We wish to thank all reviewers for the time they spent reading and commenting on our paper. We will amend the text with all the suggested clarifications and corrections.

We now reply to the most important concerns.

Momentum (or Nesterov acceleration) and variance reduction

One of the questions raised by reviewer 3 is a crucial one. While momentum is proved to help with curvature, its formulation resembles that of variance reduction. The theorems in [39] show a greater sensitivity of acceleration to noise, although this might be a limitation of the analysis. After the NeurIPS deadline, the paper "Limitations of the Empirical Fisher Approximation" was published showing that using the covariance matrix of the gradients in lieu of the Fisher (which can be thought of as a curvature matrix) was problematic. The paper "Information matrices and generalization" shows that, although the covariance and the Hessian encode different pieces of information, they are remarkably aligned in the currently used models. In conclusion, although we view deterministic momentum as helping with curvature, IGT as helping with noise and stochastic momentum as a successful method without general theoretical guarantees, there seems to be deep links between variance and curvature and the discussion of which method addresses what problem is not over.

Convergence rate of IGT

R3 mentions the O(1/t) rate. Although variance-reduction methods achieve a linear rate of convergence on the training error, the rate on the test error is still no better than O(1/t), the minimax rate. Since we are in the online setting, we are directly optimizing the test loss and thus cannot do better than O(1/t).

Weaknesses of ITA

ITA interpolates between only using the last iterate, which is robust to a violation of the identical Hessian assumption but does not decrease the variance, and using all of them, which has the lowest variance but suffers when the assumption is violated. ITA gets the best of both worlds but this come at a price, albeit small. First, the variance is greater by a constant factor than when all the iterates are used. Second, there is an additional hyperparameter: the fraction of examples to keep. While we found that our method was competitive for a wide range of values and that setting it so that one full epoch was used at the end of training worked well in all cases, it has an impact nonetheless.

Performance of Adam-ITA

To be true to our goal of limiting the number of hyperparameters of IGT, we used the same values for the parameters as the non-IGT version. While this worked well for most methods, Adam-ITA did not work well. We did not want to falsely give the impression that it cannot work well but the computational cost to fine-tune Adam-ITA parameters on these large datasets would have been prohibitive. We can however run these experiments if the reviewers think there is a lot of value in them.

Clarification about the restricted setting

We would like to point to R3 that our theorems apply in a more restrictive setting than the functions being quadratic as we require all individual Hessians to be equal. It is unclear how to extend the theorems to the general quadratic case.

Miscellaneous questions

We did not observe any numerical instabilities. Although the factor in the extrapolation step grows, subsequent iterates get closer so the extrapolation step is never too far away. HB-IGT indeed barely improves upon IGT