

Lecture 7: Speed Up Attention Via RFF

Instructors: *Alex Andoni, Edo Liberty*Scribe: *Roie Gilad*

1 Motivation

As introduced in Lecture 6, the core part of the Attention Mechanism is

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad Q, K, V \in \mathbb{R}^{n \times d}.$$

Throughout these notes we absorb the $1/\sqrt{d}$ factor into the inputs (replace q, k by $q/d^{1/4}, k/d^{1/4}$) and write $\text{Softmax}(QK^\top)V$. This computation consists of several computational and memory bottlenecks as the context windows (n) keep growing, and the scale of these dot products is about to explode. We have seen in Lecture 6 that FlashAttention and Online Softmax reduce the IO traffic (HBM accesses) from $\Theta(n^2)$ to $\Theta(n^2d^2/M)$, where M is the size of our SRAM. However, we still perform all $\Theta(n^2d)$ arithmetic FLOPs; more specifically, computing $e^{QK^\top}V$ takes $\Theta(nd)$ for each query.

2 Problem: How to do it fast, and save FLOPs

Recall that the softmax here is applied row-wise, so with $e^{(\cdot)}$ taken entrywise,

$$\text{Softmax}(QK^\top)V = D^{-1}e^{QK^\top}V, \quad D = \text{diag}(e^{QK^\top}\mathbf{1}).$$

So it suffices to compute $e^{QK^\top}V$ and the normalizer $e^{QK^\top}\mathbf{1}$ fast. Suppose we have a ‘magic’ function $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$, applied row-wise to Q and K (so $f(Q), f(K) \in \mathbb{R}^{n \times m}$), such that $e^{q^\top k} \approx f(q)^\top f(k)$ for every query q and key k , i.e.

$$e^{QK^\top} \approx f(Q)f(K)^\top.$$

Then, the following holds:

$$e^{QK^\top}V \approx (f(Q)f(K)^\top)V.$$

Since matrix multiplication is associative, we can regroup:

$$e^{QK^\top}V \approx (f(Q)f(K)^\top)V = f(Q)(f(K)^\top V).$$

Define A as follows:

$$A = \sum_{j=1}^n f(k_j) v_j^\top = f(K)^\top V \in \mathbb{R}^{m \times d}.$$

This is a fixed-size $m \times d$ summary of all key-value pairs. Each query just reads its output from it: $f(q_i)^\top A$. We handle the softmax normalization in the same way:

$$z = f(K)^\top \mathbf{1} = \sum_{j=1}^n f(k_j) \in \mathbb{R}^m.$$

Definition 1 (Linear-attention). *Given a feature map $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$, the linear-attention output for query q_i is*

$$\text{output}_i = \frac{f(q_i)^\top A}{f(q_i)^\top z} \in \mathbb{R}^{1 \times d}, \quad A = \sum_{j=1}^n f(k_j) v_j^\top, \quad z = \sum_{j=1}^n f(k_j).$$

If $f(q)^\top f(k) = e^{q^\top k}$ exactly, this equals row i of $\text{Softmax}(QK^\top)V$.

The contribution of this method is even greater during inference, where each token depends only on the past (causal attention). Then A and z are replaced by prefix sums, and we can work in an incremental manner:

$$A_i = A_{i-1} + f(k_i) v_i^\top, \quad z_i = z_{i-1} + f(k_i), \quad \text{output}_i = \frac{f(q_i)^\top A_i}{f(q_i)^\top z_i}.$$

Now, each new token costs $O(md)$, with no recomputation over history.

2.1 What Do We Actually Save?

Let n be the sequence length, d the head dimension, and m the feature dimension of f .

Full sequence (training / prefill).

	$\underbrace{(e^{QK^\top})V}_{\text{standard}}$	vs.	$\underbrace{f(Q)(f(K)^\top V)}_{\text{linear}}$
	Standard		Linear
Time	$O(n^2 d)$		$O(n m d)$
Memory	$O(n^2)$ (attention matrix)		$O(n m + m d)$ (features and state A)

With FlashAttention, standard attention needs only $O(n)$ extra memory, but the time stays $O(n^2 d)$. With a causal mask, the prefix states $A_1, \dots, A_n$ are computed by a single prefix-sum pass in the same $O(nmd)$ time. Storing all of them would take $O(nmd)$ memory, so in practice they are computed on the fly (or chunk by chunk) and only the running state is kept.

Per generated token (autoregressive decoding). As noted, each step in this setting costs $O(md)$ with $O(md)$ memory, which is *constant in n* . The saving comes from never forming the $n \times n$ matrix: the quadratic dependence on n becomes linear, and decoding becomes constant-time per token. This pays off when $n \gg m$. The price is that $f(Q)f(K)^\top$ only approximates $e^{QK^\top}$, and a better approximation needs a larger m (we quantify this in Theorems 8 and 10).

3 Constructing the Magic Function

Goal. We want a feature map $f : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that

$$e^{q^\top k} \approx f(q)^\top f(k).$$

The strategy is to write $e^{q^\top k}$ as an *expectation of a product* $\mathbb{E}[g(q)g(k)]$, and then approximate the expectation by an average over a finite (but large) number of samples. Such an average is exactly a dot product. We denote the number of random samples by m . The feature dimension is then $D = 2m$ (Step 3) or $D = m$ (Step 4); this constant factor does not affect Section 2.1.

Random features keep every token and act on each q and k separately, so the prefix-sum state of Section 2 works. Their dimension depends on the accuracy we want and only logarithmically on n . Concretely, for the trigonometric features of Step 3, taking

$$m = O\left(\frac{\log(n/\delta)}{\varepsilon^2}\right)$$

guarantees that, with probability at least $1 - \delta$, simultaneously for all n^2 query-key pairs,

$$\left|f(q)^\top f(k) - e^{q^\top k}\right| \leq \varepsilon e^{(\|q\|^2 + \|k\|^2)/2}$$

(this follows from Theorem 8 and Corollary 9). The error is additive at the scale $e^{(\|q\|^2 + \|k\|^2)/2}$. Step 4 improves this to an error relative to $e^{q^\top k}$ (Theorem 10).

Notation. Throughout this section, $w \sim \mathcal{N}(0, I_d)$ is a random vector with i.i.d. entries $w_i \sim \mathcal{N}(0, 1)$, and $Z \sim \mathcal{N}(0, 1)$ is a standard normal scalar.

3.1 Three Observations about Gaussians

3.1.1 Observation A: A Gaussian projection only sees the length of x

Lemma 2. For any fixed vector $x \in \mathbb{R}^d$, $w^\top x \stackrel{d}{=} \|x\| Z$.

Proof. If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $c \in \mathbb{R}$, then $cX \sim \mathcal{N}(c\mu, c^2\sigma^2)$. Since x_i is a fixed number,

$$w_i x_i \sim \mathcal{N}(0, x_i^2).$$

A sum of independent Gaussians is Gaussian, and the means and variances add. Therefore:

$$w^\top x = \sum_{i=1}^d w_i x_i \sim \mathcal{N}\left(0, \sum_{i=1}^d x_i^2\right) = \mathcal{N}(0, \|x\|^2).$$

A $\mathcal{N}(0, \|x\|^2)$ variable has the same distribution as $\|x\|$ times a standard normal:

$$\boxed{w^\top x \stackrel{d}{=} \|x\| Z}$$

□

Interpretation. The random scalar $w^\top x$ depends on x only through its norm. The direction of x does not matter, because the Gaussian distribution of w looks the same in every direction (rotational invariance).

3.1.2 Observation B: The moment generating function of Z

Lemma 3. For every $t \in \mathbb{R}$, $\mathbb{E}[e^{tZ}] = e^{t^2/2}$.

Proof. Write the expectation using the density function $\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$:

$$\mathbb{E}[e^{tZ}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz - z^2/2} dz.$$

If we complete the square, we get:

$$tz - \frac{1}{2}z^2 = -\frac{1}{2}(z^2 - 2tz + t^2) + \frac{1}{2}t^2 = -\frac{1}{2}(z - t)^2 + \frac{1}{2}t^2.$$

The factor $e^{t^2/2}$ does not depend on z , we can extract it from the integral:

$$\mathbb{E}[e^{tZ}] = e^{t^2/2} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2} dz}_{\text{density of } \mathcal{N}(t,1), \text{ always integrates to 1}}.$$

Overall we conclude:

$$\boxed{\mathbb{E}[e^{tZ}] = e^{t^2/2}}$$

□

3.1.3 Observation C: The characteristic function of Z

Lemma 4. For every $t \in \mathbb{R}$, $\mathbb{E}[e^{itZ}] = \mathbb{E}[\cos(tZ)] = e^{-t^2/2}$.

Proof. The identity from Lemma 3 also holds for complex t : the integral converges for every complex t , and both sides are analytic in t . Substituting $t \rightarrow it$ and using $(it)^2 = -t^2$ gives

$$\boxed{\mathbb{E}[e^{itZ}] = e^{-t^2/2}}$$

Write $e^{itZ} = \cos(tZ) + i \sin(tZ)$. The distribution of Z is symmetric and $\sin(\theta)$ is an odd function, so $\mathbb{E}[\sin(tZ)] = 0$. Hence,

$$\mathbb{E}[\cos(tZ)] = e^{-t^2/2}.$$

This is what will allow only real-valued features later in the proof. □

Sanity check - At $t = 0$ we get $\mathbb{E}[e^0] = 1$. As $t \rightarrow \infty$ the expectation tends to 0. Intuition: e^{itZ} is a point on the unit circle at angle tZ . For small t the angles concentrate near 0, so the average is close to 1. For large t the angles spread around the whole circle and cancel out.

3.2 Building the Magic Function

3.2.1 Step 1: The Gaussian kernel as an expectation of a product

Lemma 5. For all $x, y \in \mathbb{R}^d$,

$$e^{-\|x-y\|^2/2} = \mathbb{E}_w \left[\cos(w^\top x) \cos(w^\top y) + \sin(w^\top x) \sin(w^\top y) \right].$$

This result shows each factor depends on only one of x or y . Therefore, The kernel is the expectation of a dot product of two separate feature vectors.

Proof. Apply Lemma 2 to the vector $x - y$:

$$w^\top (x - y) = \|x - y\| Z.$$

Then apply Lemma 4 with $t = \|x - y\|$:

$$\mathbb{E}_w[e^{i w^\top (x-y)}] = \mathbb{E}[e^{i \|x-y\| Z}] = e^{-\|x-y\|^2/2}.$$

Notice that the right-hand side is the *Gaussian kernel*. By Lemma 4 the imaginary part averages to zero, so

$$e^{-\|x-y\|^2/2} = \mathbb{E}_w[\cos(w^\top x - w^\top y)].$$

Now, by using $\cos(a - b) = \cos a \cos b + \sin a \sin b$:

$$\cos(w^\top x - w^\top y) = \cos(w^\top x) \cos(w^\top y) + \sin(w^\top x) \sin(w^\top y) = \begin{bmatrix} \cos(w^\top x) \\ \sin(w^\top x) \end{bmatrix}^\top \begin{bmatrix} \cos(w^\top y) \\ \sin(w^\top y) \end{bmatrix}.$$

□

3.2.2 Step 2: Replace the expectation with a finite average

Definition 6 (Random Fourier Features). *Draw $w_1, \dots, w_m \stackrel{iid}{\sim} \mathcal{N}(0, I_d)$ once, and reuse the same samples for every query and every key. Define the $2m$ -dimensional vector*

$$\varphi(x) = \frac{1}{\sqrt{m}} \begin{bmatrix} \cos(w_1^\top x), \sin(w_1^\top x), \dots, \cos(w_m^\top x), \sin(w_m^\top x) \end{bmatrix}^\top \in \mathbb{R}^{2m}$$

The map φ is called the Random Fourier Features of x .

Fact 7 (Hoeffding). *If $Y_1, \dots, Y_m$ are i.i.d. with values in $[a, b]$ and mean μ , then $\Pr \left[\left| \frac{1}{m} \sum_j Y_j - \mu \right| \geq \varepsilon \right] \leq 2 \exp(-2m\varepsilon^2/(b-a)^2)$.*

Theorem 8 (Rahimi–Recht). *For all $x, y \in \mathbb{R}^d$:*

1. $\mathbb{E}[\varphi(x)^\top \varphi(y)] = e^{-\|x-y\|^2/2}$;
2. $\text{Var}[\varphi(x)^\top \varphi(y)] = \frac{1}{2m} (1 - e^{-\|x-y\|^2})^2 \leq \frac{1}{2m}$;
3. $\Pr[|\varphi(x)^\top \varphi(y) - e^{-\|x-y\|^2/2}| \geq \varepsilon] \leq 2e^{-m\varepsilon^2/2}$.

In particular, by a union bound over all n^2 query–key pairs, $m = O(\log(n/\delta)/\varepsilon^2)$ samples suffice for every pair to be ε -accurate with probability $1 - \delta$.

Proof. The factor $\frac{1}{\sqrt{m}}$ in each vector multiplies to $\frac{1}{m}$ in the product, so if we write it as an average:

$$\varphi(x)^\top \varphi(y) = \frac{1}{m} \sum_{j=1}^m Y_j, \quad Y_j = \cos(w_j^\top x) \cos(w_j^\top y) + \sin(w_j^\top x) \sin(w_j^\top y) = \cos(w_j^\top (x - y)).$$

The Y_j are i.i.d., and by Lemma 5 each has mean $e^{-\|x-y\|^2/2}$. This proves (1); by the law of large numbers the average converges to the kernel.

To prove (2), let $t = \|x - y\|$, so $Y_j \stackrel{d}{=} \cos(tZ)$ by Lemma 2. Using $\cos^2 a = \frac{1+\cos 2a}{2}$ and Lemma 4 with $2t$,

$$\mathbb{E}[Y_j^2] = \frac{1 + e^{-2t^2}}{2}, \quad \text{Var}[Y_j] = \frac{1 + e^{-2t^2}}{2} - e^{-t^2} = \frac{(1 - e^{-t^2})^2}{2}.$$

Averaging m i.i.d. terms divides the variance by m , which gives (2).

For proving (3), we notice that each $Y_j \in [-1, 1]$, so by Fact 7 with $b - a = 2$ gives (3). Setting $2e^{-m\varepsilon^2/2} = \delta/n^2$ gives the bound on m . □

3.2.3 Step 3: From the Gaussian kernel to $e^{q^\top k}$

Corollary 9 (Trigonometric features for $e^{q^\top k}$). *Let*

$$f(x) = e^{\|x\|^2/2} \varphi(x) = \frac{e^{\|x\|^2/2}}{\sqrt{m}} \left[\cos(w_1^\top x), \sin(w_1^\top x), \dots, \cos(w_m^\top x), \sin(w_m^\top x) \right]^\top \in \mathbb{R}^{2m}$$

Then $\mathbb{E}[f(q)^\top f(k)] = e^{q^\top k}$, and

$$\text{Var} [f(q)^\top f(k)] = \frac{1}{2m} e^{\|q\|^2 + \|k\|^2} (1 - e^{-\|q-k\|^2})^2.$$

Proof. Step 2 approximates $e^{-\|q-k\|^2/2}$, but attention needs $e^{q^\top k}$.

First, expand the squared distance

$$\|q - k\|^2 = \|q\|^2 + \|k\|^2 - 2q^\top k.$$

Then, separate $q^\top k$

$$q^\top k = \frac{1}{2}\|q\|^2 + \frac{1}{2}\|k\|^2 - \frac{1}{2}\|q - k\|^2.$$

Exponentiate the equation,

$$e^{q^\top k} = e^{\|q\|^2/2} \cdot e^{-\|q-k\|^2/2} \cdot e^{\|k\|^2/2}.$$

The middle factor is the Gaussian kernel, which is approximately $\varphi(q)^\top \varphi(k)$. Each outer factor depends on one vector only, so it can be absorbed into that vector's features:

$$e^{q^\top k} \approx (e^{\|q\|^2/2} \varphi(q))^\top (e^{\|k\|^2/2} \varphi(k)) = f(q)^\top f(k).$$

Now, let's conclude the properties of this result. Since $f(q)^\top f(k) = e^{(\|q\|^2 + \|k\|^2)/2} \varphi(q)^\top \varphi(k)$ is a deterministic multiple of the estimator of Theorem 8, it means $e^{(\|q\|^2 + \|k\|^2)/2} e^{-\|q-k\|^2/2} = e^{q^\top k}$, and its variance is $e^{\|q\|^2 + \|k\|^2}$ times the variance in Theorem 8. □

3.2.4 Step 4: Positive random features (Performer)

The problem with Step 3.

- **Negative estimates.** The entries of f involve $\cos$ and $\sin$, which take negative values. An individual estimate $f(q)^\top f(k)$ can therefore be negative, even though $e^{q^\top k} > 0$ always. In the softmax normalizer $f(q)^\top \sum_j f(k_j)$ this can produce values near zero or below zero, which makes the output unstable.
- **Large error exactly where softmax is small.** By Corollary 9, the error is *additive* at scale $e^{(\|q\|^2 + \|k\|^2)/2}$, no matter how small the target $e^{q^\top k}$ is. In attention most entries are small (each query attends to few keys), so their estimates are dominated by noise. For example, if $k = -q$, the target is $e^{-\|q\|^2}$ while the standard deviation is about $e^{\|q\|^2}/\sqrt{2m}$.

The idea. Skip the characteristic function and use the moment generating function of Lemma 3 directly. Its integrand $e^{w^\top x}$ is always positive.

Theorem 10 (Positive random features, Performer). *With the same samples $w_1, \dots, w_m$ as in Step 2, let*

$$f_+(x) = \frac{e^{-\|x\|^2/2}}{\sqrt{m}} \left[e^{w_1^\top x}, \dots, e^{w_m^\top x} \right]^\top \in \mathbb{R}^m$$

Then every entry of f_+ is positive, $\mathbb{E}[f_+(q)^\top f_+(k)] = e^{q^\top k}$, and

$$\text{Var} [f_+(q)^\top f_+(k)] = \frac{1}{m} e^{2q^\top k} (e^{\|q+k\|^2} - 1).$$

Proof. As proved in Lemma 2 and Lemma 3, for any vector x , apply Lemma 2 and then Lemma 3 with $t = \|x\|$:

$$\mathbb{E}_w [e^{w^\top x}] = \mathbb{E} [e^{\|x\|Z}] = e^{\|x\|^2/2}.$$

Plug in $x = q + k$,

$$\mathbb{E}_w [e^{w^\top q} e^{w^\top k}] = e^{\|q+k\|^2/2}.$$

Then, if we expand the norm, since $\|q + k\|^2 = \|q\|^2 + \|k\|^2 + 2q^\top k$ we get,

$$\mathbb{E}_w [e^{w^\top q} e^{w^\top k}] = e^{\|q\|^2/2} e^{\|k\|^2/2} e^{q^\top k}.$$

If we isolate $e^{q^\top k}$ and divide both sides by $e^{\|q\|^2/2} e^{\|k\|^2/2}$ and move these constants inside the expectation:

$$e^{q^\top k} = \mathbb{E}_w \left[e^{w^\top q - \|q\|^2/2} \cdot e^{w^\top k - \|k\|^2/2} \right].$$

This is again an expectation of a product $g(q)g(k)$.

Then, if we conclude previous equation with the definition of $f_+(x)$ we get $f_+(q)^\top f_+(k) = \frac{1}{m} \sum_{j=1}^m X_j$ with $X_j = e^{w_j^\top (q+k) - (\|q\|^2 + \|k\|^2)/2}$, which are i.i.d. with mean $e^{q^\top k}$. So the estimator is unbiased.

Then if we set, $x = 2(q + k)$, $\mathbb{E}[X_j^2] = e^{-\|q\|^2 - \|k\|^2} e^{2\|q+k\|^2} = e^{\|q+k\|^2 + 2q^\top k}$, using $2\|q + k\|^2 - \|q\|^2 - \|k\|^2 = \|q + k\|^2 + 2q^\top k$. Hence,

$$\text{Var}[X_j] = e^{2q^\top k} (e^{\|q+k\|^2} - 1),$$

and averaging m i.i.d. terms divides it by m . □

Note. Every entry of f_+ is positive, so both the estimate and the normalizer are always positive. This version also needs only m features instead of $2m$. Moreover, its variance scales with $(e^{q^\top k})^2$, so the error is *relative* to the target and vanishes as $e^{q^\top k} \rightarrow 0$. In the example $k = -q$, the variance is exactly 0, while the trigonometric estimator has variance about $e^{2\|q\|^2}/(2m)$. This is why the positive features are the ones used in practice.

4 Conclusion

Softmax attention costs $\Theta(n^2d)$ because it forms the $n \times n$ matrix $e^{QK^\top}$. If we find a feature map with $e^{q^\top k} \approx f(q)^\top f(k)$, associativity lets us compute $f(Q)(f(K)^\top V)$ in $O(nmd)$ time, and decoding needs only an $O(md)$ state per step. Such features exist because Gaussian projections only see the norm (Lemma 2) and the Gaussian’s MGF and characteristic function have closed forms (Lemmas 3, 4). The characteristic function gives Random Fourier Features (Theorem 8, Corollary 9); the MGF gives the positive features of the Performer (Theorem 10), which are always positive and have relative rather than additive error.

5 Orthogonal Random Features [not covered in class]

Motivation. In Step 2 we drew $w_1, \dots, w_m$ *independently*. This is simple, but wasteful: two samples may point in almost the same direction and measure almost the same thing. *Orthogonal Random Features* (ORF) [5] make the samples exactly orthogonal, while keeping each one, on its own, a standard Gaussian. Unbiasedness depends only on the marginal of each w_j , so it is kept for free. For positive features the variance always goes down. ORF together with positive features is the Performer’s FAVOR+ mechanism [3].

5.1 The Construction

Definition 11 (Orthogonal Random Features). *Let $m \leq d$. Draw $G \in \mathbb{R}^{m \times d}$ with i.i.d. $\mathcal{N}(0, 1)$ entries and apply Gram–Schmidt to its rows to get orthonormal $u_1, \dots, u_m$. Independently draw $g'_1, \dots, g'_m \stackrel{iid}{\sim} \mathcal{N}(0, I_d)$ and set*

$$w_j^{\text{ort}} = \|g'_j\| u_j.$$

If $m > d$, use $\lceil m/d \rceil$ independent blocks.

Why rescale? Gram–Schmidt gives unit vectors, but a Gaussian vector is a uniform direction times a random length ($\approx \sqrt{d}$). Multiplying u_j by an independent Gaussian length $\|g'_j\|$ restores exactly the $\mathcal{N}(0, I_d)$ distribution, which Lemma 2 and Theorem 10 need for unbiasedness.

Lemma 12 (Same marginals). *Each $w_j^{\text{ort}} \sim \mathcal{N}(0, I_d)$. Hence the estimators of Corollary 9 and Theorem 10 remain unbiased with orthogonal samples.*

Proof.

- **(Directions are uniform).** For any rotation R , the rotated matrix $GR^\top$ has the same distribution as G , since its rows Rg_i are again i.i.d. $\mathcal{N}(0, I_d)$. Gram–Schmidt commutes with rotations: on $GR^\top$ it outputs $Ru_1, \dots, Ru_m$. Hence $(Ru_1, \dots, Ru_m) \stackrel{d}{=} (u_1, \dots, u_m)$, so each u_j is a unit vector whose distribution is unchanged by every rotation, i.e. uniform on the sphere.
- **(Adding the length).** A Gaussian vector g is a uniform direction $g/\|g\|$ times an independent length $\|g\|$. Since $w_j^{\text{ort}} = \|g'_j\| u_j$ has exactly this form, $w_j^{\text{ort}} \sim \mathcal{N}(0, I_d)$.
- **(Unbiasedness).** Each estimator is an average $\frac{1}{m} \sum_j X_j$, where X_j depends only on w_j . By linearity, its mean depends only on the distribution of each w_j , not on how they are coupled, so it is the same as with i.i.d. samples.

□

5.2 ORF Reduces the Variance of Positive Features

Intuition. With $x = q + k$, a positive feature $e^{w_j^\top x}$ is large when w_j is aligned with x . If w_1 is strongly aligned with x , then an orthogonal w_2 cannot be, since $(u_1^\top \hat{x})^2 + (u_2^\top \hat{x})^2 \leq 1$. So a large sample tends to be paired with a smaller one: the samples are negatively correlated and their errors partly cancel.

Fact 13. Let $g \sim \mathcal{N}(0, I_d)$. Then $v = g/\|g\|$ is a uniformly random unit vector (uniform on the sphere $S^{d-1} = \{v : \|v\| = 1\}$), independent of $\|g\|$. Moreover,

$$\mathbb{E}[\|g\|^{2r}] = \prod_{i=0}^{r-1} (d + 2i).$$

Proof. The density of g is proportional to $e^{-\|g\|^2/2}$, which depends only on $\|g\|$. Hence, conditioned on $\|g\| = r$, g is uniform on the sphere of radius r , so its direction v is uniform and its law does not depend on r . For the moments, each coordinate satisfies $\mathbb{E}[e^{tg_i^2}] = (1 - 2t)^{-1/2}$, so by independence

$$\mathbb{E}[e^{t\|g\|^2}] = (1 - 2t)^{-d/2} = \sum_{r \geq 0} \frac{t^r}{r!} \prod_{i=0}^{r-1} (d + 2i),$$

and $\mathbb{E}[\|g\|^{2r}]$ is the coefficient of $t^r/r!$. □

Lemma 14 (Moments on the sphere). Let v be a uniformly random unit vector in $\mathbb{R}^d$, and let v_1, v_2 be its first two coordinates. Then for all integers $a, b \geq 0$,

$$\mathbb{E}[v_1^{2a} v_2^{2b}] \leq \mathbb{E}[v_1^{2a}] \mathbb{E}[v_2^{2b}],$$

and for $a = b = 1$ the left side equals $\frac{d}{d+2}$ times the right side.

Intuitively, $v_1^2 + v_2^2 \leq 1$, so when one coordinate is large the other must be small.

Proof.

Take $g \sim \mathcal{N}(0, I_d)$ and $v = g/\|g\|$. Then $g_1 = \|g\|v_1$ and $g_2 = \|g\|v_2$, so

$$g_1^{2a} g_2^{2b} = \|g\|^{2(a+b)} v_1^{2a} v_2^{2b}.$$

The coordinates g_1, g_2 are independent, and $\|g\|$ is independent of v (Fact 13). Hence

$$\mathbb{E}[g_1^{2a}] \mathbb{E}[g_2^{2b}] = \mathbb{E}[\|g\|^{2(a+b)}] \mathbb{E}[v_1^{2a} v_2^{2b}].$$

The same argument with $b = 0$, and with $a = 0$, gives

$$\mathbb{E}[g_1^{2a}] = \mathbb{E}[\|g\|^{2a}] \mathbb{E}[v_1^{2a}], \quad \mathbb{E}[g_2^{2b}] = \mathbb{E}[\|g\|^{2b}] \mathbb{E}[v_2^{2b}].$$

Solving the three equations for the v -moments and dividing, the factor $\mathbb{E}[g_1^{2a}] \mathbb{E}[g_2^{2b}]$ cancels:

$$\frac{\mathbb{E}[v_1^{2a} v_2^{2b}]}{\mathbb{E}[v_1^{2a}] \mathbb{E}[v_2^{2b}]} = \frac{\mathbb{E}[\|g\|^{2a}] \mathbb{E}[\|g\|^{2b}]}{\mathbb{E}[\|g\|^{2(a+b)}]}.$$

By Fact 13, the numerator is $\prod_{i < a} (d + 2i) \prod_{i < b} (d + 2i)$, and the denominator splits as $\prod_{i < b} (d + 2i) \prod_{i < a} (d + 2b + 2i)$. Cancelling the common factor,

$$\frac{\mathbb{E}[v_1^{2a} v_2^{2b}]}{\mathbb{E}[v_1^{2a}] \mathbb{E}[v_2^{2b}]} = \prod_{i=0}^{a-1} \frac{d + 2i}{d + 2b + 2i} \leq 1,$$

The inequality holds because each factor has a larger denominator. In the case $a = b = 1$, the product has a single factor, $\frac{d}{d+2}$. $\square$

Theorem 15. *Let $m \leq d$, and let $\hat{X}_{iid}$, $\hat{X}_{ort}$ be the positive-feature estimator $f_+(q)^\top f_+(k)$ built from i.i.d. and from orthogonal samples. Both are unbiased for $e^{q^\top k}$, and*

$$\text{Var}[\hat{X}_{ort}] \leq \text{Var}[\hat{X}_{iid}] - \frac{m-1}{m} \cdot \frac{e^{-\|q\|^2 - \|k\|^2} \|q+k\|^4}{2(d+2)}.$$

Proof. Write $\hat{X} = \frac{1}{m} \sum_j X_j$ with $X_j = c e^{w_j^\top x}$, where $x = q+k$ and $c = e^{-(\|q\|^2 + \|k\|^2)/2}$. By symmetry all pairs (X_i, X_j) , $i \neq j$, have the same covariance, so

$$\text{Var}[\hat{X}] = \frac{\text{Var}[X_1]}{m} + \frac{m-1}{m} \text{Cov}(X_1, X_2).$$

By Lemma 12, $\text{Var}[X_1]$ is the same in both cases, and the covariance is 0 for i.i.d. samples. So it suffices to show $\text{Cov}(X_1, X_2) \leq -\frac{c^2 \|x\|^4}{2(d+2)}$ for orthogonal samples. Let $A = w_1^\top x$ and $B = w_2^\top x$. Since $\mathbb{E}[X_j] = c e^{\|x\|^2/2}$, we have $\text{Cov}(X_1, X_2) = c^2 (\mathbb{E}[e^{A+B}] - e^{\|x\|^2})$, so it suffices to show

$$\mathbb{E}[e^{A+B}] \leq e^{\|x\|^2} - \frac{\|x\|^4}{2(d+2)}.$$

Let $\hat{x} = x/\|x\|$ (assume $x \neq 0$; otherwise there is nothing to prove) and $s_j = \|g'_j\|$, so $w_j = s_j u_j$. By rotation invariance, $(u_1^\top \hat{x}, u_2^\top \hat{x}) \stackrel{d}{=} (v_1, v_2)$ for v uniform on the sphere, independent of s_1, s_2 . Hence $A \stackrel{d}{=} s_1 \|x\| v_1$ and $B \stackrel{d}{=} s_2 \|x\| v_2$. Expanding $e^{A+B} = \sum_{i,j} A^i B^j / (i! j!)$, the terms with i or j odd vanish, because $(v_1, v_2) \mapsto (\pm v_1, \pm v_2)$ preserves. For $i = 2a$, $j = 2b$, each remaining term is

$$\frac{\|x\|^{2a+2b}}{(2a)!(2b)!} \mathbb{E}[s_1^{2a}] \mathbb{E}[s_2^{2b}] \mathbb{E}[v_1^{2a} v_2^{2b}].$$

For i.i.d. samples the same expansion holds with $\mathbb{E}[v_1^{2a}] \mathbb{E}[v_2^{2b}]$ in place of $\mathbb{E}[v_1^{2a} v_2^{2b}]$, and it sums to $\mathbb{E}[e^A] \mathbb{E}[e^B] = e^{\|x\|^2}$. By Lemma 14, every orthogonal term is at most its i.i.d. counterpart. The $a = b = 1$ term alone is smaller by

$$\frac{1}{4} \left(1 - \frac{d}{d+2}\right) \|x\|^4 = \frac{\|x\|^4}{2(d+2)},$$

using $\mathbb{E}[A^2] \mathbb{E}[B^2] = \|x\|^4$ for i.i.d. samples. Summing over all terms gives $\mathbb{E}[e^{A+B}] \leq e^{\|x\|^2} - \frac{\|x\|^4}{2(d+2)}$, as required. $\square$

5.3 Consequences

1. **A free improvement.** ORF changes only how the w_j are drawn. The feature map, the linear-attention algorithm, and the costs of Section 2.1 are unchanged, and the estimator stays unbiased (Lemma 12). By Theorem 15, the variance can only decrease, and it strictly decreases whenever $m \geq 2$ and $q+k \neq 0$.
2. **The gain is largest for small head dimension.** The improvement scales like $1/d$. Attention heads are small ($d = 64$ or 128), which is exactly where ORF helps most.

3. **It helps where positive features are weakest.** By Theorem 10, the variance of positive features grows with $\|q + k\|$. The reduction in Theorem 15 also grows with $\|q + k\|$ (as $\|q + k\|^4$), so ORF targets the hardest query–key pairs.
4. **More features than dimensions.** For $m > d$, samples in different blocks are independent. Only pairs within the same block are negatively correlated, so the benefit per pair is the same, but a smaller fraction of pairs gets it.
5. **Cost of drawing.** Gram–Schmidt costs $O(md^2)$, paid once when the features are drawn (or redrawn periodically during training), not per token.

References

- [1] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *NeurIPS*, 2007.
- [2] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear-attention. *ICML*, 2020.
- [3] K. Choromanski et al. Rethinking attention with Performers. *ICLR*, 2021.
- [4] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *NeurIPS*, 2022.
- [5] F. X. Yu, A. T. Suresh, K. Choromanski, D. Holtmann-Rice, and S. Kumar. Orthogonal random features. *NeurIPS*, 2016.