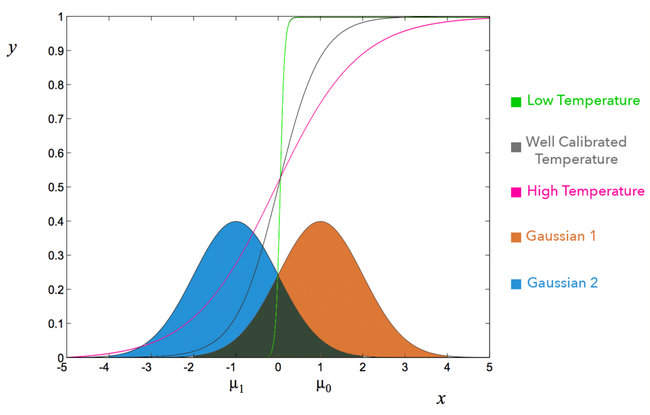
Probabilities from neural networks?

Dogma: Softmax sums to 1 + training on cross entropy loss => probabilities.



Binning?

Logistic?

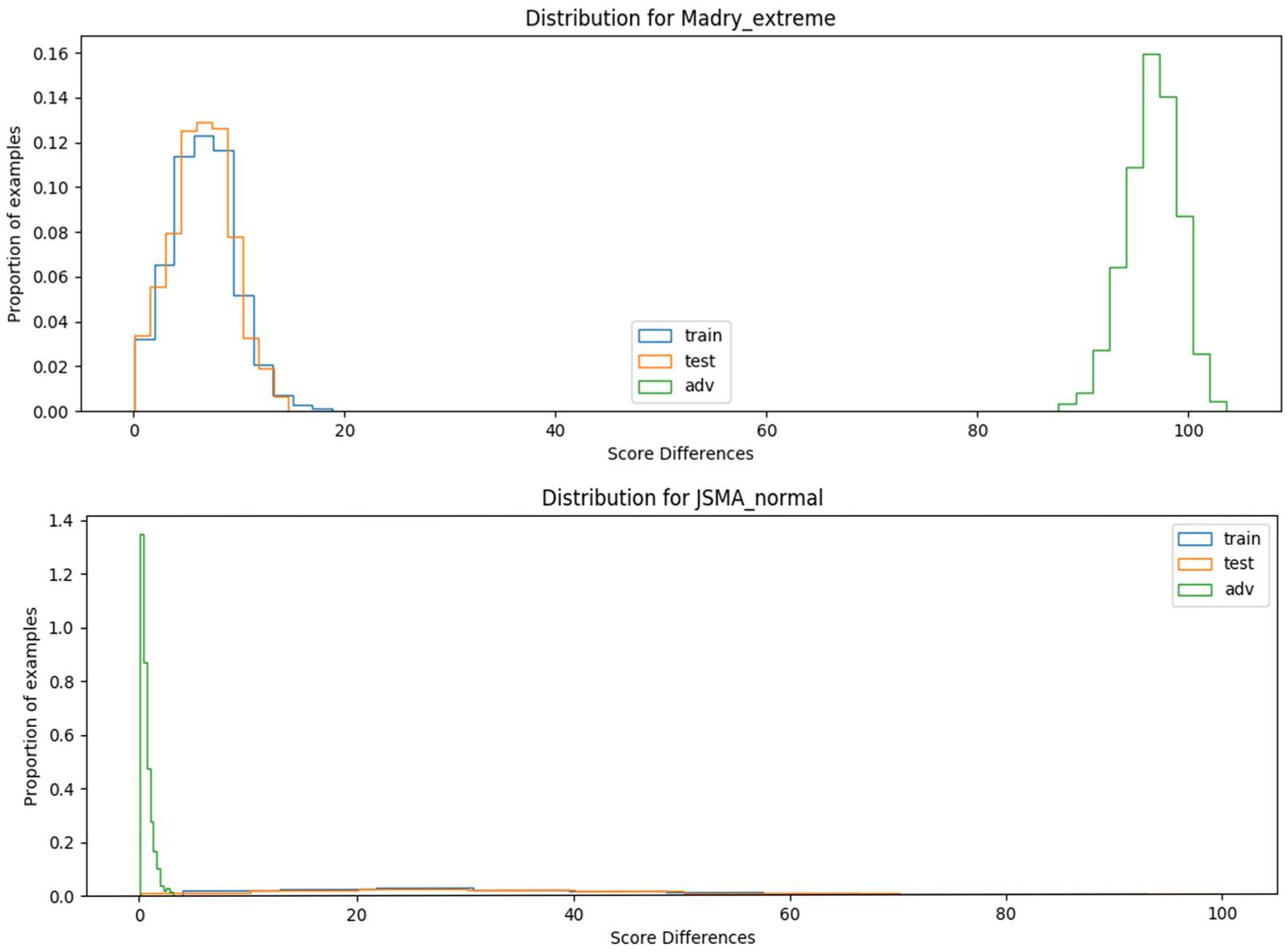
Beta?

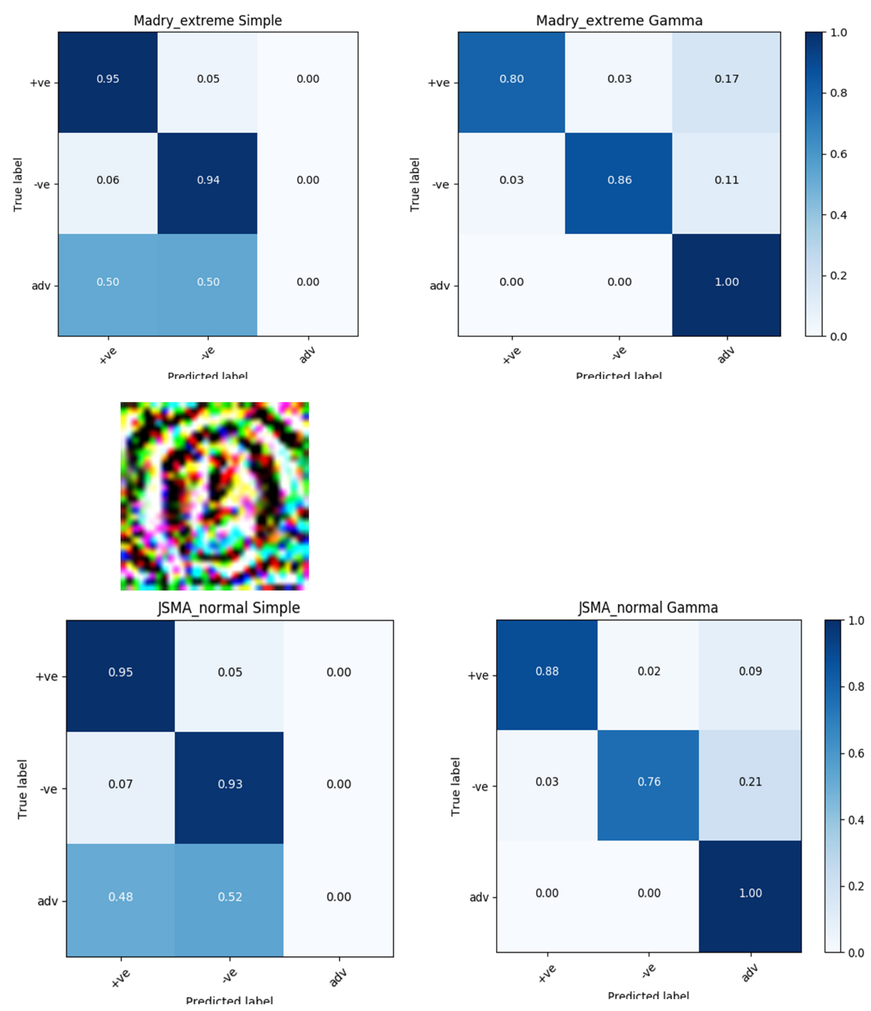
Platt?

But Lippman only gives us the guarantee of good discriminative posterior on training data.

Adversarial Examples

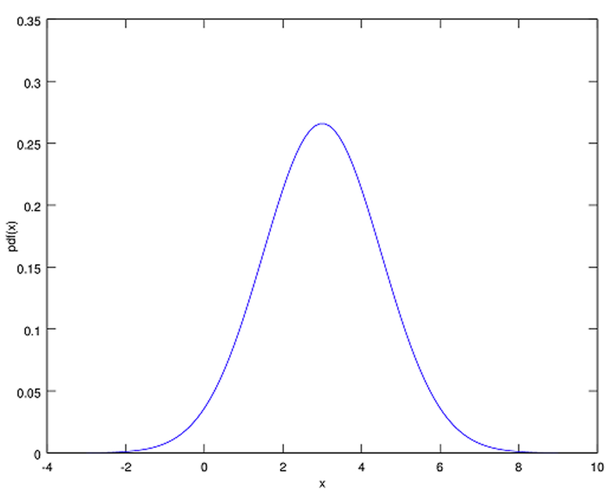
Problem 1:
Softmax logit feature space not quantified wrt. shape and meaning. Multiple Realisability?

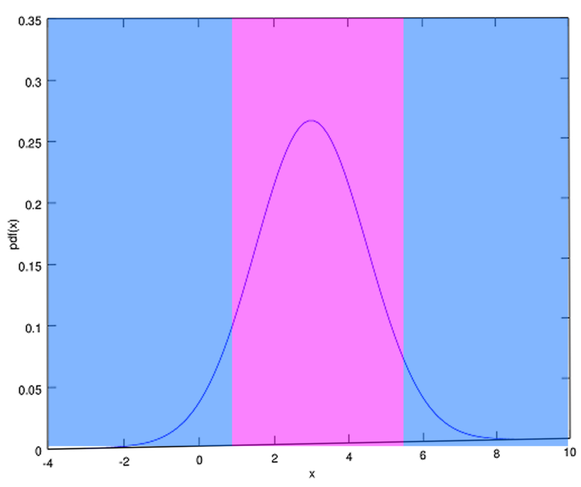




State of the art (almost) adversarial defense: Ad Hoc
State of the art neural network calibration: Ad Hoc
Conclusion : Softmax has little meaning. Logits also have little meaning.

First obvious solution: Quantify uncertainty : Background Check (Shown above)
Feature space has 16.5m ^ 1024 examples for 32x32. Prior?





Future work:
Leveraging uncertainty in discriminative / generative models to combat adversarials.