We thank all the Reviewers for their feedback and their service to the community. We are glad that you have all understood the relevance of our work and, in general, appreciated it. In the following, we try to comment on all the raised issues, and how we will address them, which will certainly improve our work.

Localization We agree that the relationship between empirical centralization and localization is important and we gave it much thought. We explained this relationship in an earlier submission, but had to omit it in the NeurIPS submission due to space limitations. We summarize it in the following paragraphs.

Empirical centralization is *complementary and orthogonal with localization*, as it fixes a different issue: localization is akin to a second-moment normalization (i.e., dividing out variance), whereas centralization is a (complementary) first-moment normalization technique. Because both localized and centralized Rademacher averages are themselves Rademacher averages, they are mutually compatible. We recommend centralizing before localizing, as this approach addresses several issues with standard localization methods.

In particular, taking localization to mean "analysis of the variance-constrained star-convex hull", function families containing $\{x \mapsto 0, x \mapsto 1\}$ (for example, any case of realizable symmetric classification with 0–1 loss must include these two functions) suffer from $\Omega(1/\sqrt{m})$ convergence rates, since constant functions have variance 0, whereas with centralization, these two functions are the same, removing this lower bound.

Instead taking localization to mean "constrain by raw-variance," the above issue is solved, but now the localized Rademacher average bounds are $\Omega(V/\sqrt{m})$, where V is still only the smallest *raw*-variance.

By centralizing before localizing, the difference vanishes, as with mean 0, raw and centralized variances coincide. Furthermore, both of these bottlenecks are resolved: with empirical centralization, in many instances we reduce the asymptotic gap between upper and lower SD bounds, and in this sense, our SD, and thus excess risk, bounds are more transparent. Thus centralization and localization are complementary, and akin to first and second moment normalizations. We intend to fully explore this connection in future work.

Monte-Carlo approach As mentioned on line 183, a consequence of Theorem 5 is that $n = 1$ trials are sufficient to asymptotically match the rate for the SD shown by Bousquet (2002) (reported in Theorem 3), whereas a use of McDiarmid's inequality (as in the state of the art w.r.t. Monte-Carlo Rademacher Averages (Bartlett and Mendelson 2002)) would require high n for low wimpy variance. Matching Bousquet's rate only occurs with centralization: non-centralized Rademacher averages have inferior Monte-Carlo concentration properties. We have additional experimental results showing a comparison between our results and the one using McDiarmid's inequality, which we did not include due space limitations but we will include in the updated supplementary materials.

RA without absolute values The identity

$$\mathbb{E}_{\sigma} \sup_{f} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) = \mathbb{E}_{\sigma} \sup_{f} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (f(x_i) - \hat{\mathbb{E}}_x[f])$$

does not hold in general. It does hold by linearity when all functions in the family have equal expectation, but consider the counterexample family $\{x \mapsto -1, x \mapsto 1\}$: any centralized Rademacher average (with or without absolute values) is then 0, whereas without centralization, the value is $\Theta(1/\sqrt{m})$ (see Eq. 5). Other less trivial counterexamples are possible but too convoluted to be explained in the limited space of this response.

Thus even the notion of Rademacher averages without absolute values benefits from centralization. Furthermore because the 2-sided symmetrization inequality (with centralization and absolute values, eq. 4) is factor-4 sharp, and by applying the non-absolute symmetrization inequality once in both directions, we recover a 2-sided guarantee, we believe that any gains obtained by removing the absolute value are marginal.

We also have evidence that, when using Rademacher averages for non-statistical-learning-theory tasks, such as approximation algorithms for data analytics tasks (e.g., references [17,18,20,21]), the absolute value is important, and should be considered as early as possible.

We omitted a discussion of non-absolute Rademacher averages for clarity of presentation, but we could include this information if it is deemed sufficiently important.

Other comments As requested by Reviewer 4, we will give additional details about applications; in particular, we will describe how our linear family bounds can be applied to get sharper bounds for selecting the optimal expert in a batch-learning panel-of-experts setting (and thus reduced regret in online settings). Reviewer 2 asks how our methods extend to model selection, and we note that, for example, this panel-of-experts setting can be extended to *structural risk minimization* if the experts are organized into concentric groups. We will use Big-Oh notation whenever possible, and clarify the statement of Thm. 4. We will fix the typos pointed out by the Reviewers and do a deep editing pass for any other spelling, grammar, or syntax issues.