We thank the reviewers for their helpful suggestions and clarifying questions. In summary, our work makes theoretical progress on the challenging and practically important question of oblivious kernel embeddings. It leaves open some interesting questions which we would be excited to work on or see solved by other researchers. As the reviewers note, the extra $\sqrt{s}$ in our Gaussian kernel upper bound and extensions to higher dimensions are particularly enticing.

Responses to specific points are below.

Extension to higher dimensions: Our results can be directly extended to higher dimensions, but with an exponential in dimension cost. We believe this is unnecessary, and proving such a result is a major open question. [AKK+20b] gives a dimension independent result, but with an additional dependence on the data set radius. The goal would be to remove the radius dependence and only depend on the statistical dimension s_λ.

Reviewer 1:

The authors mention preconditioning/spectral approximation as a motivation...some experiments that perform linear algebra based on preconditioning methods would offer more convincing evidence...

See Figure 4a. We show faster convergence in solving a standard kernel ridge regression problem with preconditioned CG based on our modified RFF method vs. the traditional RFF method.

How does the equation after line 518 follow? Lemma 7 bounds...

Thanks very much for pointing out this confusion. We will clarify the discussion and some typos in the paper.

First, note that Lemma 6 is given as Theorem 7.1 in [Erd17] with an additional factor 2 in it. This factor does not significantly affect any results, and thus we could just use this theorem directly. We also seemed to have dropped a factor of two from [BE06] in Lemma 7 – thus, in retrospect we should have just cited [Erd17] in the first place.

To clarify our own proof: as pointed out by the reviewer, Lemma 7 should be stated with $\|f\|_{L_\infty[a+\delta,b-\delta]}$ in place of $\|f\|_{[a+\delta,b-\delta]}$ as in [BE06]. Then, to prove Lemma 6 (focusing on the first bound since the second is symmetric), we set $\delta = b - x$ and thus have:

$$\frac{|f(x)|^2}{\|f\|^2_{[a,b]}} \leq \frac{\|f\|^2_{L_\infty[a+b-x,x]}}{\|f\|^2_{[a,b]}} = \frac{\|f\|^2_{L_\infty[a+\delta,b-\delta]}}{\|f\|^2_{[a,b]}} \leq \frac{s}{\delta} = \frac{s}{b-x}.$$

Note that in the first inequality, we use $a + b - x \leq x$ which follows from the assumption in this case that $x \geq (a + b)/2$.

Reviewer 2: We hope that the $\sqrt{s}$ gap in the Gaussian density leverage score bound can be closed. This is an exciting open question. Note that prior to our work, no polynomial in s bound was known.

Reviewer 3:

The proposed approach to sampling random features implicitly assumes that the λ-statistical dimension is known...

This is a good point – in theory s_λ needs to be known. In practice simple heuristics seem to suffice – see Appendix E of the supplement for more details on our implementation.

Theorem 3: what is the form of the embedding? The theorem only claims its existence...

See Corollary 27 in the appendix for a full statement of Theorem 3, which gives the explicit random features construction. This also explicitly discusses and formally proves the post-processing via random projection step, which the reviewer asks about. We will try to expand discussion in the main body, subject to space constraints.

In regards to the connection with Towards A Unified Analysis of Random Fourier Features: our results can be directly plugged into that work, which nicely applies to any method that samples random features with an upper bound on the ridge leverage scores. That work gives a leverage score approximation method but 1) it is not oblivious and 2) it has a $1/\lambda$ dependence, which for small λ can be worse than our s_λ dependence.

Reviewer 4: In regards to obtaining tighter upper bounds, see our response to Reviewer 2. We do not yet have an approach to closing the $O(\sqrt{s})$ gap but are working on it.

The authors mention that they use kernel approximation as a preconditioner to accelerate the iterative solution of the original problem. They say they get similar empirical results as [AKM+17] did by using approximation in place of K...

To clarify, our embedding method (not results) are very similar to that of [AKM+17]. Due to this similarity, we test a different method to solve linear systems, using a preconditioner rather than a direct approximation (as AKM+17 already tests the direct approximation approach). These methods are somewhat incomparable. Preconditioning leads to much higher accuracy at the expense of possibly slower runtimes when n is large. Either embedding method can be used with either system solving approach.