NeurIPS 2026

Approximate Matrix–Vectors Under a Bounded ℓ_1 Assumption and Applications to Kernel Matrices

Rikhav Shah[1] Sandeep Silwal[2] Tony Chang Wang[2]

[1]MIT [2]University of Wisconsin–Madison

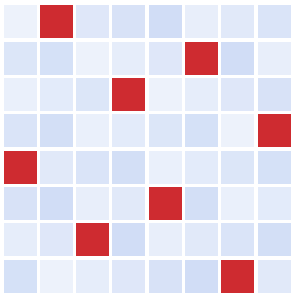
Massachusetts Institute of Technology

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Light matrices can still be dense

Given $A \in [-1, 1]^{n \times n}$, approximate Ax with

$$\|y - Ax\|_2 \leq \varepsilon \|x\|_2.$$



Many small, nonzero entries

Bound the total entry mass

$$\sum_{i,j} |a_{ij}| \leq \gamma n.$$

γ is the **lightness** parameter.

Example: normalized attention

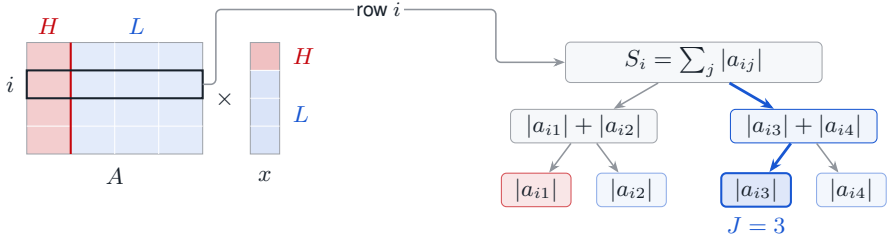
A dense row-stochastic matrix has

$$\sum_{i,j} P_{ij} = n \quad \Longrightarrow \quad \gamma = 1.$$

General matrices: split, then sample

Theorem. Any $A \in [-1,1]^{n\times n}$ with $\|A\|_{1,1} \le \gamma n$.
$O(n^2)$ preprocessing; $\widetilde{O}(\gamma n^{3/2}/\varepsilon)$ time per query.
$\|y - Ax\|_2 \le \varepsilon \|x\|_2$ with probability at least $1-\delta$ per query.

$$\|x\|_2 = 1, \qquad x = x_H + x_L, \qquad H = \{j : |x_j| \ge T\}.$$



Heavy: exact Ax_H, $|H| \le 1/T^2$.

Light: sample $p_{iJ} = |a_{iJ}|/S_i$.

Main result: Faster kernel MVPs

Theorem. Let $K_{ij} = \kappa(r_i, c_j) \in [0,1]$, with $r_i, c_j \in \mathbb{R}^d$ and $\|K\|_{1,1} \leq \gamma n$. If κ admits a Fast-KDE data structure with exponent $p \in (0,1)$, our algorithm returns y satisfying

$$\|y - Kx\|_2 \leq \varepsilon \|x\|_2 \quad \text{with probability at least } 1 - \delta.$$

Total time, including KDE construction:

$$\widetilde{O}\Big(d\,\gamma^{\frac{4}{p+4}}\,\varepsilon^{-\frac{2(p+2)}{p+4}}\,n^{\frac{3}{2}+\frac{5p}{2(p+4)}}\Big).$$

Gaussian kernel, $\gamma = O(1)$:

Indyk et al. (ICLR 2025)		**This work**
$\widetilde{O}\left(\frac{dn^{1.89}}{\varepsilon^2}\right)$	$\longrightarrow$	$\widetilde{O}\left(\frac{dn^{1.604}}{\varepsilon^{1.041}}\right)$

Signed x; asymmetric K; never constructs the full matrix.

Shah, Silwal, and Wang, Theorem 2.2. Gaussian exponents are approximate. Prior work: Indyk, Kapralov, Sheth and Wagner, ICLR 2025.

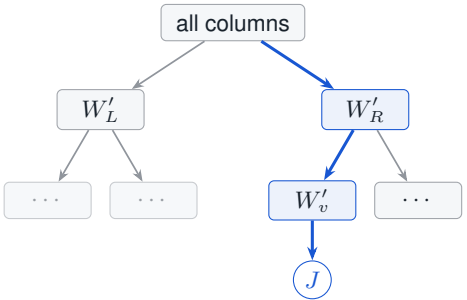
KDE estimates the masses we need

Kernel density estimation (KDE) estimates sums of kernel similarities.

For a fixed row i, node v groups m_v column points.

$$W(v) = \sum_{j \in \text{leaves}(v)} K_{ij}.$$

Query estimates $W'(v)$ **only when needed**.



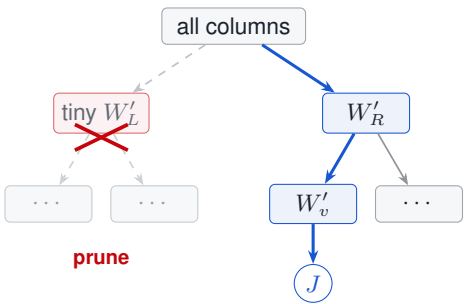
$$W(v) \le W'(v) \le (1+\eta)W(v) + \underbrace{m_v \mu}_{\text{additive error}}.$$

The obstacle: additive error can overwhelm tiny masses.

Fast-KDE interface: Definition 2.2 of our paper. Gaussian KDE ingredient: Charikar, Kapralov, Nouri and Siminelakis, FOCS 2020.

Why pruning works

Prune subtrees whose estimated mass is too small relative to additive KDE error.



1. Reliable surviving mass

For every surviving subtree v, $m_v\mu$ is small relative to $W(v)$. The estimate $W'(v)$ reliably guides sampling.

2. Small discarded mass

Each row loses little total kernel weight. Each removed entry is multiplied by a coordinate of x_L of magnitude at most T.

pruning error in row i $\leq$ $T \times$ discarded mass in row i.

Improvements across kernels

The theorem specializes to four kernel families at constant lightness.

Kernel	**KDE exponent** p	**Total time** / d
Gaussian	$0.173 + o(1)$	$\widetilde{O}(n^{1.604}/\varepsilon^{1.041})$
Exponential	$0.1 + o(1)$	$\widetilde{O}(n^{1.561}/\varepsilon^{1.024})$
Laplacian	$1/2$	$\widetilde{O}(n^{16/9}/\varepsilon^{10/9})$
Rational quadratic	0	$\widetilde{O}(n^{3/2}/\varepsilon)$

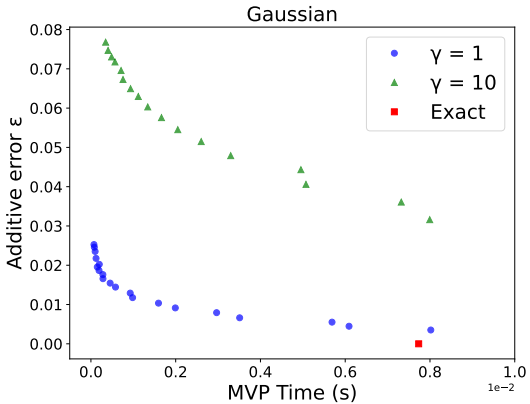
$\gamma = O(1)$; decimal exponents are approximate; rational quadratic has constant shape parameter β.

Smaller KDE exponent p gives faster kernel MVPs.

As $p \to 0$, the exponent of n approaches $3/2$.

All bounds include KDE construction. Shah, Silwal, and Wang, Theorem 2.2 and Table 1.

Empirical evaluation: Gaussian kernel



Shah, Silwal, and Wang, Appendix D, Figure 3 (original Gaussian panel).

Deterministic sparse + low rank
Theorem C.1.

Data
Synthetic Gaussian points:
10^4 rows, 10^4 columns; $d = 50$.

Evaluation
Gaussian kernel; 10 random unit query vectors per setting.

Query time after preprocessing.

Lower lightness improves the error–time tradeoff.

Small total mass enables fast MVPs

General matrices

$$\widetilde{O}\left(\frac{\gamma n^{3/2}}{\varepsilon}\right)$$

query time after $O(n^2)$ preprocessing.

Gaussian kernel matrices

$$\widetilde{O}\left(\frac{dn^{1.604}}{\varepsilon^{1.041}}\right)$$

total time, for $\gamma = O(1)$.

Better KDE gives faster kernel MVPs.

Gaussian, Exponential, Laplacian, and Rational Quadratic kernels.

Gaussian decimal exponents are approximate. Shah, Silwal, and Wang, Theorems 2.1–2.2.