We thank the reviewers for their careful reading and constructive comments. We will include their suggestions in the final version, and we will release our code. In the remainder, we want to address the main points raised in the reviews.

"It is not true that this is the first algorithm where the sketch-size depends only on the effective dimension [reference]." Thank you for pointing us this reference. We will include a detailed comparison. However, we believe that our work provides the first sketching algorithm with sketch size no larger than the effective dimension and *without a priori knowledge or estimation of the latter*. In the reference, Algorithm 2 takes as input a sketch size $m \approx d_e \log d_e$, which requires estimation of d_e. The latter can be done efficiently, but in the restricted setting $d_e \leq (n+d)^{\frac{1}{3}}/\text{poly}(\log(n+d))$ (please see, for instance, Theorem 60 in [2]). We believe that our method is more simple, and, has the advantage of starting with an arbitrarily small sketch size. As illustrated in Figure 1, the sketch size can also remain smaller than $\mathcal{O}(d_e/\rho)$. Lastly, our method applies to both the overdetermined and undetermined cases (please see Appendix B for the latter).

"The guarantee [of Theorem 7] can be trivially found [using CG with a randomized preconditioner]." We emphasize that our key contribution is to propose an adaptive sketch size algorithm. Using the IHS, we are able to monitor the progress of our algorithm and adapt the sketch size accordingly. Extending such ideas to different methods based on preconditioning + iterative refinement is an open question beyond the scope of our work.

"The algorithm [...] doesn't exploit the sparsity of the data." Although we aim to address the case of dense data matrices which is standard in the literature, our method can be extended to sparse embeddings, for which similar concentration bounds exist in the literature. We will include these additional results in the final version.

"Many hyperparameters.", "Choice of step sizes?", "Choice of target improvement rate c^*?", "K might be very large if η gets close to 1 or ρ to 0" We emphasize that ρ and η are the only users' choice parameters, which will be clarified in the revision with suggested values. Other hyperparameters are chosen as in Theorem 5, and specified by the values of ρ and η. One should choose a small η for the concentration bounds to be tighter, and a typical value which preserves a small failure probability is $\eta = o(1/\sqrt{m})$. If one picks a small ρ to get a fast convergence rate, then, to avoid many rejection steps (which cannot exceed $\mathcal{O}(\log(d_e/\rho))$ according to Theorem 5), one can either choose a larger initial value of m, or, multiply m by a constant larger than 2 at each rejection. In numerical experiments, we chose $\rho = 0.2$ for MNIST and $\rho = 0.5$ for CIFAR10, and $\eta = 0.05$ which results in at most 5 sketch size rejections.

"[Unclear] benefits of switching between GD and Polyak steps", "Is it really advantageous to run Gradient-IHS and Polyak-IHS in parallel?", "Can we have a graph that compares the best methods?". We emphasize that the Polyak-IHS update is guaranteed to perform at least as well as the gradient-IHS update, at the expense of just one gradient computation but theoretically guaranteed convergence. We have carried out additional numerical comparisons of the time and flops of running both in parallel, for both Gaussian and SRHT embeddings. In a nutshell, we typically observe that either most Polyak steps are accepted, or, most of them are rejected, and this holds for both embeddings. Based on all our numerical evaluations, the hybrid method with SRHT embeddings is most often the most efficient both in terms of time and memory (RAM) usage as well as provably convergent with the specified sharp rate.

"How does the sketch size evolve during iterations?", "Could you compare with IHS with fixed sketching dimension chosen heuristically and in hindsight?" Figures 1.(e-h) show how the adaptive sketch size changes throughout the algorithm. Importantly, it can remain much smaller than the effective dimension. Without adaptation, for small sketch sizes, the IHS fails to converge. We will include a detailed numerical comparison.

"Is the preconditioning [of pCG] done from scratch for each value of the penalty ν?" We leverage previous iterates for the preconditioning as we move along the regularization path.

"In Figure 3, [are there] error bars?" Error bars are reported on Figures 3(a–d). We will make them more readable. *"How does $a(\rho,\eta)$ appears in the bounds when applying Theorem 3 in the proof of Theorem 5."* In the proof of Theorem 5, we apply Theorem 3 with m greater than $d_e\, a(\rho,\eta)/\rho$. We will make these details clearer.

"It seems to me that theoretical claims no longer hold for [the improvement ratio] C_t." Our algorithm is guaranteed to converge by monitoring the ratio c_t, although there is indeed a gap between the C_t and c_t, which is controlled by the condition number of the matrix C_S. Consequently, we pay an additional factor $(1+\sigma_1^2/\nu^2)$ in the convergence guarantee. Please see the proof of Theorem 5 for more details.

"Formulas for d_e in lines 63 and 83 are different" We define formally d_e on line 83. We will clarify this.

"Similar theorems are already known when d_e is replaced with d." We will highlight differences between our analysis techniques which result in sharper bounds and existing ones.

"Precise reference for [...] $m = d\log(d)/\rho$ for p-CG ?" Given a target convergence rate ρ, Lemma 1 in [24] specifies $m = d^2/\rho$ for the SRHT. Tighter (and more recent) concentration bounds on the SRHT (see Lemma 3.4 [26]) suggest to use $m = d\log d/\rho$.