Dear reviewers, thank you for taking the time to review our paper. We have addressed your main questions in **A1**, **A2** and **A3**, and your remaining questions below. We especially thank **Rev1** for his/her thoughtful and encouraging remarks. All issues raised are easy to address. We will incorporate all of your suggestions.

A1: Rank assumption on the Hessian. Corollary 1 gives an example where the Hessian need not have full rank to satisfy our assumptions. Indeed whenever the sketching matrix $\mathbf{S}_k = s_k \in \mathbb{R}^d$ is a column vector, eq (9) and (20) hold trivially so long as $s_k^\top \mathbf{H}_k s_k \neq 0$. This can hold for rank deficient Hessian matrices for instance when the diagonal has no zero elements and when s_k are random unit coordinate vectors.

A2: Novelty related to Pilanci/Wainwright's work. There are many key differences between RSN and Newton Sketch (NS) [20]. First, they are simply different algorithms. The sketching method underlying NS relies on having at hand the square root of the Hessian. In contrast, RSN uses a random subspace constraint to sketch the Hessian and thus needs no square root. Furthermore, NS requires full gradient and function evaluations, while RSN only needs a sketched gradient and requires no function evaluations. The convergence proof of NS requires a sketch size proportional to ϵ^{-2}, an unknown universal constant and global spectral properties of the Hessian; see equations (12) and (19) in [30]. Thus this required sketch size could be as large as d (or larger, which makes the results vacuous). This is because the theory of NS builds upon the theory of "one shot" sketching techniques. Furthermore, they do not establish linear convergence rates[1]. In contrast, we establish linear convergence which can hold in the rank deficient case and for every sketch size. We achieve this by entirely bypassing the theory of one shot sketches, showing it is not at all necessary. This in turn gives us the freedom of choosing the sketch size arbitrarily and allows us to apply RSN to large scale problems no matter how large the dimension. We will include this discussion in the paper.

A3: Bounding $0 < c \leq \rho \leq 1$. The bound $\rho \leq 1$ follows from Lemma 7 since $\rho(x_k) \leq 1$ for all k. We can guarantee that ρ is bounded away from zero $\rho \geq c > 0$ if we use the common assumption that $f(x)$ is L–smooth and m–strongly convex. This follows under the conditions of Lemma 7 since: $\rho(x) \geq m\lambda_{\min}^{+}\left(\mathbb{E}\left[\mathbf{S}(\mathbf{S}^\top \mathbf{H}(x)\mathbf{S})^\dagger \mathbf{S}^\top\right]\right) \geq \frac{m}{L}\beta$, where $\beta := \lambda_{\min}^{+}\left(\mathbb{E}\left[\mathbf{S}(\mathbf{S}^\top \mathbf{S})^\dagger \mathbf{S}^\top\right]\right)$. The right-hand side is a fixed positive constant independent on x, thus $\rho \geq \frac{m}{L}\beta > 0$. We can even relax the strongly convex assumption since only $\lambda_{\min}^{+}(\mathbf{H}(x))$ needs to be uniformly bounded away from zero (the spectral gap must be lower bounded). Furthermore, β is known for many distributions, e.g. for Gaussians and the family of randomized orthogonal sketches (Section A.1 in [12]) we have $\beta = \frac{s}{d}$, where s is the sketch size and d the dimension. Thus $\rho \geq \frac{m}{L}\frac{s}{d}$ and ρ is at least linearly increasing in s. We will now include this lower bound.

Rev2. *Theorem 2 is not surprised ... This has been stated in [20] as an inexact Newton method.* Our RSN method is not an inexact Newton method since we do not need to guarantee that the quadratic upper bound is minimized to within a given accuracy threshold. In no way has RSN been stated/analysed in [20].

Q1. Assumption (17) ... seems to be too strong. For convex functions, this assumption is equivalent to f being lower bounded, which is a trivial assumption, since otherwise f is a linear function.

Q2. Eq (52) and Eq (76). Thank you, we have fixed the squared norms and (76) should be an inequality.

Q3. Same assumptions as [6, 28]. Why not compare? S-Newton in [6] is based on subsampling and has no dimension reduction: it is targeted at large n and small d. The opposite setting of RSN. Also, subsampling and be applied in conjunction with our technique. The method in [28] is for solving constraint linear least square, not general optimization smooth and convex optimization.

Rev3. *"Newton should be q-quadratic. Therefore ... not super impressive."* There exists only semi-local quadratic convergence for Newton based methods. For global convergence, linear rates are as good as it gets.

Q1. *Comment about parameter tuning.* We apologize, but we did not understand your question/comment.

Q2. *Where is the proposed line search strategy.* It is in Algorithm 3 in the supp. material as stated on lines 242–243 of the main paper. Our line search does not require function evaluations, but only sketched gradients and sketched Hessian. Since the sketched Hessian is already available from the RSN update, our line search is computationally much cheaper than the standard Armijo method.

Q3. *Assumption 2 seems too strong.* Assumption 2 does not hold for $x_1^2 + \text{huber}_1(x_2)$ but neither is this a twice differentiable function, thus one cannot apply Newton type methods. Assumption 2 is necessary to guarantee that the Newton direction is well defined, see Lemma 9.

Q4. *The assumption does not hold for generalized linear models if ...* This assumption holds for all convex generalized linear models independently of the rank of $\mathbf{A}$ and the regularization parameter. This follows from examining the gradient and Hessian in (77) and (78) in the supp material and using standard linear algebra results such as Lemma 10. We will clarify this point and include the proof of this claim in the supp material.

Q5. *Theorem 2: Is this not the usual gradient descent rate?* Please see lines 219–225. In particular, Theorems 2 and 3 rely on relative smoothness and convex assumptions. Under these assumptions, it is not known if gradient descent converges.

Q6. *Is it possible that $\lambda_{\min}$ ... is larger with sketching than without?* Yes if $\mathbf{D}$ is a preconditioner $\mathbf{D} \approx \mathbf{U}^{-1}$.

[1] See Theorem 2 in [10], where in the number of iterations T is lower bounded by a constant term