ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Summer 2025

Matteo Gätzner

Anytime-valid Neural Uncertainty Quantification for SPECT Imaging

Submission Date: 29.09.2025

Co-Advisor Dr. Johannes Kirschner
Advisor: Prof. Dr. Jonas Peters

To my parents, my sister, and my girlfriend, for their endless love and support.

Acknowledgements

I would like to express my sincere gratitude to those who have supported me during this work. In particular, I am deeply grateful to my supervisor, Prof. Dr. Jonas Peters, for his valuable feedback on the writing of this thesis; to my co-supervisor, Dr. Johannes Kirschner, for his advice regarding experiments and clear explanations of the theoretical background; and to Dr. Luis Barba Flores, for his support in helping me understand tomographic imaging and the implementation of neural methods.

Code and data availability

Due to licensing restrictions, the dataset used in the experiments cannot be publicly shared. A paper based on this work is currently in preparation. Upon acceptance, the code for simulating SPECT measurements, training models, computing confidence sequences, and visualizing results will be made available at https://github.com/MatteoGaetzner/anytime-nuq.

Abstract

We develop anytime-valid methods for uncertainty quantification in tomographic imaging, with a focus on single-photon emission tomography (SPECT). In SPECT, sequentially acquired data is used to reconstruct images representing the radioactivity distribution inside the object. In addition to producing image reconstructions, our approach constructs *confidence sequences*: collections of confidence sets that contain the true but unknown image with high probability simultaneously across all acquisition steps.

We investigate two variants: *prior likelihood mixing* and *sequential likelihood mixing*. Both employ likelihood-based constructions, but differ in how they use user-defined distributions. We parameterize these distributions using classical statistical estimators (MLE, MAP) as well as neural methods, namely U-Net ensembles and diffusion models.

In numerical experiments, we simulate SPECT data and compare the tightness and empirical coverage rate of different confidence sequences. Empirically, sequential likelihood mixing proves to be a particularly effective method for constructing confidence sequences. The performance of this method depends on the image predictor used: U-Net ensembles often yield tight and reliable confidence sets, while in some settings classical estimators (MLE, MAP) perform best. We also present strategies for generating uncertainty visualizations. Our results suggest that combining statistical theory with neural predictors enables principled, real-time uncertainty quantification, which may support clinical decision-making in SPECT and related modalities.

Contents

Chapter 1

Introduction

Tomography is a widely used class of imaging techniques that guide critical decisions in medicine, manufacturing, and chip metrology, for example, whether and where in the patient to perform cancer treatment (Kak and Slaney, 2001; Gerlier et al., 2022; De Chiffre et al., 2014; Pacheco and Goyal, 2010; Bhargava et al., 2012). Such imaging techniques recover a high-dimensional 3D volume or 2D slice from noisy lower-dimensional measurements – whether that object is human tissue, an industrial part, or the layers of a microchip (Kak and Slaney, 2001; De Chiffre et al., 2014; Brown, 2014).

Examples of tomographic imaging technologies are Positron Emission Tomography (PET) (Sweet, 1951; Wrenn et al., 1951; Kuhl and Edwards, 1963), Single-Photon Emission Computed Tomography (SPECT) (Kuhl and Edwards, 1963), Magnetic Resonance Tomography (MRT) (Rabi, 1937; Lauterbur, 1973), X-ray Computed Tomography (CT) (Cormack, 1963; Hounsfield, 1973). In the case of PET and SPECT the reconstructed images or volumes represent the activity levels of an administered radioisotope and can be used to diagnose brain, bone, and heart diseases (Bhargava et al., 2012). Although all tomographic reconstruction techniques use noisy measurements, most, like *filtered backprojection algorithm* (FBP) (Bracewell and Riddle, 1967), *penalized likelihood image reconstruction* (Fessler and Rogers, 1996), algorithms based on expectation-maximization (Shepp and Vardi, 1982; Hudson and Larkin, 1994), and many machine learning-based approaches (Kiss et al., 2025) do not quantify how much we can trust reconstructions and only give a point prediction.

Quantifying uncertainty in reconstructions instead of only giving point estimates has several advantages: it can prevent misinterpretation of noise artifacts as real structures, reduce patient or component exposure by stopping acquisition once an uncertainty threshold has been reached, optimize scanning time allocation by choosing angles that are expected to improve reconstruction quality the most (Barba et al., 2024), and increase trust and interpretability by giving operators visual cues about reconstruction reliability.

Before introducing the proposed constructions, it is important to clarify the mathematical setting under which they can be applied. In particular, our confidence sequence methods require certain structural assumptions on the data generating process. We now state these requirements formally.

The first requirement is that we need measurable spaces (Definition A.17) $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, and a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ (Definition A.19) that is composed of a sample space

$\Omega = (\mathcal{X} \times \mathcal{Y})^\infty$, σ-algebra (Definition A.5) $\mathcal{F} = (\mathcal{F}_\mathcal{X} \otimes \mathcal{F}_\mathcal{Y})^{\otimes\infty}$, and probability measure (Definition A.24) $\mathbf{P}$ on measurable space $(\Omega, \mathcal{F})$.

For all $i \in \mathbb{N}$,[1] define random variables (Definition A.20)

$$\begin{aligned} X_i &: \Omega \to \mathcal{X} & X_i((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots) &= \mathbf{x}_i \\ Y_i &: \Omega \to \mathcal{Y} & Y_i((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots) &= \mathbf{y}_i. \end{aligned}$$

The second requirement is that a known parameter space Θ and $\boldsymbol{\theta}^* \in \Theta$ exist such that for all covariates $\mathbf{x} \in \mathcal{X}$ a known family of conditional distributions on $\mathcal{Y}$, $\{P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x}) : \boldsymbol{\theta} \in \Theta\}$ with Radon-Nikodym derivative (Corollary A.48)

$$p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x}) = \frac{dP_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x})}{d\xi}$$

and σ-finite measure (Definition A.26) ξ on $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$, exists such that for all possible data sequences

$$((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots) \in (\mathcal{X} \times \mathcal{Y})^\infty$$

and all $t \in \mathbb{N}$ we have

$$(Y_t \mid X_t = \mathbf{x}_t) \sim P_{\boldsymbol{\theta}^*}(\cdot \mid \mathbf{x}_t).$$

This requirement is called *realizability* (Kirschner et al., 2025). Example 1.1 shows that realizability does not necessarily hold for all families of conditional distributions.

Example 1.1. Let $x_1, x_2, \dots \in (0, \infty)$, $\Theta = (0, \infty)$ and $(Y_t \mid X_t = x_t) \sim \mathcal{N}(\mathbf{0}, x_t)$. For all covariates $x \in (0, \infty)$, define the family of conditional distributions on $\mathbb{R}$ as

$$\mathscr{F}_x \coloneqq \{\mathcal{N}(1, \theta x) : \theta \in \Theta\}.$$

Then, no $\theta \in \Theta$ exists such that for all $t \in \mathbb{N}$

$$(Y_t \mid X_t = x_t) \sim \mathcal{N}(\mathbf{0}, \theta x_t)$$

since the mean of $\mathcal{N}(\mathbf{0}, x_t)$ is 0 and all members of $\mathscr{F}_x$ have mean 1. Hence, realizability does not hold if we use $\mathscr{F}_x$ as our family of conditional distributions.

The last requirement is *conditional independence of labels* $(Y_i, i \in \mathbb{N})$ given covariates $(X_i, i \in \mathbb{N})$ and parameter $\boldsymbol{\theta} \in \Theta$: for all $t \in \mathbb{N}$, $\mathbf{x}_{1:t} \coloneqq (\mathbf{x}_1, \dots, \mathbf{x}_t) \in \mathcal{X}^t$, all $\mathbf{y}_{1:t} \coloneqq (\mathbf{y}_1, \dots, \mathbf{y}_t) \in \mathcal{Y}^t$ and all $\boldsymbol{\theta} \in \Theta$, we require that the joint density satisfies

$$p_{\boldsymbol{\theta}}(\mathbf{y}_{1:t} \mid \mathbf{x}_{1:t}) = \prod_{s=1}^{t} p_{\boldsymbol{\theta}}(\mathbf{y}_s \mid \mathbf{x}_s).$$

These three requirements need to be satisfied in order to construct the confidence sequences discussed in Chapter 2. We assume that they hold for the remainder of the thesis.

[1] We use convention $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, \dots\}$.

Experimental Setting

To demonstrate and evaluate our uncertainty quantification methods, we choose a concrete tomography method, namely Single-Photon Emission Computed Tomography (SPECT). At acquisition step $t \in \mathbb{N}$, reconstruction task is to use a finite sequence of measured Poisson counts $\mathbf{y}_1, \ldots, \mathbf{y}_t \in \mathbb{N}_0^r$, $r \in \mathbb{N}$ to recover the (flattened) unknown activity distribution $\boldsymbol{\theta}^* \in [0,1]^{r^2}$ that caused them. We assume a simplified 2D parallel beam geometry and a fixed activity image resolution of 64×64, so $r = 64$ in our experiments. During each acquisition step $t \in \mathbb{N}$ a detector measures photon counts $\mathbf{y}_t \in \mathbb{N}_0^{64}$ along a projection angle $x_t \in [0, 180]$ (measured in degrees).

In what follows, we formalize the measurement process, describe the forward model underlying the data, and clarify the simplifying assumptions we adopt to focus on the uncertainty quantification aspect of the problem.

Formally, the measurable spaces are

$$(\mathcal{X}, \mathcal{F}_\mathcal{X}) = \big([0,180], \mathcal{B}([0,180])\big), \qquad (\mathcal{Y}, \mathcal{F}_\mathcal{Y}) = \big(\mathbb{N}_0^r, \mathcal{P}(\mathbb{N}_0^r)\big),$$

with product sample space

$$\Omega = \big([0,180] \times \mathbb{N}_0^r\big)^\infty, \qquad \mathcal{F} = (\mathcal{F}_\mathcal{X} \otimes \mathcal{F}_\mathcal{Y})^{\otimes\infty}.$$

We assume a fixed design of projection angles $(x_t, t \in \mathbb{N}) \in [0,180]^\infty$, so randomness arises only from the photon count sequence $(Y_t, t \in \mathbb{N})$.

Forward Model

For all acquisition steps $t \in \mathbb{N}$ and $x_t \in [0,180]$, let $A_{x_t} \in \{0,1\}^{64 \times 64^2}$ be the projection matrix encoding which pixels contribute to which detector channels. Given the parameter $\boldsymbol{\theta} \in \Theta = [0,1]^{64^2}$, the mean photon counts are

$$\boldsymbol{\lambda}(\boldsymbol{\theta}, x_t) \coloneqq A_{x_t}\boldsymbol{\theta} \in [0,\infty)^{64}.$$

Conditional on x_t, the observed photon counts follow a product-Poisson law:

$$(Y_t \mid X_t = x_t) \sim \mathrm{Pois}(\boldsymbol{\lambda}(\boldsymbol{\theta}^*, x_t)),$$

i.e., for all $\mathbf{y}_t \in \mathbb{N}_0^r$,

$$p_{\boldsymbol{\theta}}(\mathbf{y}_t \mid x_t) = \prod_{i=1}^{64^2} \exp\left(-\lambda_i(\boldsymbol{\theta}, x_t)\right) \frac{\lambda_i(\boldsymbol{\theta}, x_t)^{y_{t,i}}}{y_{t,i}!}.$$

Equivalently, the i-th projection entry $\lambda_i(\boldsymbol{\theta}^*, x_t)$ can be expressed via the discrete Radon transform as the sum of activities along the line of response (LOR) at angle $x_t \in [0,180]$:

$$\lambda_i(\boldsymbol{\theta}^*, x_t) = \mathcal{R}_i(\boldsymbol{\theta}^*, x_t) = \sum_{(k,l) \in \mathrm{LOR}(i, x_t)} \theta^*_{k,l}.$$

Thus, the matrix formulation $A_{x_t}\boldsymbol{\theta}^*$ and the LOR-sum formulation are equivalent. Figure 1.1 illustrates projections for angles $x \in \{0, 90\}$ along with Poisson count measurements.

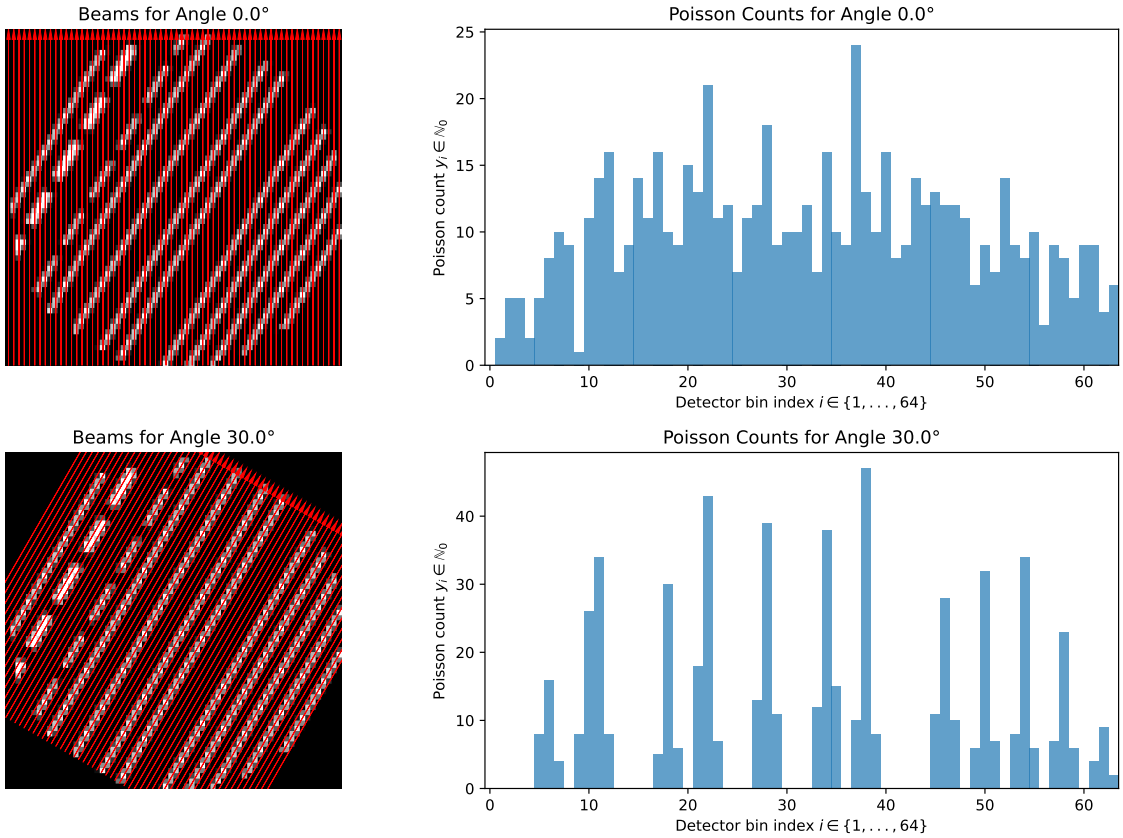


Figure 1.1: Visualization of Poisson count measurements $\mathbf{y} \sim \text{Pois}\left(\mathcal{R}(\boldsymbol{\theta}^*, x)\right)$ of first test set image at angles $x \in \{0, 30\}$.

Remarks

This forward model is deliberately simplified. For all $t \in \mathbb{N}$ and $x_t \in [0, 180]$ we use binary projection matrices $A_{x_t} \in \{0,1\}^{64\times 64^2}$, implying that each pixel intersected by a ray contributes equally and deterministically to the corresponding detector channel. In realistic SPECT models, the entries of A_{x_t} take values in $[0, 1]$, encoding detection probabilities that depend on the scattering, attenuation, and isotropic emission of gamma rays (Kak and Slaney, 2001). Our simplified setting is intended to isolate and evaluate the performance of confidence sequence methods for uncertainty quantification.

Uncertainty Quantification in Tomography

Although the above forward model specifies how measurements arise from an underlying image, it does not resolve the central question of how to quantify uncertainty in the resulting reconstructions. Addressing this question has been the subject of extensive research, and a variety of approaches have been proposed in the tomographic imaging literature. These include analytic methods (Qi and Leahy, 2000) asymptotically valid Bayesian methods that approximately sample the posterior (Zhou et al., 2020; Pedersen et al., 2022; Lee et al., 2024), bootstrap methods (Dahlbom, 2001) and dropout plus ensembling methods (Vasconcelos et al., 2023). Some authors employ conformal prediction to give prediction sets instead of point predictions which sidestep reliance on approximations or asymptotics and come with a theoretically guaranteed marginal coverage: for all covariates $\mathbf{x} \in \mathcal{X}$, error levels $\delta \in (0, 1)$ and true but unknown parameters $\boldsymbol{\theta}^* \in \Theta$ they give a covariate dependent set $C(\mathbf{x}) \subseteq \Theta$ for which

$$\mathbf{P}(\boldsymbol{\theta}^* \in C(\mathbf{x})) \geq 1 - \delta$$

holds (Kutiel et al., 2023; Ekmekci and Cetin, 2025).

Similar to the aforementioned conformal approaches, our approach also yields sets instead of points in image space. Concretely, we are computing *confidence sequences* (Darling and Robbins, 1967).

Definition 1.2 (Confidence Sequence)**.** *Let $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ be measurable spaces, and let Θ be a parameter space. Let $(\Omega, \mathcal{F})$ be a measurable space equipped with a family of probability measures $(\mathbf{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ (Definition A.19), representing candidate data-generating distributions under parameter $\boldsymbol{\theta} \in \Theta$. Let true but unknown parameter $\boldsymbol{\theta}^* \in \Theta$.*

Let $I \subseteq \mathbb{N}$ be a time index set, and let $Z := ((X_s, Y_s) : s \in I)$ be a stochastic process (Definition A.39) on probability space $(\Omega, \mathcal{F}, \mathbf{P}_{\boldsymbol{\theta}^*})$ with filtration generated by Z, $\mathbb{F} := \sigma(Z)$ (Definition A.41).

If a sequence $S = (S_t, t \in I)$ at level $\delta \in (0, 1)$ with confidence set *$S_t \subseteq \Theta$ for all $t \in I$ is a $\mathcal{P}(\Theta)$-valued, $\mathbb{F}$-adapted stochastic process (Definition A.39) for which*

$$\mathbf{P}_{\boldsymbol{\theta}^*}(\forall t \in I : \boldsymbol{\theta}^* \in S_t) \geq 1 - \delta$$

holds, then we call S a confidence sequence*.*

Let $\boldsymbol{\theta}^* \in \Theta$ and let $S = (S_s, s \in \mathbb{N})$ be a confidence sequence for some $\boldsymbol{\theta}^*$ at level $\delta \in (0, 1)$. For all acquisition steps $t \in \mathbb{N}$, the confidence set S_t depends only on data up to (and including) the data collected at step t. Since for all $t \in \mathbb{N}$, the confidence sets simultaneously contain $\boldsymbol{\theta}^*$ with high probability, the confidence sequence enables us

to make valid inferences at *any* time during the data acquisition process, hence they are called *anytime valid*. In contrast, the marginal coverage property from conformal prediction does not give use anytime validity, e.g. assume that for all $S'_1, S'_2, \dots \subseteq \Theta$ we have that for all $t \in \mathbb{N}$

$$\mathbf{P}(\boldsymbol{\theta}^* \in S'_t) \geq 1 - \delta$$

then, without further assumptions, the Fréchet inequalities give the tightest lower bound

$$\mathbf{P}(\boldsymbol{\theta}^* \in S'_1 \cap S'_2) \geq \max(0, 1 - 2\delta)$$

whereas if $S' = (S'_1, S'_2, \dots)$ were a confidence sequence for $\boldsymbol{\theta}^*$ at level δ, we would have the stronger guarantee

$$\mathbf{P}(\boldsymbol{\theta}^* \in S'_1 \cap S'_2) \geq 1 - \delta.$$

Practical Uses of Confidence Sequences

Once a confidence sequence $S = (S_s, s \in \mathbb{N})$ has been constructed, it can be turned into several practical tools. One possibility is to derive *pixelwise uncertainty images*: for all pixels, we determine the largest and smallest intensity values that remain plausible given the confidence set at step $t \in \mathbb{N}$, and use their difference as a measure of uncertainty. Another possibility is *distance-based uncertainty images*, where a given point estimate is compared to the most different image still consistent with the confidence sequence. This yields per-pixel deviation images and overall uncertainty scores that quantify how far the reconstruction might plausibly deviate from the estimate.

Uncertainty images can also be aggregated into simple scalar summaries. For instance, averaging per-pixel uncertainties across the whole image produces a single score that can drive *early stopping*: data acquisition is terminated once the uncertainty has fallen below a pre-specified threshold.

In Section 4.2, we discuss not only pixelwise and distance-based uncertainty images, but also other approaches, such as global and prediction-based uncertainty images. The details of these methods, along with empirical comparisons, are presented in Sections 4.2.1 to 4.2.4. At this point, the key message is that confidence sequences not only provide abstract coverage guarantees but can also be operationalized into concrete, task-specific visualizations that support interpretation and decision making in tomographic imaging.

Overall, the goal of this thesis is to assess, through experiments, how confidence sequence approaches can provide practical and informative uncertainty quantification in tomographic imaging.

Chapter 2

Likelihood-Based Confidence Sequences

As discussed in Chapter 1, the central mathematical objects in our approach are confidence sequences (Darling and Robbins, 1967; Robbins and Siegmund, 1970; Lai, 1976b,a). In this chapter, we (i) motivate *tight* confidence sets and relate tightness to level sets of the negative log-likelihood, (ii) present two likelihood-based constructions, *prior likelihood mixing* and *sequential likelihood mixing*, (iii) present a Laplace's method based approximation of prior likelihood mixing confidence sequences, (iv) discuss an equivalence result linking the two constructions, and (v) show how sequential likelihood mixing confidence sequences can be instantiated using mixing distributions that leverage MLE, MAP, U-Net ensembles, and diffusion-based predictors.

2.1 Tightness and Negative Log Likelihood

We now clarify what it means for a confidence sequence $S = (S_s, s \in \mathbb{N})$ to be *informative*. Although Definition 1.2 guarantees anytime-valid coverage, it says nothing about the size of the sets S_t. In addition to the anytime validity, we want S_t to shrink rapidly around the true parameter $\boldsymbol{\theta}^*$ as t increases and more data are observed, rather than remaining large and uninformative.

Let $S = (S_s, s \in \mathbb{N})$ be a confidence sequence for $\boldsymbol{\theta}^* \in \Theta$. Recall that $\boldsymbol{\theta}^*$ is unknown and we are interested in inferring it based on sequentially collected data. At step $t \in \mathbb{N}$, it is desirable for the confidence set S_t to be a small set of plausible parameters, based on the data seen up to step t. It is important that S_t is small because very large S_ts are uninformative, see Example 2.1.

Example 2.1. Let $\boldsymbol{\theta}^* \in \Theta$ and $S = (\Theta, \Theta, \dots)$. Since we assume $\boldsymbol{\theta}^* \in \Theta$, for any $\delta \in (0, 1)$,

$$\mathbf{P}(\forall t \in \mathbb{N} : \boldsymbol{\theta}^* \in \Theta) = 1 \geq 1 - \delta,$$

so S is a sound confidence sequence (Definition 1.2). However, no confidence sequence was necessary to determine that $\boldsymbol{\theta}^* \in \Theta$ as this is already assumed true. Hence, S is not informative.

As Example 2.1 shows, a confidence sequence is not necessarily informative. We prefer confidence sequences that concentrate around $\boldsymbol{\theta}^*$ and do so quickly with increasing t.

The confidence sequence constructions we consider are defined in terms of level sets of the negative log-likelihood. Hence, we define this concept next.

Definition 2.2 (Negative Log-Likelihood)**.** *Let $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ be measurable spaces, and let Θ be a parameter space. Suppose that for all $\boldsymbol{\theta} \in \Theta$ and covariate $\mathbf{x} \in \mathcal{X}$ we are given a conditional distribution $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x})$ on $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$, with Radon-Nikodym derivative $p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x})$ with respect to some reference measure. For all $t \in \mathbb{N}$, data sequences*

$$((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_t, \mathbf{y}_t)) \in (\mathcal{X} \times \mathcal{Y})^t$$

and parameters $\boldsymbol{\theta} \in \Theta$, the negative log-likelihood *of $\boldsymbol{\theta}$ at step t is defined as*

$$L_t(\boldsymbol{\theta}) \coloneqq -\sum_{s=1}^{t} \log p_{\boldsymbol{\theta}}(\mathbf{y}_s \mid \mathbf{x}_s).$$

For all $s \in \mathbb{N}$, $\boldsymbol{\theta} \in \Theta$, $\mathbf{x}_s \in \mathcal{X}$, and $\mathbf{y}_s \in \mathcal{Y}$, we introduce the shorthand notation

$$p_s(\mathbf{y}_s \mid \boldsymbol{\theta}) \coloneqq p_{\boldsymbol{\theta}}(\mathbf{y}_s \mid \mathbf{x}_s).$$

Let $\mathscr{P}(\Theta)$ denote the space of probability measures over Θ. Moreover, for each $t \in \mathbb{N}$ and every $\beta \in \mathbb{R}$, we define the set-valued function

$$C_t(\beta) \coloneqq \{\boldsymbol{\theta} \in \Theta \mid L_t(\boldsymbol{\theta}) \leq \beta\},$$

and refer to β as the *confidence coefficient*.

2.2 Prior Likelihood Mixing

In this section we define the prior likelihood mixing confidence sequence construction and Laplace's method-based approximations thereof. The prior likelihood mixing construction is given by the following theorem.

Theorem 2.3 (Prior Likelihood Mixing (Kirschner et al., 2025))**.** *For all distributions $\mu_0 \in \mathscr{P}(\Theta)$ that are independent of the observed data sequence $((\mathbf{x}_s, \mathbf{y}_s), s \in \mathbb{N}) \in (\mathcal{X} \times \mathcal{Y})^\infty$ and all error levels $\delta \in (0, 1)$, $(C_t(\beta_t^{\mathrm{plm}}(\delta)), t \in \mathbb{N})$ with confidence coefficient*

$$\beta_t^{\mathrm{plm}}(\delta) = \log \frac{1}{\delta} - \log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})$$

defines a confidence sequence for $\boldsymbol{\theta}^*$ at level δ.

A proof of Theorem 2.3 is provided in Appendix B.2.

Theorem 2.3 leaves open a design choice, namely, prior to $\mu_0 \in \mathscr{P}(\Theta)$. This is a critical design choice and influences the rate at which the confidence set $C_t(\beta_t^{\mathrm{plm}}(\delta)) \subseteq \Theta$ shrinks as $t \in \mathbb{N}$ increases. We generally desire μ_0 to be as concentrated around $\boldsymbol{\theta}^*$ as possible, so a natural choice would be $\mu_0(\boldsymbol{\nu}) = \delta_{\boldsymbol{\theta}^*}(\boldsymbol{\nu})$ where $\delta_{\boldsymbol{\theta}^*}$ is the Dirac measure centered on $\boldsymbol{\theta}^*$ with the property

$$\int_\Theta f(\boldsymbol{\nu}) \, d\delta_{\boldsymbol{\theta}^*}(\boldsymbol{\nu}) = f(\boldsymbol{\theta}^*)$$

for all continuous, compactly supported functions f. In this case, for all $t \in \mathbb{N}$, the integral inside $\beta_t^{\mathrm{plm}}(\delta)$ simplifies and we have

$$C_t(\beta_t^{\mathrm{plm}}(\delta)) = \left\{\boldsymbol{\theta} \in \Theta \mid L_t(\boldsymbol{\theta}) \leq \log \frac{1}{\delta} + L_t(\boldsymbol{\theta}^*)\right\}.$$

Unfortunately, we cannot construct $\delta_{\boldsymbol{\theta}^*}$ since $\boldsymbol{\theta}^*$ is unknown. Instead, we either construct uninformative or data-driven priors that leverage an auxiliary dataset collected prior to the start of the sequence.

2.2.1 Approximate Prior Likelihood Mixing

Let step $t \in \mathbb{N}$, error level $\delta \in (0,1)$ and data sequence $(\mathbf{x}_s, \mathbf{y}_s)_{s=1}^t \in (\mathcal{X} \times \mathcal{Y})^t$. In order to determine confidence coefficient $\beta_t^{\text{plm}}(\delta)$ it is necessary to evaluate potentially high-dimensional integral

$$\int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}).$$

Needless to say, performing this operation exactly is intractable for most $\mu_0 \in \mathscr{P}(\Theta)$. Fortunately, *Laplace's Method* can be used to approximate this integral (Laplace, 1878).

The multivariate version of Laplace's method is characterized by the following theorem.

Theorem 2.4 (Laplace's Method (Laplace, 1878))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ have two continuous derivatives on $K \subseteq \mathbb{R}^d$ and let $h : \mathbb{R}^d \to \mathbb{R}$ be continuous. Moreover, assume that f has a strict global minimizer $\mathbf{x}_* \in K$. Then gradient $\nabla f(\mathbf{x}_*) = 0$ and the Hessian at $\mathbf{x}_*$, $\nabla^2 f(\mathbf{x}_*)$, is positive semi-definite. Furthermore, assume that $h(\mathbf{x}_*) \neq 0$. Let $t \in \mathbb{R}$. For the integral*

$$I(t) = \int_K h(\mathbf{x}) \exp(-t f(\mathbf{x})) \, d\mathbf{x}$$

we have asymptotic equivalence (Definition A.2)

$$I(t) \sim \frac{h(\mathbf{x}_*)}{|\det \nabla^2 f(\mathbf{x}_*)|^{1/2}} \left(\frac{2\pi}{t}\right)^{d/2} \exp(-t f(\mathbf{x}_*)).$$

Theorem 2.4 and its proof, given in Appendix B.1, follow the exposition in Bach (2021), with additional details filled in where steps were omitted.

2.2.2 Application of Laplace's Method

We use Theorem 2.3 and Theorem 2.4 to define the approximations of prior likelihood mixing confidence sequences. For all $s \in \mathbb{N}$ and all $\boldsymbol{\theta} \in \Theta$, assume that a density f_0 of μ_0 exists with respect to some appropriate base measure. Moreover, define

$$L_t^+(\boldsymbol{\nu}) \coloneqq L_t(\boldsymbol{\nu}) - \log f_0(\boldsymbol{\nu}).$$

Let $t \in \mathbb{N}$. Then, approximate $-\log\left(\int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_0(\boldsymbol{\nu})\right)$ via Theorem 2.4:

$$\begin{aligned}
&-\log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_0(\boldsymbol{\nu}) \\
&= -\log \int f_0(\boldsymbol{\nu}) \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\boldsymbol{\nu} \\
&= -\log \int \exp\left(-(-\log f_0(\boldsymbol{\nu}) + L_t(\boldsymbol{\nu}))\right)\, d\boldsymbol{\nu} \\
&= -\log \int \exp\left(-L_t^+(\boldsymbol{\nu})\right)\, d\boldsymbol{\nu} \\
&\approx -\log \left(\frac{1}{\left|\det \nabla^2_{\boldsymbol{\theta}} L_t^+(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t^{\mathrm{MAP}}}\right|^{1/2}} (2\pi)^{d/2} \exp(-L_t^+(\boldsymbol{\theta}_t^{\mathrm{MAP}}) \right) \\
&= L_t^+(\boldsymbol{\theta}_t^{\mathrm{MAP}}) + \frac{1}{2}\log\left|\det \nabla^2_{\boldsymbol{\theta}} L_t^+(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t^{\mathrm{MAP}}}\right| - \frac{d}{2}\log(2\pi).
\end{aligned}$$

This approximation yields the confidence coefficient

$$\tilde{\beta}_t(\delta) \coloneqq \log\frac{1}{\delta} + L_t^+(\boldsymbol{\theta}_t^{\mathrm{MAP}}) + \frac{1}{2}\log\left|\det \nabla^2_{\boldsymbol{\theta}} L_t^+(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t^{\mathrm{MAP}}}\right| - \frac{d}{2}\log(2\pi),$$

approximate confidence set $C_t(\tilde{\beta}_t(\delta))$ and approximate confidence sequence

$$\tilde{S} \coloneqq (C_s(\tilde{\beta}_s(\delta), s \in \mathbb{N}).$$

Using $C_t(\tilde{\beta}_t(\delta))$ in favor of $C_t(\beta_t(\delta))$ avoids having to compute a high-dimensional integral. However, since $\tilde{S}$ is only an approximation of $(C_t(\beta_t^{\mathrm{plm}}(\delta)), t \in \mathbb{N})$ so

$$\mathbf{P}(\forall t \in \mathbb{N} : \boldsymbol{\theta}^* \in C_t(\tilde{\beta}_t(\delta))) \geq 1 - \delta$$

is not guaranteed. In fact, empirical results in Section 4.1.4 indicate that

$$\mathbf{P}(\forall t \in \mathbb{N} : \boldsymbol{\theta}^* \in C_t(\tilde{\beta}_t(\delta))) < 1 - \delta$$

in many settings.

Next, we analyze a powerful alternative confidence sequence construction that exploits predictive models in order to construct tight confidence coefficients.

2.3 Sequential Likelihood Mixing

The *sequential likelihood mixing* construction is defined by the following theorem.

Theorem 2.5 (Sequential Likelihood Mixing (Kirschner et al., 2025)). *Let $\mathcal{F}' = (\mathcal{F}'_s)_{s\in\mathbb{N}_0}$ be a filtration (Definition A.40) of the underlying probability space. For all $t \in \mathbb{N}$, let*

$$\mathcal{F}'_t = \sigma\left((X_s, Y_s), s \in [t]\right) \quad \textit{(Definition A.23)}$$

and $\mathcal{F}'_0 = \{\emptyset, \Omega\}$. Let the sequence of mixing distributions $(\mu_s, s \in \mathbb{N}_0)$ be a $\mathscr{P}(\Theta)$-valued $\mathcal{F}'$-adapted stochastic process (Definition A.41). For all $t \in \mathbb{N}$ and all error levels $\delta \in (0,1)$, define sequential likelihood mixing confidence coefficient

$$\beta_t^{\mathrm{slm}}(\delta) = \log\frac{1}{\delta} - \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_{s-1}(\boldsymbol{\nu}).$$

Then $(C_t(\beta_t^{\mathrm{slm}}(\delta)), t \in \mathbb{N})$ is a confidence sequence at level δ.

A proof of this theorem can be found in Appendix B.3.

Interestingly, the sequential likelihood mixing (Theorem 2.5) and prior likelihood mixing (Theorem 2.3) constructions are *equivalent* for specific choices of mixing distributions. Theorem 2.6 formalizes this statement.

Theorem 2.6 (Mixing Equivalence (Kirschner et al., 2025))**.** *Let $\mathcal{F}' = (\mathcal{F}'_s)_{s\in\mathbb{N}_0}$ be a filtration (Definition A.40) of the underlying probability space. For all $t \in \mathbb{N}$, let*

$$\mathcal{F}'_t = \sigma\left((X_s, Y_s), s \in [t]\right) \quad \textit{(Definition A.23)}$$

and $\mathcal{F}'_0 = \{\emptyset, \Omega\}$. Let the sequence of distributions $(\mu_s, s \in \mathbb{N}_0)$ be a $\mathscr{P}(\Theta)$-valued $\mathcal{F}'$-adapted stochastic process (Definition A.41) with

$$\mu_s(A) \propto \int_A \exp(-L_s(\boldsymbol{\theta}))\, d\mu_0(\boldsymbol{\theta}).$$

for all $s \in \mathbb{N}_0$ and $A \in \mathcal{F}'_s$. If for all $t \in \mathbb{N}$ and all error levels $\delta \in (0,1)$, we construct $\beta_t^{\mathrm{slm}}(\delta)$ using $\mu_0, \dots, \mu_t$, and $\beta_t^{\mathrm{plm}}(\delta)$ using μ_0, then $\beta_t^{\mathrm{plm}}(\delta) = \beta_t^{\mathrm{slm}}(\delta)$.

A proof of this statement can be found in Appendix B.4.

2.3.1 Mixing Distributions

For all error levels $\delta \in (0,1)$, all $t \in \mathbb{N}$ and mixing distributions $((\mu_0, \mu_1, \dots \in \mathscr{P}(\Theta)$ that match the conditions in Theorem 2.5,

$$\beta_t^{\mathrm{slm}}(\delta) = -\log\frac{1}{\delta} - \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_{s-1}(\boldsymbol{\nu}).$$

For all $s \in \mathbb{N}$, define

$$b_s^{\mathrm{slm}} := -\log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_{s-1}(\boldsymbol{\nu}).$$

Then, for all $t \in \mathbb{N}$

$$\beta_t^{\mathrm{slm}}(\delta) = \log\frac{1}{\delta} + \sum_{s=1}^{t} b_s^{\mathrm{slm}}.$$

For all $s \in \mathbb{N}$, b_s^{slm} can be interpreted as measuring how surprising observation $(\mathbf{y}_s, \mathbf{x}_s)$ is under μ_{s-1}. The more surprising $(\mathbf{y}_s, \mathbf{x}_s)$ is under μ_{s-1} the larger b_s^{slm}.

Compared to prior μ_0, which has to be independent of all data in the sequence, for all μ_t with $t \in \mathbb{N}$ *can* depend on $((\mathbf{x}_s, \mathbf{y}_s), s \in [t])$. In this work, we specialize μ_t as a uniform mixture of $k \in \mathbb{N}$ Dirac distributions: For all t, choose

$$\mu_t = \frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{t,i}}$$

centered on predictions $\hat{\boldsymbol{\theta}}_{t,1}, \dots, \hat{\boldsymbol{\theta}}_{t,k} \in \Theta$ of $\boldsymbol{\theta}^*$ based on observed data $((\mathbf{x}_s, \mathbf{y}_s), s \in [t])$. Then

$$\beta_t^{\mathrm{slm}}(\delta) = \log\frac{1}{\delta} - \sum_{s=1}^{t} \log\left(\frac{1}{k}\sum_{i=1}^{k} p_s(\mathbf{y}_s \mid \hat{\boldsymbol{\theta}}_{s,i})\right) \tag{2.1}$$

$$= \log\frac{1}{\delta} + t\log k - \sum_{s=1}^{t} \mathrm{LSE}(\log p_s(\mathbf{y}_s, \hat{\boldsymbol{\theta}}_{s,1}), \dots, \log p_s(\mathbf{y}_s, \hat{\boldsymbol{\theta}}_{s,k})) \tag{2.2}$$

where LSE denotes the log-sum-exp function:

$$\mathrm{LSE}(z_1, \ldots, z_k) \coloneqq z^* + \log(\exp(z_1 - z^*) + \cdots + \exp(z_k - z^*))$$

and $z^* = \max(z_1, \ldots, z_k)$. In our experiments, we use Equation (2.2) in favor of Equation (2.1) because it is more numerically stable.

We also investigate pre-processing $\hat{\boldsymbol{\theta}}_{t,1}, \ldots, \hat{\boldsymbol{\theta}}_{t,k} \in \Theta$ first by applying an aggregation map $A : \Theta^k \to \Theta$ to them and then setting

$$\mu_s = \delta_{\hat{\boldsymbol{\theta}}_s}, \qquad \hat{\boldsymbol{\theta}}_s = A(\hat{\boldsymbol{\theta}}_{s,1}, \ldots, \hat{\boldsymbol{\theta}}_{s,k}).$$

For example, one may take the coordinate median or average as A. This yields confidence coefficient

$$\beta_t^{\mathrm{slm}}(\delta) = \log \frac{1}{\delta} - \sum_{s=1}^{t} \log p_s(\mathbf{y}_s \mid \hat{\boldsymbol{\theta}}_s).$$

Such an approach can mitigate the effect of individual predictors that assign very low likelihood to $(\mathbf{x}_s, \mathbf{y}_s)$, potentially leading to a more stable confidence sequence.

We generate predictions $\hat{\boldsymbol{\theta}}_{s,1}, \ldots, \hat{\boldsymbol{\theta}}_{s,k}$ with four different methods, *Maximum Likelihood Estimation (MLE)*, *Maximum A Posteriori Estimation (MAP)*, *U-Net Ensembling*, *Diffusion*. The first two are classical statistical estimators. MLE relies solely on the observed data, while MAP incorporates both data and prior information. The latter two are neural methods: U-Net ensembles provide image-to-image mappings, while diffusion models enable flexible generative sampling.

- **Maximum Likelihood Estimation (MLE)**: We set $k = 1$ and define $\mu_s = \delta_{\hat{\boldsymbol{\theta}}_s^{\mathrm{MLE}}}$ with $\hat{\boldsymbol{\theta}}_s^{\mathrm{MLE}}$ an approximate maximum likelihood estimate based on the data up to step s. Details about how we compute $\hat{\boldsymbol{\theta}}_s^{\mathrm{MLE}}$ are provided in Section 4.1.3.
- **Maximum A Posteriori Estimation (MAP)**: Similarly, we set $k = 1$ and define $\mu_s = \delta_{\hat{\boldsymbol{\theta}}_s^{\mathrm{MAP}}}$ with $\hat{\boldsymbol{\theta}}_s^{\mathrm{MAP}}$ an approximate MAP estimate that incorporates both data and prior information. See Section 4.1.3 for the exact optimization formulation.
- **U-Net Ensemble**: Let $k \in \mathbb{N}$ represent the number of ensemble members. Each member is a U-Net (Ronneberger et al., 2015) and maps FBP reconstructions to the final prediction (Hansen and Salamon, 1990; Jin et al., 2017). For all $t \in \mathbb{N}$, let $\hat{\boldsymbol{\theta}}_{t,1}, \ldots, \hat{\boldsymbol{\theta}}_{t,k} \in \Theta$ be predictions from k different U-Net ensemble members where each mapped the same FBP (based on data sequence $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_t, \mathbf{y}_t)$) to its respective prediction. Then for experiments using this method we choose $\mu_t = \frac{1}{k} \sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{t,i}}$.
- **Diffusion**: In this approach, let $k \in \mathbb{N}$ correspond to the number of samples we generate. We assume the existence of an underlying distribution $P(\boldsymbol{\theta})$ on the parameter space Θ, representing the population distribution of parameters, and first train an unconditional diffusion model (Ho et al., 2020) to approximate it. Without further modification, the model then allows us to approximately draw samples from $P(\boldsymbol{\theta})$. However, these samples may be inconsistent with the observed data sequence. Instead, to construct μ_t with $t \in \mathbb{N}$, we generate samples $\hat{\boldsymbol{\theta}}_{t,1}, \ldots, \hat{\boldsymbol{\theta}}_{t,k} \in \Theta$ that are consistent with the observed data sequence $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{X} \times \mathcal{Y}$ by interleaving gradient descent steps on the objective L_t with ordinary diffusion

denoising steps (Barba et al., 2024). Then for all $t \in \mathbb{N}$, the experiments choose $\mu_t = \frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{t,i}}$.

For all methods except diffusion, we construct μ_0 as a Gaussian distribution based on an empirical dataset of $n \in \mathbb{N}$ parameters $\{\boldsymbol{\theta}^{(j)}, j \in [n]\} \subseteq \Theta$. Specifically, we set

$$\mu_0 = \mathcal{N}\Big(\bar{\boldsymbol{\theta}},\, \mathrm{diag}(s_1^2, \ldots, s_d^2)\Big)\,,$$

where $d \in \mathbb{N}$ is the dimension of the parameter space, $\bar{\boldsymbol{\theta}}$ denotes the sample mean, and s_i^2 is the sample variance of the i-th coordinate of the dataset. In experiments involving diffusion-based predictions, we instead set

$$\mu_0 = \frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{0,i}},$$

where $\hat{\boldsymbol{\theta}}_{0,1}, \ldots, \hat{\boldsymbol{\theta}}_{0,k}$ are unconditional samples drawn from the diffusion model. Details about the experimental setup are provided in later sections.

2.4 Delayed Construction with Data Splitting and Burn-in

Having introduced prior and sequential likelihood mixing and their dependence on the chosen prior and mixing distributions, we now study a refinement to improve tightness: delaying the start of the confidence sequence by using an initial burn-in phase based on data splitting. Let $t \in \mathbb{N}$ be the total number of observations, $t_0 \in [t]$ be a chosen starting time, and the observations be

$$(\mathbf{x}'_1, \mathbf{y}'_1), \ldots, (\mathbf{x}'_t, \mathbf{y}'_t) \in \mathcal{X} \times \mathcal{Y}.$$

For all $s \in \{1, \ldots, t - t_0 + 1\}$, define

$$\mathbf{x}_s \coloneqq \mathbf{x}'_{t_0-1+s}, \quad \mathbf{y}_s = \mathbf{y}'_{t_0-1+s}.$$

Then use *burn-in* data

$$\mathcal{D}_{\text{burn-in}} \coloneqq (\mathbf{x}'_1, \mathbf{y}'_1), \ldots, (\mathbf{x}'_{t_0-1}, \mathbf{y}'_{t_0-1})$$

to construct prior or mixing distributions, in Theorems 2.3 and 2.5 respectively, that are more concentrated around $\boldsymbol{\theta}^*$ and use the remaining data

$$(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_{t-t_0+1}, \mathbf{y}_{t-t_0+1})$$

to constrain the confidence set through negative log-likelihood L_{t-t_0+1}.

To apply prior likelihood mixing, we may construct prior μ_0 using $\mathcal{D}_{\text{burn-in}}$. To apply sequential likelihood mixing, for all $s \in \{1, \ldots, t - t_0 + 1\}$, we may use

$$(\mathbf{x}'_1, \mathbf{y}'_1), \ldots, (\mathbf{x}_{t_0+s-1}, \mathbf{y}_{t_0+s-1})$$

to construct μ_s.

Choosing a larger t_0 has the beneficial effect that we can condition prior and mixing distributions on more data, but also has the negative effect that we have less data to constrain $C_{t-t_0+1}(\cdot)$ through L_{t-t_0+1}, so it is not obvious which choice of t_0 is optimal.

A guaranteed disadvantage of larger t_0 is that we have to wait until the data sequence $(\mathbf{x}'_1, \mathbf{y}'_1), \ldots, (\mathbf{x}'_{t_0}, \mathbf{y}'_{t_0})$ has been observed before the first confidence set can be constructed.

Our experiments suggest that the optimal choice of t_0 to obtain the tightest confidence set depends not only on t, but also on experimental conditions such as acquisition time $T_a \in (0, \infty)$. Chapter 4 goes into more detail about that.

This concludes our discussion of likelihood-based confidence sequences. We now turn to a detailed account of the neural methods, including their motivation, training, and use for generating predictions.

Chapter 3

Neural Methods

In the Chapter 2, we introduced likelihood-based confidence sequences, discussed their role in uncertainty quantification, and presented two general constructions: prior likelihood mixing in Section 2.2 and sequential likelihood mixing in Theorem 2.5. Both approaches require prior distributions, and in the sequential case a sequence of mixing distributions. The tightness of the resulting confidence sets, measured by the size of the confidence coefficient, depends critically on how strongly these distributions concentrate around the true parameter $\boldsymbol{\theta}^* \in \Theta$.

Neural predictors can be used to construct such distributions. For all steps $t \in \mathbb{N}$, predictors produce candidate estimates

$$\hat{\boldsymbol{\theta}}_{t,1}, \ldots, \hat{\boldsymbol{\theta}}_{t,k} \in \Theta,$$

based on data $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_t, \mathbf{y}_t)$, which, if close to $\boldsymbol{\theta}^*$ in a distance measure such as

$$d_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \coloneqq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,$$

can be used to form concentrated distributions, for instance,

$$\mu_t = \frac{1}{k} \sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{t,i}} \quad \text{or} \quad \mu_t = \delta_{\hat{\boldsymbol{\theta}}_t},$$

where $\hat{\boldsymbol{\theta}}_t = \text{median}(\hat{\boldsymbol{\theta}}_{t,1}, \ldots, \hat{\boldsymbol{\theta}}_{t,k})$ is the coordinate-wise median.

In tomographic imaging, two neural approaches are particularly prominent: deterministic U-Nets that map filtered backprojections directly to ground truth images (Ronneberger et al., 2015; Jin et al., 2017; Kang et al., 2017; Han and Ye, 2018), and guided diffusion models that generate conditional samples from learned image distributions (Ho et al., 2020; Dhariwal and Nichol, 2021; Nichol et al., 2022). Both approaches have been shown to work well for the reconstruction and synthesis of medical images (Kiss et al., 2025; Yang et al., 2023), and we use them as predictors in our confidence sequence constructions.

Since both U-Nets and image-generating diffusion models output images, we assume throughout this chapter that the parameter space is an image domain $\Theta \subseteq \mathbb{R}^{C \times W \times H}$.

3.1 U-Nets

U-Nets are encoder–decoder architectures with skip connections, originally introduced for biomedical image segmentation (Ronneberger et al., 2015). Since then, they have become a standard tool for image-to-image regression, denoising, and inverse problems such as tomography (Jin et al., 2017; Kang et al., 2017; Han and Ye, 2018). Their strength lies in combining global contextual information with local spatial detail through the interaction of the contracting path and skip connections.

3.1.1 General Structure

The *contracting path* (or *encoder*) consists of repeated downsampling and feature extraction steps. Each stage reduces spatial resolution while increasing the number of channels, yielding progressively more abstract features. The final output of the encoder has size $C_{\max} \times W_{\min} \times H_{\min}$ with $W_{\min} \ll W$, $H_{\min} \ll H$, and $C_{\max} \gg C$. This latent representation constitutes the *bottleneck* and encodes the global context extracted from the input image.

The *expanding path* (or *decoder*) then increases the spatial resolution step by step using upsampling operations (e.g., transposed convolutions or interpolation) followed by feature transformations that preserve spatial dimensions. Skip connections link encoder and decoder stages at matching resolutions, allowing the decoder to reuse fine-grained spatial information discarded during downsampling and thereby improve detail reconstruction (Ronneberger et al., 2015). Figure 3.1 illustrates the general structure.

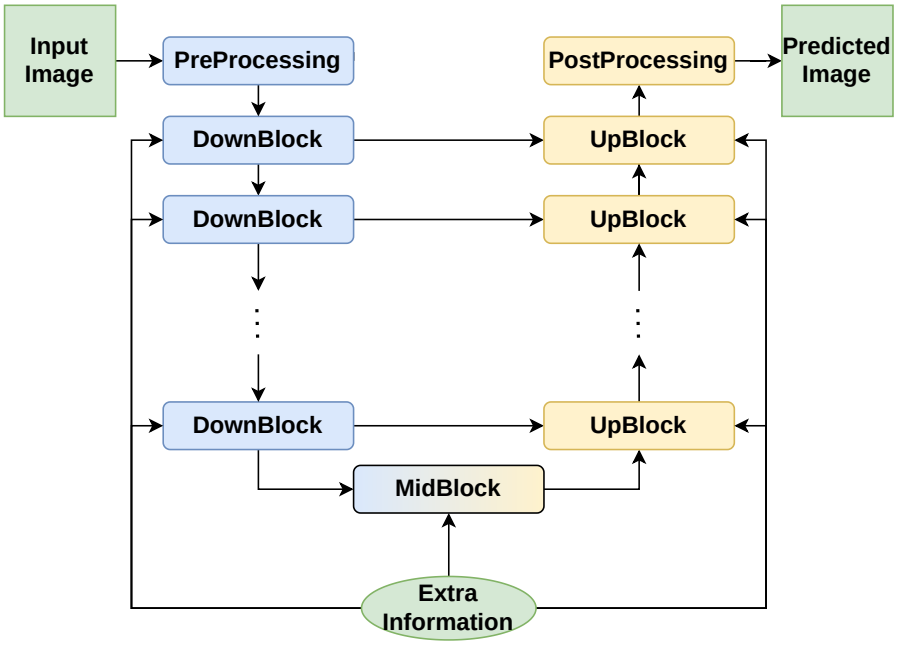


Figure 3.1: General structure of a U-Net. Green shapes denote inputs and outputs. Blue blocks form the contracting path (encoder), yellow blocks the expanding path (decoder). Additional information, if available, may be routed into the network, represented here by the *Extra Information* oval.

Numerous variants of the U-Net architecture have been proposed. *Residual U-Nets* incorporate residual connections to improve gradient flow and training stability (Drozdzal

et al., 2016; Khanna et al., 2020). *Attention U-Nets* augment skip connections with attention gates that selectively pass information from the encoder (Oktay et al., 2022). More recently introduced *Transformer–U-Net hybrids* integrate transformer blocks into the bottleneck to capture long-range dependencies and improve contextual reasoning (Vaswani et al., 2017; Chen et al., 2021, 2024).

In this work, we employ a Transformer–U-Net hybrid as a *post-processing U-Net* (Kiss et al., 2025). The precise architecture is detailed in Chapter 4.

3.1.2 Training

Although many loss functions and training strategies exist in practice, we summarize the standard training pipeline here.

1. Collect a dataset of input image, target image, extra information tuples, e.g. (noisy image, noiseless image, class of the image).
2. Optionally increase dataset size via data augmentation
3. For $N_{\text{epochs}} \in \mathbb{N}$ epochs iterate over each batch of examples. For each batch do:
   (a) Predict the target image based on the input image and auxiliary information (if available).
   (b) Compute a loss, e.g. MSE over all pixels and (predicted image, target image) pairs in the batch.
   (c) Backpropagate the loss to get the gradient w.r.t the networks parameters.
   (d) Update network parameters.
   (e) If a stopping criterion (e.g., patience based) is satisfied, stop training

   .

In our case, we train U-Nets by minimizing the MSE between predictions and ground-truth noiseless activity images. As mentioned earlier, the input images are FBP reconstructions and the only auxiliary information provided is a time step. Chapter 4 covers the exact training procedure.

Although U-Nets provide deterministic image-to-image mappings, diffusion models allow stochastic but steerable image generation. We describe these in the next section.

3.2 Diffusion Models

Diffusion models are a powerful family of deep generative models that generate samples by reversing a noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020). They have recently achieved state-of-the-art (SOTA) performance in including image and audio generation (Nichol and Dhariwal, 2021; Rombach et al., 2022; Esser et al., 2024; Fuest et al., 2024). They furthermore show competitive or SOTA performance on various natural language modeling benchmarks (Fuest et al., 2024; Li et al., 2025). Intuitively, what they do is *denoise* corrupted (or noisy) data sequentially until the clean, noiseless signal is recovered.

As mentioned earlier, in this work we use *Denoising Diffusion Probabilistic Models* (DDPM) from (Ho et al., 2020) to implement an unconditional image distribution sampler. We next describe DDPM.

3.2.1 Roadmap

Our goal is to generate samples from the target image distribution Q_0 on Θ, with density q_0. Let $\boldsymbol{\vartheta}^{(0)} \sim Q_0$ denote the corresponding random variable, and let $\boldsymbol{\theta}^{(0)} \in \Theta$ denote an observation. Since Q_0 is unknown, we approximate it by learning a joint distribution P_ϕ with density p_ϕ over length $T+1 \in \{2, 3, \dots\}$ sequences of random variables

$$\boldsymbol{\vartheta}^{(0:T)} \coloneqq (\boldsymbol{\vartheta}^{(0)}, \boldsymbol{\vartheta}^{(1)}, \dots, \boldsymbol{\vartheta}^{(T)}),$$

where $\boldsymbol{\vartheta}^{(0)}$ corresponds to an image and $\boldsymbol{\vartheta}^{(1:T)}$ are auxiliary latent variables. Sampling from p_ϕ then produces realizations $\boldsymbol{\theta}^{(0:T)} = (\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(T)}) \in \Theta^{T+1}$, and retaining only $\boldsymbol{\theta}^{(0)}$ yields an unconditional image sample from (approximately) Q_0.

In our experiments, we further guide the sampling procedure to enforce consistency with the observed data sequence and thereby obtain predictions for $\boldsymbol{\theta}^*$. Details of this guided sampling approach are deferred to Section 4.1.2.

3.2.2 Forward Process

The forward process is the distribution Q on Θ^{T+1} with density q. For all $T \in \mathbb{N}$, $\boldsymbol{\theta}^{(1:T)} \in \Theta^T$ and $\boldsymbol{\theta}^{(0)} \in \Theta$ the conditional density at $\boldsymbol{\theta}^{(1:T)}$ given $\boldsymbol{\theta}^{(0)}$ is defined such that it factorizes as a Gaussian Markov chain:

$$q(\boldsymbol{\theta}^{(1:T)} \mid \boldsymbol{\theta}^{(0)}) \coloneqq \prod_{\tau=1}^{T} q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(\tau-1)}),$$

with

$$q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(\tau-1)}) \coloneqq \mathcal{N}\big(\boldsymbol{\theta}^{(\tau)}; \sqrt{1-\beta_\tau}\,\boldsymbol{\theta}^{(\tau-1)}, \beta_\tau I\big).$$

Note that the expression $q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(\tau-1)})$ should be understood as a shorthand: while the functional form is Gaussian for all $\tau \in [T]$, the parameters $(\sqrt{1-\beta_\tau}, \beta_\tau)$ depend explicitly on τ, so each conditional density is in general different.

The variance schedule $(\beta_1, \dots, \beta_T) \in (0, \infty)^T$ is typically chosen to increase linearly: for all $\tau \in [T]$,

$$\beta_\tau = \frac{\tau - 1}{T - 1}\beta_T + \left(1 - \frac{\tau - 1}{T - 1}\right)\beta_1,$$

although improved schedules (e.g. cosine) were later introduced (Ho et al., 2020; Nichol and Dhariwal, 2021). The variances $\beta_1, \beta_T \in (0, \infty)$ are treated as hyperparameters.

3.2.3 Reverse Process

The learned reverse process is the distribution P_ϕ on Θ^{T+1} with density p_ϕ. It is defined as a Gaussian Markov chain whose mean function $\mu_\phi : \Theta \times [T] \to \Theta$ is parameterized by a U-Net. For all $T \in \mathbb{N}$, $\tau \in [T]$ and $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(T)} \in \Theta$ they it is defined as

$$p_\phi(\boldsymbol{\theta}^{(0:T)}) \coloneqq p(\boldsymbol{\theta}^{(T)}) \prod_{\tau=1}^{T} p_\phi(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}),$$

with conditional transitions

$$p_\phi(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}) \coloneqq \mathcal{N}(\boldsymbol{\theta}^{(\tau-1)}; \mu_\phi(\boldsymbol{\theta}^{(\tau)}, \tau), \sigma_\tau^2 I),$$

and prior

$$p(\boldsymbol{\theta}^{(T)}) \coloneqq \mathcal{N}(\boldsymbol{\theta}^{(T)}; \mathbf{0}, I).$$

As in the forward process Markov chain, for all $\tau \in [T]$, the shorthand $p_\phi(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)})$ denotes a Gaussian conditional density whose parameters depend on the time index τ through $\mu_\phi(\cdot, \tau)$ and σ_τ^2. Here, $\boldsymbol{\phi} \in \mathbb{R}^{d_{\text{learn}}}$ with $d_{\text{learn}} \in \mathbb{N}$ are learnable parameters. The reverse process variances σ_τ^2 are tied to the forward process schedule β_τ. For all $\tau \in [T]$, two choices are commonly used:

$$\sigma_\tau^2 = \beta_\tau \quad \text{or} \quad \sigma_\tau^2 = \frac{1 - \bar{\alpha}_{\tau-1}}{1 - \bar{\alpha}_\tau} \beta_\tau,$$

with $\bar{\alpha}_\tau \coloneqq \prod_{s=1}^{\tau} \alpha_s$ and $\alpha_\tau \coloneqq 1 - \beta_\tau$. Empirically, this choice has only a minor effect on performance (Ho et al., 2020).

3.2.4 Training Objective

We aim to minimize the expected negative log-likelihood $\mathbf{E}_{\boldsymbol{\vartheta}^{(0)}}[-\log p_\phi(\boldsymbol{\vartheta}^{(0)})]$. In the following, we re-derive the simplified training objective in (Ho et al., 2020).

Let $\boldsymbol{\theta}^{(0)} \in \Theta$ and $T \in \mathbb{N}$. Then

$$p_\phi(\boldsymbol{\theta}^{(0)}) = \mathbf{E}_{\boldsymbol{\vartheta}^{(1:T)}} \left[\frac{p_\phi(\boldsymbol{\theta}^{(0)}, \boldsymbol{\vartheta}^{(1:T)})}{q(\boldsymbol{\vartheta}^{(1:T)} \mid \boldsymbol{\theta}^{(0)})} \,\middle|\, \boldsymbol{\vartheta}^{(0)} = \boldsymbol{\theta}^{(0)} \right].$$

Jensen's inequality and convexity of $-\log$ imply that

$$\begin{aligned}
&-\log p_\phi(\boldsymbol{\theta}^{(0)}) \\
&\quad \leq \mathbf{E}_{\boldsymbol{\vartheta}^{(1:T)}} \left[-\log \frac{p_\phi(\boldsymbol{\theta}^{(0)}, \boldsymbol{\vartheta}^{(1:T)})}{q(\boldsymbol{\vartheta}^{(1:T)} \mid \boldsymbol{\theta}^{(0)})} \,\middle|\, \boldsymbol{\vartheta}^{0)} = \boldsymbol{\theta}^{(0)} \right] \qquad (3.1) \\
&\quad = \mathbf{E}_{\boldsymbol{\vartheta}^{(1:T)}} \left[-\log p(\boldsymbol{\vartheta}^{(T)}) - \sum_{\tau=2}^{T} \log \frac{p_\phi(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)})}{q(\boldsymbol{\vartheta}^{(\tau)} \mid \boldsymbol{\vartheta}^{(\tau-1)})} - \frac{p_\phi(\boldsymbol{\theta}^{(0)} \mid \boldsymbol{\vartheta}^{(1)})}{q(\boldsymbol{\vartheta}^{(1)} \mid \boldsymbol{\theta}^{(0)})} \,\middle|\, \boldsymbol{\vartheta}^{(0)} = \boldsymbol{\theta}^{(0)} \right] \\
&\quad =: \mathcal{L}.
\end{aligned}$$

For all $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(T)} \in \Theta$, define $q(\boldsymbol{\theta}^{(0:T)}) \coloneqq q(\boldsymbol{\theta}^{(0)})\, q(\boldsymbol{\theta}^{(1:T)} \mid \boldsymbol{\theta}^{(0)})$, then taking an expectation over $\boldsymbol{\vartheta}^{(0)}$ on both sides of Equation (3.1) yields the variational upper bound

$$\mathbf{E}_{\boldsymbol{\vartheta}^{(0)}}[-\log p_\phi(\boldsymbol{\vartheta}^{(0)})] \leq \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}} \left[-\log \frac{p_\phi(\boldsymbol{\vartheta}^{(0:T)})}{q(\boldsymbol{\vartheta}^{(1:T)} \mid \boldsymbol{\vartheta}^{(0)})} \right].$$

Next, rewrite $\mathcal{L}$ as (Ho et al., 2020; Sohl-Dickstein et al., 2015):

$$\begin{aligned}
\mathcal{L} &= \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\left[-\log p(\boldsymbol{\vartheta}^{(T)}) - \sum_{\tau=2}^{T} \log \frac{p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)})}{q(\boldsymbol{\vartheta}^{(\tau)} \mid \boldsymbol{\vartheta}^{(\tau-1)})} - \log \frac{p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(0)} \mid \boldsymbol{\vartheta}^{(1)})}{q(\boldsymbol{\vartheta}^{(1)} \mid \boldsymbol{\vartheta}^{(0)})}\right] \\
&= \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\Bigg[-\log p(\boldsymbol{\vartheta}^{(T)}) - \sum_{\tau=2}^{T} \log \frac{p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)})}{q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)})} \\
&\quad - \sum_{\tau=2}^{T} \log \frac{q(\boldsymbol{\vartheta}^{(\tau)} \mid \boldsymbol{\vartheta}^{(0)})}{q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(0)})} - \log \frac{p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(0)} \mid \boldsymbol{\vartheta}^{(1)})}{q(\boldsymbol{\vartheta}^{(1)} \mid \boldsymbol{\vartheta}^{(0)})}\Bigg] \\
&= \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\Bigg[-\log \frac{p(\boldsymbol{\vartheta}^{(T)})}{q(\boldsymbol{\vartheta}^{(T)} \mid \boldsymbol{\vartheta}^{(0)})} \\
&\quad - \sum_{\tau=2}^{T} \log \frac{p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)})}{q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)})} - \log p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(0)} \mid \boldsymbol{\vartheta}^{(1)})\Bigg] \\
&= \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\Bigg[D_{\mathrm{KL}}(q(\boldsymbol{\vartheta}^{(T)} \mid \boldsymbol{\vartheta}^{(0)}) \,\|\, p(\boldsymbol{\vartheta}^{(T)})) \\
&\quad + \sum_{\tau=2}^{T} D_{\mathrm{KL}}(q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}) \,\|\, p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)})) - \log p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(0)} \mid \boldsymbol{\vartheta}^{(1)})\Bigg] \\
&= \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\left[\mathcal{L}_T + \sum_{\tau=2}^{T} \mathcal{L}_{\tau-1} - \mathcal{L}_0\right]
\end{aligned}$$

with for all $\tau \in \{2, \ldots, T\}$

$$\begin{aligned}
\mathcal{L}_T &\coloneqq \mathbf{E}_{\boldsymbol{\vartheta}^{(0)}, \boldsymbol{\vartheta}^{(T)}}\left[D_{\mathrm{KL}}(q(\boldsymbol{\vartheta}^{(T)} \mid \boldsymbol{\vartheta}^{(0)}) \,\|\, p(\boldsymbol{\vartheta}^{(T)}))\right] \\
\mathcal{L}_{\tau-1} &\coloneqq \mathbf{E}_{\boldsymbol{\vartheta}^{(0)}, \boldsymbol{\vartheta}^{(\tau-1)}, \boldsymbol{\vartheta}^{(\tau)}}\left[D_{\mathrm{KL}}(q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}) \,\|\, p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}))\right] \\
\mathcal{L}_0 &\coloneqq \mathbf{E}_{\boldsymbol{\vartheta}^{(0)}, \boldsymbol{\vartheta}^{(1)}}\left[-\log p_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(0)} \mid \boldsymbol{\vartheta}^{(1)})\right].
\end{aligned}$$

Since $\mathcal{L}_T$ does not involve learnable parameters, it can be ignored. For simplicity, we also ignore $\mathcal{L}_0$, although it involves learnable parameters.

Next, we derive the simplified version of $\mathcal{L}_{\tau-1}$, following the steps in (Zhang, 2025).The derivation has been adapted to align with our notation and to explicitly distinguish between random variables and deterministic quantities, with the aim of improving clarity. The first step is to show Lemma 3.1 and then use it to show Lemma 3.2.

Lemma 3.1. *For all $\tau \in \{2, \ldots, T\}$ and all $\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(\tau-1)} \in \Theta$ we have*

$$q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(0)}) = \mathcal{N}(\boldsymbol{\theta}^{(\tau)}; \sqrt{\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(0)},\ (1-\bar{\alpha}_\tau)I).$$

with $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$.

Proof. See Appendix B.5. □

Lemma 3.2. *For all $\tau \in \{2, \ldots, T\}$ and $\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(\tau-1)}, \boldsymbol{\theta}^{(0)} \in \Theta$*

$$q(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}) = \mathcal{N}(\boldsymbol{\theta}^{(\tau-1)}; \tilde{\mu}_\tau(\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}), \tilde{\beta}_\tau I)$$

with

$$\tilde{\mu}_\tau(\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}) = \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau}\boldsymbol{\theta}^{(0)} + \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\boldsymbol{\theta}^{(\tau)},$$
$$\tilde{\beta}_\tau = \frac{1-\bar{\alpha}_{\tau-1}}{1-\bar{\alpha}_\tau}\beta_\tau.$$

Proof. See Appendix B.6. The claim is proven using Lemma 3.1. □

Motivated by Lemma 3.2 we choose $\sigma_\tau^2 = \tilde{\beta}_\tau$. The next lemma significantly simplifies $\mathcal{L}_{\tau-1}$.

Lemma 3.3. *For* $\tau \in \{2, \dots, T\}$

$$\mathcal{L}_{\tau-1} = \frac{1}{2\sigma_\tau^2}\,\mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\boldsymbol{\vartheta}^{(\tau)}}\left[\|\mu_{\boldsymbol{\phi}}(\boldsymbol{\vartheta}^{(\tau)}, \tau) - \tilde{\mu}_\tau(\boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)})\|^2\right] + C,$$

where C *is a constant independent of* $\boldsymbol{\phi}$.

Proof. See Appendix B.7. The claim is proven using Lemma 3.2. □

For all $\tau \in [T]$, define

$$\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)}, \epsilon) \coloneqq \sqrt{\bar{\alpha}_\tau}\boldsymbol{\vartheta}^{(0)} + \sqrt{1-\bar{\alpha}_\tau}\epsilon$$

with $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. Then, Lemma 3.1 implies

$$\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)}, \epsilon) = \boldsymbol{\vartheta}^{(\tau)}. \tag{3.2}$$

Rearranging yields

$$\boldsymbol{\vartheta}^{(0)} = \frac{1}{\sqrt{\bar{\alpha}_\tau}}(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)}, \epsilon) - \sqrt{1-\bar{\alpha}_\tau}\epsilon) \tag{3.3}$$

Plugging Equations (3.2) and (3.3) into the equation in Lemma 3.3 and using the definition

of $\tilde{\mu}_\tau$ yields

$$\begin{aligned}
\mathcal{L}_{\tau-1} \propto \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\Bigg[\frac{1}{2\sigma_\tau^2}\Bigg\|\mu_\phi\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau\right) \\
- \tilde{\mu}_\tau\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon), \frac{1}{\sqrt{\bar{\alpha}_\tau}}(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \sqrt{1-\bar{\alpha}_\tau}\epsilon)\right)\Bigg\|^2\Bigg] \\
= \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\Bigg[\frac{1}{2\sigma_\tau^2}\Bigg\|\mu_\phi\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau\right) \\
- \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau}\frac{1}{\sqrt{\bar{\alpha}_\tau}}(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \sqrt{1-\bar{\alpha}_\tau}\epsilon) - \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon)\Bigg\|^2\Bigg] \\
= \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\Bigg[\frac{1}{2\sigma_\tau^2}\Bigg\|\mu_\phi\Big(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau\Big) \\
- \frac{1}{1-\bar{\alpha}_\tau}\left(\frac{\beta_\tau}{\sqrt{\alpha_\tau}}(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \sqrt{1-\bar{\alpha}_\tau}\,\epsilon) + \sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon)\right)\Bigg\|^2\Bigg] \\
= \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\Bigg[\frac{1}{2\sigma_\tau^2}\Bigg\|\mu_\phi\Big(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau\Big) \\
- \frac{1}{1-\bar{\alpha}_\tau}\left(\Big(\frac{\beta_\tau}{\sqrt{\alpha_\tau}} + \sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})\Big)\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \tfrac{\beta_\tau}{\sqrt{\alpha_\tau}}\sqrt{1-\bar{\alpha}_\tau}\,\epsilon\right)\Bigg\|^2\Bigg] \\
= \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\Bigg[\frac{1}{2\sigma_\tau^2}\Bigg\|\mu_\phi\Big(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau\Big) - \left(\frac{1}{\sqrt{\alpha_\tau}}\,\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \frac{\beta_\tau}{\sqrt{\alpha_\tau}}\cdot\frac{\sqrt{1-\bar{\alpha}_\tau}}{1-\bar{\alpha}_\tau}\,\epsilon\right)\Bigg\|^2\Bigg] \\
= \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\Bigg[\frac{1}{2\sigma_\tau^2}\Bigg\|\mu_\phi\Big(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau\Big) - \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\,\epsilon\right)\Bigg\|^2\Bigg].
\end{aligned}$$

Let $\boldsymbol{\theta}^{(0)} \in \Theta$, $\tau \in \{2,\dots,T\}$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. The last equation implies that

$$\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\theta}^{(0)},\epsilon),\tau)$$

should, in expectation, be close to

$$\frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\theta}^{(0)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\,\epsilon\right)$$

in order to minimize $\mathcal{L}_{\tau-1}$. At the τ-th reverse step the only unknown in it is ϵ, we may instead train our model to predict ϵ based on $(\boldsymbol{\theta}^{(\tau)},\tau)$. To that end, for all $\tau \in [T]$ and $\boldsymbol{\theta}^{(\tau)} \in \Theta$ define

$$\mu_\phi(\boldsymbol{\theta}^{(\tau)},\tau) \coloneqq \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\theta}^{(\tau)} - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon_\phi(\boldsymbol{\theta}^{(\tau)},\tau)\right) \tag{3.4}$$

where ϵ_ϕ is the trainable model that predicts ϵ from $(\boldsymbol{\theta}^{(\tau)},\tau)$. During inference time, to sample $\boldsymbol{\theta}^{(\tau-1)} \sim p_\phi(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)})$ with $\tau \in [T-1]$ we sample $z \sim \mathcal{N}(\mathbf{0}, I)$ and set

$$\boldsymbol{\theta}^{(\tau-1)} = \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\theta}^{(\tau)} - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon_\phi(\boldsymbol{\theta}^{(\tau)},\tau)\right) + \sigma_\tau z.$$

Remember that for all $\tau \in \{2, \dots, T\}$

$$\mathcal{L}_{\tau-1} \propto \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\left[\frac{1}{2\sigma_\tau^2}\left\|\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau) - \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon\right)\right\|^2\right].$$

Now, consider the inner difference. Plugging in Equation (3.4) and simplifying it yields

$$\begin{aligned}
&\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(\tau-1)},\epsilon),\tau) - \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(\tau-1)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon\right)\\
&\quad= \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(\tau-1)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon_\phi\right) - \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(\tau-1)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon\right)\\
&\quad= \frac{\beta_\tau}{\sqrt{\alpha_\tau}\sqrt{1-\bar{\alpha}_\tau}}\left(\epsilon_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(\tau-1)},\epsilon),\tau) - \epsilon\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau) - \frac{1}{\sqrt{\alpha_\tau}}\left(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon) - \frac{\beta_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\epsilon\right)\right\|^2\\
&\quad= \frac{\beta_\tau^2}{\alpha_\tau(1-\bar{\alpha}_\tau)}\left\|\epsilon_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau) - \epsilon\right\|^2. \qquad (3.5)
\end{aligned}$$

Since $\sigma_\tau^2 = \tilde{\beta}_\tau = \frac{1-\bar{\alpha}_{\tau-1}}{1-\bar{\alpha}_\tau}\beta_\tau$, we have

$$\frac{\beta_\tau^2}{2\,\sigma_\tau^2\,\alpha_\tau\,(1-\bar{\alpha}_\tau)} = \frac{\beta_\tau}{2\,\alpha_\tau\,(1-\bar{\alpha}_{\tau-1})}. \qquad (3.6)$$

Substituting Equation (3.5) into $\mathcal{L}_{\tau-1}$ and applying Equation (3.6) yields

$$\mathcal{L}_{\tau-1} \propto \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\left[\frac{\beta_\tau^2}{2\,\sigma_\tau^2\,\alpha_\tau\,(1-\bar{\alpha}_\tau)}\left\|\epsilon_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau) - \epsilon\right\|^2\right].$$

Dropping the scalar weighting factor as in (Ho et al., 2020) and training with the simplified objective

$$\mathcal{L}_{\text{simple}} = \mathbf{E}_{\boldsymbol{\vartheta}^{(0)},\epsilon}\left[\left\|\epsilon_\phi(\boldsymbol{\vartheta}^{(\tau)}(\boldsymbol{\vartheta}^{(0)},\epsilon),\tau) - \epsilon\right\|^2\right]$$

works empirically. In our implementation, we use $\mathcal{L}_{\text{simple}}$ for training.

3.2.5 Training and Unconditional Sampling Algorithms

Algorithms 1 and 2 show the unconditional DDPM training and sampling procedures, respectively (Zhang, 2025).

Algorithm 1 Unconditional Training

repeat
  Sample $\boldsymbol{\theta}^{(0)} \in \Theta$ from dataset $\mathcal{D}$
  Sample $\tau \sim \text{Uniform}(\{1, \ldots, T\})$
  Sample $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$
  Take a gradient descent step on

$$\nabla_{\phi} \left\| \epsilon_{\phi}\left(\sqrt{\bar{\alpha}_{\tau}}\, \boldsymbol{\theta}^{(0)} + \sqrt{1-\bar{\alpha}_{\tau}}\, \epsilon, \tau\right) - \epsilon \right\|^2$$

until converged

Algorithm 2 Unconditional Sampling

Sample $\boldsymbol{\theta}^{(T)} \sim \mathcal{N}(\mathbf{0}, I)$
for $\tau \leftarrow T, \ldots, 1$ **do**
  Sample $z \sim \mathcal{N}(\mathbf{0}, I)$ if $\tau > 1$, else set $z \leftarrow 0$
  Update

$$\boldsymbol{\theta}^{(\tau-1)} \leftarrow \frac{1}{\sqrt{\alpha_{\tau}}} \left(\boldsymbol{\theta}^{(\tau)} - \frac{\beta_{\tau}}{\sqrt{1-\bar{\alpha}_{\tau}}} \epsilon_{\phi}(\boldsymbol{\theta}^{(\tau)}, \tau) \right) + \sigma_{\tau} z$$

end for
return $\boldsymbol{\theta}^{(0)}$

Chapter 4

Experiments

In our experiments, we address three central questions.

1. *Which confidence sequences yield the tightest confidence sets?*
2. *What is their empirical coverage rate, i.e. how often does the true parameter lie within all confidence sets of the sequence?*
3. *How can we visualize uncertainty?*

We start our discussion by describing how we specialize prior likelihood mixing (Theorem 2.3) and sequential likelihood mixing (Theorem 2.5) for evaluation.

4.1 Comparison of Confidence Sequences

The first set of experiments investigates which confidence sequence constructions yield the tightest confidence sets. Since tight confidence sets correspond to low confidence coefficients, we compare the confidence coefficients of the considered constructions for different acquisition times $T_a \in (0, \infty)$.

For *prior likelihood mixing*, we consider both exact and approximate specializations. The approximate ones rely on the Laplace method (Theorem 2.4) to approximate the integral in Theorem 2.3, with priors given by the standard normal distribution or a learned diagonal normal distribution. The exact ones use Dirac deltas or mixtures thereof, each centered on predictions from a guided diffusion model or a U-Net ensemble member.

For *sequential likelihood mixing*, we restrict our attention to exact specializations. Again, the mixing distributions are Dirac deltas or mixtures thereof, centered on diffusion model or U-Net predictions. Table 4.1 provides an overview of all prior and mixing distributions used in the experiments.

We vary two experimental factors: the *acquisition time* per angle and the *starting time* of the confidence sequence (see Section 2.4). In SPECT, a rotating gamma camera head (detector head) collects photon counts in *step-and-shoot mode*, where the detector rotates to a given angle, halts for a specified acquisition time T_a, collects photons, and then proceeds to the next angle. Longer acquisition times generally yield more accurate reconstructions and lower uncertainty, so we evaluate $T_a \in \{1, 100, 10000\}$.

Table 4.1: Prior likelihood mixing distributions μ_{t_0-1} and sequential likelihood mixing distributions μ_{s-1}. For non-diffusion methods, the initial distribution at $s = 0$ is $\mu_0(\boldsymbol{\nu}) = \mathcal{N}(\boldsymbol{\nu}; \bar{\boldsymbol{\theta}}, \operatorname{diag}(s_1^2, \dots, s_{r^2}^2))$ with image dataset sample mean $\bar{\boldsymbol{\theta}}$ and sample variances $s_i \in (0, \infty)$, $i \in [r^2]$. In prior likelihood mixing experiments with diffusion we sample unconditionally (Algorithm 2) if $t_0 = 1$ and L_{t_0-1}-guided (Algorithm 3) if $t_0 \in \{2, 3, \dots\}$. In sequential likelihood mixing experiments with $t_0 = 1$ and diffusion, μ_0 is constructed using unconditional sampling, otherwise we use L_{s-1}-guidance to construct μ_{s-1} with $s \in \{2, 3, \dots\}$. Approximate confidence sequence use rely on Laplace's method (Theorem 2.4).

Specialization	**Shorthand**	**Definition**	**Exact**
		Prior Likelihood Mixing	
Standard normal	P-SN	$\mathcal{N}(\mathbf{0}, I)$	No
Learned normal	P-LN	$\mathcal{N}(\bar{\boldsymbol{\theta}}, \operatorname{diag}(s_1^2, \dots, s_r^2))$	No
U-Net mixture	P-UMix	$\frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{t_0-1,i}}$	Yes
U-Net median	P-UMed	$\delta_{\hat{\boldsymbol{\theta}}_{t_0-1}}$ with $\hat{\boldsymbol{\theta}}_{t_0-1} = \operatorname{median}(\hat{\boldsymbol{\theta}}_{t_0-1,1}, \dots, \hat{\boldsymbol{\theta}}_{t_0-1,k})$	Yes
Diffusion mixture	P-DMix	$\frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{t_0-1,i}}$	Yes
Diffusion median	P-DMed	$\delta_{\hat{\boldsymbol{\theta}}_{t_0-1}}$ with $\hat{\boldsymbol{\theta}}_{t_0-1} = \operatorname{median}(\hat{\boldsymbol{\theta}}_{t_0-1,1}, \dots, \hat{\boldsymbol{\theta}}_{t_0-1,k})$	Yes
		Sequential Likelihood Mixing	
MLE	S-MLE	$\delta_{\hat{\boldsymbol{\theta}}_{s-1}^{\mathrm{MLE}}}$	Yes
MAP	S-MAP	$\delta_{\hat{\boldsymbol{\theta}}_{s-1}^{\mathrm{MAP}}}$, prior $\mathcal{N}(\mathbf{0}, I)$	Yes
U-Net mixture	S-UMix	$\frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{s-1,i}}$	Yes
U-Net median	S-UMed	$\delta_{\hat{\boldsymbol{\theta}}_{s-1}}$ with $\hat{\boldsymbol{\theta}}_{s-1} = \operatorname{median}(\hat{\boldsymbol{\theta}}_{s-1,1}, \dots, \hat{\boldsymbol{\theta}}_{s-1,k})$	Yes
Diffusion mixture	S-DMix	$\frac{1}{k}\sum_{i=1}^{k} \delta_{\hat{\boldsymbol{\theta}}_{s-1,i}}$	Yes
Diffusion median	S-DMed	$\delta_{\hat{\boldsymbol{\theta}}_{s-1}}$ with $\hat{\boldsymbol{\theta}}_{s-1} = \operatorname{median}(\hat{\boldsymbol{\theta}}_{s-1,1}, \dots, \hat{\boldsymbol{\theta}}_{s-1,k})$	Yes

We also investigate the effect of delaying the construction of the confidence sequence by choosing different starting steps t_0. If $t_0 \in \{2, 3, \dots\}$, the first $t_0 - 1$ samples are used only to inform prior or mixing distributions, and no confidence sets are available until t_0, at which point we construct C_{t_0}. From then on we abusively write for all $t_0 \in \mathbb{N}$ and $t \in \mathbb{N}_0$

$$C_{t_0+t}(\beta_{t_0+t}(\delta)) = \left\{ \boldsymbol{\theta} \in \Theta \;\middle|\; -\sum_{s=t_0}^{t_0+t} \log p_s(\mathbf{y}_s \mid \boldsymbol{\theta}) \leq \beta_{t_0+t}(\delta) \right\},$$

with the understanding that the sets are based only on post–burn-in data $(\mathbf{x}_s, \mathbf{y}_s)_{s=t_0}^{t_0+t}$. This notation highlights the trade-off: larger t_0 allows priors and mixing distributions to be more concentrated, but reduces the amount of data available for constraining the confidence sets later on. We therefore consider $t_0 \in \{1, 60, 120\}$.

Overall, with 100 test set images, 12 confidence sequence specializations, 3 acquisition times, and 3 starting steps, we simulate a total of 10 800 confidence sequences.

Since many of the confidence sequence specializations rely on predictions from learned models, we first describe these in more detail. In particular, we focus on post-processing U-Nets and diffusion models. Their performance is critically influenced by several factors, including dataset generation, pre-processing, training procedures, hyperparameter choices, and, in the case of diffusion models, the inference algorithm. We discuss post-processing U-Nets first, before turning to diffusion models.

4.1.1 Post-processing U-Net Architecture

In Section 3.1 a general overview of U-Nets has been presented. We now focus on design choices and implementation details.

In this work, we used the `UNet2DModel` class from the Python library *diffusers* (von Platen et al., 2022), version 0.33.1, to implement Transformer-U-Net hybrids that chain residual and Self-Attention blocks. An illustration of the overall architecture of the *Post-processing U-Nets*[1] is provided in Figure 4.1, while Figure 4.2 details the individual blocks.

The FBP first gets processed by a `Conv2d` block which represents a 2D convolution layer (Fukushima, 2013; Schmidhuber, 2015).[2]

Then several `AttnDownBlocks` process the resulting image, together with a time embedding.[3] The `AttnDownBlocks` apply three `ResnetBlock2D` and `Attention` blocks in an interleaving fashion. After that a strided convolution with stride 2 is applied, cutting spatial dimensions by half. Intermediate hidden states are stored for later use by its corresponding `AttnUpBlock`. See `Skip 1`, ..., `Skip 4` in Figure 4.2a and Figure 4.2b. The `Self-Attention` block in Figure 4.2e dynamically selects one of the three *scaled dot product attention* implementations, depending on what it considers optimal for the workload (Paszke et al., 2017; Lefaudeux et al., 2022; Dao, 2023). `GroupNorm`, `SiLU` and `Dropout` blocks correspond to *Group Normalization* (Wu and He, 2018), *SiLU* activation function (Hendrycks and Gimpel, 2023) and *Dropout* layers (Hinton et al., 2012),

[1] Adopting the terminology in (Kiss et al., 2025) we call our U-Nets *post-processing* U-Nets since they post-process FBP reconstructions.

[2] All `Conv2d` blocks have a kernel size of 3, stride 1 and 1 pixel zero padding and preserve spatial dimensions.

[3] The advantage of giving the network time information is that it informs it about the quality of the FBP reconstruction.

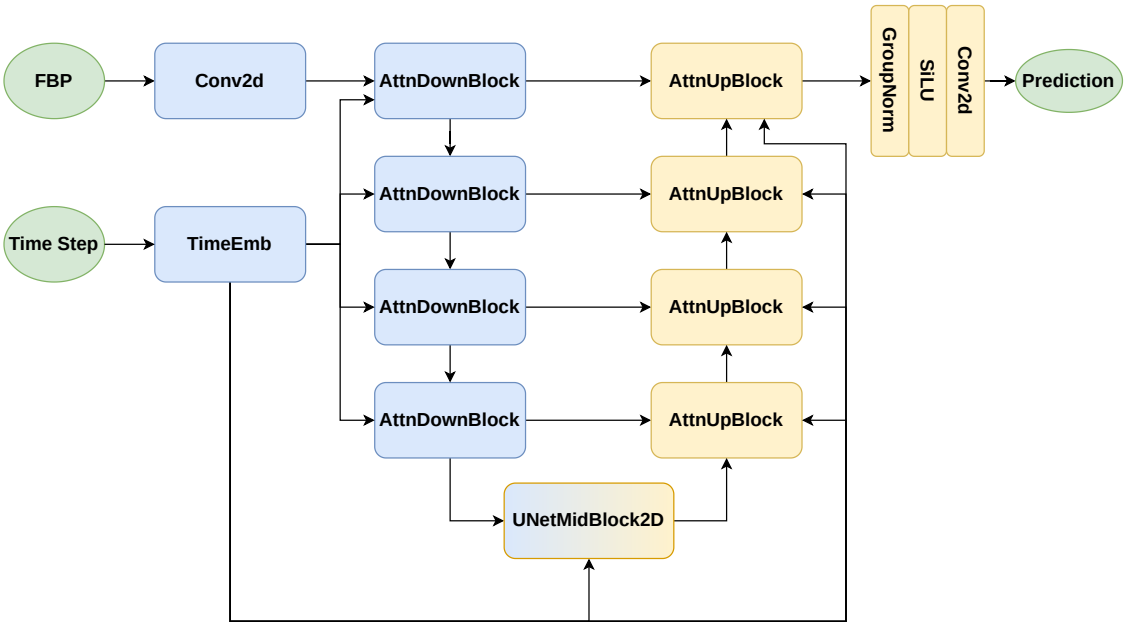
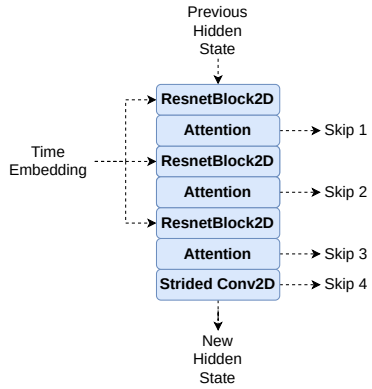
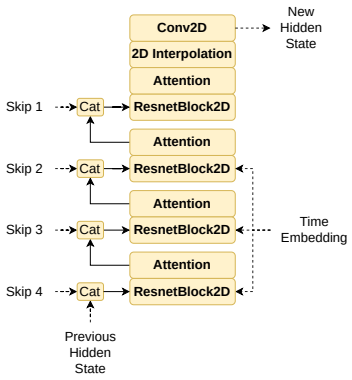


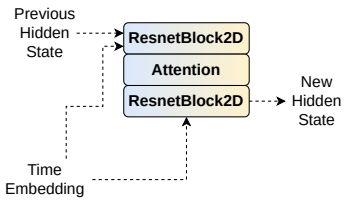
Figure 4.1: Post-processing U-Net architecture overview.



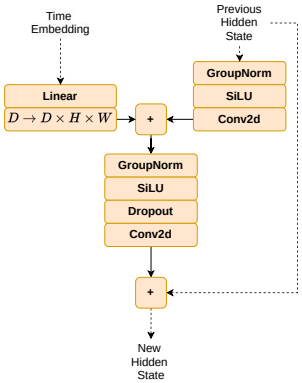
(a) AttnDownBlock.



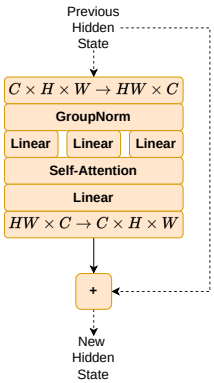
(b) AttnUpBlock.



(c) UNetMidBlock2D.



(d) ResnetBlock2D.



(e) Attention.

Figure 4.2: Detailed view of U-Net blocks: AttnDownBlock and AttnUpBlock (top row), UNetMidBlock2D, ResnetBlock2D, and Attention (bottom row).

respectively. $D \to D \times H \times W$ blocks expand D dimensional vectors into a $D \times H \times W$ dimensional space by repeating its coordinates across them. $C \times H \times W \to HW \times C$ blocks unroll dimensions H and W into a single dimension and transposes the result afterward. $HW \times C \to C \times H \times W$ blocks do the inverse operation.

`UNetMidBlock2D` takes a time embedding and hidden state produced by the last `AttnDownBlock` as input, processes it using an `Attention` block and two `ResnetBlock2D`. The result is a hidden state that encodes global information about the FBP reconstruction.

Afterwards, four `AttnUpBlock`s repeatedly upsample and enrich this hidden state with information encoded in intermediate hidden states from adjacent `AttnDownBlock`s. Lastly, a `GroupNorm`, `SiLU` and `Conv2d` layer gets applied, which yields the final image prediction.

Training

We first split the dataset of 1000 square 64×64 greyscale chip images $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{1000}^* \in \Theta = [0,1]^{64^2}$ into a 900 images training set $\mathcal{D}_{\text{train}} \coloneqq \{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{900}^*\}$ and a 100 images test set $\mathcal{D}_{\text{test}} \coloneqq \{\boldsymbol{\theta}_{901}^*, \ldots, \boldsymbol{\theta}_{1000}^*\}$.[4] Then, split the training set into 9 folds $F_1, \ldots, F_9$ with fold F_j, $j \in [9]$ consisting of a training and validation set, symbolically fold $F_j \coloneqq (\mathcal{D}_{\text{train}}^{(j)}, \mathcal{D}_{\text{val}}^{(j)})$. Define the index set

$$I_j \coloneqq \{100(j-1)+1, \ldots, 100(j-1)+100\}$$

for all $j \in [9]$. Then the training set indices $[900] = I_1 \cup \cdots \cup I_9$. Next, define

$$\mathcal{D}_{\text{train}}^{(j)} \coloneqq \left\{ \boldsymbol{\theta}_i^* \in \mathcal{D}_{\text{train}} \mid i \in \bigcup_{l \neq j} I_l \right\} \text{ and } \mathcal{D}_{\text{val}}^{(j)} \coloneqq \{\boldsymbol{\theta}_i^* \in \mathcal{D}_{\text{val}} \mid i \in I_j\}.$$

We train 27 U-Nets in total; one each fold $F_j \in \{F_1, \ldots, F_9\}$ and acquisition time $T_a \in \{1, 100, 10000\}$.

Consider some fixed F_j, $j \in [9]$ and acquisition time $T_a \in \{1, 100, 10000\}$. We train the corresponding U-Net on $\mathcal{D}_{\text{train}}^{(j)}$ and pick the best checkpoint based on its performance on $\mathcal{D}_{\text{val}}^{(j)}$. The loss function and performance measure is the coordinate-wise mean squared error (MSE). For all ground-truth images $\boldsymbol{\theta}^* \in \Theta = [0,1]^{64 \times 64}$ and predictions $\hat{\boldsymbol{\theta}} \in \Theta$, define the mean squared error (MSE) as

$$\text{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \coloneqq \frac{1}{64^2} \sum_{i=1}^{64} \sum_{j=1}^{64} (\hat{\theta}_{i,j} - \theta_{i,j}^*)^2.$$

For all batches of predictions $\hat{\boldsymbol{\theta}}_{1:n} \in \Theta^n$ with corresponding targets $\boldsymbol{\theta}_{1:n}^* \in \Theta^n$, we extend this by averaging over images:

$$\text{MSE}(\hat{\boldsymbol{\theta}}_{1:n}, \boldsymbol{\theta}_{1:n}^*) \coloneqq \frac{1}{n} \sum_{k=1}^{n} \text{MSE}(\hat{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*).$$

In order to make a prediction, the network takes a filtered backprojection, computed using the Shepp-Logan filter, and a time step as input. FBPs are in turn computed based on a sinogram

$$\mathcal{S}_t = (\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_t}) \in \mathbb{N}_0^{t \times 64}$$

[4] In this chapter, we occasionally treat images as elements of $[0,1]^{64 \times 64}$ and at other times as elements of $[0,1]^{64^2}$, depending on whether spatial structure or vectorized form is more convenient.

and corresponding angles $x_1, \ldots, x_t$ with $x_{i_1} \leq \cdots \leq x_{i_t}$.

We simulate ten data sequences for each image. Denote the h-th data sequence, with $h \in [10]$, corresponding to the i-th image, with $i \in [800]$, in $\mathcal{D}^{(j)}_{\text{train}}$ as $(x_{h,s}, \mathbf{y}_{j,i,h,s})_{s=1}^{180}$ and its corresponding FBP as $\boldsymbol{\theta}^{\text{FBP}}_{j,i,h} \in \Theta$. Denote the i-th image in $\mathcal{D}^{(j)}_{\text{train}}$ as $\boldsymbol{\theta}^*_{j,i}$. Then

$$\mathbf{y}_{j,i,h,s} \sim \text{Pois}\left(\mathcal{R}(\boldsymbol{\theta}^*_{j,i}, x_{h,s})\right).$$

To ensure different angles $x_{\cdot,s}$ and photon counts $\mathbf{y}_{j,i,\cdot,s}$ across data sequences, we use a different random seed for each one, i.e. the random seed for the data sequence $h \in [10]$ is h.

This setup results in 1440000 different FBP-target image pairs for each fold F_j, $j \in [9]$; one for each combination of data sequence index $h \in [10]$, training image index $i \in [800]$ and acquisition step $t \in [180]$.

We train for two epochs with AdamW, parameterized with an initial learning rate of 0.0003 and a weight decay factor of 0.042 (Loshchilov and Hutter, 2019). We use a batch size of 32 and a dropout probability of 0.29. The dimensionality of time embeddings is 32. As we have grayscale images, the input FBPs have shape $1 \times 64 \times 64$. The images that the `AttnDownBlock`s output have shape $64 \times 32 \times 32$, $128 \times 16 \times 16$, $256 \times 8 \times 8$ and $512 \times 4 \times 4$.

To reduce gradient variance with respect to neural network parameters, we randomly sample (FBP, acquisition step, target image) batches in a stratified manner. Fix a permutation $\pi^1 := (i_{1,1}, \ldots, i_{1,8000})$ of $[8000]$, and for all $k \in [8000]$ fix a permutation $\pi^{2,k} := (i_{2,k,1}, \ldots, i_{2,k,180})$ of $[180]$. These permutations are drawn uniformly at random once and then treated as fixed throughout the following. Define

$$\boldsymbol{\pi}^1 := (\underbrace{\pi^1_1, \ldots \pi^1_{8000}, \ldots, \pi^1_1, \ldots, \pi^1_{8000}}_{180 \text{ times}}) \in [8000]^{1440000}$$
$$\boldsymbol{\pi}^2 := (\pi^{2,1}_1, \pi^{2,2}_1, \ldots, \pi^{2,8000}_1, \pi^{2,1}_2, \ldots, \pi^{2,8000}_{179}, \pi^{2,1}_{180}, \pi^{2,2}_{180}, \ldots, \pi^{2,8000}_{180}) \in [180]^{1440000}.$$

For all batch indices $l \in [1440000/32] = [45000]$, we use $\boldsymbol{\pi}^{1,l} := \boldsymbol{\pi}^1_{l:l+32}$ and $\boldsymbol{\pi}^{2,l} := \boldsymbol{\pi}^2_{l:l+32}$. Hence, batch $l \in [45000]$ consists of images

$$\boldsymbol{\theta}^*_{j,\mathbf{i}^l} := \left(\boldsymbol{\theta}^*_{j,i^l_1}, \ldots, \boldsymbol{\theta}^*_{j,i^l_{32}}\right) \in \Theta^{32}$$

with

$$\mathbf{i}^l := ((\pi^{1,l}_1 - 1 \mod 800) + 1, \ldots, (\pi^{1,l}_{32} - 1 \mod 800) + 1) \in [800],$$

and FBPs

$$\boldsymbol{\theta}^{\text{FBP}}_{j,\mathbf{i}^l,\mathbf{h}^l} := \left(\boldsymbol{\theta}^{\text{FBP}}_{j,i^l_1,h^l_1}, \ldots \boldsymbol{\theta}^{\text{FBP}}_{j,i^l_1,h^l_{32}}, \ldots, \boldsymbol{\theta}^{\text{FBP}}_{j,i^l_2,h^l_1}, \ldots, \boldsymbol{\theta}^{\text{FBP}}_{j,i^l_{32},h^l_{32}}\right) \in \Theta^{32}$$

with

$$\mathbf{h}^l := (\lceil \boldsymbol{\pi}^{1,l}_1 / 800 \rceil, \ldots, \lceil \boldsymbol{\pi}^{1,l}_{32} / 800 \rceil) \in [180]$$

and acquisition steps $\boldsymbol{\pi}^{2,l} \in [180]$.

In addition, we employ the *cosine schedule with warmup* learning rate schedule from the diffusers library (von Platen et al., 2022). For the first 10% of training steps it linearly increases the learning rate from 0 to 0.0003. Afterwards, it drops following a half-cosine curve back to 0 over the remaining steps.

The PSNR of U-Net ensemble predictions is summarized in Figure 4.3. In addition to peak signal-to-noise ratio (PSNR), we report several complementary metrics in Table 4.3 and Appendix C.2. For all ground-truth images $\boldsymbol{\theta}^* \in [0,1]^{H \times W}$ and predictions $\hat{\boldsymbol{\theta}} \in [0,1]^{H \times W}$, let $\mathrm{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ be defined as above. Then

$$\begin{aligned}
\mathrm{PSNR}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &\coloneqq 10 \cdot \log_{10}\left(\frac{1}{\mathrm{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)}\right), \\
\mathrm{RMSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &\coloneqq \sqrt{\mathrm{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)}, \\
\mathrm{L1}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &\coloneqq \frac{1}{HW} \sum_{i,j \in [H] \times [W]} |\hat{\theta}_{i,j} - \theta^*_{i,j}|, \\
\mathrm{ZeroOne}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &\coloneqq \frac{1}{HW} \sum_{i,j \in [H] \times [W]} \mathbf{1}\left\{\theta^*_{i,j} \neq \mathrm{round}(\hat{\theta}_{i,j})\right\}.
\end{aligned}$$

These metrics capture pixel-wise error (RMSE, ℓ^1), binary misclassification (ZeroOne), and perceptual similarity (SS) in addition to PSNR. For structural similarity (SS), we use the function `skimage.metrics.structural_similarity` from the *scikit-image* Python library (version 0.25.2) with `data_range=1`, following the formulation of Wang et al. (2004); see also Wang and Bovik (2009) for a broader discussion of perceptual fidelity measures.

Next, we turn to our implementation of guided diffusion.

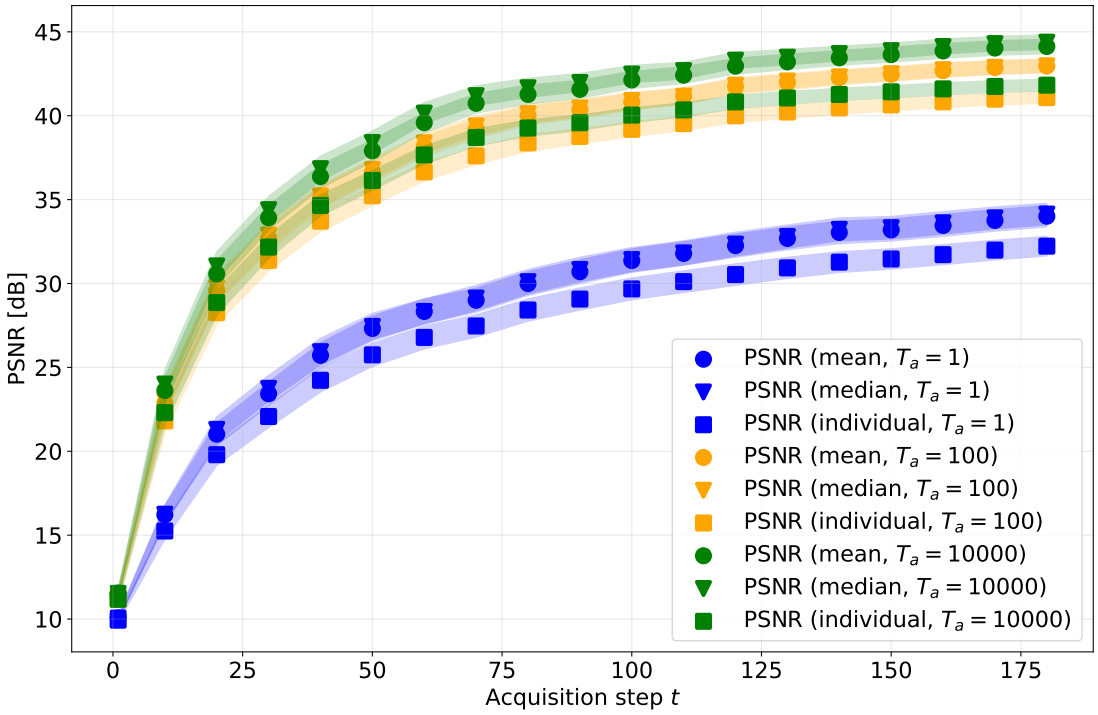


Figure 4.3: PSNR of mean, median as well as individual U-Net ensemble member predictions against acquisition steps. Acquisition steps $t \in \mathbb{N}$ corresponds to data sequence $((x_s, \mathbf{y}_s), s \in [t])$ being used.

4.1.2 L_t-Guided Diffusion

Following Section 3.2, we train an unconditional DDPM to capture the prior over images. Algorithm 1 describes the training procedure. Algorithm 2 is designed to yield samples from the underlying distribution of images. Importantly, this algorithm does not consider observed data $((\mathbf{x}_s, \mathbf{y}_s), s \in [t])$, $t \in \mathbb{N}$. To use the observed data, we employ Algorithm 3 which interweaves denoising with *consistency steps*.

Algorithm 3 L_t-Guided Sampling

Require: Denoiser ϵ_ϕ, number of outer steps $N \in \mathbb{N}$, number of inner steps $k \in \mathbb{N}$, time indices $\tau_1, \ldots, \tau_N \in [T]$, learning rates $\eta_1, \ldots, \eta_k \in (0, \infty)$, data sequence $(x_1, \mathbf{y}_1), \ldots, (x_t, \mathbf{y}_t) \in \mathcal{X} \times \mathcal{Y}$, guidance interval $g \in [N]$, number of cleanup steps $c_t \in \{0, \ldots, N\}$

Ensure: Data- and prior-consistent reconstruction $\boldsymbol{\theta}^{(\tau_{N+1})} \in \Theta$

Initialize $\boldsymbol{\theta}^{(\tau_1)} \sim \mathcal{N}(\mathbf{0}, I)$
for $n \leftarrow 1, \ldots, N$ **do**
  Sample $z \sim \mathcal{N}(\mathbf{0}, I)$ if $n < N$, else set $z \leftarrow 0$
  Set

$$\boldsymbol{\mu}_0^{(\tau_n)} \leftarrow \frac{1}{\sqrt{\alpha_{\tau_n}}} \left(\boldsymbol{\theta}^{(\tau_n)} - \frac{\beta_{\tau_n}}{\sqrt{1 - \bar{\alpha}_{\tau_n}}} \epsilon_\phi(\boldsymbol{\theta}^{(\tau_n)}, \tau_n) \right)$$

  if $n > 1$ and $n \mod g = 0$ and $n \leq N - c_t$ **then**
    Initialize optimizer: $\texttt{optim} \leftarrow \texttt{init}\left(\boldsymbol{\mu}_0^{(\tau_n)}, \eta_0\right)$
    for $i \leftarrow 1, \ldots, k$ **do**
      Take consistency step $\boldsymbol{\mu}_i^{(\tau_n)} \leftarrow \texttt{step}\left(\texttt{optim}, \eta_i, \boldsymbol{\mu}_{i-1}^{(\tau_n)}, (x_s, \mathbf{y}_s)_{s=1}^t\right)$
    end for
  else
    $\boldsymbol{\mu}_k^{(\tau_n)} \leftarrow \boldsymbol{\mu}_0^{(\tau_n)}$
  end if
  Update $\boldsymbol{\theta}^{(\tau_{n+1})} \leftarrow \boldsymbol{\mu}_k^{(\tau_n)} + \sigma_{\tau_n} z$
end for
return $\boldsymbol{\theta}^{(\tau_{N+1})}$

Algorithm 3 differs from Algorithm 2 in multiple ways. One difference is that the outer loop iterates from τ_1 to τ_N where $T \approx \tau_1 > \tau_2 > \cdots > \tau_N = 0$ and $N \ll T$, $N \in \mathbb{N}$. This speeds up inference by reducing the number of denoising and consistency steps. The second difference is the existence of the inner loop. In it, we optimize the current prediction of the mean $\boldsymbol{\mu}_0^{(\tau_n)}$ towards an image that has a reduced negative log-likelihood L_t, $\boldsymbol{\mu}_k^{(\tau_n)}$, thereby making it more consistent with the observed data sequence $((x_s, \mathbf{y}_s), s \in [t])$. In our experiments, we choose $T = 1000$ in Algorithm 1 and $N = 100$, $k = 50$ and a constant learning rate of $\eta_i = 0.05$ for all $i \in [k]$ for Algorithm 3.

Consistency steps use the entire available sinogram (all acquired angles and detector bins). For all $t \in \mathbb{N}$, outer iterations $n \in [N]$, inner iterations $i \in [k]$, iterates $\boldsymbol{\mu}_{i-1}^{(\tau_n)} \in \Theta$, measured sinograms

$$\mathcal{S}_t \coloneqq (\mathbf{y}_s, s \in [t]) \in \mathbb{N}_0^{t \times 64},$$

predicted sinogram

$$\hat{\mathcal{S}}_t \coloneqq (\hat{\mathbf{y}}_s, s \in [t]) = (\mathcal{R}(\boldsymbol{\mu}_{i-1}^{(\tau_n)}, x_1), \ldots, \mathcal{R}(\boldsymbol{\mu}_{i-1}^{(\tau_n)}, x_t)) \in [0, \infty)^{t \times 64},$$

and angles $x_1, \dots, x_t \in [0, 180]$, the objective that

$$\operatorname{step}\left(\text{optim}, \eta_i, \boldsymbol{\mu}_{i-1}^{(\tau_n)}, ((x_s, \mathbf{y}_s), s \in [t])\right)$$

minimizes is

$$\mathcal{L}_t(\boldsymbol{\theta}) = \frac{1}{t \cdot 64} \sum_{s=1}^{t} \sum_{d=1}^{64} \Big(\hat{y}_{s,d} \; - \; y_{s,d} \log(\hat{y}_{s,d} + \varepsilon) \; + \; \log \Gamma(y_{s,d} + 1) \Big),$$

where $\varepsilon = 10^{-8}$ is a constant for numerical stability. $\mathcal{L}_t$ is essentially a numerically stabilized and averaged Poisson negative log-likelihood.[5] Dividing by the number of detector bins (64) and step (t) ensures that we can use the same learning rate across resolutions and acquisition steps. If $t = 0$ we do not use guidance, i.e. $\mathcal{L}_t(\cdot) = 0$. To take consistency steps we use the Adam optimizer (Kingma and Ba, 2017), so $\texttt{init}\left(\boldsymbol{\mu}_0^{(\tau_n)}, \eta_0\right)$ returns a corresponding initialized Adam optimizer instance. For stability `step` clamps pixel values to $[0, 1]$, predicted sinogram values to $[0, \infty)$, replaces NaNs in gradients with zero and afterwards clips the by projecting them onto the ℓ^2 ball with radius 1. We choose the number of cleanup steps $c \in [N]$ heuristically:

$$c_t \coloneqq \lfloor \max(0, 1 - t/180) \cdot 5 \rfloor + 5.$$

Intuitively, the more measurements we have (i.e. the larger t is), the fewer cleanup steps we need since gradient steps w.r.t. L_t inject less noise into the reverse process. So for larger t we need to clean up less noise at the end of the reverse process, justifying a lower number of cleanup steps.

For simplicity we did not include normalization and denormalization operations in Algorithm 3. Normalizing images to $[-1, 1]^{64\times64}$ is however important in practice since it increases training stability. For this reason, ϵ_ϕ is trained on normalized images in $[-1, 1]^{64\times64}$ in our experiments. This necessitates `step` to denormalize and clip $\boldsymbol{\mu}_{i-1}^{(\tau_n)}$ to $[0, 1]^{64\times64}$ before computing $\hat{\mathcal{S}}_t$. After performing the gradient step, `step` also needs to normalize the updated the image back to $\mathbb{R}^{64\times64}$ such that the next denoising step works as expected. Finally, before returning $\boldsymbol{\theta}^{(\tau_{N+1})}$, denormalization and clipping have to be performed as well.

The noise predictor ϵ_ϕ is a U-Net trained with learning rate 0.0025, batch size 32, dropout 0.37, and weight decay 0.0043 for 500 epochs. We implement it with `diffusers`' `UNet2DModel`, but with an architecture distinct from the post-processing U-Nets (see Figures 4.4 and 4.5).

4.1.3 MLE and MAP Estimates

To approximate the maximum likelihood and maximum a posteriori estimates introduced in Section 2.3.1, we solve the optimization problems

$$\begin{aligned} \boldsymbol{\theta}_t^{\text{MLE}} &\in \arg\min_{\boldsymbol{\theta} \in \Theta} L_t(\boldsymbol{\theta}), \\ \boldsymbol{\theta}_t^{\text{MAP}} &\in \arg\min_{\boldsymbol{\theta} \in \Theta} \big(L_t(\boldsymbol{\theta}) - \log f_0(\boldsymbol{\theta})\big), \end{aligned}$$

where f_0 denotes the density of the prior distribution μ_0. In practice, we obtain approximations $\hat{\boldsymbol{\theta}}_t^{\text{MLE}} \approx \boldsymbol{\theta}_t^{\text{MLE}}$ and $\hat{\boldsymbol{\theta}}_t^{\text{MAP}} \approx \boldsymbol{\theta}_t^{\text{MAP}}$ using Algorithm 4. We generally parameterize it with maxSteps $= 20000$ and patience $= 100$.

[5] The $\log \Gamma$ term may be dropped since it does not depend on $\boldsymbol{\mu}_{i-1}^{(\tau_n)}$.

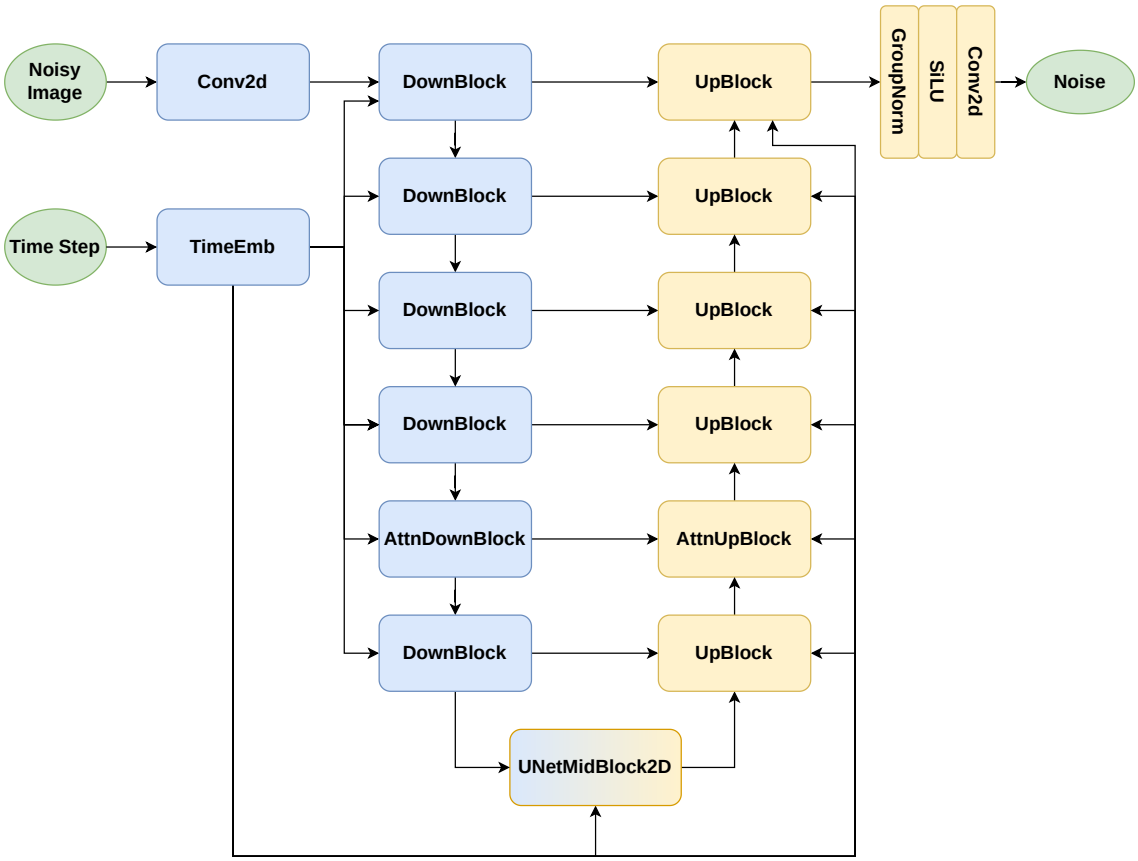
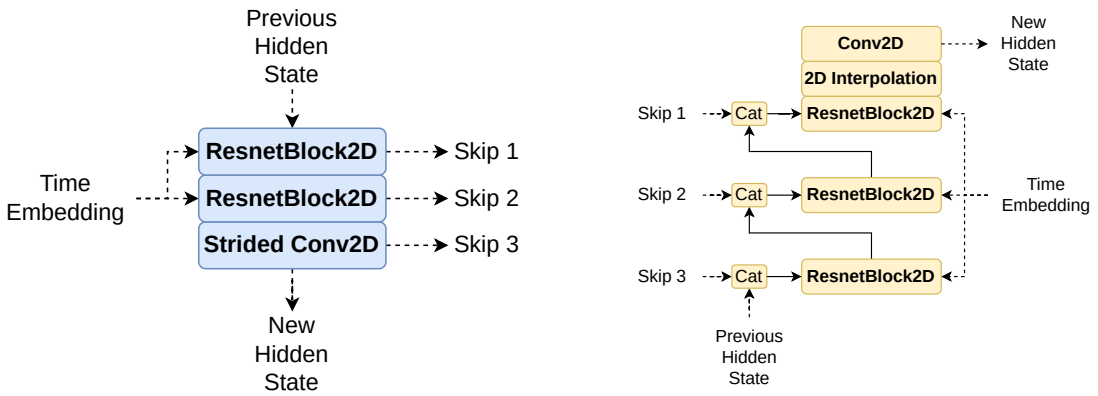


Figure 4.4: Architectural overview of DDPM noise predictor ϵ_ϕ.



(a) DownBlock. (b) UpBlock.

Figure 4.5: `DownBlock` and `UpBlock` blocks used in DDPM noise predictor ϵ_ϕ. Other blocks can be found in Figure 4.2.

4.1.4 Results

As mentioned in the beginning of this section, we compare confidence sequences via their confidence coefficients $\beta_t(\delta)$ with $t \in \mathbb{N}$ and $\delta \in (0,1)$.

Let error level $\delta \in (0,1)$, acquisition step $t \in \mathbb{N}$ and confidence coefficients $\beta_{t,1}(\delta), \beta_{t,2}(\delta) \in (0,\infty)$ with same starting step $t_0 \in \mathbb{N}$, $t \geq t_0$ and acquisition time $T_a \in (0,\infty)$. If $\beta_{t,1}(\delta) < \beta_{t,2}(\delta)$, the confidence set corresponding to $\beta_{t,1}(\delta)$ is smaller than the one corresponding to $\beta_{t,2}(\delta)$, that is, $C_t(\beta_{t,1}(\delta)) \subset C_t(\beta_{t,2}(\delta))$.

For all error levels $\delta \in (0,1)$, acquisition steps $t \in \mathbb{N}$, and confidence coefficients $\beta_t(\delta) \in (0,\infty)$, define *normalized confidence coefficient*

$$\tilde{\beta}_t(\delta) := \frac{\beta_t(\delta) - L_t(\boldsymbol{\theta}^*)}{L_t(\boldsymbol{\theta}^*)}.$$

Since for all data sequences $(x_1, \mathbf{y}_1), \ldots, (x_t, \mathbf{y}_t) \in \mathcal{X} \times \mathcal{Y}$, the negative log-likelihood satisfies $L_t(\boldsymbol{\theta}^*) > 0$ we have

$$\begin{aligned} \beta_{t,1}(\delta) < \beta_{t,2}(\delta) &\iff \beta_{t,1}(\delta) - L_t(\boldsymbol{\theta}^*) < \beta_{t,2}(\delta) - L_t(\boldsymbol{\theta}^*) \\ &\iff \frac{\beta_{t,1}(\delta) - L_t(\boldsymbol{\theta}^*)}{L_t(\boldsymbol{\theta}^*)} < \frac{\beta_{t,2}(\delta) - L_t(\boldsymbol{\theta}^*)}{L_t(\boldsymbol{\theta}^*)} \\ &\iff \tilde{\beta}_{t,1}(\delta) < \tilde{\beta}_{t,2}(\delta). \end{aligned}$$

In words, smaller normalized confidence coefficients correspond to smaller confidence sets. However, normalized confidence coefficients are more interpretable than their raw counterparts since $\tilde{\beta}_t(\delta) = 0$ is always optimal[6] and we can check whether it approaches zero as we increase t and does so quickly. If $\tilde{\beta}_t(\delta)$ approaches zero quickly we can infer that $C_t(\beta_t(\delta))$ tightens quickly too.

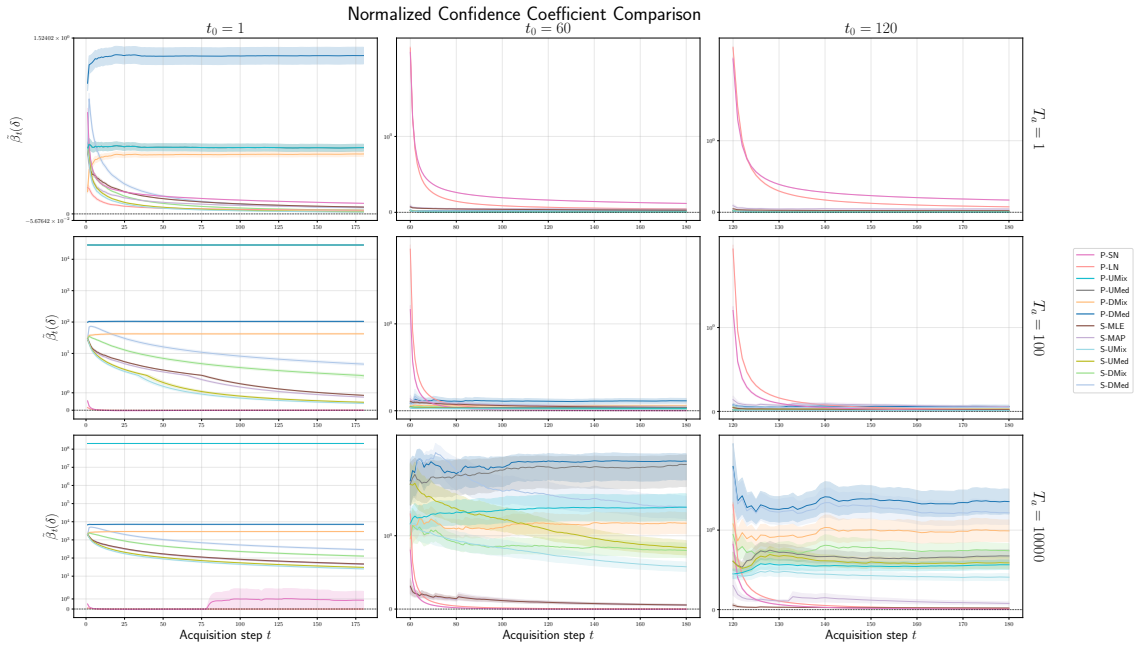
Table 4.2: Average confidence coefficient by method and configuration. Overall minimal average confidence coefficients are **bold**. Additionally minimal *exact* average confidence coefficients are **bold** and marked with an asterisk (*****).

	$T_a = 1$			$T_a = 100$			$T_a = 10000$		
Method	$t_0 = 1$	$t_0 = 60$	$t_0 = 120$	$t_0 = 1$	$t_0 = 60$	$t_0 = 120$	$t_0 = 1$	$t_0 = 60$	$t_0 = 120$
P-SN	17132	11938	6437	**28486**	**19460**	10113	74758	**28401**	**14661**
P-LN	15964	11040	5919	28614	19635	10314	**41942**	28627	14889
P-UMix	24218	10418	5277	8.2×10^8	19882	9929	8.6×10^{12}	67413	22364
P-UMed	24218	10432	5279	8.2×10^8	20024	9949	8.6×10^{12}	83241	23957
P-DMix	23206	10466	5296	1.2×10^6	20417	10125	1.2×10^8	60974	28451
P-DMed	35978	10521	5312	3.0×10^6	21460	10394	3.0×10^8	85265	33723
S-MLE	16674	10707	5388	74284	20243	10054	3.8×10^6	30561	**14682***
S-MAP	16502	10814	5506	68706	20361	10507	3.6×10^6	**30557***	15830
S-UMix	**15790**	**10395**	**5276**	**50285***	**19640***	**9919**	**2.1×10^6***	47709	20391
S-UMed	15864	10401	5276	53640	19732	9943	2.5×10^6	56492	22883
S-DMix	16164	10454	5311	1.3×10^5	19957	10079	1.0×10^7	51741	25050
S-DMed	16674	10479	5327	2.7×10^5	20550	10258	2.3×10^7	70609	32092

Figure 4.6 displays normalized confidence coefficients $\tilde{\beta}_t(\delta)$ over acquisition steps for all considered confidence sequences. Shaded bands around the curves indicate the standard

[6] Here, optimal means that $\beta_t(\delta)$ is as close as possible to $L_t(\boldsymbol{\theta}^*)$ while satisfying $\boldsymbol{\theta}^* \in C_t(\beta_t(\delta))$. In this case, $C_t(\beta_t(\delta))$ is the tightest likelihood-based confidence set containing $\boldsymbol{\theta}^*$, given that L_t depends on $((\mathbf{x}_s, \mathbf{y}_s), s \in [t])$.

Figure 4.6: Normalized confidence coefficients $\tilde{\beta}_t(\delta)$ across different methods, acquisition times, and starting steps. Shaded regions denote standard error of the mean (SEM). The vertical axis is on a logarithmic scale.



error of the mean (SEM) at each acquisition step. Complementary plots of raw coefficients $\beta_t(\delta)$, their differences $\beta_t(\delta) - L_t(\boldsymbol{\theta}^*)$, and normalized coefficients are provided in Appendix C.1.

For easier comparison between methods and configurations, Table 4.2 reports average confidence coefficients, aggregated over test set images and steps, for all combinations of acquisition time, starting step, and method. For low acquisition time ($T_a = 1$), sequential likelihood mixing with U-Net ensembles (S-UMix) yields the smallest average coefficients. At medium acquisition time ($T_a = 100$), prior likelihood mixing with a standard normal prior (P-SN) achieves the lowest coefficients overall, while among the exact methods, S-UMix performs best. For high acquisition time ($T_a = 10000$), P-SN and P-LN again yield the smallest averages, whereas among exact methods, S-UMix, S-MLE, and S-MAP attain the lowest values.

While all exact constructions enjoy formal anytime-valid coverage guarantees, the approximate prior likelihood mixing methods (P-SN, P-LN) do not. To assess whether these methods nevertheless achieve acceptable coverage in practice, we evaluate sequence-level violations: for each test image, we count it as a violation if at least one of the confidence sets in the sequence fails to contain the true parameter. We then average these indicators over the test set to obtain an empirical violation rate. To make this precise, we now introduce the necessary notation and definitions.

For all error levels $\delta \in (0, 1)$, starting steps $t_0 \in \{1, 60, 120\}$, acquisition times $T_a \in \{1, 100, 10000\}$, confidence coefficient specializations

$$m \in \{\text{P-SN}, \text{P-LN}, \text{P-UMix}, \text{P-UMed}, \text{P-DMix}, \text{P-DMed},\\ \text{S-MLE}, \text{S-MAP}, \text{S-UMix}, \text{S-UMed}, \text{S-DMix}, \text{S-DMed}\} =: M$$

and test set images $\boldsymbol{\theta}_i^* \in \{\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{|\mathcal{D}_{\text{test}}|}^*\}$, $i \in [|\mathcal{D}_{\text{test}}|]$, denote the corresponding (approximate) confidence coefficient at step $t_0 + t$, with $t \in \{0, 1, \dots, 180 - t_0\}$, by $\beta_{t_0+t,m,i}(\delta)$.

Associated with each quadruple (t_0, t, m, i), the (approximate) confidence set is

$$C_{t_0+t,m,i}(\beta_{t_0+t,m,i}(\delta)) \coloneqq \Big\{\boldsymbol{\theta} \in \Theta \;\Big|\; -\sum_{s=t_0}^{t_0+t} \log p_s(\mathbf{y}_{i,s} \mid \boldsymbol{\theta}) \leq \beta_{t_0+t,m,i}(\delta)\Big\},$$

where $(x_{i,1}, \mathbf{y}_{i,1}), \dots, (x_{i,180}, \mathbf{y}_{i,180})$ denotes the observation sequence corresponding to the image $\boldsymbol{\theta}_i^*$.

For acquisition time $T_a \in \{1, 100, 10000\}$, method $m \in M$, and starting step $t_0 \in \{1, 60, 120\}$, we define the *empirical violation rate* $\hat{v}_{t_0,T_a,m} \in [0, 1]$ of method m as

$$\hat{v}_{t_0,T_a,m} \coloneqq \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \mathbf{1}\{\exists\, t \in \{0, \dots, 180 - t_0\} : \boldsymbol{\theta}_i^* \notin C_{t_0+t,m,i}(\beta_{t_0+t,m,i}(\delta))\}\,,$$

where $\mathcal{D}_{\text{test}}$ is the test set of ground-truth images. This is a Monte Carlo estimate of

$$v_{t_0,T_a,m} \coloneqq \mathbf{E}_{\boldsymbol{\vartheta}}\Big[\mathbf{1}\{\exists\, t \in \{0, \dots, 180 - t_0\} : \boldsymbol{\vartheta} \notin C_{t_0+t,m}(\beta_{t_0+t,m}(\delta))\}\Big],$$

where $\mathbf{E}_{\boldsymbol{\vartheta}}$ denotes an expectation over random images from the underlying image distribution.

Figure 4.7 shows that the exact confidence sequences (all except P-LN and P-SN) have empirical violation rates that do not exceed $\delta = 0.05$ in most settings. The only exception is P-DMix using starting step $t_0 = 1$ and acquisition time $T_a = 1$. In contrast, P-LN and P-SN frequently exceed δ, particularly for $T_a \in \{100, 10000\}$. We emphasize that $v_{t_0,T_a,m} \leq \delta$ is a weaker guarantee than anytime validity: it is (i) a guarantee over finitely many confidence sets rather than over a infinite number of them, and (ii) an average over random images $\boldsymbol{\vartheta}$, whereas anytime validity is a point-wise guarantee ensuring that, for each fixed parameter, the violation probability is at most δ simultaneous for all confidence sets. Given the observed high empirical violation rate of P-LN and P-SN, we do not recommend Laplace-based approximations for prior likelihood mixing in practice.

4.2 Uncertainty Images

In this section, we present confidence-sequence-based *uncertainty images*. These images visualize which locations in a reconstruction are more or less reliable. A pixel value of 0 in the uncertainty image indicates full reliability of the corresponding reconstructed pixel, whereas a value of 1 indicates complete unreliability. As mentioned before, there are multiple ways to construct such images. We begin with the most reliable but also most conservative approach, namely *pixelwise uncertainty images*.

4.2.1 Pixelwise Uncertainty Images

Let $t \in \mathbb{N}$, $\delta \in (0, 1)$, and $\beta_t(\delta) \in (0, \infty)$ be a confidence coefficient. We define *pixelwise uncertainty image* $\mathbf{u}_t^{\text{pixel}} \in [0, 1]^{r \times r}$ by setting, for each pixel $(i, j) \in [r]^2$,

$$u_{t,i,j}^{\text{pixel}} \coloneqq \overline{\theta}_{t,i,j}^{\text{pixel}} - \underline{\theta}_{t,i,j}^{\text{pixel}}\,,$$

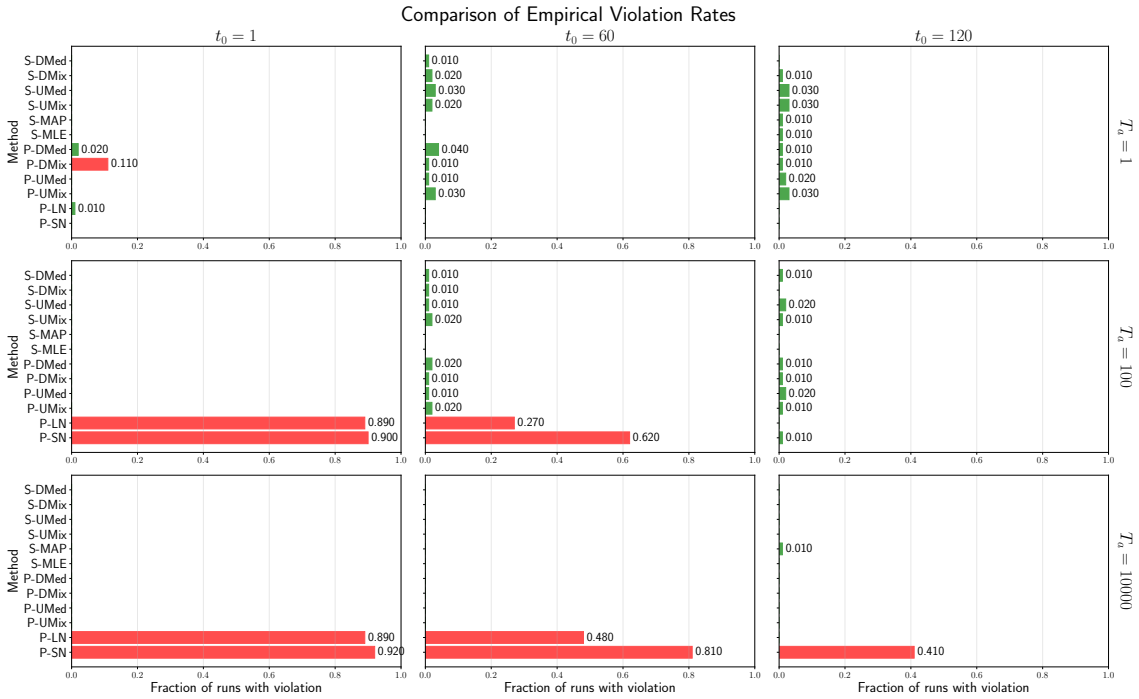


Figure 4.7: Empirical violation rates $\hat{v}_{t_0,T_a,m}$ for investigated methods $m \in M$, starting steps $t_0 \in \{1, 60, 120\}$ and acquisition times $T_a \in \{1, 100, 10000\}$.

where

$$\overline{\theta}^{\text{pixel}}_{t,i,j} := \max_{\boldsymbol{\theta} \in C_t(\beta_t(\delta))} \theta_{i,j}, \qquad \underline{\theta}^{\text{pixel}}_{t,i,j} := \min_{\boldsymbol{\theta} \in C_t(\beta_t(\delta))} \theta_{i,j}.$$

If $\beta_t(\delta)$ is taken from an exact confidence sequence for $\boldsymbol{\theta}^*$, then with probability at least $1 - \delta$

$$\boldsymbol{\theta}^* \in \left\{ \boldsymbol{\theta} \in \Theta : \forall (i,j) \in [r]^2,\ \theta_{i,j} \in [\underline{\theta}^{\text{pixel}}_{t,i,j}, \overline{\theta}^{\text{pixel}}_{t,i,j}] \right\}.$$

This guarantee strongly motivates the use of $\mathbf{u}^{\text{pixel}}_t$ to visualize the uncertainty across pixels of $\boldsymbol{\theta}^*$.

However, Figure 4.8 shows that the average pixelwise uncertainty,

$$\bar{u}^{\text{pixel}}_t := \frac{1}{r^2} \sum_{(i,j) \in [r]^2} u^{\text{pixel}}_{t,i,j},$$

decreases only slowly with acquisition time and over 180 steps. Only approximate confidence sequences, which lack anytime-validity guarantees both theoretically and empirically, yield notable reductions for acquisition times $T_a \in \{1, 100\}$. For $T_a = 10000$, S-MLE and S-MAP confidence sequences provide some reduction, but in light of the strong predictive performance of U-Net ensembles in this regime (see Table 4.3), the pixelwise uncertainty images remain highly conservative.

In Table 4.4 the best performing confidence sequence constructions utilizing 180 data points $(x_1, \mathbf{y}_1), \ldots, (x_{180}, \mathbf{y}_{180} \in \mathcal{X} \times \mathcal{Y}$ are displayed. In Figures 4.9 to 4.11 you can find the corresponding pixelwise uncertainty images for the first test set image.

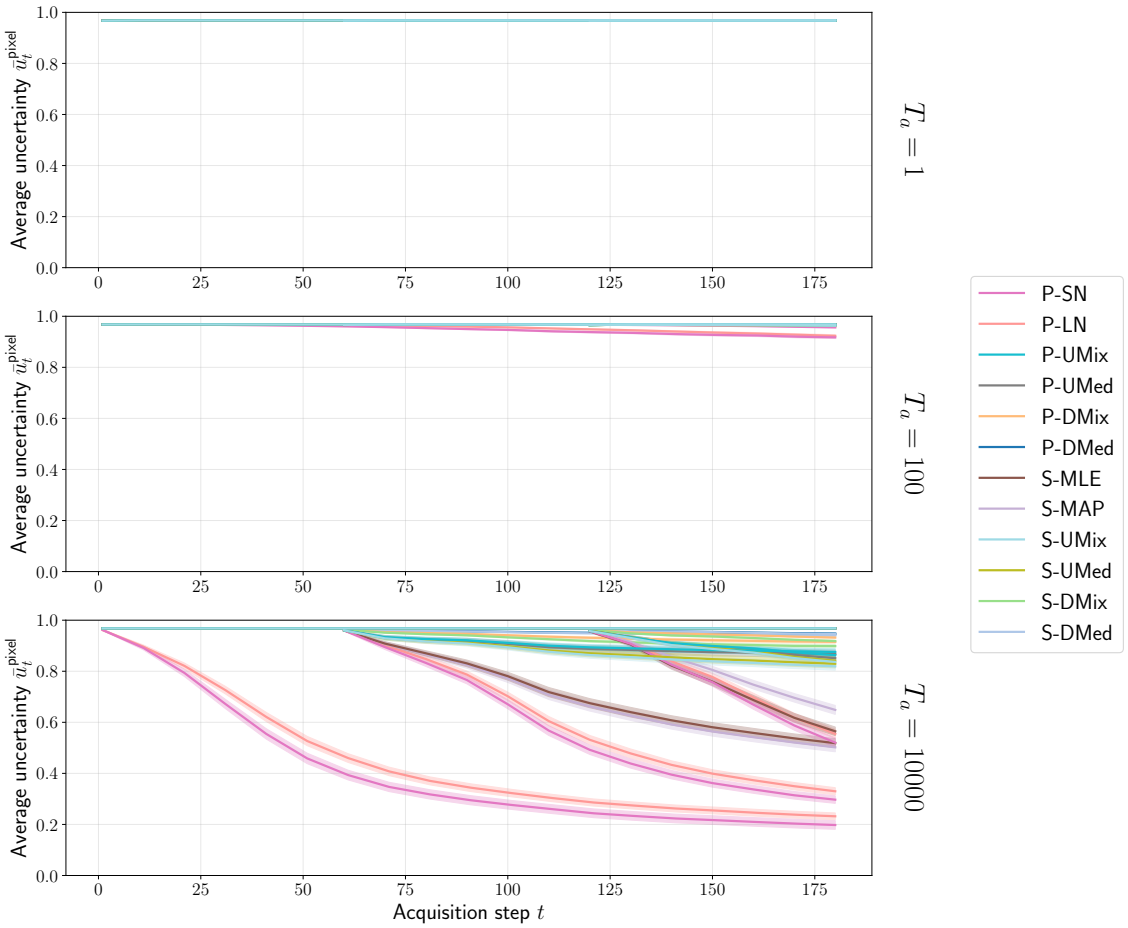


Figure 4.8: Pixelwise uncertainty averaged across all pixels and 100 test set images for different confidence sequences, grouped by acquisition time.

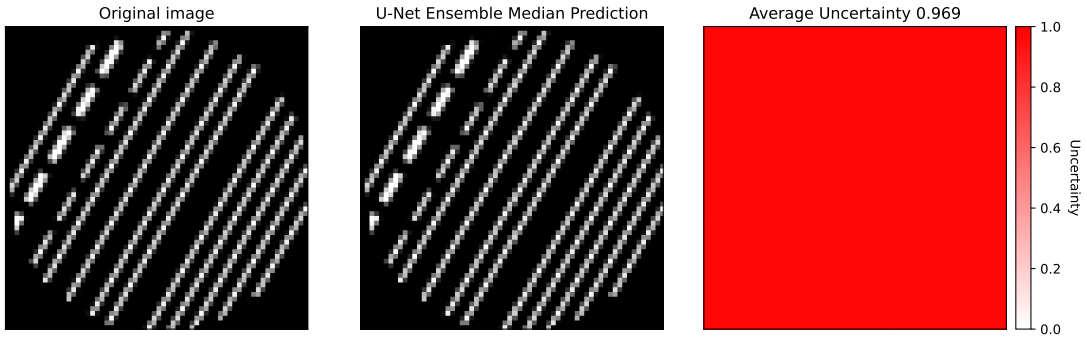


Figure 4.9: Pixelwise uncertainty image for P-DMix CS with $T_a = 1$, $t_0 = 1$ and $t = 180$.

Table 4.3: Performance metrics for U-Net ensemble median predictions across different steps and acquisition times. Values shown as mean ± SEM across test images.

Metric	**Step** (t)	**Acquisition Time** (T_a)		
		1	**100**	**10000**
PSNR (dB)	1	10.06 ± 0.14	11.36 ± 0.23	11.53 ± 0.23
	90	30.83 ± 0.73	40.46 ± 0.52	41.98 ± 0.54
	180	34.14 ± 0.64	42.98 ± 0.40	44.38 ± 0.43
RMSE	1	0.3181 ± 0.0051	0.2791 ± 0.0067	0.2736 ± 0.0066
	90	0.0383 ± 0.0026	0.0113 ± 0.0007	0.0097 ± 0.0007
	180	0.0249 ± 0.0016	0.0079 ± 0.0004	0.0069 ± 0.0004
L1	1	0.2432 ± 0.0046	0.1922 ± 0.0063	0.1882 ± 0.0062
	90	0.0077 ± 0.0007	0.0023 ± 0.0001	0.0019 ± 0.0001
	180	0.0047 ± 0.0003	0.0019 ± 0.0001	0.0016 ± 0.0001
ZeroOne	1	0.2201 ± 0.0079	0.1857 ± 0.0067	0.1824 ± 0.0065
	90	0.0737 ± 0.0016	0.0711 ± 0.0015	0.0710 ± 0.0015
	180	0.0720 ± 0.0015	0.0710 ± 0.0015	0.0710 ± 0.0015
SS	1	0.1466 ± 0.0136	0.3313 ± 0.0258	0.3478 ± 0.0253
	90	0.9802 ± 0.0023	0.9974 ± 0.0003	0.9980 ± 0.0003
	180	0.9908 ± 0.0010	0.9987 ± 0.0002	0.9990 ± 0.0001

Table 4.4: Average pixelwise uncertainty across test set images of top three exact methods-starting step combinations using 180 data points $(x_1, \mathbf{y}_1), \ldots, (x_{180}, \mathbf{y}_{180}) \in \mathcal{X} \times \mathcal{Y}$.

Acquisition Time	Best Method	Second Best	Third Best
$T_a = 1$	P-DMix, $t_0 = 1$ 0.968 ± 0.000	P-UMix, $t_0 = 1$ 0.968 ± 0.000	S-UMed, $t_0 = 1$ 0.968 ± 0.000
$T_a = 100$	S-UMed, $t_0 = 60$ 0.962 ± 0.011	S-UMix, $t_0 = 60$ 0.963 ± 0.010	P-UMed, $t_0 = 60$ 0.965 ± 0.007
$T_a = 10000$	S-MAP, $t_0 = 60$ 0.502 ± 0.172	S-MLE, $t_0 = 60$ 0.518 ± 0.170	S-MLE, $t_0 = 120$ 0.565 ± 0.164

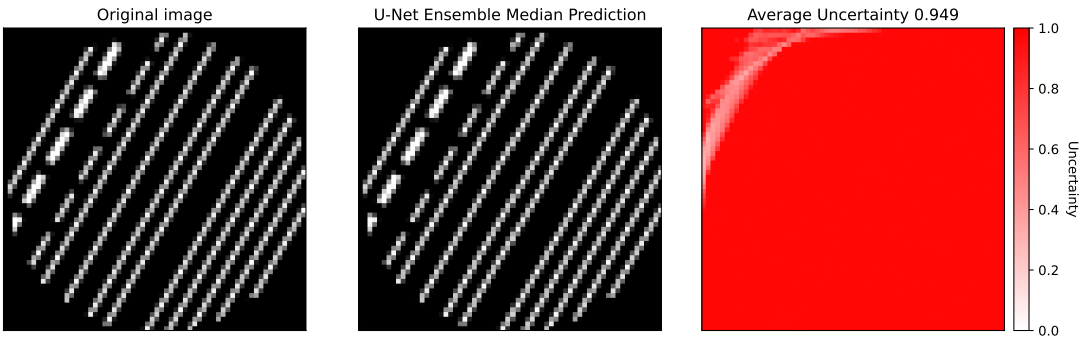


Figure 4.10: Pixelwise uncertainty image for S-UMed CS with $T_a = 100$, $t_0 = 60$ and $t = 180$.

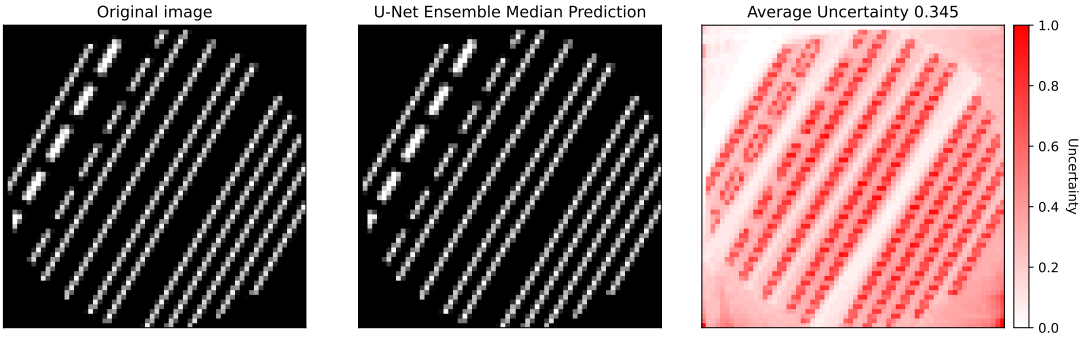


Figure 4.11: Pixelwise uncertainty image for S-MAP CS with $T_a = 10000$, $t_0 = 60$ and $t = 180$.

4.2.2 Global Uncertainty Images

Next, we consider an alternative, less conservative construction called *global uncertainty images*, denoted $\mathbf{u}_t^{\text{global}} \in [0,1]^{r \times r}$. Here, we implicitly assume that all pixels vary together, which justifies the term 'global'. Although this assumption is almost certainly unrealistic, it can make the resulting uncertainty images less conservative and in some cases more informative.

For $t \in \mathbb{N}$ and $(i,j) \in [r]^2$, define

$$u_{t,i,j}^{\text{global}} := \overline{\theta}_{t,i,j}^{\text{global}} - \underline{\theta}_{t,i,j}^{\text{global}},$$

with $\overline{\boldsymbol{\theta}}_t^{\text{global}}, \underline{\boldsymbol{\theta}}_t^{\text{global}} \in \Theta$ such that there exist

$$\overline{\boldsymbol{\nu}}_t^{\text{global}} \in \arg\max_{\boldsymbol{\theta} \in C_t} \sum_{(i,j) \in [r]^2} \theta_{i,j}, \qquad \underline{\boldsymbol{\nu}}_t^{\text{global}} \in \arg\min_{\boldsymbol{\theta} \in C_t} \sum_{(i,j) \in [r]^2} \theta_{i,j},$$

for which

$$\overline{\boldsymbol{\theta}}_t^{\text{global}} \approx \overline{\boldsymbol{\nu}}_t^{\text{global}}, \qquad \underline{\boldsymbol{\theta}}_t^{\text{global}} \approx \underline{\boldsymbol{\nu}}_t^{\text{global}}.$$

We obtain these approximations using gradient-based constrained optimization.

Examples are shown in Figures 4.12 to 4.14. These images yield less conservative uncertainty estimates but are informative only for larger acquisition times, $T_a \in \{100, 10000\}$.

4.2.3 Prediction-based Uncertainty Images

As shown in the previous sections, pixelwise and global uncertainty images are often too conservative to be informative, especially in low-acquisition-time or low-data settings. To obtain tighter uncertainty estimates, we instead use *data-consistent predictions*.

At step $t \in \mathbb{N}$, we generate $k \in \mathbb{N}$ predictions $\hat{\boldsymbol{\theta}}_{t,1}, \ldots, \hat{\boldsymbol{\theta}}_{t,k} \in \Theta$. Predictions that do not lie within the corresponding confidence set are projected back into it by gradient steps that minimize the negative log-likelihood L_t. If the projection fails, the prediction is replaced with $\hat{\boldsymbol{\theta}}_t^{\text{MLE}}$. Denote the modified predictions by $\hat{\boldsymbol{\theta}}'_{t,1}, \ldots, \hat{\boldsymbol{\theta}}'_{t,k} \in \Theta$.

We then define *prediction-based uncertainty image* $\mathbf{u}_t^{\text{pred}} \in [0,1]^{r \times r}$ with pixel values

$$u_{t,i,j}^{\text{pred}} := \max_{l \in \{1,\ldots,k\}} \hat{\boldsymbol{\theta}}'_{t,l,i,j} - \min_{l \in \{1,\ldots,k\}} \hat{\boldsymbol{\theta}}'_{t,l,i,j}, \qquad (i,j) \in [r]^2.$$

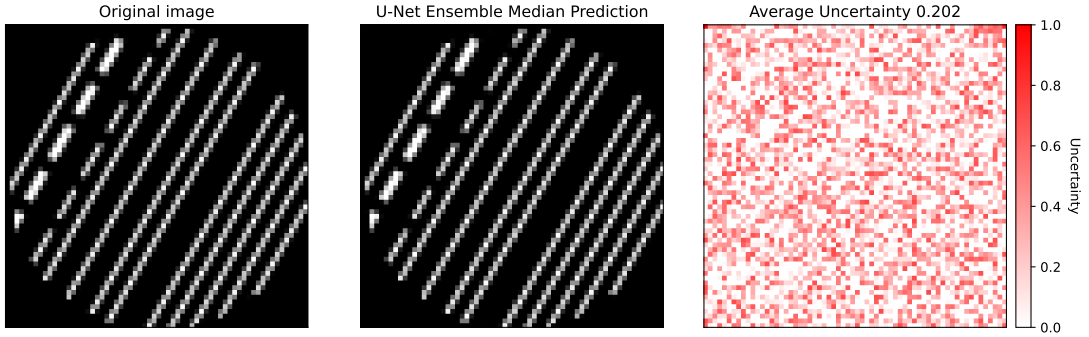


Figure 4.12: Global uncertainty image for P-DMix CS with $T_a = 1$, $t_0 = 1$ and $t = 180$.

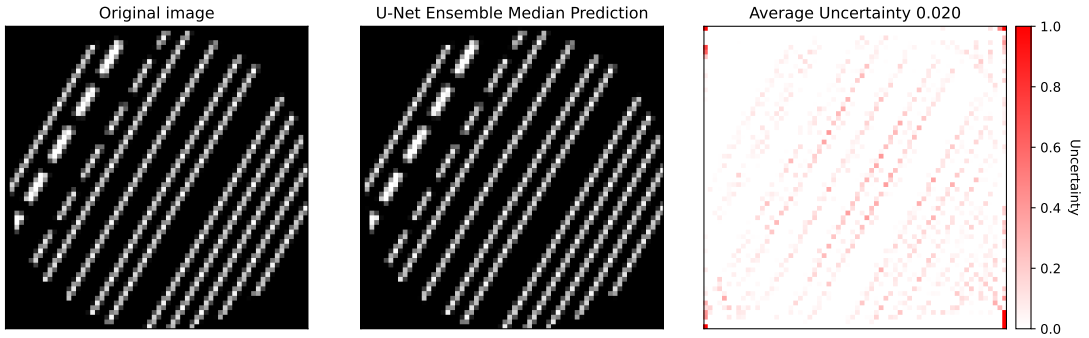


Figure 4.13: Global uncertainty image for S-UMed CS with $T_a = 100$, $t_0 = 60$ and $t = 180$.

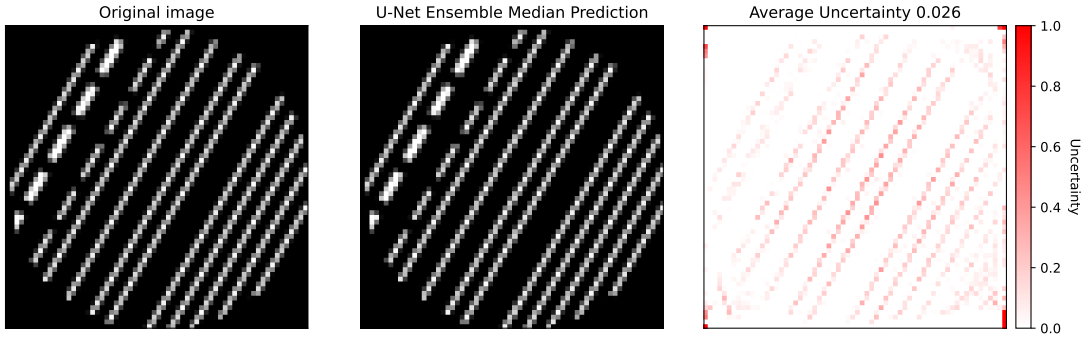


Figure 4.14: Global uncertainty image for S-MAP CS with $T_a = 10000$, $t_0 = 60$ and $t = 180$.

Figure 4.15: Prediction-based uncertainty images across acquisition steps for P-DMix CS with $T_a = 1$, $t_0 = 1$, and U-Net ensemble predictions.

Figure 4.15 illustrates this approach for an example confidence sequence. It exploits deep-learning-based predictions to produce tighter and more informative uncertainty estimates. For early acquisition steps, when predictions collapse to the average of the training set images due to insufficient data, the resulting uncertainties are less reliable. However, afterwards the average uncertainty

$$\bar{u}_t^{\text{pred}} := \frac{1}{r^2} \sum_{i,j \in [r]} u_{t,i,j}^{\text{pred}}$$

consistently upper bounds the true mean absolute error (MAE); see Figure 4.16. This property can be exploited to derive practical early stopping rules, e.g. stopping data acquisition once $\bar{u}_t^{\text{pred}}$ falls below a threshold such as 0.015. Of course, for robustness, such thresholds must be validated across additional examples.

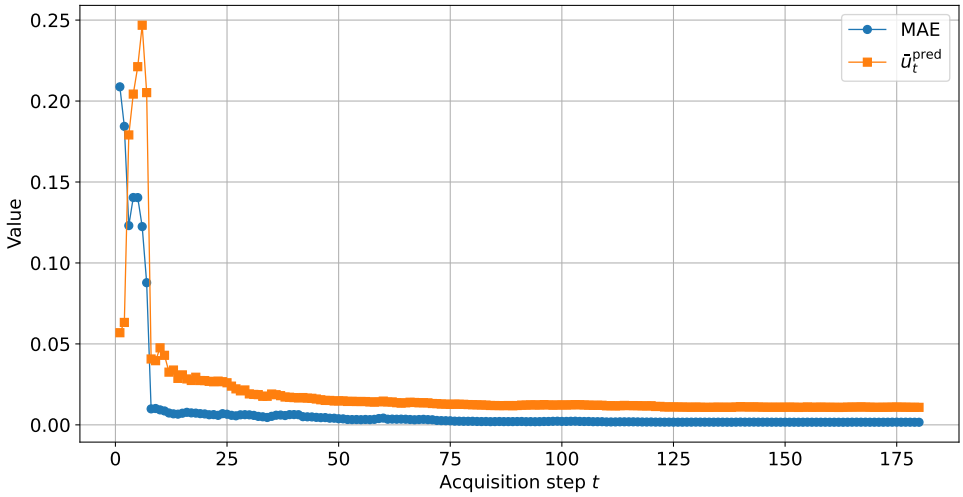


Figure 4.16: Prediction-based average uncertainty and mean absolute error (MAE) across pixels for P-DMix CS with $T_a = 1$, $t_0 = 1$, U-Net ensemble predictions, and acquisition steps $t \in [180]$.

4.2.4 Distance-based Uncertainty Images

Another approach to defining uncertainty scores is to take a given prediction $\hat{\boldsymbol{\theta}} \in \Theta$ and, within the confidence set, find an image that is *most different* from $\hat{\boldsymbol{\theta}}$ according to a chosen distance. Formally, for a distance function $d : \mathbb{R}^{r^2} \times \mathbb{R}^{r^2} \to [0, \infty)$, and whenever $C_t(\beta_t(\delta)) \neq \emptyset$, define the set of distance maximizers

$$\mathcal{M}_t(\hat{\boldsymbol{\theta}}) := \left\{ \boldsymbol{\theta} \in C_t(\beta_t(\delta)) \;\middle|\; d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}' \in C_t(\beta_t(\delta))} d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}') \right\}.$$

In our visualizations, we use the ℓ^2-distance, but other choices are possible.

Let $\boldsymbol{\nu}_t^{\max} \in \mathcal{M}_t(\hat{\boldsymbol{\theta}})$ be one such maximizer and let $\boldsymbol{\theta}_t^{\max} \in \Theta$ denote an approximation obtained in practice (e.g., via gradient-based optimization) such that

$$\boldsymbol{\theta}_t^{\max} \approx \boldsymbol{\nu}_t^{\max}.$$

Given $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_t^{\max}$, define *distance-based uncertainty image* $\mathbf{u}_t^d \in [0,1]^{r\times r}$ with pixel values

$$u_{t,i,j}^d := |\hat{\theta}_{i,j} - \theta_{t,i,j}^{\max}|, \qquad (i,j) \in [r]^2.$$

Because $\Theta = [0,1]^{r^2}$, each $u_{t,i,j}^d \in [0,1]$. The mean uncertainty score is

$$\bar{u}_t^d := \frac{1}{r^2} \sum_{(i,j)\in[r]^2} u_{t,i,j}^d.$$

This construction quantifies uncertainty *relative* to the prediction $\hat{\boldsymbol{\theta}}$: small $u_{t,i,j}^d$ indicate pixels where the prediction is close to at least one extremal element of the confidence set. Conversely, if $\hat{\boldsymbol{\theta}}$ is far from the true parameter, there will typically exist an element of the confidence set that is far from $\hat{\boldsymbol{\theta}}$, resulting in large values of $u_{t,i,j}^d$. In this way, the overconfidence observed in prediction-based uncertainty images (see Section 4.2.3) is mitigated, since the distance-based construction reacts by assigning greater uncertainty whenever predictions deviate substantially from the confidence set. Example distance-based uncertainty images can be seen in Figure 4.17. As the figure shows, the average distance-based uncertainty scores $\bar{u}_t^d$ decrease substantially over time, indicating that these uncertainty images not only avoid overconfidence, but also avoid being overly conservative.

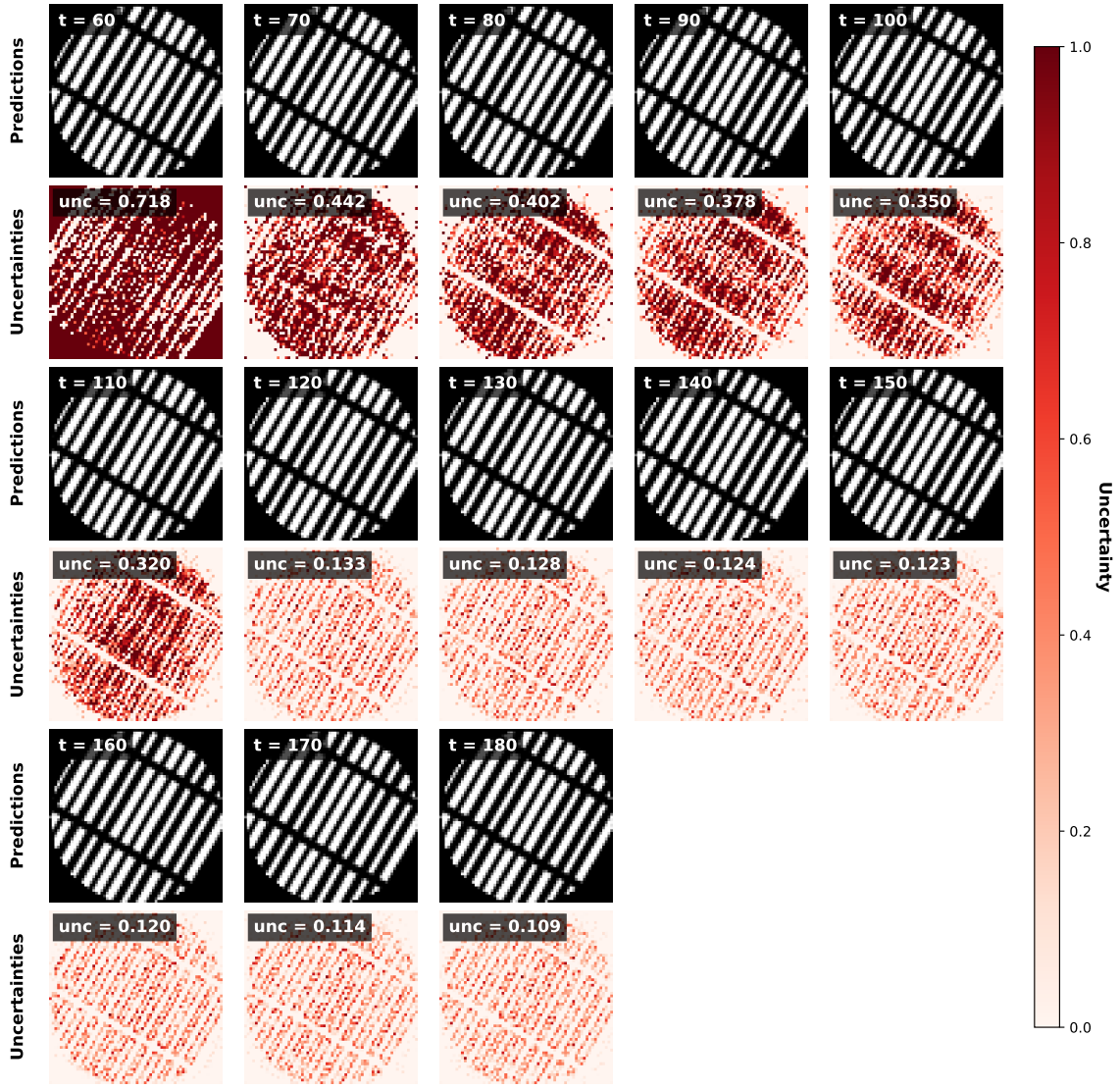


Figure 4.17: Distance-based uncertainty images across acquisition steps for S-UMix CS with $T_a = 100$, $t_0 = 60$, and U-Net ensemble predictions. Here the ℓ^2-distance is used to construct $\boldsymbol{\theta}^{\max}$.

Chapter 5

Summary

The goal of this thesis was to analyze, develop, and experimentally compare confidence-sequence–based methods for uncertainty quantification in tomographic imaging. Although most established approaches provide only point predictions without indicating their reliability, the methods presented here overcome this limitation through mathematically rigorous confidence sequences based on probability theory.

The work focused primarily on two constructions, prior likelihood mixing and sequential likelihood mixing, and further examined Laplace-based approximations of prior likelihood mixing. We specialized these constructions using U-Net ensembles, diffusion models, and statistical estimators.

The experimental results reveal that the different confidence sequence variants offer complementary strengths. U-Net ensembles and diffusion models achieve highly accurate predictions even under short acquisition times, while classical statistical estimators such as MLE and MAP are particularly effective in high-count regimes, producing tighter and more informative confidence sequences. Beyond numerical performance, the thesis also introduced uncertainty images, which visualize local reconstruction reliability and offer an intuitive way to assess prediction trustworthiness. While pixelwise uncertainty images are the most reliable, we found that distance-based uncertainty images are less conservative and avoid the disadvantages that global and prediction-based uncertainty images suffer.

Together, these contributions demonstrate that the combination of statistical theory and modern machine learning methods enables rigorous uncertainty quantification for tomographic reconstruction. In the longer term, the presented methods have the potential to support safer and more reliable decision-making in medical imaging, industrial applications, and beyond.

Bibliography

Bach, F. (2021, July). Approximating integrals with Laplace's method. Machine Learning Research Blog (blog). Posted on July 23, 2021.

Barba, L., J. Kirschner, T. Aidukas, M. Guizar-Sicairos, and B. Béjar (2024, October). Diffusion Active Learning: Towards Data-Driven Experimental Design in Computed Tomography.

Bhargava, P., G. He, A. Samarghandi, and E. S. Delpassand (2012, March). Pictorial review of SPECT/CT imaging applications in clinical nuclear medicine. *American Journal of Nuclear Medicine and Molecular Imaging 2*(2), 221–231.

Bracewell, R. N. and A. C. Riddle (1967, November). Inversion of Fan-Beam Scans in Radio Astronomy. *The Astrophysical Journal 150*, 427. Publisher: IOP ADS Bibcode: 1967ApJ...150..427B.

Brown, R. W. (2014). *Magnetic resonance imaging: physical principles and sequence design* (Second edition. ed.). Hoboken, New Jersey: Wiley-Blackwell.

Chen, J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou (2021). TransUNet: Transformers make strong encoders for medical image segmentation.

Chen, J., J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, Z. Shaoting, L. Xing, L. Lu, A. Yuille, and Y. Zhou (2024). TransUNet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *97*, 103280. Publisher: Elsevier.

Cormack, A. M. (1963, September). Representation of a Function by Its Line Integrals, with Some Radiological Applications. *Journal of Applied Physics 34*(9), 2722–2727.

Dahlbom, M. (2001, November). Estimation of image noise in PET using the bootstrap method. In *2001 IEEE Nuclear Science Symposium Conference Record (Cat. No.01CH37310)*, Volume 4, pp. 2075–2079 vol.4. ISSN: 1082-3654.

Dao, T. (2023). FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Darling, D. A. and H. Robbins (1967, July). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences 58*(1), 66–68. Publisher: Proceedings of the National Academy of Sciences.

de Bruijn, N. (1970). *Asymptotic methods in analysis* (3rd ed.). Bibliotheca Mathematica. Netherlands: North-Holland Publishing Company.

De Chiffre, L., S. Carmignato, J. P. Kruth, R. Schmitt, and A. Weckenmann (2014,

January). Industrial applications of computed tomography. *CIRP Annals 63*(2), 655–677.

Dhariwal, P. and A. Nichol (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, Volume 34, pp. 8780–8794. Curran Associates, Inc.

Drozdzal, M., E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal (2016). The importance of skip connections in biomedical image segmentation. In G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, and J. Cornebise (Eds.), *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187. Springer International Publishing.

Ekmekci, C. and M. Cetin (2025, April). Conformalized Generative Bayesian Imaging: An Uncertainty Quantification Framework for Computational Imaging. arXiv:2504.07696 [eess].

Esser, P., S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633. PMLR. ISSN: 2640-3498.

Fessler, J. and W. Rogers (1996, September). Spatial resolution properties of penalized-likelihood image reconstruction: space-invariant tomographs. *IEEE Transactions on Image Processing 5*(9), 1346–1358.

Fuest, M., P. Ma, M. Gui, J. Schusterbauer, V. T. Hu, and B. Ommer (2024). Diffusion models and representation learning: A survey.

Fukushima, K. (2013). Training multi-layered neural network neocognitron. *40*, 18–31. Publisher: Pergamon.

Gerlier, C., M. Forster, A. Fels, M. Zins, G. Chatellier, and O. Ganansia (2022, November). Head computed tomography for elderly patients with acute altered mental status in the emergency setting: value for decision-making and predictors of abnormal findings. *Clinical and Experimental Emergency Medicine 9*(4), 333–344.

Han, Y. and J. C. Ye (2018). Framing u-net via deep convolutional framelets: Application to sparse-view CT. *37*(6), 1418–1429.

Hansen, L. and P. Salamon (1990, October). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(10), 993–1001.

Hendrycks, D. and K. Gimpel (2023). Gaussian error linear units (GELUs).

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors.

Ho, J., A. Jain, and P. Abbeel (2020, December). Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, pp. 6840–6851. Curran Associates Inc.

Hounsfield, G. N. (1973, December). Computerized transverse axial scanning (tomography). 1. Description of system. *The British Journal of Radiology 46*(552), 1016–1022.

Hudson, H. and R. Larkin (1994, December). Accelerated image reconstruction using ordered subsets of projection data. *IEEE Transactions on Medical Imaging 13*(4), 601–609.

Jin, K. H., M. T. McCann, E. Froustey, and M. Unser (2017, September). Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Transactions on Image Processing 26*(9), 4509–4522.

Kak, A. C. and M. Slaney (2001, January). *Principles of Computerized Tomographic Imaging*. Society for Industrial and Applied Mathematics. Publication Title: Principles of Computerized Tomographic Imaging.

Kang, E., J. Min, and J. C. Ye (2017). A deep convolutional neural network using directional wavelets for low-dose x-ray CT reconstruction. *44*(10), e360–e375.

Khanna, A., N. D. Londhe, S. Gupta, A. Semwal, and N (2020). A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images. *40*(3), 1314–1327. Publisher: Elsevier.

Kingma, D. P. and J. Ba (2017). Adam: A method for stochastic optimization.

Kirschner, J., A. Krause, M. Meziu, and M. Mojmir (2025, February). Confidence Estimation via Sequential Likelihood Mixing.

Kiss, M. B., A. Biguri, Z. Shumaylov, F. Sherry, K. J. Batenburg, C.-B. Schönlieb, and F. Lucka (2025, February). Benchmarking learned algorithms for computed tomography image reconstruction tasks. *Applied Mathematics for Modern Challenges 3*(0), 1–43. Publisher: Applied Mathematics for Modern Challenges.

Klenke, A. (2020, oct). *Probability Theory* (3 ed.). Universitext. Cham: Springer International Publishing.

Kuhl, D. E. and R. Q. Edwards (1963, April). Image Separation Radioisotope Scanning. *Radiology 80*(4), 653–662. Publisher: Radiological Society of North America.

Kutiel, G., R. Cohen, M. Elad, D. Freedman, and E. Rivlin (2023). Conformal Prediction Masks: Visualizing Uncertainty in Medical Imaging. In H. Chen and L. Luo (Eds.), *Trustworthy Machine Learning for Healthcare*, Cham, pp. 163–176. Springer Nature Switzerland.

Königsberger, K. (2004). *Analysis 2*. Springer-Lehrbuch. Springer.

Lai, T. L. (1976a, April). Boundary Crossing Probabilities for Sample Sums and Confidence Sequences. *The Annals of Probability 4*(2), 299–312. Publisher: Institute of Mathematical Statistics.

Lai, T. L. (1976b, March). On Confidence Sequences. *The Annals of Statistics 4*(2), 265–280. Publisher: Institute of Mathematical Statistics.

Laplace, P.-S. d. .-. A. d. t. (1878). *Oeuvres complètes de Laplace. Tome 8 / publiées sous les auspices de l'Académie des sciences, par MM. les secrétaires perpétuels*. Académie des sciences.

Lauterbur, P. C. (1973, March). Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance. *Nature 242*(5394), 190–191. Publisher: Nature Publishing Group.

Lee, J., T. H. Joshi, M. S. Bandstra, D. L. Gunter, B. J. Quiter, R. J. Cooper, and K. Vetter (2024, October). Radiation image reconstruction and uncertainty quantification using a Gaussian process prior. *Scientific Reports 14*(1), 22958. Publisher: Nature Publishing Group.

Lefaudeux, B., F. Massa, D. Liskovich, W. Xiong, V. Caggiano, S. Naren, M. Xu, J. Hu, M. Tintore, S. Zhang, P. Labatut, D. Haziza, L. Wehrstedt, J. Reizenstein, and G. Sizov (2022). xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers.

Li, T., M. Chen, B. Guo, and Z. Shen (2025). A survey on diffusion language models.

Loshchilov, I. and F. Hutter (2019). Decoupled weight decay regularization.

Nichol, A. Q. and P. Dhariwal (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8162–8171. PMLR. ISSN: 2640-3498.

Nichol, A. Q., P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen (2022). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16784–16804. PMLR. ISSN: 2640-3498.

Oktay, O., J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert (2022). Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*.

Pacheco, M. and D. Goyal (2010, May). X-ray computed tomography for non-destructive failure analysis in microelectronics. In *2010 IEEE International Reliability Physics Symposium*, pp. 252–258. ISSN: 1938-1891.

Paszke, A., S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer (2017). Automatic differentiation in pytorch. In *NIPS-W*.

Pedersen, F. H., J. S. Jørgensen, and M. S. Andersen (2022, March). A Bayesian Approach to CT Reconstruction with Uncertain Geometry. arXiv:2203.01045 [math].

Qi, J. and R. Leahy (2000, May). Resolution and noise properties of MAP reconstruction for fully 3-D PET. *IEEE Transactions on Medical Imaging 19*(5), 493–506.

Rabi, I. I. (1937, April). Space Quantization in a Gyrating Magnetic Field. *Physical Review 51*(8), 652–654. Publisher: American Physical Society.

Robbins, H. and D. Siegmund (1970). Boundary Crossing Probabilities for the Wiener Process and Sample Sums. *The Annals of Mathematical Statistics 41*(5), 1410–1429. Publisher: Institute of Mathematical Statistics.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer (2022). High-resolution image synthesis with latent diffusion models. pp. 10684–10695.

Ronneberger, O., P. Fischer, and T. Brox (2015, May). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs].

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *61*, 85–117. Publisher: Pergamon.

Shepp, L. A. and Y. Vardi (1982, October). Maximum Likelihood Reconstruction for Emission Tomography. *IEEE Transactions on Medical Imaging 1*(2), 113–122.

Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan, and S. Ganguli (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265. PMLR. ISSN: 1938-7228.

Sweet, W. H. (1951, December). The Uses of Nuclear Disintegration in the Diagnosis and Treatment of Brain Tumor. *New England Journal of Medicine 245*(23), 875–878. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJM195112062452301.

Vasconcelos, F., B. He, N. Singh, and Y. W. Teh (2023, May). UncertaINR: Uncertainty Quantification of End-to-End Implicit Neural Representations for Computed Tomography. arXiv:2202.10847 [eess].

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Ville, J. (1939). *Étude critique de la notion de collectif*. Doctoral dissertation (phd thesis), Université de Paris.

von Platen, P., S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf (2022). Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Wang, Z., A. Bovik, H. Sheikh, and E. Simoncelli (2004). Image quality assessment: from error visibility to structural similarity. *13*(4), 600–612.

Wang, Z. and A. C. Bovik (2009). Mean squared error: Love it or leave it? a new look at signal fidelity measures. *26*(1), 98–117.

Wrenn, F. R., M. L. Good, and P. Handler (1951, May). The Use of Positron-emitting Radioisotopes for the Localization of Brain Tumors. *Science 113*(2940), 525–527. Publisher: American Association for the Advancement of Science.

Wu, Y. and K. He (2018). Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19.

Yang, L., Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang (2023). Diffusion models: A comprehensive survey of methods and applications. *56*(4), 105:1–105:39.

Zhang, W. (2025). Lecture 8: Denoising diffusion probabilistic models and normalizing flows. https://weizhang.userpage.fu-berlin.de/teaching/2025-summer/lecture_08.pdf. Accessed: 2025-09-14.

Zhou, Q., T. Yu, X. Zhang, and J. Li (2020, January). Bayesian Inference and Uncertainty Quantification for Medical Image Reconstruction with Poisson Data. *SIAM Journal on Imaging Sciences 13*(1), 29–52.

Appendix A

Definitions and Theorems

All probability theory related definitions, theorems, lemmas and examples in this appendix come from Klenke (2020) unless stated otherwise. Some of them are copied word for word.

Definition A.1 (Asymptotic dominance)**.** *We say that* $f : \mathbb{R}^d \to \mathbb{R}$ asymptotically dominates $g : \mathbb{R}^d \to \mathbb{R}$ *(with limit* $\mathbf{x}_0$*), symbolically* $f(\mathbf{x}) = o(g(\mathbf{x}))$*, if and only if*

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{f(\mathbf{x})}{g(\mathbf{x})} = 0.$$

Definition A.2 (Asymptotic equivalence (de Bruijn, 1970))**.** *We say that* $f : \mathbb{R} \to \mathbb{R}$ *is* asymptotically equivalent *to* $g : \mathbb{R} \to \mathbb{R}$*, symbolically* $f \sim g$*, if*

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1.$$

Definition A.3 (Multi-index (Königsberger, 2004))**.** *Let* $\alpha_1, \dots, \alpha_k \in \{0, 1, \dots\}$ *and* $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$*, then*

(i) $|\boldsymbol{\alpha}| := \alpha_1 + \alpha_2 + \cdots + \alpha_n$,

(ii) $\boldsymbol{\alpha}! := \alpha_1! \cdot \alpha_2! \cdots \alpha_n!$,

(iii) for $\mathbf{x} \in \mathbb{R}^d$*, define* $\mathbf{x}^{\boldsymbol{\alpha}} := x_1^{\alpha^1} x_2^{\alpha^2} \cdots x_d^{\alpha^d}$,

(iv) for $f : \mathbb{R}^k \to \mathbb{R}$ *for which all* $(k+1)$*-th order partial derivatives exist, define*

$$D^{\boldsymbol{\alpha}} f := \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Theorem A.4 (Taylor's theorem (Königsberger, 2004))**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$*. If all* $(k+1)$*-th order partial derivatives of* f *exist and form continuous functions, that is* $f \in C^{k+1}(\mathbb{R}^d)$*, then for all* $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^d$ *there exists a* $\boldsymbol{\xi} \in \mathbb{R}^d$ *that lies on the line segment between* $\mathbf{x}$ *and* $\mathbf{x}_0$ *such that*

$$f(\mathbf{x}) = T_k(\mathbf{x}, \mathbf{x}_0) + R_k(\mathbf{x}, \mathbf{x}_0)$$

with k*-th order Taylor polynomial*

$$T_k(\mathbf{x}, \mathbf{x}_0) = \sum_{\boldsymbol{\alpha}\in\{1,2,\dots\}^d, |\boldsymbol{\alpha}|\le k} \frac{D^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}}$$

and remainder term

$$R_k(\mathbf{x}, \mathbf{x}_0) = \sum_{\boldsymbol{\alpha} \in \{1,2,\dots\}^d, |\boldsymbol{\alpha}| = k+1} \frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{\xi})}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}}.$$

Remark. The remainder term R_k in Theorem A.4 is asymptotically dominated (Definition A.1) by $(\mathbf{x} - \mathbf{x})^k$ with respect to limit $\mathbf{x}_0$, symbolically $R_k = o((\mathbf{x} - \mathbf{x}_0)^k)$.

Definition A.5 (σ-algebra)**.** *A collection of subsets $\mathcal{A} \subseteq 2^\Omega$ is called a σ-algebra if it satisfies:*

(i) $\Omega \in \mathcal{A}$.

(ii) For every $A \in \mathcal{A}$, its complement A^c is also in $\mathcal{A}$.

(iii) $\mathcal{A}$ is closed under countable unions.

Theorem A.6 (Generated σ-algebra)**.** *Let $\mathcal{E} \subseteq 2^\Omega$. Then there exists a smallest σ-algebra $\sigma(\mathcal{E})$ with $\mathcal{E} \subseteq \sigma(\mathcal{E})$:*

$$\sigma(\mathcal{E}) \coloneqq \bigcap_{\substack{\mathcal{A} \subseteq 2^\Omega \text{ is a } \sigma\text{-algebra} \\ \mathcal{E} \subseteq \mathcal{A}}} \mathcal{A}.$$

$\sigma(\mathcal{E})$ is called the σ-algebra generated by *$\mathcal{E}$. $\mathcal{E}$ is called the* generator *of $\sigma(\mathcal{E})$.*

Definition A.7 (Topological space)**.** *A* topological space *is a pair (X, τ), where X is a set and τ is a collection of subsets of X, called* open sets*, satisfying:*

1. *$\varnothing$ and X are in τ.*
2. *The union of any collection of sets in τ is also in τ.*
3. *The intersection of any finite number of sets in τ is also in τ.*

Remark. The usual (or standard) topology on $\mathbb{R}^n$, $(\mathbb{R}^n, \tau)$ is the one induced by the standard Euclidean metric. Here collection $\tau = \{B_r(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n, r \in \mathbb{R}\}$ and open balls $B_r(\mathbf{x}) = \{\mathbf{x}' \in \mathbb{R}^n : \|\mathbf{x}' - \mathbf{x}\|_2 < r\}$.

Definition A.8 (Borel σ-algebra)**.** *Let (Ω, τ) be a topological space (Definition A.7). Then σ-algebra*

$$\mathcal{B}(\Omega) \coloneqq \mathcal{B}(\Omega, \tau) \coloneqq \sigma(\tau)$$

that is generated by the open sets (Definition A.7) is called the Borel σ-algebra *on Ω. The elements $A \in \mathcal{B}(\Omega, \tau)$ are called* Borel sets *or* Borel measurable sets*.*

Often we refer to $\mathcal{B}(\mathbb{R}^n) = \mathcal{B}(\mathbb{R}^n, \tau)$ with τ containing the open Euclidean balls (open intervals $\mathbb{R}$) as the Borel σ-algebra*.*

Definition A.9 (Metric space)**.** *A* metric space *is a pair (X, d) where X is a set and $d : X \times X \to [0, \infty)$ is a function (called a* metric*) satisfying:*

1. *$d(x, y) = 0$ if and only if $x = y$.*
2. *$d(x, y) = d(y, x)$ for all $x, y \in X$ (symmetry).*
3. *$d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (triangle inequality).*

Definition A.10 (Cauchy sequence)**.** *Let (X, d) be a metric space (Definition A.9). A sequence (x_n) in X is called a* Cauchy sequence *if for every $\epsilon > 0$ there exists an $N \in \mathbb{N}$ such that for all $m, n \geq N$,*

$$d(x_m, x_n) < \epsilon.$$

This concept is central to the definition of completeness (Definition A.11).

Definition A.11 (Complete metric space)**.** *A metric space (X, d) (Definition A.9) is* complete *if every Cauchy sequence in X (Definition A.10) converges to a limit that is also in X.*

Definition A.12 (Dense subset)**.** *Let (X, τ) be a topological space (Definition A.7). A subset $D \subseteq X$ is* dense *in X if every non-empty open set $U \subseteq X$ intersects D; that is,*

$$U \cap D \neq \varnothing.$$

Equivalently, the closure of D is X.

Definition A.13 (Separable space)**.** *A topological space (X, τ) (Definition A.7) is* separable *if there exists a countable dense subset (Definition A.12) $D \subseteq X$.*

Example A.14. $\mathbb{Q}^n$ is a countable dense subset (Definition A.12) of $\mathbb{R}^n$, so $\mathbb{R}^n$ with the usual topology (Appendix A) is separable (Definition A.13).

Definition A.15 (Polish space)**.** *A topological space X is called a* Polish space *if there exists a metric d on X such that:*

1. *The metric d induces the topology on X (i.e. the open sets of X coincide with those given by d; see Definition A.9).*
2. *The metric space (X, d) is complete (Definition A.11).*
3. *The metric space (X, d) is separable (Definition A.13).*

Example A.16. $\mathbb{R}^n$ with the usual metric (Appendix A) is a Polish space (Definition A.15) since it is separable by $\mathbb{Q}^n$ (Definition A.13) and complete (Definition A.11).

Definition A.17 (Measurable Space)**.** *A pair $(\Omega, \mathcal{A})$ consisting of a nonempty set Ω and a σ-algebra $\mathcal{A} \subseteq 2^\Omega$ (Definition A.5) is called a* measureable space*. The sets $A \in \mathcal{A}$ are called* measurable sets*. If Ω is at most countably finite and if $\mathcal{A} = 2^\Omega$, then the measurable space $(\Omega, 2^\Omega)$ is called* discrete*.*

Remark. It is common to slightly abuse notation and call Ω *a measurable space* (Definition A.17), omitting the definition of the associated σ-algebra (Definition A.5). In those cases Ω is the first entry of the actual measurable space.

Definition A.18 (Measure space)**.** *A triple $(\Omega, \mathcal{A}, \mu)$ is called a* measure space *if $(\Omega, \mathcal{A})$ is a measurable space and μ is a measure (Definition A.24) on $\mathcal{A}$.*

Definition A.19 (Probability space)**.** *A measure space $(\Omega, \mathcal{A}, \mathbf{P})$ (Definition A.18) with $\mathbf{P}(\Omega) = 1$ is called a* probability space *and sets $A \in \mathcal{A}$ are called* events*.*

Definition A.20 (Random variables)**.** *Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $(\Omega', \mathcal{A}')$ be a measurable space. A function $X : \Omega \to \Omega'$ is called a* random variable *if X is a measurable map (Definition A.21). In the common case where $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, X is called a real random variable. For any $A' \in \mathcal{A}'$ we write*

$$\{X \in A'\} \coloneqq X^{-1}(A') \quad \textit{and} \quad \mathbf{P}(X \in A') \coloneqq \mathbf{P}(X^{-1}(A')).$$

Definition A.21 (Measurable maps)**.** *Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces. A function $X : \Omega \to \Omega'$ is said to be* $\mathcal{A}$-$\mathcal{A}'$-measurable *if for every $A' \in \mathcal{A}'$, the preimage $X^{-1}(A')$ belongs to $\mathcal{A}$. In the special case when $\Omega' = \mathbb{R}$ and $\mathcal{A}'$ is the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ (Definition A.8), X is called an* $\mathcal{A}$-measurable real map*.*

Definition A.22 (Generated σ-algebra)**.** *Let $(\Omega', \mathcal{A}')$ be a measurable space and let Ω be a nonempty set. Let $X \colon \Omega \to \Omega'$ be a map. The preimage*

$$X^{-1}(\mathcal{A}') \coloneqq \left\{ X^{-1}(A') : A' \in \mathcal{A}' \right\}$$

is the smallest σ-algebra with respect to which X is measurable. We say that $\sigma(X) \coloneqq X^{-1}(\mathcal{A}')$ is the σ-algebra on Ω that is generated *by X.*

Definition A.23 (Generated σ-algebra)**.** *Let Ω be a nonempty set. Let I be an arbitrary index set. For any $i \in I$, let $(\Omega_i, \mathcal{A}_i)$ be a measurable space (Definition A.17) and let $X_i \colon \Omega \to \Omega_i$ be an arbitrary map. Then*

$$\sigma(X_i, i \in I) \coloneqq \sigma\left(\bigcup_{i \in I} \sigma(X_i) \right) = \sigma\left(\bigcup_{i \in I} X_i^{-1}(\mathcal{A}_i) \right)$$

is called the σ-algebra (Definition A.5) on Ω that is generated *by $(X_i, i \in I)$. This is the smallest σ-algebra with respect to which all X_i are measurable (Definition A.21).*

Definition A.24 (Content, Premeasure, Measure, Probability Measure)**.** *Let $\mathcal{A}$ be a semiring of subsets of Ω and let $\mu : \mathcal{A} \to [0, \infty]$ be a set function with $\mu(\emptyset) = 0$. Then:*

- *μ is a* content *if for any finitely many disjoint sets $A_1, \ldots, A_n \in \mathcal{A}$ with $\bigcup_{i=1}^n A_i \in \mathcal{A}$,*

$$\mu\Big(\biguplus_{i=1}^{n} A_i \Big) = \sum_{i=1}^{n} \mu(A_i).$$

- *μ is a* premeasure *if for any countable collection of disjoint sets $A_1, A_2, \cdots \in \mathcal{A}$ with $\bigcup_{i=1}^\infty A_i \in \mathcal{A}$,*

$$\mu\Big(\biguplus_{i=1}^{\infty} A_i \Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

- *μ is a* measure *if it is a premeasure and $\mathcal{A}$ is a σ-algebra (Definition A.5).*
- *μ is a* probability measure *if it is a measure and $\mu(\Omega) = 1$.*

Definition A.25 (Finite, σ-finite measure)**.** *Let $\mathcal{A}$ be a semiring. A content μ on $\mathcal{A}$ measurs is called a*

(i) finite *if $\mu(A) < \infty$ for every $A \in \mathcal{A}$ and*

(ii) σ-finite if there exists a sequence of sets $\Omega_1, \Omega_2, \cdots \in \mathcal{A}$ such that $\Omega = \bigcup_{n=1}^\infty \Omega_n$ and such that $\mu(\Omega_n) < \infty$ for all $n \in \mathbb{N}$.

Remark. A measure (Definition A.24), as opposed to a measurable map (Definition A.21), is defined on a measurable space (Definition A.17) and outputs a nonnegative number. It *measures* the size or mass of sets. A measurable maps map between measurable spaces and have the nice property that their preimages $X^{-1}(A)$ are measurable, i.e. we do not end up with sets we cannot measure when we look at what sets *cause* X to take certain values.

Definition A.26 (σ-finite measure space)**.** *A measure space (Definition A.18) with σ-finite measure (Definition A.25) is called a* σ-finite measure space*.*

Example A.27. Lebesgue measure (Theorem A.33) on the real numbers is not finite but σ-finite since

$$\bigcup_{k\in\mathbb{Z}} [k, k+1] = \mathbb{R} \text{ and } \lambda(\mathbb{R}) = 1.$$

Hence, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ is a σ-finite measure space (Definition A.26).

Definition A.28 (Simple function)**.** *Let $(\Omega, \mathcal{A})$ be a measurable space (Definition A.17). A map $f\colon \Omega \to \mathbb{R}$ is called a* simple function *if there is an $n \in \mathbb{N}$ and mutually disjoint measurable sets $A_i, \ldots, A_n \in \mathcal{A}$ (Definition A.17), as well as numbers $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, such that $f = \sum_{i=1}^n \alpha_i 1_{A_i}$.*

Remark. A measurable map that assumes only finitely many values is a simple function.

Definition A.29 (Simple function spaces)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Denote by $\mathbb{E}$ the vector space of simple functions (Definition A.28) on $(\Omega, \mathcal{A})$ and by*

$$\mathbb{E}^+ := \{f \in \mathbb{E} : f \geq 0\}$$

the cone of nonnegative simple functions.

Definition A.30 (Normal representation)**.** *If*

$$f = \sum_{i=1}^m \alpha_i 1_{A_i} \tag{A.1}$$

for some $m \in \mathbb{N}$ and for $\alpha_1, \ldots, \alpha_m \in (0, \infty)$, and for mutually disjoint sets $A_1, \ldots, A_m \in \mathcal{A}$, then Equation (A.1) is said to be a normal representation of f.

Definition A.31. *Define the map $I\colon \mathbb{E}^+ \to [0, \infty]$ by*

$$I(f) = \sum_{i=1}^m \alpha_i \mu(A_i)$$

if f has normal representation $f = \sum_{i=1} \alpha_i 1_{A_i}$ (Definition A.30).

Definition A.32 (Integral)**.** *If $f\colon \Omega \to [0, \infty]$ is measurable (Definition A.21), then we define the* integral *of f with respect to μ by*

$$\int f \, d\mu := \sup\left\{I(g) : g \in \mathbb{E}^+, g \leq f\right\} \quad \text{(Definition A.31).}$$

Theorem A.33 (Lebesgue measure)**.** *There exists a uniquely determined measure λ^n on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with property*

$$\lambda^n((a, b]) = \prod_{i=1}^n (b_i - a_i) \quad \text{for all } a, b \in \mathbb{R}^n \text{ with } a < b.$$

λ^n is called the Lebesgue measure *on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ or* Lebesgue-Borel measure*.*

Definition A.34 (Lebesgue σ-algebra)**.** *The Lebesgue σ-algebra is*

$$\mathcal{B}^*(\mathbb{R}^n) = \sigma(\mathcal{B}(\mathbb{R}^n) \cup \mathcal{N})$$

where $\mathcal{N}$ is the class of all Lebesgue-Borel null sets (sets where the Lebesgue-Borel measure (Theorem A.33) evaluates to zero).

Definition A.35 (Integral of a measurable function)**.** *A measurable function* $f\colon \Omega \to \overline{\mathbb{R}}$ *with* $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ *is* μ-integrable *if* $\int |f|\, d\mu < \infty$*. We write*

$$\mathcal{L}^1(\mu) := \mathcal{L}^1(\Omega, \mathcal{A}, \mu) := \left\{ f\colon \Omega \to \overline{\mathbb{R}} : f \text{ is measurable and and } \int |f|\, d\mu < \infty \right\}$$

Let $f^-(x) = \max(-f(x), 0)$ *and* $f^+ = \max(f(x), 0)$*. For* $f \in \mathcal{L}^1$*, we define the integral of* f *with respect to* μ *by*

$$\int f(\omega)\ \mu(d\omega) := \int f\ d\mu := \int f^+\ d\mu - \int f^- d\mu.$$

If we only have $\int f^-\ d\mu < \infty$ *or* $\int f^+\ d\mu < \infty$*, then we also define* $\int f\ d\mu$ *in the same way. In these cases* $+\infty$ *and* $-\infty$*, respectively, are possible.*

For $A \in \mathcal{A}$*, we define* $\int_A f\ d\mu := \int (f 1_A)\ d\mu$*.*

Definition A.36 (Lebesgue integral)**.** *Let* λ *be the Lebesgue measure (Theorem A.33) on* $\mathbb{R}^n$*. Let* $f\colon \mathbb{R}^n \to \mathbb{R}$ *be measurable with respect to* $\mathcal{B}^*(\mathbb{R}^n) - \mathcal{B}(\mathbb{R})$ *(Definition A.21) and* λ*-integretable (Definition A.35). Here* $\mathcal{B}^*(\mathbb{R}^n)$ *is the Lebesgue* σ*-algebra (Definition A.34) and* $\mathcal{B}(\mathbb{R})$ *is the Borel* σ*-algebra (Definition A.8). Then we call*

$$\int f\ d\lambda$$

the Lebesgue integral *of* f*. If* $A \in \mathcal{B}(\mathbb{R}^n)$ *and* $f\colon \mathbb{R}^n \to \mathbb{R}$ *is measurable, then we write*

$$\int_A f\ d\lambda := \int f\ 1_A\ d\lambda \quad .$$

Definition A.37 (Density)**.** *Let* μ *be a measure (Definition A.24) on* $(\Omega, \mathcal{A})$ *and let* $f : \Omega \to [0, \infty)$ *be a measurable map (Definition A.21). If* ν *is a measure that can be expressed as integral (Definition A.32)*

$$\nu(A) := \int (1_A f)\ d\mu \qquad \text{for } A \in \mathcal{A},$$

we say that ν *has* density f *with respect to* μ*.*

Definition A.38 (Transition kernel, Markov kernel)**.** *Let* $(\Omega_1, \mathcal{A}_1)$ *and* $(\Omega_2, \mathcal{A}_2)$ *be measurable spaces. A map*

$$\kappa : \Omega_1 \times \mathcal{A}_2 \to [0, \infty)$$

is called a σ*-finite* transition kernel *if:*

(i) For every $A_2 \in \mathcal{A}_2$*, the function* $\omega \mapsto \kappa(\omega, A_2)$ *is* $\mathcal{A}_1$*-measurable (Definition A.21).*

(ii) For every $\omega \in \Omega_1$*, the set function* $A_2 \mapsto \kappa(\omega, A_2)$ *is a* σ*-finite measure on* $(\Omega_2, \mathcal{A}_2)$ *(Definition A.25).*

If, in (ii), $\kappa(\omega, \cdot)$ *is a probability measure for every* ω *(i.e.* $\kappa(\omega, \Omega_2) = 1$*), then* κ *is called a* stochastic kernel *or* Markov kernel*. If instead* $\kappa(\omega, \Omega_2) \leq 1$ *for all* ω*, then* κ *is called* sub-Markov *or* sub-stochastic*.*

Definition A.39 (Stochastic process)**.** *Let* $I \subseteq \mathbb{R}$*. A family of random variables* $X = (X_t, t \in I)$ *(Definition A.20) on probability space* $(\Omega, \mathcal{A}, \mathbf{P})$ *(Definition A.17) with values in Polish space* $(E, \mathcal{E})$ *(Definition A.15) is called a stochastic process with index set (or time set)* I *and range* E*.*

Definition A.40 (Filtration)**.** *Let $\{\mathcal{F}_t\}_{t\in I}$ be a family of σ-algebras (Definition A.5) on Ω. The family is called a* filtration *if for all $s,t\in I$ with $s\le t$ we have*

$$\mathcal{F}_s\subseteq\mathcal{F}_t.$$

Definition A.41 (Adapted stochastic process)**.** *A stochastic process $X=(X_t,t\in I)$ (Definition A.39) is called* adapted *to filtration $\mathbb{F}=\{\mathcal{F}_t\}_{t\in I}$ (Definition A.40) if X_t is $\mathcal{F}_t$-measurable (Definition A.21) for all $t\in I$. If $\mathcal{F}_t=\sigma(X_s,s\le t)$ for all $t\in I$, then we denote by $\mathbb{F}=\sigma(X)$ the* filtration generated by *X.*

Definition A.42 (Martingales)**.** *Let $(\Omega,\mathcal{F},\mathbf{P})$ be a probability space (Definition A.19), $I\subseteq\mathbb{R}$, and let $\mathbb{F}$ be a filtration (Definition A.40). Let $X=(X_t)_{t\in I}$ be a real-valued, adapted stochastic process (Definition A.41) with $\mathbf{E}[|X_t|]<\infty$ for all $t\in I$. X is called (with respect to $\mathbb{F}$) a*

$$\begin{aligned}\text{martingale } &\text{if } \mathbf{E}[X_t\mid\mathcal{F}_s]=X_s \text{ for all } s,t\in I \text{ with } t>s,\\ \text{submartingale } &\text{if } \mathbf{E}[X_t\mid\mathcal{F}_s]\ge X_s \text{ for all } s,t\in I \text{ with } t>s,\\ \text{supermartingale } &\text{if } \mathbf{E}[X_t\mid\mathcal{F}_s]\le X_s \text{ for all } s,t\in I \text{ with } t>s.\end{aligned}$$

If $X_t\ge 0$ P-almost surely for all $t\in I$, X is called non-negative*.*

Theorem A.43 (Fubini)**.** *For $i\in\{1,2\}$, let $(\Omega_i,\mathcal{A}_i,\mu_i)$ be σ-finite measure spaces (Definition A.26). Let $f\colon\Omega_1\times\Omega_2\to\overline{R}$ be measurable with respect to product-σ-algebra $\mathcal{A}_1\otimes\mathcal{A}_2$. If $f\ge 0$ or $f\in\mathcal{L}^1(\mu_1\otimes\mu_2)$, then*

$$\begin{aligned}\omega_1&\mapsto\int f(\omega_1,\omega_2)\,\mu_2(d\omega_2) \quad \text{is } \mu_1\text{-a.e. defined and } \mathcal{A}_1\text{-measurable},\\ \omega_2&\mapsto\int f(\omega_1,\omega_2)\,\mu_1(d\omega_1) \quad \text{is } \mu_2\text{-a.e. defined and } \mathcal{A}_2\text{-measurable},\end{aligned}$$

and

$$\begin{aligned}\int_{\Omega_1\times\Omega_2} f\,d(\mu_1\otimes\mu_2)&=\int\left(\int f(\omega_1,\omega_2)\,\mu_2(d\omega_2)\right)\mu_1(d\omega_1)\\&=\int\left(\int f(\omega_1,\omega_2)\,\mu_1(d\omega_1)\right)\mu_2(d\omega_2)\quad.\end{aligned}$$

Definition A.44 (Absolute continuity, singularity)**.** *Let $\mu:\mathcal{A}\to[0,\infty]$ and $\nu:\mathcal{A}\to[0,\infty]$ be two measures (Definition A.24) on $(\Omega,\mathcal{A})$.*

(i) ν is called absolutely *continuous with respect to μ (symbolically $\nu\ll\mu$) if*

$$\forall A\in\mathcal{A}.\,\nu(A)=0\implies\mu(A)=0.$$

The measures μ and ν are called equivalent *(symbolically $\mu\approx\nu$) if $\nu\ll\mu$ and $\mu\ll\nu$.*

(ii) μ is called singular *to ν (symbolically $\mu\perp\nu$) if there exists an $A\in\mathcal{A}$ such that $\mu(A)=0$ and $\nu(A^c)=0$.*

Example A.45. Dirac measure δ_0 (remember $\delta_0(\{0\})=1$) is singular with respect to the Lebesgue measure λ (Theorem A.33), symbolically $\delta_0\perp\lambda$, since $\lambda(\{0\})=\delta_0(\mathbb{R}\setminus\{0\})=0$.

Example A.46. Measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, \infty]$ is absolutely continuous with respect to reference $\lambda([a, b)) = b - a$, the Lebesgue measure (Theorem A.33), if

$$\forall A \in \mathcal{B}(\mathbb{R}).\, \lambda(A) = 0 \implies \mu(A) = 0.$$

Theorem A.47 (Lebesgue's decomposition theorem)**.** *Let μ and ν be σ-finite measures $(\Omega, \mathcal{A})$. Then ν can be uniquely decomposed into an absolutely continuous part ν_a and a singular part ν_s (with respect to μ)*

$$\nu = \nu_a + \nu_s, \quad \text{where } \nu_a \ll \mu \text{ and } \nu_s \perp \mu.$$

ν_a has a density with respect to μ, and $\frac{d\nu_a}{d\mu}$ is a $\mathcal{A}$-measurable and finite μ-a.e..

Corollary A.48 (Radon-Nikodym theorem)**.** *Let μ and ν be σ-finite measures on $(\Omega, \mathcal{A})$. Then*

$$\nu \text{ has density w.r.t. } \mu \iff \nu \ll \mu.$$

In this case, $\frac{d\nu}{d\mu}$ is $\mathcal{A}$-measurable and finite μ-a.e. $\frac{d\nu}{d\mu}$ is called the Radon-Nikodym derivative *of ν with respect to μ.*

Lemma A.49 (Ville's Inequality (Ville, 1939))**.** *Let $(M_t)_{t \geq 1}$ be a non-negative supermartingale (Definition A.42). Then, for any real number $\alpha > 0$,*

$$\mathbf{P}\left(\sup_{t \geq 1} M_t \geq \alpha\right) \leq \frac{\mathbf{E}[M_1]}{\alpha}.$$

Appendix B

Proofs

B.1 Proof of Laplace's Method

Assume the conditions in Theorem 2.4.

Let $\mathbf{x} \in \mathbb{R}^d$ and $t \in (0, \infty)$. Define $\tilde{f}(\mathbf{x}) \coloneqq f(\mathbf{x}) - f(\mathbf{x}_*)$ and $\tilde{h}(\mathbf{x}) \coloneqq h(\mathbf{x}_*)h(\mathbf{x})$, then $\min_{\mathbf{x}\in K} \tilde{f}(\mathbf{x}) = \tilde{f}(\mathbf{x}_*) = 0$. Furthermore, define

$$\mathbf{y} \coloneqq t^{1/2}(\nabla^2 f(\mathbf{x}_*))^{1/2}(\mathbf{x} - \mathbf{x}_*).$$

Then

$$\mathbf{x} = t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*$$

and

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}.$$

To simplify notation define

$$\bar{f}_t(\mathbf{y}) \coloneqq \tilde{f}(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*) = f(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*) - f(\mathbf{x}_*)$$

and

$$\bar{h}_t(\mathbf{y}) \coloneqq h(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*).$$

We will use that

$$\begin{aligned}
\nabla \bar{f}_t(\mathbf{y}) &= t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\nabla f(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*), \\
\nabla \bar{f}_t(0) &= 0,
\end{aligned}$$

and that

$$\begin{aligned}
(\nabla^2 \bar{f}_t(\mathbf{y}))|_{\mathbf{y}=0} &= t^{-1}(\nabla^2 f(\mathbf{x}_*))^{-1}\nabla^2 f(t^{-1/2}\nabla^2 f(\mathbf{x}_*)^{-1/2}0 + \mathbf{x}_*) \\
&= t^{-1}(\nabla^2 f(\mathbf{x}_*))^{-1}\nabla^2 f(\mathbf{x}_*) \\
&= t^{-1}I.
\end{aligned}$$

Using a Taylor expansion (Theorem A.4) of $\bar{f}_t$ around 0 we have that

$$\bar{f}_t(\mathbf{y}) = \frac{1}{2t}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0) = \frac{1}{2t}\|\mathbf{y}\|^2 + o(1/t),$$

with $o(1/t)$ approaching zero at a rate of at least t.

Define

$$K_t \coloneqq \left\{ t^{1/2}(\nabla^2 f(\mathbf{x}_*))^{1/2}(\mathbf{x} - \mathbf{x}_*) : \mathbf{x} \in K \right\},$$

then we may rewrite $I(t)$ as follows:

$$\begin{aligned}
I(t) &= \int_K h(\mathbf{x}) \exp(-t f(\mathbf{x}))\, d\mathbf{x} \\
&= \int_K h(\mathbf{x}) \exp(-t(\tilde{f}(\mathbf{x}) + f(\mathbf{x}_*)))\, d\mathbf{x} \\
&= e^{-tf(\mathbf{x}_*)} \int_K h(\mathbf{x}) \exp(-t\tilde{f}(\mathbf{x})))\, d\mathbf{x} \\
&= e^{-tf(\mathbf{x}_*)} \int_K h(t^{1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*) \\
&\qquad\qquad \cdot \exp(-t\tilde{f}(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*)) \left| \det \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| \, d\mathbf{y} \\
&= \frac{\exp(-tf(\mathbf{x}_*))}{t^{d/2} |\det \nabla^2 f(\mathbf{x}_*)|^{1/2}} \int_{t^{1/2}(\nabla^2 f(\mathbf{x}_*))^{1/2}(K - \mathbf{x}_*)} \bar{h}_t(\mathbf{y}) \exp(-t\bar{f}(\mathbf{y}))\, d\mathbf{y} \\
&= \frac{\exp(-tf(\mathbf{x}_*))}{t^{d/2} |\det \nabla^2 f(\mathbf{x}_*)|^{1/2}} \int_{t^{1/2}(\nabla^2 f(\mathbf{x}_*))^{1/2}(K - \mathbf{x}_*)} \bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0) \right) \, d\mathbf{y} \\
&= \frac{\exp(-tf(\mathbf{x}_*))}{t^{d/2} |\det \nabla^2 f(\mathbf{x}_*)|^{1/2}} \int_{\mathbb{R}^d} \mathbf{1}\{\mathbf{y} \in K_t\} \bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0) \right) \, d\mathbf{y}.
\end{aligned}$$

For all $\mathbf{y} \in \mathbb{R}^d$, $R_t(\mathbf{y}, 0) = o(1/t)$ and $\lim_{t\to\infty} \bar{h}_t(\mathbf{y}) = h(\mathbf{x}_*)$, so we have limit

$$\lim_{t\to\infty} \mathbf{1}\{\mathbf{y} \in K_t\} \bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0) \right) = h(\mathbf{x}_*) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 \right).$$

Furthermore, for all $t \in (0, \infty)$ and all $\mathbf{y} \in \mathbb{R}^d$ we can bound

$$\begin{aligned}
\left| 1\{\mathbf{y} \in K_t\} \bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0) \right) \right| &= \left| 1\{\mathbf{y} \in K_t\} \bar{h}_t(\mathbf{y}) \exp\left(-t\bar{f}(\mathbf{y}) \right) \right| \\
&= \left| 1\{\mathbf{y} \in K_t\} \bar{h}_t(\mathbf{y}) \right| \cdot \left| \exp\left(-t\bar{f}(\mathbf{y}) \right) \right| \\
&\le \left| 1\{\mathbf{y} \in K_t\} \bar{h}_t(\mathbf{y}) \right| \\
&\le \left| 1\{\mathbf{y} \in K_t\} \max_{z \in K_t} \bar{h}_t(z) \right|.
\end{aligned}$$

Since

$$\begin{aligned}
\max_{\mathbf{y} \in K_t} \bar{h}_t(\mathbf{y}) &= \max_{\mathbf{y} \in K_t} h(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2}\mathbf{y} + \mathbf{x}_*) \\
&= \max_{\mathbf{x} \in K} h(t^{-1/2}(\nabla^2 f(\mathbf{x}_*))^{-1/2} t^{1/2} (\nabla^2 f(\mathbf{x}_*))^{1/2}(\mathbf{x} - \mathbf{x}_*) + \mathbf{x}_*) \\
&= \max_{\mathbf{x} \in K} h(\mathbf{x} - \mathbf{x}_* + \mathbf{x}_*) \\
&= \max_{\mathbf{x} \in K} h(\mathbf{x}),
\end{aligned}$$

h is continuous and $K \subseteq \mathbb{R}$ is compact,

$$\int_{\mathbb{R}^d} |1\{\mathbf{y} \in K_t\}| \max_{z \in K_t} \left| \bar{h}_t(z)\, d\mathbf{y} \right| = \int_K \left| \max_{\mathbf{x} \in K} h(\mathbf{x}) \right| < \infty,$$

that is

$$|1\{\mathbf{y} \in K_t\}| \max_{z \in K_t} \left|\bar{h}_t(z)\right|$$

is integrable in the Lebesgue-sense.

By the dominated convergence theorem and the fact that

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2\right)\, d\mathbf{y} = (2\pi)^{d/2}$$

is the normalizer of the d-dimensional standard normal density we have

$$\begin{aligned}
&\lim_{t\to\infty} \int_{\mathbb{R}^d} \mathbf{1}\{\mathbf{y} \in K_t\}\bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0)\right)\, d\mathbf{y} \\
&\quad = \int_{\mathbb{R}^d} \lim_{t\to\infty} \mathbf{1}\{\mathbf{y} \in K_t\}\bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0)\right)\, d\mathbf{y} \\
&\quad = \int_{\mathbb{R}^d} h(\mathbf{x}_*) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2\right)\, d\mathbf{y} \\
&\quad = h(\mathbf{x}_*)(2\pi)^{d/2}.
\end{aligned}$$

Define

$$I'(t) \coloneqq \frac{h(\mathbf{x}_*)}{|\det \nabla^2 f(\mathbf{x}_*)|^{1/2}} \left(\frac{2\pi}{t}\right)^{d/2} \exp(-tf(\mathbf{x}_*)).$$

Then

$$\begin{aligned}
&\lim_{t\to\infty} \frac{I(t)}{I'(t)} \\
&\quad = \lim_{t\to\infty} \left(\frac{\exp(-tf(\mathbf{x}_*))}{t^{d/2}|\det \nabla^2 f(\mathbf{x}_*)|^{1/2}} \int_{\mathbb{R}^d} \mathbf{1}\{\mathbf{y} \in K_t\}\bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0)\right)\, d\mathbf{y}\right) \\
&\qquad \frac{t^{d/2}|\det \nabla^2 f(\mathbf{x}_*)|^{1/2}}{\exp(-tf(\mathbf{x}_*))h(\mathbf{x}_*)(2\pi)^{d/2}} \\
&\quad = \lim_{t\to\infty} \frac{1}{h(\mathbf{x}_*)(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathbf{1}\{\mathbf{y} \in K_t\}\bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0)\right)\, d\mathbf{y} \\
&\quad = \frac{1}{h(\mathbf{x}_*)(2\pi)^{d/2}} \lim_{t\to\infty} \int_{\mathbb{R}^d} \mathbf{1}\{\mathbf{y} \in K_t\}\bar{h}_t(\mathbf{y}) \exp\left(-\frac{1}{2}\|\mathbf{y}\|^2 + R_t(\mathbf{y}, 0)\right)\, d\mathbf{y} \\
&\quad = \frac{h(\mathbf{x}_*)(2\pi)^{d/2}}{h(\mathbf{x}_*)(2\pi)^{d/2}} = 1.
\end{aligned}$$

Hence, by Definition A.2 $I(t)$ and $I'(t)$ are asymptotically equivalent, symbolically

$$I(t) \sim I'(t).$$

B.2 Proof of the Prior Likelihood Mixing Theorem

Assume the conditions in Theorem 2.3.

Next, define the concept of a *likelihood ratio* and *marginal likelihood ratio.*

Definition B.1 (Likelihood Ratio (Kirschner et al., 2025))**.** *For all $\boldsymbol{\nu}, \boldsymbol{\theta} \in \Theta$ and all steps $t \in \mathbb{N}$ we call*

$$R_t(\boldsymbol{\nu}, \boldsymbol{\theta}) \coloneqq \prod_{s=1}^{t} \frac{p_s(\mathbf{y}_s \mid \boldsymbol{\nu})}{p_s(\mathbf{y}_s \mid \boldsymbol{\theta})}$$

their likelihood ratio *at step t.*

Definition B.2 (Marginal Likelihood Ratio (Kirschner et al., 2025))**.** *For all* $\boldsymbol{\nu}, \boldsymbol{\theta} \in \Theta$*, steps* $t \in \mathbb{N}$ *and data-independent prior distributions* $\mu_0 \in \mathscr{P}(\Theta)$*. We call*

$$Q_t(\boldsymbol{\theta}) \coloneqq \int R_t(\boldsymbol{\nu}; \boldsymbol{\theta}) \, d\mu_0(\boldsymbol{\nu}) = \frac{\int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})}{\prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\theta})}$$

the marginal likelihood ratio *of* $\boldsymbol{\theta} \in \Theta$ *with respect to* μ_0 *and* $(Q_s(\boldsymbol{\theta}), s \in \mathbb{N})$ *the* marginal likelihood ratio process *with respect to* μ_0*.*

Let $\mathcal{F}' = (\mathcal{F}'_t, t \in \mathbb{N})$ be a filtration (Definition A.40) on Ω for which for all $t \in \mathbb{N}$

$$\mathcal{F}'_t = \sigma\left((X_s, Y_s), s \in [t]\right) \quad \text{(Definition A.23)}$$

By construction, $(Q_s, s \in \mathbb{N})$ is an adapted stochastic process (Definition A.41) with respect to $\mathcal{F}'$. We first show that the stochastic process (Definition A.39) is a nonnegative martingale (Definition A.42) with $\mathbf{E}[Q_1(\boldsymbol{\theta}^*)] = 1$.

Using Fubini's theorem (Theorem A.43) we have

$$\begin{aligned}
\mathbf{E}[Q_t(\boldsymbol{\theta}^*) \mid \mathcal{F}'_{t-1}] &= \int \frac{\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})}{p_t(\mathbf{y} \mid \boldsymbol{\theta}^*) \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \, dP_{\boldsymbol{\theta}^*}(\mathbf{y} \mid \mathbf{x}_t) \\
&= \int \frac{\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})}{\prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \, d\xi(\mathbf{y}) \\
&= \frac{1}{\prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \int \left(\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) \right) d\xi(\mathbf{y}) \\
&= \frac{1}{\prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \int \left(\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\xi(\mathbf{y}) \right) d\mu_0(\boldsymbol{\nu}) \\
&= \frac{1}{\prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \int \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \underbrace{\left(\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \, d\xi(\mathbf{y}) \right)}_{=1} d\mu_0(\boldsymbol{\nu}) \\
&= \frac{\int \prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})}{\prod_{s=1}^{t-1} p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} = Q_{t-1}(\boldsymbol{\theta}^*).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbf{E}[Q_1(\boldsymbol{\theta}^*)] &= \int \frac{\int p_1(\mathbf{y} \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})}{p_1(\mathbf{y} \mid \boldsymbol{\theta}^*)} \, dP_{\boldsymbol{\theta}^*}(\mathbf{y}_1 \mid \mathbf{x}_1) \\
&= \int \frac{\int p_1(\mathbf{y} \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})}{p_1(\mathbf{y} \mid \boldsymbol{\theta}^*)} \, dP_{\boldsymbol{\theta}^*}(\mathbf{y}_1 \mid \mathbf{x}_1) \\
&= \int \left(\int p_1(\mathbf{y} \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) \right) d\xi(\mathbf{y}) \\
&= \int \underbrace{\left(\int p_1(\mathbf{y} \mid \boldsymbol{\nu}) \, d\xi(\mathbf{y}) \right)}_{=1} d\mu_0(\boldsymbol{\nu}) = 1
\end{aligned}$$

Since for all $\boldsymbol{\nu} \in \Theta$, $\mathbf{y} \in \mathcal{Y}$, $t \in \mathbb{N}$ and $\mathbf{x} \in \mathcal{X}$

$$p_{\boldsymbol{\nu}}(\mathbf{y} \mid \mathbf{x}) \geq 0 \text{ and } \mu_0 \geq 0,$$

we have that $Q_t(\boldsymbol{\theta}^*) \geq 0$ $\mathbf{P}$-almost surely. Hence, by Definition A.42, $(Q_t(\boldsymbol{\theta}^*), t \in \mathbb{N})$ is a non-negative martingale.

Since every martingale is a supermartingale (Definition A.42), $(Q_t(\boldsymbol{\theta}^*), t \in \mathbb{N})$ is a non-negative supermartingale and we may apply Ville's inequality (Lemma A.49).

Applying Ville's inequality yields

$$\mathbf{P}\left(\sup_{t\in\mathbb{N}} Q_t(\boldsymbol{\theta}^*) \geq \frac{1}{\delta}\right) \leq \delta.$$

Since

$$\begin{aligned}
&\mathbf{P}\left(\sup_{t\in\mathbb{N}} Q_t(\boldsymbol{\theta}^*) \geq \frac{1}{\delta}\right) \leq \delta \\
\implies &\mathbf{P}\left(\sup_{t\in\mathbb{N}} Q_t(\boldsymbol{\theta}^*) > \frac{1}{\delta}\right) \leq \delta \\
\iff &\mathbf{P}\left(\sup_{t\in\mathbb{N}} Q_t(\boldsymbol{\theta}^*) \leq \frac{1}{\delta}\right) \geq 1-\delta \\
\iff &\mathbf{P}\left(\forall t \in \mathbb{N}:\ Q_t(\boldsymbol{\theta}^*) \leq \frac{1}{\delta}\right) \geq 1-\delta \\
\iff &\mathbf{P}\left(\forall t \in \mathbb{N}:\ \log Q_t(\boldsymbol{\theta}^*) \leq \log\frac{1}{\delta}\right) \geq 1-\delta \\
\iff &\mathbf{P}\left(\forall t \in \mathbb{N}:\ \log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_0(\boldsymbol{\nu}) + L_t(\boldsymbol{\theta}^*) \leq \log\frac{1}{\delta}\right) \geq 1-\delta \\
\iff &\mathbf{P}\left(\forall t \in \mathbb{N}:\ L_t(\boldsymbol{\theta}^*) \leq \log\frac{1}{\delta} - \log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_0(\boldsymbol{\nu})\right) \geq 1-\delta,
\end{aligned}$$

the sequence $(C_t(\beta_t^{\mathrm{plm}}(\delta), t \in \mathbb{N})$ with

$$\beta_t^{\mathrm{plm}}(\delta) = \log\frac{1}{\delta} - \log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_0(\boldsymbol{\nu})$$

is a confidence sequence for $\boldsymbol{\theta}^*$ at level δ.

B.3 Proof of the Sequential Likelihood Mixing Theorem

Assume the conditions in Theorem 2.5.

Definition B.3. *For all steps $t \in \mathbb{N}$ and parameters $\boldsymbol{\theta} \in \Theta$ define the* sequential marginal likelihood ratio $\mathcal{S}_t(\boldsymbol{\theta})$ *with respect to mixing distributions $\mu_0, \mu_1, \cdots \in \mathscr{P}(\Theta)$ as follows*

$$\mathcal{S}_t(\boldsymbol{\theta}) \coloneqq \prod_{s=1}^{t} \frac{\int p_s(\mathbf{y}_s \mid \boldsymbol{\nu})\, d\mu_{s-1}(\boldsymbol{\nu})}{p_s(\mathbf{y}_s \mid \boldsymbol{\theta})}.$$

Since the product of zero terms is 1, $\mathbf{E}[\mathcal{S}_0(\boldsymbol{\theta}^*)] = 1$.

By Fubini's theorem (Theorem A.43), for all steps $t \in \mathbb{N}$

$$\begin{aligned}
\mathbf{E}[\mathcal{S}_t(\boldsymbol{\theta}^*) \mid \mathcal{F}'_t] &= \int \left(\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \, d\mu_{t-1}(\boldsymbol{\nu}) \right) \prod_{s=1}^{t-1} \frac{\int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu})}{p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \, d\xi(\mathbf{y}) \\
&= \prod_{s=1}^{t-1} \frac{\int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu})}{p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \int \left(\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \, d\mu_{t-1}(\boldsymbol{\nu}) \right) \, d\xi(\mathbf{y}) \\
&= \prod_{s=1}^{t-1} \frac{\int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu})}{p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} \int \left(\int p_t(\mathbf{y} \mid \boldsymbol{\nu}) \, d\xi(\mathbf{y}) \right) \, d\mu_{t-1}(\boldsymbol{\nu}) \\
&= \prod_{s=1}^{t-1} \frac{\int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu})}{p_s(\mathbf{y}_s \mid \boldsymbol{\theta}^*)} = \mathcal{S}_{t-1}(\boldsymbol{\theta}^*).
\end{aligned}$$

It follows from Definition A.42 that $(\mathcal{S}_s(\boldsymbol{\theta}^*), s \in \mathbb{N})$ is a martingale. Moreover, since for all steps $t \in \mathbb{N}$, parameters $\boldsymbol{\nu} \in \Theta$ and all $(\mathbf{x}_t, \mathbf{y}_t) \in (\mathcal{X} \times \mathcal{Y})$,

$$\mu_{t-1} \geq 0 \text{ and } p_t(\mathbf{y}_t \mid \boldsymbol{\nu}) \geq 0,$$

$(\mathcal{S}_t(\boldsymbol{\theta}^*), t \in \mathbb{N})$ is a nonnegative martingale (Definition A.42).

Ville's inequality (Lemma A.49) implies that

$$\mathbf{P}\left(\exists t \in \mathbb{N} : \ \mathcal{S}_t(\boldsymbol{\theta}^*) \geq \frac{1}{\delta} \right) \leq \delta. \tag{B.1}$$

Equation (B.1) implies

$$\mathbf{P}\left(\forall t \in \mathbb{N} : \ \mathcal{S}_t(\boldsymbol{\theta}^*) \leq \frac{1}{\delta} \right) \geq 1 - \delta.$$

and

$$\underbrace{\mathbf{P}\left(\forall t \in \mathbb{N} : \ \log \mathcal{S}_t(\boldsymbol{\theta}^*) \leq \log \frac{1}{\delta} \right)}_{=(*)} \geq 1 - \delta.$$

Since for all $t \in \mathbb{N}$

$$\log \mathcal{S}_t(\boldsymbol{\theta}^*) = L_t(\boldsymbol{\theta}^*) + \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu}),$$

we have

$$(*) = \mathbf{P}\left(\forall t \in \mathbb{N} : \ L_t(\boldsymbol{\theta}^*) \leq \log \frac{1}{\delta} - \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu}) \right) \geq 1 - \delta.$$

Hence, $(C_t(\beta_t^{\mathrm{slm}}(\delta)), t \in \mathbb{N})$ with

$$\beta_t^{\mathrm{slm}}(\delta) = \log \frac{1}{\delta} - \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu})$$

is a confidence sequence for $\boldsymbol{\theta}^*$ at level δ.

B.4 Proof of the Mixing Equivalence

Assume the conditions in Theorem 2.6.

Furthermore, for all $s \in \mathbb{N}_0$, define

$$M_s \coloneqq \int \prod_{u=1}^{s} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}).$$

Then for all $t \in \mathbb{N}$

$$\begin{aligned}
\sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu}) &= \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) M_{s-1}^{-1} \prod_{u=1}^{s-1} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) \\
&= \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \prod_{u=1}^{s-1} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) - \sum_{s=0}^{t-1} \log M_s \\
&= \sum_{s=1}^{t} \log \int \prod_{u=1}^{s} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) - \sum_{s=0}^{t-1} \log M_s \\
&= \sum_{s=1}^{t} \log M_s - \sum_{s=0}^{t-1} \log M_s \\
&= \log M_t - \log M_0 \\
&= \log \int \prod_{u=1}^{t} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) - \log \int \prod_{u=1}^{0} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) \\
&= \log \int \prod_{u=1}^{t} p_u(\mathbf{y}_u \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu})
\end{aligned}$$

implies that

$$\sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu}) = \log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}). \tag{B.2}$$

Theorems 2.3 and 2.5 and Equation (B.2) imply that for all $t \in \mathbb{N}$

$$\begin{aligned}
\beta_t^{\mathrm{plm}}(\delta) &= \log \frac{1}{\delta} - \log \int \prod_{s=1}^{t} p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_0(\boldsymbol{\nu}) \\
&= \log \frac{1}{\delta} - \sum_{s=1}^{t} \log \int p_s(\mathbf{y}_s \mid \boldsymbol{\nu}) \, d\mu_{s-1}(\boldsymbol{\nu}) = \beta_t^{\mathrm{slm}}(\delta).
\end{aligned}$$

B.5 Proof of First DDPM Lemma

Given a fixed $T \in \mathbb{N}$, we prove Lemma 3.1 by induction over $\tau \in [T]$. We want to show that for all $\tau \in \mathbb{N}$ and $\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(\tau-1)} \in \Theta$

$$q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(0)}) = \mathcal{N}(\boldsymbol{\theta}^{(\tau)}; \sqrt{\bar{\alpha}_\tau}\, \boldsymbol{\theta}^{(0)}, \ (1 - \bar{\alpha}_\tau) I).$$

with $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$.

Base case. If $\tau = 1$, by definition of the forward transition, for all $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(0)} \in \Theta$

$$q(\boldsymbol{\theta}^{(1)} \mid \boldsymbol{\theta}^{(0)}) = \mathcal{N}\left(\boldsymbol{\theta}^{(1)}; \sqrt{\alpha_1}\, \boldsymbol{\theta}^{(0)}, \ (1 - \alpha_1) I\right),$$

which coincides with the desired form since $\bar{\alpha}_1 = \alpha_1$ and $1 - \bar{\alpha}_1 = 1 - \alpha_1$.

Induction step. Assume that for some $\tau \in \{2, \dots, T\}$ the statement holds at $\tau - 1$, i.e., for all $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(\tau-1)} \in \Theta$,

$$q(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(0)}) = \mathcal{N}\left(\boldsymbol{\theta}^{(\tau-1)}; \sqrt{\bar{\alpha}_{\tau-1}}\boldsymbol{\theta}^{(0)},\ (1 - \bar{\alpha}_{\tau-1})I\right).$$

Then we may reparameterize the random variable $\boldsymbol{\vartheta}^{(\tau-1)}$ as follows:

$$\boldsymbol{\vartheta}^{(\tau-1)} \overset{d}{=} \sqrt{\bar{\alpha}_{\tau-1}}\boldsymbol{\vartheta}^{(0)} + \sqrt{1 - \bar{\alpha}_{\tau-1}}\,\mathbf{r}^{(\tau-1)}, \quad \mathbf{r}^{(\tau-1)} \sim \mathcal{N}(\mathbf{0}, I) \tag{B.3}$$

with $\mathbf{r}^{(\tau-1)}$ independent of $\boldsymbol{\vartheta}^{(0)}$. The definition of forward transitions implies that for all $\tau \in [T]$

$$\boldsymbol{\vartheta}^{(\tau)} \overset{d}{=} \sqrt{\alpha_\tau}\,\boldsymbol{\vartheta}^{(\tau-1)} + \sqrt{1 - \alpha_\tau}\,\tilde{\mathbf{r}}^{(\tau-1)}, \quad \tilde{\mathbf{r}}^{(\tau-1)} \sim \mathcal{N}(\mathbf{0}, I), \tag{B.4}$$

with $\tilde{\mathbf{r}}^{(\tau-1)}$ independent of $\boldsymbol{\vartheta}^{(\tau-1)}$.

Plugging Equation (B.3) into Equation (B.4) yields

$$\begin{aligned}\boldsymbol{\vartheta}^{(\tau)} &\overset{d}{=} \sqrt{\alpha_\tau}\left(\sqrt{\bar{\alpha}_{\tau-1}}\boldsymbol{\vartheta}^{(0)} + \sqrt{1 - \bar{\alpha}_{\tau-1}}\,\mathbf{r}^{(\tau-1)}\right) + \sqrt{1 - \alpha_\tau}\,\tilde{\mathbf{r}}^{(\tau-1)} \\ &= \sqrt{\bar{\alpha}_\tau}\,\boldsymbol{\vartheta}^{(0)} + \sqrt{\alpha_\tau(1 - \bar{\alpha}_{\tau-1})}\,\mathbf{r}^{(\tau-1)} + \sqrt{1 - \alpha_\tau}\,\tilde{\mathbf{r}}^{(\tau-1)}.\end{aligned}$$

with $\boldsymbol{\vartheta}^{(0)}, \mathbf{r}^{(\tau-1)}, \tilde{\mathbf{r}}^{(\tau-1)}$ mutually independent and $\mathbf{r}^{(\tau-1)}, \tilde{\mathbf{r}}^{(\tau-1)} \sim \mathcal{N}(\mathbf{0}, I)$. This shows that for all $\boldsymbol{\theta}^{(0)} \in \Theta$, given $\boldsymbol{\vartheta}^{(0)} = \boldsymbol{\theta}^{(0)}$, the random variable $\boldsymbol{\vartheta}^{(\tau)}$ is Gaussian. Its conditional mean is

$$\mathbf{E}_{\boldsymbol{\vartheta}^{(\tau)}}\left[\boldsymbol{\vartheta}^{(\tau)} \,\middle|\, \boldsymbol{\vartheta}^{(0)} = \boldsymbol{\theta}^{(0)}\right] = \sqrt{\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(0)}.$$

Using the independence of $\mathbf{r}^{(\tau-1)}$ and $\tilde{\mathbf{r}}^{(\tau-1)}$, its conditional covariance is

$$\begin{aligned}\mathrm{Cov}_{\boldsymbol{\vartheta}^{(\tau)}}\left(\boldsymbol{\vartheta}^{(\tau)} \mid \boldsymbol{\vartheta}^{(0)} = \boldsymbol{\theta}^{(0)}\right) &= \alpha_\tau(1 - \bar{\alpha}_{\tau-1})I + (1 - \alpha_\tau)I \\ &= (1 - \alpha_\tau\bar{\alpha}_{\tau-1})I = (1 - \bar{\alpha}_\tau)I.\end{aligned}$$

Therefore, for all $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(\tau)} \in \Theta$, we have

$$q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(0)}) = \mathcal{N}\left(\boldsymbol{\theta}^{(\tau)}; \sqrt{\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(0)},\ (1 - \bar{\alpha}_\tau)I\right).$$

B.6 Proof of Second DDPM Lemma

We prove Lemma 3.2. Let $\tau \in [T]$ and $\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(\tau-1)}, \boldsymbol{\theta}^{(0)} \in \Theta$. Using Bayes' rule (first step), the definition of the forward process, and Lemma 3.1 (second step) we obtain

$$
\begin{aligned}
&q(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}) \\
&\quad = q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(\tau-1)}) \, \frac{q(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(0)})}{q(\boldsymbol{\theta}^{(\tau)} \mid \boldsymbol{\theta}^{(0)})} \\
&\quad = \mathcal{N}(\boldsymbol{\theta}^{(\tau)}; \sqrt{\alpha_\tau}\, \boldsymbol{\theta}^{(\tau-1)}, \beta_\tau I) \, \frac{\mathcal{N}(\boldsymbol{\theta}^{(\tau-1)}; \sqrt{\bar{\alpha}_{\tau-1}}\, \boldsymbol{\theta}^{(0)}, (1-\bar{\alpha}_{\tau-1})I)}{\mathcal{N}\left(\boldsymbol{\theta}^{(\tau)}; \sqrt{\bar{\alpha}_\tau}\boldsymbol{\theta}^{(0)}, (1-\bar{\alpha}_\tau)I\right)} \\
&\quad \propto \exp\left(-\frac{\|\boldsymbol{\theta}^{(\tau)} - \sqrt{1-\beta_\tau}\boldsymbol{\theta}^{(\tau-1)}\|^2}{2\beta_\tau} - \frac{\|\boldsymbol{\theta}^{(\tau-1)} - \sqrt{\bar{\alpha}_{\tau-1}}\boldsymbol{\theta}^{(0)}\|^2}{2(1-\bar{\alpha}_{\tau-1})} + \frac{\|\boldsymbol{\theta}^{(\tau)} - \sqrt{\bar{\alpha}_\tau}\boldsymbol{\theta}^{(0)}\|^2}{2(1-\bar{\alpha}_\tau)} \right) \\
&\quad \propto \exp\left(-\frac{1}{2} \underbrace{\left(\frac{\|\boldsymbol{\theta}^{(\tau)} - \sqrt{1-\beta_\tau}\boldsymbol{\theta}^{(\tau-1)}\|^2}{\beta_\tau} + \frac{\|\boldsymbol{\theta}^{(\tau-1)} - \sqrt{\bar{\alpha}_{\tau-1}}\boldsymbol{\theta}^{(0)}\|^2}{1-\bar{\alpha}_{\tau-1}} \right)}_{=:(*)} \right).
\end{aligned}
$$

Rewriting $(*)$

$$
\begin{aligned}
(*) &= \frac{\|\boldsymbol{\theta}^{(\tau)} - \sqrt{1-\beta_\tau}\,\boldsymbol{\theta}^{(\tau-1)}\|^2}{\beta_\tau} + \frac{\|\boldsymbol{\theta}^{(\tau-1)} - \sqrt{\bar{\alpha}_{\tau-1}}\,\boldsymbol{\theta}^{(0)}\|^2}{1-\bar{\alpha}_{\tau-1}} \\
&= \frac{1}{\beta_\tau}\Big(\|\boldsymbol{\theta}^{(\tau)}\|^2 - 2\sqrt{1-\beta_\tau}\,\boldsymbol{\theta}^{(\tau)\top}\boldsymbol{\theta}^{(\tau-1)} + (1-\beta_\tau)\|\boldsymbol{\theta}^{(\tau-1)}\|^2\Big) \\
&\quad + \frac{1}{1-\bar{\alpha}_{\tau-1}}\Big(\|\boldsymbol{\theta}^{(\tau-1)}\|^2 - 2\sqrt{\bar{\alpha}_{\tau-1}}\,\boldsymbol{\theta}^{(\tau-1)\top}\boldsymbol{\theta}^{(0)} + \bar{\alpha}_{\tau-1}\|\boldsymbol{\theta}^{(0)}\|^2\Big) \\
&= \left(\frac{1-\beta_\tau}{\beta_\tau} + \frac{1}{1-\bar{\alpha}_{\tau-1}}\right)\|\boldsymbol{\theta}^{(\tau-1)}\|^2 \\
&\quad - 2\boldsymbol{\theta}^{(\tau-1)\top}\left(\frac{\sqrt{1-\beta_\tau}}{\beta_\tau}\,\boldsymbol{\theta}^{(\tau)} + \frac{\sqrt{\bar{\alpha}_{\tau-1}}}{1-\bar{\alpha}_{\tau-1}}\,\boldsymbol{\theta}^{(0)}\right) + \text{const.} \\
&= \boldsymbol{\theta}^{(\tau-1)\top} \underbrace{\left(\frac{1-\beta_\tau}{\beta_\tau} + \frac{1}{1-\bar{\alpha}_{\tau-1}}\right) I}_{=:A_\tau} \boldsymbol{\theta}^{(\tau-1)} - 2 \underbrace{\left(\frac{\sqrt{1-\beta_\tau}}{\beta_\tau}\,\boldsymbol{\theta}^{(\tau)} + \frac{\sqrt{\bar{\alpha}_{\tau-1}}}{1-\bar{\alpha}_{\tau-1}}\,\boldsymbol{\theta}^{(0)}\right)^\top}_{=:b_\tau^\top} \boldsymbol{\theta}^{(\tau-1)} + \text{const.}
\end{aligned}
$$

Hence,

$$
(*) = \left(\boldsymbol{\theta}^{(\tau-1)} - A_\tau^{-1}b_\tau\right)^\top A_\tau \left(\boldsymbol{\theta}^{(\tau-1)} - A_\tau^{-1}b_\tau\right) \; + \; \text{const.}
$$

Using $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \alpha_\tau \bar{\alpha}_{\tau-1}$,

$$
A_\tau = \left(\frac{1-\beta_\tau}{\beta_\tau} + \frac{1}{1-\bar{\alpha}_{\tau-1}}\right) I = \frac{1-\bar{\alpha}_\tau}{\beta_\tau(1-\bar{\alpha}_{\tau-1})} I, \qquad A_\tau^{-1} = \frac{\beta_\tau(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau} I.
$$

Hence

$$
A_\tau^{-1}b_\tau = \frac{\beta_\tau(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\left(\frac{\sqrt{1-\beta_\tau}}{\beta_\tau}\,\boldsymbol{\theta}^{(\tau)} + \frac{\sqrt{\bar{\alpha}_{\tau-1}}}{1-\bar{\alpha}_{\tau-1}}\,\boldsymbol{\theta}^{(0)}\right) = \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(\tau)} + \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(0)}.
$$

Define

$$\tilde{\mu}_\tau(\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}) \coloneqq \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(0)} + \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\,\boldsymbol{\theta}^{(\tau)}, \qquad \tilde{\beta}_\tau \coloneqq \frac{\beta_\tau(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}.$$

Then

$$\begin{aligned} &q\Big(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}\Big) \\ &\quad\propto \left(\boldsymbol{\theta}^{(\tau-1)} - A_\tau^{-1} b_\tau\right)^\top A_\tau \left(\boldsymbol{\theta}^{(\tau-1)} - A_\tau^{-1} b_\tau\right) \\ &\quad\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}^{(\tau-1)} - \tilde{\mu}(\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)})\right)^\top \tilde{\beta}_\tau^{-1} I (\boldsymbol{\theta}^{(\tau-1)} - \tilde{\mu}(\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}))\right). \end{aligned}$$

Thus

$$q\Big(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}\Big) = \mathcal{N}\Big(\boldsymbol{\theta}^{(\tau-1)};\, \tilde{\mu}_\tau(\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(0)}),\, \tilde{\beta}_\tau I\Big).$$

B.7 Proof of Third DDPM Lemma

We prove Lemma 3.3. Let $\tau \in \{2, \ldots, T\}$. By definition,

$$\mathcal{L}_{\tau-1} = \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\Big[D_{\mathrm{KL}}(q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}) \,\|\, p_\phi(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}))\Big].$$

Write the KL divergence as an inner expectation:

$$\begin{aligned} \mathcal{L}_{\tau-1} &= \mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\left[\mathbf{E}_{\boldsymbol{\vartheta}^{(\tau-1)}}\left[\log q(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}) - \log p_\phi(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}) \,\middle|\, \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}\right]\right] \\ &= -\mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\left[\mathbf{E}_{\boldsymbol{\vartheta}^{(\tau-1)}}\left[\log p_\phi(\boldsymbol{\vartheta}^{(\tau-1)} \mid \boldsymbol{\vartheta}^{(\tau)}) \,\middle|\, \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}\right]\right] + \text{const.}. \end{aligned}$$

The Gaussian parameterization of the reverse transition implies that for all $\boldsymbol{\theta}^{(\tau)}, \boldsymbol{\theta}^{(\tau-1)} \in \Theta$

$$-\log p_\phi(\boldsymbol{\theta}^{(\tau-1)} \mid \boldsymbol{\theta}^{(\tau)}) = \frac{1}{2\sigma_\tau^2}\left\|\boldsymbol{\theta}^{(\tau-1)} - \mu_\phi(\boldsymbol{\theta}^{(\tau)}, \tau)\right\|^2 + \text{const.}.$$

Hence

$$\mathcal{L}_{\tau-1} = \frac{1}{2\sigma_\tau^2}\,\mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\left[\mathbf{E}_{\boldsymbol{\vartheta}^{(\tau-1)}}\left[\left\|\boldsymbol{\vartheta}^{(\tau-1)} - \mu_\phi(\boldsymbol{\vartheta}^{(\tau)}, \tau)\right\|^2 \,\middle|\, \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}\right]\right] + \text{const.} \tag{B.5}$$

Lemma 3.2 allows us to simplify the inner expectation

$$\begin{aligned} &\mathbf{E}_{\boldsymbol{\vartheta}^{(\tau-1)}}\left[\left\|\boldsymbol{\vartheta}^{(\tau-1)} - \mu_\phi(\boldsymbol{\vartheta}^{(\tau)}, \tau)\right\|^2 \,\middle|\, \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}\right] \\ &\quad= \|\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}, \tau)\|^2 - 2\mu_\phi^T(\boldsymbol{\vartheta}^{(\tau)}, \tau)\mathbf{E}_{\boldsymbol{\vartheta}^{(\tau-1)}}\left[\boldsymbol{\vartheta}^{(\tau-1)} \,\middle|\, \boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}\right] + \text{const.} \\ &\quad= \|\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}, \tau)\|^2 - 2\mu_\phi^T(\boldsymbol{\vartheta}^{(\tau)}, \tau)\tilde{\mu}_\tau(\boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)}) + \text{const.} \\ &\quad= \|\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}, \tau) - \tilde{\mu}_\tau(\boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)})\|^2 + \text{const.} \end{aligned} \tag{B.6}$$

Substituting Equation (B.6) into Equation (B.5) yields

$$\mathcal{L}_{\tau-1} = \frac{1}{2\sigma_\tau^2}\,\mathbf{E}_{\boldsymbol{\vartheta}^{(0:T)}}\left[\left\|\mu_\phi(\boldsymbol{\vartheta}^{(\tau)}, \tau) - \tilde{\mu}_\tau(\boldsymbol{\vartheta}^{(\tau)}, \boldsymbol{\vartheta}^{(0)})\right\|^2\right] + \text{const.},$$

which is the claim.

Appendix C

Plots and Algorithms

C.1 Confidence Coefficients

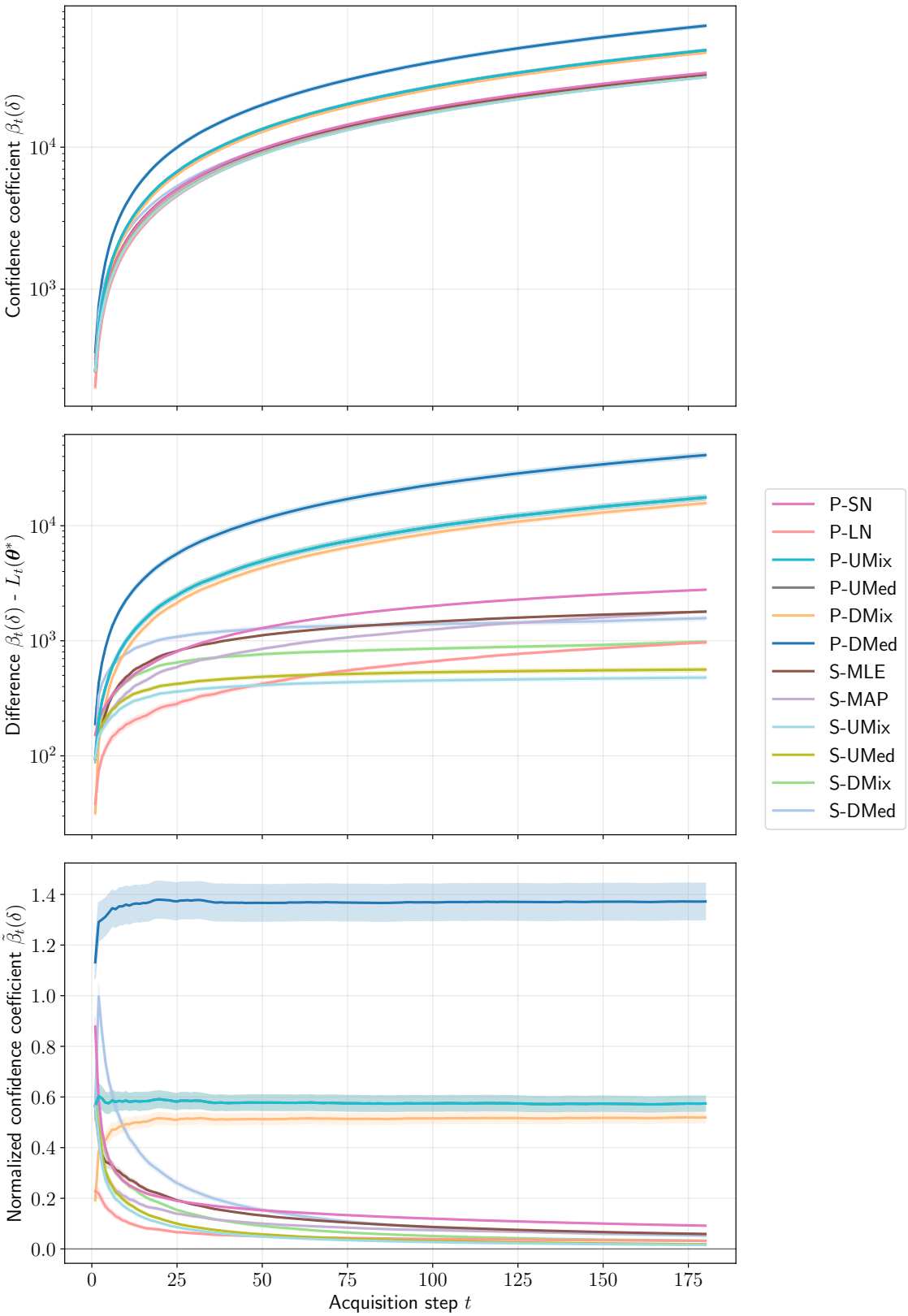


Figure C.1: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 1$ and starting step $t_0 = 1$.

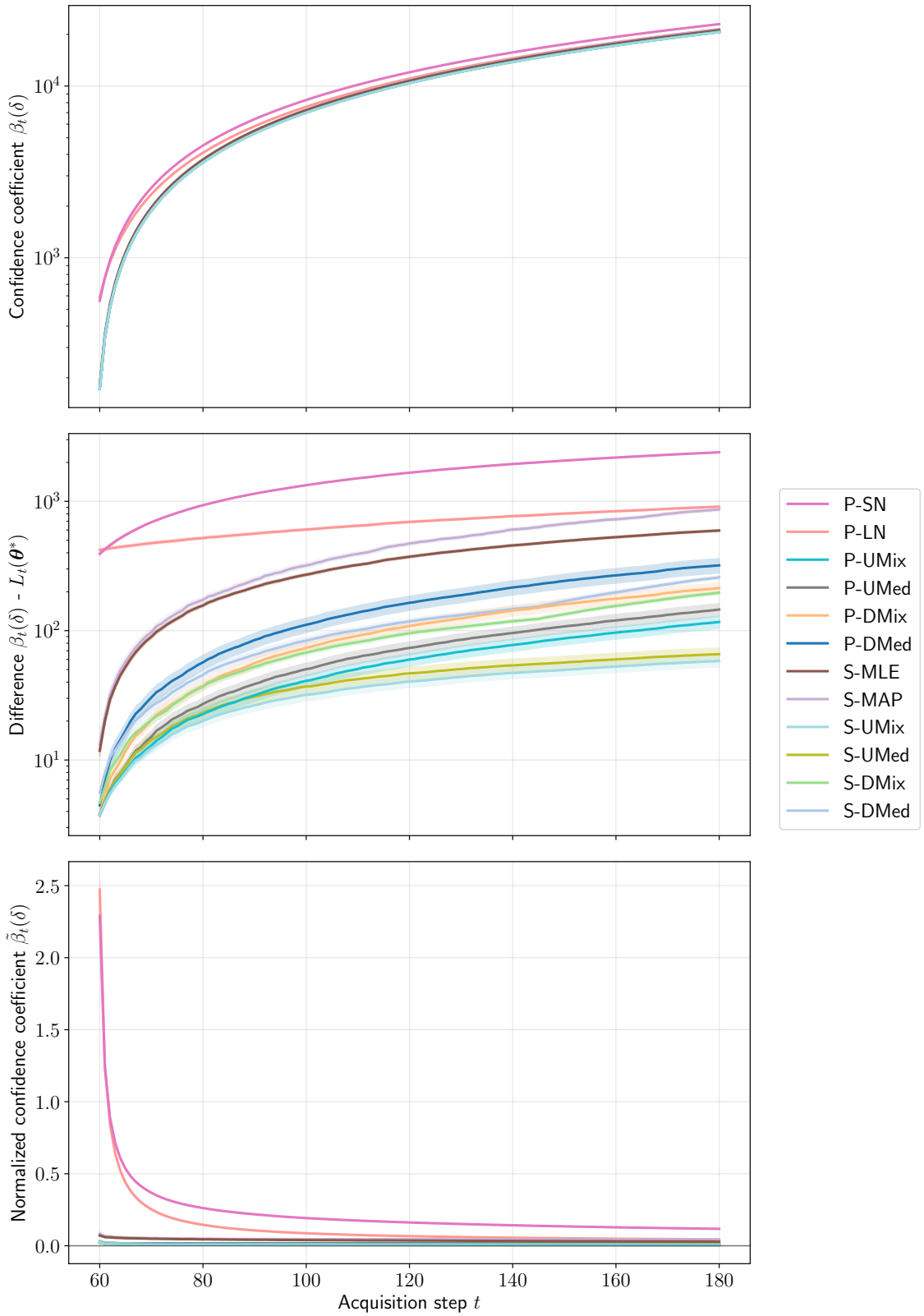


Figure C.2: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 1$ and starting step $t_0 = 60$.

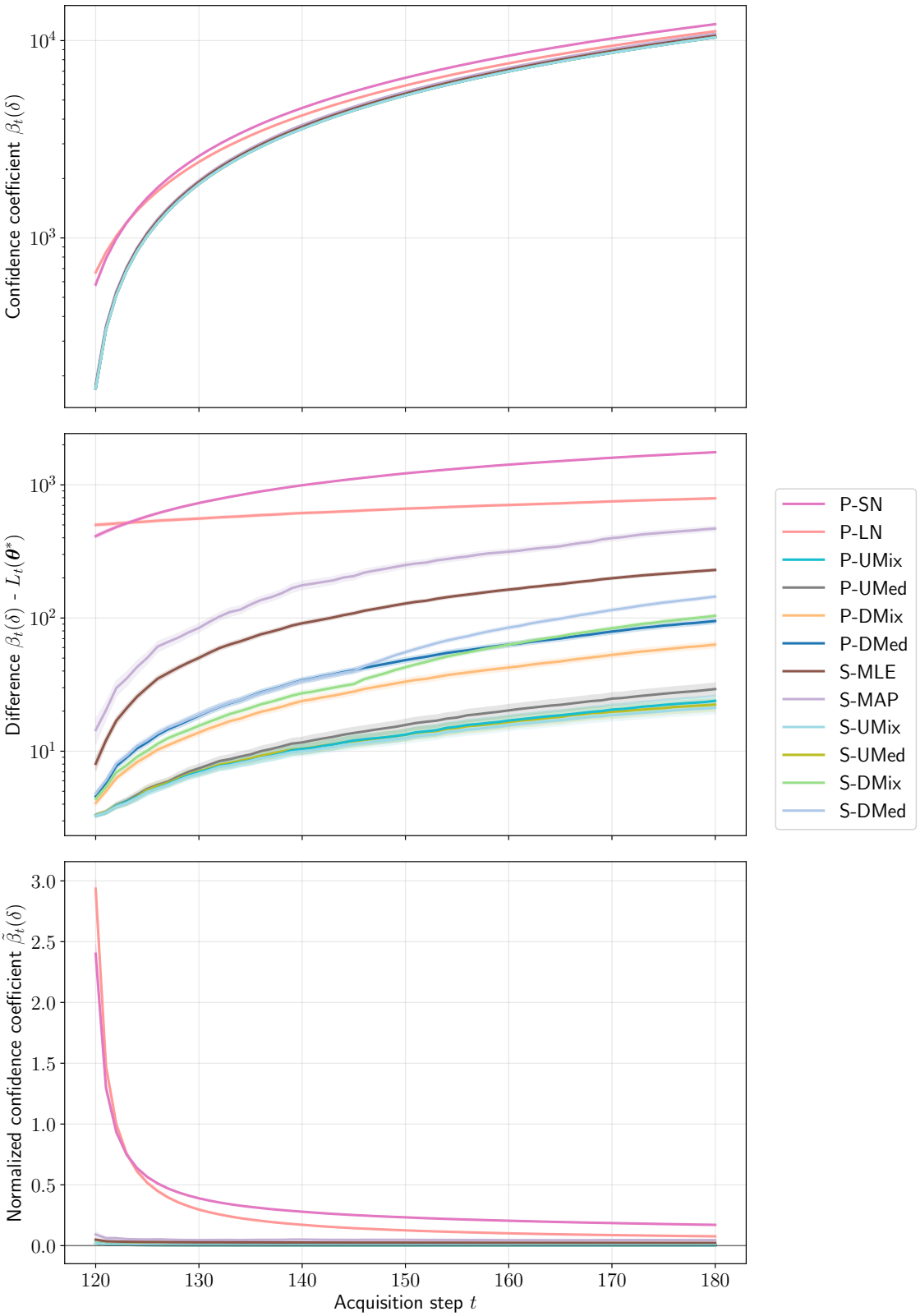


Figure C.3: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 1$ and starting step $t_0 = 120$.

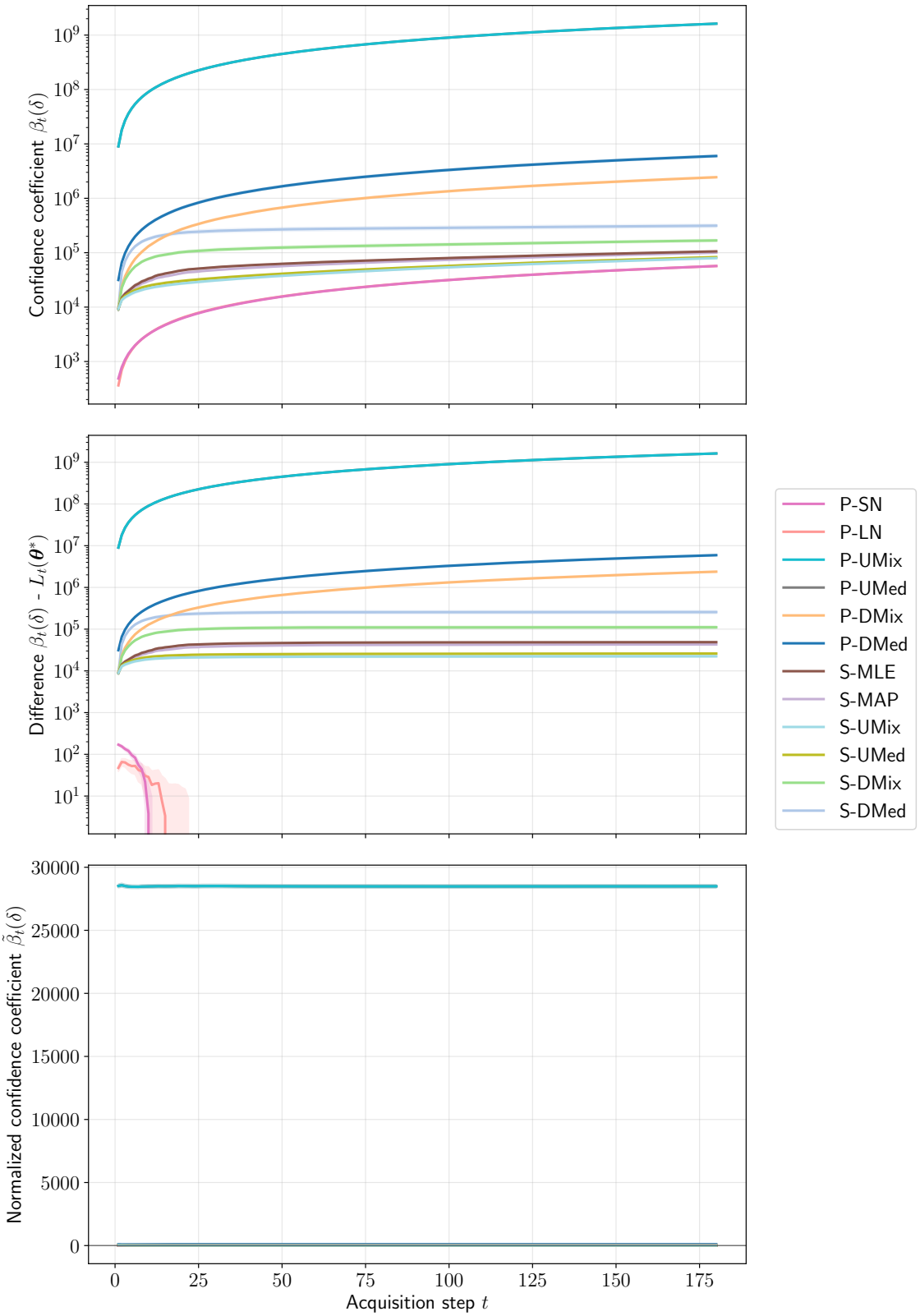


Figure C.4: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 100$ and starting step $t_0 = 1$.

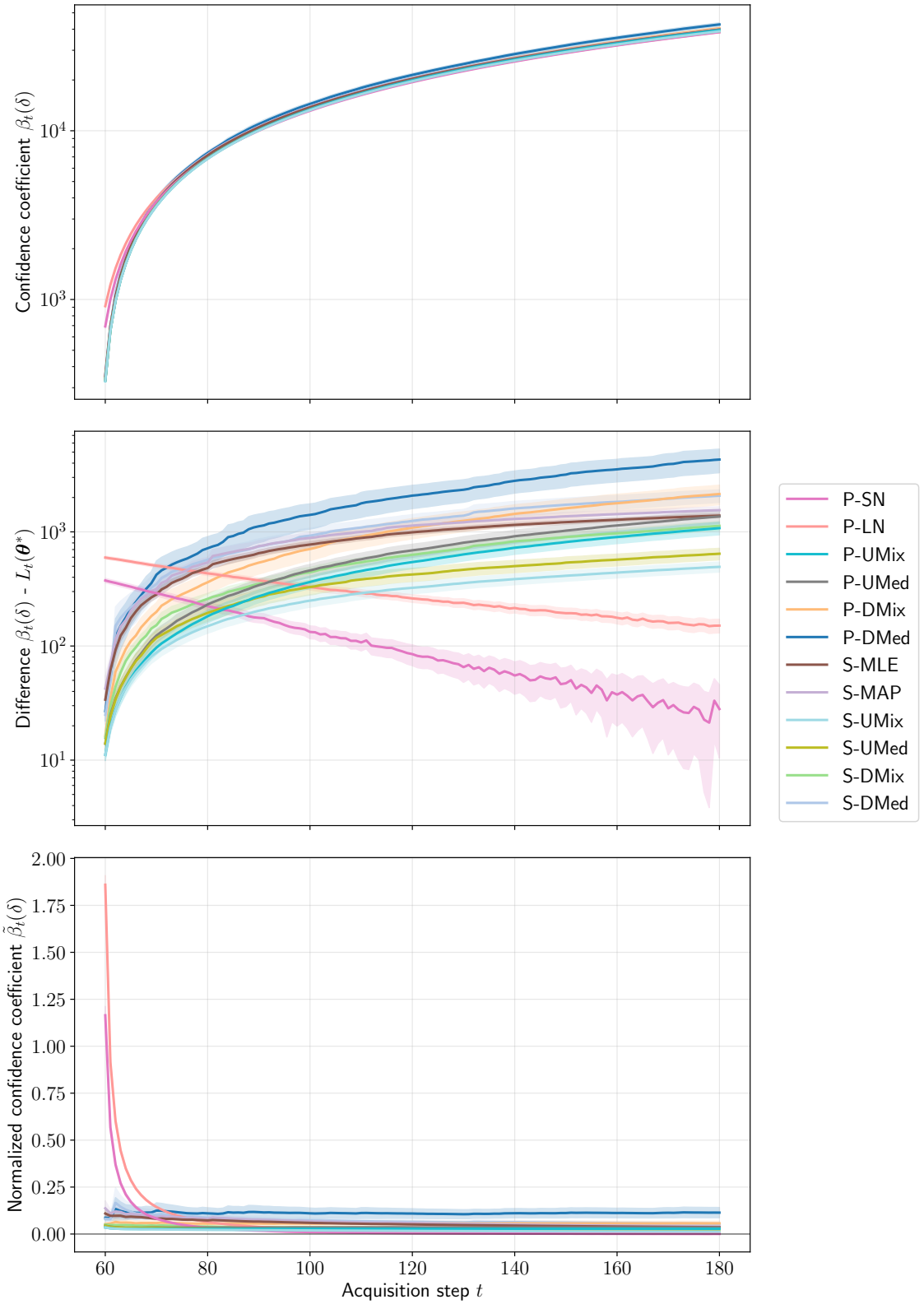


Figure C.5: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 100$ and starting step $t_0 = 60$.

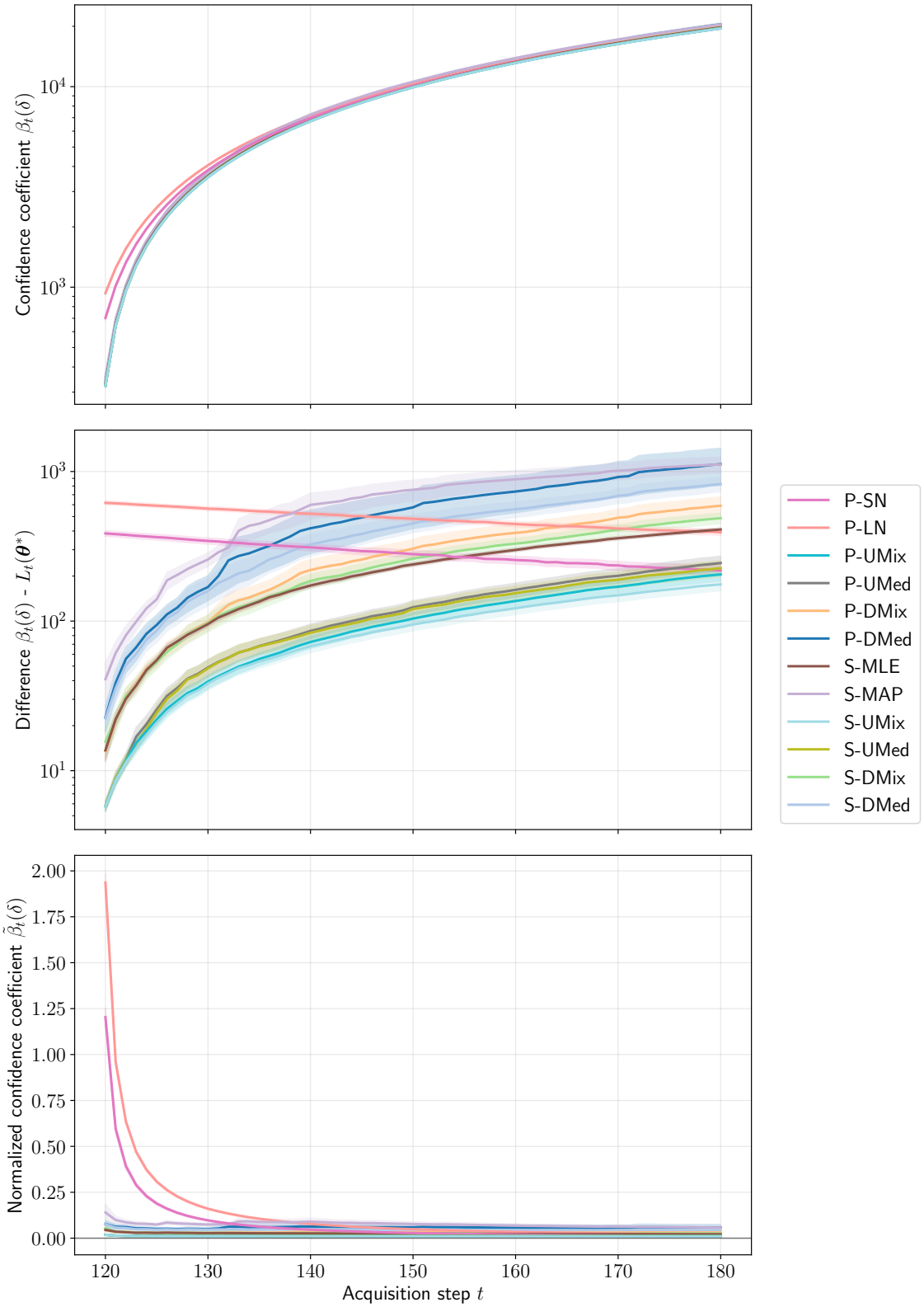


Figure C.6: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 100$ and starting step $t_0 = 120$.

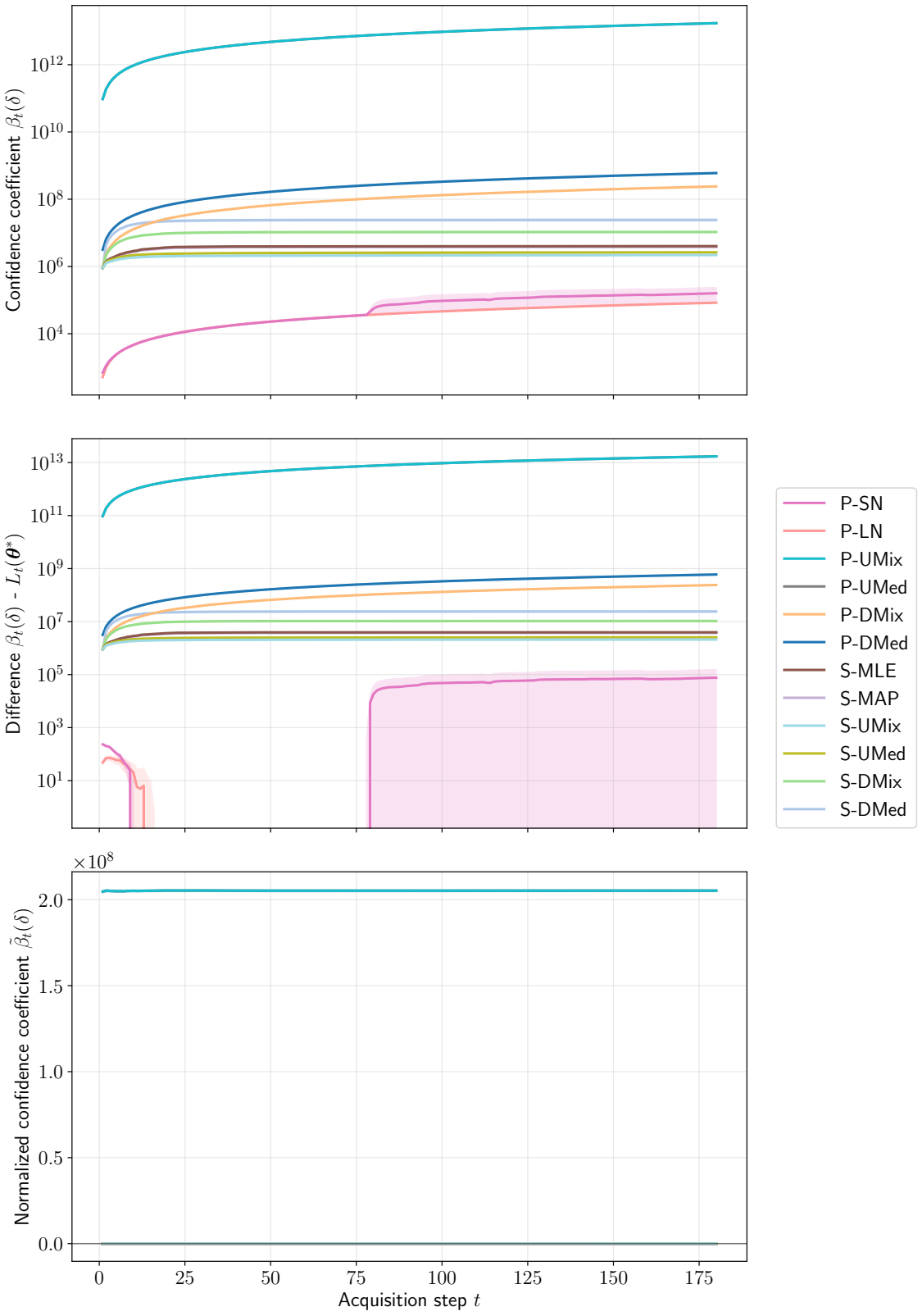


Figure C.7: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 10000$ and starting step $t_0 = 1$.

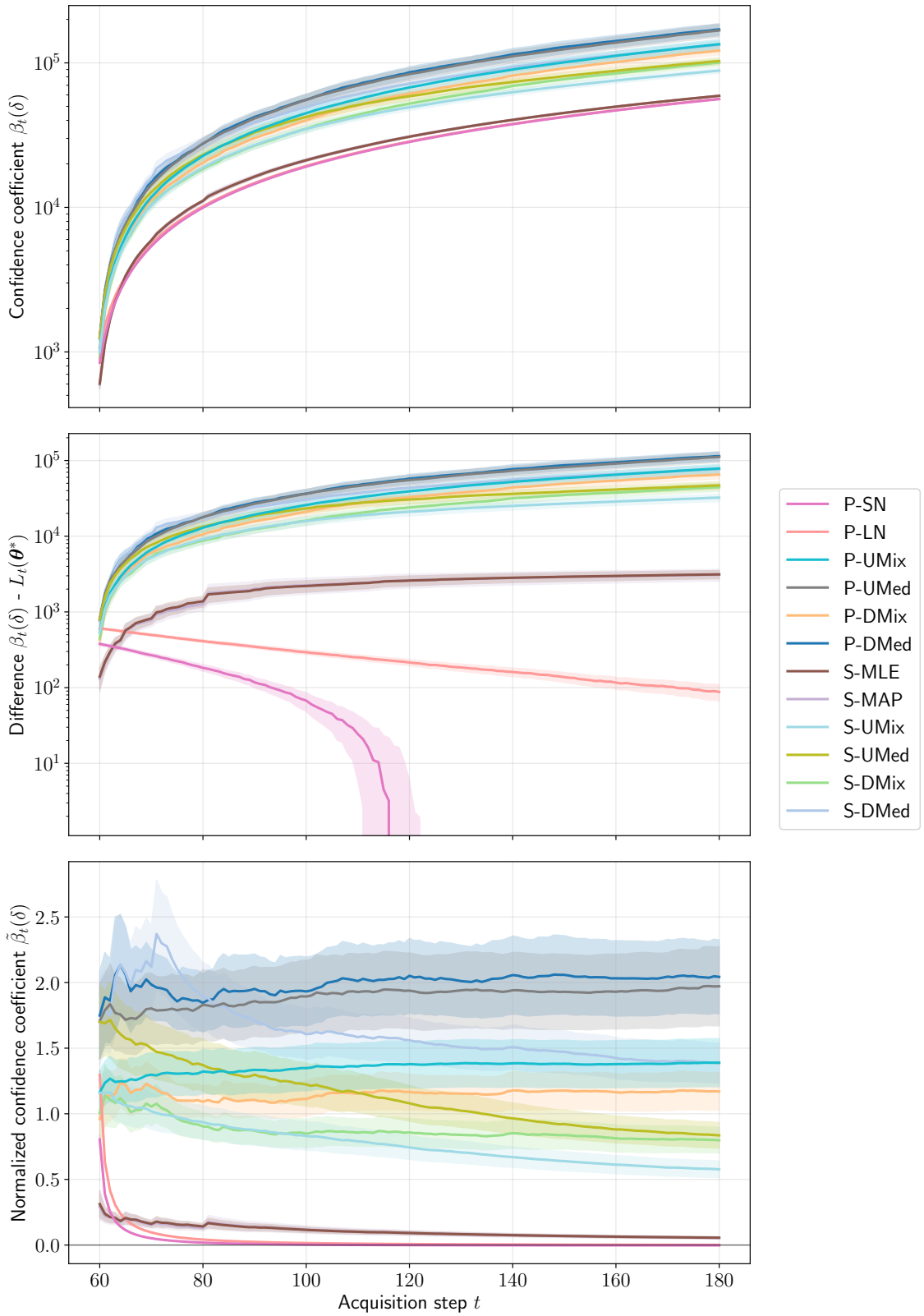


Figure C.8: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 10000$ and starting step $t_0 = 60$.

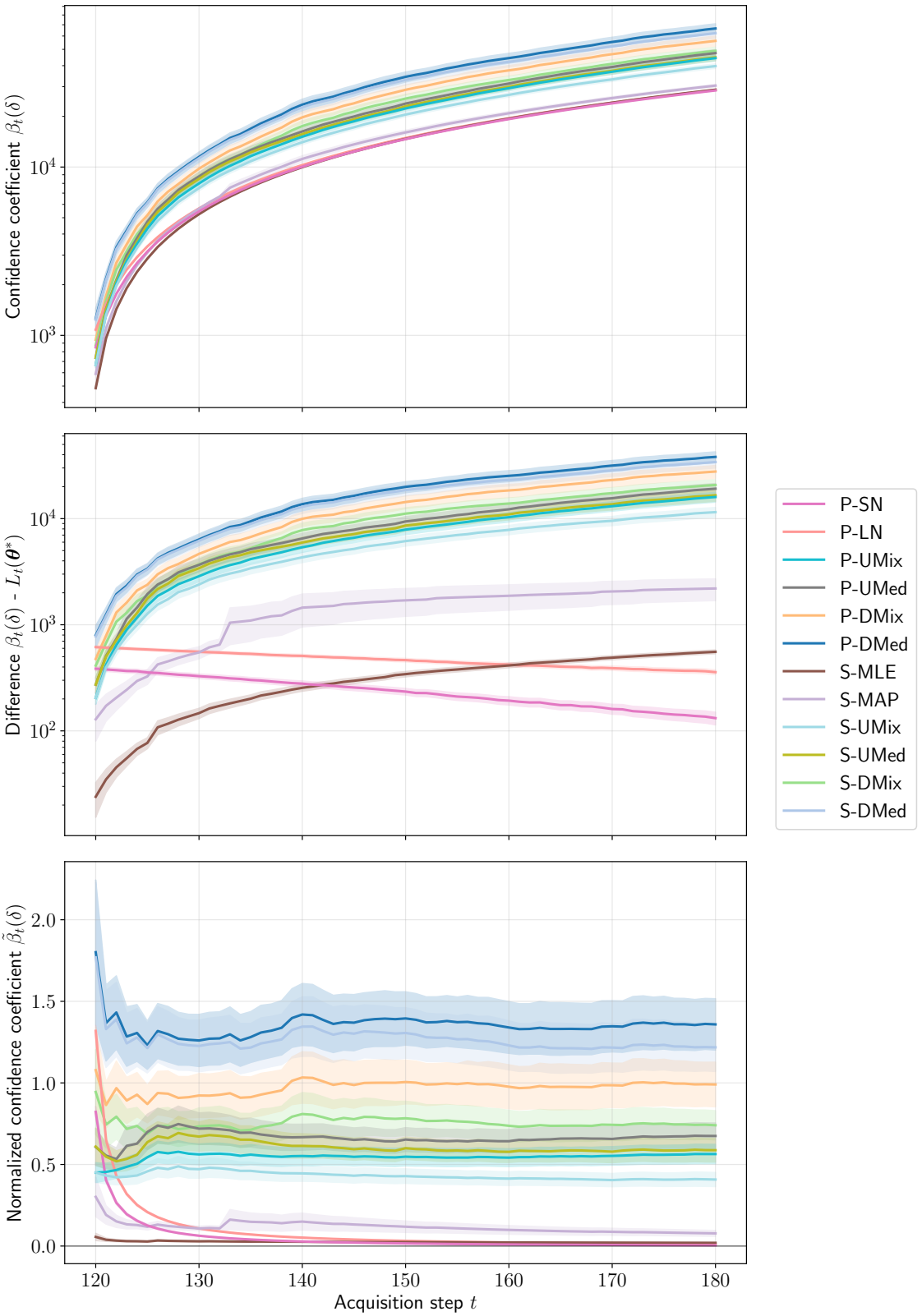


Figure C.9: Confidence coefficients, differences between confidence coefficients and negative log-likelihood of the true image, and normalized confidence coefficients over acquisition steps. Here, acquisition time $T_a = 10000$ and starting step $t_0 = 120$.

C.2 U-Net Performance

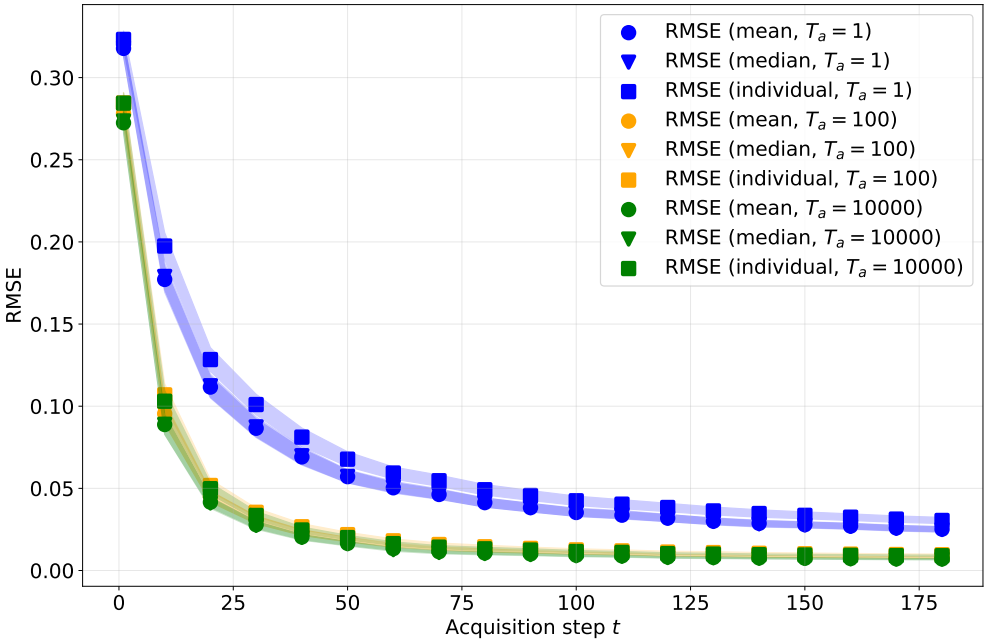


Figure C.10: RMSE of mean, median as well as individual U-Net ensemble member predictions against acquisition steps. Steps $t \in \mathbb{N}$ corresponds to data sequence $((x_s, \mathbf{y}_s), s \in [t])$ being used.

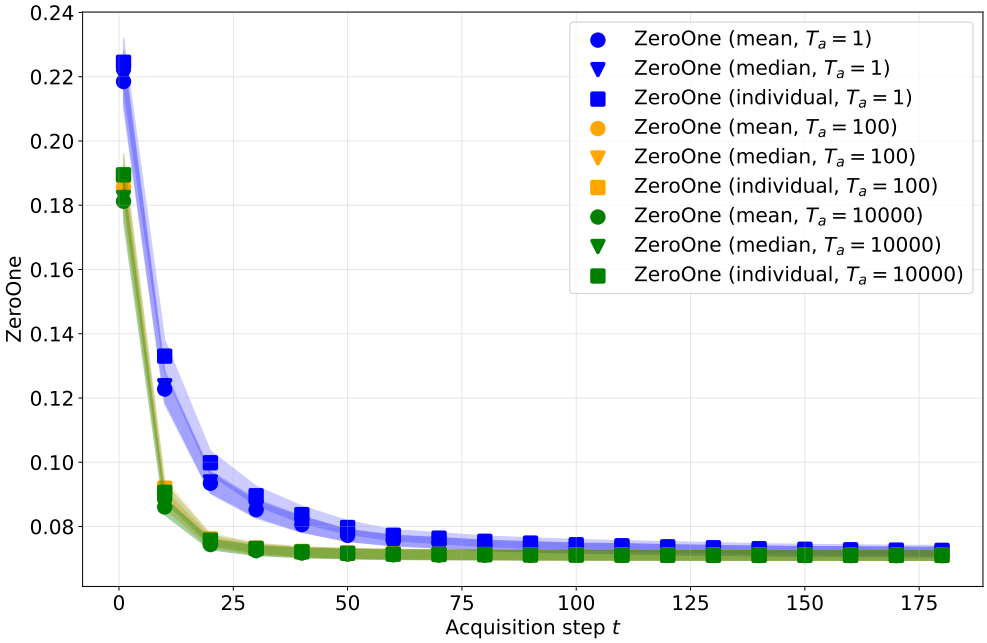


Figure C.11: ZeroOne loss of mean, median as well as individual U-Net ensemble member predictions against acquisition steps. Steps $t \in \mathbb{N}$ corresponds to data sequence $((x_s, \mathbf{y}_s), s \in [t])$ being used.

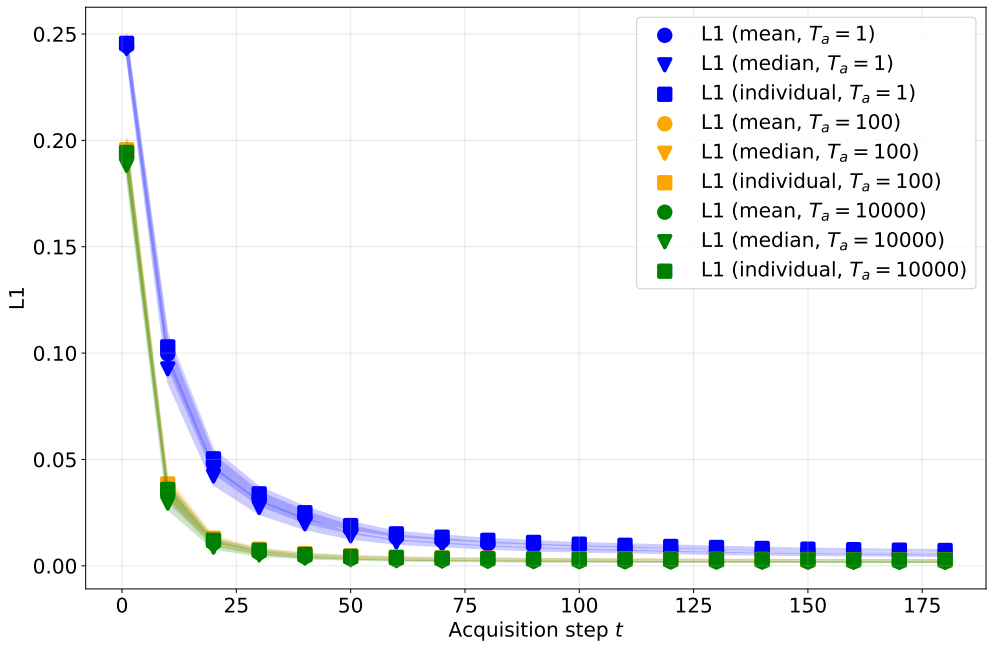


Figure C.12: ℓ^1 loss of mean, median as well as individual U-Net ensemble member predictions against acquisition steps. Steps $t \in \mathbb{N}$ corresponds to data sequence $((x_s, \mathbf{y}_s), s \in [t])$ being used.

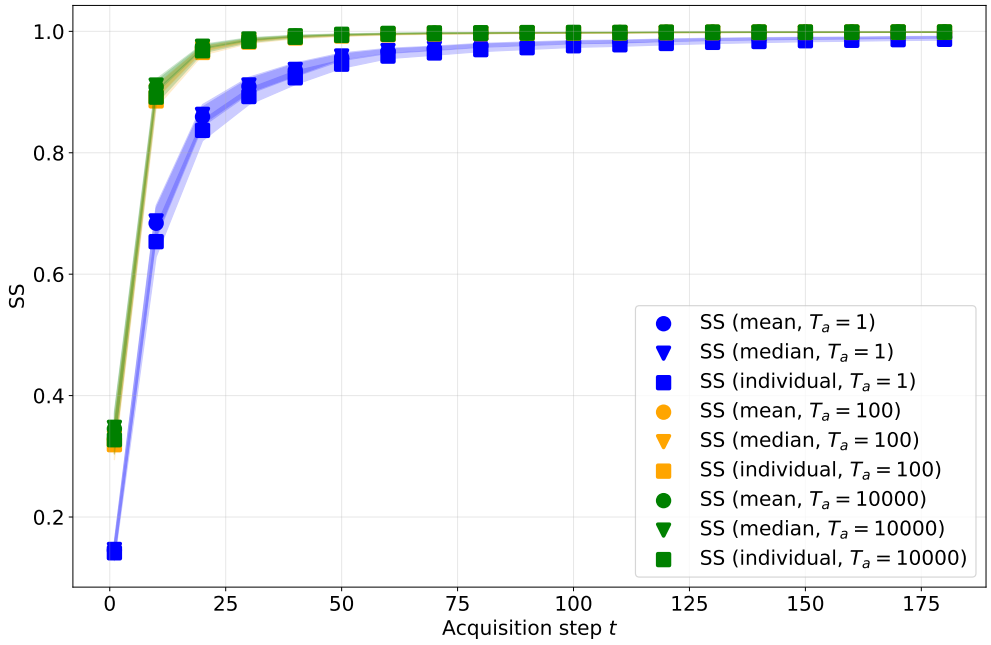


Figure C.13: Structural similarity of mean, median as well as individual U-Net ensemble member predictions against acquisition steps. Steps $t \in \mathbb{N}$ corresponds to data sequence $((x_s, \mathbf{y}_s), s \in [t])$ being used.

C.3 MLE and MAP Estimation

Algorithm 4 Batched MLE/MAP Estimation with Early Stopping

Require: Sinograms $\mathcal{S}_{1:T}$, angles $\mathbf{x}_{1:T}$, initial FBP estimates $\boldsymbol{\theta}_{1:T}^{\text{FBP}}$
Require: Prior parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$ (for MAP only), acquisition time T_a
Require: Optimization parameters: maxSteps, patience, initial learning rate η_0
Initialize parameters: $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^{\text{FBP}}$ for $t = 1, \ldots, T$
Initialize active set: $\mathcal{A} \leftarrow \{1, 2, \ldots, T\}$
Initialize best images: $\boldsymbol{\theta}_t^* \leftarrow \boldsymbol{\theta}_t$ for all t
Initialize minimum losses: $L_t^* \leftarrow +\infty$ for all t
Initialize patience counters: $p_t \leftarrow$ patience for all t
Initialize Adam optimizer (Kingma and Ba, 2017) with per-slice parameter groups
for $k = 1$ to maxSteps **do**
    if then$\mathcal{A} = \emptyset$
        break
    end if
    Clamp active images: $\boldsymbol{\theta}_t \leftarrow \text{clamp}(\boldsymbol{\theta}_t, 0, T_a)$ for $t \in \mathcal{A}$
    $L_{\text{total}} \leftarrow 0$
    for each $t \in \mathcal{A}$ **do**
        Compute forward projection: $\hat{\mathcal{S}}_t \leftarrow \mathcal{R}(\boldsymbol{\theta}_t, \mathbf{x}_{1:t})$
        Compute negative log-likelihood $L_t \leftarrow \sum_{i=1}^{t} \left[\hat{\mathcal{S}}_{t,i} - \mathcal{S}_{t,i} \log(\hat{\mathcal{S}}_{t,i} + \epsilon) + \log \Gamma(\mathcal{S}_{t,i} + 1)\right]$
        if MAP mode **then**
            Add prior term: $L_t \leftarrow L_t - \sum_j \log \mathcal{N}(\theta_{t,j} | \mu_j, \sigma_j^2)$
        end if
        $L_{\text{total}} \leftarrow L_{\text{total}} + L_t$
        if $L_t < L_t^*$ **then**
            $L_t^* \leftarrow L_t$, $\boldsymbol{\theta}_t^* \leftarrow \boldsymbol{\theta}_t$, $p_t \leftarrow$ patience
        else
            $p_t \leftarrow p_t - 1$
        end if
    end for
    Compute gradients: ∇L_{total}
    Clean NaN/Inf gradients and clip gradients
    Update parameters using optimizer
    Handle NaN/Inf in parameters by reverting to best images
    for each $t \in \mathcal{A}$ **do**
        if divergence detected: $p_t <$ patience$/2$ and $L_t > 1.05 \cdot L_t^*$ **then**
            Reset: $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^*$, $p_t \leftarrow$ patience
            Halve learning rate: $\eta_t \leftarrow \max(\eta_t/2, \eta_0/10)$
        else if $L_t > 1.01 \cdot L_t^*$ **then**
            $p_t \leftarrow$ patience
        end if
        if $p_t \leq 0$ **then**
            Freeze slice: remove t from $\mathcal{A}$
        end if
    end for
end for
return Best reconstructions $\{\frac{1}{T_a}\boldsymbol{\theta}_t^*\}_{t=1}^{T}$

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

Anytime-valid Neural Uncertainty Quantification for SPECT Imaging

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Gätzner	Matteo

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zürich, 29.09.2025	Matteo Gätzner

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).