Posterior Temperature Optimization in Variational Inference for Inverse Problems

Max-Heinrich Laves
Hamburg University of Technology
max-heinrich.laves@tuhh.de

Malte Tölle
Heidelberg University Hospital
malte.toelle@med.uni-heidelberg.de

Alexander Schlaefer
Hamburg University of Technology
schlaefer@tuhh.de

Sandy Engelhardt
Heidelberg University Hospital
sandy.engelhardt@med.uni-heidelberg.de

Abstract

Bayesian methods feature useful properties for solving inverse problems, such as tomographic reconstruction. The prior distribution introduces regularization, which helps solving the ill-posed problem and reduces overfitting. In practice, this often results in a suboptimal posterior temperature and the full potential of the Bayesian approach is not realized. In this paper, we optimize both the parameters of the prior distribution and the posterior temperature using Bayesian optimization. Well-tempered posteriors lead to better predictive performance and improved uncertainty calibration, which we demonstrate for the task of sparse-view CT reconstruction. Our source code is publicly available at github.com/Cardio-AI/mfvi-dip-mia.

1 Introduction

Reconstructing a tomography from a finite number of X-ray projections requires solving an inverse problem. The unknown tomography $\boldsymbol{x}$ can only be observed through projections $\boldsymbol{y} = \mathcal{F}[\boldsymbol{x}]$, affected by the forward Radon transform $\mathcal{F}$, which is not directly invertible. The reconstruction can be found by minimization of the ill-posed objective $\hat{\boldsymbol{x}} = \arg\min \{\mathcal{L}(\boldsymbol{y}, \mathcal{F}[\hat{\boldsymbol{x}}]) + \lambda \mathcal{R}(\hat{\boldsymbol{x}})\}$, with similarity measure $\mathcal{L}$ and regularization $\mathcal{R}$, weighted by λ [1]. Common regularization is manually engineered, such as penalization of spatial derivatives, or implicitly learned from a large data set. However, obtaining ground truth pairs $\{\boldsymbol{x}, \boldsymbol{y}\}$ is impossible in computed tomography (CT), especially in sparse-view CT, where only a limited number of projections are obtained to reduce radiation exposure.

Deep image prior (DIP) has shown promising results in solving inverse problems by optimizing a randomly-initialized convolutional network as neural representation of the reconstruction [2, 3]. To overcome the overfitting behavior of DIP, different Bayesian approaches have been proposed [4, 5]. In Bayesian deep learning, a prior distribution $p(\boldsymbol{w} \,|\, \alpha)$ is placed over the weights $\boldsymbol{w}$ of a neural network, governed by a hyperparameter α. After observing the data $\mathcal{D}$, we are interested in the posterior $p(\boldsymbol{w} \,|\, \mathcal{D}, \alpha) = p(\mathcal{D} \,|\, \boldsymbol{w}, \alpha) p(\boldsymbol{w} \,|\, \alpha) / p(\mathcal{D})$. However, this distribution is not tractable in general as the normalizing factor involves marginalization of the model likelihood over the prior $p(\mathcal{D}) = \int p(\mathcal{D} \,|\, \boldsymbol{w}, \alpha) p(\boldsymbol{w} \,|\, \alpha) \,\mathrm{d}\boldsymbol{w}$. A common way to approximate the posterior is variational inference (VI), which uses optimization to find the member $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ of a family of distributions that is close to the exact posterior, defined by the variational parameters $\boldsymbol{\phi}$. $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ is optimized w.r.t. $\boldsymbol{\phi}$, such that the Kullback-Leibler divergence is minimized with regard to the true posterior [6]. A practical implementations of VI is *Bayes by backprop*, where a fully factorized Gaussian distribution $w_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$ is used as variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{w})$, also known as mean-field distribution, which treats the mean and variance of each weight as learnable parameters $\phi_{ij} = \{\mu_{ij}, \sigma_{ij}^2\}$ [7].

Bayesian Deep Learning workshop, NeurIPS 2021.

Cold Posteriors Cold posteriors have been reported to perform better in practice in the context of Bayesian deep learning [8]. In order to bring the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ close to the true posterior, a lower bound on the log-evidence (ELBO) is derived and maximized. Graves [9] already suggested to reweight the complexity term in the ELBO using a factor λ to balance both terms in case of discrepancy between number of weights and training samples:

$$\mathrm{ELBO}(q_{\boldsymbol{\phi}}(\boldsymbol{w})) = \mathbb{E}_{\boldsymbol{w}\sim q}[\log p(\mathcal{D} \,|\, \boldsymbol{w})] - \lambda \,\mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{w}) \,\|\, p(\boldsymbol{w})] \,. \tag{1}$$

It is common for Bayesian deep learning researchers to employ values of $\lambda < 1$ to achieve better predictive performance [7]. While their main motivation was to qualitatively balance out discrepancies between number of model parameter and dataset size, the reweighting has recently been studied in more detail and described as the "cold posterior" effect [10]. Wenzel et al. [8] derived the tempered Bayesian posterior $p(\boldsymbol{w} \,|\, \mathcal{D}) \propto \exp(-U(\boldsymbol{w})/T)$ with posterior energy function $U(\boldsymbol{w}) = -\log p(\mathcal{D} \,|\, \boldsymbol{w}) - \log p(\boldsymbol{w})$ and have shown empirically that cold posteriors with $T < 1$ perform considerably better. The authors also recover Eq. (1) and show that introducing λ into the ELBO is equivalent to a partially tempered posterior, where only the likelihood term is scaled.

In this paper, we will not argue whether cold posteriors invalidate Bayesian principles, as there is disagreement among researchers [8, 10, 11], but use it in a directed way to increase predictive performance and uncertainty calibration of unsupervised sparse-view CT reconstruction with deep image prior. This workshop paper is based on our recent journal submission [12] and extends it by additional experiments on CIFAR-10/100 (see Appendix C).

2 Temperature-scaled Posterior

The ELBO for a fully temperature-scaled posterior in VI is given by (derivation in Appendix B):

$$\begin{aligned}\mathrm{ELBO}_T(q_{\boldsymbol{\phi}}(\boldsymbol{w})) &= -\,\mathbb{E}_{\boldsymbol{w}}\left[\log q_{\phi}(\boldsymbol{w}) - \tfrac{1}{T}\log p(\boldsymbol{w})\right] + \mathbb{E}_{\boldsymbol{w}}\left[\tfrac{1}{T}\log p(\mathcal{D} \,|\, \boldsymbol{w})\right] &(2)\\ &= -\,\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{w}) \,\|\, p(\boldsymbol{w})^{1/T}\right] + \mathbb{E}_{\boldsymbol{w}}\left[\tfrac{1}{T}\log p(\mathcal{D} \,|\, \boldsymbol{w})\right] \,. &(3)\end{aligned}$$

The KL contains the scaled prior $p_T(\boldsymbol{w}) \propto p(\boldsymbol{w})^{1/T}$, which will have the same mean, but different variance as the unscaled prior. In case of a Gaussian prior $p(\boldsymbol{w}) \propto \exp(-\|\boldsymbol{w}\|^2/2\sigma^2)$, this is equivalent to a scaled prior variance $p(\boldsymbol{w})^{1/T} \propto \exp(-\|\boldsymbol{w}\|^2/2\sigma_T^2)$ with $\sigma_T = \sqrt{T}\sigma$ [13]. Therefore, we set $p_T(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \frac{\sigma^2}{T}\boldsymbol{I}^2)$, which results in the following minimization criterion

$$\arg\min_{\boldsymbol{\phi}} \; T \cdot \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{w}) \,\|\, p_T(\boldsymbol{w} \,|\, T)\right] - \mathbb{E}_{\boldsymbol{w}}\left[\log p(\mathcal{D} \,|\, \boldsymbol{w})\right] \,, \tag{4}$$

which, in contrast to Eq. (1) and Wenzel et al. [8], optimizes the fully temperature-scaled ELBO_T.

3 Posterior Temperature Optimization

Instead of manually selecting the optimal posterior temperature using heuristics or inefficient grid search, we employ Bayesian optimization (BO) to jointly find the posterior temperature T and prior scale σ. BO allows us to optimize functions that are expensive to evaluate, e.g., the training of a deep network [14]. It uses a computationally inexpensive surrogate to retrieve a distribution over functions.

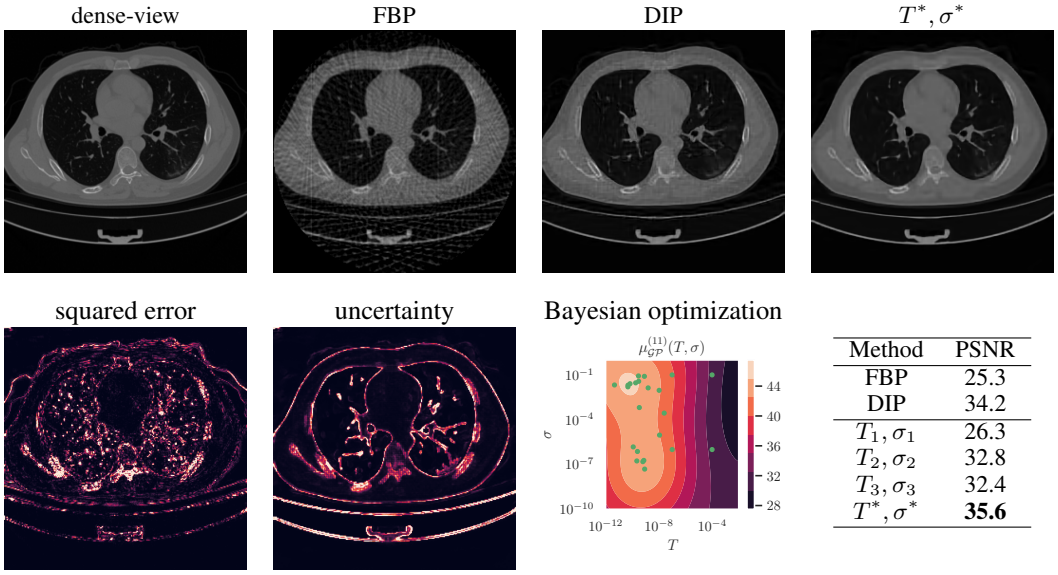
We apply optimization of the posterior temperature to maximize the peak signal-to-noise ratio (PSNR) between the sparse-view reconstruction $\hat{\boldsymbol{x}}$ and the dense-view image $\boldsymbol{x}$ as a function of T and σ

$$\max_{T\in\mathcal{T},\sigma\in\mathcal{S}} \; f(T,\sigma) = \max_{T\in\mathcal{T},\sigma\in\mathcal{S}} \; \mathrm{PSNR}(\hat{\boldsymbol{x}}(T,\sigma), \boldsymbol{x}) \tag{5}$$

using a Gaussian process (GP) as surrogate $f \sim \mathcal{GP}$. In each step of the BO, we evaluate our objective function f at the current candidates T^* and σ^* to increase the set of observations $\mathcal{D}_{\mathrm{BO}}$ and update the posterior of the surrogate model. Next, we maximize an acquisition function $a(T, \sigma; \mu_{\mathcal{GP}}, \sigma^2_{\mathcal{GP}})$ using the current GP posterior mean $\mu_{\mathcal{GP}}$ and variance $\sigma^2_{\mathcal{GP}}$. Its maximizing arguments $T^*, \sigma^* \leftarrow \arg\max \; a(T, \sigma; \mu_{\mathcal{GP}}, \sigma^2_{\mathcal{GP}})$ are used as candidates for the next iteration [15]. We choose the commonly accepted expected improvement (EI) as acquisition function

$$a_{\mathrm{EI}}(T, \sigma; \mu_{\mathcal{GP}}, \sigma^2_{\mathcal{GP}}) = \mathbb{E}\left[\max(y - f^*), 0) \,|\, y \sim \mathcal{N}(\mu_{\mathcal{GP}}(T,\sigma), \sigma^2_{\mathcal{GP}}(T,\sigma))\right] \,, \tag{6}$$

where $f^* = f(T_{\mathrm{best}}, \sigma_{\mathrm{best}})$ is the minimal value of the objective function observed so far. Eq. (6) can be solved analytically as shown in [16]. We utilize automatic differentiation from modern deep learning frameworks to optimize the acquisition function to get the next candidates T^* and σ^* [17].



Method	PSNR
FBP	25.3
DIP	34.2
T_1, σ_1	26.3
T_2, σ_2	32.8
T_3, σ_3	32.4
T^*, σ^*	**35.6**

Figure 1: Posterior temperature optimization for sparse-view CT reconstruction: (Top) Dense-view ground truth and sparse-view test reconstruction from FBP, non-Bayesian DIP and Bayesian DIP at optimal posterior temperature T^* and prior scale σ^*. (Bottom) Predictive error and uncertainty at $\{T^*, \sigma^*\}$, mean of the GP from BO, and PSNR for different methods and values of $\{T, \sigma\}$.

4 Experiments

To evaluate posterior temperature optimization in Bayesian inversion, we simulate sparse-view CT by computing only 45 projections from dense-view lung CTs of COVID-19 patients[1] using the forward Radon transform. We use mean-field VI (MFVI) as Bayesian approach to DIP for solving the inverse task (see Fig. 2 in the appendix). The Bayesian network is used as parameterization of the reconstruction $\hat{\boldsymbol{x}}$ and its variational parameters are optimized by minimizing Eq. (4) using the squared error $\|\mathcal{F}[\hat{\boldsymbol{x}}] - \boldsymbol{y}\|^2$ as likelihood. BO is used to find optimal values for $\{T, \sigma\}$ as described below.

Finding the Optimal Posterior Temperature The Gaussian process regressor from § 3 is implemented in GPyTorch [17] using a constant mean function with prior $\mathcal{N}(15, 4^2)$, a scaled radial basis function kernel as covariance function and a prior length-scale $\ell = 0.3$. The surrogate model is trained on observations $\{(\log T_i, \log \sigma_i), \mathrm{PSNR}(\hat{\boldsymbol{x}}_{T_i,\sigma_i}, \boldsymbol{x})\}$ to impose a non-negativity constraint on T and σ. A Gaussian likelihood with a homoscedastic noise model with prior $\Gamma(0.1, 100)$ is used. We limit the search space to $T \in [1\mathrm{e}{-12}, 1\mathrm{e}{-2}]$ and $\sigma \in [1\mathrm{e}{-10}, 1]$ and initialize the BO with four candidate pairs with $T \in \{1\mathrm{e}{-7}, 1\mathrm{e}{-4}\}$ and $\sigma \in \{1\mathrm{e}{-6}, 1\mathrm{e}{-1}\}$. If the acquisition function from Eq. (6) has multiple local maxima, we select the best four candidates for the next iteration.

Results The results for a test image are summarized in Fig. 1. At optimal temperature T^*, the Bayesian reconstruction outperforms filtered back-projection (FBP) and non-Bayesian DIP by means of PSNR. From the GP mean, we see that the posterior temperature has a considerable effect on the reconstruction, with $T^* \ll 1$. The effect of the prior scale is less prominent, with optimal value $\sigma^* \approx 1\mathrm{e}{-2}$. We observe similar findings for classification experiments on CIFAR-10/100 with Bayesian ResNets (see Appendix C). The uncertainty calibration is improved at optimal temperature.

5 Conclusion

We optimized the ELBO for a fully tempered posterior to exploit the cold posterior effect in Bayesian deep learning. For ill-posed inverse problems, the optimized posterior temperature introduces the right amount of regularization to allow enough flexibility but to avoid overfitting. This can be used in many medical applications such as CT reconstruction, registration, denoising, or artifact removal.

[1] We use publicly available data from https://coronacases.org

Acknowledgments and Disclosure of Funding

MT is supported by Informatics for Life founded by the Klaus Tschira Foundation. ML and AS are partially funded by the Interdisciplinary Competence Center for Interface Research (ICCIR).

References

[1] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013. doi: 10.1109/TMI.2013.2265603.

[2] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. Deep Image Prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. doi: 10.1109/CVPR.2018.00984.

[3] 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.

[4] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A bayesian perspective on the deep image prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5451, 2019.

[5] 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, 2020.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[8] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, volume 119, pages 10248–10259, 2020.

[9] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2011.

[10] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

[11] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? In *International Conference on Machine Learning*, 2021. arXiv preprint arXiv:2104.14421.

[12] Max-Heinrich Laves, Malte Tölle, Alexander Schlaefer, and Sandy Engelhardt. Posterior temperature optimized bayesian models for inverse problems in medical imaging. *under review*, 2021.

[13] Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2021.

[14] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, pages 2171–2180, 2015.

[15] Peter I. Frazier. A tutorial on bayesian optimization. In *arXiv Preprint*, 2018. arXiv:1807.02811.

[16] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[17] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[18] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. In *Bayesian Deep Learning Workshop (NeurIPS)*, 2019. arXiv:1909.13550.

A Conceptual Overview

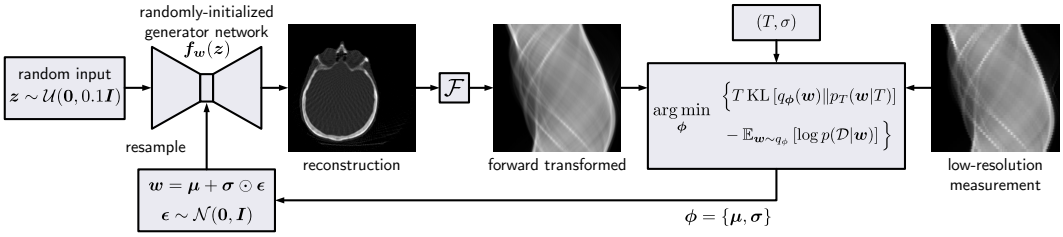


Figure 2: Conceptual overview. A randomly-initialized MFVI autoencoder network fed with uniform noise outputs a CT. The image reconstruction is performed iteratively by applying the forward Radon transform $\mathcal{F}$ and minimizing the fully tempered negative ELBO w.r.t. the variational parameters $\boldsymbol{\phi} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$ using gradient descent. The posterior temperature T and prior standard deviation σ are found using Bayesian optimization.

B Derivation of Fully Tempered ELBO

In the following, the ELBO for a fully temperature-scaled Bayesian posterior in variational inference is derived. Let $p_T(\boldsymbol{w} \mid \mathcal{D})$ be the fully tempered posterior [8]:

$$\mathrm{KL}\left[q_\phi(\boldsymbol{w}) \,\|\, p_T(\boldsymbol{w} \mid \mathcal{D})\right] \tag{7}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left[\log q_\phi(\boldsymbol{w}) - \log p_T(\boldsymbol{w} \mid \mathcal{D})\right] \tag{8}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left[\log q_\phi(\boldsymbol{w}) - \log \frac{(p(\boldsymbol{w} \mid \mathcal{D})p(\boldsymbol{w}))^{1/T}}{\int (p(\boldsymbol{w}' \mid \mathcal{D})p(\boldsymbol{w}'))^{1/T}\,\mathrm{d}\boldsymbol{w}'}\right] \tag{9}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left[\log q_\phi(\boldsymbol{w}) - \log(p(\boldsymbol{w} \mid \mathcal{D})p(\boldsymbol{w}))^{1/T}\right] + \underbrace{\log \int (p(\boldsymbol{w} \mid \mathcal{D})p(\boldsymbol{w}))^{1/T}\,\mathrm{d}\boldsymbol{w}}_{\text{const. w.r.t. } \boldsymbol{w},\ =:\log E_T} \tag{10}$$

$$= \underbrace{\mathbb{E}_{\boldsymbol{w}}\left[\log q_\phi(\boldsymbol{w}) - \tfrac{1}{T}\log p(\boldsymbol{w})\right] - \mathbb{E}_{\boldsymbol{w}}\left[\tfrac{1}{T}\log p(\mathcal{D} \mid \boldsymbol{w})\right]}_{=:\mathrm{ELBO}_T(q_\phi(\boldsymbol{w}))} + \log E_T \tag{11}$$

$$\Rightarrow \log E_T = \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{w}) \,\|\, p_T(\boldsymbol{w} \mid \mathcal{D})\right] + \mathrm{ELBO}_T(q_{\boldsymbol{\phi}}(\boldsymbol{w})) \tag{12}$$

As the tempered evidence E_T is constant, maximizing ELBO_T minimizes the KL, thus bringing the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ closer to the fully tempered posterior $p_T(\boldsymbol{w} \mid \mathcal{D})$.

C CIFAR-10/100 Experiments

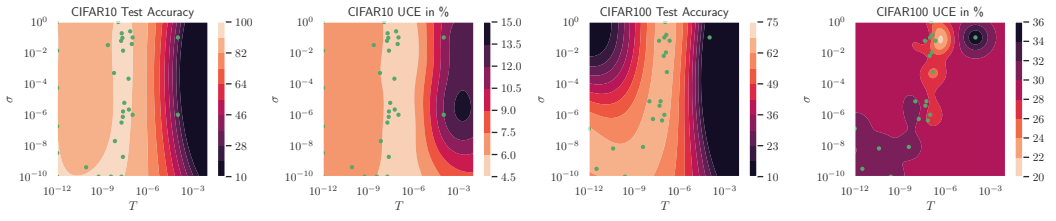


Figure 3: We additionally perform classification experiments on CIFAR-10 (ResNet-34) and CIFAR-100 (ResNet-50). The figures show estimated accuracy and uncertainty calibration error (UCE) [18] landscapes. Green dots denote observed points during BO. As for CT reconstruction, the posterior temperature T has a considerable effect on both the accuracy and calibration. On CIFAR-100, the effect of the prior scale σ on the calibration can not be neglected. We measure uncertainty as the entropy of the softmax vector after Monte Carlo integration $\mathcal{H}\left[1/N \sum_{i=1}^{N} \boldsymbol{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_i)\right]$.

D BO Steps

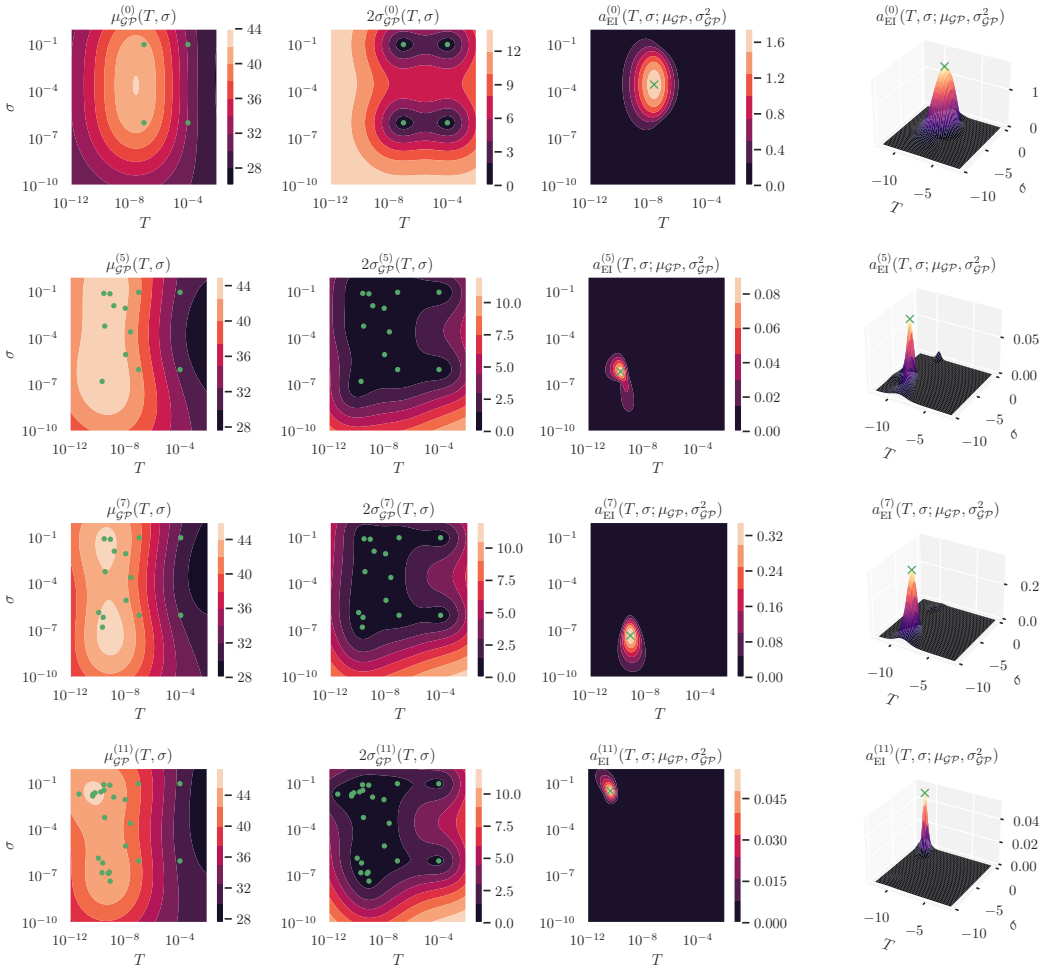


Figure 4: Posterior temperature optimization for CT reconstruction: GP mean, confidence (2 standard deviations) and expected improvement acquisition function after BO iteration $i \in \{0, 5, 7, 11\}$. Green dots denote observed points and green crosses show candidates for the next BO iteration. Note that per BO step, up to 4 candidates are evaluated in parallel.

E Implementation Details

- Code for training pipeline and evaluation is available at github.com/Cardio-AI/mfvi-dip-mia.
- For CT reconstruction, we use the same architecture as described by Lempitsky et al. [2] and optimize the network for 1e5 iterations.
- The final CT is sampled from the probabilistic neural representation using Monte Carlo integration $\hat{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{x}}_i$, where $\hat{\boldsymbol{x}}_i$ is a sample from the posterior predictive $p(\boldsymbol{x} \mid \boldsymbol{w}_i, \boldsymbol{y})$.
- We estimate reconstruction uncertainty using the predictive variance from Monte Carlo samples $\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{x}}_i - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{x}}_i)^2$.