We would like to thank all the reviewers for all the suggestive comments. Since all of the reviewers have different questions and doubts we decide to answer individually. Because of a lack of space we cite the papers using the references of the submitted work.

Reviewer #1:

1) Only upper bound! In this paper we only present upper bounds, but we also observe that our rate of convergence is indeed optimal since it matches corresponding lower bounds [3,7]. We will stress this important fact.

2) Extend the result to general convex loss function! Yes, we analyze the least-squares setting because of the easier structure. We expect similar results could hold for other loss (e.g. convex-Lipshitz, or even just smooth twice differentiable), but proofs, as well as details, to be different.

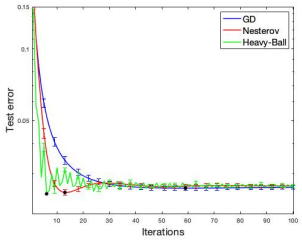
Reviewer #2:

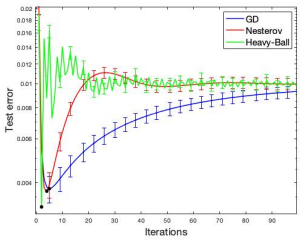
1) What's new and what's known? The filtering properties of ν-method and Nesterov accelerated algorithm for inverse problems have been studied respectively in [10] and [22]. We use these results together with probabilistic tools to derive their learning properties. Concerning the novelty in Theorem 1, implicit regularization properties of gradient descent algorithm have been largely studied both in the learning context (see e.g. [30,6]) and in the inverse problem scenario (see e.g. [10]). For what concern the optimization properties of accelerated methods, they have been studied since [22,10] but only in inverse problems.

2) Do the constants C_1 and C_2 depends on R and c_γ? Yes, they do, the dependence can be tracked in the proof of the general result and is at most linear for both R and c_γ. We will ass a comment on this point.

3) How the qualification affects the learning bound? Theorem 2 holds if the parameter r of the source condition is smaller than the qualification parameter of the chosen optimization algorithm, hence a higher qualification can adapt to better properties of the unknown target function. In this paper, we didn't focus on the effects of the qualification, but we will add some plots to illustrate this. See below.

In these simulations (where we chose the same parameters of those in the paper) it can be observed that increasing the parameter r of the source condition gradient descent (qualification ∞) can recover the behavior of the other methods.





4) Infimum over $\mathcal{H}$? It's a typos, the infimum is over $\mathcal{X}$, but extending the expected risk to $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ and denoting with $\mathcal{H}$ the subspace of linear function then they are the same.

Reviewer #3:

1) Advantages of accelerated methods? Considering the large-scale scenario or problems of learning on a budget, one cannot chose to pay with time, for example if the optimal iterations are 10^6 for GD and 10^3 for the accelerated versions but we have a limited budget of 10^2 iterations then acceleration is to prefer because of the faster decrease. Accelerated methods allow to reach the same accuracy with less computation, so they are to prefer. It is true that in practice it is difficult to stop at the optimal iteration and the plateau of "good iterations" is more strict for the accelerated methods, but this problem concern how to tune hyperparameters not the algorithm itself. On the other hand gradient descent can exploit it's higher qualification (see point 3 of reviewer #2), but again in practice one do not know nothing about source condition.

2) Accelerated methods are more unstable! We call "stability" the error which derive from running the algorithm with finite data instead of infinite, it can be seen at line 434 of Supplementary material that this term turns out to be the second addend of the learning bound. For accelerated methods this term is the square of the gradient descent one, so accelerated methods are more unstable.

3) Why placing $1/t$ and $1/t^2$? Main mathematical contribution in the appendix! The proof holds in general for all spectral algorithms defined by a filter function g_λ. In Section 3 we prove that gradient descent and the accelerated methods are filter functions by choosing the regularization parameter λ respectively as $1/t$ and $1/t^2$. Yes, the theorem in the appendix is more general but in this paper we wanted to focus on the statistical properties of the acceleration rather that general spectral filtering.

4) Missing definitions! It's true, the function g_t can be extended through spectral calculus to a function of operators by defining it on the eigenvalues. The operator $x \otimes x$ is defined as the operator form $\mathcal{X}$ to $\mathcal{X}$ such that $x \otimes x(v) = \langle x, v \rangle x$, and $\Sigma = \mathbb{E}[x \otimes x]$.

We thank again the reviewers and hope to have clarified all their doubts.