We thank the reviewers for valuable comments and questions. We address each reviewer individually below.

Response to Reviewer #1. Below we address the three weaknesses put forth by the Reviewer #1.

1. *On the linearity assumption.* The reviewer points out that the linearity assumption removes much difficulty in the analysis. On the contrary, even for linear problems, the statistical/computational analysis of early-stopped iterative algorithms (typically gradient descent) is rather involved and usually results in rather long proofs (see, e.g., [10, 37, 44, 47, 49, 50])). In our work, the analysis is so simple due to the novel connection that we develop between offset Rademacher complexities and mirror descent algorithms. Given the fundamental nature of this connection, we hope that our work will motivate follow-up research in this direction, extending our framework to non-linear algorithms and also extending the framework of offset Rademacher complexities beyond the quadratic loss (see also response to R#3).

In addition, given the number of papers whose setups are special cases of ours (cf. Section 2.1 and Appendix D), we do not see the linearity/kernel assumption as particularly limiting. While there is on-going implicit regularization research in the context of neural networks (as we briefly review in Section 2.1), such works, in contrast to our work, do not focus on excess risk guarantees. Finally, we emphasize that linear models are fundamental in statistics and machine learning, and there is growing literature that uses them to also describe key features in non-linear neural networks. For instance, see the double descent phenomenon and the Neural Tangent Kernel literature).

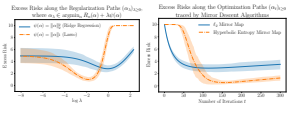
2. *On the boundedness assumption and comparison with [WYW, 19].* We remark that Theorems 1 and 2 can be applied to also obtain high-probability bounds (cf. footnote 1 on page 3 and Appendix D.1). We opt to present bounds in expectation as such results are the simplest to state and also require minimal assumptions on the data generating mechanism, namely boundedness. Note that, in contrast to our results, [WYW] prove high-probability bounds under a well-specified model assumption, which is not present in our work. In addition, the analysis of [WYW] does not apply to derive results such as Theorem 4 (updates are not gradient descent) and Theorem 5 (statistical guarantees along the whole optimization path) that can be easily proved within our framework. We refer to Appendix D.1 for an extended discussion, where we also discuss why our results obtained in Theorem 3 are nevertheless similar to the ones in [WYW].

3. *On Theorems 3 and 4.* The reviewer's comments (both in summary and in weaknesses sections) on Theorems 3 and 4 is a misunderstanding since both theorems are **discrete** time results that follow via an application of Theorem 2. We kindly ask the reviewer if they could have another look at Theorems 3 and 4 and the surrounding discussions.

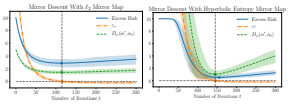
Response to Reviewer #2. The primary limitations are reliance on the quadratic loss and linearity of the model. We hope that both can be addressed in future work. Regarding the linearity assumption, please see our response to the Reviewer #1. Regarding the quadratic loss, please see our response to the Reviewer #3. In the applications that we present, the only assumption on the data-generating distribution is boundedness, which is necessary in the distribution-free setting (i.e., in contrast to the related works [37, 47], we do not assume a true model generating the data).

Response to Reviewer #3. Regarding the linearity assumption, we point to our answer to the R#1 above. At this point, it is indeed unclear under what conditions our arguments generalize to other loss functions. We believe that there would not be significant difficulties in extending our arguments to smooth and strongly convex loss functions considered in [47]. At the same time, we believe that the fundamental connection that we have observed between mirror descent and offset Rademacher complexities *can offer a dual view on localization via offset Rademacher processes*, facilitating future research in this direction. Regarding the reviewer's suggestions, we agree with all of them.

Response to Reviewer #4. Regarding the information-theoretic discussions, localized complexity measures are known to yield minimax-optimal rates for various problems. In addition to the early works on localized complexity measures [9,23], Corollary 12 in [25] shows that offset Rademacher complexities capture correct rates for non-parametric regression. In Section 3.3 in [47], minimax optimality of localized complexities is established for regular kernels. We will add these pointers. The function $g_{\mathcal{F}}$ such that $R(g_{\mathcal{F}}) = \inf_{g \in \mathcal{G}} R(g)$ is defined as a notational convenience (f.note 2 on p. 3). Finally, we address the reviewer's questions regarding the stopping time $t^\star$ and the request for a simple example demonstrating the key concepts that appear in our paper. Intuitively $t \ll t^\star$ (where $t^\star$ is the prescribed stopping time) results in a sub-optimal estimator for the same reason that poorly tuned explicitly regularized estimators are sub-optimal (i.e., the number of iterations t play a similar role to the regularization parameter λ in explicitly penalized problems). Consider a data model $x_i \sim N(0, I_d)$ and $y_i | x_i \sim \langle \alpha', x_i \rangle + N(0, \sigma^2)$ and assume that α' is a sparse vector. Due to the sparsity of α', lasso is superior to ridge regression (cf. Fig. (a)); similarly, mirror descent with the hyperbolic entropy potential is superior to gradient descent (cf. Fig. (b)). Also note the similarities between Fig. (a) and (b). Finally, Fig. (c) and (d) graphically demonstrate the main idea behind our proof techniques: up to the vertical dotted line that denotes $t^\star$ (cf. line 277), the Bregman divergence $D_\psi(\alpha', \alpha_t)$ is non-increasing (green lines). **[Zoom in on the figure for details.]**



(a) (b)



(c) (d)