Appendix A

Demonstration of Average Precision Calculation

To demonstrate the overly optimistic results from computing the AP of a raw sigmoidal output in comparison to a 0.5-thresholded prediction (binarized output), we provide the following example:

Let $y = [y_0, y_1]$ be the ground truth annotation for a truth array of batch size two (containing two samples) such that:

- $y_0 = [0, 0, 1, 1]$; and
- $y_1 = [0, 0, 0, 1]$.

A model then predicts for each sample sigmoid outputs $\hat{y}$ such that:

- $\hat{y_0} = [0.1, 0.7, 0.75, 0.8]$; and
- $\hat{y_1} = [0.3, 0.6, 0.2, 0.8]$.

The binarized output $\tilde{y}$ is then:

- $\tilde{y_0} = \hat{y_0} \geq 0.5 = [0, 1, 1, 1]$; and
- $\tilde{y_1} = \hat{y_1} \geq 0.5 = [0, 1, 0, 1]$.

Next, we can compare the calculated AP predicted for $\hat{y}$ and $\tilde{y}$, by applying Algorithm 1, shown below. The AP score we get for $\hat{y}$ is then 1.00, while the score for $\tilde{y}$ is 0.60.

Clearly, the prediction arrays in our example show high confidence (if we are to interpret the sigmoid outputs as confidence-based probabilities) for erroneous bits in comparison to the ground truth in y. Yet, the score for the sigmoid output reflects a perfect prediction, giving an overly optimistic impression of model performance. Only the binarized output score reasonably penalizes the model.

Another way of interpreting the difference is by observing how the rectangular approximation of the precision-recall (PR) curve changes when we calculate AP for $\hat{y}$ versus $\tilde{y}$. For $\hat{y}$, we obtain the following precisions and recalls for each threshold:

Algorithm 1 Average Precision Computation

This function is written explicitly for clarity and reproduces the results of scikit-learn's average precision calculation [24] with the averaging method set to "micro".

```
function AveragePrecision($y_{\text{true}}, y_{\text{pred}}$)
    $flatPredicted \leftarrow$ Flatten($y_{\text{pred}}$)
    $uniquePredicted \leftarrow$ Unique($flatPredicted$)
    $thresholds \leftarrow$ SortDescending($uniquePredicted$)
    $precisions \leftarrow list()$
    $recalls \leftarrow list(0)$
    for all $threshold \in thresholds$ do
        $\gamma_{\text{pred}} \leftarrow \{$if $(pred \geq threshold)$ then 1 else 0 for each $pred$ in $y_{\text{pred}}\}$
        $tp \leftarrow$ Sum($\{$1 for $(b_{\text{pred}}, b_{\text{true}})$ in $(\gamma_{\text{pred}}, y_{\text{true}})$ if $b_{\text{pred}} = 1$ and $b_{\text{true}} = 1\}$)
        $fp \leftarrow$ Sum($\{$1 for $(b_{\text{pred}}, b_{\text{true}})$ in $(\gamma_{\text{pred}}, y_{\text{true}})$ if $b_{\text{pred}} = 1$ and $b_{\text{true}} = 0\}$)
        $fn \leftarrow$ Sum($\{$1 for $(b_{\text{pred}}, b_{\text{true}})$ in $(\gamma_{\text{pred}}, y_{\text{true}})$ if $b_{\text{pred}} = 0$ and $b_{\text{true}} = 1\}$)
        $tp \leftarrow tp + tp$
        $fp \leftarrow fp + fp$
        $fn \leftarrow fn + fn$
        $precision \leftarrow tp/(tp + fp)$
        $recall \leftarrow tp/(tp + fn)$
        Append($precisions, precision$)
        Append($recalls, recall$)
    end for
    $recallRange \leftarrow list()$
    for $i \leftarrow 1$ to Length($recalls$) do
        Append($recallRange, recalls[i] - recalls[i-1]$)
    end for
    $ap \leftarrow$ Sum($\{p \cdot r$ for $(p, r)$ in $(precisions, recallRange)\}$)
    return $ap$
end function
```

- Precision $= [1.00, 1.00, 0.75, 0.60, 0.50, 0.43, 0.38]$; and
- Recall $= [0.00, 0.67, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00]$.

For $\tilde{y}$:

- Precision $= [0.6, 0.38]$; and
- Recall $= [0.00, 1.00, 1.00]$.

When we plot these, we get the curves displayed in Figure A.1. Since the second component in this case is a rectangle with a width of zero, this approach collapses to the micro-precision score. However, generally, the AP of a binarized output is not equivalent to its precision score.

Lastly, we may wonder if the AP for the binarized output is always strictly lesser than the AP score for its sigmoid variant. In fact, there are rare situations (at least with regard to learned models) where this presumption fails. In the scenario where the predictions are largely the mirror opposite of the ground truth, then the AP score of the binarized output may exceed that of the sigmoid output. However, as we are dealing with the evaluation of learned models and random models, it is almost always the case that the sigmoid output's AP scores are lower-bounded by the binarized output's AP scores. We demonstrate this conclusion in Figure A.2, using data generated from a sampling procedure detailed in Algorithm 2.

Algorithm 2 Noise Factor Experiment

This sampling process entails generating 100,000 synthetic annotations. The goal is to add increasing amounts of Gaussian noise to these annotations to mimic the predictions of a "learned model". As the amount added of noise approaches infinity, the predictions approach that of a random model. In this manner, we can simulate and compare the AP scores of sigmoid and binarized outputs from a spectrum ranging from the perfect model to a random model.

```
$nSamples \leftarrow 100000$
$totalIncrements \leftarrow 20$
$multiplier \leftarrow 7.5$
$noiseFactors \leftarrow list()$
$sigScores \leftarrow list()$
$binScores \leftarrow list()$
for $i \leftarrow 0$ to $totalIncrements$ do
    $noiseFactor \leftarrow i \cdot multiplier / totalIncrements$
    $annotations \leftarrow \textsc{RandomIntegers}(0, 1, (nSamples, 10))$
    $annotationsShape \leftarrow \textsc{Shape}(annotations)$
    $pred \leftarrow annotations + \textsc{RandomNormal}(0, noiseFactor, annotationsShape)$
    $sigPred \leftarrow \textsc{Clip}(pred, 0, 1)$
    $binPred \leftarrow sigPred > 0.5$
    $sigAP \leftarrow \textsc{AveragePrecision}(annotations, sigPred)$
    $binAP \leftarrow \textsc{AveragePrecision}(annotations, binPred)$
    $\textsc{Append}(noiseFactors, noiseFactor)$
    $\textsc{Append}(sigScores, sigAP)$
    $\textsc{Append}(binScores, binAP)$
end for
```

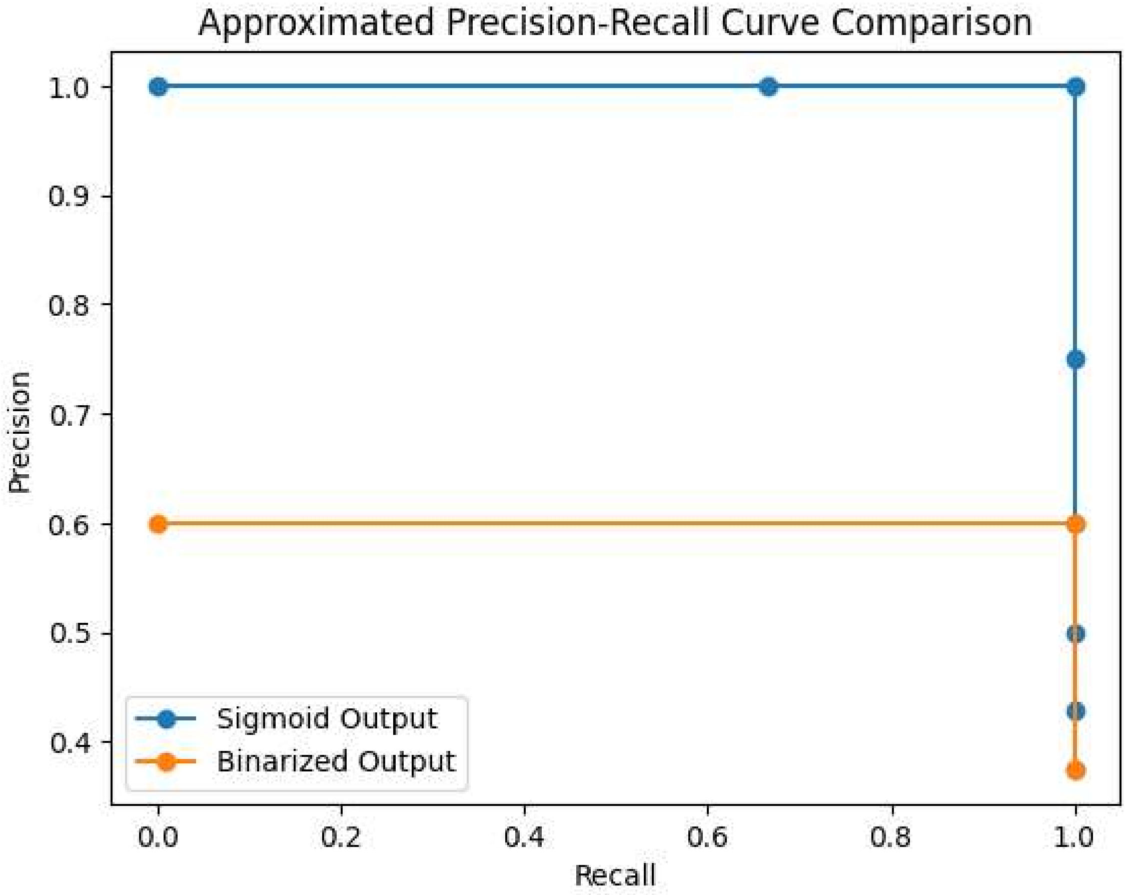

Figure A.1: **Comparison between approximated PR curves.** In our example, the precisions and recalls obtained in the intermediate steps of Algorithm 1 produce the displayed PR curves. In this particular example, both curves are comprised of only a single rectangle with non-zero width. Generally, this observation does not hold true.

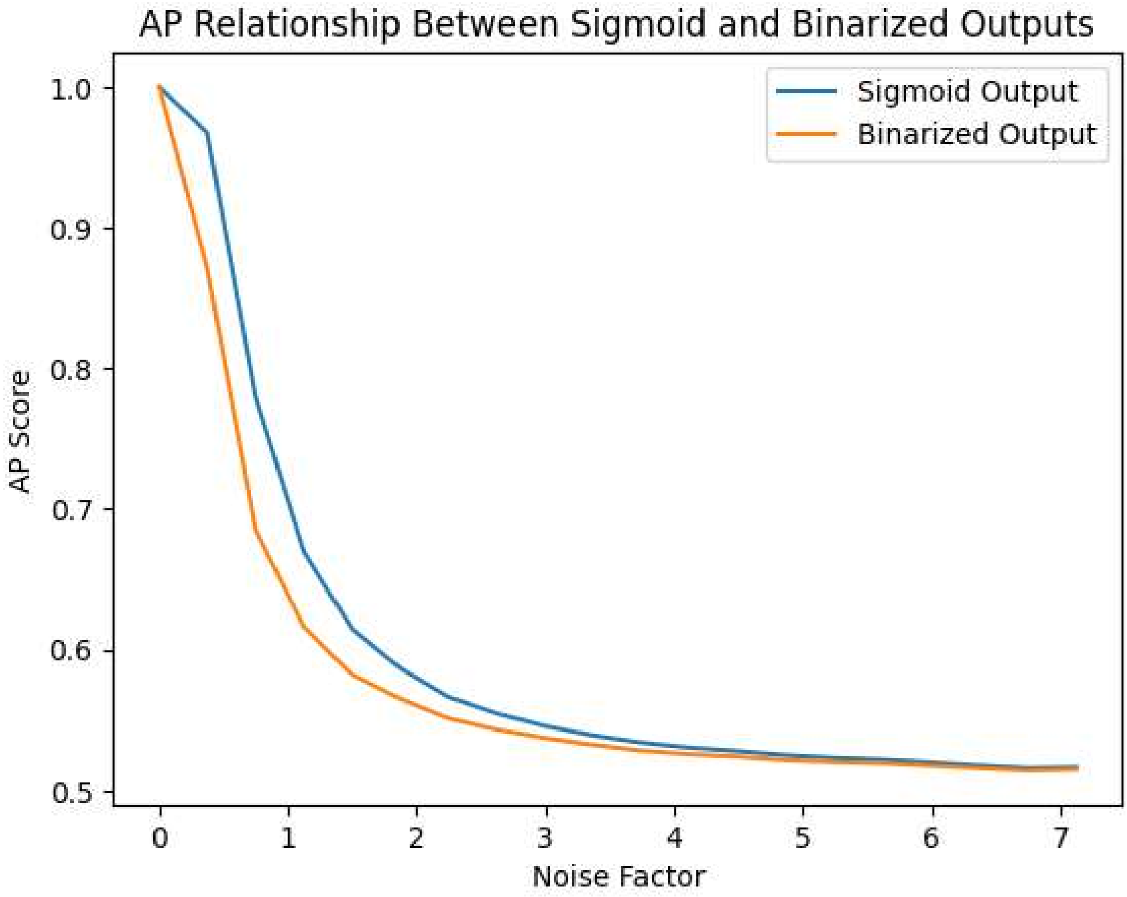

Figure A.2: **Comparison between calculated AP scores with increasing noise.** The plotted displays the output of Algorithm 2. In this experiment, as Gaussian noise is supplied to the system to mimic the descent of a perfect model into a random one, both the sigmoid and the binarized output AP scores approach 0.5, as expected. Most notably, during this entire process, the sigmoid AP scores are lower-bounded by the binarized AP scores outside of small fluctuations near convergence.