We thank all reviewers for their time and expertise. We are particularly happy that R1 and R3 are positive about our paper. R1's summary of our contributions resonates strongly with what we wanted to achieve.

We are worried that R2 may have misunderstood main points of the paper. Some of their comments seem to suggest that we endorse the EF. The exact opposite is the case. The EF, which doesn't exist in the Stats literature, is widely used in ML (see our citations). We argue that it should *not* be used. Let us clarify some of your points:

> (R2) *"for applying the exact natural gradient [...] there is no reason to use the EF over the Fisher."*
We agree! However, a large portion of the literature applying natural gradient ideas to machine learning uses the EF and many researchers are under the impression that the EF and the Fisher are interchangeable. We cite 15 papers that use the EF, 9 of them from the past 3 years, and we know of at least 4 additional papers, accepted or under review at ICML, UAI and NeurIPS 2019, perpetuating this confusion. Basically, our paper makes the points that you levy against us in the review. Our goal is to point out the differences between the EF and the Fisher and clear the confusion, in the hope that future applications of natural gradient methods to machine learning will be more impactful.

> (R2) *"the authors used the term "empirical Fisher" [... did the authors coin] the term?"*
We did not coin this term and, as noted in L265-267, we believe it is a misnomer. To the best of our knowledge, the term "empirical Fisher" first appears in the manuscript of Martens [2014]. Altough the origin of the concept might be Le Roux et al. [2007], who confused the covariance and the Fisher information, and Graves [2011] who proposed it as an approximation of the Hessian. The idea was heavily popularized by the Adam optimizer [Kingma and Ba, 2015].

> (R2) *"This should be connected to observed Fisher information which is a well-defined concept in statistics."*
This is a fantastic point that we definitely need to address. While this terminology is widely used in the literature we cite, it is true that the concept might sound strange to statisticians as the terms "observed Fisher" and, even more confusing, "empirical Fisher information", are used, in statistics, to refer to what the community we are trying to reach calls the Fisher information. We will use the additional page to make the context clearer to a larger audience and give an overview of how the EF approximation came to be.

> (R2) *"the experiments are mainly performed on toy examples. It is not clear how [to generalize] to real networks"*
Note again that we are arguing *against* the use of the EF concept. If it doesn't even work on toy problems, why rely on it in "real networks" that are much harder to understand?

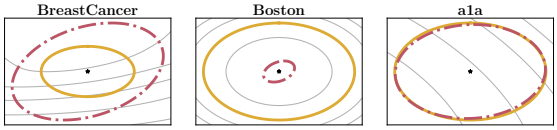
Questions raised by R3 & R1 (thanks for these!):
> (R3) *"the plots in fig 3 show that the cosine [...] is close to 1 near the end of the training, which seem to contradict the claims in subsection 3.1 and 3.2: in practice it seems that the cases discussed in 3.1 and 3.2 are not met."*
The plots for §3.1 and 3.2 are in Fig.2, showing strong differences at the minimum. Fig. 3 showcases §3.3, about preconditioning, showing that updates can be opposite. While the cosine gets close to 1 in Fig.3, this metric captures only a small part of the relationship and the two matrices can still be very different (see figure at end of this rebuttal).

> (R3) *"how do you obtain eq. 12? [...] line 184 the Fisher does not match eq. 2"* Eq.12 and L184 show the equations for the empirical Fisher (Eq.3) and the Fisher (Eq.2) applied to the linear regression example of L178-179.

> (R3) *"line 192 if b is arbitrarily close to 1 then $\nabla_b \log b$ is arbitrarily close to 0 [...] I think you wanted to write that using the definition line 179 with a variance 1 normal distribution, $b_n(\theta)$ can not get arbitrarily close to 1."* and
> (R1) *"For classification [when] all labels are predicted correctly with probability 1, wouldn't [F = EF = 0?]"*
The subtelty of the proposed change to the definition of the GGN (Prop.3) relies on this key point: for the guarantee in Prop.2 to hold, $\nabla_b a(b)$ needs to go to 0, which can not happen for $a(b) = \log b$ as $\nabla_b \log b = 1/b$ (for $b \in [0, 1]$). As both reviewers noted, if b goes to 1 and is correct, the EF is 0 (the gradients are 0), the Fisher is 0 (the model is no longer probabilistic, meaning new samples can not give more information) and the Hessian is also 0 (some parameters, the weights for classification or the precision–if learned–for regression, need to reach ∞ and the landscape is infinitely flat). We will add a paragraph to explain this special case.

> (R3) *"in eq 15 I think the multiplicative N should be in front of the covariance matrix if we follow [...] eq 3"*
We defined Σ as the covariance a uniformly distributed gradient scaled by N (L184) to be consistent with the definition of the loss as a sum in Eq.1. Sorry for the confusion; we will try to make it clearer.

Typos found by R3: Will all be addressed, sincere thanks for the close reading. (Eq.5, L.183-185 and Prop.2)
Sorry about the lost pointer to the proof of Prop.2 - it is in the supplementary §C.3.



Quadratic approximations (as in Fig.2) using the projection of F (yellow) and EF (red) on the two largest eigenvectors of F, at the end of training with the EF, using the settings of Fig.3. While the cosines are close to 1, the matrices can still be very different in terms of directions and scaling.