empirical mean $\widetilde{X} = \frac{1}{s}\sum_{i=1}^{s} X_i$ satisfies $\mathbb{E}[\|L - \widetilde{X}\|] \leq \frac{O(1)}{s} \cdot \max\{\alpha\sqrt{s \log n}, r \log n\}$, and $\Pr[\|L - \widetilde{X}\| \geq \varepsilon] \leq n \cdot \exp\left(-\frac{s \cdot \varepsilon^2}{2\alpha^2 + 2r\varepsilon}\right)$ for all $\varepsilon > 0$.

We thus need to establish eq. (6) for $M \sim \mathcal{M}$ with appropriate r, α. Fixing $k \in \mathbb{N}$, we have by the sub-additivity and sub-multiplicativity of the spectral norm, and a binomial expansion,

$$\|\mathbb{E}[(M-L^\star)^k]\| \leq \mathbb{E}\left[(\|M\| + \|L^\star\|)^k\right] = \sum_{j=0}^{k} \binom{k}{j} \mathbb{E}[(\|M\|-1)^j]\,(\|L^\star\|+1)^{k-j}\,. \tag{7}$$

Thus we need to bound the moments $\mathbb{E}[(\|M\|-1)^j]$ and the powers $(\|L^\star\|+1)^j$.

Lemma 6 $\mathbb{E}[(\|M\|-1)^k] \leq \sigma^k(k+1)!$ *for all $k \in \mathbb{N}$.*

Proof M is a block diagonal matrix with all-1 blocks. Let b_M denote the size of its largest block. It is straightforward to see that the largest eigenvalue of M is b_M, thus $\|M\| = b_M$. Next, recall that $b_M \leq \lceil b \rceil \leq b+1$ where $b \sim \Gamma(2, \sigma)$. Hence, $(\|M\|-1)^k \leq b^k$. Taking expectation of both sides, the lemma follows from the formula for Gamma moments in Fact 1. ■

Lemma 7 $\|L^\star\| \leq (e^{1/\sigma}+1)/(e^{1/\sigma}-1)$.

Proof We show two ways to prove the lemma. The first is a more generic approach based on Fourier analysis of Toeplitz matrices, which has the potential advantage of possibly generalizing better to other positional bias matrices beyond ALiBi. Recall that a matrix T is Toeplitz if there is a function $g : \mathbb{Z} \to \mathbb{R}$ such that $T_{ij} = g(i-j)$ for all i, j. The ALiBi matrix $L^\star$ is Toeplitz with $g(z) = e^{-|z|/\sigma}$. The following lemma offers a way to bound the spectral norm of a Toeplitz matrix.

Lemma 8 ([15], Lemma 4.1) *Let T be a Toeplitz matrix with $T_{ij} = g(i-j)$. Suppose g satisfies $\sum_{i=-\infty}^{\infty} |g(i)| < \infty$. Then $\|T\| \le \sup f$ where f is the discrete-time Fourier transform (DTFT) of g.*

Thus, to bound the spectral norm of $L^\star$ with Theorem 8 we need the DTFT of $g(z) = e^{-|z|/\sigma}$. A standard fact, which can be verified by direct derivation, is that the DTFT of $g_\phi(z) = \phi^{|z|}$ for $\phi \in (-1, 1)$ is $f(\omega) = (1-\phi^2)/(1-2\phi\cos(\omega)+\phi^2)$ and its supremum is $(1+\phi)/(1-\phi)$ (attained at $\cos(\omega) = 1$). For $L^\star$ we have $\phi = e^{-1/\sigma}$, and thus Theorem 7 follows from Theorem 8.

The second proof for Theorem 7 is more direct, albeit more specialized to ALiBi. Recall the matrix norm definitions $\|M\|_1 = \max_j \sum_i |M_{ij}|$ and $\|M\|_\infty = \max_i \sum_j |M_{ij}|$. By a standard Hölder inequality for matrix norms, $\|L^\star\| \le \sqrt{\|L^\star\|_1 \|L^\star\|_\infty}$. Since $L^\star$ is symmetric, $\|L^\star\|_1 = \|L^\star\|_\infty$, hence $\|L^\star\| \le \|L^\star\|_\infty$. Thus, it suffices to bound the row sums in $L^\star$. The sum of row i is

$$\sum_{j=1}^{n} L^\star_{ij} = 1 + \sum_{t=1}^{i-1} e^{-t/\sigma} + \sum_{t=1}^{n-i} e^{-t/\sigma} \le 1 + 2\sum_{t=1}^{\infty} e^{-t/\sigma} = \frac{e^{1/\sigma}+1}{e^{1/\sigma}-1},$$

where the final equality is established by calculating the sum of the geometric series. ■

We plug Theorems 6 and 7 into eq. (7) and get

$$\|\mathbb{E}[(M-L^\star)^k]\| \le \sum_{j=0}^{k} \binom{k}{j} \cdot \sigma^j (j+1)! \cdot \left(\frac{e^{1/\sigma}+1}{e^{1/\sigma}-1} + 1\right)^{k-j} \le 2^k k! \sum_{j=0}^{k} \binom{k}{j} \cdot \sigma^j \cdot \left(\frac{e^{1/\sigma}+1}{e^{1/\sigma}-1} + 1\right)^{k-j}$$

$$= 2^k k! \cdot \left(1 + \sigma + \frac{e^{1/\sigma}+1}{e^{1/\sigma}-1}\right)^k = (2\Psi_\sigma)^k k! = \frac{k!}{2} \cdot (2\Psi_\sigma)^{k-2} \cdot (2\sqrt{2}\Psi_\sigma)^2,$$

where we have used that $(j+1)! \le (k+1)! \le 2^k k!$ for all $k \ge j \ge 0$. Thus, eq. (6) holds with $r = 2\Psi_\sigma$ and $\alpha = 2\sqrt{2}\Psi_\sigma$, and thus eqs. (1) and (2) in Theorem 1 now follow from Theorem 5.

B.3. Maximal Block Size

We proceed to proving eq. (4) in Theorem 1. By the previous sections, this amounts to bounding the maximal bin size in RBFs. To this end, we recall subgamma concentration (see, e.g., [37]).

Definition 9 *A real-valued random variable Z is called* (ν, κ)-right-subgamma *if for all $t \in (0, 1/\kappa)$ it holds that $\mathbb{E}\left[e^{t(Z-\mathbb{E}[Z])}\right] \le \exp\left(\nu t^2/(2-2\kappa t)\right)$.*

Fact 2 *if Z is (ν, κ)-right-subgamma, then for all $t > 0$, $\Pr[Z \ge \mathbb{E}[Z] + \sqrt{2\nu t} + \kappa t] \le e^{-t}$.*

Lemma 10 (see Theorem 12 in the appendix) *$Z \sim \Gamma(2, \sigma)$ is $(2\sigma^2, \sigma)$-right-subgamma.*

In Theorem 1 we draw i.i.d. samples $M_1, \ldots, M_s \sim \mathcal{M}$ and denote by $b_{\max}$ their maximal block size. Let $b_1, \ldots, b_s \sim \Gamma(2, \sigma)$ be the corresponding samples of their RBFs. By the previous sections, each M_j is block-diagonal with maximal bin size $\lceil b_j \rceil \leq b_j + 1$. Therefore, $b_{\max} \leq 1 + \max_j b_j$. We use Fact 2 with $t = \ln(s/\delta)$, $\mathbb{E}[b] = 2\sigma$ from Fact 1, and ν, κ from Theorem 10. We get

$$\forall j, \quad \Pr[b_j > 1 + 2\sigma + 2\sigma\sqrt{\ln(s/\delta)} + \sigma \ln(s/\delta)] < \delta/s.$$

Since $s \geq 1$ and $\delta < 1/e$, eq. (4) follows by a union bound over $j = 1, \ldots, s$.

B.4. Near-Linear time Approximation Algorithm

In this section, we turn to Theorem 2. The algorithm in the theorem replaces the exact ALiBi matrix $L^\star$ with the approximation $\widetilde{M}$. Given an attention instance (K, Q, V), in the notation from Section 2, we have $A^\star = A \odot L^\star$ (non-causal case) or $A^\star = A \odot J \odot L^\star$ (causal case), $P^\star = (D^{[A^\star]})^{-1} A^\star$, and the exact output is $T^\star = P^\star V$. The algorithm in Theorem 2 computes $\widetilde{A}^\star = A \odot \widetilde{M}$ (non-causal case) or $\widetilde{A}^\star = A \odot J \odot \widetilde{M}$ (causal case), then $\widetilde{P}^\star = (D^{[\widetilde{A}^\star]})^{-1} \widetilde{A}^\star$, and returns $\widetilde{T}^\star = \widetilde{P}^\star V$.

Running time analysis. Let M be a sample from $\mathcal{M}$ and let $I_1, \ldots, I_t \subset \{1, \ldots, n\}$ be its contiguous blocks of indices. Let $\widetilde{A} = A \odot M$. Since M and thus $\widetilde{A}$ is block-diagonal, we can compute $\widetilde{A}V$ block by block as $\widetilde{A}V = \sum_{j=1}^t \widetilde{A}_{[I_j, I_j]} V_{[I_j, :]}$. Since the blocks in M are all-1 matrices, $\widetilde{A}_{[I_j, I_j]} = (A \odot M)_{[I_j, I_j]} = A_{[I_j, I_j]}$, and hence $\widetilde{A}V = \sum_{j=1}^t A_{[I_j, I_j]} V_{[I_j, :]}$. Similarly, $D^{[\widetilde{A}]}$ can be computed as $D^{[\widetilde{A}]} = \sum_{j=1}^t D^{[A_{[I_j, I_j]}]}$. The upshot is that $\widetilde{A}$ and $D^{[\widetilde{A}]}$ can be computed by computing only the restriction of A to the blocks of M. Note that A is the regular (positionally unbiased) attention matrix, and therefore, computing $\widetilde{A}$ and $D^{[\widetilde{A}]}$ is reduced to a collection of smaller regular attention computations on the subsets of keys and queries that correspond to the blocks $\{I_j\}$. The time complexity is $d \sum_{j=1}^t |I_j|^2$. Observe that $\sum_j |I_j| = n$ since the blocks form a partition of $\{1, \ldots, n\}$, and therefore $d \sum_{j=1}^t |I_j|^2 \leq d \max_j |I_j| \sum_j |I_j| = nd \max_j |I_j|$. With the empirical mean $\widetilde{M} = \frac{1}{s} \sum_{i=1}^s M_i$ of s samples with block sizes $\{I_{ij}\}$, the overall time complexity for computing $\widetilde{T}^*$ is thus $\sum_{i=1}^s nd \max_j |I_{ij}| \leq ndsb_{\max}$. The probabilistic bound on $b_{\max}$ in Theorem 1 (eq. (4)) now establishes the running time in Theorem 2 in the non-causal case. The same arguments hold in the causal case with $A \odot J$ instead of A.

Approximation guarantee. Here we only outline how matrix norm approximation yields attention approximation, and leave the full details to the appendix. For all i, j one can verify that $T^\star_{ij} = (\sum_k P_{ik} L^\star_{ik} V_{kj}) / (\sum_k P_{ik} L^\star_{ik})$ and $\widetilde{T}^\star_{ij} = (\sum_k P_{ik} \widetilde{M}_{ik} V_{kj}) / (\sum_k P_{ik} \widetilde{M}_{ik})$, where recall that P is the row-stochastic matrix $P = (D^{[A]})^{-1} A$. The error in the numerators can be bounded either with the max-norm of $L^\star - \widetilde{M}$ as

$$\left| \sum_k P_{ik} (L^\star - \widetilde{M})_{ik} V_{kj} \right| \leq \|L^\star - \widetilde{M}\|_{\max} \|V\|_{\max} \left| \sum_k P_{ik} \right| = \|L^\star - \widetilde{M}\|_{\max} \|V\|_{\max}, \tag{8}$$

or with the spectral norm of $L^\star - \widetilde{M}$ as

$$\left| \sum_k P_{ik} (L^\star - \widetilde{M})_{ik} V_{kj} \right| = \left| (L^\star - \widetilde{M})^\top_{i,*} (P_{i,*} \odot V_{*,j}) \right| \leq \|L^\star - \widetilde{M}\| \cdot \|V\|_{\max} \|P\|_{2,\infty}. \tag{9}$$

A similar argument applies to the denominators with an all-1 matrix instead of V. From Theorem 1 we now get that the numerator error is bounded by $\Delta_P \|V\|_{\max}$ and the denominator error is bounded by Δ_P. The bound on $\|T^\star - \widetilde{T}^\star\|_{\max}$ now follows by manipulation. See Appendix B for full details.

B.5. Omitted Lemmas for Theorem 1

Lemma 11 (max-norm bound in Theorem 1) *Under the assumptions of Theorem 1, the max-norm bound in eq. (3) holds.*

Proof Let $i, j \in [n]$. By Theorem 4 and the positional LSH construction in Section 3, we have $L^\star_{ij} = \Pr_{M\sim\mathcal{M}}[M_{ij} = 1]$. Therefore, $\widetilde{M}_{ij}$ is the average of s Bernoulli random variables with expectation $L^\star_{ij}$. Therefore, by the Chernoff-Hoeffding inequality,

$$\Pr\left[\left|L^\star_{ij} - \widetilde{M}_{ij}\right| > \varepsilon\right] < 2e^{-2s\varepsilon^2}$$

The lemma follows by a union bound over all i, j. ■

Lemma 12 (Theorem 10 restated) *$Z \sim \Gamma(2, \sigma)$ is $(2\sigma^2, \sigma)$-right-subgamma.*

Proof Let $t \in (0, 1/\sigma)$. By plugging the expectation and MGF from Fact 1,

$$\ln \mathbb{E}[e^{t(Z-\mathbb{E}[Z])}] = \ln\left(\mathbb{E}[e^{tZ}] \cdot e^{-t\mathbb{E}[Z]}\right) = \ln((1-\sigma t)^{-2} \cdot e^{-2\sigma t}) = 2(-\ln(1-\sigma t) - \sigma t).$$

From the Taylor expansion of $-\ln(1-x)$ at 0 we have, for every $x \in (0, 1)$,

$$-\ln(1-x) - x = \sum_{i=1}^{\infty} \frac{x^i}{i} - x = \sum_{i=2}^{\infty} \frac{x^i}{i} \le \frac{x^2}{2} \sum_{i=0}^{\infty} x^i = \frac{x^2}{2} \cdot \frac{1}{1-x}.$$

Plugging this above with $x = \sigma t$,

$$\ln \mathbb{E}[e^{t(Z-\mathbb{E}[Z])}] \le 2 \cdot \frac{\sigma^2 t^2}{2} \cdot \frac{1}{1-\sigma t} = \frac{2\sigma^2 t^2}{2(1-\sigma t)}.$$

Exponentiating both sides proves the claim. ■

B.6. Proof of Theorem 2

The running time was proven in Section B.4. Here we prove the approximation guarantee. Let $A' = A$ in the non-causal case and $A' = A \odot J$ in the causal case. In either case we have $P = (D^{[A']})^{-1}A'$, and P is row-stochastic.

Observe that the exact target output $T^\star$ and the returned approximation $\widetilde{T}^\star$ are computed as

$$T^\star_{ij} = \frac{\sum_{k=1}^n A'_{ik} L^\star_{ik} V_{kj}}{\sum_{k=1}^n A'_{ik} L^\star_{ik}} \quad \text{and} \quad \widetilde{T}^\star_{ij} = \frac{\sum_{k=1}^n A'_{ik} \widetilde{M}_{ik} V_{kj}}{\sum_{k=1}^n A'_{ik} \widetilde{M}_{ik}}.$$

We may divide both the numerators and the denominators by $D_{ii}^{[A']}$ and write,

$$T^\star_{ij} = \frac{\sum_{k=1}^n P_{ik} L^\star_{ik} V_{kj}}{\sum_{k=1}^n P_{ik} L^\star_{ik}} \quad \text{and} \quad \widetilde{T}^\star_{ij} = \frac{\sum_{k=1}^n P_{ik} \widetilde{M}_{ik} V_{kj}}{\sum_{k=1}^n P_{ik} \widetilde{M}_{ik}}.$$

For convenience let us denote the numerators and the denominators as

$$N_{ij} = \sum_{k=1}^n P_{ik} L^\star_{ik} V_{kj}, \quad \widetilde{N}_{ij} = \sum_{k=1}^n P_{ik} \widetilde{M}_{ik} V_{kj}, \quad D_i = \sum_{k=1}^n P_{ik} L^\star_{ik}, \quad \widetilde{D}_i = \sum_{k=1}^n P_{ik} \widetilde{M}_{ik}.$$

Equations (8) and (9) in Section B.4 show that

$$\forall i,j \quad \left| N_{ij} - \widetilde{N}_{ij} \right| \le \min\{ \|L^\star - \widetilde{M}\|_{\max}, \|P\|_{2,\infty} \|L^\star - \widetilde{M}\| \} \cdot \|V\|_{\max},$$

and the same arguments with an all-1 matrix instead of V yield

$$\forall i \quad \left| D_i - \widetilde{D}_i \right| \le \min\{ \|L^\star - \widetilde{M}\|_{\max}, \|P\|_{2,\infty} \|L^\star - \widetilde{M}\| \}.$$

With probability $1-\delta$ we have by Equations (2) and (3) in Theorem 1,

$$\|L^\star - \widetilde{M}\|_{\max} \le O\left(\sqrt{\frac{\log(n/\delta)}{s}} \right) \quad \text{and} \quad \|L^\star - \widetilde{M}\| \le O\left(\Psi_\sigma \sqrt{\frac{\log(n/\delta)}{s}} + \frac{\log(n/\delta)}{s} \right).$$

In the statement of Theorem 2 we assume that $s \ge \log(n/\delta)$ and hence $\frac{\log(n/\delta)}{s} \le \sqrt{\frac{\log(n/\delta)}{s}}$. Plugging this above, we get

$$\forall i,j \quad \left| N_{ij} - \widetilde{N}_{ij} \right| \le \Delta_P \|V\|_{\max} \quad \text{and} \quad \forall i \quad \left| D_i - \widetilde{D}_i \right| \le \Delta_P. \tag{10}$$

Furthermore, we have

$$\forall i,j \quad \left| \frac{N_{ij}}{D_i} \right| = \left| \frac{\sum_{k=1}^n P_{ik} L^\star_{ik} V_{kj}}{\sum_{k=1}^n P_{ik} L^\star_{ik}} \right| \le \left| \frac{\sum_{k=1}^n P_{ik} L^\star_{ik}}{\sum_{k=1}^n P_{ik} L^\star_{ik}} \right| \cdot \|V\|_{\max} = \|V\|_{\max}. \tag{11}$$

From Equations (10) and (11),

$$\left| \widetilde{T}^\star_{ij} - T^\star_{ij} \right| = \left| \frac{\widetilde{N}_{ij}}{\widetilde{D}_i} - \frac{N_{ij}}{D_i} \right| = \left| \frac{\widetilde{N}_{ij} - N_{ij}}{\widetilde{D}_i} + \frac{N_{ij}}{D_i} \cdot \frac{D_i - \widetilde{D}_i}{\widetilde{D}_i} \right| \le \frac{2\Delta_P \|V\|_{\max}}{\widetilde{D}_i}. \tag{12}$$

Finally, observe that $\beta^\star_P = \min_i D_i$ in the notation of Theorem 2. Therefore, if the condition of item 2 in the theorem holds, then by Equation (10) we have for all i that $\widetilde{D}_i \ge D_i - \left| \widetilde{D}_i - D_i \right| \ge \beta^\star_P - \Delta_P$. Plugging this in Equation (12) yields the theorem. □

Appendix C. Generality: Positional LSH beyond ALiBi

So far we have focused on ALiBi as a concrete case of a popular positional bias that adheres to our positional LSH framework. In this section, we draw a distinction between the results that extend to any positional bias scheme that adheres to Section 3 and the results specific to ALiBi.

What extends to any positional LSH: Let L be any positional bias matrix that satisfies eq. (5) in Section 3. Then the identity $\mathbb{E}_{M\sim\mathcal{M}}[M] = L$ in Theorem 1 immediately holds. Furthermore, the max-norm convergence bound, eq. (3) in Theorem 1, also holds, since its proof relies only on scalar Bernoulli concentration (Theorem 11). Consequently, the uniform attention approximation guarantee of Theorem 2 holds in a slightly weaker form, with $\Delta'_P = C\sqrt{\log(n/\delta)/s}$ instead of Δ_P.

What is specific to ALiBi: In Theorem 1, the spectral norm convergence bounds eqs. (1) and (2) in Theorem 1 and the block size bound eq. (4) rely on the specific properties of the RBFs LSH scheme for ALiBi (Theorem 3), particularly on properties of its underlying Gamma distribution and its interplay with the space of indices in a sequence (Sections B.2 and B.3). This also includes the contiguity of the blocks (see Sections B.1 and 3). For other positional bias and LSH schemes, the analogous results would require their own proofs. Note that the bound on the block size is key to the running time of the positionally-biased attention approximation algorithm in Theorem 2.

It is worth remarking that our results extend fully to natural extensions of ALiBi. For example, a two-dimensional variant of ALiBi may be defined over pixels in an image as $L^{\star\star}_{ij} = \exp(-\|\bar{x}-\bar{y}\|_1)$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are pixel coordinates. Since RBFs form an LSH scheme for the Laplacian kernel $\exp(-\|x-y\|_1)$ of any dimension, our analysis will go through.

Appendix D. Full Experiments and Experimental Details

D.1. Convergence in Matrix Norms

We validate our theoretical results on convergence in matrix norms in Theorems 1 and 2 directly on attention operations in pre-trained models. We use the publicly available Mistral-7B model. We use as inputs 30 text segments from the Wikitext-103 dataset [27] with context lengths of 4k-4.5k and record the output of the models. We then retrieve the Q,K,V matrices for each model per input, layer and attention head. We compute $\|L^\star - \widetilde{M}\|$, $\|L^\star - \widetilde{M}\|_{\max}$ and $\|T^\star - \widetilde{T}^\star\|_{\max}$, for varying sample sizes.

The results are reported in Figure 2. Results for additional σ values is displayed in Appendix D. The results show that for both models, norm errors decay towards zero as the sample size increases.

Figure 3 depicts plots for additional σ values, complementing the experiments described in subsection 4.

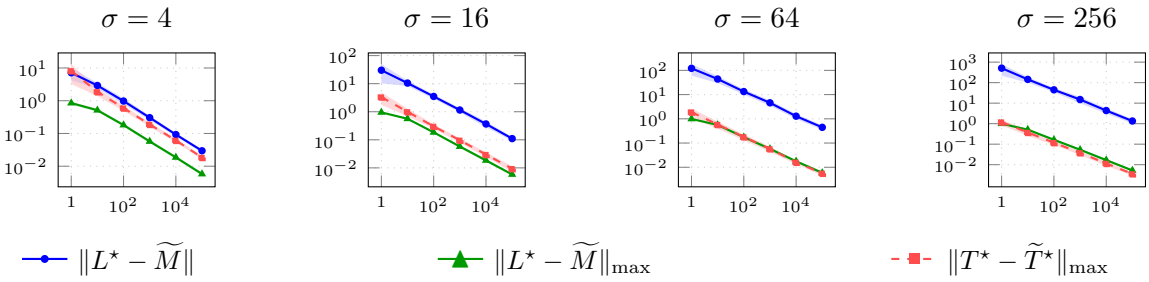

Figure 3: $\|L^\star - \widetilde{M}\|$, $\|L^\star - \widetilde{M}\|_{\max}$ and $\|T^\star - \widetilde{T}^\star\|_{\max}$ as functions of the sample size s, for four attention heads (with σ values: $\sigma \in \{4, 16, 64, 256\}$), for Mistral-7B [21]. Both axes show values on a log scale. Shaded regions span [mean $-$ std, mean $+$ std] across inputs, and they do not appear if the standard deviation is negligible.

The results show that all norms decrease towards zero as the sample size increases.

D.2. Model Training Experiments

We evaluate the effects of our approximation method on downstream model performance. To this end, we implemented a prototype of our positional LSH method for ALiBi in FlashAttention-2 [12]. We note that our implementation is not optimized for performance from an engineering perspective (see also the discussion of our limitations in Section E). The goal of the experiments in this section is to gain empirical insight into the approximation quality of our method rather than demonstrate efficiency gains. We use the Wikitext-103 dataset [27] and train two publicly available models:

- Qwen3-0.6B [34] for pre-training from scratch experiments.
- Mistral-7B [21] for fine-tuning experiments from its publicly available pre-trained checkpoint.

We evaluate perplexity (PPL) and next-token prediction accuracy (Acc), both on the original context length used in training and under context length extrapolation (where the model is evaluated on longer context lengths than it was trained on). Qwen3-0.6B is trained with an 8k context length and evaluated at an extrapolated length of 16k. Mistral-7B is trained with a 4k context length and evaluated at extrapolated lengths of 8k and 16k. We evaluate Positional LSH with different sample sizes s. We compare it with the original models without positional bias and with ALiBi. As a baseline for comparison, we also include a "fixed blocks" baseline with a single block-diagonal mask matrix with a fixed block size of 512.

The results are displayed in Tables 1 and 2. For both models, as the number of samples in Positional LSH increases, performance improves, approaching that of ALiBi. Furthermore, Positional LSH outperforms the fixed blocks baseline in both models. Finally, for Qwen, Positional LSH outperforms the original model, even with a small sample size. On Mistral-7B, on the other hand, neither positional bias method outperforms the original model in length extrapolation.

Experimental details and parameters are provided in Appendix D.

Experimental details. Experiments were run on an NVIDIA H100 Tensor Core GPU (with 80GB VRAM). The AdamW optimizer [26] was used for training, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The hyperparameters were selected via evaluations performed on three different values for each parameter, spanning conventional value ranges. For Qwen3-0.6B, a learning rate of 5×10^{-4} is used, while for Mistral-7B the learning rate is 1×10^{-4}. The remaining hyperparameters

are shared across both models: 5 epochs, a weight decay of 0.01, and a warmup ratio of 0.1. A batch size of 1 is used for training and evaluation in both models, as larger batch sizes would exceed GPU memory limits and would compromise context length. The results for the matrix norm convergence experiments are averaged over 3 independent runs. Mistral-7B was fine-tuned using LoRA [17] with the following configuration: rank 16, $\alpha = 32$, dropout 0.05, no bias, targeting the query, key, value, and output projection matrices.

Appendix E. Conclusion and Limitations

We introduced a positional LSH framework for attention with bias, establishing a formal connection between positional bias, binary attention masks, and positional embeddings. For ALiBi, this yields both a structural approximation of the bias matrix by randomized contiguous block masks and a uniform near-linear time approximation theorem for ALiBi-biased attention. Our experiments support this approximation perspective: as the number of samples increases, the approximation improves and downstream behavior approaches exact ALiBi.

Our contribution is theoretical and algorithmic rather than a systems result, and our experiments are of limited scale meant only to validate the theory. Although our analysis gives a near-linear asymptotic algorithm, we do not demonstrate wall-clock speedups over hardware-optimized exact ALiBi implementations. Some of this gap may be due to our implementation being under-optimized, but it likely also reflects the fact that current hardware and low-level kernels are highly optimized for large dense operations, so asymptotic gains from decomposing attention into many smaller local computations need not translate into speedups at the context lengths we experimented with. Identifying regimes where the favorable asymptotics of positional LSH lead to practical gains may involve much longer context lengths and is left an intriguing direction for future work.