We are very grateful to the reviewers for their invested time and expertise. Thank you very much for positive evaluation! We hope that all raised issues are properly addressed in our rebuttal. All minor corrections are implemented.

R1: *Potential applications and significance.* In our opinion, one of the most notable applications of second-order methods are ill-conditioned problems, which appear very often in practice. In particular, we are interested in minimization of SoftMax objective (the log-sum-exp function). This is a smooth approximation of the pointwise maximum of linear functions, and it is hard to solve this problem, when a smoothing parameter is small. At the same time, the classical Newton method works badly on this problem in practice (having no global convergence guarantees), when the starting point is far away from the optimum. We are happy to add our experimental results with this objective into the supplementary part. Additionally, the current experiments with Logistic regression problem show that our stochastic methods can be more efficient not only for reaching high-accuracy, but for low accuracy level ($10^{-2} - 10^{-4}$) as well (see Fig. 6). We also believe, that our theoretical developments can help in constructing new second-order algorithms.

Subsolver and Initialization. We have a brief description of our subsolver in the suppl. part (lines 380-385). We will add more details there. The code with our implementation will be available. The starting point was the origin. The first step of FW algorithm (with the standard step-size $\gamma_k := \frac{2}{k+2}$) is on the border of the domain. We think, this is the reason of the function value increase. The use of line-search in FW seems to be reasonable in this case. We will try that, thanks.

R2: *Contracting-Domain Newton vs. Aggregating Newton.* We agree that the first algorithm seems to have better performance in practice. From our experience, the aggregating method is more stable though. We have a brief comparison of these two methods in the supplementary section E (lines 373-379, and Fig. 5).

Additional feedback. Strong convexity of the composite term is assumed only in Theorem 2. One of the main contributions of [29] was the extension of first-order conditional gradient method onto the case of composite optimization problems. Additionally, the second-order Contracting Trust-Region method was proposed, which has the form of Algorithm 2 from our work (however, only $O(1/k)$ rate was established). We will highlight these contributions in our paper, thanks.

R3: *Stochastic methods* are both based on Algorithm 1. The difference between Algorithm 1 and 2 is in how we treat the composite part (see the remark on line 123). We will try to make our presentation of stochastic variants more clear.

The advantage over first-order methods. The complexity of one step for simple sets can be estimated as $O(n^3)$ + the cost of computing the Hessian. For Logistic regression, the cost of computing the gradient is $O(mn)$, and the Hessian is $O(mn^2)$, where m is the dataset size. Hence, when $m \gg n$, our methods can benefit significantly. Comparing the convergence rates, our complexity parameter H_ν depends only on the variation of the Hessian (in arbitrary norm). It can be much smaller than the max. eigenvalue of the Hessian, which typically appears in the rates of first-order methods.

Numerical simulations. We will provide more details regarding the experiments and the average performance of the methods, thanks. There was a typo in (23), this rate should be the same as that one in (13). Note, that on Fig.3 and Fig.6 we compare the number of Epochs (total data accesses), not the iterations.

R4: *Bounded domain.* We think, that the assumption on the boundness of the problem domain is not very strong. From the practical perspective, it seems natural to consider a bound for the variables (otherwise, the problem may appear to be non-feasible). The domain can be often introduced on a stage of creating a machine learning model (for example: regularization by ℓ_p-ball, as in (4); or the standard simplex, which is the domain of the probabilities for finite distributions). Considering the class of constrained optimization problems with bounded domain, our method has the same global convergence rate as that one of Cubic Newton, for convex objectives with Lipschitz continuous Hessian. At the same time, the subproblem at each iteration is simpler (no cubic terms). Thus, we think that our results are not weak.

Better convergence rate. Our methods can be considered as a second-order generalization of the conditional gradient method. Up to our knowledge, the best convergence rates for the methods of this type are known to be $O(1/k)$, for the class of convex functions with Lipschitz continuous derivatives (for example, [29]). The rate $O(1/k^2)$, which we prove in our work, seems to be significantly better.

R5: *Hölder continuity.* The claims of our paper are valid for any $\nu \in [0, 1]$. It is important, that our methods do not use a specific value of this parameter. Therefore, they may automatically adapt to the level of smoothness, achieving the best complexity guarantee among all $\nu \in [0, 1]$. *The proof of (6)* is based on the Newton-Leibniz formula:

$$||\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)||_* \quad = \quad \| \int_0^1 (\nabla^2 f(x + \tau(y-x)) - \nabla^2 f(x))(y-x) d\tau \|_* \overset{(5)}{\leq} \frac{H_\nu \|y-x\|^{1+\nu}}{1+\nu}.$$

Missed references to Katyusha and SARAH: thanks, we add them.

R6: *Empirical comparison with second-order methods with line search.* Indeed, it seems to be a reasonable and interesting comparison. We are happy to add more experiments into the supplementary part. *A typo in the statement of Theorem 4:* the correct rate should be $O(1/k^{1+\nu})$. This is fixed, thanks.