THE PREDICTION ADVANTAGE: A UNIVERSALLY MEANINGFUL PERFORMANCE

Ran El-Yaniv, Yonatan Geifman, Yair Wiener

OUTLINE

- Introduction and motivation
- The prediction advantage
- Bayesian marginal prediction
- PA for several loss functions
- Related measures
- Empirical results
- Future research and open questions
- Conclusion

INTRODUCTION

- Consider an imbalanced problem
    - Does 99% accuracy is good enough?
    - When the minority class is only 0.5%?
- Can 70% accuracy on multi-class with 3 classes can be compared to 70% with 4 classes?
- Haberman – a dataset with 26.4% of minority class with reported results of 27%
- We are looking for a universal measure that can obtain the complexity and the bias of the problem.

MAIN IDEA

- Lets obtain the performance advantage of the prediction function over the "random" function
- Challenges:
  - What is the "random classifier"
  - How can we compare 2 classifiers? Which loss? Subtract? Divide?
  - Does it general for regression and classification? For any loss function?

PREDICTION ADVANTAGE

$$\mathrm{PA}_\ell(f) = 1 - \frac{R_\ell(f)}{R_\ell(f_0)} = 1 - \frac{E_{X,Y}(\ell(f(X),Y))}{E_{X,Y}(\ell(f_0(X),Y)}.$$

BAYESIAN MARGINAL PREDICTION (BMP)

- The optimal prediction function with respect to the marginal distribution of Y.
- The BMP predicts a constant value/class while being oblivious to X and P(Y|X).
- we expect the BMP to obtain only the complexity of the problem latent in P(Y).

THE BMP IS CONSTANT

- Why the BMP is a constant?
  - Yaw principal
  - Lemma: Consider a general function g~Q and a convex loss function

$$R_\ell(g) = E_Y E_Q \ell(Y, g) \geq E_Y \ell(Y, E_Q\{g\}) = R_\ell(E_Q\{g\}).$$

PREDICTION ADVANTAGE – PROPERTIES

- Order preservation - The PA forms a weak ordering of the functions, similar to the order formed by the loss function
- Boundedness - the PA is bounded by 1. PA=1 achieved only by the perfect classifier.
- Meaningfulness - PA=0 when f has no advantage over the BMP

PA FOR CROSS ENTROPY LOSS

- Cross-entropy loss -

$$\ell(f(X), Y) = -\sum_{i \in C} \Pr\{Y = i\} \log\left(\Pr\{f(X) = i\}\right)$$

- Multi class problem with k classes $f(x) : \mathcal{X} \to R^k$
- The BMP is the marginal probabilities for each class

$$f_0(X)_i = P\{Y = e_i\}$$

- Labels are given in one-hot representation

PA FOR CROSS ENTROPY LOSS – PROOF

- Lets define an arbitrary distribution Q and $f_Q(X) \sim Q$

$$
\begin{aligned}
R_\ell(f_0) &= E\ell(f_0(X), Y) \\
&= \sum_{i \in C} \Pr\{Y = e_i\} \ell(f_0(X), e_i) \\
&= \sum_{i \in C} -\Pr\{Y = e_i\} \log\left(\Pr\{Y = e_i\}\right) \\
&= H(Y)
\end{aligned}
$$

$$
\begin{aligned}
R_\ell(f_Q) &= E\ell(f_Q(X), Y) \\
&= \sum_{i \in C} \Pr\{Y = e_i\} \ell(f_Q(X), e_i) \\
&= \sum_{i \in C} -\Pr\{Y = e_i\} \log\left(f_{Qi}(X)\right)
\end{aligned}
$$

PA FOR CROSS ENTROPY LOSS – PROOF

- We calculate: $R_\ell(f_Q) - R_\ell(f_0)$

$$
\begin{aligned}
R_\ell(f_Q) - R_\ell(f_0) &= \sum_{i \in C} -\Pr\{Y = e_i\} \log\left(f_{Qi}\right) + \sum_{i \in C} \Pr\{Y = e_i\} \log\left(\Pr\{Y = e_i\}\right) \\
&= \sum_{i \in C} \Pr\{Y = e_i\} \log\left(\Pr\{Y = e_i\} / f_{Qi}(X)\right) \\
&= D_{kl}(f_0(X) || f_Q(X)) \\
&\geq 0.
\end{aligned}
$$

- The BMP loss: $R_\ell(f_0) = H(P(Y))$
- The PA: $\mathrm{PA}_\ell(f) = 1 - \dfrac{R_\ell(f)}{H(P(Y))}.$

PA FOR 0/1 LOSS

- The BMP: $f_0 = \text{argmax}_i(\Pr\{Y = i\})$
- The BMP risk:

$$R_{\ell_{0-1}}(f_0) = 1 - \max_{i \in C}(\Pr\{Y = i\}) = 1 - \Pr\{Y = j\}.$$

- The PA:

$$\text{PA}_\ell(f) = 1 - \frac{R_\ell(f)}{R_\ell(f_0)} = 1 - \frac{R_\ell(f)}{1 - \max_{i \in C}(\Pr\{Y = i\})}$$

PA FOR SQUARED LOSS

- The BMP $f_0 = E[Y]$
- The BMP risk:

$$R_\ell(f_0) = E_Y[(Y - f_0)^2] = E_Y[(Y - E[Y])^2] = var(Y)$$

- The PA:

$$\mathrm{PA}_\ell(f) = 1 - \frac{R_\ell(f)}{R_\ell(f_0)} = 1 - \frac{R_\ell(f)}{var(Y)}$$

PA FOR ABSOLUTE LOSS

- The BMP for absolute loss: $f_0 = median(Y)$
- The BMP risk:

$$R_\ell(f_0) = E_Y[|Y - median(Y)|] = D_{med}$$

- The PA:

$$\mathrm{PA}_\ell(f) = 1 - \frac{R_\ell(f)}{R_\ell(f_0)} = 1 - \frac{R_\ell(f)}{D_{med}}.$$

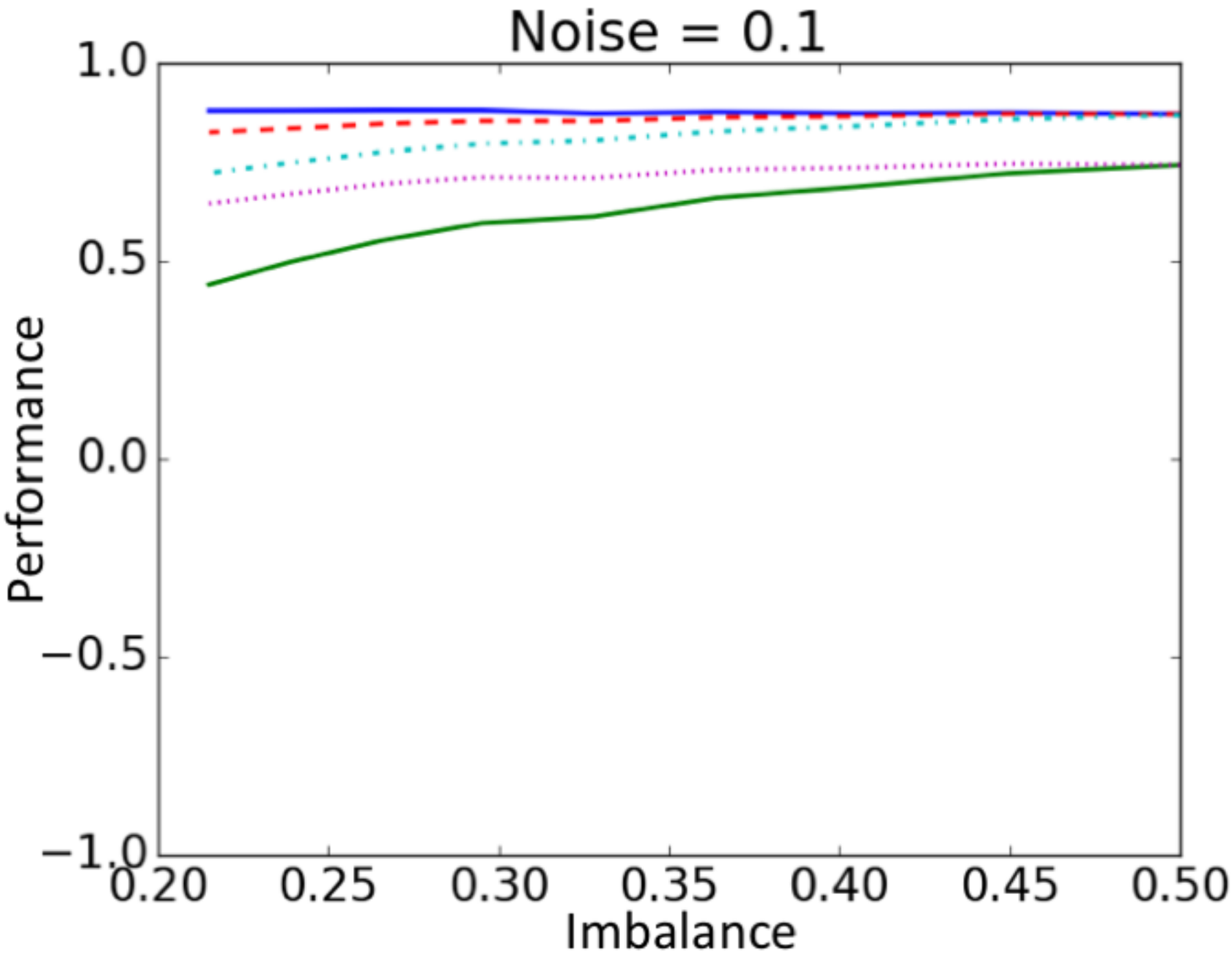
RELATION TO OTHER MEASURES

- Some other measures defined as two numbers (e.g., precision recall), we look for one number
- We compared to F-score, Cohen's kappa, and balanced accuracy
- The PA bounds from below all the other measures

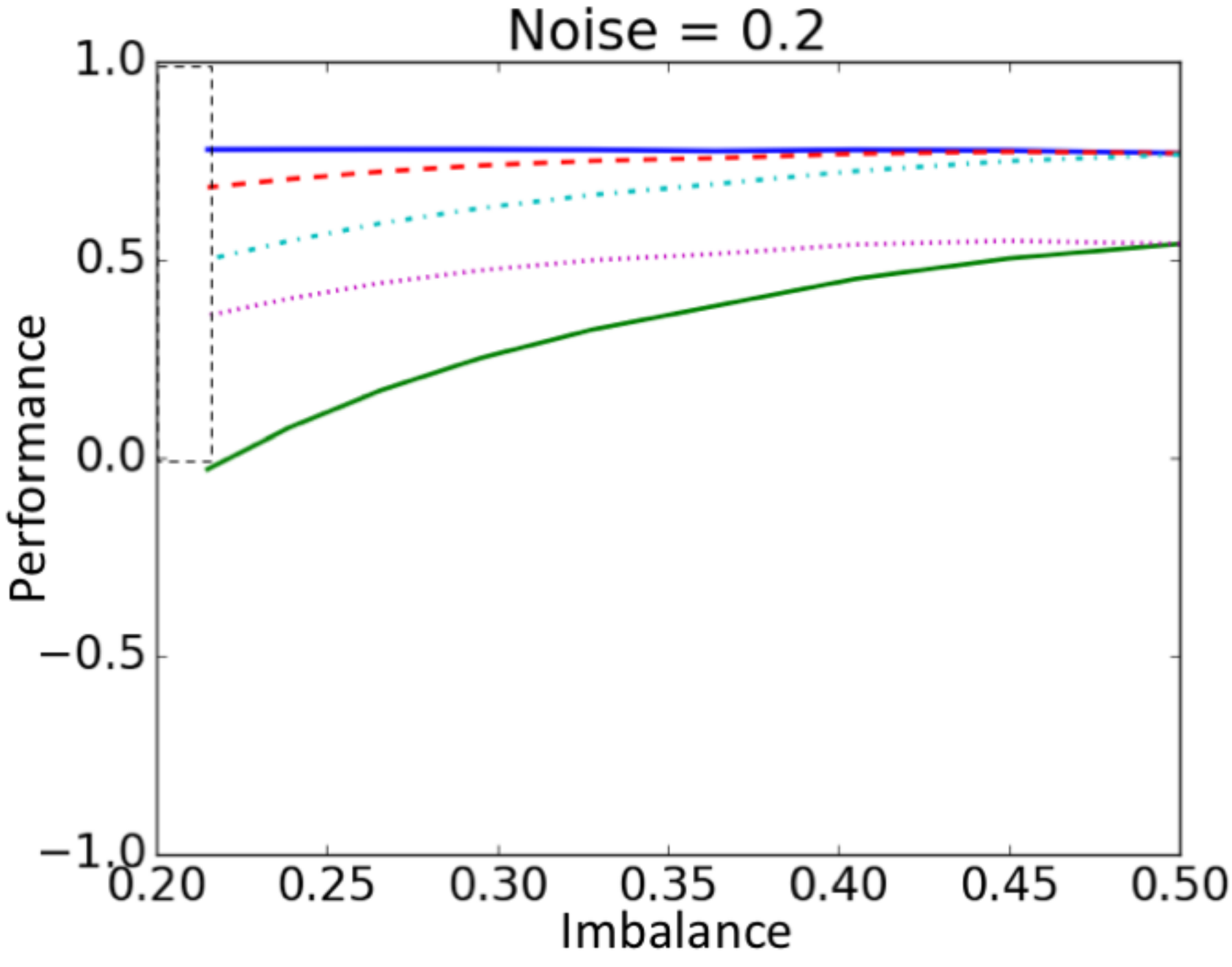
EMPIRICAL RESULTS

- We compared some relevant performance measure on different noise levels and imbalance levels on the breast cancer dataset
- Measures:
  - Balanced accuracy - (TP+TN)/2
  - F-measure - harmonic mean of precision and recall
  - Cohen’s kappa - inter-rater agreement measure

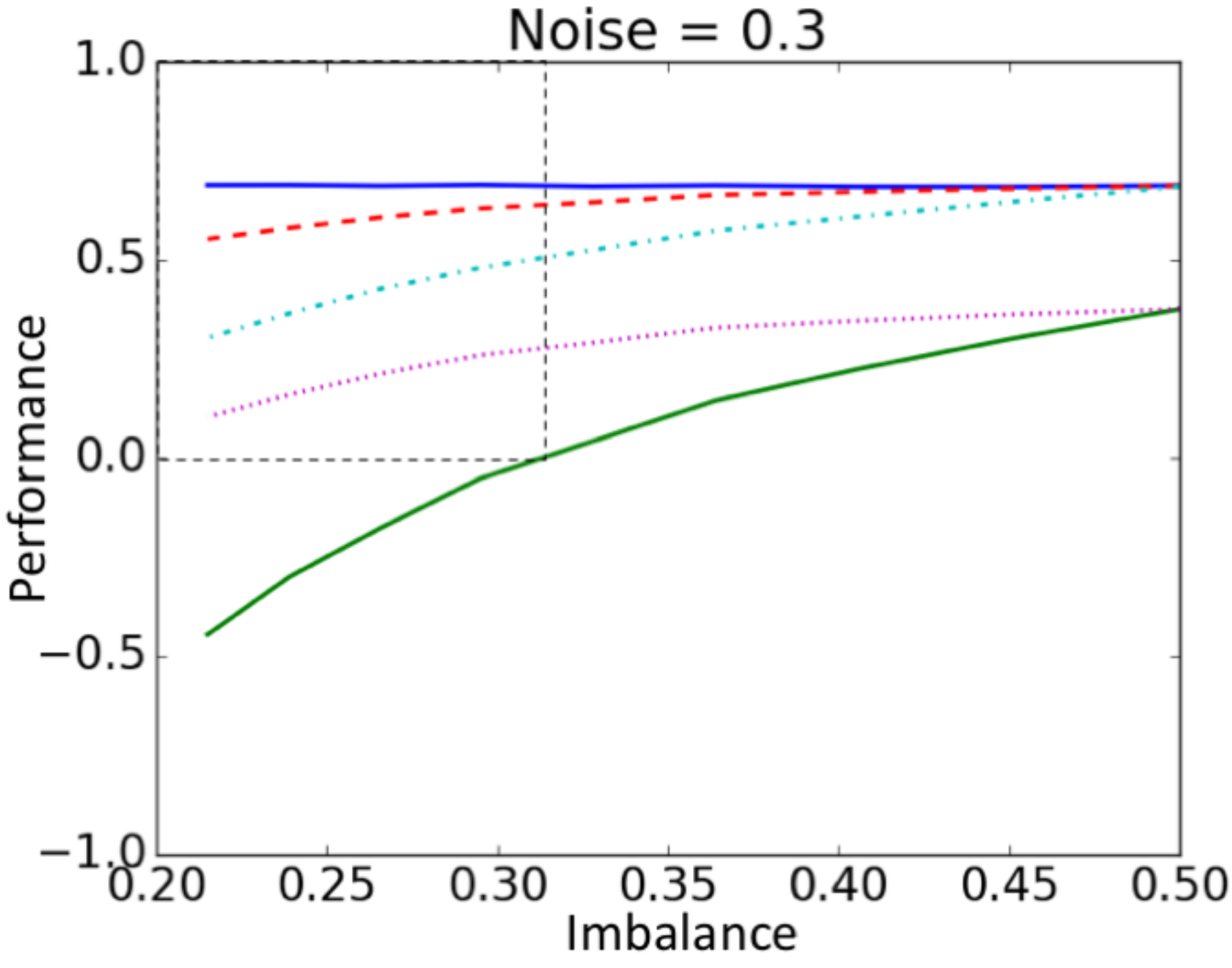
EMPIRICAL RESULTS



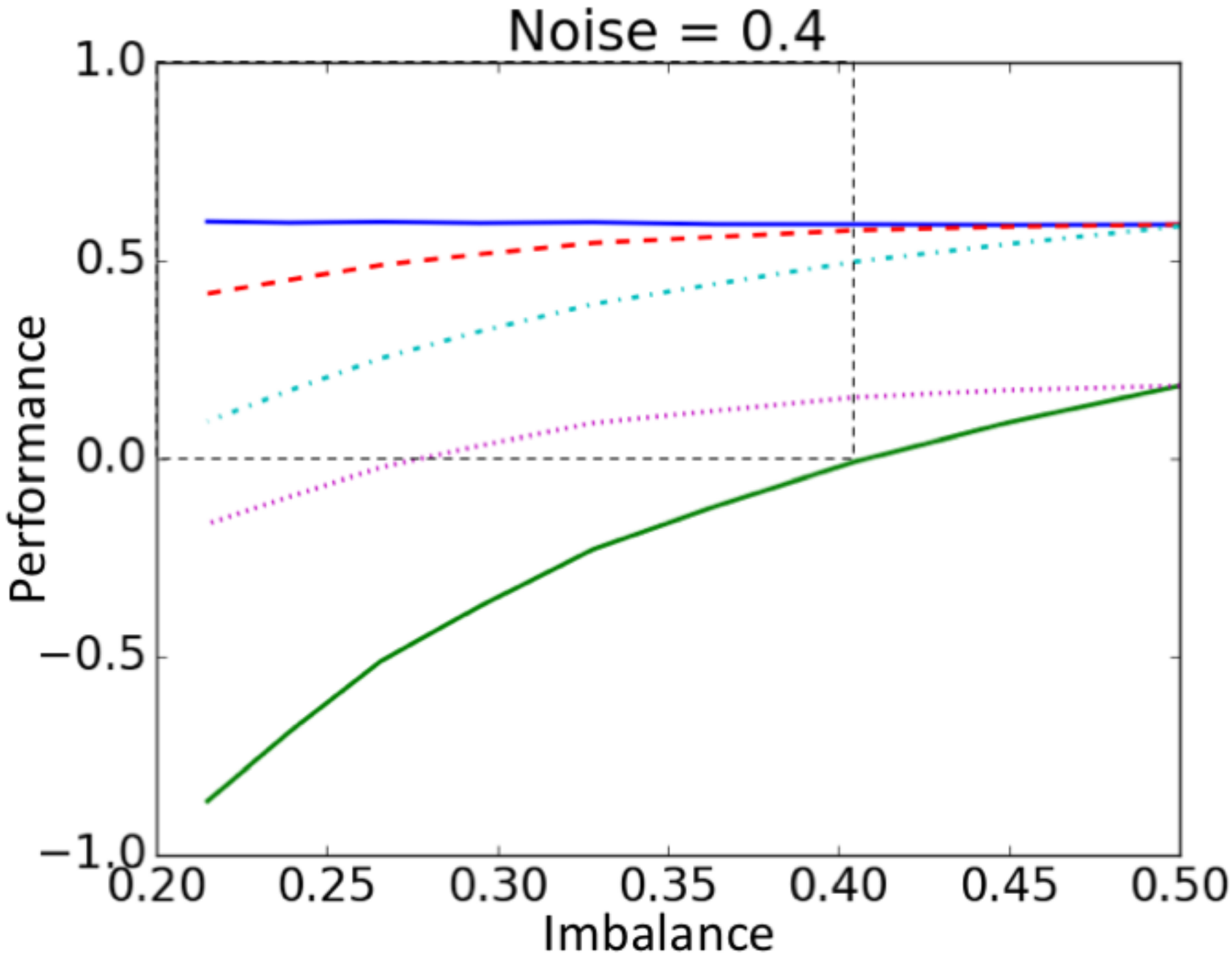
EMPIRICAL RESULTS



EMPIRICAL RESULTS



EMPIRICAL RESULTS



PA AND SELECTIVE PREDICTION

- In selective prediction for every coverage rate we have different P(Y)
- Risk-coverage curves are misleading
- We argue that in this case the objective has to be the PA and we should measure the PA-coverage curve
- Still not clear how to construct a reject mechanism which optimize PA

CONCLUSION AND FUTURE WORK

- We presented a universal performance measure
- It is still not clear how to best estimate some of the measures (entropy, median, etc...)
- Does the PA can be used as an optimization objective? where is it needed? how to optimize it? (non convex)