Research goals: In most learning problems in practice, optimizing the original loss is computationally intractable. One common method for addressing this problem is to instead optimize a surrogate loss function. Design a surrogate loss with good learning guarantees is a fundamental question in learning theory that directly relates to the design of algorithms. One desirable property of a surrogate loss function is consistency (also know as Bayes-consistency), which requires that asymptotically nearly optimal minimizers of the surrogate excess error also nearly optimally minimize the target excess error.

This paper establishes a systematic approach (an embedding framework) to design and analysis consistency of a kind of convex surrogates – polyhedral (piecewise-linear convex) surrogate losses. The approach consists of two steps: finding an embedding or embedded loss, and constructing a calibrated link function. The authors not only establish a series of fruitful theory underpinning the framework, but also describe in detail the applicability of the framework to the concrete examples. Moreover, the consistent polyhedral surrogates obtained by this systematic approach benefit from a linear excess error bound, which is the best rate one can achieve for general surrogate losses.

Significance: The paper constitutes a significant, technically correct contribution to the learning theory (more precisely, the statistical consistency theory) of loss functions. The most prior published work in this field has figured out ad-hoc derivations of consistent surrogate loss functions for many leaning problems, while an systematic and general approach is urgently needed to the best of my knowledge. The paper advances the current state of understanding of surrogate losses from the two points of view: the use of polyhedral losses and the systematic construction of consistency for such a kind of losses. In particular, the authors show the existence of consistent polyhedral losses with linear excess error bound for general discreet target loss, which motives the use of polyhedral losses from the statistical consistency view. The authors also provide an constructive procedures for finding such losses, which can be widely applied to the learning problems in practice.

Although currently, in contrast to the differentiable and smooth losses, the optimization of the polyhedral losses (in particular with the neural networks) remains a open question. Also, the complexity and challenges of using this systemic construction procedures depend on the specific problems. I believe the paper has enough and solid contributions for JMLR and the study of theses questions can be left as future work.

Evaluation: All claims are clearly articulated and supported by theoretical analyses. I have gone through all the proofs at various levels of depth, and the results/proofs are sensible and sound. Besides, I found that the paper will be more accessible to readers after replacing the informal analysis in the paragraph with a formal proof for some important corollaries, such as Corollary 20 as suggested in "Miscellaneous minor issues". In particular, adding such a formal proof will help understand the relation of different theoretical results,

e.g., Lemma 10, Lemma 19, Proposition 11, etc.

Clarity: Although this paper looks theoretically dense at glance, it was a very enjoyable experience to read this paper. The authors did a commendable job to discuss the strong motivation and give intuitive explanations for most theoretical results presented in this paper, such as before or after all their main theorems. I believe an interested reader with a background in machine learning, but no special knowledge of the paper's subject, could understand and appreciate the paper's results. Besides, I suggest a list of minor changes in "Miscellaneous minor issues " that might help further improve the quality.

Related work and discussion: I appreciate that authors have done an extensive literature research on the field – calibration and consistency of loss functions. Unfortunately, some recent works listed below in this field are missing and should be added in the revised version.

1. P. Long and R. Servedio. Consistency versus realizable H-consistency for multiclass classification. In International Conference on Machine Learning, pages 801–809, 2013.

2. Kuznetsov, V., Mohri, M., and Syed, U. Multi-class deep boosting. In Advances in Neural Information Processing Systems, pp. 2501–2509, 2014.

3. Zhang, M. and Agarwal, S. Bayes consistency vs. H-consistency: The interplay between surrogate loss func-tions and the scoring function class. In Advances in Neu-ral Information Processing Systems, pp. 16927–16936, 2020.

4. P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. In Advances in Neural Information Processing Systems, pages 9804–9815, 2021a.

5. P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. A finer calibration analysis for adversarial robustness. arXiv preprint arXiv:2105.01550, 2021b.

6. P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds for surrogate loss minimizers. In International Conference on Machine Learning, pages 1117–1174, 2022a.

7. P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Multi-class H-consistency bounds. In Advances in Neural Information Processing Systems, 2022b.

These works study H-consistency, that is, consistency accounting for the hypothesis sets adopted. More precisely, [1, 2] studied H-consistency in the realizable case. [3] designed a new family of functions that could help H-consistency in view of the phenomenon discussed in [1]. [4,5] studied H-calibration and H-consistency for adversarial robustness. [6,7] proposed and studied the corresponding quantitative relation of H-consistency bounds in both non-adversarial

and adversarial cases. In summary, these works all shed light on the design and analysis of calibrated and consistent surrogate losses and should be included in the related work.

In these cases, calibration and consistency are not equivalent in general. It would be interesting to see if the framework in the paper could be helpful and be generalized to the H-consistency case. It would be great if authors could add a discussion on this, in parallel to that on the other interesting directions listed in the conclusion section.

Questions and comments:

1. More explanation on why a weak inequality in the definition of separated link (Definition 15) is more natural in applications such as hinge loss for binary classification.

2. Excess error (regret) bounds vs consistency. It is interesting to see that any asymptotic consistent polyhedral loss admits a quantitative excess error bound (as shown in Theorem 18). I feel in general we can say the quantitative excess error bound can imply consistency while the opposite direction seems not true. Is it possible to give an example of the case where a surrogate loss (certainly not polyhedral) is consistent but do not admit an excess error bound? If so, I think Theorem 18 can be viewed as a very crucial property for the polyhedral losses and can be valued more in the learning theory of loss functions.

3. Optimization of polyhedral losses. It is interesting to see that the polyhedral losses would be a good choice from the statistical consistency view. But, a polyhedral loss like hinge loss may not be differentiable and the optimization of such losses would be more difficult than the smooth losses, in particular when using the modern neural networks in practice. I wonder what authors think about this. It would be interesting to incorporate the optimization to compare the choice of surrogate losses. The paragraph in the conclusion does not answer this question very well.

4. The example in section 5.4 and non-smoothness of polyhedral losses. The Weston-Watkins hinge loss is not consistent with respect to the multi-class zero-one loss because hinge loss is not differentiable at zero [1,2]. I am wondering if the embedding framework and the construction (such as the proof of Theorem 14 and Theorem 2) proposed in this paper are able to obtain a Weston-Watkins polyhedral loss that is consistent with respect to the zero-one loss without any assumptions on the distribution in spite of the non-smoothness of polyhedral losses.

   In the example 5.4, it is interesting to see that the embedding framework can be used to show that Weston-Watkins hinge loss is consistent with respect to the ordered partition loss. But, it seems not clear to see if the framework can be used to obtain a new Weston-Watkins polyhedral loss consistent with respect to the original zero-one loss.

[1] T. Zhang. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 2004.

[2] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. Journal of Machine Learning Research, 2007.

5. Proposition 21 and upper bounding the original loss. Proposition 21 shows that the equivalent condition to an embedding is matching Bayes risks. This result is a bit surprising in view that the common surrogates always upper bound the original loss and their Bayes risks can not match. For example, the hinge loss uppers bounds the zero-one loss and their Bayes risks are not equal. I am wondering if proposition 21 could imply that the consistent surrogate polyhedral losses derived using the embedding framework in the paper actually do not upper bound the original loss?

6. Theorem 33 shows that every possible calibrated link must be produced from construction 2. I am wondering if a similar result holds for construction 1 and embedding?

Miscellaneous minor issues and suggestions:

- Change "label distribution" to "conditional distribution" at the beginning of section 2.2. The "label distribution" often refers to the marginal distribution on the label space.
- Add a definition of "link function"which first appears at the beginning of section 2.3.
- Change "expected loss" to "conditional loss", e.g., at the beginning of section 2.3. The term "expected loss" often refers to the one taking expectation on both input space and label space.
- Change the notation of hypothesis H_1, H_2 above the Definition 7 to h_1, h_2. H is often used to denote the hypothesis set instead of hypothesis.
- Add a formal definition of "consistency" instead of describing it in the paragraph above Definition 7.
- Add the definition of sgn appearing at the end of page 6, specifying its value on 0.
- Restate Theorem 1 after Theorem 14. The `thm-restate` package may be helpful here. Also, restate Theorem 2 in Section 4.
- Add a formal definition of the "calibrated link" which first appears at the end of page 10.
- Change the notation of excess error R to ΔR, e.g., at the beginning of section 4.3. R often refers to the generalization error instead of the excess error.

- Change “a variant of multi-class classification” to “a variant of binary and multi-class classification” at the beginning of section 5.2. Some papers cited here are only for the binary case.
- Correct the typo - “Yu and Blaschko” appears twice continuously at the end of page 15.
- Add a formal proof for Corollary 20. This is a very important result. A formal proof will help understand the relation of Lemma 10, Lemma 19, Proposition 11, etc.
- Proposition 21 and 23 also provide a equivalent condition that L embeds a discrete loss. These conditions should be summarized together into Corollary 20.

Recommendation: Overall, I believe the paper did a great job except some minor issues. Therefore, I recommend conditional accept with the following list of changes that can be checked upon resubmission.

- Add the missing references along with a discussion on them as suggested in Related work and discussion.
- Some replies to questions and comments listed above.
- Address the miscellaneous minor issues listed above.