A Probabilistic Deep Image Prior over Image Space

Riccardo Barbano[†] RICCARDO.BARBANO.19@UCL.AC.UK
University College London

Javier Antorán[†] JA666@CAM.AC.UK
University of Cambridge

José Miguel Hernández-Lobato JMH233@CAM.AC.UK
University of Cambridge

Bangti Jin B.JIN@UCL.AC.UK
University College London

Abstract

The deep image prior regularises under-specified image reconstruction problems by reparametrising the target image as the output of a CNN. We induce a prior over images by scoring CNN outputs using a classical image reconstruction regulariser. We translate this functional prior into weight space using a change of variables and propose an efficient linearised Laplace inference algorithm. Hyperparameters are optimised with Type-II MAP. We obtain pixel-wise uncertainty estimates, which we show to be calibrated to the reconstruction error.

1. Introduction

Inverse problems in imaging centre around the recovery of an unknown image $\mathbf{x} \in \mathbb{R}^{d_x}$ from the corrupted measurement

$$\mathbf{y}_\delta = \mathrm{A}\mathbf{x} + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{y}_\delta \in \mathbb{R}^{d_y}$ is the noisy measurement data, $\mathrm{A} \in \mathbb{R}^{d_y \times d_x}$ a linear forward operator, and $\boldsymbol{\eta}$ an i.i.d. noise (e.g., Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma_y^2 \mathrm{I})$). Many tomographic reconstruction problems take this form (e.g., computed tomography (CT), magnetic resonance imaging and magnetic particle imaging). Due to the inherent ill-posedness of the reconstruction task (e.g., in the under-determined case $d_y \ll d_x$), suitable regularisation is crucial and is key for a successful recovery of $\mathbf{x}$ (Tikhonov and Arsenin, 1977; Engl et al., 1996; Ito and Jin, 2014). A successful approach is the deep image prior (DIP), introduced by Ulyanov et al. (2018), which regularises the reconstruction by reparametrising $\mathbf{x}$ as the output of a deep convolutional neural network (CNN). Liu et al. (2019) and Baguer et al. (2020) combine DIP with the total variation (TV) regulariser, one of the most popular and well-established penalties for image reconstruction (Rudin et al., 1992).

In this work, we re-cast the deep image prior as a probabilistic prior over images. Within the probabilistic framework, a natural approach to regularise the ill-posed inverse imaging problem is to place a prior over the reconstruction. Our prior scores the outputs of a neural network according to their TV seminorm. We then translate this functional prior into weight space using the predictive complexity prior (PredCP) framework of Nalisnick et al. (2021). We provide a linearised Laplace inference algorithm with computational

[†] Equal contribution.

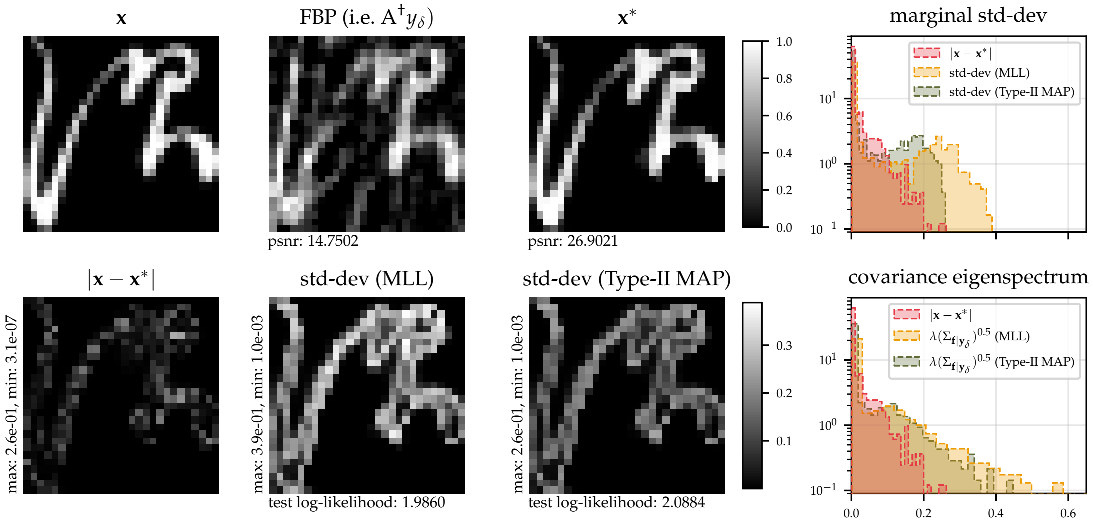

Figure 1: Example of an original image $\mathbf{x}$. Its noisy measurement $\mathrm{A}\mathbf{x}+\boldsymbol{\eta}$, is reconstructed by FBP, $\mathrm{A}^\dagger\mathbf{y}_\delta$. The DIP provides a higher quality image $\mathbf{x}^*$. The pixel-wise DIP reconstruction error $|\mathbf{x}-\mathbf{x}^*|$ visually correlates with the uncertainties provided by our method. The scale of the pixel-wise standard deviation (std-dev) obtained when employing the proposed prior (Type-II MAP) matches the error more closely than when the hyperparameters are optimised with standard marginal likelihood.

cost scaling in $O(d_x d_y^2)$, and select hyperparameters via a Type-II MAP objective with a complexity $O(d_y^3)$. Our method provides pixel-wise uncertainty estimates, which we show to be predictive of the reconstruction error. We show that our TV-PredCP prior results in increased calibration, cf. fig. 1. A distinct feature of the proposed approach is that it keeps the mean reconstruction obtained using the traditional regularised DIP formulation, which allows us to leverage state of the art DIP training methods (Baguer et al. (2020); Barbano et al. (2021)).

2. Deep Image Prior and Total Variation Regulariser

DIP (Ulyanov et al., 2018, 2020) aims at finding the minimiser of the fidelity $\|\mathrm{A}\mathbf{x}-\mathbf{y}_\delta\|^2$, by assuming that the unknown $\mathbf{x}$ is the output of a CNN, $\mathbf{x}=\mathbf{f}(\mathrm{A}^\dagger\mathbf{y}_\delta,\boldsymbol{\theta})$. As input it takes an approximate reconstruction $A^\dagger\mathbf{y}_\delta$ of $\mathbf{x}$ (Barbano et al., 2021), which is constant and we can thus abbreviate our notation $\mathbf{f}(\mathrm{A}^\dagger\mathbf{y}_\delta,\boldsymbol{\theta})=\mathbf{f}(\boldsymbol{\theta})$. $A^\dagger$ is an approximate inverse, e.g., filter back-projection (FBP) in CT (Dudgeon and Mersereau, 1984). $\boldsymbol{\theta}\in\mathbb{R}^{d_\theta}$ denotes the network's parameters to be learned. The standard choice of the network architecture $\mathbf{f}$ is U-Net (Ronneberger et al., 2015). Thus, DIP solves

$$\boldsymbol{\theta}^*\in\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^{d_\theta}}\|\mathrm{A}\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}_\delta\|^2, \tag{2}$$

and presents $\mathbf{f}(\boldsymbol{\theta}^*)$ as the recovered image. Ulyanov et al. (2018) use early stopping to avoid overfitting to noise in $\mathbf{y}_\delta$. However, the need for early stopping can be avoided by properly regularising the TV of the network output, leading to an objective

$$\boldsymbol{\theta}^*\in\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^{d_\theta}}\|\mathrm{A}\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}_\delta\|^2+\lambda\mathrm{TV}(\mathbf{f}(\boldsymbol{\theta})), \tag{3}$$

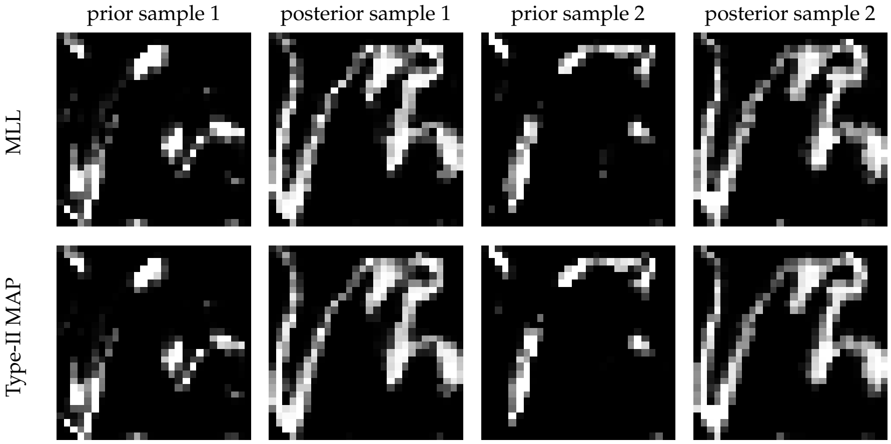

Figure 2: Samples from the linearised prior $\mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{f}})$ and posterior $\mathcal{N}(\mathbf{f}(\boldsymbol{\theta}^*), \Sigma_{\mathbf{f}|\mathbf{y}_\delta})$ for the image from fig. 1. The top row hyperparameters have been optimised with marginal likelihood (MLL) while the bottom with Type-II MAP using our proposed TV-PredCP prior. Although the difference is subtle, the latter are smoother and present less artefacts.

where $\lambda > 0$ balances the two terms. $\mathrm{TV}(\mathbf{f}(\boldsymbol{\theta}))$ denotes the anisotropic TV seminorm of the network output $\mathbf{f}(\boldsymbol{\theta})$

$$\mathrm{TV}(\mathbf{x}) = \sum_i \sum_j |X_{i,j} - X_{i+1,j}| + \sum_i \sum_j |X_{i,j} - X_{i,j+1}|,$$

where $X \in \mathbb{R}^{h \times w}$ denotes the image vector $\mathbf{x}$ reshaped into a matrix of height h by width w. TV has been widely used in image processing (Chambolle et al., 2010), and more recently also found to improve performance of DIP methods (Liu et al., 2019; Baguer et al., 2020). See Appendix C for a more in-depth discussion on TV as a regulariser.

3. Transforming the DIP into a Bayesian Prior over Images

We can interpret the loss in eq. (3) as a maximum a posteriori (MAP) objective, where we have imposed a probabilistic prior over neural functions $\mathbf{f}$ that favour smoothness in a TV sense

$$p(\mathbf{f}) \propto \exp(-\lambda \mathrm{TV}(\mathbf{f})), \quad \text{with} \quad \mathbf{f} \in \{\mathbf{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^{d_\theta}\}. \tag{4}$$

We combine this prior with sum of squares projection loss between $\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{y}_\delta$. The latter is obtained from a Gaussian conditional likelihood: $p(\mathbf{y}_\delta|\mathbf{f}) = \mathcal{N}(\mathbf{y}_\delta; \mathrm{A}\mathbf{f}, \sigma_y^2 \mathrm{I})$ with $\sigma_y^2{=}1$. Despite the space of functions providing a convenient canvas in which to express our prior beliefs, it significantly complicates subsequent probabilistic reasoning with NNs (Sun et al., 2019; Burt et al., 2020). Instead, we consider the following weight-space hierarchical model

$$\mathbf{y}_\delta|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{y}_\delta; \mathrm{A}\mathbf{f}(\boldsymbol{\theta}), \sigma_y^2 \mathrm{I}), \quad \boldsymbol{\theta}|\boldsymbol{\ell} \sim \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \Sigma_{\boldsymbol{\theta}}(\boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})), \quad \boldsymbol{\ell} \sim p(\boldsymbol{\ell}), \tag{5}$$

in which, building upon Fortuin et al. (2021), the spatial covariance $\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})$ among parameters is specified as

$$[\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})]_{ij,i'j'} = \mathrm{cov}(\boldsymbol{\theta}_{i,j}, \boldsymbol{\theta}_{i',j'}) = \begin{cases} \sigma_d^2 \exp\left(-\frac{\Delta(j,j')}{\ell_d}\right), & \text{if } i = i', \\ 0, & \text{else,} \end{cases} \tag{6}$$

where i indexes a convolutional filter and j the spatial location of a specific parameter within a filter. $\Delta(\cdot,\cdot)$ is the Euclidean distance between filter pixels. ℓ_d acts as a characteristic length-scale and σ_d^2 is the marginal prior variance. These two parameters are defined per convolutional block $d \in \{1, 2, \dots, D\}$ in the CNN such that $\boldsymbol{\ell} = [\ell_1, \ell_2, \dots, \ell_D]$ and $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}} = [\sigma_1^2, \sigma_2^2, \dots, \sigma_D^2]$. We do not place a prior over the latter, instead treating it as a hyperparameter. A diagram showing the correspondences between blocks $1, \dots, D$ and U-Net architectural components is included in appendix A.

By choosing a larger $\boldsymbol{\ell}$, we constrain the parameters in the convolutional filters and thus effectively enforce smoothness in the output. Thus, a prior placed over the filter length-scale $\boldsymbol{\ell}$ can act as a surrogate for the TV prior in eq. (4). To make this connection explicit, we construct a predictive complexity prior (PredCP) (Nalisnick et al., 2021) over $\boldsymbol{\ell}$:

$$p(\boldsymbol{\ell}) = p(\ell_1)p(\ell_2)\dots p(\ell_D) = \prod_{d=1}^{D} \pi(\kappa_d) \left| \frac{\partial \kappa_d}{\partial \ell_d} \right|, \quad \text{with} \tag{7}$$

$$\kappa_d := \mathbb{E}_{p(\boldsymbol{\theta}_d|\ell_d)\prod_{i=1, i\neq d}^{D} \delta(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)} \left[\lambda \mathrm{TV}(\mathbf{f}(\boldsymbol{\theta}))\right]. \tag{8}$$

Here κ_d is the expected smoothness (in a TV sense) of the network output $\mathbf{f}(\boldsymbol{\theta})$ induced by the prior over parameters of block d, when the parameters of all other blocks are set to some reference value $\boldsymbol{\theta}^*$. The random variable κ_d is modelled by an exponential distribution $\pi(\kappa_d) = \mathrm{Exp}(\kappa_d; 1)$ and is related to the filter lengthscale ℓ_d by means of the change of variable formula. The separable factorisation of $p(\boldsymbol{\ell})$ expresses our belief that blocks placed at different positions within the network will have different sized contributions to $\mathrm{TV}(\mathbf{f}(\boldsymbol{\theta}))$. The definition of κ_d ensures that we can measure the contribution of every block independently of the rest, ensuring the dimensionality match formally needed in the change of variables. We discuss the bijectivity of the mapping between ℓ_d and κ_d in Appendix E. In summary, by using the predictive complexity prior, we translate the image space smoothness prior in eq. (4) into a prior on the spatial smoothness of our convolutional filter parameters.

4. Linearised Laplace Inference in the Low-Dimensional Dual Space

Even when working in parameter space, inference in the model from eq. (5) is computationally intractable. We resort to locally linearising the U-Net around $\boldsymbol{\theta}^*$

$$\mathbf{h}(\boldsymbol{\theta}) := \mathbf{f}(\boldsymbol{\theta}^*) + \mathrm{J}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad \text{with} \quad \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \Sigma_{\boldsymbol{\theta}}(\boldsymbol{\ell}, \boldsymbol{\sigma^2_{\theta}})), \tag{9}$$

where $\mathrm{J} = \nabla_{\boldsymbol{\theta}} \mathbf{f}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \in \mathbb{R}^{d_x \times d_\theta}$ denotes the Jacobian matrix of the network with respect to $\boldsymbol{\theta}$. The covariance structure of this model matches that of a generalised linear model with

J as a design matrix. It can be written over image space as $\Sigma_{\mathbf{f}} = \mathrm{J}\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})\mathrm{J}^\top \in \mathbb{R}^{d_x \times d_x}$. The Jacobian term enforces the structure of CNN while the weight prior covariance enforces TV smoothness in eq. (7). We display draws from this model in fig. 2. For brevity, we hereon omit the dependence on kernel parameters $(\boldsymbol{\ell}, \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\theta}})$.

4.1. Efficient Posterior Predictive Computation

The linearised posterior over the image space is obtained in closed form using linear-Gaussian conjugacy $p(\mathbf{f}|\mathbf{y}_\delta) = \mathcal{N}(\mathbf{f}; \mu_{\mathbf{f}|\mathbf{y}_\delta}, \Sigma_{\mathbf{f}|\mathbf{y}_\delta}) \propto \mathcal{N}(\mathbf{y}_\delta; \mathrm{A}\mathbf{f}, \sigma_y^2 \mathrm{I})\mathcal{N}(\mathbf{f}; \mathbf{h}(\mathbf{0}), \Sigma_{\mathbf{f}})$. Its mean is $\mu_{\mathbf{f}|\mathbf{y}_\delta} = \mathbf{f}(\boldsymbol{\theta}^*)$, leaving the DIP's output unchanged. Its covariance is $\Sigma_{\mathbf{f}|\mathbf{y}_\delta} = (\Sigma_{\mathbf{f}}^{-1} + \sigma_y^{-2}\mathrm{A}^\top\mathrm{A})^{-1}$. Using the Woodbury matrix identity, we adopt the dual form

$$\Sigma_{\mathbf{f}|\mathbf{y}_\delta} = (\Sigma_{\mathbf{f}}^{-1} + \sigma_y^{-2}\mathrm{A}^\top\mathrm{A})^{-1} = \Sigma_{\mathbf{f}} - \Sigma_{\mathbf{f}}\mathrm{A}^\top(\mathrm{A}\Sigma_{\mathbf{f}}\mathrm{A}^\top + \sigma_y^2\mathrm{I})^{-1}\mathrm{A}\Sigma_{\mathbf{f}}^\top, \tag{10}$$

and obtain an expression in terms of the inverse of the observation space posterior covariance $\Sigma_y = \mathrm{A}\Sigma_{\mathbf{f}}\mathrm{A}^\top + \sigma_y^2\mathrm{I} \in \mathbb{R}^{d_y \times d_y}$. The computational complexity of eq. (10) scales as $\mathcal{O}(d_x d_y^2)$ as opposed to $\mathcal{O}(d_x^3)$ or $\mathcal{O}(d_\theta^3)$ for the image space or parameter space views, respectively.

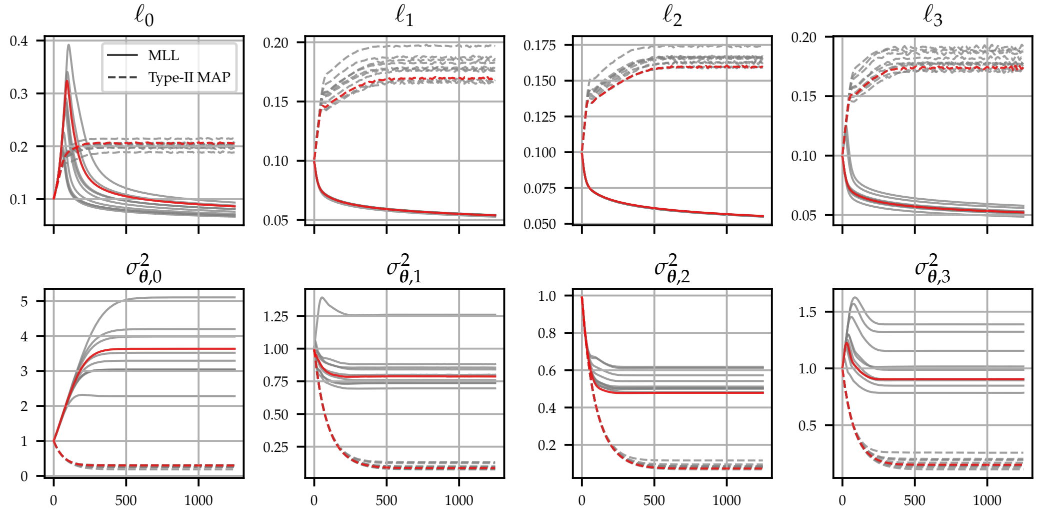

Figure 3: Optimisation of $(\boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})$ via MLL and Type-II MAP. Traces in red refer to the optimisation of the exemplary reconstruction shown in fig. 1. The TV-PredCP leads to larger prior lengthscales $\boldsymbol{\ell}$ and lower variances $\boldsymbol{\sigma}^2_\theta$.

4.2. Type-II MAP Learning of Hyperparamters with the TV-PredCP

Once a mode $\boldsymbol{\theta}^*$ of the posterior $p(\boldsymbol{\theta}|\mathbf{y}_\delta)$ has been found with the MAP criterion eq. (3), we select hyperparameters $(\sigma_y^2, \boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})$ that will provide well calibrated error-bars by maximising the volume of the posterior mode. This is done with the linearised Type-II MAP objective:

$$\begin{aligned}\mathcal{G}(\sigma_y^2, \boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) &= \log p(\mathbf{y}_\delta, \boldsymbol{\ell}; \sigma_y^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) = \log \mathcal{N}(\mathbf{y}_\delta; \mathbf{0}, \Sigma_y(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) + \sigma_y^2 \mathrm{I}) + \log p(\boldsymbol{\ell}; \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) \\ &= \frac{1}{2}\left(-\sigma_y^{-2} ||\mathbf{y}_\delta - \mathrm{A}\mathbf{f}(\boldsymbol{\theta}^*)||_2^2 - ||\boldsymbol{\theta}^*||^2_{\Sigma_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)} - \log|\Sigma_y(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) + \sigma_y^2 \mathrm{I}|\right) \\ &\qquad + \sum_{d=1}^{D} -\kappa_d(\ell_d, \sigma_d^2) + \log\left|\frac{\partial \kappa_d(\ell_d, \sigma_d^2)}{\partial \ell_d}\right| + C, \end{aligned} \tag{11}$$

where C is independent of our hyperparameters. See Appendix D.2 for a derivation of (11). We again employ the dual formulation to keep our update cost $\mathcal{O}(d_y^3)$. As shown in fig. 3, the PredCP acts not only on the convolutional filters' lengthscale $\boldsymbol{\ell}$, but also informs their prior marginal variances $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2$ which parametrise the bijection in eq. (7). Additionally, as described in Appendix D, the linearised model allows us to compute κ_d in closed form.

Table 1: Test log-likelihood on KMNIST averaged over 10 randomly chosen characters.

#directions (Sparse)	20		10	
noise (%)	5	10	5	10
DIP($\sigma_y^2 = 1$)	1.4188±0.1140	0.0007±0.2971	0.4562±0.1991	−2.6569±0.5695
DIP (MLL σ_y^2)	1.5118±0.1157	0.0552±0.2997	0.5566±0.2008	−2.5967±0.5716
Bayes DIP MLL	2.1782±0.0367	**2.0132±0.0411**	2.0407±0.0370	**1.7386±0.1032**
Bayes DIP TV-MAP	**2.2372±0.0371**	**2.0407±0.0430**	**2.1011±0.0449**	**1.7516±0.1139**

5. Experimental Evaluation

We reconstruct CT measurement data $\mathbf{y}_\delta$ simulated from the Kuzushiji-MNIST (KMNIST) dataset consisting of 28×28 grayscale images of Hiragana characters (Clanuwat et al., 2018). The linear forward map A is given by the discrete Radon transform. For each simulated measurement $\mathbf{y}_\delta$, we train $\mathbf{f}(\boldsymbol{\theta})$ via (3), fixing the input to a coarse FBP reconstruction, for 25k iterations. We then optimise the hyperparameters $(\sigma_y^2, \boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)$ with Type-II MAP using (11) for 1.25k iterations. We also consider omitting the PredCP in (11), reverting to the standard marginal likelihood (MLL).

Qualitative Results Figure 1 shows an exemplary character which we recover with both FBP and DIP. The reconstruction problem is highly ill-posed (10 directions), thus FBP reconstructions exhibit strong artefacts. As expected, the DIP reconstruction obtains 12dB higher peak noise-to-signal ratio (psnr). The pixel-wise standard deviation provided by our method correlates strongly with the DIP reconstruction error. However, the hyperparameters found via MLL lead to excessively large uncertainty. TV-PredCP regularisation leads to hyperparameters that produce better-calibrated uncertainty. In fig. 3 we show how the TV-PredCP drives $\boldsymbol{\ell}$ to larger values and $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2$ to be smaller. The use of the PredCP restricts our prior, and thus posterior, to functions that are smooth in a TV sense, resulting in smaller error-bars. This can also be seen in fig. 2, where samples from the TV-PredCP model are smoother and present less artefacts. See Appendix A for additional experimental figures.

Quantitative Results We consider 20 and 10 directions taken uniformly from 0° to 180°. We add 5% (low-noise) and 10% (high-noise) Gaussian noise to $\mathrm{A}\mathbf{x}$. We conduct each experiment using the same 10 randomly chosen KMNIST test images. We find some images to include spurious large valued pixels far away from the region of interest, violating the modelling assumption that $\mathbf{x}$ is noiseless eq. (1). DIP treats these pixels as noise, does not reconstruct them and the hyperparameter optimisation eq. (11) converges to larger values of σ_y^2. We thus add the correction term $\sigma_y^2(\mathrm{A}^\top \mathrm{A})^\dagger$ to our predictive variance and provide further discussion in appendix D.1. We study the standalone effect of this correction by treating it as a baseline, labelled "MLL σ_y^2". We also consider an isotropic unit variance noise model. Table 1 show test log-density, obtained by method (predictive posterior given in section 4) and baselines, for all experimental settings. All methods share the same network parameters and thus the same mean reconstruction. Consequently, higher values in log-density indicate better uncertainty calibration. In all settings, capturing model uncertainty significantly improves performance. There is a clear but smaller benefit in employing the TV-PredCP prior which is most notable in the low-noise setting.

6. Limitations and Conclusion

We have demonstrated how a probabilistic formulation of the DIP can yield well-calibrated uncertainty in a computational tomography setting. Our results suggest that the TV, apart from acting as a regulariser, can yield a performant probabilistic prior over images for the inverse problem. Open questions are whether performance can be improved by giving a fully probabilistic treatment to $\boldsymbol{\ell}$ and whether the strength of the TV can be chosen with an MLL objective. The KMNIST setting under consideration is quite small. We will investigate how to scale our method to larger output spaces, where Jacobian computation might be intractable and the observation dimension d_y may still be large.

Acknowledgments

The authors would like to thank Marine Schimel, Alexander Terenin, Eric Nalisnick, Erik Daxberger and James Allingham for the fruitful discussions. We also thank Johannes Leuschner for his help in the preparation of the manuscript. R.B. acknowledges support from the i4health PhD studentship (UK EPSRC EP/S021930/1), and from The Alan Turing Institute (UK EPSRC EP/N510129/1). The work of BJ is partially supported by UK EPSRC grant EP/T000864/1. JA acknowledges support from Microsoft Research, through its PhD Scholarship Programme, and from the EPSRC. This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (http://www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant.

References

Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on*

Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/781877bda0783aac5f1cf765c128b437-Abstract.html.

Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A method for explaining uncertainty estimates. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XSLF1XFq5h.

Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9):094004, 2020.

Riccardo Barbano, Johannes Leuschner, Maximilian Schmidt, Alexander Denker, Andreas Hauptmann, Peter Maaß, and Bangti Jin. Is deep image prior in need of a good education? *arXiv preprint arXiv:2111.11926*, 2021.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan, editors, *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 401–413. ACM, 2021. doi: 10.1145/3461702.3462571. URL https://doi.org/10.1145/3461702.3462571.

David R. Burt, Sebastian W. Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. *CoRR*, abs/2011.09421, 2020. URL https://arxiv.org/abs/2011.09421.

Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. In *Theoretical foundations and numerical methods for sparse recovery*, pages 263–340. de Gruyter, 2010.

Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A bayesian perspective on the deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5443–5451. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00559. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Cheng_A_Bayesian_Perspective_on_the_Deep_Image_Prior_CVPR_2019_paper.html.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Erik A. Daxberger, Eric T. Nalisnick, James Urquhart Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In

Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2510–2521. PMLR, 2021. URL http://proceedings.mlr.press/v139/daxberger21a.html.

Dan E Dudgeon and Russell M Mersereau. *Multidimensional digital signal processing.* Prentice-hall, 1984.

HW Engl, M Hanke, and A Neubauer. Regularization of inverse problems, ser. *Mathematics and its Applications. Dordrecht: Kluwer Academic Publishers Group*, 375, 1996.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Ratsch, Richard E Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL https://openreview.net/forum?id=xaqKWHcoOGP.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 703–711. PMLR, 2021. URL http://proceedings.mlr.press/v130/immer21a.html.

Kazufumi Ito and Bangti Jin. *Inverse problems: Tikhonov theory and algorithms*, volume 22. World Scientific, 2014.

Max-Heinrich Laves, Malte Tölle, and Tobias Ortmaier. Uncertainty estimation in medical image denoising with bayesian deep image prior. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 81–96. Springer, 2020.

Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S Kamilov. Image restoration using total variation regularized deep image prior. In *ICASSP 2019*, 2019. doi: 10.1109/ICASSP.2019.8682856.

David John Cameron Mackay. *Bayesian Methods for Adaptive Models.* PhD thesis, USA, 1992. UMI Order No. GAX92-32200.

Eric T. Nalisnick, Jonathan Gordon, and José Miguel Hernández-Lobato. Predictive complexity priors. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 694–702. PMLR, 2021. URL http://proceedings.mlr.press/v130/nalisnick21a.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Matthias W. Seeger and Hannes Nickisch. Large scale bayesian inference and experimental design for sparse linear models. *SIAM J. Imaging Sci.*, 4(1):166–199, 2011. doi: 10.1137/090758775. URL https://doi.org/10.1137/090758775.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13969–13980, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger B. Grosse. Functional variational bayesian neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rkxacs0qY7.

Andrey N Tikhonov and Vasiliy Y Arsenin. Solutions of ill-posed problems. *New York*, 1 (30):487, 1977.

Malte Tölle, Max-Heinrich Laves, and Alexander Schlaefer. A mean-field variational inference approach to deep image prior for inverse problems in medical imaging. In Mattias P. Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schlaefer, and Floris Ernst, editors, *Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany*, volume 143 of *Proceedings of Machine Learning Research*, pages 745–760. PMLR, 2021. URL https://proceedings.mlr.press/v143/tolle21a.html.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9446–9454, 2018.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *Int. J. Comput. Vis.*, 128(7):1867–1888, 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01303-4.

Appendix A. Additional Experimental Details

In our experiments we use a down-sized version of U-Net. The reduced output dimension d_x, and the simplicity of the reconstruction problem allow us to adopt a shallow architecture without compromising on reconstruction quality. Figure 4 shows the network architecture used for this work. We adopt the U-Net architecture proposed by Barbano et al. (2021), with the only difference being that we remove group-normalisation layers. Each light-blue box corresponds to a multi-channel feature map. The number of channels is set to 32 at every scale. The arrows denote the different operations. We give a Bayesian treatment to all parameters in convolutional layers (approx. 78k parameters). We denote with $*$ the operations that do not receive a Bayesian treatment (1x1 convolution mixing layers). We identify 4 blocks: Down_0 ($d = 0$), Down_1 ($d = 1$), Up_0 ($d = 2$) and Up_1 ($d = 3$). The Down blocks consists of a 3×3 convolution followed by a 3×3 convolution operation; the Up blocks instead consists of two successive 3×3 convolutional operations.

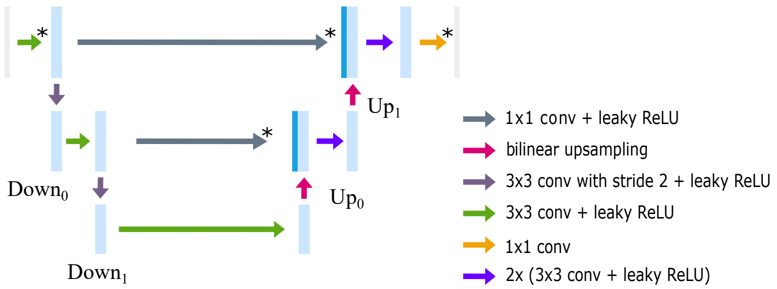

Figure 4: U-Net architecture diagram.

A.1. Additional Experimental Figures

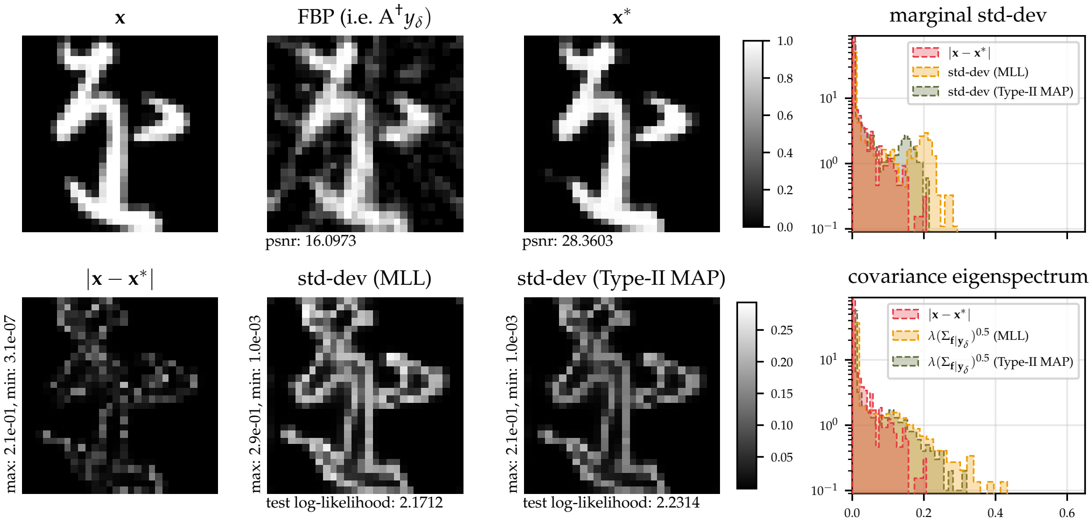

Figure 5: Exemplary reconstruction of a Hiragana character.

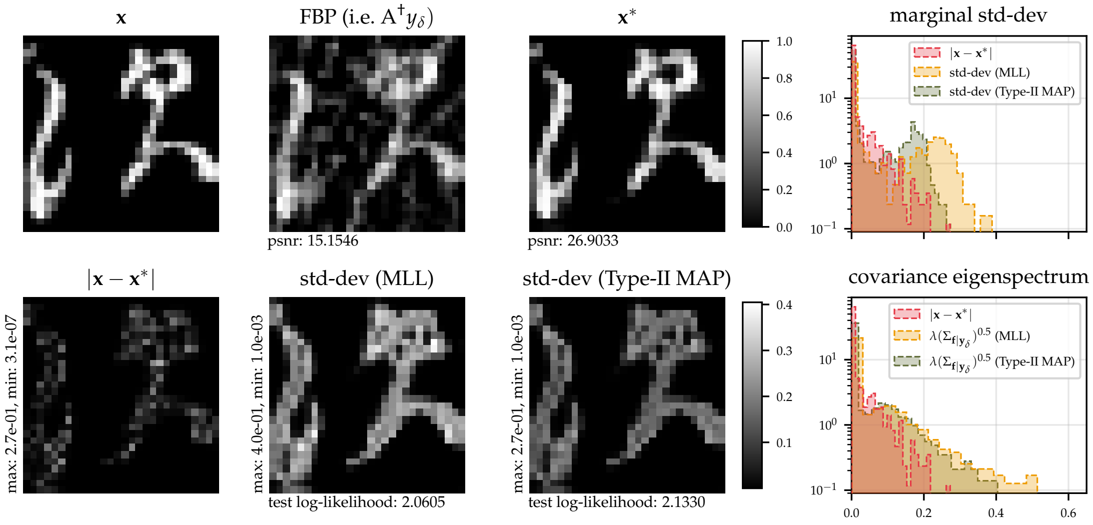

Figure 6: Exemplary reconstruction of a Hiragana character.

Figure 5 and fig. 6 show additional reconstructions of two exemplary characters taken from the KMNIST dataset. We also include additional samples from both the linearised prior $\mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{f}})$ and posterior $\mathcal{N}(\mathbf{f}(\boldsymbol{\theta}^*), \Sigma_{\mathbf{f}|\mathbf{y}_\delta})$ for the image from fig. 1 in fig. 7 and fig. 8, respectively. A careful visual inspection suggests that the sample obtained using the hyperparameters optimised with Type-II MAP presents higher degree of smoothness, along with less artefacts, as one may expect from the construction of the PredCP prior.

Appendix B. Discussion on Previous Work on Bayesian DIP

Cheng et al. (2019) perform stochastic gradient Langevin dynamics (SGLD) inference with an isotropic Gaussian prior over the network parameters. Laves et al. (2020); Tölle et al. (2021) use both Monte Carlo (MC) dropout inference, and then Gaussian mean-field variational inference. These works resorted to a Bayesian treatment of DIP as a way to alleviate the need for early-stopping. However, Baguer et al. (2020) show that the TV regulariser is a simpler yet more effective approach.

Our proposed probabilistic method aims at producing prior and predictive distributions in the image space, which is in its goal essentially different from existing probabilistic treatment of DIP (Cheng et al., 2019; Tölle et al., 2021). Indeed, Dropout and SGLD have been found to not be very robust methods for uncertainty estimation (Snoek et al., 2019; Antorán et al., 2020; Daxberger et al., 2021). Furthermore, the previous work is sampling based, and does not place an explicit likelihood function over the image space.

To the best of our knowledge there is little existing work on the application of predictive uncertainty to deep tomographic reconstructions, with the most notable being (Seeger and Nickisch, 2011). However, there is a rapidly growing body of literature on the utility of uncertainty for human computer interaction (Antorán et al., 2021; Bhatt et al., 2021). With our work we hope to take steps towards building a foundation for more research in this area.

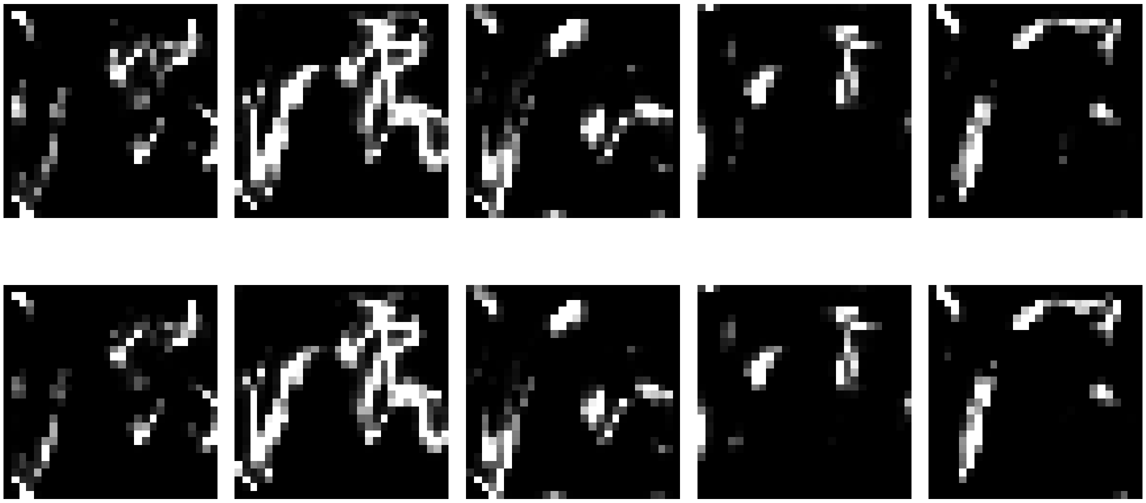

Figure 7: Drawing sample from the linearised prior over images. Each sample is drawn from a multivariate Gaussian with zero mean and prior covariance $\Sigma_{\mathbf{f}}$. Samples in the top row are obtained with hyperparameters optimised with MLL. Samples in the bottom row, instead, are collected using the hyperparameters optimised with Type-II MAP.

Appendix C. Discussion on the TV Regulariser as a Prior

We may question whether TV is the ideal regulariser for the KMIST dataset. The TV regulariser enforces sparsity in the overall image gradients, which in turn induces smoothness in the image. A TV regulariser is highly recommended when we observe sparsity in the edges present in an image, especially when the edges are a small fraction of the overall image pixels. That is often the case in high-resolution medical images. Intuitively, the higher the resolution of the image is, the higher the sparsity level of the edges is. However, in the KMIST dataset, due to the low resolution of the images, the edges constitute a considerable fraction of the total pixels. Therefore, a TV regulariser could be sub-optimal. In KMNIST, it is difficult to clearly distinguish (in TV sense) what is part of the image structure, what is part of the background. The stroke is only a few pixels wide, and ground-truth pixel values are generated through interpolation (Clanuwat et al., 2018). In future work, we plan to compare different regularisers, such as an L^1 norm.

It is tempting to think that we do not need the PredCP machinery to translate the TV regulariser into the weight space. Indeed, the Laplace approximation simply involves a quadratic approximation around a mode of the log posterior, without placing any requirements on the prior used to induce said posterior. Along this line of reasoning, we can decompose the Hessian of the log posterior $\log p(\boldsymbol{\theta}|\mathbf{y}_\delta)$ into the contributions from the

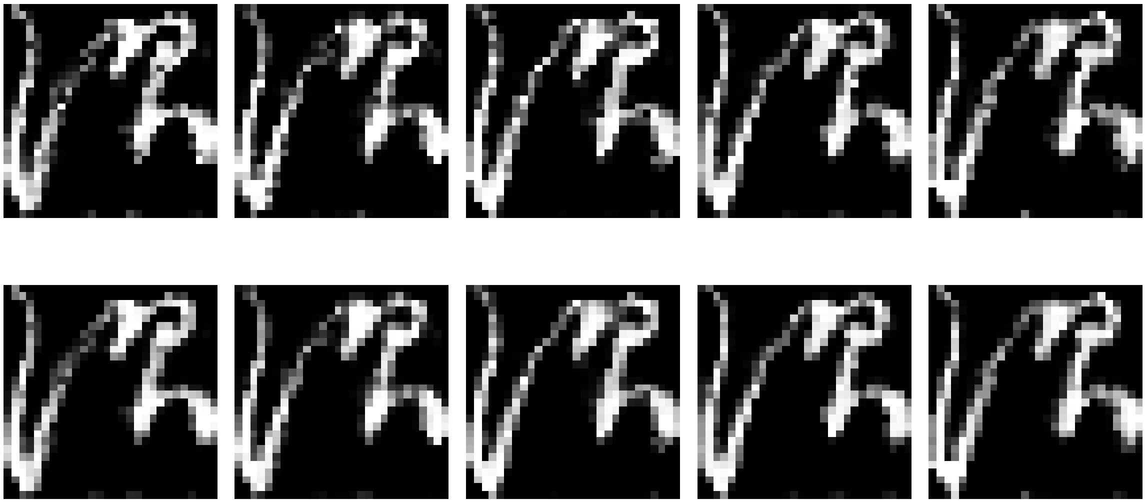

Figure 8: Drawing sample from the linearised posterior over images. Each sample is drawn from a multivariate Gaussian with mean $\mu_{\mathbf{f}|\mathbf{y}_\delta}$ and posterior covariance $\Sigma_{\mathbf{f}|\mathbf{y}_\delta}$. Samples in the top row are obtained with hyperparameters optimised with MLL. Samples in the bottom row are, instead, collected using the hyperparameters optimised with Type-II MAP.

likelihood and the prior as

$$\nabla^2_{\boldsymbol{\theta}} \left(\log p(\mathbf{y}_\delta | \mathrm{A}\mathbf{f}(\boldsymbol{\theta})) + \log p(\mathbf{f}(\boldsymbol{\theta})) \right) |_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

and quickly realise that the log of the anisotropic TV prior chosen to be $p(\mathbf{f}) \propto \exp(-\lambda \mathrm{TV}(\mathbf{f}))$ as in eq. (4) is only once differentiable. Ignoring the the origin (where the absolute value function is non-differentiable), we obtain:

$$\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{f}(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \propto -\nabla^2_{\boldsymbol{\theta}} \mathrm{TV}(\mathbf{f}(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = 0.$$

Thus, a naive application of the Laplace approximation would eliminate the effect of the prior, leaving the posterior ill defined.

Appendix D. Derivation of Method Proposed in Section 4

D.1. Posterior Predictive Covariance

We provide an alternative derivation for the posterior predictive covariance, where the probabilistic reasoning is performed in the weight space. We start from the model introduced in eq. (5):

$$\mathbf{y}_\delta | \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{y}_\delta; \mathrm{A}\mathbf{f}(\boldsymbol{\theta}), \sigma^2_y \mathrm{I}), \quad \boldsymbol{\theta} | \boldsymbol{\ell} \sim \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \Sigma_{\boldsymbol{\theta}})$$

and write down the linearized Laplace approximate posterior distribution over weights:

$$p(\boldsymbol{\theta}|\mathbf{y}_\delta) \approx \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, \Sigma_{\boldsymbol{\theta}|\mathbf{y}_\delta}) \quad \text{with} \quad \Sigma_{\boldsymbol{\theta}|\mathbf{y}_\delta} = \left(\sigma_y^{-2}\mathrm{A}^\top \mathrm{J}^\top \mathrm{J}\mathrm{A} + \Sigma_{\boldsymbol{\theta}}^{-1}\right)^{-1}. \tag{12}$$

This expression is obtained by substituting the NN $\mathbf{f}(\boldsymbol{\theta})$ with the linearised surrogate $\mathbf{h}$ and computing the product of Gaussian prior and posterior. We refer to (Mackay, 1992; Immer et al., 2021) for a derivation. Then we rewrite the above expression using the Woodbury matrix inversion identity:

$$\Sigma_{\boldsymbol{\theta}|\mathbf{y}_\delta} = \left(\sigma_y^{-2}\mathrm{A}^\top \mathrm{J}^\top \mathrm{J}\mathrm{A} + \Sigma_{\boldsymbol{\theta}}^{-1}\right)^{-1} = \Sigma_{\boldsymbol{\theta}} - \Sigma_{\boldsymbol{\theta}}\mathrm{J}^\top \mathrm{A}^\top(\sigma_y^2 \mathrm{I} + \mathrm{A}\mathrm{J}\Sigma_{\boldsymbol{\theta}}^{-1}\mathrm{J}^\top \mathrm{A}^\top)^{-1}\mathrm{A}\mathrm{J}\Sigma_{\boldsymbol{\theta}}^\top \tag{13}$$

The predictive distribution over images can be built by marginalising the NN parameters in the conditional likelihood $p(\mathbf{x}|\mathbf{y}_\delta) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_\delta)\, d\boldsymbol{\theta}$.

As mentioned in section 5, KMNIST presents spurious high valued pixels away from the region containing the handwritten character. This presents a problem due to our modelling assumptions in eq. (1), assuming $\mathbf{x}$ is noiseless, and thus our likelihood function from eq. (5) being defined over the space of observations $\mathbf{y}_\delta$. We translate the uncertainty induced by the observation noise to the space of images by computing the conditional log-likelihood Hessian with respect to $\mathbf{f}$: $-\nabla^2_{\mathbf{f}} \log p(\mathbf{y}_\delta|\mathbf{f}) = \sigma_y^{-2}\mathrm{A}^\top\mathrm{A} \in \mathbb{R}^{d_x \times d_x}$. This matrix is of rank at most d_y, which can be much smaller than d_x due to the ill-conditioning of the reconstruction problem, and therefore cannot act as a proper Gaussian precision matrix on its own. However, we incorporate the noise uncertainty from the observation subspace into the image space $\mathbf{x}$ by adding the pseudoinverse $\sigma_y^2(\mathrm{A}^\top\mathrm{A})^\dagger$ to the predictive covariance. We move forward with weight marginalisation when using the linear model to recover the predictive distribution:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}_\delta) = &\int \mathcal{N}(\mathbf{x}; \mathbf{f}(\boldsymbol{\theta}^*) + \mathrm{J}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \sigma_y^2(\mathrm{A}^\top\mathrm{A})^\dagger)\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, \Sigma_{\boldsymbol{\theta}|\mathbf{y}_\delta})\, d\boldsymbol{\theta} \\ =&\mathcal{N}(\mathbf{x}; \mathbf{f}(\boldsymbol{\theta}^*), \mathrm{J}\Sigma_{\boldsymbol{\theta}|\mathbf{y}_\delta}\mathrm{J}^\top + \sigma_y^2(\mathrm{A}^\top\mathrm{A})^\dagger) \end{aligned}$$

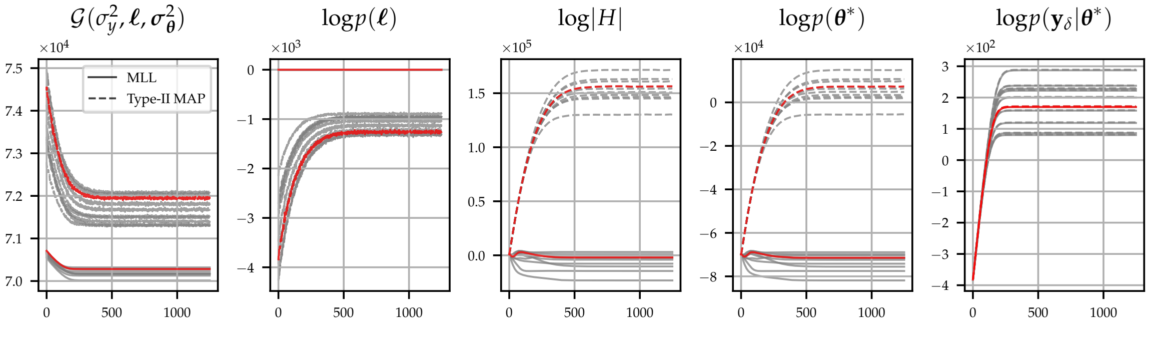

Figure 9: Evolution of standard marginal likelihood (MLL) and Type-II MAP during optimisation along with their individual components. The log TV-PredCP $\log p(\boldsymbol{\ell})$, the log-determinant of the posterior Hessian $\log|H|$, the weight-mode log prior density $\log p(\boldsymbol{\theta}^*)$, and the observation conditional log-density $\log p(\mathbf{y}_\delta|\boldsymbol{\theta}^*)$. Traces in red refer to the optimisation of the exemplary reconstruction shown in fig. 1.

D.2. Laplace Marginal Likelihood and Type-II MAP eq. (11)

The Laplace method provides an estimate of the model evidence, also known as marginal likelihood (MLL) or Type-II maximum likelihood, based on a quadratic approximation to the volume of a mode $\boldsymbol{\theta}^*$ of the log joint $\log p(\mathbf{y}_\delta, \boldsymbol{\theta})$ (Mackay, 1992):

$$\log p(\mathbf{y}_\delta) = \log \int p(\mathbf{y}_\delta, \boldsymbol{\theta})\, d\boldsymbol{\theta} \approx \log \left(p(\mathbf{y}_\delta, \boldsymbol{\theta}^*)(2\pi)^{d_\theta/2} | - \nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}_\delta, \boldsymbol{\theta})_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}|^{-\frac{1}{2}} \right), \tag{14}$$

where the Hessian determinant of the joint log-density with respect to the model parameters $H = -\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}_\delta, \boldsymbol{\theta})_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ captures the width of the mode $\boldsymbol{\theta}^*$ in question. Expanding the above expression we have:

$$\log p(\mathbf{y}_\delta) \approx \log p(\mathbf{y}_\delta|\boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \frac{1}{2}\log |H| + \frac{d_\theta}{2}\log(2\pi). \tag{15}$$

Note that the observation conditional log-density $\log p(\mathbf{y}_\delta|\boldsymbol{\theta}^*)$ and the weight-mode log prior density $\log p(\boldsymbol{\theta}^*)$ are given respectively by

$$\begin{aligned} \log p(\mathbf{y}_\delta|\boldsymbol{\theta}^*) &= -\frac{d_y}{2}\log(2\pi) - \frac{1}{2}\log|\sigma_y^2 \mathrm{I}| - \frac{1}{2\sigma_y^2}||\mathbf{y}_\delta - \mathrm{A}\mathbf{f}(\boldsymbol{\theta}^*)||_2^2, \\ \log p(\boldsymbol{\theta}^*) &= -\frac{d_\theta}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\boldsymbol{\theta}}| - \frac{1}{2}||\boldsymbol{\theta}^*||^2_{\Sigma^{-1}_{\boldsymbol{\theta}}}. \end{aligned}$$

Accordingly, the Hessian term H can be decomposed into

$$H = -\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}_\delta, \boldsymbol{\theta})_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = -\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}_\delta|\boldsymbol{\theta}^*) - \nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}^*).$$

Under the linearised approximation (i.e., substituting $\mathbf{f}(\boldsymbol{\theta})$ with its first-order Taylor expansion $\mathbf{h} = \mathbf{f}(\boldsymbol{\theta}^*) + \mathrm{J}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$), we obtain:

$$-\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}_\delta|\boldsymbol{\theta}^*) = \nabla^2_{\boldsymbol{\theta}} \frac{1}{2\sigma_y^2}||\mathbf{y}_\delta - \mathrm{A}(\mathbf{f}(\boldsymbol{\theta}^*) + \mathrm{J}(\boldsymbol{\theta} - \boldsymbol{\theta}^*))||_2^2 = \frac{1}{\sigma_y^2}\mathrm{J}^\top \mathrm{A}^\top \mathrm{A}\mathrm{J}.$$

Trivially, we have $-\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}^*) = \Sigma^{-1}_{\boldsymbol{\theta}}$. By the matrix determinant lemma, the determinant of H is given by

$$|H| = |\frac{1}{\sigma_y^2}\mathrm{J}^\top \mathrm{A}^\top \mathrm{A}\mathrm{J} + \Sigma^{-1}_{\boldsymbol{\theta}}| = |\mathrm{A}\mathrm{J}\Sigma_{\boldsymbol{\theta}}\mathrm{J}^\top \mathrm{A}^\top + \sigma_y^2 \mathrm{I}||\Sigma^{-1}_{\boldsymbol{\theta}}||\frac{1}{\sigma_y^2}\mathrm{I}|. \tag{16}$$

Finally, recall from eq. (6) that $\Sigma_{\boldsymbol{\theta}}$ depends on hyperparameters $(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})$, which is explicitly indicated for the purpose of derivation. Thus, the expanded linearised Laplace model evidence is given by

$$\begin{aligned} &\log p(\mathbf{y}_\delta; \sigma_y^2, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell}) \\ &= -\frac{1}{2}\log|\sigma_y^2 \mathrm{I}| - \frac{1}{2\sigma_y^2}||\mathbf{y}_\delta - \mathrm{A}\mathbf{f}(\boldsymbol{\theta}^*)||_2^2 - \frac{1}{2}\log|\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})| - \frac{1}{2}||\boldsymbol{\theta}^*||^2_{\Sigma^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})} \\ &\quad - \frac{1}{2}\log|\mathrm{A}\mathrm{J}\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})\mathrm{J}^\top \mathrm{A}^\top + \sigma_y^2 \mathrm{I}| + \frac{1}{2}\log|\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})| + \frac{1}{2}\log|\sigma_y^2 \mathrm{I}| + C \\ &= -\frac{1}{2\sigma_y^2}||\mathbf{y}_\delta - \mathrm{A}\mathbf{f}(\boldsymbol{\theta}^*)||_2^2 - \frac{1}{2}||\boldsymbol{\theta}^*||^2_{\Sigma^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})} - \frac{1}{2}\log|\mathrm{A}\mathrm{J}\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}, \boldsymbol{\ell})\mathrm{J}^\top \mathrm{A}^\top + \sigma_y^2 \mathrm{I}| + C \end{aligned} \tag{17}$$

where C captures all terms constant with respect to $(\sigma_y^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\ell})$. Recall that $\Sigma_y(\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\ell}) = \mathrm{A}\mathrm{J}\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\ell})\mathrm{J}^\top\mathrm{A}^\top$. Next we turn to the Pred-CP TV prior over $\boldsymbol{\ell}$:

$$\log p(\boldsymbol{\ell}; \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) = \sum_{d=1}^{D} -\kappa_d + \log\left|\frac{\partial \kappa_d}{\partial \ell_d}\right|, \text{ with } \kappa_d := \mathbb{E}_{p(\boldsymbol{\theta}_d|\ell_d;\sigma_d^2)\prod_{i=1,i\neq d}^{D}\delta(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)}\left[\lambda \mathrm{TV}(\mathbf{f}(\boldsymbol{\theta}))\right].$$

Combining this with eq. (17), we obtain a Type-II maximum a posteriori (MAP) objective from eq. (3):

$$\begin{aligned}\mathcal{G}(\sigma_y^2, \boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) =& \log p(\mathbf{y}_\delta, \boldsymbol{\ell}; \sigma_y^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) = \log \mathcal{N}(\mathbf{y}_\delta; \mathbf{0}, \Sigma_y(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) + \sigma_y^2 \mathrm{I}) + \log p(\boldsymbol{\ell}; \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) \\ =& \frac{1}{2}\left(-\sigma_y^{-2}||\mathbf{y}_\delta - \mathrm{A}\mathbf{f}(\boldsymbol{\theta}^*)||_2^2 - ||\boldsymbol{\theta}^*||^2_{\Sigma_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)} - \log|\Sigma_y(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) + \sigma_y^2 \mathrm{I}|\right) \\ &+ \sum_{d=1}^{D} -\kappa_d(\ell_d, \sigma_d^2) + \log\left|\frac{\partial \kappa_d(\ell_d, \sigma_d^2)}{\partial \ell_d}\right| + C.\end{aligned}$$

Figure 9 shows the evolution of standard marginal likelihood (MLL) and Type-II MAP during optimisation along with their individual components. The TV-PredCP introduces additional constraints into the model by encouraging the prior to contract (stronger parameter correlations and smaller marginal variances as shown in fig. 3). This prior concentrates a higher density on $\boldsymbol{\theta}^*$. In turn, this results in a more contracted posterior which we observe as a larger Hessian determinant. The conditional data density is unaffected by the TV-PredCP.

Appendix E. Monotonicity of the TV in the Prior Lengthscales

In order to apply the change of variables formula in eq. (7), we require bijectivity in the relationship between ℓ_d and $\kappa_d := \mathbb{E}_{p(\boldsymbol{\theta}_d|\ell_d)\prod_{i=1,i\neq d}^{D}\delta(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)}\left[\mathrm{TV}(\mathbf{f}(\boldsymbol{\theta}))\right]$. In the current setting, both variables are one-dimensional, making this constraint easier to satisfy. In fact, it suffices to show monotonicity between the two. We estimate the relationship between this variable pair empirically for every convolutional block in the U-net using Monte Carlo.

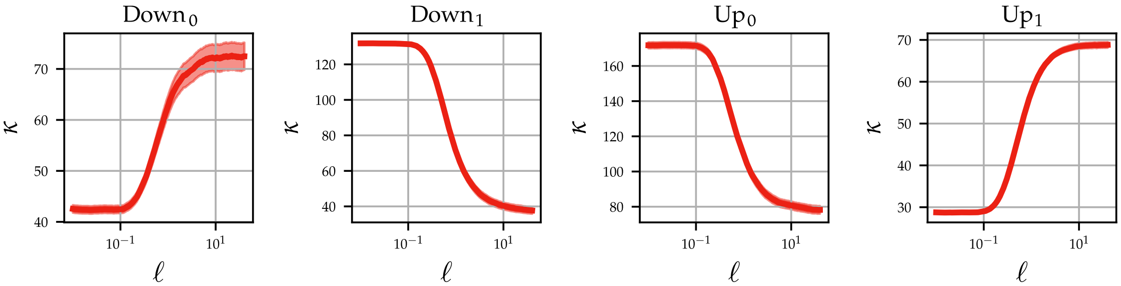

Figure 10: Experimental evidence of monotonicity in the non-linearised network. κ is estimated with 1000 Monte Carlo samples.

The results are displayed in fig. 10. In practise, we work with the linearised model in eq. (9) for the purposes of inference. In fig. 11, we show very compelling numerical evidence for the monotonicity. We observe that κ increases in $\boldsymbol{\ell}$ since large values for $\boldsymbol{\ell}$ lead to an increased marginal variance $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}$ over images. Just as expected, after fixing the marginal variance to 1, we observe that the lengthscales have a monotonically decreasing relationship with the expected TV.

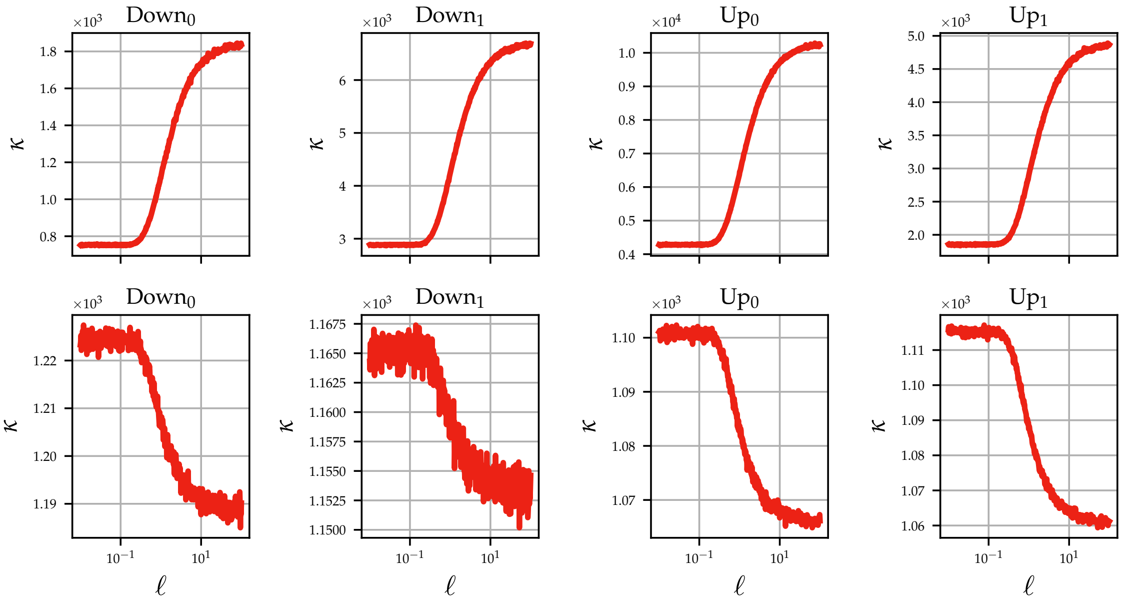

Figure 11: Experimental evidence of monotonicity for the linearised network. κ is estimated with 500 Monte Carlo samples. In the bottom row we fix the marginal variance in image space to be 1. This allows us to observe the smoothing effect from $\boldsymbol{\ell}$.

However, analytically studying the monotonicity can be a delicate matter. A simple analysis follows on the bijectivity of the mapping in the linear setting, which is of great interest as it matches our experimental setup.

$$\begin{aligned}\kappa_d =&\mathbb{E}_{p(\boldsymbol{\theta}_d|\ell_d)\prod_{j=1,j\neq d}^{D}\delta(\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*)}[\mathrm{TV}(\mathbf{h}(\boldsymbol{\theta}))]\\ =&\mathbb{E}_{p(\boldsymbol{\theta}_d|\ell_d)\prod_{j=1,j\neq d}^{D}\delta(\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*)}\Big[\sum_i|\mathbf{h}(\boldsymbol{\theta})_i-\mathbf{h}(\boldsymbol{\theta})_{i+1}|\Big],\end{aligned} \tag{18}$$

assuming that the output is a one-dimensional signal so there is only one derivative to simplify the discussion. First we interpret the distribution of $\mathbf{h}(\boldsymbol{\theta})_i - \mathbf{h}(\boldsymbol{\theta})_{i+1}$. Note that $\mathbf{h}(\boldsymbol{\theta})$ can be written as $\mathbf{h}(\boldsymbol{\theta}) = C + \mathrm{J}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, by slightly abusing the notation C to denote the terms constant with respect to ℓ_d and i indexes an entry of $(\mathrm{J}\boldsymbol{\theta}) \in \mathbb{R}^{d_x}$. Thus, we can rewrite this expression as an inner product between two vectors:

$$\mathbf{h}(\boldsymbol{\theta})_i - \mathbf{h}(\boldsymbol{\theta})_{i+1} = (\mathrm{J}\boldsymbol{\theta})_i - (\mathrm{J}\boldsymbol{\theta})_{i+1} = (\mathrm{J}_i - \mathrm{J}_{i+1})\boldsymbol{\theta}_d = \mathbf{v}_i\boldsymbol{\theta}_d,$$

where $\mathrm{J}_i \in \mathbb{R}^{1\times d_{\theta_d}}$ denotes our NN's Jacobian for a single output pixel i (i.e., the ith row of the Jacoian matrix J, corresponding to the block parameters $\boldsymbol{\theta}_d$, which has a length d_{θ_d})

and $\mathbf{v}_i = \mathrm{J}_i - \mathrm{J}_{i+1} \in \mathbb{R}^{1\times d_{\theta_d}}$, $i = 1, \ldots, d_x - 1$. Now, recall that the block parameters $\boldsymbol{\theta}_d$ is distributed as

$$\boldsymbol{\theta}_d \sim \mathcal{N}(\boldsymbol{\theta}_d; \mathbf{0}, \Sigma_d(\ell_d, \sigma_d^2)),$$

in the expectation in (18), whereas the remaining parameters are fixed at the mode $\boldsymbol{\theta}_j^*$, $j \neq d$, i.e., $\prod_{j=1, j\neq d}^{D} \delta(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)$. Let $\mathbf{V}_d \in \mathbb{R}^{(d_x-1)\times d_{\theta_d}}$ correspond to the stacking of the vectors $\mathbf{v}_i \in \mathbb{R}^{1\times d_{\theta_d}}$, i.e., the Jacobian of the network output with respect to the weights in convolutional group d. Since the affine transformation of a Gaussian distribution remains Gaussian, $\mathbf{V}_d\boldsymbol{\theta}_d$ is distributed according to

$$\mathbf{V}_d\boldsymbol{\theta}_d \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_d\Sigma_d\mathbf{V}_d^\top).$$

Note that the matrix $\mathbf{V}_d\Sigma_d\mathbf{V}_d^\top$ is not necessarily invertible, and if not, as usual, the inverse covariance should be interpreted in the sense of pseudo-inverse. Let $\mathbf{a} =: \mathbf{V}_d\boldsymbol{\theta}_d \in \mathbb{R}^{d_x-1}$. Then we can rewrite our quantity of interest as

$$\kappa_d = \mathbb{E}_{\mathbf{a}\sim\mathcal{N}(\mathbf{0}, \mathbf{V}_d\Sigma_d\mathbf{V}_d^\top)}\Big[\sum_i |a_i|\Big] = \sum_i \mathbb{E}_{a_i\sim\mathcal{N}(\mathbf{0}, \mathbf{v}_i\Sigma_d\mathbf{v}_i^\top)}[|a_i|]. \tag{19}$$

The distribution of $|a_i|$ can be expressed in terms of the truncated normal distributions. It is known that the expectation increases monotonically with the variance $\mathbf{v}_i\Sigma_d\mathbf{v}_i^\top$. It remains to examine the monotonicity of $\mathbf{v}_i\Sigma_d\mathbf{v}_i^\top$ in ℓ_d. Indeed, by the definition of Σ_d, we have

$$\frac{\partial}{\partial \ell_d}\mathbf{v}_i\Sigma_d\mathbf{v}_i^\top = \mathbf{v}_i\frac{\partial}{\partial \ell_d}\Sigma_d\mathbf{v}_i^\top.$$

Direct computation gives

$$\frac{\partial}{\partial \ell_d}[\Sigma_d(\ell_d)]_{j,j'} = \frac{\partial}{\partial \ell_d}\sigma_d^2 \exp\Big(-\frac{\Delta(j,j')}{\ell_d}\Big) = \frac{\sigma_d^2\Delta(j,j')}{\ell_d^2}\exp\Big(-\frac{\Delta(j,j')}{\ell_d}\Big),$$

and thus

$$\frac{\partial}{\partial \ell_d}\mathbf{v}_i\Sigma_d\mathbf{v}_i^\top = \frac{\sigma_d^2}{\ell_d^2}\sum_j\sum_{j'}\mathbf{v}_{i,j}\Delta(j,j')\exp\Big(-\frac{\Delta(j,j')}{\ell_d}\Big)\mathbf{v}_{i,j'}.$$

Then it follows that if the vectors $\mathbf{v}_i$ were arbitrarily, the monotonicity issue essentially rests on the positive definiteness of the associated derive kernel. For example, for the standard Gaussian kernel $e^{-\frac{(x-y)^2}{\ell_d}}$ (i.e., Δ is the squared Euclidean distance), we need to verify the positive definiteness of the kernel $k(x,y) = (x-y)^2 e^{-\frac{(x-y)^2}{\ell_d}}$. This issue seems generally challenging to verify directly, since $(x-y)^2$ is not a positive semidefinite kernel by itself on $\mathbb{R}$, even though the Gaussian kernel $e^{-\frac{(x-y)^2}{\ell_d}}$ is indeed positive semidefinite. Thus, one cannot use the standard Schur product theorem to conclude the positive definiteness. Alternatively, one can also compute the Fourier transform of the kernel $k(x) = x^2e^{-x^2}$ directly, which is given by

$$\mathcal{F}[k(x)](\omega) = \frac{2-\omega^2}{4}\frac{1}{\sqrt{2}}e^{-\frac{\omega^2}{4}}.$$

Clearly, the Fourier transform is not positive over the whole real line $\mathbb{R}$. By Bochner's theorem, this kernel is actually not positive. The fact that the kernel is no longer positive definite makes the analytical study challenging. This indicates the risk for a potential non-monotonicity in $\boldsymbol{\ell}$. Nonetheless, we emphasise that this condition is only sufficient, but not necessary. We leave a full investigation of the monotonicity to a future work, given the the compelling empirical evidence for monotonicity in both the NN and linearised settings.