1) Comparison with Catoni's Bound (R1 & R2 & R3): Catoni's bound has the form $(1+c_c)\hat{\mathcal{L}}_S Q + \frac{\mathcal{C}_c}{m}(\mathrm{KL}(Q\|P) + \log\frac{1}{\delta})$, while our bound (Eq. (25)) can be written $(1+c_r)\hat{\mathcal{L}}_S Q - \frac{c_r(1-h^2)}{m}\sum_{i=1}^m(\mathbb{E}_Q f(z_i))^2 + \frac{\mathcal{C}_r}{m}(\mathrm{KL}(Q\|P)+\log\frac{1}{\delta}+1)$. Here c_c, c_r inflate the empirical risk and $\mathcal{C}_c, \mathcal{C}_r$ are constants. Let T_m be $\frac{c_r(1-h^2)}{m}\sum_{i=1}^m(\mathbb{E}_Q f(z_i))^2$. Note that c_c and c_r must be fixed before seeing the data. Assuming we equate the inflation of the empirical risk, i.e., $c_c = c_r$, the proposed bound is tighter than Catoni's bound provided $m > \frac{1}{T_m}\left((\mathcal{C}_r - \mathcal{C}_c)\left(\mathrm{KL}(Q\|P) + \log\frac{1}{\delta}\right) + \mathcal{C}_r\right)$. If T_m converges to a positive number (a reasonable assumption), then our proposed bound will be tighter for sufficiently many samples. If we assume $c_c \neq c_r$, our bound can still be tighter than Catoni's bound under more involved conditions.

2) Talagrand-type concentration inequalities (R1): We thank the reviewer for this valuable comment. As the reviewer anticipates, some adaptation of Talagrand's result seems to be necessary. We can raise this as an open question.

3) From binary loss to bounded or unbounded losses (R2): Our main results for binary loss can be extended to $[0,1]$-valued (i.e., bounded) loss, but with a different constant $C = \frac{4h^2 c}{9(1+h^2 c)(1+h^2 c/9)}$ that has the same interpretation as in Eq. (14). Briefly, the proof relies on Jensen's inequality $\mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q) - \frac{c}{m}\sum_{i=1}^m \mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2 \leq \mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q) - \frac{ch^2}{m}\sum_{i=1}^m(\mathbb{E}_Q f(z_i))^2$. Then, by a symmetrization technique similar to that in Eq. (17), we get

$$\frac{1}{4}\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa}\mathcal{L}_{\mathcal{D}}(g) - \hat{\mathcal{L}}_S(g) - ch^2\hat{\mathcal{L}}_S(g^2) \geq t\right) \leq \mathbb{P}_{S,\boldsymbol{\epsilon}}\left(\sup_{g\in\mathcal{G}_\kappa}\left[\frac{1}{m}\sum_{i=1}^m \epsilon_i\left(g(z_i) + c'h^2 g^2(z_i)\right) - c''h^2\hat{\mathcal{L}}_S\left(g^2\right)\right] \geq \frac{t}{4}\right)$$

where $c' = \frac{c+c_2}{2}, c'' = \frac{c-c_2}{2}$. This inequality enables the use of KL's Legendre transform, as in Eq. (23). The bound for $[0,1]$-valued loss can then be derived following similar techniques in the paper. We're happy to include these details or leave them out as the reviewers see fit. For unbounded losses, it might be possible to extend our results under a sub-Gaussian assumption, but we prefer not to speculate. In the revised paper, we discuss extensions to general loss functions and cite Alquier and Guedj (2018).

4) Connections between Eq. (9) and Lemma 1 of (Guedj, 2019) (R2): Thanks for pointing out this connection. We now cite Guedj (2019) in our revision.

5) Comparison with Tolstikhin and Seldin, 2013 (R2): Thank you for pointing out this missing reference. Note that the empirical variance (as appears in Thm. 4 of Tolstikhin and Seldin, 2013) and our "flatness" are distinct. The former is $\mathbb{E}_Q\frac{1}{m}\sum_{i=1}^m[f(z_i) - \frac{1}{m}\sum_{i=1}^m f(z_i)]^2$, while our "flatness" is $\frac{1}{m}\sum_{i=1}^m\mathbb{E}_Q[f(z_i) - \mathbb{E}_Q f(z_i)]^2$. It is possible for the second quantity ("flatness") to be zero, even when the first quantity is large. We now cite Tolstikhin and Seldin (2013) and highlight the relationship.

6) Connections to Grunwald and Mehta, 2019 (R3): Grunwald and Mehta propose a novel notion of complexity in terms of a "luckiness function", which generalizes the "prior" in PAC-Bayes and unifies the classical Rademacher complexity bound for ERM into the same framework. On the other hand, the paper is not directly related to deriving PAC-Bayes bounds using Rademacher-process approaches, hence not comparable to our work. However, it is certainly of great interest to study if our PAC-Bayes work can be extended to their more general framework in terms of "luckiness functions". We will add the discussions in the revised paper.

7) Connections to Dziugaite and Roy, 2017 (R2 & R3): We modified the code of Dziugaite and Roy (2017) and determined that the posterior they find is not "h-flat" in our sense. After some investigation, we believe the reason is that they are optimizing a PAC-Bayes bound and due to the poor prior choice, they underfit, and as a result, the posterior they find corresponds to a Gaussian with large variance for many parameters that are essentially "useless". We think investigating this and other empirical questions further is an interesting and open avenue of research, though well beyond the scope of this paper.

8) How is the proposed approach "more appropriate than "classical" approaches"? (R3): We're not entirely clear on the question, but here is our best attempt. We're happy to add further discussion if the reviewer can expand their question in their update. We cannot at present derive our "flatness" bound by a direct PAC-Bayes approach, without going through the Rademacher argument. We now raise this as an open problem.

9) Choosing a Dirac mass as a posterior would enable comparison with ERM (R2): In order for there to exist a (data-independent) prior P that, with high probability, dominates a Dirac mass concentrated on a random point η (thus yielding a finite KL divergence term), η must lie in a countable set with high probability. In general, ERMs do not satisfy this property. In order to study ERM using PAC-Bayes bounds, one usually relates the risk and empirical risk of a Gibbs classifier to the ERM. Standard approaches exploit margin. Herbrich and Graepel (2001) is a classical reference.

10) Minor Issues and Missing Citations (R1 & R2 & R3): We thank the reviewers for their comments and suggestions. We have corrected all typos and missing citations in our revisions.