We thank reviewers for the time they took to read the paper and provide comments and suggestions.

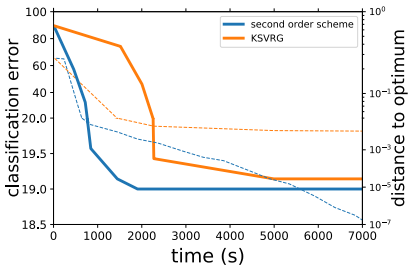
Reviewer 1. As the reviewer correctly points out one crucial aspect of the proposed algorithm is exactly the independence from the condition number of the problem. More generally, note that the paper provides an optimization algorithm that at the same time is: (a) independent (except for logarithmic terms) from the condition number of the problem, (b) globally convergent (unlike [6,7], the method will reach the optimum point from zero), (c) efficient from a computational viewpoint. Essentially all the known optimization algorithms nowadays satisfy only a subset of the points above. Indeed first order methods don't satisfy (a); Second order methods for general losses only (a); Second order with approximation methods as in Def. 2 satisfy (a)+(c). However the fact that they are not globally convergent make them not usable on real problems, since a first order method is needed to arrive very close to the optimum, losing (a). Finally there exists the very peculiar class of self-concordant function for which it has been proven that it is possible to achieve (a)+(b)+(c). However such result is more of theoretical interest in ML since very few objective functions there are self concordant. The novel contribution of the paper is a globally convergent optimization scheme that allows to achieve (a)+(b)+(c) for a larger class of functions i.e. *generalized self concordant* functions (GSC), which instead contains many losses of interest for regression, multivalued regression, robust regression, multiclass classification and multilabeling.

Relevance of GSC for ML and kernel methods. We would like to stress the significance of our result for part of ML and related fields. Indeed, while GSC is a subset of convex functions, it still covers many widely used approaches in machine learning, model estimation, statistics and finance, which are still framed in terms of large scale generalized linear models, like logistic and soft-max regression. More specifically, kernel regression and classification methods have the peculiar property to lead to large scale convex problem even if they are learning non-linear functions. In the community of kernel methods it is still an open problem how to find efficient methods to solve such big convex problems, without depending on the condition number, and our result is can constitute a key step in this direction.

Second order skepticism. We also understand the skepticism towards second order methods. Note however that (1) they become crucial to solve severely ill-conditioned problems, where the condition number is $\gg n$ i.e. 10^{10}, since first order methods can't avoid the dependence on the condition number. (2) the fact that second order methods have been labeled as inefficient in the '90 exactly because proper approximation techniques *and* globalization schemes where missing. In the past few years many approximation schemes have been developed as discussed in the introduction using novel results from randomized linear algebra. Nowadays as discussed in Section 2, the cost per iteration (epoch) of one approximated second order method is essentially the same as gradient descent (or stochastic gradient descent). Crucially the globalization scheme provided in this paper guarantees that the number or required iterations does not depend polynomially on the condition number (as in gradient descent or the stochastic methods where it is linear, or a square root with acceleration), but only logarithmically. So we need way less iterations than a first order method.

In particular we will make the introduction more clear putting elements of the discussion above; we will add a paragraph after the definition of GSC function to clarify their role; we will summarize the computational costs of the various algorithms of Section 2 via an additional table and moved some of the crucial aspects on how we improved over the state of the art from Section 2 to the contribution paragraph in the spirit of the discussion above.

Experiments. The proposed paper is clearly of theoretical flavour. We provide a new algorithm, and crucial part of the contribution is the development of the mathematical theory in terms of techniques, theorems and proofs to guarantee the correct behaviour of the algorithm. Additionally to show the practical validity of the approach, we performed a large scale experiment in the field of kernel methods (see above). In particular, in the experiments we compared our approach with KSVRG that is nowadays one of the fastest accelerated first order stochastic methods and is widely used to train kernel methods. We will add this plot showing the classification error for the experiment, which shows interestingly that it is still hard for first order methods, to achieve good performance in classification, for ill conditioned problems.



Reviewer 2. We thank Reviewer 2 the very thorough reading of the paper and for pointing out many typos which we will correct. We agree that FALKON is a very related algorithm; Indeed, the aim of the paper is to extend FALKON to more general losses. In particular the proposed approach recovers it in the case of squared loss. We are currently working on a minimal usable implementation for python and multicore-GPU, we will make the code open source and available on github this fall. Finally we will correct the format of the equation and the typos as suggested.

Reviewer 3. We thank Reviewer 3 for the very thorough understanding of the paper. We will add a note about the difference between GSC [6] and self concordant functions [24] after the definition of GSC functions. The difference with Newton sketch and our paper is that Newton sketch's globally convergent scheme applies to self-concordant functions while ours applies to GSC functions. Moreover as suggested, in Section 1.1 we will emphasize the advantages/disadvantages of our method. Indeed, its aim is to extend the performance of quadratic solvers to GSC functions, up to a log factor (only quadratic solvers are "condition number independent").