We would like to thank all four reviewers for the close look given at the paper. We believe that the paper gained clarity and readability by taking their comments into account. We now answer the different issues and comments.

Introduction (REVIEWER 1 AND 2) *Accelerated optimization*: this choice of wording was made to emphasize the fact that optimizing a smoothed version of the non-smooth objective function would lead faster to a (better) result. However, following your advice and to avoid confusion with the sense commonly given in the field, we will replace this term by *faster optimization*.
Problem formulation: we will introduce the problem tackled more formally in the introduction, introducing the objective function f and describing Table 1 as the error bounds on the norm of the difference between the objective function and its resulted smoothed form.
Poor generalization capabilities: our sentence is misleading, "resulting in poor generalization capabilities [8]" will be deleted.

Widening of local minima (REVIEWER 4) The widening of local minima was indeed not sufficiently developped, and we have thus decided to add the following result to Sec. 2.3 (using the extra page of the camera-ready version).

Proposition 1 (Widening of local minima)**.** *Let $\mu > 0$ and $y \in \mathbb{R}^d$. If f is L-Lipschitz, then there exists $z \in \mathbb{R}^d$ such that $\|y - z\| \leq \mu$ and, $\forall x \in \mathbb{R}^d$,*

$$f_{\gamma,\alpha}(x) \quad \leq \quad f(y) + \tfrac{L}{2\mu}\|x - z\|^2 + \tfrac{d}{2\alpha}\left(1 + \ln\left(1 + \tfrac{\alpha\mu L}{d}\right)\right) , \tag{1}$$

where $\gamma = \min\left\{\mu, \sqrt{\frac{\mu}{L\alpha}}\right\}$.

In other words, for every local minima $y \in \mathbb{R}^d$ of the objective function f, $f_{\gamma,\alpha}(x)$ will be small (i.e., approximately $f(y)$) in a neighborhood of y of size μ (see Fig. 1.c for an example). As a consequence, a good but thin local minimum will have its basin of attraction increased, and thus be easier to reach by GD even when the starting point x_0 is far from the local minimum. For example, consider $f(x) = \min\{1, |x|\}$. Its gradient is zero for any $x \notin [-1, 1]$, which means that GD initialized outside this region will be stationary. Moreover, for large smoothing parameters γ, RS will tend to flatten the objective and thus lead to the same behavior. However, BMR (with a sufficiently large α) will create an almost quadratic function in a region $x \in [-\mu, \mu]$, thus allowing GD to converge even when initialized at distance μ from the origin.

Experiments (REVIEWER 2, 3 AND 4) *Adding the comparison against ME would have been helpful*: Moreau Envelope is a great tool when the inner optimization problem (*i.e.*, the proximal operator) can be solved efficiently, ideally in closed form. Here we focus on settings where there is no closed form, thus directly using Moreau Envelope would add a full optimization problem at each step. The results would eventually need to be analyzed w.r.t. the precision achieved when solving the ME problem, adding a layer of complexity to the analysis. We chose not to follow that direction as a first go, in order to illustrate the core concepts unequivocally.
Optimization for DL, 10 dimensions is a bit disappointing: Our goal was here to emphasize the improvements over RS on simple functions. Scaling the method to big ML problems such as the training of deep neural networks and policy optimization in Reinforcement Learning is currently investigating and left for future work.
BMR presents some oscillations on Figure 3: this is due to the exploratory nature of the method in addition to the particular choice of α in this experiment, leading to an opportunistic behavior. Lowering α would decrease the oscillations observed toward the end of the process.

Typos and references (REVIEWER 1, 2 AND 3) Thank you for pointing to the relevant works [1, 3]. We will include them in the paper. We will also fix the reference [2] in the proofs. Finally, the bias in $O(\sqrt{K})$ is indeed a typo, it will be modified to $O(1/\sqrt{K})$.

General comments (REVIEWER 3) Our method bridges the gap between two different smoothing methods commonly used in optimization, both having their advantages and drawbacks. It must not be viewed as a trivial combination of ME and RS but as an interpolation between them. This interpolation allows to take the best of both worlds and leads to faster optimization, as shown in the experiments.

References

[1] A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 2012.

[2] C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. John Wiley & Sons, 2011.

[3] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 2005.