We thank all the reviewers for their feedback and suggestions. We first address some commonly raised points:

1. Computing the effective dimension d_λ (raised by Reviewers 1 and 2). In the case of the surrogate leverage score sampling sketch, the effective dimension does not need to be explicitly calculated. In particular, this means that the input sparsity time algorithm from Theorem 2 does *not* require any knowledge of d_λ (this applies to both sketching and scaling). When using scaled regularization with a standard sketch (e.g., Gaussian), a rough estimate of d_λ is sufficient to significantly reduce the bias. We will discuss this in detail in the final version of the paper.

2. The assumption that $\alpha \geq d$ (raised by Reviewers 2 and 4). As discussed in Remark 11, the assumption that $\alpha \geq d$ in Theorem 2 (and Table 1) is introduced *only* to account for the additional computational cost of the surrogate sketch when comparing to prior work (the cost is related to DPP sampling), but is not needed for the convergence rate. In fact, unlike in Determinantal Averaging [10], our convergence rate bound holds for any $\alpha \geq 1$, and it is by a factor of d better than in [10] *for all* α (Table 1). It is worth pointing out that time complexity of DPP sampling is an ongoing area of research, and we believe that further improvements in this area will eliminate the need for the $\alpha \geq d$ assumption.

3. Condition number dependence of the convergence rate result (raised by Reviewers 3 and 4). Dependence on the condition number is standard in the analysis of distributed Newton-type methods, including GIANT [25], Determinantal Averaging [10], DANE [22], AIDE [21], DiSCO [26] and others. Our approach enjoys a merely logarithmic dependence of the number of iterations on the condition number (see Remark 3), which is at least as good as any of the listed methods. Following Reviewer 3's suggestion, we empirically compared the convergence rate of distributed Newton using the surrogate sketch (on the logistic regression task from Figure 3a) to a full unsketched gradient descent (GD) with backtracking line search, observing that GD required over 3x more iterations, compared to our method, until reaching 10^{-9} relative function value error. We will add these additional experiments to the final version.

Reviewer 1

- **Comparison of [25] and our work:** The main difference between [25] and our work is that our method gives unbiased estimates for the Newton step, which is not the case for [25]. This makes the most difference in the high q (many workers) and high accuracy (small ϵ) regime, which is important in massively parallel computing and federated learning. Also, from a practical point of view, simply applying rescaled regularization to the method of [25] (without using surrogate sketching) improves the error performance of [25] greatly, as illustrated in Figure 1.

Reviewer 2

- **Does a result analogous to Lemma 8 hold for other surrogate sketches?** Yes, we believe that a version of Lemma 8 can be shown for any surrogate sketch which admits a spectral approximation of $\mathbf{A}^\top\mathbf{A}$ (most of our proofs extend naturally to this general case, however some technical difficulties remain).
- **Advantage over using a single server with a preconditioner:** When we desire an accurate solution (i.e., small ϵ), then the cost of performing $\log(1/\epsilon)$ iterations over all data can be very significant. By averaging the outputs of multiple workers, we can obtain better estimates for the Newton direction, and thus require far fewer than $\log(1/\epsilon)$ iterations over all data (see Remark 3).
- **Difference between this work and [9]:** While our proof techniques for showing the expectation formulas in Lemma 4 are indeed based on [9], the scaled regularization phenomenon is specific to our analysis ([9] did not observe this scaling). Furthermore, unlike us, [9] considers neither the algorithmic nor the measure concentration aspects of determinantal distributions. For example, they do not have an analogue of Lemma 8.
- **What is capital** E **in Definition 2?** It is any event measurable with respect to the random (matrix) variable $\mathbf{X}$.
- **Is a surrogate sketch for a general random projection useful?** In terms of sampling from, say, a surrogate Gaussian sketch, we expect that this can be done efficiently (e.g., [12] shows how to sample from a slightly different determinant-rescaled Gaussian distribution). Furthermore, we believe that by bounding the discrepancy between a standard sketch and its surrogate, it will be possible to bound the bias of the standard sketch.

Reviewer 3

- **Linear/non-linear models:** We would like to clarify that the model is not necessarily linear but the update direction is obtained through solving a linear system. Sketched Newton's method works for convex cost functions in general, and our surrogate sketches apply there as well. For instance, Figure 7 in the appendix shows the error curves for distributed Newton sketch applied to a convex objective formulated using the log-barrier method.

Reviewer 4

- **Comparison of the least squares case and the general convex case:** The convergence rate in Theorem 2 (least squares case) is stated in terms of the ℓ_2 norm *squared*, whereas Theorem 10 (general convex case) is stated in terms of the ℓ_2 norm (not squared). Squaring the expression in Theorem 10 shows that the dependence on κ is the same in both cases (also, it matches [10] and [25]). We will clarify this in the final version.