A step towards understanding why classification helps regression

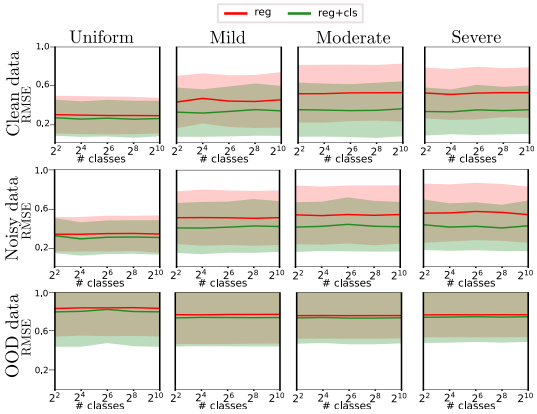

Figure 1. RMSE across 10 functions and 5 repetitions. Classification helps on imbalanced data for *Clean* and *Noisy* scenarios.

Balanced data	IMDB-WIKI-DIR (MSE)	NYUD2-DIR (RMSE)
reg	192.54 (±1.65)	1.452 (±0.031)
reg+cls (100)	192.04 (±2.43)	1.462 (±0.115)

Table 1. Balanced data: For balanced data adding a classification loss has limited effect. (Averaged over 3 repetitions).

Thank you R1, R3, R6, R7 for the constructive feedback! We will incorporate all clarity/typo comments.

Synthetic 1D OOD scenario:

[R3]: *'$P(x|y)$ is not the same as we may have domain shift'*
[R6]: *'How this "Out of distribution" case relates to the "sota" experiments?'*
We investigate three possible scenarios: *Clean*, *Noisy*, *OOD*. From these three, we observe that *classification helps on imbalanced data*, and formalize this specific setting, starting from [29] and making the same assumptions (We will explicitly add these). Our focus is *imbalanced regression* not OOD, so the derivations and "sota" experiments do not have to apply to OOD. We will clarify this.

Extended 1D analysis:

[R1]: *'I would have liked .. at least a second 1D dataset'*
[R7]: *'The synthetic 1D dataset is not sufficient'*
Great suggestion! In Fig. 1 here, we extend the 1D analysis to a family of functions of the form: $f(x)=a\sin(cx)+b\sin(dx)$, where the $f(x)\in[-1.5, 1.5]$ and $x\in[-1, 1]$. We sample 10 functions and average over 5 randoms seeds. Except for the OOD scenario, classification helps regression when the data is imbalanced, which supports our hypothesis and derivations. We will update the 1D experiments.

Balanced real data:

[R1]: *'.. also include at least one balanced dataset'*
[R7]: *'interesting .. with balanced data sampling'*
In Tab. 1 here, for *IMDB-WIKI-DIR* and *NYUD2-DIR* we keep a subset of the training data such that the target distribution is balanced. As expected, classification has limited benefits on balanced data. We will add this.

Imbalanced data	IMDB-WIKI-DIR (MSE)	NYUD2-DIR (RMSE)
reg	138.53 (±1.17)	1.515 (±0.038)
reg+cls (2)	134.91 (±1.20)	1.485 (±0.018)

Table 2. On imbalanced data: Even 2 classes can add useful information. (Averaged over 3 repetitions).

Number of classes on real data:

[R1]: *'Does using just 2 classes work'*
[R6]: *'how many classification tasks.. how it influence the performance'*
From Fig. 1 varying the number of classes (on the x-axis) has limited impact. Next to 10 and 100 classes, we evaluated also 2 classes on real data. In Tab. 2, classification helps regression even with few classes. We will add this.

Additional answers R1:

[R1]: *'Section 3.2.. If balanced classes are used, does this not correspond to reg+cls bal'*
Sorry, we will clarify. We mean the *reg+cls* variants outperform more sophisticated baselines.

Additional answers R3:

[R3]: *'re-balancing .. is known in classification literature'; '..related with the works in [A1]?'*
Yes, classification is easier for balanced/appropriately grouped classes, as in [A1]. Yet, the regression is only applied on imbalanced targets. So the relation between the classification and regression is essential. We will add [A1].
[R3]: *'NYUD2, reg-MSE 1.515 is better than reg+cls(10 cls) 1.521. This hinders the hypothesis'*
Our *reg+cls(10 cls)* is on par with the baseline, including std. This is because classification fails to converge on certain runs, so it cannot help the regression. On *NYUD2-DIR* the learning is unstable and varies with the random seeds. Yet, our other three variants of *reg+cls* outperform *reg*.

Additional answers R6:

[R6]: *'only step imbalance was considered.. whereas "sota" methods are usually closer to exponential'*
We follow prior works [27,49] when adding classification to regression. And on real data we keep the natural data distribution (see supp. material). Yet, we expect the findings to hold for other imbalanced distributions.

Additional answers R7:

[R7]: *'a clear description of the architecture used'*
We keep the architectures of [47], and add a single linear layer for predicting classes, which we remove at inference time. [47] uses ResNet-50 (He et al.,2016) for IMDB-WIKI-DIR the ResNet-50-based model (Hu et al., 2019) for NYUD2-DIR. We will add this.
[R7]: *'Comparing parameters.. questions about fairness'*
All models have *precisely the same architecture and number of parameters* during inference. The only difference is in the gradients they received during training. We will clarify.