@Reviewer 1. Your detailed comments were much appreciated. • The irreducible term involving V_n' does not necessarily prevent a fast rate in the realizable setting as we will clarify now: First, suppose that Q is chosen to satisfy (6) (this assumption will simplify the argument that follows at the price of weakening our bound slightly). With this choice of Q, one can now see that V_n' is just a measure of the average performance of hypotheses drawn from the "alternate posterior" $Q_{\leq m}$ [resp. $Q_{>m}$] on the *unseen* i.i.d. sample $Z_{>m}$ [resp. $Z_{\leq m}$]. Thus, when $|\ell| \leq 1$ and $m = n/2$, the first sum in the expression of V_n' satisfies $\sum_{j=1}^{m} \mathbb{E}_{h\sim Q_{>m}}[\ell_h(Z_j)^2]/n \leq \sum_{j=1}^{m} \mathbb{E}_{h\sim Q_{>m}}[\ell_h(Z_j)]/n \leq L(Q_{>m})/2 + \tilde{O}(1/\sqrt{n})$, where the last inequality holds w.h.p. due to Hoeffding. By applying the same treatment to the second sum in V_n', we arrive at (*): $V_n' \leq (L(Q_{\leq m}) + L(Q_{>m}))/2 + \tilde{O}(1/\sqrt{n})$ (w.h.p.). If Q is chosen to be ERM, for example, then the risks $L(Q_{\leq m})$ and $L(Q_{>m})$ are often very small in the realizable setting (as they approach the zero Bayes risk). If these risks are of order $1/\sqrt{n}$, then due to (*), our irreducible term $\mathcal{C}\sqrt{R_n' \cdot \mathrm{s}_{\delta,n}/n} = \mathcal{C}\sqrt{V_n' \cdot \mathrm{s}_{\delta,n}/n}$ in (1) is of order $\tilde{O}(1/n)$. In contrast, McAllester's bound can be of order $\tilde{O}(\mathrm{KL}/n)$ in the realizable setting, and so in this case our irreducible term is of comparable size or even smaller since it is "KL"-free. The same conclusion can be drawn in the *non-realizable* setting, since in this case, for n larger enough, $\sqrt{L_n(P_n) \cdot \mathrm{KL}/n}$ becomes the dominant term in McAllester's bound which, again due to the KL and (*) above, can be larger than our irreducible term. The argument above should also answer your question about whether the upper-bound $V_n' \leq b^2$ (or the $O(b/\sqrt{n})$ upper-bound on the irreducible term) is too pessimistic—indeed it is as revealed by (*) above. • Though it may be more interpretable, McAllester's bound, which is the result of the kl analysis of Maurer's bound (see, e.g., (3) in TS paper), is far from being competitive with our bound—in our experiments it performs worse than Catoni's bound; the bound due to Maurer that we report on in our experiments is exactly (2) without any relaxation (see lines 152-153), which is much tighter than McAllester's—this is also why we only reported results for the former. • We will modify line 22 to reflect your point. • From (*), one would expect $m = n/2$ to be the optimal choice for the data splitting. However, with the convention that $Q(\varnothing)$ equals some prior, say P_0, the bound would still be meaningful for $m \in \{0, n\}$. • It can be shown that TS's bound can equivalently be written as (constants and log-factors omitted): $L(P_n) - L_n(P_n) \leq \max(\sqrt{\mathbb{V}_n \cdot \mathrm{KL}/n}, \mathrm{KL}/n)$, where $\mathbb{V}_n$ is the empirical loss variance—the case distinction they consider is merely another way of writing the same inequality. This inequality can be relaxed using $\max(a, b) \leq a + b, \forall a, b > 0$, to recover a bound of the form (1). We will explain this point in the appendix and point to it in Sec. 2. • Lines 142-144: Consider the second term in V_n, in the Corollary of Thm. 1, involving the sum from $j = m + 1$ to n; looking inside the square, we see that to predict the data point Z_j, the estimator $\hat{h}$ only uses the sample $Z_{\leq m}$ and suffers loss $\ell_{\hat{h}(Z_{\leq m})}(Z_j)$. In contrast, in Thm. 1, one can use the posterior $Q_{<j} \equiv \delta(\hat{h}(Z_{<j}))$ which depends on the extra points $Z_{m+1}, \ldots, Z_{j-1}$ and suffers loss $\ell(\hat{h}(Z_{<j}), Z_j)$. This loss gets closer and closer (as $j \to n$) to the loss that would be incurred by the estimator $\hat{h}(Z_{\leq n})$ "trained" on the whole sample $Z_{\leq n}$. • In the experiments, $m = n/2$ and in (5) we considered (p, q) equal to $(1, m)$,$(m + 1, n)$, and $(1, n)$, to compute $Q_{\leq m} \equiv \delta(\hat{h}(Z_{\leq m}))$, $Q_{>m} \equiv \delta(\hat{h}(Z_{>m}))$, and $\hat{h}(Z_{\leq n})$ (the latter is used for P_n). • In our experiments the posterior P_n is a Gaussian centered at $\hat{h}(Z_{\leq n})$. So, to see why it is important for $\hat{h}(Z_{\leq m})$ and $\hat{h}(Z_{>m})$ to be close to $\hat{h}(Z_{\leq n})$, consider the extreme case where P_n has zero variance, i.e. $P_n \equiv \delta(\hat{h}(Z_{\leq n}))$. In this case, it is clear from the expression of V_n that, if $\hat{h}(Z_{\leq m}) = \hat{h}(Z_{>m}) = \hat{h}(Z_{\leq n})$, then $V_n = 0$. So, when P_n has non-zero but small enough variance, one would still expect $V_n \simeq 0$, when $\hat{h}(Z_{\leq m}) \simeq \hat{h}(Z_{>m}) \simeq \hat{h}(Z_{\leq n})$, which can make the first term on the RHS of (1) small.

@Reviewer 2. Thank you for your feedback on the experiments. • When running the synthetic experiments, we found that the bounds were highly sensitive to one particular parameter—the variance of the Gaussian posterior P_n. For this reason, the variance was optimized for every bound separately (see lines 165-167). The sensitivity w.r.t. the Bayes Optimal Predictor (BOP) was weak; varying the BOP did not change the relative ranking of the bounds or affect the gap between them by much, and so due to space we did not include results for different BOPs. Nevertheless, we will now add the cases where the Bayes error is equal to 0.05 and 0.2 for randomly generated BOPs in the appendix. • In Figure 1, the Bayes error is 0.1 (the true labels are flipped with probability $1 - 0.9$, see line 173); we will make this clearer by adding it to the Figure directly. • We will cite and briefly discuss Rivasplata et al.'s paper in the relevant section.

@Reviewer 3. Thank you for pointing out Rivasplata et al.'s paper. • Although their title might suggest otherwise, the ideas, techniques, and results of their paper are substantially different from ours. In fact, even though the title of their paper mentions "instance-based", to compute their bound, one needs to know the uniform (worst case over *all* possible samples, not just the observed one) stability for the learning algorithm involved. For many popular algorithms, such as gradient descent for example, no non-trivial bound on this worst-case stability parameter is known. In contrast, we can get non-trivial bounds for *any* algorithm as soon as our *empirical* notion of stability V_n—which can be calculated on the data—is small. We will cite Rivasplata et al. and explain this difference in the relevant section. • The experiments in our paper are based on λ-penalized logistic regression, which happens to be an algorithm for which Rivasplata's worst-case stability parameter β_n can be calculated after all, giving $\beta_n = 1/(\lambda n)$. We tested Rivasplata's bound with this value of β_n, and found that in our experiments it performs worse than the bounds we currently compare ours against (i.e. Catoni, Maurer, and TS). For completeness, we will add these additional results with discussion in the appendix.