We thank all reviewers for their time and feedback; we address common and individual comments in turn.

(R3, R1) Non-asymptotic intervals, improved widths: In the revision we will highlight that non-asymptotic CIs can be derived from the CV concentration inequalities of [10,11,2,3,Cornec arXiv:1011.5133]. These CIs are more difficult to deploy as they require (1) stronger stability than loss stability, (2) a known upper bound on stability, and (3) either a known upper bound on the loss or a known uniform bound on the covariates and a known sub-Gaussianity constant for the response variable. In addition, the reliance on somewhat loose inequalities typically leads to overly large, relatively uninformative CIs. For example, we implemented the ridge regression CI from [Thm. 3, 11] for our `FlightDelays` experiment (an implementable CI is not provided for any other learning algorithm). This CI takes as input the maximum absolute value of the target y ($B_Y = 8.03$ after mean-centering) and the maximum ℓ_2 norm of a feature vector x (after mean-centering, $B_X = 13.17$ with standardization or $B_X = 4200$ without). When standardizing as in Fig. 2, the smallest width produced by [Thm. 3, 11] for any value of n is 90.2; that is 86 times larger than the largest width of our CLT intervals (equal to 1.04). When not standardizing as in App. K Fig. 3, our maximum width is 1.03, but the minimum [Thm. 3, 11] width is 5×10^{14}. We will emphasize this important advantage of CLT intervals in the revision.

(R1) Stability clarifications: We will clarify in the revision that our stability assumptions

1. Do not require that h_n be convex (in fact, many past stability results are for a 0-1 validation loss [12,15,16,19,4]) and do not require that h_n be related to a loss function used to train a learning method
2. Cover k-nearest neighbor methods [12], decision tree methods [4], and ensemble methods [16] in addition to non-convex SGD and strongly convex ERM
3. Hold even when training error is a poor proxy for test error due to overfitting (e.g., 1-nearest neighbor has training error 0 but is still suitably stable [12])

We are not aware of other approaches that provide CLTs for such a broad class of learning algorithms and losses, but we would appreciate any pointers to literature that we have missed.

(R1) Experiments: We appreciate the suggestions to improve our presentation and will introduce a new experiment with synthetic data generated from a known model. We feel it is important to also maintain our existing real-data experiments, as these best reflect how the competing CIs and tests perform in practice, under the eccentricities of real data which are hard to capture with synthetic data. For example, it is common in real data to have one method dominate for smaller sample sizes and the other dominate for larger sample sizes; this is precisely what we see in the right column of Fig. 2. We will clarify that the aim of this assessment is not to establish power convergence or to assess power in an absolute sense but rather to verify that for a diversity of settings encountered in the wild (e.g., random forest much better than logistic regression in Fig. 2 left and ridge regression barely better than neural network in Fig. 2 right), our tests provide power as good as (and often better than) the most popular heuristics from the literature. We will clarify that the reported sizes $= \frac{\text{\# of rejected } H_0 \text{ simulations}}{\text{\# } H_0 \text{ simulations}}$ and powers $= \frac{\text{\# of rejected } H_1 \text{ simulations}}{\text{\# } H_1 \text{ simulations}}$, where one of the 500 simulations is declared H_0 if the test error of $\mathcal{A}_2 \leq$ test error of $\mathcal{A}_1$ and H_1 otherwise. Notably, we only see size estimates exceeding the level when the number of H_0 simulations is very small (when $\mathcal{A}_2$ improves upon $\mathcal{A}_1$ in so few simulation replications that the Monte Carlo error in the size estimate is large).

(R1) Known R_n: In addition, we have rerun all `FlightDelays` regression experiments using an exact known R_n (so that R_n need not be estimated). In these new experiments, we take the population distribution to be the empirical distribution over our entire `FlightDelays` dataset (so that R_n is an expectation over 5.8M datapoints) and sample training sets independently from this population. With this setup, we can exactly determine which of two algorithms has better k-fold test error, and the results are comfortingly very similar to those reported in the submission.

(R1) $\mathbf{V_n}$: We have only studied the conditional setting, but [5] recently proved an unconditional CLT under much more restrictive assumptions than their conditional CLT and found consistent variance estimation to be more elusive.

(R1) Terminology: In the revision, we will clarify the formal definitions of "asymptotically exact" (coverage converging to *exactly* $1-\alpha$) and "asymptotically valid" (coverage asymptotically $\geq 1-\alpha$), as the former is a stronger property.

(R3) Non-asymptotic linearity: In the revision, we will highlight that (3.2) in Thm. 2 already implies a "non-asymptotic linearity" statement by providing an explicit non-asymptotic bound on the departure from linearity in terms of the algorithm's loss stability: $\mathbb{E}[(\frac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n) - \frac{1}{\sigma_n\sqrt{n}}\sum_{i=1}^n(\bar{h}_n(Z_i) - \mathbb{E}[\bar{h}_n(Z_i)]))^2] \leq \frac{3}{2\sigma_n^2} n(1-\frac{1}{k})\gamma_{loss}(h_n)$.

(R3) AMSE: We will clarify in the revision that we have no particular interest in unbiasedness; rather, our interest in CV comes from its popularity: consumers and developers of ML methods are already using CV to estimate test error, and we aim to turn those readily available estimates into valid inferences about test error without requiring any new expensive computation (i.e., using only standard CV outputs). In addition, for estimating the mean of a univariate normal, the best unbiased estimator is admissible and minimax optimal, so while bias can improve MSE for some values of the true mean, no alternative estimator will have better MSE for all values of the unknown true mean.

(R4) Our weaker assumptions: In the revision, we will endeavor to improve intuition, highlighting that past results exclude asymmetric (like SGD), inconsistent, and less stable learning algorithms and heavy-tailed data distributions (see Apps. F & G for detailed examples of simple learning problems excluded by past work but covered by ours).