Supplementary Material : Uncertainty Quantification in Deep Binary Classification via Quantiles

No Author Given

No Institute Given

1 Proofs

This section covers the proofs of the Lemmas in Section 2 of the main manuscript
Lemma 2.1[1] The Lipschitz constant of the BQR loss is $\max(\tau, 1-\tau)$

Proof. Recall that, the empirical risk under the BQR loss is:

$$L(y,z) = -(y \log p_z + (1-y)\log(1-p_z))$$

where

$$p_z \equiv \begin{cases} 1-\tau\exp((\tau-1)z) & z \geq 0 \\ (1-\tau)\exp(\tau z)) & z < 0 \end{cases}$$

Let us consider the following cases.
Case-1a: $0 < z_1 < z_2, y = 1$

$$\frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} = \frac{\log(1-\tau e^{(\tau-1)z_2}) - \log(1-\tau e^{(\tau-1)z_1})}{z_2 - z_1}$$

The RHS approaches maximum as $z_2, z_1 \to 0$. Taking the limit w.r.t z_1 first, we get,

$$\lim_{z1\to 0} \frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} = \frac{\log(1-\tau e^{(\tau-1)z_1}) - \log(1-\tau)}{z_2}$$

and then taking the limit w.r.t z_2 later, we get

$$\lim_{z_2,z_1\to 0} \frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} = \tau$$

Therefore,

$$\frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} \leq \tau$$

[1] Visualizations can be seen here (clickable hyperlinks) :- Y=0 and Y=1

Case-1b: $0 < z_1 < z_2, y = 0$

$$\frac{|L(0,z_2) - L(0,z_1)|}{|z_2 - z_1|} = \frac{\log\left(\tau e^{(\tau-1)z_2}\right) - \log\left(\tau e^{(\tau-1)z_1}\right)}{z_2 - z_1}$$

In this case, the RHS simplifies to,

$$\begin{aligned}\frac{|L(0,z_2) - L(0,z_1)|}{|z_2 - z_1|} &= \frac{(z_2 - z_1)(1-\tau)}{z_2 - z_1} \\ &= 1 - \tau\end{aligned}$$

Therefore,

$$\frac{|L(0,z_2) - L(0,z_1)|}{|z_2 - z_1|} \leq 1 - \tau$$

Case-2a: $z_1 < 0 < z_2, y = 1$

$$\frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} = \frac{\log\left(1 - \tau e^{(\tau-1)z_2}\right) - \log\left((1-\tau)e^{\tau z_1}\right)}{z_2 - z_1}$$

The RHS approaches maximum as $z_2, z_1 \to 0$. Taking the limit w.r.t z_1 first, we get,

$$\lim_{z1 \to 0} \frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} = \frac{\log\left(1 - \tau e^{(\tau-1)z_2}\right) - \log\left(1-\tau\right)}{z_2}$$

and then taking the limit w.r.t z_2, we get

$$\lim_{z_2, z_1 \to 0} \frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} = \tau$$

Therefore,

$$\frac{|L(1,z_2) - L(1,z_1)|}{|z_2 - z_1|} \leq \tau$$

Case-2b: $z_1 < 0 < z_2, y = 0$

$$\frac{|L(0,z_2) - L(0,z_1)|}{|z_2 - z_1|} = \frac{\log\left(\tau e^{(\tau-1)z_2}\right) - \log\left(1 - (1-\tau)e^{\tau z_1}\right)}{z_2 - z_1}$$

The RHS approaches maximum as $z_2, z_1 \to 0$. Taking the limit w.r.t z_1 first, we get,

$$\lim_{z1 \to 0} \frac{|L(0,z_2) - L(0,z_1)|}{|z_2 - z_1|} = \frac{\log\left(1 - \tau e^{(\tau-1)z_2}\right) - \log\left(\tau\right)}{z_2}$$

and then taking the limit w.r.t z_2, we get

$$\lim_{z_2, z_1 \to 0} \frac{|L(0,z_2) - L(0,z_1)|}{|z_2 - z_1|} = 1 - \tau$$

Therefore,

$$\frac{|L(0,z_2)-L(0,z_1)|}{|z_2-z_1|} \leq 1-\tau$$

Case-3a: $z_1 < z_2 < 0, y = 1$

$$\frac{|L(1,z_2)-L(1,z_1)|}{|z_2-z_1|} = \frac{\log\left((1-\tau)e^{\tau z_2}\right) - \log\left((1-\tau)e^{\tau z_1}\right)}{z_2-z_1}$$

The RHS simplifies to,

$$\frac{|L(1,z_2)-L(1,z_1)|}{|z_2-z_1|} = \frac{\tau(z_2-z_1)}{z_2-z_1}$$

Therefore,

$$\frac{|L(0,z_2)-L(0,z_1)|}{|z_2-z_1|} \leq \tau$$

Case-3b: $z_1 < z_2 < 0, y = 0$

$$\frac{|L(0,z_2)-L(0,z_1)|}{|z_2-z_1|} = \frac{\log\left(1-(1-\tau)e^{\tau z_2}\right) - \log\left(1-(1-\tau)e^{\tau z_1}\right)}{z_1-z_2}$$

The RHS approaches maximum as $z_1, z_2 \to 0$. Taking the limit w.r.t z_2 first, we get,

$$\lim_{z_1\to 0} \frac{|L(0,z_2)-L(0,z_1)|}{|z_2-z_1|} = \frac{\log(1-\tau) - \log\left(1-\tau e^{(\tau-1)z_1}\right)}{z_1}$$

and then taking the limit w.r.t z_1, we get

$$\lim_{z_1,z_2\to 0} \frac{|L(0,z_2)-L(0,z_1)|}{|z_2-z_1|} = 1-\tau$$

Therefore,

$$\frac{|L(0,z_2)-L(0,z_1)|}{|z_2-z_1|} \leq 1-\tau$$

Hence, $\forall z_1, z_2 \in R, y \in \{0,1\}$

$$\frac{|L(y,z_2)-L(y,z_1)|}{|z_2-z_1|} \leq \max(1-\tau, \tau)$$

□

Lemma 2.2 BQR also admits a bound in terms of the curvature of the function f^*. That is

$$c_1 E((f-f^*)^2) \leq E(L(y,f)-L(y,f^*)) \leq c_2 E((f-f^*)^2)$$

where c_1 and c_2 constants, bounded away from 0.

Proof. Recall that, the empirical risk under the BQR loss is:

$$L(y,z) = -(y \log p_z + (1-y)\log(1-p_z))$$

where

$$p_z \equiv \begin{cases} 1-\tau\exp((\tau-1)z) & z>0 \\ (1-\tau)\exp(\tau z)) & z \le 0 \end{cases}$$

Using Taylor series expansion of $h_a(b) = L(b,y) - L(a,y)$, with $a = f, b = f^*$, we can write,

$$h_a(b) = h_a(a) + h'_a(a)(b-a) + \frac{1}{2}h''_a(a)(b-a)^2$$

We will be looking at h''_a to determine the bounds for the curvature of the loss function. Let us consider the following cases.
Case-1: $b \ge 0, a \ge 0$

$$\begin{aligned} h_a(b) &= -(1-te^{-(1-t)a})\log(1-te^{-(1-t)b}) - (te^{-(1-t)a})\log(te^{-(1-t)b}) + g(a) \\ &= -(1-te^{-(1-t)a})\log(1-te^{-(1-t)b}) - (te^{-(1-t)a})(\log(t) - (1-t)b)) + g(a) \end{aligned}$$

$$h'_a(b) = -(1-te^{-(1-t)a})\frac{t(1-t)e^{-(1-t)b}}{(1-te^{-(1-t)b})} - (te^{-(1-t)a})(1-t)$$

$$h''_a(b) = (1-te^{-(1-t)a})\frac{t(1-t)^2e^{-(1-t)b}}{(1-te^{-(1-t)b})^2}$$

$h''_a(b)$ is maximum at $a=0, b=0$, and minimum at $a=M, b=M$, therefore

$$\begin{aligned} A_1 \equiv h''_a(b) &\ge \frac{t(1-t)^2e^{-(1-t)M}}{1-te^{-(1-t)M}} \\ h''_a(b) &\le t(1-t) \end{aligned}$$

Case-2: $b \le 0, a \le 0$

$$\begin{aligned} h_a(b) &= -(1-t)e^{ta}\log((1-t)e^{tb}) - (1-(1-t)e^{ta})\log(1-(1-t)e^{tb}) + g(a) \\ &= -(1-t)e^{ta}(\log((1-t)+tb) - (1-(1-t)e^{ta})\log(1-(1-t)e^{tb}) + g(a) \end{aligned}$$

$$h'_a(b) = -(1-t)te^{ta} + t(1-t)(1-(1-t)e^{ta})\frac{e^{tb}}{1-(1-t)e^{tb}}$$

$$h''_a(b) = t^2(1-t)(1-(1-t)e^{ta})\frac{e^{tb}}{(1-(1-t)e^{tb})^2}$$

$h''_a(b)$ is maximum at $a = 0, b = 0$, and minimum at $a = -M, b = -M$, therefore

$$A_2 \equiv h''_a(b) \geq \frac{t^2(1-t)e^{-tM}}{1-(1-t)e^{-tM}}$$
$$h''_a(b) \leq t(1-t)$$

Case-3: $b \geq 0, a \leq 0$

$$\begin{aligned} h_a(b) &= -(1-t)e^{ta}\log(1-te^{-(1-t)b}) - (1-(1-t)e^{ta})\log(te^{-(1-t)b}) + g(a) \\ &= -(1-t)e^{ta} - \log(1-te^{-(1-t)b}) - (1-(1-t)e^{ta})(\log(t) + -(1-t)b) + g(a) \end{aligned}$$

$$h'_a(b) = -t(1-t)^2e^{ta}\frac{e^{-(1-t)b}}{1-te^{-(1-t)b}} + (1-(1-t)e^{ta}(1-t)$$

$$h''_a(b) = t(\text{`} - t)^3e^{ta}\frac{e^{-(1-t)b}}{(1-te^{-(1-t)b})^2}$$

$h''_a(b)$ is maximum at $a = 0, b = 0$, and minimum at $a = -M, b = M$, therefore

$$A_3 \equiv h''_a(b) \geq \frac{t(1-t)^3e^{-M}}{(1-te^{-(1-t)M})^2}$$
$$h''_a(b) \leq t(1-t)$$

Case-4: $b \leq 0, a \geq 0$

$$\begin{aligned} h_a(b) &= -(1-te^{-(1-t)a})\log((1-t)e^{tb}) - te^{-(1-t)a}\log(1-(1-t)e^{tb}) + g(a) \\ &= -(1-te^{-(1-t)a})\log((1-t)e^{tb}) - (1-te^{-(1-t)a})(\log(1-t) + tb) \end{aligned}$$

$$h'_a(b) = -(1-te^{-(1-t)a})t - t^2(1-t)e^{-(1-t)a}\frac{e^{tb}}{1-(1-t)e^tb}$$

$$h''_a(b) = t^3(1-t)e^{-(1-t)a}\frac{e^{tb}}{(1-(1-t)e^{tb})^2}$$

$h''_a(b)$ is maximum at $a = 0, b = 0$, and minimum at $a = M, b = -M$, therefore

$$A_4 \equiv h''_a(b) \geq \frac{t^3(1-t)e^{-M}}{(1-(1-t)e^{-M})^2}$$
$$h''_a(b) \leq t(1-t)$$

Therefore,

$$c_1 = 0.5\min(A_1, A_2, A_3, A_4)$$
$$c_2 = 0.5t(1-t)$$

□

Theorem 2.3 Suppose Assumptions 2.1-2.3 hold. Let f be the deep ReLU network with W number of parameters. Under BQR, with probability at least $1 - e^{-\gamma}$, for large enough n, for some $C > 0$,

$$\|f - f^*\|^2_{L_2(x)} = E((f - f^*)^2) \leq B$$

for $B = C\left(\frac{W \log(W)}{n} \log n + \frac{\log \log n + r}{n} + \epsilon^2_{f^*}\right)$

Proof. See Theorem 2 of [1] □

2 R^2 values in Table III

Using our expected estimate of the misclassification rate given a confidence score, in Table 1 (and Table III of the main manuscript), we compute the R^2 score between the obtained values and the expected values. The values for the Yacht dataset may seem incorrect on first glance, however they are not. This is due to two factors. The first is that this R^2 score is computed using Scikit Learn's `r2_score` method[2]. As such, the value of R^2 ranges from $-\inf$ to 1.0, as as constant model that disregards inputs is given a score of 0.0. This occurs in the Yacht dataset, as it extremely easy to classify, and as such the number of misclassifications are zero for all δ scores above 0.2

References

1. Farrell, M.H., Liang, T., Misra, S.: Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. ArXiv arXiv:1809.09953 (2018)

[2] https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

			δ-Score					
Dataset	t	Rate	0.1	0.2	0.3	0.4	0.5	R^2
Abalone	9	Mis.	0.40	0.35	0.25	0.15	0.04	0.89
		Ret.	1.00	0.84	0.68	0.52	0.33	
	7	Mis.	0.45	0.33	0.23	0.13	0.02	0.95
		Ret.	1.00	0.94	0.89	0.80	0.63	
Boston	22	Mis.	0.44	0.33	0.23	0.13	0.02	0.96
		Ret.	1.00	0.96	0.91	0.86	0.74	
	18	Mis.	0.30	0.27	0.18	0.10	0.02	0.78
		Ret.	1.00	0.97	0.92	0.87	0.78	
California	1.8	Mis.	0.41	0.35	0.24	0.13	0.03	0.94
		Ret.	1.00	0.92	0.84	0.74	0.60	
	2.0	Mis.	0.43	0.37	0.26	0.15	0.03	0.88
		Ret.	1.00	0.92	0.84	0.76	0.61	
Concrete	35	Mis.	0.42	0.31	0.23	0.11	0.04	0.97
		Ret.	1.00	0.94	0.87	0.79	0.66	
	50	Mis.	0.46	0.34	0.20	0.15	0.01	0.94
		Ret.	1.00	0.97	0.92	0.88	0.81	
Energy	20	Mis.	0.38	0.21	0.23	0.08	0.00	0.89
		Ret.	1.00	0.99	0.99	0.99	0.97	
	15	Mis.	0.42	0.41	0.39	0.21	0.00	0.54
		Ret.	1.00	0.96	0.93	0.89	0.84	
Protein	5	Mis.	0.37	0.36	0.24	0.13	0.04	0.90
		Ret.	1.00	0.88	0.76	0.61	0.40	
	9	Mis.	0.41	0.37	0.26	0.15	0.04	0.87
		Ret.	1.00	0.88	0.75	0.61	0.40	
Redshift	0.65	Mis.	0.40	0.32	0.20	0.12	0.02	0.99
		Ret.	1.00	0.96	0.92	0.87	0.78	
	0.9	Mis.	0.39	0.28	0.19	0.19	0.01	0.89
		Ret.	1.00	0.97	0.93	0.87	0.81	
Wine	5	Mis.	0.43	0.39	0.29	0.17	0.07	0.70
		Ret.	1.00	0.87	0.74	0.57	0.33	
	6	Mis.	0.47	0.36	0.26	0.18	0.03	0.83
		Ret.	1.00	0.95	0.90	0.83	0.74	
Yacht	2	Mis.	0.13	0.10	0.03	0.00	0.00	-9.6
		Ret.	1.00	1.00	0.99	0.97	0.89	
	7.5	Mis.	0.24	0.10	0.00	0.00	0.00	-1.6
		Ret.	1.00	0.99	0.98	0.94	0.86	

Table 1: Misclassiciation and Retention rates per δ-score for Thresholded UCI datasets