We thank the reviewers for their careful reading and constructive comments. We feel that the reviews are largely positive. In the remainder, we want to address some of the issues raised, and we will address them in detail in the revision. We will also release the code for the numerical experiments.

"Why would one care about the variable sketches, rather than the fixed sketch", "It's fine if this paper's method and analysis is not the best for chronological and historical reasons, however please state of the art clear and provide a reference." *This reference appeared after our work. It is based on the IHS with a fixed SRHT embedding, and their convergence rate appears currently the best known (in the asymptotic sense) for the SRHT. Differently, we emphasize that our analysis provides an* ***exact and closed-form formula of the convergence rate****. From a practical standpoint, their optimal fixed sketch algorithm critically relies on a momentum term, and this can be sensitive to noise and rounding. Without momentum, their fixed sketch algorithm has worse convergence rate than ours.*

"Did you experiment applying the truncated Walsh-Hadamard transform when using SRHT ?", "No numerical comparisons against state-of-the-art randomized methods are given.", "Could you [...] empirically investigate how [many refreshing steps] could be skipped?" *We'll include these additional experiments in the final version, and comparisons to state-of-the-art methods. In particular, our algorithm is faster than the pre-conditioned conjugate gradient method [24]. Further, it's more robust to noise in the gradients compared to the aforementioned fixed sketch algorithm. Although we do not have theoretical guarantees for skipping refreshing the embedding, we observe in practice that refreshing at each iteration can be omitted at the cost of convergence rate. Detailed numerical results will be included.*

"Practical performance improvement by using orthogonal transforms is slight", "The benefit of orthogonal occurs as $\xi \to 1$ [but we] are interested with ξ as small as possible, close to γ.", "[The algorithm] might be slower [when] ε cannot really be treated as a constant." *We emphasize one of our key findings, that is, the SRHT has the remarkable benefit of a fast projection method compared to Gaussian embeddings, along with always improving the convergence rate. For $\gamma = 0.1$, $\xi = 0.5$, the limiting convergence ratio between ρ_h and ρ_g is about 60%. So orthogonal still has benefit. Even with ξ being close to γ, the ratio is still strictly less than one. Indeed, we improve theoretical time complexity when ε is treated as a constant, which is a reasonable setting, and we will discuss more general cases in the final version.*

"Could you please precisely give the references claiming that $m \approx d\log(d)$ is prescribed for state-of-the-art algorithms and for which algorithm?" *Up to constant factors, the authors of [24] originally prescribed $m \geq d^2$ (see Lemma 1). Improved concentration bounds on the SRHT [27] can be used to improve this lower bound to $d \log d$. See also [6], Thm 3.1, where the bound $m \geq C \log d[\sqrt{d} + \sqrt{\log n}]^2$ is stated.*

"'For SRHT, we use the optimal step sizes'. I thought it was not proven to be optimal for SRHT. Did I miss something?" *We will make clearer that this step size is optimal conditional on $\beta = 0$.*

"A comment on how to deal in the case where n is not a power of 2." *One can use padding with zeros, which increases the value of n. This slightly increases the convergence rate. Or one can take a random subset of coordinates of SRHT, which empirically does not increase convergence rate, but is somewhat slower to compute.*

"How was equation (2) with inverse of sketched Hessian solved?" *In practice, the fastest method is to solve approximately the linear system with an iterative solver such as conjugate gradient.*

"Can this analysis be also extended to accommodate count-sketch or any other sparse sketching matrices?" *We rely on recent results in random matrix theory (RMT) which, to our knowledge, have not been derived for sparse embeddings. But there is recent work that analyzes both SRHT/Haar and some sparse embeddings in the same asymptotic framework, for PCA: arxiv.org/abs/2005.00511. It may work here but possibly with stronger assumptions on the data.*

"The SRHT in this paper incorporates an additional permutation.", "The optimality of IHS is only proved for Haar transform." *We do consider this additional permutation as we leverage recent results from the RMT. We cannot think of drawbacks of the extra permutation (computational or otherwise). For proving optimality for the SRHT, we would need results currently unknown in the RMT. Please see Remark A.1 for more details.*

"'SRHT [...] contains less randomness, but is more structured and faster to generate' than Haar matrix." *SRHT relies on a random permutation and n sign-flips, while constructing Haar matrices usually needs order n^2 random Gaussian variables. Thanks to the matrix decomposition of the SRHT, multiplications are faster to perform via FFT.*

"Analysis is asymptotic and holds in the infinite limit only.", "... not very standard for sketching results", "Paper assumes that the ratio $d/n = \gamma$ is fixed.", "The language of the paper oversells things [...] without qualifying that everything is asymptotic." *The asymptotic framework is a good fit as one will only use sketching when the dataset is large. It also leads to clean theoretical results. Finite-sample results sometimes hide large constants. Our results are down to the constant and thus can be used easily by practitioners. The asymptotic result is also powerful enough to illustrate that Hadamard projection is superior to Gaussian projection. Moreover, the asymptotic results agree with simulations well with a few thousands samples. We will make our claims more precise and use asymptotically optimal wherever needed.*