Supplementary material for: "On the value of weak information of supervision: An empirical study"

Jerónimo HERNÁNDEZ-GONZÁLEZ [a], Aritz PÉREZ [b]

[a] *Dept. of Mathematics and Computer Science, University of Barcelona, Gran Via de les Corts Catalanes 585, Barcelona, Spain*

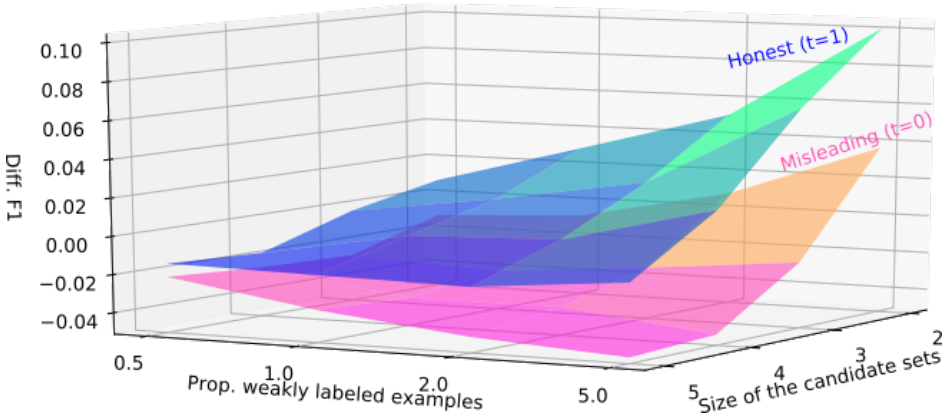
[b] *Basque Center for Applied Mathematics, Al. Mazarredo 14, Bilbao, Spain*

1. Additional figures

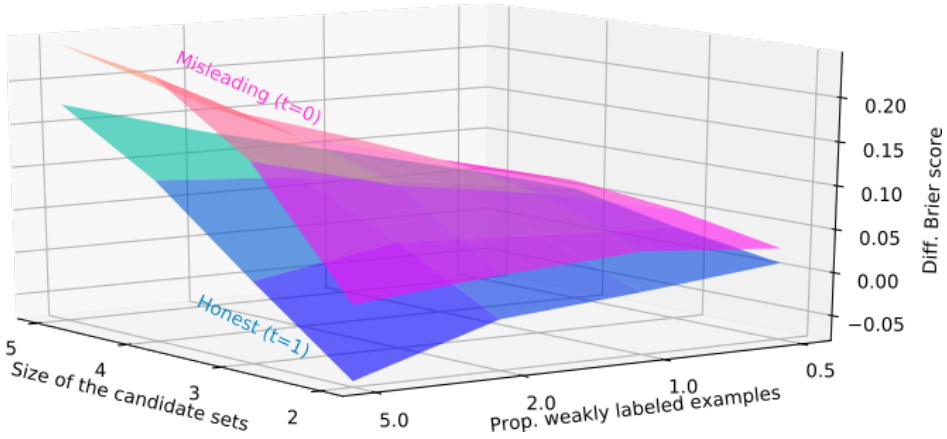
Following the same experimental setup than the one explained in the paper, we have carried out more experiments questioning the "correct model" assumption.

1.1. Experiments with complex generative models, and simple learning models

In the following figures, data was synthetically generated from a generative model of large complexity ($K = 4$), whereas the learnt model was simpler ($K = 1$). Firstly, in Figure 1 we show the ability to learn with weak supervision as the proportion of weakly supervised examples increases, and also as the size of the candidate sets increases. Figure 2 shows the performance of the learnt classifiers as the size of the labeled subset and that of the weakly supervised subset are increased.

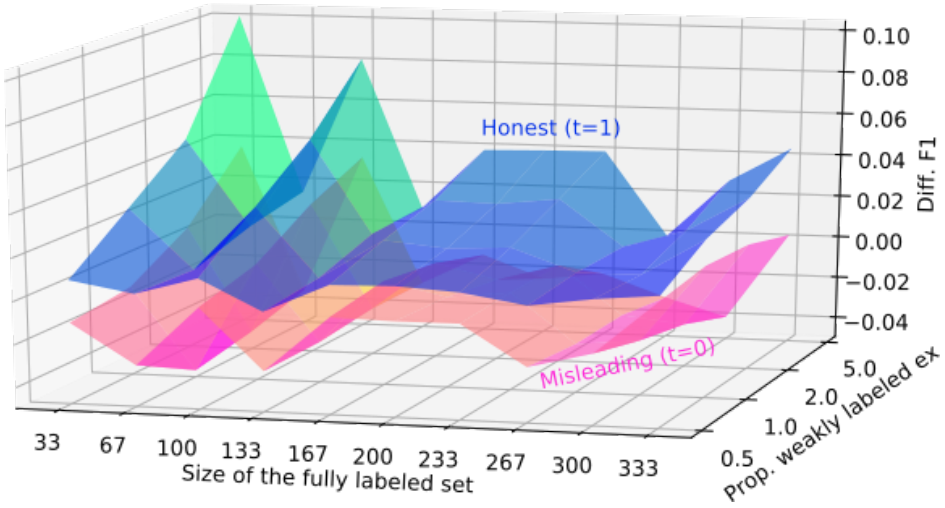


(a) Results in terms of macro F1

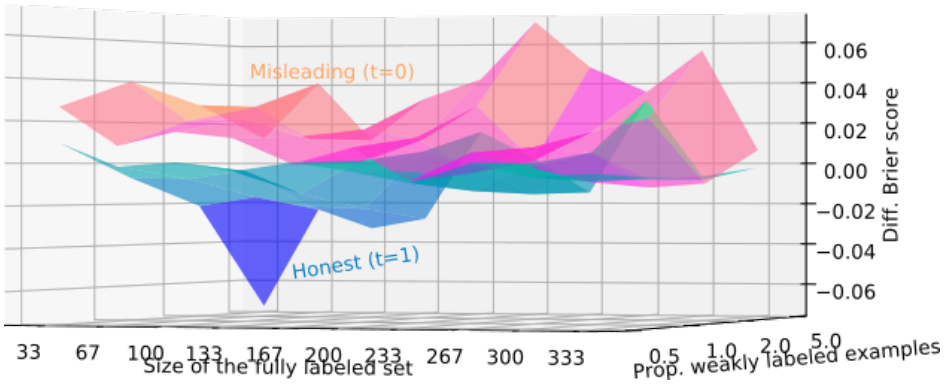


(b) Results in terms of Brier score

Figure 1. Performance of the classifiers learnt in different scenarios obtained from combining different sizes of the candidate sets and proportions of weakly labeled examples. In both figures, each surface represents experiments with misleading (labels in S are really improbable) or honest (labels in S are probable) labels, and shows the difference in terms of macro F1 or Brier score with respect to the results of a similar classifier learnt only with the labeled subset. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), simple generative and learnt models ($K = 1$), and no forced consistency ($s = 1$).



(a) Results in terms of macro F1

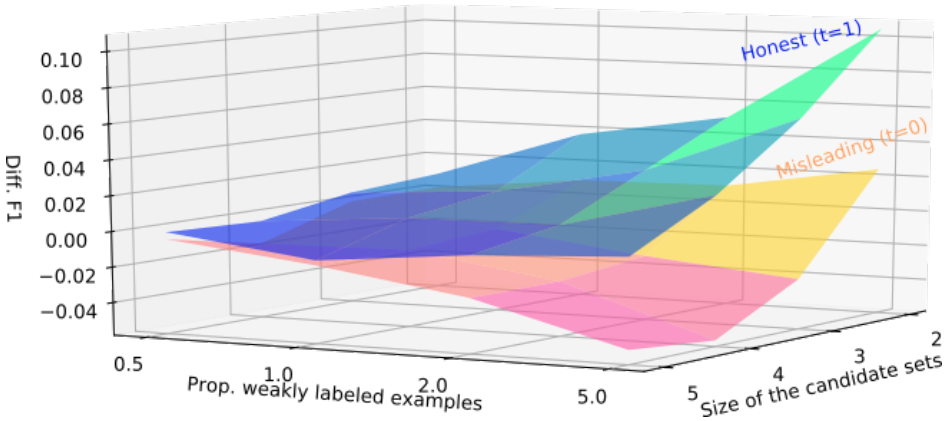


(b) Results in terms of Brier score

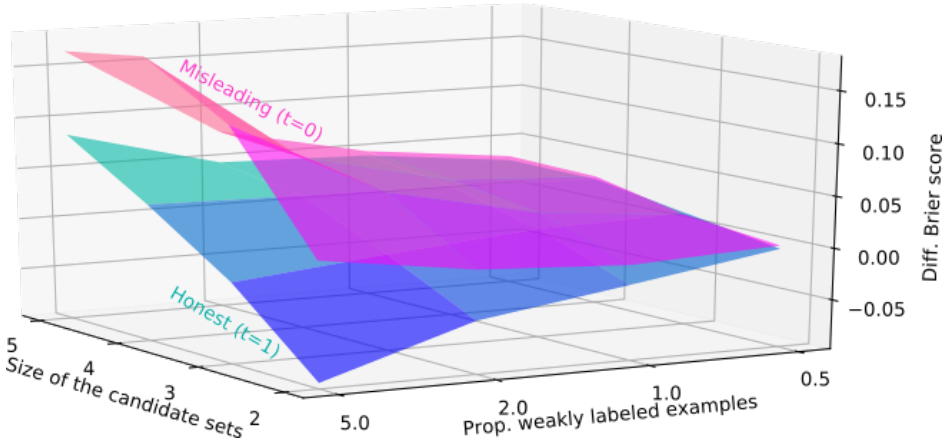
Figure 2. Performance of the classifiers learnt in different scenarios obtained from combining different amounts of fully labeled data and proportions of weakly labeled examples. In both figures, each surface represents experiments with misleading (labels in S are really improbable) or honest (labels in S are probable) labels, and shows the difference in terms of macro F1 or Brier score with respect to the results of a similar classifier learnt only with the labeled subset. Other parameters are fixed to small labelsets ($|S| = 2$), simple generative and learnt models ($K = 1$), and no forced consistency ($s = 1$).

1.2. Experiments with simple generative models, and large learning models

In the following figures, data was synthetically generated from a generative model of simple complexity ($K = 1$), whereas the learnt model was more complex ($K = 4$). Firstly, in Figure 3 we show the ability to learn with weak supervision as the proportion of weakly supervised examples increases, and also as the size of the candidate sets increases. Figure 4 shows the performance of the learnt classifiers as the size of the labeled subset and that of the weakly supervised subset are increased.

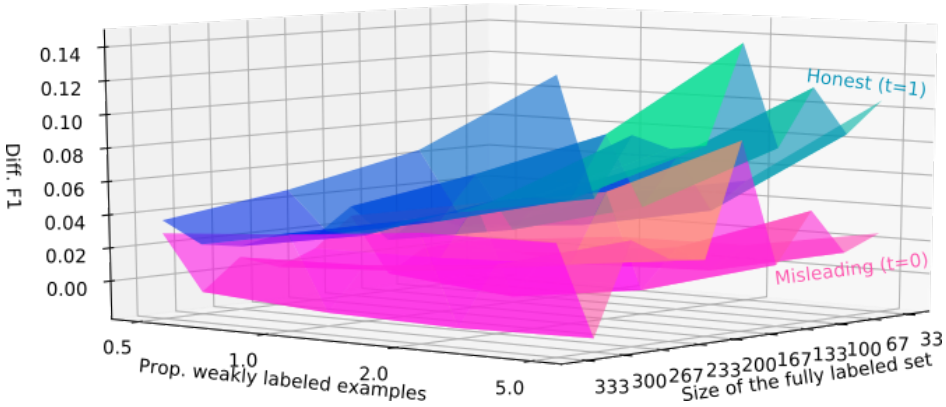


(a) Results in terms of macro F1

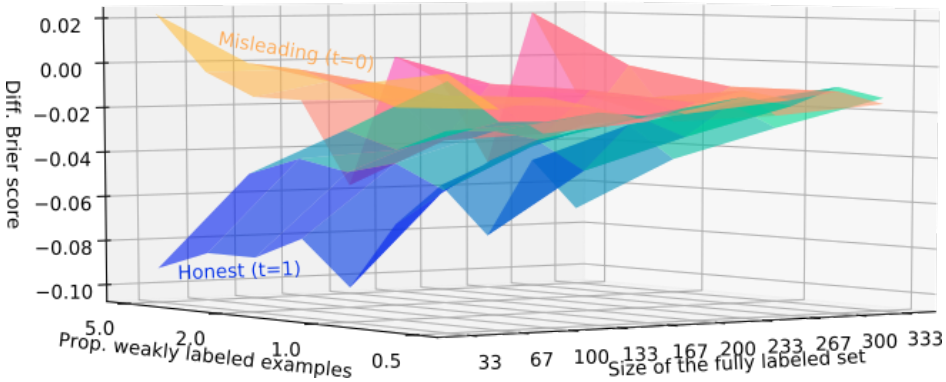


(b) Results in terms of Brier score

Figure 3. Performance of the classifiers learnt in different scenarios obtained from combining different sizes of the candidate sets and proportions of weakly labeled examples. In both figures, each surface represents experiments with misleading (labels in S are really improbable) or honest (labels in S are probable) labels, and shows the difference in terms of macro F1 or Brier score with respect to the results of a similar classifier learnt only with the labeled subset. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), simple generative and learnt models ($K = 1$), and no forced consistency ($s = 1$).



(a) Results in terms of macro F1

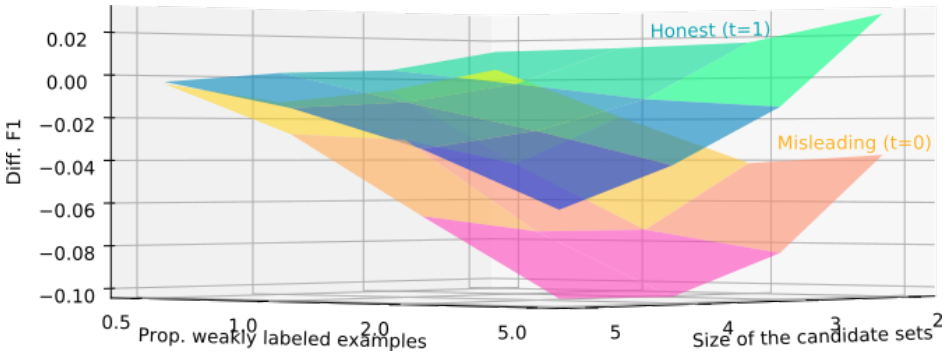


(b) Results in terms of Brier score

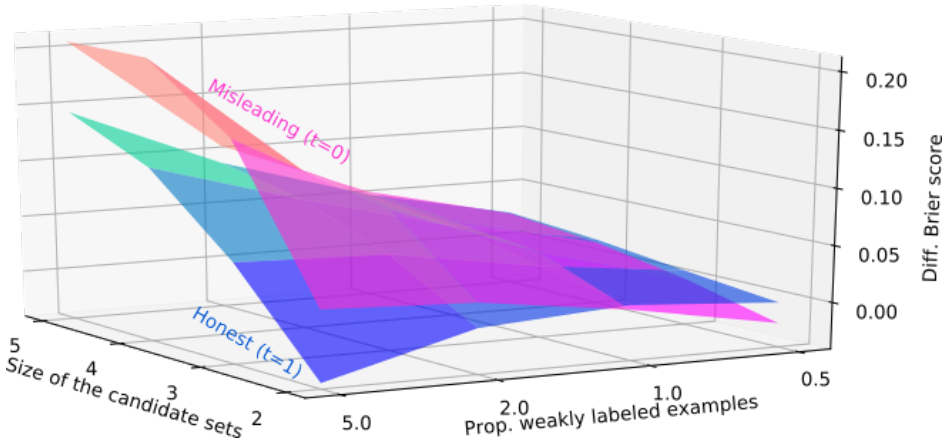
Figure 4. Performance of the classifiers learnt in different scenarios obtained from combining different amounts of fully labeled data and proportions of weakly labeled examples. In both figures, each surface represents experiments with misleading (labels in S are really improbable) or honest (labels in S are probable) labels, and shows the difference in terms of macro F1 or Brier score with respect to the results of a similar classifier learnt only with the labeled subset. Other parameters are fixed to small labelsets ($|S| = 2$), simple generative and learnt models ($K = 1$), and no forced consistency ($s = 1$).

1.3. Experiments with complex generative and learning models

In the following figures, data was synthetically generated from a generative model of large complexity ($K = 4$), and similar models were used for learning ($K = 4$). Note that in the paper we use also models of the same complexity for both generation and learning, but in that case with a simpler structure ($K = 1$). Firstly, in Figure 5 we show the ability to learn with weak supervision as the proportion of weakly supervised examples increases, and also as the size of the candidate sets increases. Figure 6 shows the performance of the learnt classifiers as the size of the labeled subset and that of the weakly supervised subset are increased.

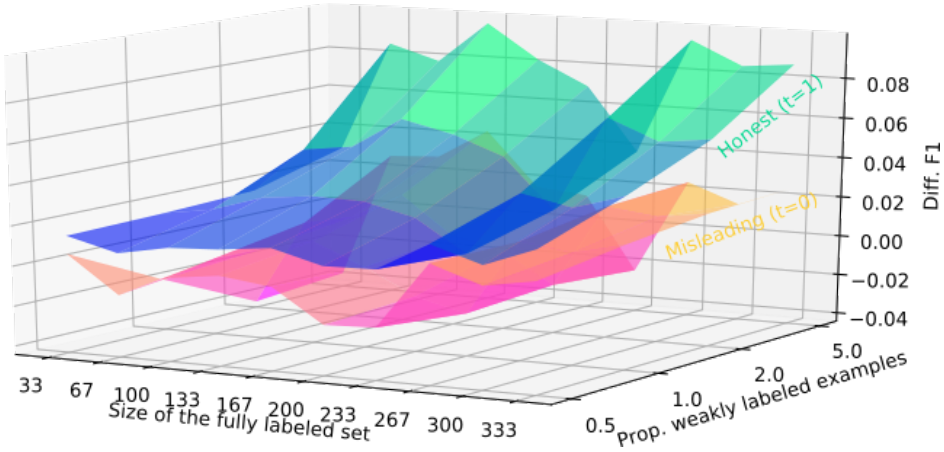


(a) Results in terms of macro F1

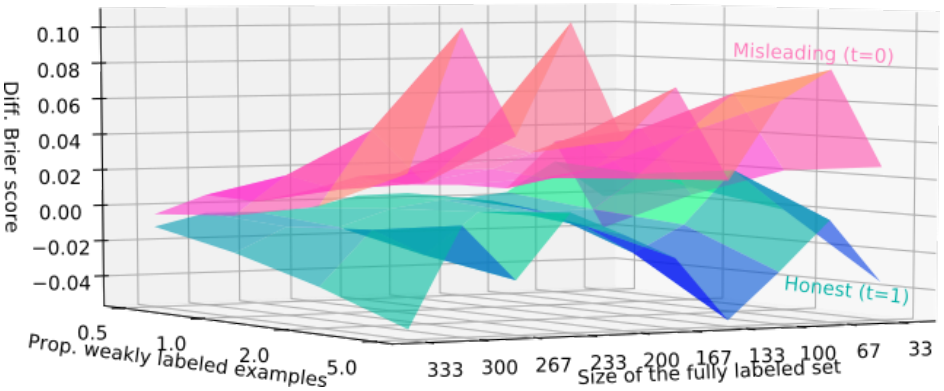


(b) Results in terms of Brier score

Figure 5. Performance of the classifiers learnt in different scenarios obtained from combining different sizes of the candidate sets and proportions of weakly labeled examples. In both figures, each surface represents experiments with misleading (labels in S are really improbable) or honest (labels in S are probable) labels, and shows the difference in terms of macro F1 or Brier score with respect to the results of a similar classifier learnt only with the labeled subset. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), simple generative and learnt models ($K = 1$), and no forced consistency ($s = 1$).



(a) Results in terms of macro F1



(b) Results in terms of Brier score

Figure 6. Performance of the classifiers learnt in different scenarios obtained from combining different amounts of fully labeled data and proportions of weakly labeled examples. In both figures, each surface represents experiments with misleading (labels in S are really improbable) or honest (labels in S are probable) labels, and shows the difference in terms of macro F1 or Brier score with respect to the results of a similar classifier learnt only with the labeled subset. Other parameters are fixed to small labelsets ($|S| = 2$), simple generative and learnt models ($K = 1$), and no forced consistency ($s = 1$).