We thank the reviewers for their detailed and constructive feedback.

R1: *It is not clear… how the assumption of bounded gradients…:* We would like to point out that the assumption of bounded gradient is only employed in the theoretical analysis of the bias. Such assumptions are quite common in the analysis of ML algorithms (see e.g., [Hazan and Kale, 2014]). As mentioned in our submission, in many settings, this holds because of the clipping of the gradients (see e.g., [Goodfellow, Bengio and Courville, 2016]).

Same as above for the normalization of the class embeddings and input embeddings…: As discussed in lines 167-174 of our submission, it is correct that we assume the embeddings to be normalized for RF-softmax, however, such normalization is widely used in practice (e.g., see references [26], [27], and [28]). Furthermore, we have empirically shown (on both NLP and extreme classification datasets) that with proper setting of τ, the normalized embeddings do not degrade the final performance of the model.

R2: *It feels that there is a skip between Sections 2 and 3 …:* We thank the reviewer for the constructive comment. We will enhance the presentation by better motivating the kernel-based sampling to approximate exponential families (i.e., RF-softmax method and its analysis) and smoothing the transition from section 2 to section 3. We have motivated the use of random fourier features by comparing its performance against two most natural candidates in Table 1. We will improve the relevant discussion in the final version.

The paper perhaps lacks a discussion on approaches … such as minimization of Fisher divergence … an approximation scheme for log-partition function was considered in [2, Section 2.3]…: We thank the reviewer for the suggestion. In [Hyvarinen 2005], the partition function Z is just the function of model parameter and thus disappears in scoring function. However, in our case, the partition function depends on the input h (which changes during the training). Therefore, while calculating the score function (taking derivative of Z with respect to (h, c)), the partition function has a non-trivial contribution. As for [Vembu et al., 2009], given ways to generate uniform samples for the set of classes, they propose a MCMC approach to sample a class with a distribution that is close to full softmax distribution. Such methods do not come with precise sample/computational complexity guarantees. We will include a discussion in the final version.

The notation is not very clear, especially in the appendix…: We will highlight the distribution/random variables with respect to which we take the expectations. Also, we will try to eliminate any inconsistency/ambiguity regarding the notation elsewhere in the paper.

In Eq. (5), I was not clear for the motivation of m until checking the appendix and reading the whole paper…: Adjusting the logits for negative classes using their expected number of occurrence [Bengio Senecal, 2008] is critical for the unbiasedness of sampled softmax loss, e.g., it ensures that Z' is an unbiased estimator of Z. We will add a comment to clarify the process of adjusting the logits in (5).

Related work can also be improved and relevance of the work in the context of extreme classification…: We will highlight relevant papers in extreme classification literature in the final version.

R3: *What do you get by applying Theorem 1 to RF-softmax …:* As pointed out in the discussion following Theorem 1, the result provides a guidance for selecting a sampling distribution with low bias (by highlighting the requirement of tight multiplicative approximation). We will combine Theorem 1 and 2 (at least in the setting with large D) in the final version to obtain the bounds for RFF.

Theorem 2 and Remark 2: Where is the dependence on D coming from in $o_D(1)$? …: Thanks for the comment. We will include a comment on how γ_2 depends on γ_1. As it's clear from the proof of Theorem 2 and the statement of Remark 2, one can choose $\gamma_1 = const\sqrt{(d \log D)/D}$. Now γ_1 (and thus γ_2) scales as $o_D(1)$ (while keeping other parameters fixed).

What happens if the inner product $\phi(c_i)^T\phi(h)$ is negative?: With normalized embeddings, for finite ν, $e^{(\nu h^T c_i)}$ is strictly positive. Therefore, for reasonably large D, with high probability, $\phi(c_i)^T\phi(h)$ should be non-negative. In general, one can replace $\phi(c_i)^T\phi(h)$ with $max(0, \phi(c_i)^T\phi(h))$ (with minor modifications in the proofs).

How does RF-softmax compare to hierarchical softmax…: As discussed in [Balnc Rendle, 2018] and references therein, in many tasks, the final solution of hierarchical softmax is worse than those of both full-softmax and sampled-softmax.

"q_j should provide a tight uniform multiplicative approximation of e^{o_j}." How tight is the bound…: We believe that the bounds presented in Theorem 1 are fairly tight. In particular, these bounds recover the unbiasedness of the gradient for full softmax distribution.

It would be good to show the calculation that verifies (15): (15) follows from the existing literature on RFF. We plan to include a citation to [Yu et al. 2016, Lemma 1].

We will fix all the typos pointed out by the reviewers and eliminate other remaining typos in the final version.