We thank the reviewers for their insightful comments and respond to their concerns and questions below.

Choice of h_k (R1). We consider h_k to be either Eq. 4 (proximal point), Eq. 9 (proximal gradient), or Eq. 10 (prox SGD). In principle, it is possible to design other surrogates, which would lead to new algorithms coming with convergence guarantees given by Prop. 1 and 4, but the three previous examples already cover important cases.

Plots in Section 5 (R1,R2,R4,R5). We thank the reviewers for noting a few problems with Fig. 1. As noted by R1, the legends are available in the appendix, and we will add them in the main text, if the paper is accepted.

Acceleration for ill-conditioned problems (R1). In the appendix, we conduct an experiment with an extremely high condition number (much higher than what is traditionally used for these problems). This is a case where the sublinear convergence rates for convex optimization ($\mu = 0$) are typically better than the linear rates for strongly convex optimization ($\mu > 0$), as long as the number of iterations is smaller than L/μ. In such a situation, a better strategy would have been not to assume the objective to be strongly convex at all. We will clarify this.

Clarity (R2). We agree that the clarity of our paper is subject to improvement and we thank the reviewer for his suggestions, which we will take into account, if the paper is accepted. Note that the norms are indeed Euclidean.

Code (R2,R5). All algorithms are implemented in C++ in the file utils/svm.h and mex files for Matlab are given in the folder mex/* Compiling requires (i) installing the Intel C++ compiler 2017, (ii) using Matlab 2016a, (iii) setting up the path to the Intel libraries in build.m, (iv) typing “build” in matlab. We admit that this procedure is not friendly and we will do our best to make it simpler in the future. Reproducing the results also requires accessing the datasets, which we will provide on an external website after publication since they were too large to be uploaded in the supplementary.

ckn-cifar (R2). We consider images encoded by unsupervised CKNs, and use our algorithms for the classification layer, which is convex. Addressing non-convex problems would be very interesting, but beyond the scope of our paper.

Dropout (R4). We use DropOut as in “Wager et al. Altitude Training: Strong Bounds for Single-Layer Dropout. NIPS 2014.”, which displays moderate gains on text classification tasks. However, the choice of DropOut in our paper (vs. more realistic data augmentation strategies for images, see [7]) is mainly motivated by the need of a simple optimization benchmark illustrating stochastic finite-sum problems, where the amount of perturbation is easy to control.

Clarity (R4). We thank the reviewer for his suggestion about adding a table for h_k/H_k, which we will do.

Significance (R4). We agree that direct acceleration methods are appealing. However, we also believe that the practical benefits of Catalyst are often underestimated. In Alg. 2 of [34], Catalyst is presented with three variants a), b) and c). Variants a) and b) require estimating optimality gaps for the sub-problems, which requires a lot of care (and pain) when implementing the method. Variant c) uses a fixed budget in the inner loops and is trivial to implement, e.g., see the file svm.h in the code we provide. Because variant c) came later, the community has focused on a) and b), often refering to Catalyst as being theoretically appealing but not practical. Yet, experiments conducted in [34] show that the simple strategy c), which we follow in our paper, is more effective in general than a) and b) in practice.

Finally, we believe that a universal acceleration framework for stochastic optimization is also a significant contribution from a conceptual/methodological point of view, but this is of course a subjective statement. It also provides a new point of view on stochastic proximal point methods, which have recently gained some attention in machine learning, see [3].

10 questions from R5. We thank the reviewer for his questions, which will lead to clarifications in the paper.
1. We will precise that [34] addresses only deterministic objectives (which may be finite sums).
2. We agree that a large variance combined with small μ leads to a large complexity, but this unfortunate combination would affect any stochastic optimization method since $O(\sigma^2/\mu\varepsilon)$ is asymptotically optimal. The goal of variance-reduction for the stochastic finite-sum problem is then precisely to reduce σ^2.
3. We will clarify that $t' = \lceil (2D/\mu)^{1/d} \rceil$ and $\tau = 1/2t'$ are fixed quantities, thus τ does not decrease. When writing $t = st'$, we simply mean that we consider (2) after s restarts, corresponding to st' iterations of $\mathcal{M}$.
4. κ is chosen as in [34] for the deterministic case $\sigma = 0$, leading to the right near-optimal complexity, even for $\sigma > 0$.
5.-6. $\mathcal{H}_2$ and $\mathcal{H}_3$ may look like strong assumptions, but they resemble classical ones used in the definition of estimate sequences by Nesterov (from the deterministic world), see Eq. (2.2.1) and (2.2.2) of [41]. We will clarify the analogy.
7. In this paragraph, we recover the results of Sec. 2 of [21] only, which treats the case with exact minimization, whereas we recover later the case with inexact minimization in Prop. 4.
8. When using the Catalyst surrogate, $\mathcal{H}_4$ implies $\mathcal{H}_3$ with $\delta_k = \varepsilon_k$. However, keeping both assumptions allows us to address more exotic cases (line 220 to 225). For clarity, it may however be better to treat this case only in the appendix.
9. Given any sequence $(\varepsilon_k)_{k \geq 0}$ and surrogate satisfying $\mathcal{H}_4$, the relation between F and $(\varepsilon_k)_{k \geq 0}$ is given in Prop. 4.
10. The goal of lines 197-200 is to illustrate our results when the objective is deterministic (which is what we mean by $\sigma = 0$). In such a case, we show that we achieve acceleration and recover the Catalyst method of [34].
We also thank the reviewer for correcting some typos and noting the problem with references [9] and [10].