Supplementary Material for the Paper: Hinge-Wasserstein: Estimating Multimodal Aleatoric Uncertainty in Regression Tasks

Ziliang Xiong[1], Arvi Jonnarth[1,3], Abdelrahman Eldesokey[2],
Joakim Johnander[4], Bastian Wandt[1], Per-Erik Forssén[1]
[1]Computer Vision Laboratory, Department of Electrical Engineering, LiU, Sweden
[2]Visual Computing Center, KAUST, Saudi Arabia
[3]Husqvarna Group, Huskvarna, Sweden, [4]Zenseact, Sweden
{name.surname}@{liu.se, kaust.edu.sa, zenseact.com}

1. Introduction

We provide in this material the contents promised in the main paper and additional results: Sec. 2 comparison between different uncertainty measures, ablation study on Gaussian smoothing and different hinge values, and additional qualitative results on HLW task (cf Sec. 3.4, Sec. 4.1, Sec. 4.2 and Sec. 4.3 in the main paper).

2. Ablation study and additional results

2.1. Comparison between different uncertainty measures

The variance, u_σ, is a common uncertainty measure for regression problems [2]:

$$u_\sigma(\hat{p}_y) = \sum_k k^2 \hat{p}_y[k] - \left(\sum_k k \hat{p}_y[k] \right)^2 \qquad (1)$$

The main disadvantage of (1) is that it can easily be dominated by secondary modes that are far from the dominant mode. Another possible choice is the inverse of the maximum bin value, defined as in

$$u_M(\hat{p}_y) = \frac{1}{\hat{p}_y[k^*]}, k^* = \arg\max_k(\hat{p}_y[k]) \ . \qquad (2)$$

As the number of modes increases, the maximum mode will also drop, indicating larger uncertainty. In Sec. 2.4 we show the effect on AUSE of these three uncertainty measures.

We show AUSE of varying hinge values for the stereo disparity task in Tab. 1 with three different uncertainty measures. All entries use the softplus activation and L_1 normalization as the final layer. Among the three measures, variance u_σ in Eq. (1) achieves the smallest AUSE, which is desired for sorting predictions by uncertainty. We still report entropy-based AUSE in the main paper to be consistent with other tasks. u_M in Eq. (2) has a slightly larger AUSE but shares the same trend, that the AUSE initially optimizes towards an optimal value, and then gets worse as the hinge increases. All three different uncertainty measures achieve the optimal AUSE at hinge, $\gamma_W = 0.0075$. This shows that the improvement on AUSE from our proposed *hinge*-W_1 is robust to various uncertainty measures.

Table 1. **Stereo disparity.** Comparison of the effect on AUSE for three different uncertainty measures, entropy, the inverse of the max bin value, and distribution variance described in Sec. 2.1.

Settings	entropy	MAX	variance
hinge-W, $\gamma_W = 0$	17.8	20.03	16.13
hinge-W, $\gamma_W = 0.0025$	17.1	18.66	16.00
hinge-W, $\gamma_W = 0.005$	15.9	17.40	15.07
hinge-W, $\gamma_W = 0.0075$	**15.9**	**17.37**	**14.87**
hinge-W, $\gamma_W = 0.01$	17.1	18.80	15.13
hinge-W, $\gamma_W = 0.0125$	17.5	19.46	14.90
hinge-W, $\gamma_W = 0.015$	17.3	19.43	15.33

Furthermore, the validity of the entropy as the scalar uncertainty measure is assessed using *kernel density estimation* (KDE) plots on the two test sets. This is done for the entropy of the slope and offset distributions, *i.e.*, $u_H(\hat{p}_\alpha)$, $u_H(\hat{p}_\rho)$. Ideally, the mode of the uncertainty measure distribution on the one-line test set should be lower and well separated from the one on the two-line test set.

Fig. 1 shows KDE plots for the one- and two-line test sets. Using the NLL loss (green) leads to a small magnitude of uncertainty for the two-line test set overlapping the peak of one-line test set. Using the plain Wasserstein loss (blue) the network cannot distinguish ambiguous images with higher aleatoric uncertainty from others. The hinge-W_1 loss (orange) improves the separation of the modes for the two distributions. Thus, we conclude that hinge-W_1

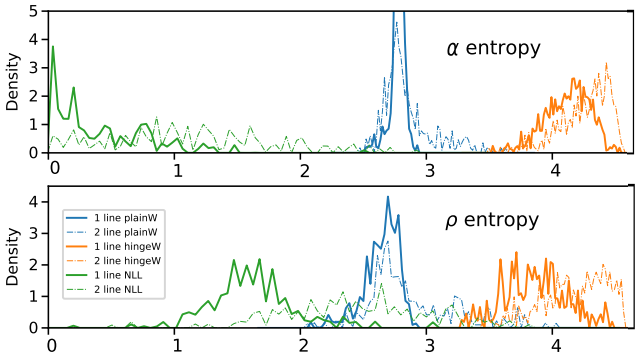

Figure 1. Entropy distribution of predictions for different losses. The models are trained with unimodal annotations. Top: KDE plots for α entropy; Bottom: KDE plots for ρ entropy; 1 line, 2 line in the legend denotes two test sets in Sec. 4.1 in the main paper. On the 2-line test set, the entropy distribution is expected to be higher than on the 1-line test set.

generates better aleatoric uncertainty estimates.

Table 2. Test results on *Horizon Lines in the Wild*. For *hinge*-W_1 and *plain Wasserstein*, Gaussian smoothed ($\sigma = 4$) training target apply. Both metrics are multiplied by 100.

Loss	AUC ↑	α AUSE ↓	ρ AUSE↓
NLL[3]	**64.13**$\pm$0.04	44.30$\pm$0.86	51.47$\pm$2.56
Ensemble NLL	**66.83**	39.50	48.00
Plain W_1	64.40$\pm$0.22	100.40$\pm$2.76	153.70$\pm$3.33
hinge-W_1,γ_W=0.0100	66.60$\pm$0.11	49.88$\pm$1.73	78.49$\pm$1.61
hinge-W_1,γ_W=0.0125	65.64$\pm$0.14	39.75$\pm$1.40	64.68$\pm$2.27
hinge-W_1,γ_W=0.0150	64.24$\pm$0.41	29.18$\pm$0.75	69.41$\pm$5.18
hinge-W_1,γ_W=0.0175	62.32$\pm$0.16	29.05$\pm$1.68	37.34$\pm$1.15
hinge-W_1,γ_W=0.0200	62.72$\pm$0.14	**26.97**$\pm$1.19	**30.82**$\pm$1.92
hinge-W_1,γ_W=0.0225	62.48$\pm$0.26	27.80$\pm$0.94	38.83$\pm$11.14
hinge-W_1,γ_W=0.0250	62.00$\pm$0.09	32.45$\pm$1.36	31.11$\pm$6.11

2.2. Synthetic dataset: ablation on Gaussian smoothing

In this section, we show that Gaussian smoothing is beneficial for line regression tasks. Tab. 3 shows the results for training with Dirac ground truth, i.e., no Gaussian smoothing. As the hinge increases, both the regression performance (in terms of AUC) and uncertainty estimation (in terms of AUSE) improve until the hinge value ($\gamma_W = 0.015$) is too large to maintain stable training. Compared with Table 1 in the main paper, we notice that Dirac ground truth has much worse AUC and AUSE at hinge value ($\gamma_W = 0.015$), indicating that Gaussian smoothing can help to maintain a stable training at a large hinge value.

2.3. HLW: ablation on different hinge

Tab. 2 shows the ablation study on different hinge values on HLW. Hinge-Wasserstein with $\gamma_W = 0.02$ achieves the best quality of uncertainty estimation, but it suffers a small drop in the regression performance. It is worth noting that as AUC keeps dropping as γ_W increases, whereas AUSE first drops and then increases. This indicates there exists an optimal value γ_W^* on the HLW dataset. When $\gamma_W > \gamma_W^*$, it will be rather hard to train the neural network, as there will rarely be any gradients from the loss. We also argue that γ_W^* depends on the number of bins in the regression by classification framework. E.g., there are 100 bins for the horizon line detection task, and thus, $\gamma_W = 1/100$ means that hinge-Wasserstein allows a random guess.

2.4. Stereo disparity: ablation on different hinge

We report the results of different hinge values for the stereo disparity task in Tab. 4. As hinge increases, both regression performance (in terms of EPE) and uncertainty estimation (in terms of AUSE) improved for both boundary pixels and all the pixels. This shows our proposed hinge-W_1 improves the challenging multimodal regression.

3. Horizon in the wild: additional qualitative results

Fig. 2 shows more examples of images and the corresponding predicted densities for α and ρ. The peak shapes are more clearly defined for α than for ρ. This is a general trend that we have observed, and it is also consistent with the more focused curves for alpha at the bottom of Fig. 2.

Table 3. Ablation study on the synthetic dataset. For all results we use unimodal ground truth as a Dirac function. Standard deviation is computed over five randomly initialized models.

Loss	AUC ↑	α AUSE ↓	ρ AUSE↓	α CRPS ↓	ρ CRPS ↓
Plain W_1	47.04±0.04	62.73±2.71	56.40±3.71	10.1±0.07	10.8±0.02
hinge-W_1 $\gamma_W = 0.005$	47.08±1.70	55.51±11.35	40.85±14.63	8.55±1.79	9.42±1.52
hinge-W_1 $\gamma_W = 0.01$	53.44±0.05	57.43±3.51	42.70±3.60	9.55±0.43	9.96±0.04
hinge-W_1 $\gamma_W = 0.015$	21.52±0.14	69.43±9.32	58.81±13.42	9.77±6.02	9.45±0.42

Table 4. Stereo disparity results on Scene Flow. Regression performance in terms of EPE, 1PE, and 3PE, and uncertainty evaluation in terms of entropy-based AUSE. MM denotes multimodal training with $k = 5$, and standard error is reported over five runs.

2*Setting	2*Loss	All pixels				Edge pixels			
		EPE ↓	1PE ↓	3PE ↓	AUSE ↓	EPE ↓	1PE ↓	3PE ↓	AUSE ↓
Softmax	Plain W_1 [1]	**0.98**±0.01	**9.44**±0.06	**4.04**±0.03	19.4±0.37	**3.05**±0.03	**17.4**±0.12	**10.1**±0.10	27.5±0.70
Softmax	*hinge*-W_1, $\gamma_W = 0.0075$ (Ours)	1.03±0.02	9.80±0.11	4.19±0.05	**18.7**±0.29	3.11±0.01	17.8±0.09	10.3±0.05	26.7±0.41
Softmax	*hinge*-W_1, $\gamma_W = 0.01$ (Ours)	0.99±0.01	9.62±0.06	4.08±0.03	**18.7**±0.43	**3.05**±0.03	17.6±0.12	**10.1**±0.06	**26.4**±0.31
Softplus	Plain W_1 [1]	1.00±0.01	9.74±0.07	4.12±0.03	18.1±0.89	3.05±0.01	17.5±0.09	10.1±0.07	27.2±1.64
Softplus	*hinge*-W_1, $\gamma_W = 0.001$ (Ours)	0.97±0.01	9.48±0.05	4.05±0.03	17.4±0.44	3.00±0.02	17.2±0.12	9.91±0.05	26.1±0.59
Softplus	*hinge*-W_1, $\gamma_W = 0.0025$ (Ours)	0.97±0.02	9.35±0.16	3.97±0.06	16.5±0.55	**2.98**±0.03	**17.1**±0.16	**9.80**±0.07	23.6±0.63
Softplus	*hinge*-W_1, $\gamma_W = 0.005$ (Ours)	**0.96**±0.01	**9.31**±0.05	**3.96**±0.03	16.0±0.39	3.00±0.03	**17.1**±0.11	9.84±0.12	23.5±0.58
Softplus	*hinge*-W_1, $\gamma_W = 0.0075$ (Ours)	1.00±0.01	9.52±0.06	4.06±0.04	15.6±0.33	3.07±0.01	17.4±0.09	10.0±0.07	23.2±0.71
Softplus	*hinge*-W_1, $\gamma_W = 0.01$ (Ours)	0.98±0.01	9.48±0.06	4.05±0.03	16.4±0.48	3.04±0.02	17.2±0.16	9.99±0.10	23.1±0.77
Softplus	*hinge*-W_1, $\gamma_W = 0.0125$ (Ours)	1.00±0.01	9.60±0.08	4.12±0.05	15.8±0.27	3.06±0.02	17.4±0.08	10.0±0.06	**21.7**±0.50
Softplus	*hinge*-W_1, $\gamma_W = 0.015$ (Ours)	0.97±0.02	9.38±0.13	3.98±0.06	16.1±0.60	3.01±0.03	17.2±0.21	9.88±0.11	22.9±0.80
Softplus	*hinge*-W_1, $\gamma_W = 0.02$ (Ours)	0.98±0.01	9.46±0.09	4.04±0.03	**15.5**±0.19	3.03±0.03	17.2±0.13	9.94±0.09	21.8±0.37
Softplus	*hinge*-W_1, $\gamma_W = 0.04$ (Ours)	1.02±0.01	9.75±0.09	4.17±0.05	**15.1**±0.26	3.13±0.02	17.7±0.14	10.3±0.08	**21.3**±0.49
Softplus, MM	Plain W_1 [1]	1.00±0.03	9.61±0.25	4.15±0.16	14.1±1.46	3.15±0.11	17.59±0.38	10.3±0.26	19.8±2.08
Softplus, MM	*hinge*-W_1, $\gamma_W = 0.0075$ (Ours)	**0.96**±0.01	**9.27**±0.11	**3.94**±0.03	13.0±0.32	**3.00**±0.03	**17.00**±0.17	**9.79**±0.13	17.2±0.01
Softplus, MM	*hinge*-W_1, $\gamma_W = 0.01$ (Ours)	0.97±0.03	9.40±0.21	4.01±0.08	**12.6**±0.01	3.04±0.03	17.20±0.08	9.98±0.03	**16.3**±0.90

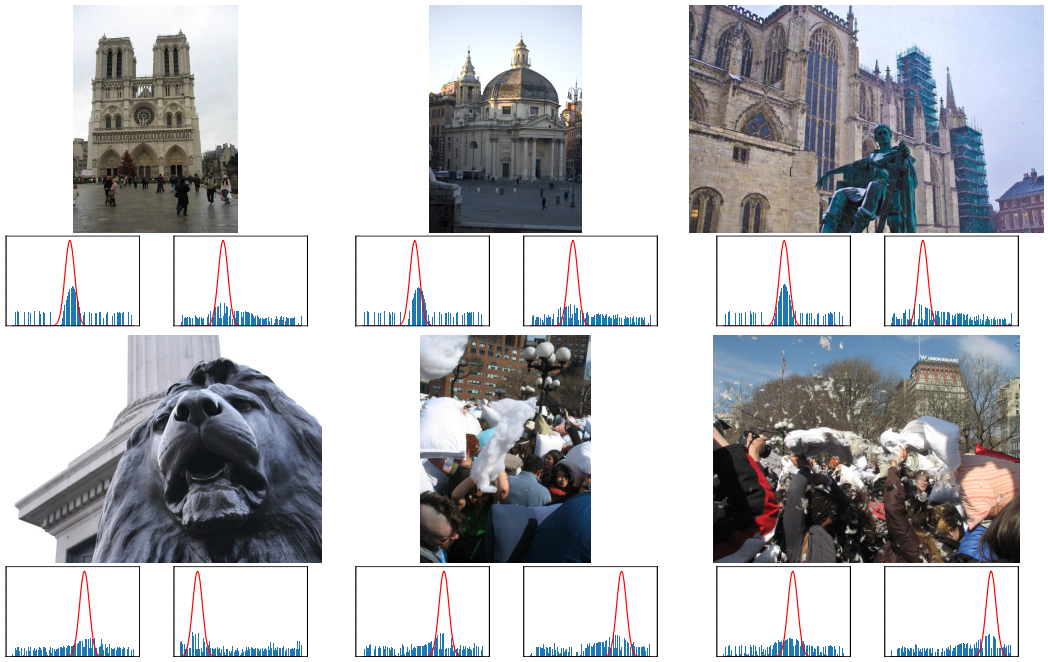

Figure 2. Horizon line detection should be framed as a probabilistic regression problem due to its inherently stochastic nature. First row: Images where horizon line detection is easy (red line) and direct regression would work. Second row: Image where horizon line detection is hard and direct regression would not work. Plots below the images show the output probability distributions for the horizon line parameters (α, ρ), from the proposed method. Red: ground truth; Blue: predicted density. Images are from the HLW dataset [3].

References

[1] Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Weinberger, and Wei-Lun Chao. Wasserstein distances for stereo disparity estimation. In *NeurIPS*, 2020. 3

[2] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667, 2018. 1

[3] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild. In *British Machine Vision Conference (BMVC)*, pages 20.1–20.12, 2016. Acceptance rate: 39.4%. 2, 3