A ESTIMATING MUTUAL INFORMATION BY BINNING

This section presents a replication of the experiments from Shwartz-Ziv & Tishby (2017) and Saxe et al. (2018), but varying the number of bins for estimating MI. Non-quantized TANH and RELU networks were trained in the standard setting, and MI was estimated using 30, 100 and 256 bins. The number of bins were chosen to match the numbers used in the earlier works (30 used by Shwartz-Ziv & Tishby, 100 used by Saxe et al.) and to match the discretization in the quantized neural networks considered in Section 4.1.

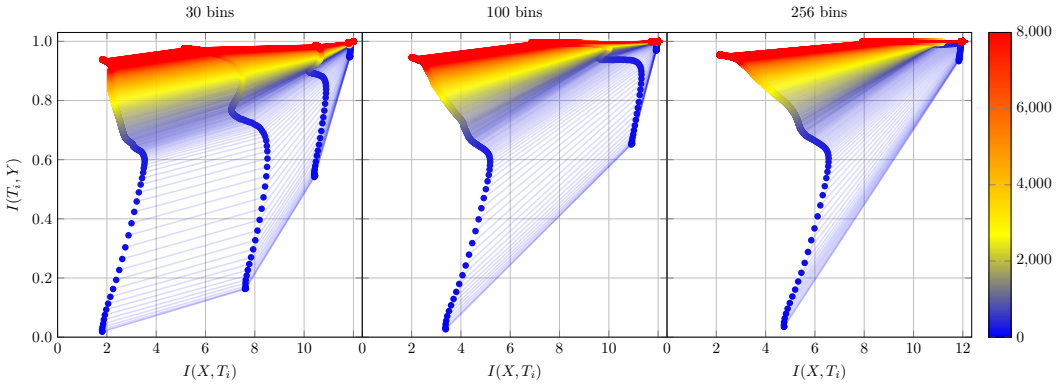

Figure A.5: Information planes for the TANH network with various number of bins. 30 bins corresponds to a replication of the results from Shwartz-Ziv & Tishby (2017). The network is trained in the standard setting, and the average of 50 runs is plotted.

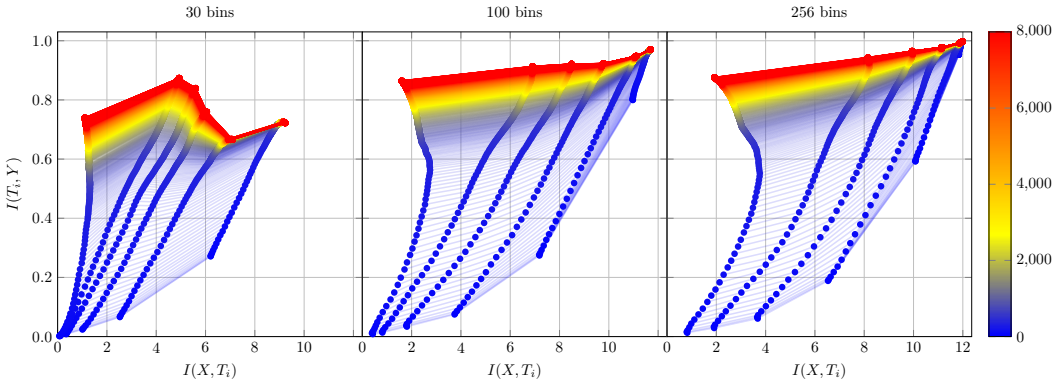

Figure A.6: Information planes for the RELU network with various number of bins. 100 bins replicate the results from Saxe et al. (2018). The network is trained in the standard setting, and the average of 50 runs is plotted.

The resulting information planes are shown for the TANH and RELU networks in Figure A.5 and Figure A.6 respectively.

From Figure A.5, we clearly see the results of Shwartz-Ziv & Tishby (2017) replicated using 30 bins for the TANH network; two phases are clearly visible in each layer. As expected, for every layer T with binned $T' = B(T)$, $\mathrm{I}(X;T')$ and $\mathrm{I}(T';Y)$ increase with the number of bins used with the phases still remaining visible in the distinguishable layers.

For the RELU network with 100 bins, the results of Saxe et al. (2018) were also replicated, that is, the compression phase was not observed. When using only 30 bins for the RELU network, the estimation broke down completely in the sense that the DPI (2) is violated: $\mathrm{I}(T';Y)$ is non-monotone, a phenomenon occurring because of the estimation, which has also been observed by Geiger (2020). Again, we see for any layer T that $\mathrm{I}(X;T')$ and $\mathrm{I}(T';Y)$ increase with m, although to a lesser degree

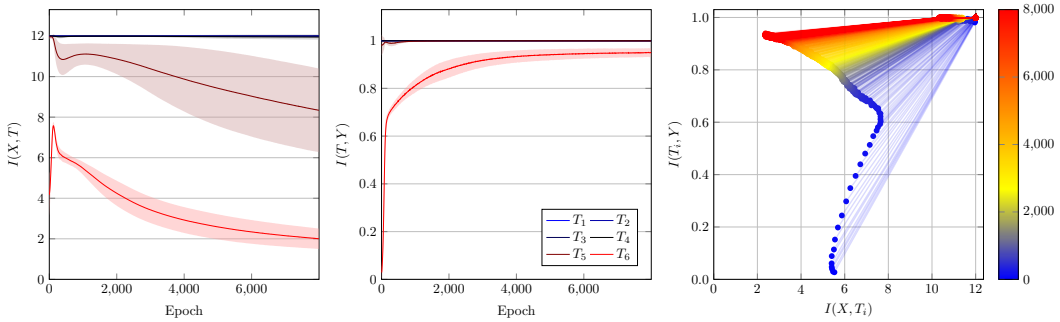

Figure B.7: Mean and variance for $\mathrm{I}(X;T)$ (left) and $\mathrm{I}(T;Y)$ (center), and the information plane for the median deviating (L2-distance from the mean) repetition (right), for the TANH network in the standard setting.

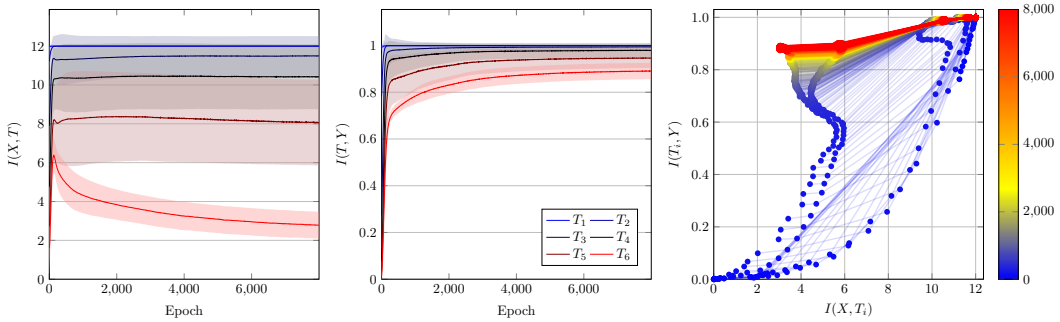

Figure B.8: Mean and variance for $\mathrm{I}(X;T)$ (left) and $\mathrm{I}(T;Y)$ (center), and the information plane for the median deviating (L2-distance from the mean) repetition (right), for the RELU network in the standard setting.

than for the TANH network, most likely due to the RELU activation functions actually dropping information (by zeroing negative activations).

In general, the estimation of MI in either network is highly dependable on the number of bins used, indicating that the interpretation of the information plane is not straight forward, when using binning for estimation.

B MI VARIANCE

New section with additional plots showing the variance across repetitions.

This section investigates how the information plane varies in the experimental trials.

Figures B.7, B.8 and B.9 show the mean and variance of $\mathrm{I}(X;T)$ (left) and $\mathrm{I}(T;Y)$ (center), as well as the information plane for a single trial (right), for the TANH and RELU networks in the standard setting, and the bottleneck network applied to MNIST, respectively. For the single trial, we selected the median deviating repetition, ranked by the L2-distance from the mean.

From the plots, we see that when the mean MI is large, the variance is usually low. The variance is similar across the different networks (note the different scaling of the y-axes), with slightly larger variance for the RELU activations. Together with the median deviating information plane, the plots suggest that the mean information plane represents individual repetitions fairly well.

Additionally, the compression phase in the TANH network can clearly be seen in Figure B.7 (left), while it is absent in the RELU network (Figure B.8).

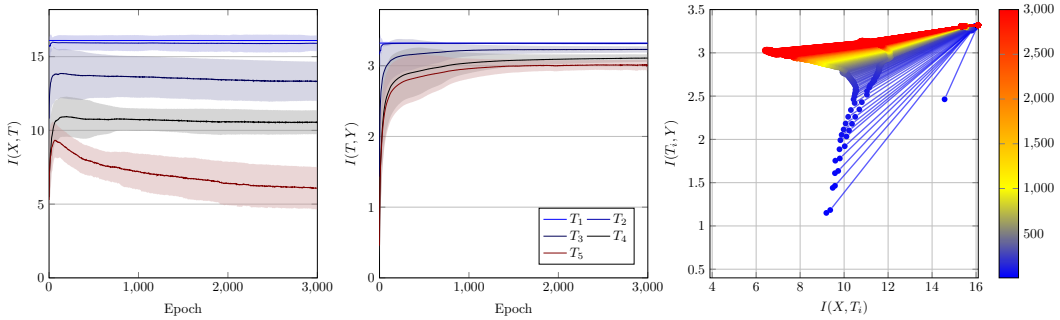

Figure B.9: Mean and variance for $\mathrm{I}(X;T)$ (left) and $\mathrm{I}(T;Y)$ (center), and the information plane for the median deviating (L2-distance from the mean) repetition (right), for the Bottleneck network applied to MNIST.

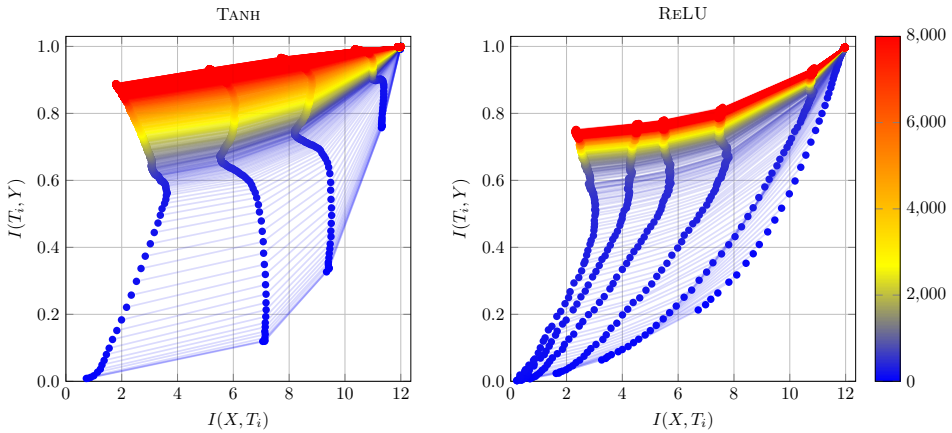

Figure C.10: Information planes for 4-bit quantized TANH network (left) and RELU network (right). The mean of 30 runs is plotted.

C EFFECT OF THE QUANTIZATION PRECISION

Merged from the 32-bit and 4-bit experiment appendix sections.

This section presents information planes obtained using 4- and 32-bit quantized neural networks trained in the standard setting, in order to investigate the effect of the precision used in the quantization.

Figure C.10 depicts the resulting information planes for the 4-bit networks. As each neuron can take only $2^4 = 16$ different values, the total number of possible states per layer decreases significantly in this setting, and as expected we see lower overall MI measurements. For the TANH network, several more layers become distinguishable. The observed information planes looks similar to those observed in the original experiments by Shwartz-Ziv & Tishby (2017) and Saxe et al. (2018). However, the network accuracy has now degraded compared to the non-quantized networks (see Supplementary Material F), which indicates that the binning used in the estimation of the MI in previous experiments has discarded information vital to the network.

Figure C.11 shows the resulting information planes for the 32-bit networks. As expected, we see an overall increase in MI; the information drops only very slowly through the network. Each layer has many possible states and – given the small data set – we we get closer to the behavior of a continuous system.

D QUANTIZED NETWORK WITH RANDOMIZED PREFITTING

Added control experiment with prefitting on random labels.

This section presents results for the standard setting using 8-bit quantized networks, but with weights prefitted to random labels. First, the network is fitted for 1000 epochs to the synthetic data set with

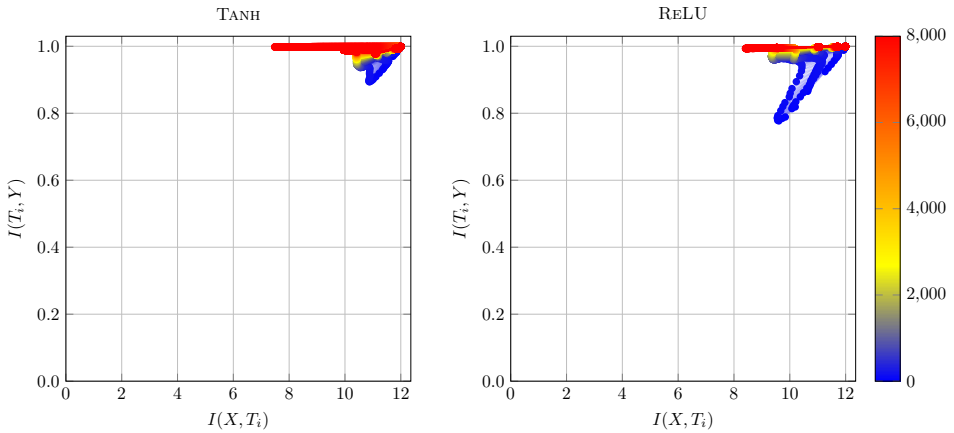

Figure C.11: Information planes for 32-bit quantized TANH network (left) and RELU network (right). The mean of 30 runs is plotted.

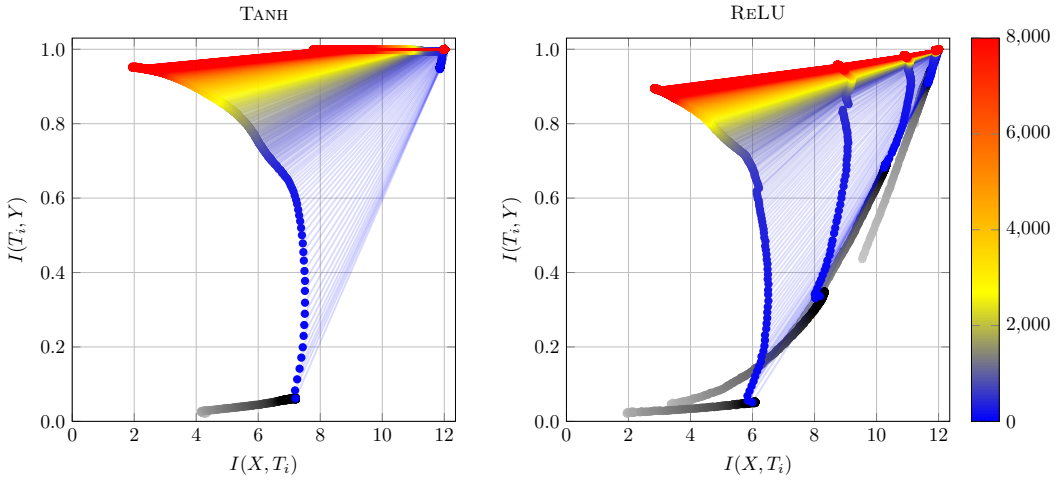

Figure D.12: The information planes for the TANH (left) and RELU (right) networks when prefitting on random labels for 1000 epochs. The means of 20 repetitions are shown. In the plots, Y always refers to the true labels.

the labels shuffled. The network is then trained for 8000 epochs with correct labels (as done in Section 4.1). The experiment is repeated 20 times. The resulting information planes for the TANH and the RELU network are presented in Figure D.12, left and right respectively.

Inspecting the plots, we see that, for all discernible layers, $\mathrm{I}(X;T)$ appears to increase when fitted to random labels. This is not surprising as the input is not shuffled.

As a consequence of the increase in $\mathrm{I}(X;T)$, the hidden layers (most notable in the RELU network) also see an increase in $\mathrm{I}(T;Y)$. As the ability to distinguish different inputs from a hidden layer T increases, so does $\mathrm{I}(T;Y)$.

As expected, the network accuracy does not increase when training on random labels, and accordingly there is no lateral movement in the information plane for the output layer.

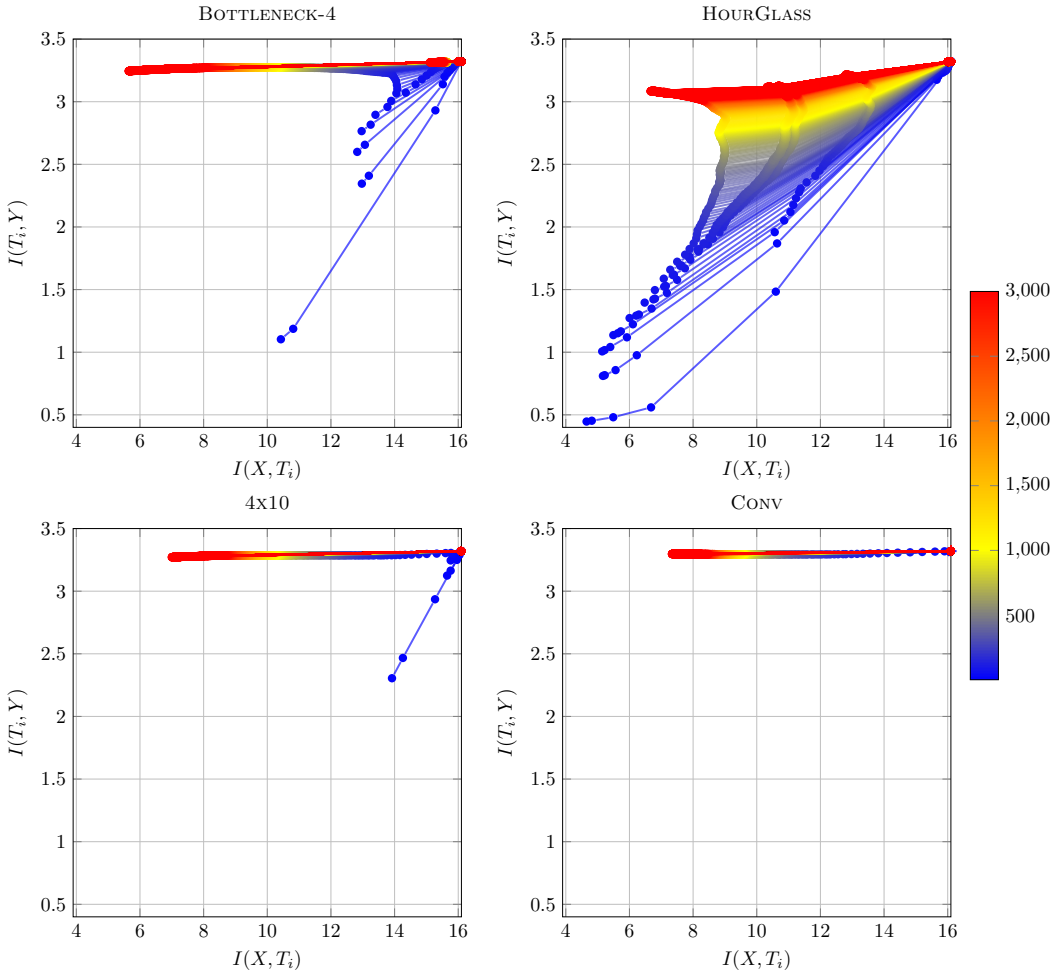

Figure E.13: Information planes for the additional networks trained on MNIST. Note for HOURGLASS that the three last hidden layers almost coincide completely.

E MNIST WITH OTHER NETWORK ARCHITECTURES

Updated with additional architectures (before the section included only the 4X10 architecture).

This section presents the results of training a number of 8-bit quantized networks with varying architectures on MNIST. We consider the same setup as in Section 4.2, but with the following architectures (all using RELU activations in the hidden layers and a final 10 layer SOFTMAX output):

- The BOTTLENECK-4 network with 4 hidden layers of widths 16, 12, 8 and 4, i.e. the bottleneck has width 4.
- The HOURGLASS network with 6 hidden layers of widths 16, 8, 4, 2, 4, 8. The bottleneck width is still 2, but the network expands again after the bottleneck, creating an hourglass shape.
- The 4X10 network with 4 hidden layers, each of width 10.
- The CONV network, a simple convolutional network with structure: CV-MP-CV-MP-FC, where CV denotes a (3,3) 2-channel convolutional layer, MP denotes a (2,2) max pooling and FC denotes a fully connected RELU layer with 20 neurons.

Below, we denote the network used in Section 4.2 by BOTTLENECK-2. Each experiment is repeated 10 times and the resulting information planes are reported in Figure E.13. As expected, given the increased complexity, all networks exhibit better performance than the bottleneck structured network considered in Section 4.2 (the accuracy of all networks are presented in Figure F.16 in Supplementary

Material F), obtaining accuracies close to or above 95% for all networks except the HOURGLASS network.

From Figure E.13, we observe the information curves for the BOTTLENECK-4 network to be similar in shape to the BOTTLENECK-2, but with higher MI measurements in general. This is not surprising, as the layers are wider and thus more expressive.

As expected, the HOURGLASS network also exhibits information curves similar to BOTTLENECK-2 (as the architecture is similar for the first few layers). Not surprisingly, the two expanding layers of widths 4 and 8 almost coincide completely with the bottleneck layer; information cannot increase but is also not decreased significantly.

The two more expressive networks without bottlenecks, 4X10 and CONV, have almost completely trivial information curves. Limited fitting for the output layer in 4X10 and compression in the output layers for both networks can be observed. The experiments indicate the limitations of exact IB analysis for complex, large scale networks.

F NETWORK ACCURACIES

Figure F.14 and Figure F.15 report the accuracies obtained by the TANH and RELU networks when fitted on the synthetic data, respectively. As can be seen, the networks suffers only slightly when

Updated figure with reruns of bad repetitions.

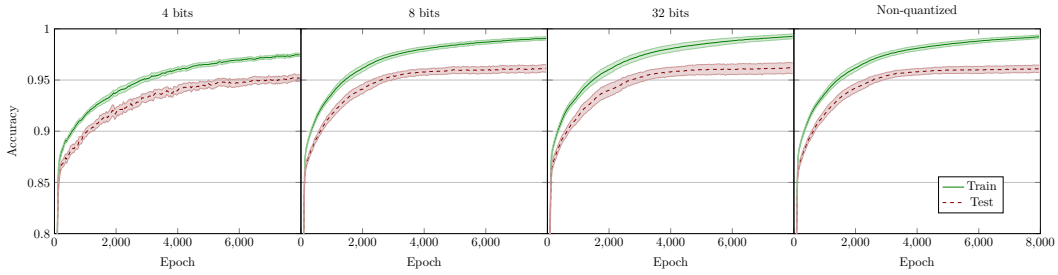

Figure F.14: Accuracies of the quantized TANH networks (three on the left) and the non-quantized TANH network (right). The means and 95% confidence intervals over 50 repetitions are reported.

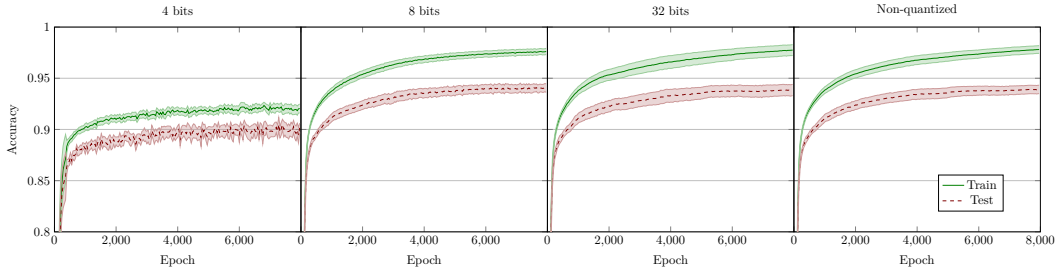

Figure F.15: Accuracies of the quantized RELU networks (three on the left) and the non-quantized RELU network (right). The means and 95% confidence intervals over 50 repetitions are reported.

using only 4 bits in the quantization.

Figure F.16 reports the accuracy of the networks fitted to MNIST. Unsurprisingly, the more complex networks (BOTTLENECK-4, 4X10, CONV) obtain better accuracy with lower variance, compared to the networks with a low-width bottleneck (BOTTLENECK-2, HOURGLASS).

Updated figure with additional architectures and with re-runs of bad repetitions.

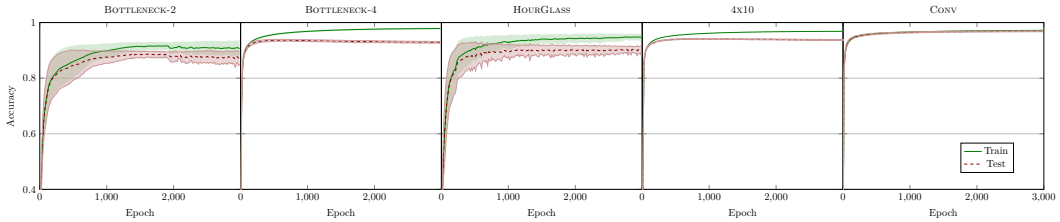

Figure F.16: Accuracies of the quantized networks applied to MNIST. The means and 95% confidence intervals over 20 (BOTTLENECK-2 network) and 10 (other networks) repetitions are reported. The decrease in training accuracy for the bottleneck structured network (left) is due to the optimization objective being cross-entropy.