We warmly thank the reviewers for their time and for sharing valuable feedback, we will use it to improve our paper. Overall, we are encouraged by the reviewers' reactions. Next we summarize our contributions at a high level, and then we address each reviewer's comments separately. We hope that our responses clear out any remaining concerns.

This paper should be regarded as a theoretical contribution that takes a critical stand on the "usual assumptions" on which PAC-Bayes inequalities are based (i.e. (a) bounded loss, (b) data-free prior, and (c) i.i.d. data observations), clarifying their role and illustrating how to obtain PAC-Bayes inequalities in cases where these assumptions are removed. Importantly, our work enables new PAC-Bayes inequalities with data-dependent priors. Furthermore, our paper contributes a unified approach to understanding PAC-Bayes inequalities, and their distinctions. Our paper also makes a case for the usefulness of formalizing data-dependent distributions as stochastic kernels, which may bring a new perspective to the PAC-Bayes literature, in spite of the fact that stochastic kernels are well-known in other literature, namely, on stochastic processes and their applications. We did also aim to write an informative and easy-to-read paper that may help the larger machine learning community to connect with the PAC-Bayes literature.

Response to Reviewer 1

We are grateful for your comments on the notation, we gladly will address this in the revised paper. Also we appreciate your feedback about the need to elaborate more on the novelty and relevance of our paper. We will add discussions about the new kinds of results one can obtain now that were not possible before, including rates where relevant. We will add a remark that there is indeed a direct relationship between moments of the loss and concentration of eigenvalues of the covariance matrix in the least-squares example. Consider a simple noise-free linear regression $Y = \mathbf{X}^\top \boldsymbol{w}^\star$, a "ridge" prior $p^0(\boldsymbol{w}) \propto e^{-\lambda \|\boldsymbol{w}\|^2}$, and let $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_i \mathbf{X}_i \mathbf{X}_i^\top - \mathbb{E}[\mathbf{X}_1 \mathbf{X}_1^\top]$. Then, for the log-exponential moment of the loss we have the identity $\log \mathbb{E} \int e^{L(\boldsymbol{w}) - \hat{L}_S(\boldsymbol{w})} p^0(\mathrm{d}\boldsymbol{w}) = \frac{d}{2}\log(2\pi) + \log \mathbb{E}|\det(\hat{\boldsymbol{\Sigma}}_0 + \lambda \mathbf{I})|^{-\frac{1}{2}}$ which shows equivalence of concentration of eigenvalues of $\hat{\boldsymbol{\Sigma}}_0$ and concentration of the loss since $\lambda_i(\hat{\boldsymbol{\Sigma}}_0) \to 0$ as $n \to \infty$ for i.i.d. instances.

Response to Reviewer 2

We will do a better job at discussing the usual formulation (for all posterior distributions) versus the formulation in terms of a stochastic kernel. Their equivalence can be illustrated by discussing specific stochastic kernels. Note that each theorem holds for an arbitrary kernel, so in that sense it holds "for all posteriors" as per the usual formulation in the literature. On the other hand, as we mentioned in our paper, we think that it is important to make explicit the data-dependence of the distributions, which is what the stochastic kernel formulation does. We will add a comment that the mathematical proof of our theorem is valid for (convex) $F : \mathbb{R}^k \to \mathbb{R}$ with an arbitrary k, while $k = 2$ is relevant for the known PAC-Bayes bounds. However, applications with $k > 2$ might emerge in the future, which is why we think it is interesting to point out our theorem's validity for arbitrary k, as this could enable new results. We are aware that the refined form of Pinsker inequality ($\mathrm{kl}(p\|q) \geq (p-q)^2/(2q)$, for $p < q$) is stronger than the other one only when $q < 0.25$, we apologize for the imprecision, this will be clarified in the revised paper. Thanks for pointing out the missing reference Blanchard & Fleuret (2007), we will make sure to include it in the revised paper.

Response to Reviewer 3

Many thanks for your positive reaction to our work! Responding to the four raised points regarding the related literature: (1) & (3) We will discuss in the introduction the work of Catoni (2007) in connection with data-dependent priors (regular conditional probability distributions), and accordingly instead of claiming "firsts" we will highlight instead that our work makes a case for the usefulness of representing data-dependent distributions as stochastic kernels, with attribution to Catoni (2007) and Alquier (2008) as predecessors. (2) We will comment on our main theorem being valid under general data-generation assumptions (i.e. not restricted to (c) i.i.d. data). (We are aware of this, sadly we missed commenting on it, but happy to fix it.) We'll expand the coverage of literature on PAC-Bayes bounds for non-i.i.d. data, including Alquier & Wittenberger (2012). The pointer to Rio's inequality is most helpful, we really appreciate this! (4) We meant that in some cases we can have concentration of the smallest eigenvalue of a sample covariance matrix even for unbounded instances, our reference is Section 5.4.2 of Vershynin (2011), discussing eigenvalue concentration of a sample covariance matrix with heavy-tailed observations. We will add discussion about Holland (2019) too.

Response to Reviewer 4

We'd like to clarify that our Theorem 2 is not restricted to convex losses, only the function F is restricted to be convex (e.g. $F(x,y) = c(x-y)^2$ was used with $x = \hat{L}_S(h)$ and $y = L(h)$), but the theorem is valid e.g. for the 01 loss or for the ramp loss, which are non-convex. We are glad for the question, we will emphasize this in the revised paper. Also, the notation "KL" (upper-case) is the for KL divergence in the generic case (between any pair of distributions), while "kl" (lower-case) is reserved for Bernoulli distributions. We think this notation is helpful as a visual aid, which has been used by other authors as well, but of course we can insert reminders throughout the paper to help the reader.