Response to Reviewer 1: We appreciate your valuable & insightful comments and suggestions.

Data normalization. Yes, it is an issue which is worth more discussion. In this paper, we assumed that all data instances have unit l_2 norm. We adopted this assumption partly due to a seemly common observation that instance normalization often improves classification/regression/clustering results. For example, if one examines the datasets in the LIBSVM (cited as [5]) website, for most of the datasets, all instances are normalized to have unit l_2 norm. There are also similar observations in the deep learning literature; see for example, the paper by Ulyanov et.al.: *Instance Normalization: The Missing Ingredient for Fast Stylization*, arXiv:1607.08022.

Even if the instance norms are not 1, one can often assume they are known because that only requires storing one real number per instance. With known norms, LM quantization is essentially the same, that is, we quantize data by scaling the quantizer according to the norm of each vector. In some application (e.g., regression), assuming instance normalization simplified the analysis as one does not have to keep track of the norms in the calculations. We will expand the discussions on the impact of the assumption of unit norm. Thank you.

Debiased variance. Yes, this is an interesting problem and we believe it should be possible to design quantizers that aim at reducing debiased variance. And yes, this is a meaningful and interesting topic to study. Thanks for pointing this out.

Additional experiments on regression. Thanks for the nice suggestion. Figure 1 is a simulated result of OLS. We report test mean squared error (MSE) of fitting OLS using different strategies. For uniform quantizer, we set the largest finite boarders equal to corresponding LM quantizer to make fair comparison. LM outperforms uniform quantization on this task. More results of this kind can be reported in the supplementary material, as you kindly suggested.

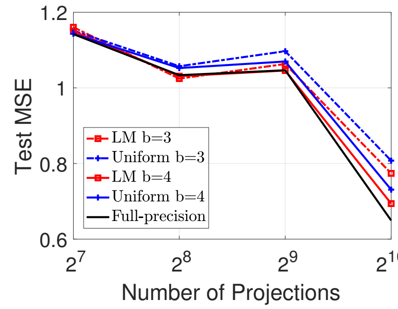

Figure 1: Simulation on OLS problem. $n = 3000$, $d = 1200$. Both X and β are generated from iid $N(0, 1)$ and X is normalized. Gaussian noise variance $\sigma^2 = 0.2$.

Response to Reviewer 2: Thanks for your valuable comments and suggestions. We would like to elaborate on the motivation of "compressed + quantized" learning. It should be now clear that "compressed learning" is a popular topic in the past 10 years, with many good papers in premier conference proceedings and journals, for a wide range of applications: similarity search, clustering, classification, regression, etc. Because one will have to store/transmit the compressed data and use them for subsequent calculations, it is a natural step to consider quantized version of compressed learning. Besides the papers written by Boufounos and collaborators, there is already a fairly rich literature on quantized random projections, often in non-machine-learning venues, for example,

Quantized Compressive Sensing, by Zymnis, Boyd and Candès, IEEE Signal Process. Lett., 2010; and
Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach by Plan and Vershynin, IEEE Trans. Information Theory, 2013, among other papers written by prominent researchers.

Before this submission, theoretical analysis, especially on learning performance using quantized compressive data, has not been conducted yet. In recent years, as data size becomes larger and larger, data compression is becoming more and more important. Thus, we hope our work will be useful both theoretically and practically.

Also, thanks for suggesting to exploit the trade-off between number of bits, number of projections, and accuracy. Since the derived bounds are functions of these parameters, we agree it is beneficial to generate plots to show the trade-off.

Response to Reviewer 3: Thanks so much for raising the interesting and very important issue regarding the definition of "generalization error". We agree the definitions in the literature are not always consistent. You are correct that Theorem 2 and Theorem 3 are about "Bayes Risk". Interestingly, "Bayes risk" in the context of near-neighbor classifiers is sometimes also referred to as "generalization error"; see reference [25], the well-known textbook in machine learning:

Understanding Machine Learning: From Theory to Algorithms, by Shai Shalev-Shwartz and Shai Ben-David. 2014

Chapter 19.2.1, entitled "*A Generalization Bound for the 1-NN Rule*" , derives the "generalization bounds" for 1-NN classifiers, where the RHS terms are indeed "Bayes Risk". The book is available online. While we are not allowed to provide links here, it is fairly easy to find. We certainly do not mean to "blame on" this nice book. We agree with Reviewer 3 that one should be more consistent with the definitions. We will think carefully what might be a more suitable title. Perhaps simply removing "Generalizing" from the current title might be an option?

The issue is similar for regression. There are quite a few papers which called regression test error as "generalization error", e.g., see *Compressed Least-Squares Regression*, by Odalric-Ambrym Maillard and Rémi Munos, NIPS 2009.

Again, we do not blame on prior papers for the inconsistency regarding definitions. We will think about this issue carefully and might also consult other experts. Thanks also for other suggestions on improving the quality of the work.