Response Letter
'Implicitly Constrained Semi-Supervised Least Squares Classification'

Jesse H. Krijthe & Marco Loog

First of all, we would like to thank all the reviewers and the editor for reading the paper and providing constructive comments that have certainly improved the paper. We are convinced that the changes we have made to the manuscript have taken away many of the concerns the reviewers had.

One of the main results in our manuscript is a proof that in the one univariate setting our proposed semi-supervised extension of the least squares classifier, which we call implicitly constrained least squares classification (ICLS), always improves the performance in terms of the squared loss, when compared to the supervised solution. An important property of this procedure is that it does not make use of additional assumptions, like the clustering or low density separation assumptions that many semi-supervised approaches do. The point of our work is to show through this result that it is possible to construct semi-supervised learners with strong performance guarantees that do not rely on additional assumptions.

During the time this work was under review we were able to extend this one dimensional result to a proof of improvement in terms of squared loss in the transductive sense for the multivariate case for an adaptation of the ICLS procedure, to complete our 1D proof in Theorem 1. We are not aware of any such results in the semi-supervised/transductive learning literature. We have updated the rest of the manuscript accordingly to reflect the addition of this important result. In particular we have rewritten the theoretical results in section 4.2 and added 4.3. The experiments section and discussion have been extended to include results for both the adapted ICLS procedure using Euclidean distances that is used in Theorem 2 and the extended ICLS procedure based on a generalized Euclidean distance used in the multivariate proof of Theorem 3. Since the proof of the multivariate case was also one of the suggestions of one of the reviewers, we hope they will appreciate our updated manuscript.

Aside from these major improvements we have updated to manuscript in many places to improve clarity and address specific comments by the reviewers. We will now cover these specific concerns brought forth by the reviewers and the subsequent changes to the manuscript one by one. The reviewers' comments are indicated in **bold**, followed by our response.

I. Response to Editors comments

I think the paper contains an interesting idea that could be further clarified and generalized. Specifically, a major limitation is the focus on least squares classification, which is perhaps antiquated compared to more "modern" classification methods. The idea and theory is quite clear and crisp for this specific model. Does it apply broadly to more models? I suspect that it depends: for example, for logistic regression this idea may not apply as the hidden response is not bounded; but what about others? Essentially, significantly expanding on the last two paragraphs of section 6, discussing them upfront, or even adding another methods beyond least squares classification that benefits from the idea, can improve the paper.

We agree with your comments that the results could be generalized further. During the time this work was under review, we have studied this question, especially in the context of likelihood based models such as linear discriminant analysis (LDA). For this specific case we found that empirically similar improvements are possible. It is important to note, however, that for LDA we have so far not been able to derive any theoretical results like those in this manuscript. For the specific case of logistic loss, it indeed seems likely that the current definition of the constrained space will not bound the solution in any way. A further complication is that logistic regression does not have a closed form solution, which is used in our current derivation for the least squares classifier.

As per your suggestion, we have extended the discussion in section 6 to discuss these issues. The point we want to get across with this work is the interesting observation that for some models we can do robust semi-supervised learning without the traditional assumptions considered to be required for semi-supervised learning. It is certainly important to emphasize least squares is just one, theoretically interesting, example of this, which we hope we have now done.

Specifically we added the following paragraphs to Section 6:

While the results presented in this work are promising for squared loss, a worthwhile extension would be to other loss functions. In this work, we were able to derive a quadratic programming formulation for ICLS because there is a closed-form solution of the supervised least squares problem. For many loss functions, closed-form solutions do not exist, which prohibits a straightforward formulation of their implicitly constrained semi-supervised counterparts. In particular, in the derivation of ICLS, we made use of the closed-form solution given an imputed labeling to derive a quadratic programming problem in terms of the labels. Without a closed form solution, one of the main difficulties is that, even if the loss considered is differentiable, one cannot straightaway apply techniques like gradient descent to the parameters as this typically leads to solutions that are outside of the set $\mathcal{C}_\beta$.

Besides these issues, there is the other open question of what loss functions could benefit from constraining the solution to a subset like $\mathcal{C}_\beta$ in the first place. For logistic loss, for instance, the use of the logistic function ensures that posteriors are always between $[0, 1]$. In that case it seems that the current definition of $\mathcal{C}_\beta$ does not constrain the solutions at all. Treating negative log likelihood as a loss function, on the other hand, does lead to interesting semi-supervised classifiers, for instance in linear discriminant analysis [1]. Even when $\mathcal{C}_\beta$ does not constrain the solution, we could still be able to construct other constrained sets with interesting performance guarantees. For instance, many semi-supervised classifiers make the additional assumption that the mean posterior labeling of the unlabeled objects is similar to the observed label prior in the labeled objects [2], [3] or is assumed to be known [4]. Rather than ensuring non degradation for every transductive set, such as the result in Theorem 3, such additions to the definition of $\mathcal{C}_\beta$ could lead to improvement over the supervised solution with high probability.

On a related note, it is worth discussing the SSL assumptions, or the lack thereof, of the proposed method further. It is perhaps interesting to consider a broader sense of "assumptions", where the hypothesis space is the assumption itself. This is highlighted in the recent book Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David.

Regarding the lack of assumptions, it is indeed generally true (as Shalev-Shwartz and Ben-David also argue [5]), that the hypothesis space is a very important part of the assumptions underlying any classifier. We have tried to emphasize the point that in the context of deriving a semi-supervised version of a supervised classifier, these assumptions that are already made by the choice of supervised classifier may already be enough to leverage them to construct a semi-supervised variant. The additional constraint that we introduce on the hypothesis space, only relies on the additional assumption that the loss of the supervised classifier would improve

if we add more labeled data. This seems very reasonable for most classifiers. If we can not hope to improve performance using labeled data, there is little hope in using the unlabeled data.

We have improved section 6 to clarify this point, specifically in this paragraph:

Some have argued that, for discriminative classifiers, semi-supervised learning is impossible without additional assumptions about the link between labeled and unlabeled objects [6], [7]. ICLS, however, is both a discriminative classifier and no explicit additional assumptions about this link are made. Any assumptions that are present follow, implicitly, from the choice of squared loss as the loss function. One could argue that constraining the solutions to $\mathcal{C}_\beta$ is an assumption as well. While this is true, it corresponds to a very weak assumption about the supervised classifier: that it will improve when we add additional labeled data. This lack of additional assumptions has another advantage: no additional parameters need to be correctly set for the results in Sections 4 and 5. There is, for instance, no weight to be chosen for the importance of the unlabeled data. Therefore, implicitly constrained semi-supervised learning is a very different approach to semi-supervised learning than the methods discussed in Section 2.

II. Response to Reviewer #1's comments

The reviewer would have expected the paper to cite and discuss "Unlabeled data: Now it helps, now it doesn't, Singh, Nowark and Zhu". The reviewer encourages you to do so in the final version.

We have included a short discussion in section 2 of how the results presented in "Unlabeled data: Now it helps, now it doesn't" by Singh, Nowark and Zhu relate to our work. While the goal of their work is similar in the way of exploring the limits of semi-supervised learning, their main contribution is studying what we may be able to prove once we make a clustering assumption, while one of the main insights offered by our work is that there are cases where we can do without this additional assumption, and rely on the choice of the supervised classifier alone. We do very much agree with their observation that the benefits of semi-supervised learning may be more fruitfully studied in a finite sample scenario. This directly relates to our results in Theorem 2 and 3, where the improvement are actually for finite samples.

Specifically, the following paragraph was added to section 2: *Some have argued unlabeled data can* only *help if $P(X)$ and $P(Y|X)$ are somehow linked [7]. The goal of our work is to show that in some cases (i.e. the least squares classifier) we do not need explicit assumptions about those links for semi-supervised learning to be possible. Instead, we leverage the implicit assumptions, including possible misspecification, that are already present in the supervised classifier. Similar to [7], however, we also study the finite sample case.*

III. Response to Reviewer #2's comments

This paper proposes a semi-supervised algorithm for least squares classification. The authors claim that existing semi-supervised methods usually use some assumptions (cluster, smoothness, low-density), and these assumptions may lead to performance degradation. To make the unlabeled data never hurt the performance in expectation, this paper tries to perform semi-supervised learning without additional assumption.

The proposed algorithm is called ICLS (implicitly constrained least squares). As in Equation (4), ICLS removes any additional assumption in the objective of semi-supervised learning, and just minimizes the squared loss on the labeled data with a constraint that the learned parameter β (i.e., the classification model) should be in a specific parameter set C (i.e., a hypothesis space). And the parameter set C consists of the β's that trained on all (labeled and unlabeled) examples going through all possible label assignments for the unlabeled examples.

I think the constraint in equation (4) is redundant. The detailed reason is as follows: First, let us denote by β^* the optimal parameter obtained by minimizing the squared loss on labeled data (the solution of equation (4) without the constraint). Then, β^* will give a prediction

for the unlabeled data, denoted by y_u, which is of cause a possible label assignment for the unlabeled data. Because β^* gets zero squared loss on (X_u, y_u), so it is still the optimal solution of minimizing squared loss on $([X, X_u], [y, y_u])$, and thus β^* belongs to C.

To summary, the β^* is always in the set C. Corresponding to Fig 1, $\hat{\beta}_{sup}$ is always in C_β. So the constraint in equation (4) is redundant, and thus the proposed algorithm is exactly the same with minimizing the squared loss on the labeled data. So indeed the proposed method does not use any assumption, but it neither use any information from the unlabeled data, it is exactly a supervised least squares classifier, which has no novelty or contribution. Based on the above analysis, I am really surprised that in the experiments ICLS performs better than the supervised least squares classifier (LS) significantly and consistently on all datasets. I think they should have the same results.

We would like to thank the reviewer for having an in depth look at the method, to identify a possible contradiction in the method and the results. We think, however, that the argument that the constraint is redundant is based on a misunderstanding of the supervised model, which we have tried to clarify in the new manuscript. The problem lies in the statement that "Then, β^* will give a prediction for the unlabeled data, denoted by y_u, which is of cause a possible label assignment for the unlabeled data. This is not true for the least squares classifier, since its output is unbounded. In essence, the output of $x^\top \beta^*$can be larger than 1 or smaller than 0. However, since we know the labels should be 0 or 1, or in the relaxed setting $[0, 1]$, we cannot assign labels that will give 0 loss on (X_u, y_u). Take for instance the very simple setting where we have two labeled objects in 1D, one point at $x = -1$ labeled 0 and one at $x = +1$ labeled 1. The supervised parameter estimate would give: $y = a + b \cdot x = 0.5 + 0.5 \cdot x$. Now suppose we have on unlabeled object at $x = 2$. There is no labeling possible such that we will again find the parameters a=0.5, b=0.5. In other words, thisβ^* will not be in the constrained set. We hope this issue is now clearer in the revised text of sections 3 and 4.

some typos

Thank you for pointing out the typos, which we have corrected in the new version of the manuscript.

IV. Response to Reviewer #3's comments

Thank you very much for the thorough review and thoughtful comments. We will address them one by one.

This work introduces a semi-supervised algorithm via optimizing the least square loss, whereas much existing work has focused on semi-supervised learning by designing efficient SVM algorithms. The authors should clearly illustrate the differences from previous work. More importantly, what's the benefit of using least square loss.

There are several benefits to least squares classification. First of all, it can be solved efficiently. Secondly, although the squared loss may not be as popular as some other techniques it has solid performance in many practical settings, see for example the works by [8], [9] and others referenced in the manuscript. In other words, in the supervised setting it leads to useful classifiers for many problems so it is worthwhile to study its extension to the semi-supervised setting. Thirdly, and most importantly for our purposes, is that it has a closed form solution. This allows for the proofs considered in the manuscript. To our knowledge, no similar results are available for other losses in the semi-supervised literature.

Regardless, the goal of our work is to show there is a classifier (in this case the least squares classifier), for which we can prove performance increases when using unlabeled data without needing additional assumptions. Our claim is that it is a very interesting observation that this semi-supervised version will always outperform, or at least not degrade the performance of, the supervised classifier. As indicated in the discussion section and in a response to one of the previous questions, it is clearly a question of interest how to extend the techniques introduced here to other loss function that do not have closed-form solutions.

In the updated manuscript we have emphasized this goal by discussing it earlier in the introduction and rewriting the abstract to make this point clearer: *We introduce a novel semi-supervised version of the least squares classifier. In implicitly constrained least squares (ICLS) classification, we minimize the squared loss on the labeled data among the set of parameters implied by all possible labelings of the unlabeled data. Unlike previous discriminative semi-supervised methods, our approach does not introduce explicit additional assumptions into the objective function, but leverages implicit assumptions already present in the choice of the supervised least squares classifier. We show this classifier can be formulated as a quadratic programming problem and its solution can be found using a simple gradient descent procedure. In a specific 1-dimensional case without intercept, we give an intuitive proof that this method can never lead to worse performance than supervised least squares classification. In the more general multidimensional case we prove that a slightly adapted procedure leads to guaranteed improvements in terms of both the parameter estimates and, more importantly, in terms of squared loss. The latter result illustrates that strong theoretical guarantees for semi-supervised procedures are possible, at least for specific classifiers and loss functions. Experimental results corroborate the theoretical results and indicate desirable properties over alternative semi-supervised approaches to least squares classification.*

The first theoretical result (Theorem 1) shows that the proposed procedure never performs worse than supervised learning in terms of squared loss, however, the theorem holds only for one dimension. The author should give a completed proof for general case, and it is a little trivial to discuss the one-dimension case.

While we agree the multidimensional case is even more interesting, we disagree the one-dimensional case is trivial. The proof gives insight into how and why this procedure works. In as far as this was not clear in the previous version of the manuscript, we hope we have improved the text to make this clearer.

While our manuscript was under review, we were able to find a proof for the general case for an extended version of our semi-supervised procedure. The biggest change in the new manuscript is that we have added this proof (Section 4.3) and updated the experiments to explore its empirical properties as well. The resulting theorem is a very strong result in that the procedure always gives lower loss than the supervised solution in the transductive setting (Theorem 3).

The second theoretical result (Theorem 2) shows improvement in parameter estimation for an adapted procedure in the multivariate case. Does Theorem 2 really show that the proposed procedure never performs worse than supervised learning in terms of squared loss? What the difference and gap?

Theorem 2 indeed did not show that there is improvement in terms of squared loss. It is an interesting question why this happens. In the new version of our manuscript we included the Euclidean projection, that is used in Theorem 2, in the cross-validation experiments. We hope it is clear from those results, that at least its solution is not the same as that of the original ICLS procedure. To illustrate why improvement in terms of Euclidean distance does not necessarily translate into improvement in terms of the squared loss, consider Figure IV in this response letter. The space in this figure is the parameter space for a 2D problem without intercept. The red point indicates the supervised solution, while the green point indicates the oracle solution, the parameters we would find if we would have the labels for all the (labeled and unlabeled) objects. The yellow point is the Euclidean projection of the supervised solution onto the constrained set (the black region). Measured from the oracle solution, the yellow point is closer to the oracle solution in terms of Euclidean distance. In terms of the squared loss, however, indicated by the blue isoline, the supervised solution is closer. This example indicates why Theorem 2 does not necessarily guarantees improvement in terms of squared loss. In the new Theorem 3, we do guarantee non-degradation in terms of squared loss. This estimator correspond to the blue point in Figure IV. In the updated manuscript, we have attempted to clarify the difference between the claims made by Theorem 2 and Theorem 3 by

also including them in the experimental section.

The authors should clearly give detailed proofs of Theorems 1 and 2, which is convenient for readers to understand this work.

We have attempted to improve the clarity of both proofs (Section 4.1 and 4.2 and the new proof in 4.3) in our new version of the manuscript, but welcome specific comments that could further improve the clarity. In particular, we have made the assumptions that we use in the proof more explicit inside Theorem 1.

Section 5 presents an empirical evaluation of the proposed approach on benchmark datasets. However, the authors do not make any comparisons with the state-of-the-art semi-supervised algorithms.

The goal of this work is to show that it is possible to build a semi-supervised classifier that is never worse than its supervised counterpart. The experiments are set up to verify to what extend the proposed procedure reaches this goal. We are not convinced that a comparison to other supervised and semi-supervised methods, that incorporate different losses, gives a clearer view of the point we explore in the manuscript. As such we think such a comparison, although interesting, is beyond the scope of this paper. We have attempted to make this goal clearer throughout the text.

V. Acknowledgements

We would like to thank all 3 reviewers and the editor for their thoughtful comments. Overal, we think the addition of non-degradation of performance in the multivariate case has improved the manuscript. We think the concept of implicitly constrained semi-supervised learning and the theorems in section 4 are important because it shows that for some classifiers, semi-supervised learning without additional constraints is possible. We are convinced that by addressing your comments, the new manuscript gets this point across more clearly.

References

[1] Krijthe, J.H., Loog, M.: Implicitly Constrained Semi-Supervised Linear Discriminant Analysis. In: International Conference on Pattern Recognition, Stockholm (2014)

[2] Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning, Morgan Kaufmann Publishers (1999) 200–209

[3] Collobert, R., Sinz, F., Weston, J., Bottou, L.: Large scale transductive SVMs. Journal of Machine Learning Research **7** (2006) 1687–1712

[4] Mann, G.S., McCallum, A.K.: Generalized expectation criteria for semi-supervised learning with weakly labeled data. Journal of Machine Learning Research **11** (2010) 955–984

[5] Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning (2014)

[6] Seeger, M.: Learning with labeled and unlabeled data. Technical report (2001)

[7] Singh, A., Nowak, R.D., Zhu, X.: Unlabeled data: Now it helps , now it doesnt. In: Advances in Neural Information Processing Systems. (2008) 1513–1520

[8] Poggio, T., Smale, S.: The Mathematics of Learning: Dealing with Data. Notices of the AMS (2003) 537–544

[9] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. (2005)

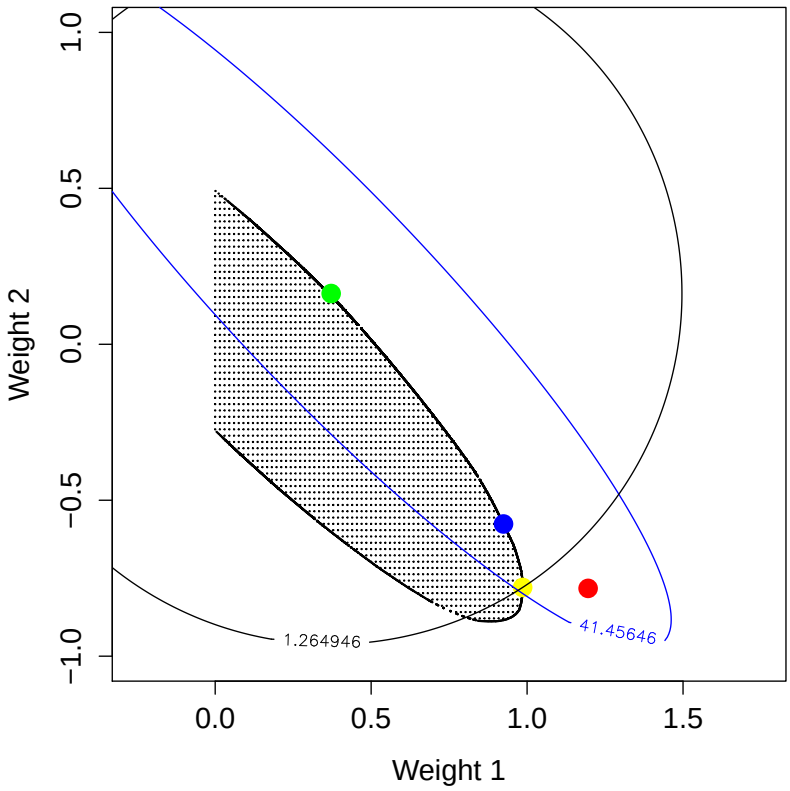

Fig. 1. An example were the Euclidean distance projection gives a worse solution. The black region is the constrained parameter space. The oracle solution, meaning the solution of the supervised classifier where all the labels are known, is shown in green. The supervised solution is indicated in red. Notice that the Euclidean projection (yellow), while having lower Euclidean distance to the oracle solution (black isoline), it has a higher loss (blue isoline) than the supervised solution.