Supplementary Material for "Sampling-free Epistemic Uncertainty Estimation Using Approximated Variance Propagation"

Subsequently we provide additional information. In section 1 we write down the Jacobians which are used to approximately propagate the covariance matrix through non-linearities. Section 2 and 3 we derive the formula for the covariance matrix of the element-wise product of independent random vectors and, respectively, the expectation and variance of ReLU given a Gaussian distribution. Section 4 compares the performance of Monte-Carlo dropout with similar runtime (few samples) with our work. Further, we show empirically in section 5 for that the sampling-based approach converges to our analytic form for the case of the synthetic dataset. Finally in section 6 we show further qualitative results of our experiments. We also refer to the video which can be found in the supplementary material for qualitative results.

1. Jacobians of Activation Functions

We show the Jacobians of the activation functions - ReLU, sigmoid and softmax - that are used throughout our experiments. For ReLU we assume its derivative in the origin to be zero. Then the Jacobian is given by

$$J_{ij}(ReLU(\vec{x})) = \begin{cases} 1, & \text{if } i = j \text{ and } x_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In case of an element-wise sigmoid function $\sigma(\vec{x})$ the Jacobian is given by

$$J_{ij}(\sigma(\vec{x})) = \begin{cases} \sigma(x_i)(1-\sigma(x_i)), & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

and for the softmax respectively

$$J_{ij} = S_i(\delta_{ij} - S_j) \tag{3}$$

where S_i and S_j are the i-th and j-th entry of the softmax output and δ_{ij} is the Kronecker delta.

2. Covariance of Hadamard Product of Random Vectors

In the following X and Z denote random variables and $\vec{X}$ and $\vec{Z}$ denote random vectors, which may each have a non-diagonal covariance matrix but do not depend on each other. Further $\Sigma_{\vec{X}}$ denote the covariance matrix of $\vec{X}$.

We are interested in the covariance matrix of $\vec{Y} = \vec{Z} \circ \vec{X}$ resulting from an element-wise multiplication of $\vec{Z}$ and $\vec{X}$. Therefore we plug $\vec{Z} \circ \vec{X}$ into the definition of the covariance matrix:

$$\Sigma_{\vec{Z}\circ\vec{X}} = E[(\vec{Z} \circ \vec{X})(\vec{Z} \circ \vec{X})^T] - E[\vec{Z} \circ \vec{X}]E[\vec{Z} \circ \vec{X}]^T \tag{4}$$

Given that $\vec{Z}$ and $\vec{X}$ are independent and that

$$(\vec{Z} \circ \vec{X})(\vec{Z} \circ \vec{X})^T = (\vec{Z}\vec{Z}^T) \circ (\vec{X}\vec{X}^T) \tag{5}$$

Eq. 4 yields:

$$\Sigma_{\vec{Z}\circ\vec{X}} = E[\vec{Z}\vec{Z}^T] \circ E[\vec{X}\vec{X}^T] - (E[\vec{Z}]E[\vec{Z}]^T) \circ (E[\vec{X}]E[\vec{X}]^T) \tag{6}$$

Now we can compare Eq. 6 with

$$\begin{aligned} &\Sigma_Z \circ \Sigma_X = \\ &(E[\vec{Z}\vec{Z}^T] - E[\vec{Z}]E[\vec{Z}]^T) \circ (E[\vec{X}\vec{X}^T] - E[\vec{X}]E[\vec{X}]^T) \\ &= E[\vec{Z}\vec{Z}^T] \circ E[\vec{X}\vec{X}^T] + E[\vec{Z}]E[\vec{Z}]^T \circ E[\vec{X}]E[\vec{X}]^T - \\ &E[\vec{Z}\vec{Z}^T] \circ E[\vec{X}]E[\vec{X}]^T - E[\vec{X}\vec{X}^T] \circ E[\vec{Z}]E[\vec{Z}]^T \end{aligned} \tag{7}$$

and see that Eq. 6 is equivalent to

$$\begin{aligned} &\Sigma_{\vec{Z}\circ\vec{X}} = \Sigma_Z \circ \Sigma_X + \\ &E[\vec{Z}\vec{Z}^T] \circ E[\vec{X}]E[\vec{X}]^T + E[\vec{X}\vec{X}^T] \circ E[\vec{Z}]E[\vec{Z}]^T - \\ &2(E[\vec{Z}]E[\vec{Z}]^T) \circ (E[\vec{X}]E[\vec{X}]^T) \\ &= \Sigma_Z \circ \Sigma_X + E[\vec{X}]E[\vec{X}]^T \circ (E[\vec{Z}\vec{Z}^T] - E[\vec{Z}]E[\vec{Z}]^T) + \\ &E[\vec{Z}]E[\vec{Z}]^T \circ (E[\vec{X}\vec{X}^T] - E[\vec{X}]E[\vec{X}]^T) \end{aligned} \tag{8}$$

Using the definition of the covariance matrix we obtain the desired form of the equation:

$$\Sigma_{\vec{Z}\circ\vec{X}} = \Sigma_Z \circ \Sigma_X + E[\vec{X}]E[\vec{X}]^T \circ \Sigma_Z + E[\vec{Z}]E[\vec{Z}]^T \circ \Sigma_X \quad (9)$$

3. Expectation and Variance of ReLU Given a Gaussian Distribution

We write down the first- and second-order moments of a $f(X) = max(0, X)$ where X is a scalar, normal distributed random variable. We are only interested in scalar inputs since we write out these formulars for the assumption of a diagonal covariance matrix.

Given a univariate Gaussian $N(\mu, \sigma)$ with mean μ and standard deviation σ, the expectation of $f(X)$ is determined by the integral

$$E_{X\propto N(\mu,\sigma)}[max(0,X)] = \frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty x exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx = \sqrt{\frac{1}{2\pi}}\sigma exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \frac{\mu}{2}\left(1 - erf\left(\frac{\mu}{\sqrt{2}\sigma}\right)\right) \quad (10)$$

where $erf(x)$ is the error function with

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x exp(-z^2)dz \quad (11)$$

The variance is then given by

$$Var_{X\propto N(\mu,\sigma)}[max(0,X)] = E_{X\propto N(\mu,\sigma)}[max(0,X)^2] - E_{X\propto N(\mu,\sigma)}[max(0,X)]^2 \quad (12)$$

Here we know $E_{X\propto N(\mu,\sigma)}[max(0,X)]^2$ via the expectation. The other term yields

$$E_{X\propto N(\mu,\sigma)}[max(0,X)^2] = \frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty x^2 exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx = \frac{1}{2}\left(\sigma+\mu^2\right)\left(1 + erf\left(\frac{\mu}{\sqrt{2}\sigma}\right)\right) + \frac{\mu\sigma}{\sqrt{2\pi}}exp\left(-\frac{\mu^2}{2\sigma^2}\right) \quad (13)$$

4. Comparison with Monte-Carlo (MC) Dropout of Similar Computational Cost

We evaluate the predicted mean and standard deviation (STD) of our approach (OUR) and MC dropout on Boston Housing treating the result with 10000 samples as ground truth (GT). OUR is of computational advantage for more than approximately 175 samples. Fig. 1 shows the Mean absolute difference of MC dropout/OUR and GT (red/blue) depending on the number of samples. Even for up to 500 samples MC dropout fails to match the accuracy of our STD approximation. The mean approximation of MC dropout performs already better for much fewer samples (>100).

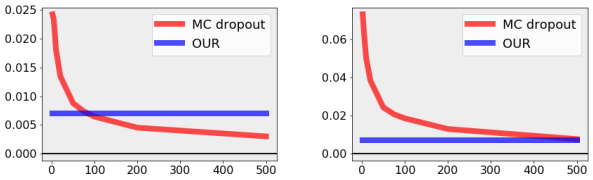

Figure 1. Mean absolute difference of MC dropout/OUR and GT (red/blue) depending on the number of samples on Boston Housing. OUR is constant without sampling. Left: Mean. Right: STD.

5. Absolute Variance Difference for Synthetic Data

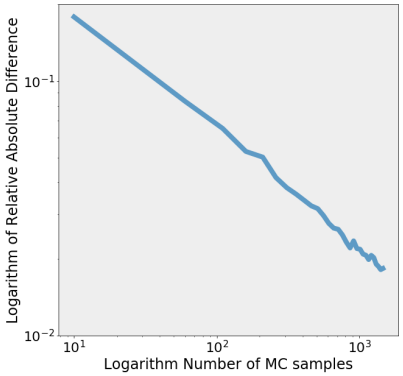

Figure 2. Relative absolute error between standard deviation obtained by Monte-Carlo dropout [2] and our approximation in a double logarithmic plot. We observe that the relative absolute difference approaches increasingly small values for larger numbers of samples.

We fit a neural network to a synthetic dataset. In 2 we show empirically that the sampling-based variance estimate converges to our analytic expression. We observe that the relative absolute difference between the sampling-based variance estimate and our approximation converges to zero for large numbers of samples.

6. Qualitative Results Including Our Prediction

6.1. Bayesian SegNet [5]

We show more qualitative results of our approximation using Bayesian SegNet[5] on CamVid dataset [1]. These

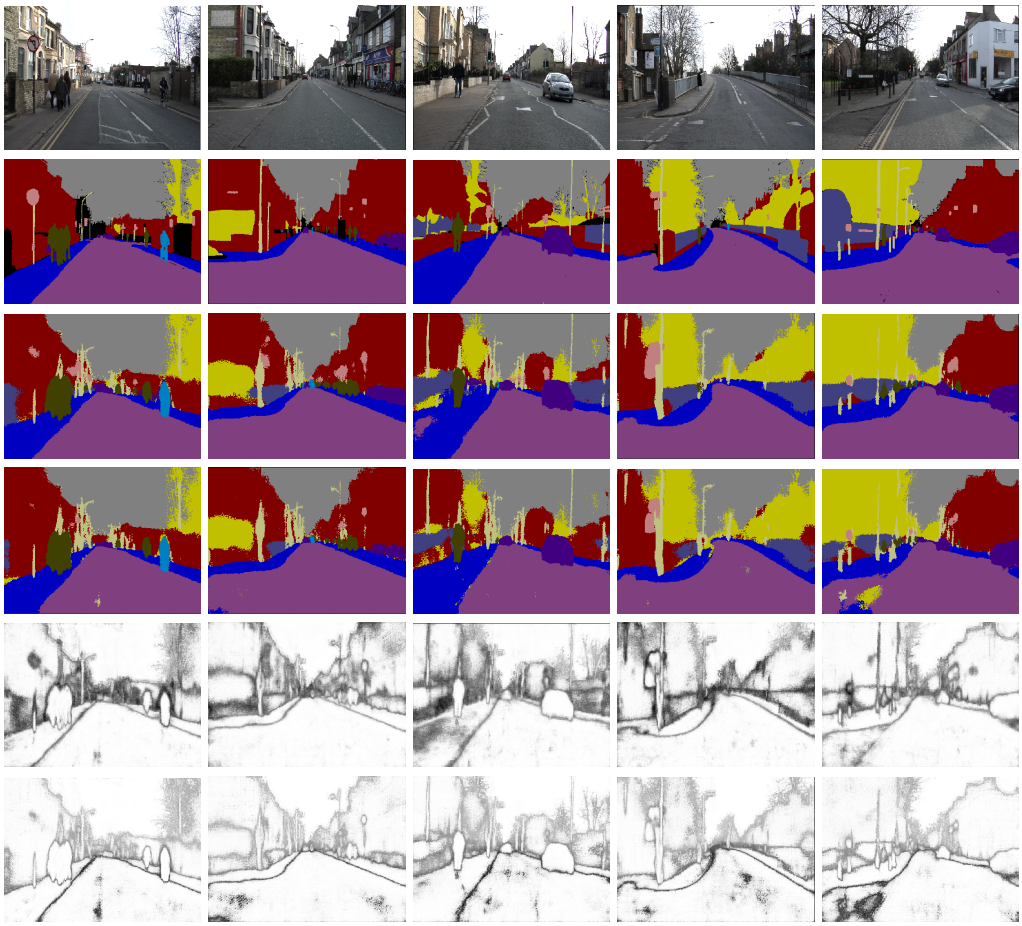

Figure 3. Qualitative results of Bayesian SegNet [5] on CamVid [1]. First row: Original images. Second Row: Ground truth. Third row: Prediction using MC dropout. Fourth row: Our prediction (normal dropout activation scaling). Fifth row: Uncertainty using MC dropout. Sixth row: Our approximation.

are shown in Fig. 3. We observe that the network is mostly uncertain about object boundaries. We refer to the video in the supplementary material for more qualitative results.

6.2. Monocular Depth Regression [4]

We show more qualitative results of our approximation using monocular depth regression [4] on KITTI dataset [3]. These are shown in Fig. 4. We observe that the network is very certain about the region of highest depth resolution and generally uncertain about the left and right border of the image. The latter results from the use of non-overlapping stereo images at training time. We refer to the video in the supplementary material for more qualitative results.

6.3. Qualitative Results of Class Hold-out

We trained BayesianSegNet[1] withholding the classes for pedestrians and cyclists. Here we show qualitative results of this experiment (see Fig. 5). We observe that the locations of withheld classes tend to be more uncertain than other regions of the image.

[1] only one dropout layer prior to the final layer

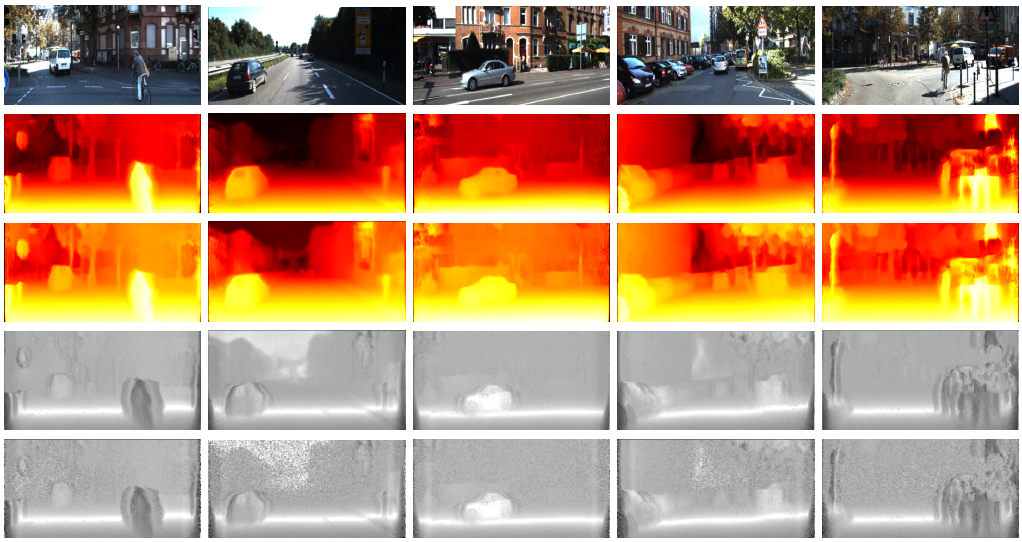

Figure 4. Qualitative results of monocular depth regression [4] on KITTI [3]. First row: Original images. Second Row: Prediction using MC dropout. Thrid row: Our prediction (normal dropout activation scaling). Fourth row: Uncertainty using MC dropout. Fifth row: Our approximation.

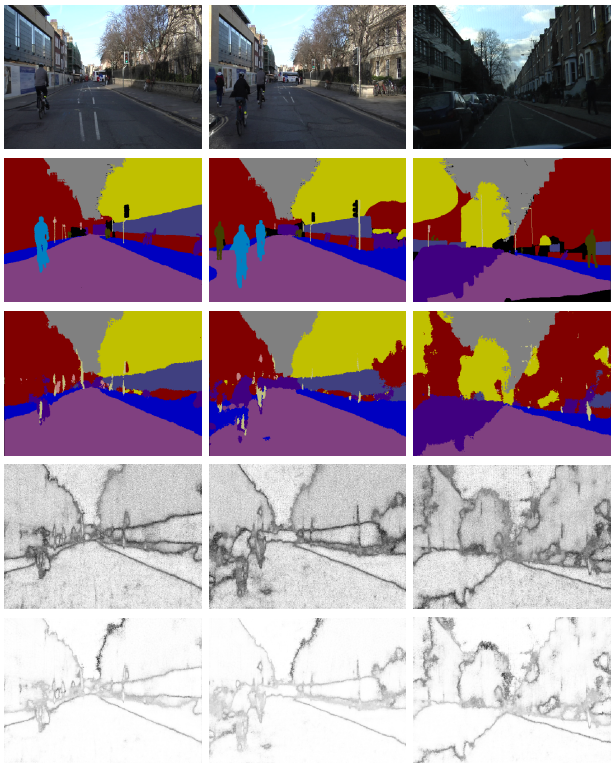

Figure 5. Qualitative results when trained without pedestrian and bicyclist classes. First row: Input image. Second row: Ground truth. Third row: Segmentation result using Monte-Carlo dropout [2]. Fourth row: Uncertainty estimate using Monte-Carlo dropout. Fifth row: Our approximation.

References

[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[2] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[5] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *CoRR*, abs/1511.02680, 2015.