Generative modeling applied to medical imaging tomographic reconstruction

Alessandro Cecchini

October 2023

Contents

1 Preliminaries / How to read this report

This report describes our current studies on the use of a specific class of generative models – *Diffusion(-type) models* for solving the inverse problem of emission tomographies (further denoted by ETs): *Positron Emission Tomography* and *Single-photon Emission Computed Tomography* (resp. PET and SPECT). Before going into details of ETs and models we want to provide some informal context and guidelines on how to read this report.

First of all, it would be impossible to make a report that would be informative, interesting and concise for *any reader* (specifically because the latter properties are actually reader-dependent). Instead of guessing the tastes we choose another path – we start with a number of informal 'claims' in which we believe and want to share prior to getting into 'the juice' of the work[1]:

1. in applied sciences there are only two types of problems – *forward problems* (FP) and *inverse problems* (IP), e.g., having operator A, finding $y = Ax$ for x given is a forward problem and finding x from y would be an inverse problem; more informally – say, if A describes an evolution operator for observed physical phenomena, x being the initial conditions and y being the measured response, then finding y from x (and A) is a FP and reconstructing x from data y is an IP; to sum up, it is assumed that in a given problem there is always a natural relationship of *causality* between input and output parameters which defines whether problem is FP or IP

2. in this work we target IPs (specifically for medical imaging, hence all examples are motivated by reconstructing images from other images)

3. usually, IPs are harder than FPs (solution of $Ax = y$ may be unstable if A is ill-conditioned or even non-unique whereas response y is continuous whenever A is a reasonably bounded operator); This can be seen as a result of using human-crafted devices with limited precision and finite response range for measurements of phenomena that are better described as vectors in 'bulky' infinite-dimensional spaces. Hence, our measurement processes act often as *compact operators* which according to Baire's theorem are *almost never continuously-invertible*

4. if measurements are stochastic – adding a statistical layer (e.g., noise over an analytical model) during modeling stage does not reduce the ill-posedness of the inverse problem (and usually does not increase it, apart from when noise is dependent on the underlying analytic response)

5. the only possible way to stabilize solutions of IPs is to restrict the space of solutions (e.g., when we look for $x \in X$ s.t. $Ax = y$ and using least-squares estimates to project solution on X); introduction of restrictions is often due to *additional information* which can be reasonably objective (positivity of signal (e.g., concentration, density, mass, etc.), geometric restrictions due to experimental setup, etc.) and very subjective (guessing the degree of 'smoothness' of a given signal); the working name of this step is *regularization* of an IP, additional information is called *prior* information

6. properly chosen regularization may significantly improve reconstruction quality (even to go beyond the direct resolution limit of an actual measurement device; super-resolution is a good example here) – as well poor regularization may introduce significant bias and make the solution completely inadmissible; in view of the above we say – *construction and choice of informative priors is of very great importance*

7. generally, for a given IP *a point estimate* (e.g. Maximum Likelihood (MLE) or Maximum Aposteriori (MAP) estimates) is *not sufficient* for a solution – due to ill-posedness there is always the risk of large deviations of the solution from the 'truth', so error-bounds is a necessary part of any report on solving an IP;[2] there is a huge number of generic ways to produce error-bounds for an estimator, for example: *frequentist*, *bayesian*, *bootstrap* (actually any reasonable methodology will do, while it is *conditional* on data and model)

[1] we assume that reader is acquainted with used notions and ideas (but do not ask to accept them deeply)

[2] to say simply, we are interested not in a single solution but in all solutions that are adequate given the data, model and our prior information

8. we prefer *bayesian approach* over others for solving IPs, that is we are interested in constructing 'efficient' (conditional) samplers from the *posterior distribution* given ET data, model and a prior[3]

9. sampling cats or faces is not the same as sampling medical images(!) – there are obvious risk differences between reporting "a cat with two tails" and "a patient with a malformed organ/unexpected tumor/etc."; moreover, "hallucinations" in natural images are easier to detect with a naked eye, while for medical scans it necessities large medical experience and self-awareness (it takes great effort to convince yourself that a smoother nice-looking image in fact can be much more erroneous)

10. ETs is not Magnetic Resonance Imaging (MRI) or Computed Tomography(CT) – it is harder(!) – ML-community proposes a great number of "super-resolution" generative methods to solve inverse problems of CT or MRI because:

    (a) these have huge, easily-accessible training datasets

    (b) noise levels in data and ill-posedness still allow very good quality reconstructions with plain classical optimization methods (even *without any informative prior*)

    Claims (1), (2) are essentially wrong for ETs (we discuss this in detail later): datasets are small, very heterogeneous and noise levels are very high - this complicates the quick embedding of recent advances in machine learning into the field

11. *Practical details matter* – ETs is a field with reasonably well established physical and mathematical models and neglecting those would be a crude step (as it may happen in some data-analysis scenarios where the origin of features is less important)

12. "Try to do at least better than it is right now." (from private discussion with prof. D. Rubin) – despite the complexity of IPs in ETs and skepticism on the use of ML/AI in medical imaging we believe that it is still valuable to test/try new approaches (sometimes poorly understood – Diffusion-type models for tomography is a perfect example) – at worst we will learn on the degree of applicability of those, at best - we may improve upon current research; moreover, *honest practical work* and *careful communication* bring intuition and right questions to those who may work on the theoretical side of yet "dirty" and "unestablished" practical methods

Having these messages in mind a reader has a baseline which we followed in our study, so in case of ambiguity or loss of storyline a helping rule may be to recall some of these messages.

From the claims above one can guess the general topic of the work – we build a bayesian sampler for the inverse problem of ETs using very recent Diffusion-type models that reached state-of-the-art in many tasks of unconditional and conditional image generation. However, this does not mean at all that these models are well-applicable in a such complex problem as PET/SPECT. Taking this conservative view we say that the vague idea of the internship – analysis of the applicability of Diffusion models to emission tomographies.

The contents of this report are organized as follows. In Section 2 we present practical and mathematical models of ETs. In Section 3 we present two problems that we tackle in this internship:

1. *construction of informative priors for guided posterior sampling using Diffusion models*

2. *dose enhancement*

In Section 4 we introduce various existing diffusion-type models, first for unconditional generation (prior learning) and then conditional generation (posterior sampling). Between the two problems mentioned above, we choose the problem of dose enhancement as viable for applying Diffusion models and discard the problem of construction of informative priors. This is discussed in detail in Section **??**. In Section 5 we present details of the implementation of the unconditional diffusion-model sampler and present current results in Section 6. We sum up our current studies in Section 7 together with perspectives for future work in the time left for the internship (2 weeks by the time of submission).

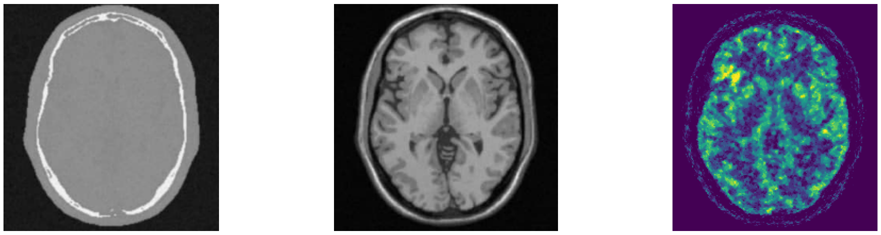
[3] we believe that introduction of aleatoric and epistemic uncertainties to final error-bounds is of great importance, especially when complex regression models are used; a reader skeptical towards semi-philosophical arguments may think that this is merely a methodology to perform sensitivity analysis of a given model

2 Emission tomographies – PET/SPECT

Positron emission tomography (PET) and single-photon emission computed tomography (SPECT) belong to the branch of nuclear medical imaging [NIH, 2016] which also includes the well-known X-ray Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). We believe that the quickest and most intuitive way to introduce ETs is to contrast them with CT or MRI (Claim 11). Then, in view of this mathematical modelization of our problems becomes natural.

2.1 Medical nuclear imaging

The purpose of CT and MRI is to provide *anatomical information* to the end-user – anatomical contrast is achieved via the response of subjects' tissues to exterior ionization process: in CT these are attenuation properties of tissues for emitted X-ray photons, for MRI these are proton densities which modulate response to strong magnetic fields produced by the scanner (see Figure 1).



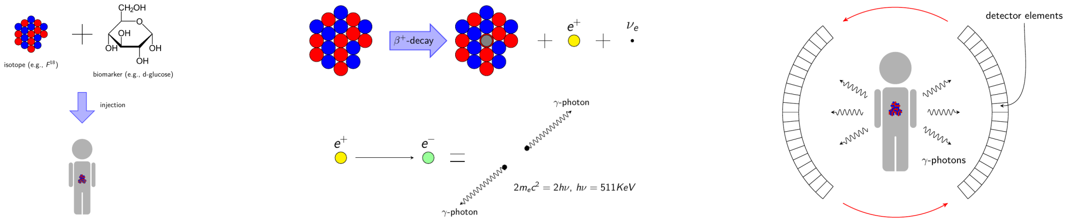
(a) CT

(b) MRI

(c) PET

Figure 1: Examples of reconstructed images in nuclear imaging of a 2d synthetic brain slice

Oppositely to the above examples, ETs provide *functional information* – activity of metabolism (e.g., glucose consumption, blood flow level, oxygen exchange rate) in areas of interest [Marcu et al., 2018]. For example, a bright spot in Figure 1(c) corresponds to increased glucose consumption which is typical for certain types of tumorous brain lesions (however, a lesion may be absolutely 'invisible' on CT or MRI scans(!) if the related densities are of little difference with surrounding healthy tissues). So how do ETs work? The generic data acquisition pipeline for PET is given in Figure 2.



(a) tracer injection

(b) nuclear decay

(c) data-acquisition

Figure 2: PET acquisition pipeline

Data acquisition in PET:

a) chemical compound of a *biological marker* and an *isotope* with a short half-life time is injected into the patient's blood flow

b) at a single nuclear decay of the isotope *a positron*[4] is emitted which by a very short diffusion falls off onto an electron in surrounding matter and both annihilate; at annihilation two γ-photons[5] are produced, which according to *law of conservation of momentum propagate along a straight line in opposite directions*

[4] a particle with a mass of electron and a positive charge

[5] α, β, γ-photons are characterized only by their energy and these are merely range conventions; γ - corresponds to relatively high energies

c) pair of γ-photons hits (almost) simultaneously two detectors and a *double event* (time and line of propagation) is recorded; recording data is performed during period T – collection of all double-events (times of hits and lines) constitutes the raw data in PET – called a *sinogram* (list-mode – if every event is recorded, histogram-mode – if all events are summarized in total number of double events along different lines)

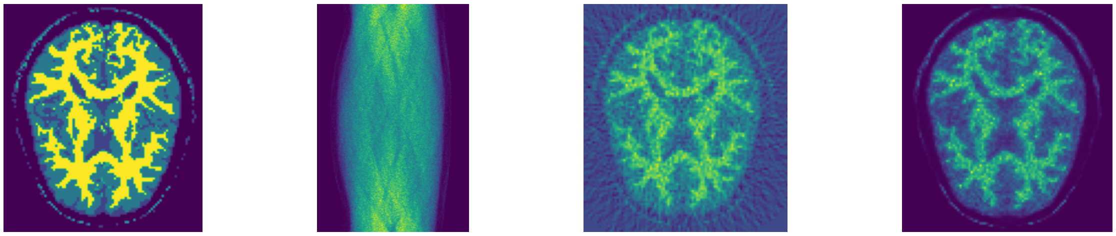
Data acquisition in SPECT: it is essentially the same as in PET, with the only difference that at decay *only one γ-photon is produced*

From the above description, it is easy to guess that the IP for ETs is to *reconstruct the spatial distribution of the tracer* from the sinogram data. The reconstruction process is based on the assumption that photons propagate along straight lines – this helps to reduce the problem to inversion of Radon-type linear integral operator (see e.g. [Toft, 1996], [Natterer, 2001]):

$$Y^t(\ell) \sim \mathrm{Po}(t \cdot R\lambda(\ell)),\ t > 0, \tag{1}$$

where $Y^T(\ell)$ is the measured number of double events along line ℓ during t, $\mathrm{Po}(\Lambda)$ is the Poisson distribution with intensity $\Lambda > 0$, R is Radon transform for PET (will be defined precisely in Subsection 2.2), $\lambda = \lambda(x)$ is the unknown distribution of the tracer for $x \in \mathbb{R}^2$ or $x \in \mathbb{R}^3$ (depending if the problem is reduced to a 2D-slice or being kept in 3D).

Reconstruction process: Many approaches exist to estimate λ from $Y^T(\ell)$, $\ell \in L$, where L is set of lines detectable within scanner geometry – the common point is the *instability* due to ill-posedness of R if no special regularization is applied (recall Claims 3, 4, 5). We prefer to list only two classical algorithms which also serve as the basis for many newer, modern and fancy ones (for recent advances see, e.g. [Reader et al., 2020]): *Filtered-Backprojection* (FBP)[Natterer, 2001] which is a single-run algorithm based on the analytical formula for R^{-1} but neglecting the Poisson model in (1) and *Maximum Likelihood EM-algorithm* (MLEM) [Shepp and Vardi, 1982] which is iterative and respects the Poisson model and some geometrical properties of R [Siddon, 1985]. To say, both FBP and MLEM are still used in real scanners and serve as a main workhorse in the industry because of their simplicity and scalability, though they still produce strong noise artifacts (see Figure 3).



(a) pet-phantom (b) sinogram (c) FBP (d) MLEM

Figure 3: Examples of reconstructed images via FBP and MLEM

Regularization within FBP is possible (e.g., via the use of low-pass filtering; see e.g., [Natterer, 2001]), however, these produce only linear shift-invariant filters which can easily be suboptimal for a specific class of images (e.g., head scans). The fact that FBP neglects Poisson distribution of Y^T results in strong streak-type artifacts that are well-studied mathematically [Krishnan and Quinto, 2015] but the problem of their correction remains open within the FBP itself. At the same time, MLEM does not produce such artifacts and allows direct construction of regularized schemes (such as MAP for various types of prior information; Claims 5, 6) since it is based on the Poisson likelihood expression of (1) – these are the main reasons why currently MLEM and its various machine learning extensions dominate in industry and in research.

In particular, in this internship in both problems (recall 1, 2) we considered MLEM(-type) reconstruction for posterior sampling (will be discussed in detail in subsection 3.1).

2.2 Mathematical model of the inverse problem for ETs

For simplicity of exposition, we restrict the model to dimension $d = 2$ which represents an axial slice of the imaging target (that is sinogram and unknown isotope densities are represented by compactly supported functions on $\mathbb{R}^2$). A classical approach to obtain full 3D reconstruction is to proceed in a slice-by-slice fashion which is also implemented in many PET/SPECT scanners.

There is a great amount of literature on mathematical models of ETs (see e.g., [Natterer, 2001], [Goncharov, 2019] and references therein). Below we keep the exposition minimalistic (to formulate later the statistical model) but not simplistic in order to preserve some intuition behind for the reader (recall Claim 11). Also, if we do not state explicitly a proposition/lemma/theorem, then it means that this result is standard in the domain.

2.2.1 Idealized inverse problem

Let $\lambda(x)$, $x \in \mathbb{R}^2$ be a non-negative compactly supported infinitely-smooth function[6], lines on $\mathbb{R}^2$ be parameterized as follows:

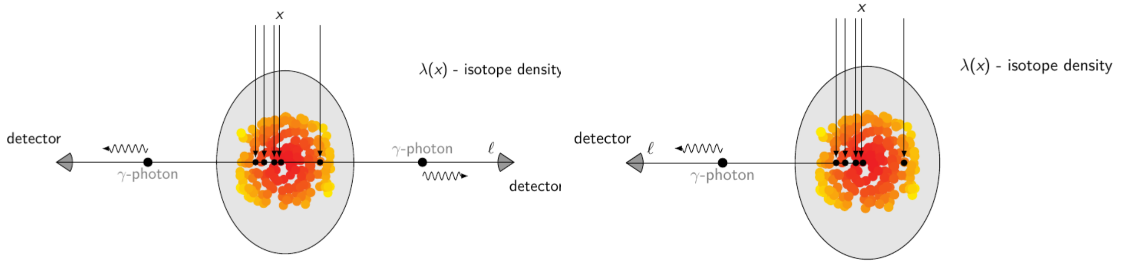
$$\ell(s,\theta) = \{x \in \mathbb{R}^2 : x = s\theta + t\theta^{\perp},\, t \in \mathbb{R}\},\; s \in \mathbb{R},\; \theta = (\cos\phi, \sin\phi),\; \theta^{\perp} = (-\sin\phi, \cos\phi). \tag{2}$$

Photons emitted from the isotope experience attenuation when crossing the patient's surrounding tissues – this is described by *attenuation coefficient* $a(x)$, $x \in \mathbb{R}^2$. We assume that $a(x)$ is also non-negative, compactly supported and smooth.

We say that the measured intensity (photon rate) along line ℓ with attenuation coefficient a is given by the line integral:

$$R_a\lambda(s,\theta) \overset{\text{def}}{=} \int\limits_{\ell(s,\theta)} \lambda(x) p_a(x; s,\theta)\, dx, \tag{3}$$

where $p_a(x; s, \theta)$ is called the *weight* and denotes the probability that a pair of photons (for PET; only one photon for SPECT) will not be attenuated during their propagation along ℓ starting at x at will reach both detectors (see Figure 4). Operator R_a maps isotope densities into functions on lines in $\mathbb{R}^2$ which correspond to intensities of the Poisson process on each line (see formula (1)), so $R_a\lambda$ is a function on lines.



(a) PET

(b) SPECT

Figure 4: photons trajectories in ETs

The weight $p_a(x; s, \theta)$ is based on the Beer-Lambert attenuation model and is given by the formulas:

$$p_a(x; s,\theta) = e^{-Ra(s,\theta)} \text{ for PET,} \tag{4}$$

$$p_a(x; s,\theta) = e^{-\int_0^{+\infty} a(x+t\theta)\, dt} \text{ for SPECT,} \tag{5}$$

where $a(x)$ is the attenuation coefficient. Note that $p_a(x; s, \theta)$ for PET *does not depend on* x, therefore it may be moved out from the integration in (3), but this is not the case for SPECT.

In ETs attenuation coefficient $a(x)$ is assumed to be known (both for forward and inverse problems) and in fact is computed prior to the actual PET/SPECT procedure[7]. Then idealized forward and inverse problems for ETs are formulated as follows:

[6] smoothness is not essential and results are valid for generalized distributions

[7] attenuation $a(x)$ is computed from a separate CT-scan prior the PET/SPECT procedure

Problem 1 (idealized FP) *Given $p_a(x;s,\theta)$ and $\lambda(x)$ find $R_a\lambda$ for all lines on $\mathbb{R}^2$.*

Problem 2 (idealized IP) *Given $p_a(x;s,\theta)$ and $R_a\lambda$ for all lines on $\mathbb{R}^2$ find $\lambda(x)$.*

The forward problem is trivial, while the inverse problem gave rise (and still gives) to a huge amount of research on studies of the operator R_a and its deep extensions. In this work, we are interested in IPs, which is why we formulate a 'meta-theorem' to summarize everything relevant that is known about IP.

Theorem 1 (tomographical folklore) *Let $\lambda(x)$, $a(x)$ be compactly supported and sufficiently smooth functions on $\mathbb{R}^2$ (not crucial). Then*

1. *in literature R_a is known as weighted Radon transform and arises in many different tomographies (X-ray CT, MRI, PET/SPECT, ultrasound, etc.) [Goncharov, 2019]*
2. *R_a is a linear bounded operator between reasonable Hilbert spaces*
3. *$R_a\lambda$ uniquely defines λ, R_a is a compact operator and hence is not continuously invertible*
4. *Since R_a is compact, it can be approximated by finite-dimensional operators (matrices) in strong operator norm; mover R_a admits singular value decomposition (SVD) and singular values decay at rate $\sigma_k \asymp k^{-1/2}$*
5. *decay rate $\sigma_k \asymp k^{-1/2}$ corresponds to the class of mildly ill-posed inverse problems, so if the noise in data is small enough, reconstructions may not even suffer too much from the noise (this is the case in X-ray CT where reconstructions are already very good without any regularization though the problem is ill-posed)*
6. *from the SVD it is obvious that inversion of R_a is numerically unstable as soon as R_a^{-1} is applied to a vector out of the range of R_a (which happens with probability 1 since the measured data is corrupted with Poisson noise)*
7. *both for PET and SPECT (weights from (4), (5)) there exist very efficient analytical formulas (for PET it is FBP for SPECT it is FBP-type) to invert R_a (only on its range, of course) which are implementable in terms of Non-Uniform Fast-Fourier Transforms (NUFFT)*

The goal of this subsection was to introduce the underlying analytical model and overview the ill-posed nature of IPs of ETs. The next step is to put a statistical layer over it (spoiler – Poisson noise) to formulate a proper statistical model. At last, it is a good moment to recall Claim 4 that we should not expect any regularization of the problem by going into the stochastic world – ill-posedness is in R_a and will always be there whenever R_a is used for the link mapping of the model.

2.2.2 Practical approach: discretization

As it was said in Subsection 2.2.1, Theorem 1, R_a can be approximated with a matrix, so the idealized IP of ETs reduces in practice to solving a linear system:

$$[R_a]\lambda = \Lambda \text{ for } \lambda \in \mathbb{R}^p_+,\ \Lambda \in \mathbb{R}^d_+, \tag{6}$$

where $[R_a]$ is the matrix approximation of R_a of size $d \times p$, $\mathbb{R}^p_+$ is the non-negative octant in $\mathbb{R}^p$. Numerical construction of $[R_a]$ can be simply achieved by discretizing the space and usually a uniform discretization grid is used: $\lambda(x)$ is supported in a unit square $[-1,1]^2$ and is locally constant within each pixel of the uniform grid (see Figure 5).

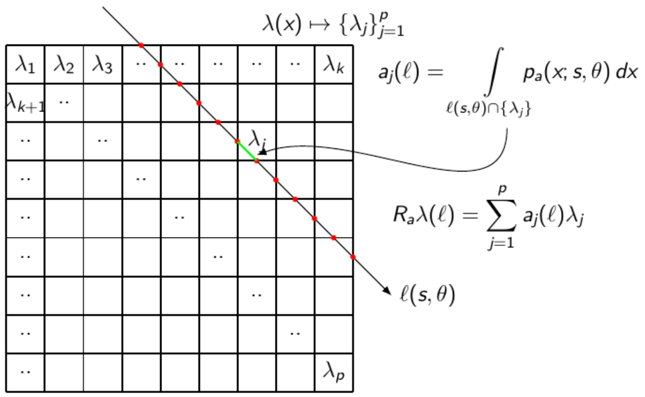

$\lambda(x) \mapsto \{\lambda_j\}_{j=1}^p$
λ_1
λ_2
λ_3
λ_k
λ_{k+1}
λ_j
λ_p
$a_j(\ell) = \int_{\ell(s,\theta)\cap\{\lambda_j\}} p_a(x; s, \theta)\, dx$
$R_a\lambda(\ell) = \sum_{j=1}^p a_j(\ell)\lambda_j$
$\ell(s,\theta)$

Figure 5: discretization of R_a

From now on we assume that the image representing isotope density consists of p pixels and d lines are used to record the data: λ_j for $j \in \{1, \dots, p\}$ corresponds to isotope concentration (measured in [Bq/mm^3]) at pixel with index j, Λ_i corresponds to measured intensity along line $l(s_i, \theta_i)$, where $\{(s_i, \theta_i)\}_{i=1}^d$ form uniform grid in $[-1, 1] \times [0, \pi]$.[8] The precise choice of d for given p is guided by the Shannon-Nyquist criterion [Natterer and Hadeler, 1980], but for the present work, it is sufficient to say that $d = O(p^2)$.

Note that after R_a has been discretized, there are essentially no fundamental differences between PET and SPECT in terms of $[R_a]$ as it was in (4), (5) for the idealized model. We may also suppress attenuation index a from all the formulas after since a is known and fixed for a patient – and in any case we solve the IP for ET for each person individually. From now and later on we will use the following notation for $[R_a]$ and will call it by *projector (or projection matrix)*:

$$[R_a] \mapsto A \in \mathrm{Mat}(d, p). \tag{7}$$

For practical reasons matrix A is also renormalized in a special manner, so that A attains *stochastic interpreptation*:

$$a_{ij} = P(\text{detection of a double event along line } i | \text{decay happened at pixel } j), \tag{8}$$

$$a_{ij} \geq 0, \ \sum_{i=1}^d a_{ij} = 1. \tag{9}$$

In view of (6), (7) the idealized IP can be rewritten as follows:

$$\text{given } A,\ \Lambda \in \mathbb{R}^p_+, \text{ find } \lambda \in \mathbb{R}^p_+ \text{ such that} A\lambda = \Lambda. \tag{10}$$

From Theorem 1 and formulas (8)-(10) one can see that we encounter a linear inverse problem with mildly ill-conditioned A and with restrictions on the positivity of the solution. Now we have all the tools to finally introduce the statistical model for ETs.

2.2.3 Statistical model for emission tomographies

Likelihood: In ETs for any finite set of lines $\ell_1, \dots, \ell_k$ it is true that

$$Y^t(\ell_1), \dots, Y^t(\ell_k) \text{ are mutually independent}, \tag{11}$$

where $Y^t(\ell)$ are the photon counts measured along ℓ and defined in (1). Note that a combination of the latter defines a Poisson process on $L \times [0, +\infty)$ (product of space of lines and time interval). Then,

[8] For the uniform grid which is standard in practice, there is a very efficient algorithm [Siddon, 1985] which can compute forward $[R_a]\lambda$ in $O(p)$ FLOPS, however, memory footprint of storing the matrix quickly explodes ($O(dp) \sim O(10^{10})$ for common image resolutions) so $[R_a]$ cannot be stored on existing GPUs at once and is evaluated on-the-fly for any matrix-vector product.

using the discretization from Subsection 2.2.2, the likelihood for a finite-dimensional statistical model with Poisson observations is given by the formula:

$$P(Y|A,\lambda,t) = \prod_{i=1}^{d} \frac{(t\Lambda_i)^{Y_i}}{Y_i!} e^{-t\Lambda_i},\ \Lambda = A\lambda, \tag{12}$$

where $\lambda \in \mathbb{R}^p_+$ is the parameter, A is the projection matrix, $Y^t = (Y_1^t, Y_2^t, \dots, Y_d^t) \in \mathbb{N}_0^d$ is the measured sinogram. The negative log-likelihood function up to terms independent of λ has the following form:

$$L(\lambda|A, Y^t, t) = -\sum_{i=1}^{d} Y_i^t \log(t\Lambda_i) + \sum_{i=1}^{d} t\Lambda_i,\ \Lambda = A\lambda. \tag{13}$$

Note that likelihood $L(\lambda|\dots)$ as a function of λ depends on it only through $A\lambda$, where A is ill-conditioned. This means that $L(\lambda|\dots)$ is flat in directions in or close to $\ker A$.

SNR: The signal-to-noise ratio for Poisson observations from (12) at line i is the following:

$$\mathrm{SNR}_i(\lambda) = \frac{\mu_i}{\sigma_i} = \frac{t\Lambda_i}{\sqrt{t\Lambda_i}} = \sqrt{t\Lambda_i}. \tag{14}$$

The formula (14) is very informative since it shows that longer measurements (higher t) or higher intensities (larger Λ) result in less noise in data and, hence, less noisy reconstructions. The reality of ETs is that both t and Λ are very-very low because:

1. patient can stay in scanner only limited time inside the scanner ($\sim 20 - 30$ minutes)
2. injected dose is low because the high dose could cause radiation damage and inflict, for example, a secondary cancer
3. large portion of emitted photons are attenuated, scattered or leave the scene without being ever registered

To clarify even better the influence of the amount of registered photons we present the following result from classical statistics

Theorem 2 (folklore for Poisson processes) *Let Y^t be the 1-dimensional Poisson process with intensity Λ. Then*

(1) Law of Large Numbers: $Y^t/t \xrightarrow{a.s.} \Lambda$ for $t \to +\infty$,

(2) Central Limit Theorem: $(Y^t - t\Lambda)/\sqrt{t\Lambda} \xrightarrow{d} \mathcal{N}(0,1)$ for $t \to +\infty$.

Case (1) is clear and corresponds to very high SNR and stable reconstruction of Λ (and also of λ via Theorem 1). Case (2) claims that there is a Gaussian approximation to the Poisson law for large $t\Lambda$, however, for ETs this absolutely not the case and any research on ETs that makes such an approximation falls out any realistic regime. By this paragraph, we wanted to highlight again Claim 10 and bring more mathematical evidence to it.

Maximum Likelihood Estimate and MLEM: Using (13), the Maximum Likelihood estimate $\widehat{\lambda}^t$ is defined by the condition of the gradient of $L(\lambda|\dots)$ being zero[9] and is given by:

$$-\sum_{i=1}^{d} \frac{Y_i^t - t\widehat{\Lambda}_i^t}{\widehat{\Lambda}_i^t} = 0,\ \widehat{\Lambda}^t = A\widehat{\lambda}^t. \tag{15}$$

[9] this is not completely true since λ is not strictly positive and may have zeros at some pixels; optimality condition in this case must involve necessary Karush-Kuhn Tucker conditions, but it is not much relevant for our work, so we omit it in our analysis

Though equation (15) is nonlinear, it can be solved via a classical MLEM algorithm whose iteration is as follows:

$$\widehat{\lambda}_j^{t,(k+1)} = \widehat{\lambda}_j^{t,(k)} \sum_{i=1}^{d} \frac{a_{ij}(Y_i^t/t)}{\widehat{\Lambda}_i^{t,(k)}},\ \widehat{\lambda}_j^{t,(0)} = (1,\dots,1),\ k \to +\infty. \tag{16}$$

Iterations in (16) are convergent to $\widehat{\lambda}^t$ but as the problem is ill-posed, for k too being large solution can be too noisy. As a result, a common regularization technique is to use *early-stopping* to obtain more smooth reconstructions. In all our numerical experiments we fixed a number of iterations (or stopped earlier if reached the tolerance for the update) which is a reasonable scenario when one has a fixed numerical budget per one reconstruction problem.

Prior and the bayesian model: Recall that Question 1 was aimed at the construction of a very informative prior $p(\lambda)$ in the form of a Diffusion-type model (further discussed in subsection 3.1). Having a prior $p(\lambda)$ one defines a bayesian solution to the IP of ETs via Bayes' formula (recall Claims 7, 8):

$$P(\lambda|A, Y^t, t) = \frac{P(Y^t|A, \lambda, t)p(\lambda)}{P(Y^t|A, t)}. \tag{17}$$

Though formula (17) is direct and even analytic (for analytic priors) – the main challenge is to construct an efficient sampler from the posterior which is complicated by the fact that the model is flat in some directions (these actually necessarily must be regularized by the prior). Despite the large amount of literature on designing posterior samplers for inverse problems, the case of ETs has much smaller attention compared to situations where the model is well approximated by a Gaussian one (see e.g. [Goncharov et al., 2023] and references therein).

Dose enhancement: Results of this paragraph refer directly to Question 2. Let Y^{T_2} be a sinogram recorded during time interval $[0, T_2)$ which corresponds to the standard injected dose (or standard measurement length). Assume that another sinogram was recorded during exact same procedure but now during time $[0, T_1)$, where $T_1 < T_2$. It is obvious that Y^{T_1} and Y^{T_2} *are correlated*, for example because

$$Y^{T_1} \le Y^{T_2} \text{ coordinate-wise.} \tag{18}$$

What is the distribution of $Y^{T_1}|Y^{T_2}$? It appears the distribution is described by the following sampling scheme:

$$Y_i^{T_1} \sim \text{Binomial}(Y_i^{T_2}, \frac{T_1}{T_2}) \text{ for each } i. \tag{19}$$

Formula (19) is based on independence between lines (see (11)) and the famous *thinning property of the Poisson process*: if for each event in the Poisson process with intensity Λ we toss a coin with probabilities $(p, 1-p)$ for heads-tails respectively, and add an event to the counter if the side of the coin is heads, then the distribution of the resulting number of counts is again Poisson with intensity $p\Lambda$. Specifically, given $k_1 \le k_2$

$$\begin{aligned}
\mathbb{P}\big(Y_i^{T_1} = k_1, Y_i^{T_2} = k_2\big) &= \mathbb{P}\big(Y_i^{T_1} = k_1, Y_i^{T_2-T_1} = k_2 - k_1\big) &(20)\\
&= \mathbb{P}\big(Y_i^{T_1} = k_1\big)\mathbb{P}\big(Y_i^{T_2-T_1} = k_2 - k_1\big) &(21)\\
&= \frac{e^{-\Lambda T_1}(\Lambda T_1)^{k_1}}{k_1!}\frac{e^{-\Lambda(T_2-T_1)}(\Lambda(T_2-T_1))^{k_2-k_1}}{(k_2-k_1)!} &(22)\\
&= \frac{e^{-\Lambda T_2}(\Lambda T_2)^{k_2}}{k_2!}\binom{k_2}{k_1}\left(\frac{T_1}{T_2}\right)^{k_1}\left(\frac{T_2-T_1}{T_2}\right)^{k_2-k_1} &(23)\\
&= \mathbb{P}\big(Y_i^{T_2} = k_2\big)\mathbb{P}\big(Y_i^{T_1} = k_1 \mid Y_i^{T_2} = k_2\big) &(24)
\end{aligned}$$

Interpreting (19) one can say that having a standard-dose sinogram Y^T one can generate directly low-dose sinograms Y^t, $t < T$ which can be used to train models that infer on Y^{T_2} while having observed only Y^{T_1}. Such a problem is well-known as *dose enhacement* in nuclear imaging (super-resolution in machine learning community) and it is discussed in subsection 3.3.

3 Problem statement

Now we formulate the two problems 1, 2 which we study in this internship. Both are related to Diffusion-type models but target completely different tasks with different complexity. Running a bit ahead – we have a negative result on task 1 and a partially positive result on task 2.

3.1 Diffusion-type models literature overview

At this stage, we do not speak of mathematical foundations of Diffusion-type models[10] but want to show that despite incredibly active development, solutions to inverse problems are only yet to come to practice. Concerning ETs the situation is even worse – there are no attempts at all to construct informative priors as it usually happens and only a few works target the *dose enhancement problem*.

Before proceeding we recall that *unconditional generative model* corresponds when a generative model learns to sample one single distribution, whereas *conditional generative models* learn from a family of distributions being conditioned on some parameter. A typical case of conditional generation (which is if our interest) is to sample from the bayesian posteriors which can be used to solve *inverse problems*.

Denoising Diffusion probabilistic models (DDPM) have recently reached the state-of-the-art on unconditional and conditional image generation tasks: image synthesis of natural images [Dhariwal and Nichol, 2021], [Ho et al., 2020], [Song et al., 2020] on the synthesis of medical images [Müller-Franzes et al., 2022] (but not nuclear imaging), subjective fusion of CT and MRI [Zhu et al., 2020] (see also Claim 10). Among conditional sampling schemes they started also penetrate to the domain of inverse problems: [Song et al., 2021b] (bayesian posterior sampling for X-ray CT), [Song et al., 2022], [Chung et al., 2023] (CT and MRI), [Chung et al., 2022] (claim of solving general inverse problems – even with Poisson data(!); but tested only on natural images on classical tasks such as inpainting, etc.), [Cardoso et al., 2023] (improvement upon the previous article – now posterior can be very far from the prior which was not the case before; Diffusion model learns the prior, but for posterior inference sequential Monte-Carlo is used to guide the diffusion model; solving linear inverse problems with Gaussian noise – this simple structure allows to reduce them to the problem of inpainting).

Schrödinger bridges constitute a very recent extension of DDPMs which at least theoretically is an exact method to transport one distribution into another exactly and in finite time: [Wang et al., 2021], [De Bortoli et al., 2021] (unconditional generation), [Liu et al., 2023] (image-to-image transport of degraded images into enhanced ones without passing by some reference distribution; numerical algorithm is efficient but theoretical results contain serious flaws), [Heng et al., 2023] (guided generative model for posterior sampling for inverse problems).

To sum up there are many works on diffusion models, however, much less on medical imaging [Kazerouni et al., 2023] and even less for Poisson observations: [Chung et al., 2022], [Jiang et al., 2023], [Gong et al., 2023]. For example, [Chung et al., 2022] contains mistakes in theoretical results, and when the Poisson model is used - it is immediately approximated by a Gaussian which is unacceptable for ET models.

3.2 Problem 1: informative priors and regularization for ETs

In principle, if a prior $p(\lambda)$ is learned via a diffusion-type model, there are many ways to organize conditional sampling from the posterior using the same diffusion model (see references in the previous section). A principled approach that we were attracted with was the one developed in [Mardani et al., 2023] – where the learned diffusion model is plugged into reconstruction in form of PnP (Plug-and-Play prior) or RED-approach (Regularization-by-denoising) [Romano et al., 2017]. Though the solution of RED is not bayesian but only a single point estimate, a nonparametric posterior sample could have been constructed using approach from [Goncharov et al., 2023], where sampling is organized as solving an optimization problem with randomized sinograms.

However, even before attempting adaptation of the model for Poisson data – we have encountered serious problems.

From the previous works two important qualitative points were extracted:

[10] at this stage these can be seen as some black-box samplers which are apparently capable of transport reference simple distributions (e.g., a standard Gaussian) into very complex ones

1. diffusion-type models require large datasets (common are MNIST, CelebA, ImageNet containing at least dozens of thousands of images each for each class; there are no such datasets) – this makes sense because the generative model efficiently learns stochastic transport between a Gaussian if dimension of the image and a specific class of images whose effective dimension is much smaller

2. noise in considered inverse problems is often (but not always) is approximated by a Gaussian

Specifically for PET the main problems are:

1. **Data scarcity**: little real PET data accessible online

2. **Data heterogeneity**: sinograms and reconstructions depend nonlinearly on patients' anatomy (recall (4), (5)); data recorded across different machines is not mappable to a single unified format; even two healthy persons may have very different distributions of injected isotope

3. **Low photons count**: the number of photons recorded is very low, even with high recording time – Gaussian approximation of Poisson noise is inadequate

Due to the above reasons and shortage of time we had to abandon prematurely an attempt to construct informative priors and embed then in bayesian reconstructions. In fact, if one analyses the literatrue in PET and its regularization methods – no priors as global PET images were essentially constructed ever, but only local regularization penalties (of TV-type) induced from additional MRI scans[Bowsher et al., 1996], [Bowsher et al., 2004], [Vunckx et al., 2011].

3.3 Problem 2: Dose reduction via conditional generative models

The dose enhancement problem was partially explained in Section 2.2.3. Below we present two types of dose enhancement problems.

Dose enhacement problem in sinogram space. Given recorded sinogram Y^{T_1} sample from the posterior $P(Y^{T_2}|Y^{T_1})$, $T_2 > T_1$, where (Y^{T_1}, Y^{T_2}) correspond to sinograms registered during time intervals $[0, T_1)$, $[T_2 - T_1, T_2)$, respectively.

Dose enhacement problem in reconstruction space. Given recorded sinogram Y^{T_1} sample from the posterior $P(\widehat{\lambda}^{T_2}(Y^{T_2})|\widehat{\lambda}^{T_1}(Y^{T_1}))$, $T_2 > T_1$, where (Y^{T_1}, Y^{T_2}) correspond to sinograms registered during time intervals $[0, T_1)$, $[T_2 - T_1, T_2)$, respectively, $\widehat{\lambda}^t$ is the MLE estimate from (15).

By the time of this report, there are only two works that apply diffusion-type models for dose enhancement in PET in reconstruction space – [Jiang et al., 2023] (DDPM over latent representations), [Gong et al., 2023] (with prior information from MRI). Dose enhancement in sinogram space is more interesting because it is free of the reconstruction algorithm (denoising happens in data space, not in reconstruction space). However, the problem is complicated by the fact that intensity (sinogram for large T) is a function on the manifold – image space of operator A in $\mathbb{R}^d$, hence there should be a separate theoretical study on how to apply existing diffusion models on manifolds [De Bortoli, 2022], [Fishman et al., 2023], to the model of PET.

Methodologically there are no problems with data problems to do enhancement for ETs since the difussion model learns merely a correction to the image to reduce the noise on it. Potentially one could even train a model (a denoiser) on arbitrary synthetic images:

1. Take arbitrary synthetic image λ

2. sample its low and high-dose sinograms: $Y^{T_2} \sim \mathrm{Po}(T_2 \cdot A\lambda)$ for high-dose, $Y^{T_1}|Y^{T_2}$ for low dose using formula (19)

3. update model until convergence

4. use *transfer learning* to specify the trained denoiser to a set of brain slices/full-body slices etc.

In view of the above, we pursued the problem of dose enhancement in reconstruction space.

4 Generative modeling

In this section, we will present a general overview of both conditional and unconditional diffusion models, and more generally of generative modeling processes. We also discuss a class of conditional generative models called *stochastic interpolants* which also seem adapted to PET/SPECT reconstruction problem.

4.1 Unconditional generative modeling

Let's fix the notations for this section. $X_0 \in \mathbb{R}^p$ represents a reconstructed brain slice obtained from fixed exposure time T_2. It is distributed according to law $\mathbb{P}_0$, which is interpreted as the general distribution of patient's brains. We will assume that $\mathbb{P}_0$ is absolutely continuous with respect to Lebesgue's measure and note its density ρ_0

The goal of an unconditional generative model is to estimate ρ_0 by learning a transport map between X_0 and a well-known distribution, which in general is the standard Gaussian law, where $\mathcal{N}\big(x;\ 0, I_p\big)$ represents its density evaluated at point x.

A naive approach consists of connecting X_0 to an independent random variable $W \sim \mathcal{N}\big(0, I_p\big)$ with a straight line. The random variable $X_t = (1-t)X_0 + tW$ with time parametrization $t \in [0,1]$, represents the linear interpolation between these two variables. The law of X_t conditioned on the data X_0 is a Gaussian with mean $(1-t)X_0$ and covariance matrix $t^2 I_p$. To obtain the law of X_t we apply Bayes rule, i.e. $\rho_t(x) = \displaystyle\int_{\mathbb{R}^p} \mathcal{N}\big(x;\ x_0, I_p\big)\rho_0(x_0)dx_0$. But how do we derive a transport map for such interpolation that does not make use of the actual data ?

One way is to identify a velocity field $u_t : \mathbb{R}^p \to \mathbb{R}^p$ such that ρ_t satisfies the Continuity Equation (CE):

$$\partial_t \rho_t + \nabla \cdot (u_t \rho_t) = 0 \tag{25}$$

We say that u_t generates ρ_t. In such case [Villani, 2009], X_t satisfies the *probability flow* ODE :

$$\frac{d}{dt} X_t = u_t(X_t) \tag{26}$$

The inverse implication is also true. If X_t is the solution to the above probability flow ODE, then u_t generates its distribution ρ_t.

Supposing we have access to such velocity field u_t, the map is then given as a sampling procedure as described in Algorithm 1, taking $T = 1$, $\pi = \mathcal{N}\big(0, I_p\big)$ and $\hat{u}_t = u_t$. Sampling a data point following Algorithm 1 is equivalent as sampling it from the data distribution ρ_0.

Algorithm 1 Flow Matching Sampling

```
Require: $\pi$, $\hat{u}_t$, T.
1: Sample $x \sim \pi$
2: Solve ODE $\frac{d}{dt} X_t = \hat{u}_t(X_t), \quad X_T = x$
3: return $X_0$
```

One may ask if such a velocity field u_t exists for our linear interpolation. The answer is yes, a slight relaxation on the assumptions defining a *stochastic interpolant*, introduced in [Albergo et al., 2023a], enables us to derive, in a similar fashion, the existence of the velocity field and a formula for deriving it: $u_t(x) = \mathbb{E}\left[W - X_0 \mid X_t = x\right]$.

Then why not directly learn the velocity field? This is a challenging task as one may notice that the definition of u_t involves conditioning X_0 and W on X_t. Nonetheless, under mild hypothesis, regressing against the velocity field $u_{t|0}(\cdot, x_0)$ which generates the data conditioned density $\rho_{t|0}(\cdot, x_0)$ on some data point $X_0 = x_0$ results in the same optima as regressing against u_t [Lipman et al., 2023, Tong et al., 2023]. To be specific, given an estimator $\hat{u}_t^\theta$ parametrized by θ (i.e. a neural network) we have the equivalence of the following loss

$$L_{FM}(\theta) = \frac{1}{2}\int_0^T \phi(t)\mathbb{E}\left\|\hat{u}_t^\theta(X_t) - u_t(X_t)\right\|_2^2 dt \tag{27}$$

$$\equiv \frac{1}{2}\int_0^T \phi(t)\mathbb{E}\left\|\hat{u}_t^\theta(X_t) - u_{t|0}(X_t \mid X_0)\right\|_2^2 dt \tag{28}$$

where $T = 1$, $\phi(t)$ is a positive time-weighting function and the last expectation is computed with respect to the joint law of (X_0, X_t), that is $\rho_{0,t}(x_0, x) = \rho_{t|0}(x \mid x_0)\rho_0(x_0)$. The class of generative models, obtained by first training an estimator by minimizing the above loss functional and then sampling following the procedure described in Algorithm 1 - where you feed the trained estimator as the input to the algorithm - is known as *Flow Matching*.

4.2 Unconditional Diffusion Models as Score-Based Generative Models (SGMs)

4.2.1 Forward diffusion process

Diffusion models are intrinsically different from the linear interpolation constructed above. While linearly interpolating between X_0 and W is done in finite time, diffusion models map data to a Gaussian in the infinite time limit. Specifically, we consider the *forward* diffusion SDE :

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t, \quad X_0 \sim \rho_0 \tag{29}$$

$$f : \mathbb{R}^p \times [0, +\infty) \to \mathbb{R}^p \quad g : [0, +\infty) \to \mathbb{R}$$

where f is the known as the *drift*, g as the *diffusion coefficient* and $(W_t)_{t\geq 0}$ is a Wiener process. Drift and diffusion coefficient are chosen such that if X_t is the solution to the above diffusion SDE, $\lim_{t\to+\infty} X_t = X_\infty \sim \mathcal{N}(0, \sigma^2 I_p)$, for some $\sigma^2 \in \mathbb{R}_{>0}$. Otherwise, if no stationary distribution exists, we would like to get for a time T large enough, $X_t \approx N(0, \sigma_t^2 I_p)$, for all $t \geq T$. To be rigorous, in the latter case we request that there exists a strictly increasing noise schedule $t \in [0, +\infty) \mapsto \sigma_t \in \mathbb{R}_{\geq 0}$, $\sigma_0 = 0$, $\lim_{t\to+\infty} \sigma_t = +\infty$ and a function $h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, such that $\lim_{t\to+\infty} h(\sigma_t) = 0$ and

$$\exists C > 0 \; \forall t > 0, \quad \mathrm{KL}\left[\rho_t(x) \;\middle|\middle|\; \mathcal{N}(x;\; 0, \sigma_t^2 I_p)\right] \leq C \cdot h(\sigma_t) \tag{30}$$

where the Kullback-Leibler divergence is a type of statistical distance - specifically a divergence - measuring how close two distributions are. In fact, the Kullback-Leibler divergence between two densities is always positive and $p = q$ implies $\mathrm{KL}\left[p(x) \mid\mid q(x)\right] = 0$.

Two classes of solutions we are interested in are the Variance Preserving (VP) and Variance Exploding (VE) processes, solutions to the SDEs:

$$\mathrm{VP} : dX_t = -\frac{1}{2}\sigma_t X_t dt + \sqrt{\sigma_t} dW_t$$
$$\mathrm{VE} : dX_t = \sqrt{2\dot{\sigma}_t \sigma_t} dW_t$$

with σ_t a noise schedule as defined above. Using the Ito integral calculus we get

$$\mathrm{VP} : X_t = X_0 e^{-\frac{1}{2}\int_0^t \sigma_t dt} + \sqrt{1 - e^{-\int_0^t \sigma_s ds}} W_t$$
$$\mathrm{VE} : X_t = X_0 + \sigma_t W_t$$

where $W_t \sim \mathcal{N}(0, I_d)$ are independent and identically distributed for all $t \geq 0$. To verify that VP processes converge to a standard Gaussian one may use the duality between forward diffusion SDE and the forward diffusion Fokker-Plank Equation (FPE) [Øksendal, 2014]. Specifically, there exists an equivalence for a process to be a solution to the forward diffusion SDE and its distribution to be a solution to the forward diffusion FPE :

$$\partial_t \rho_t(x) = -\nabla \cdot f(x,t)\rho_t(x) + \frac{g(t)^2}{2}\nabla \cdot \nabla \rho_t(x) \tag{31}$$

For VP processes the forward diffusion FPE simplify

$$\partial_t \rho_t(x) = \frac{\sigma_t}{2}\nabla \cdot (x\rho_t(x) + \nabla \rho_t(x))$$

Integrating the above equation we get [Franzese et al., 2023], for ρ_t a solution, $\lim_{t\to+\infty} \rho_t(x) = \mathcal{N}\big(x;\, 0, I_p\big)$.

Now for VE processes [Franzese et al., 2023] proved that there exists a function $h(\sigma_t) = \dfrac{1}{\sigma_t^2}$ satisfying bound (30). Hence, both VP and VE belong to the class of diffusion processes.

4.2.2 Backward diffusion process

While reversing the probability flow ODE of the linear interpolant only involved switching the initial condition sampling variable, from the data X_0 to the Gaussian W, the stochastic nature of the forward diffusion SDE renders time reversal a non-trivial task.

Let's fix some upper time limit T from which we would like to reverse the forward diffusion SDE. Under mild assumptions on the data distribution ρ_0 ([Haussmann and Pardoux, 1986, Cattiaux et al., 2022]), the backward process $\big(X_t^B\big)_{t\in[0,T]} = (X_t)_{T-t\in[0,T]}$ solves the *backward* diffusion SDE:

$$\mathrm{d}X_t^B = \Big(- f(X_t^B, t) + g(t)^2 \nabla \log \rho_{T-t}(X_t^B)\Big)\mathrm{d}t + g(t)\mathrm{d}W_t^B, \quad X_0^B \sim \rho_T, \tag{32}$$

where $(W_t^B)_{t\geq 0}$ is another Wiener process. In fact, it is easy to see that X_t^B has marginal distribution ρ_{T-t} for all $t \in [0,T]$. One may notice that the drift of the backward diffusion SDE involves the score $s_t(x) = \log \nabla \rho_t(x)$ of X_t. Therefore, if we had access to an estimator of the score $\hat{s}_t^\theta$, for upper time limit T large enough, so that $X_T \sim \rho_T \approx \mathcal{N}\big(\,\cdot\,; 0, \sigma_T^2 I_p\big)$, where σ_t is the usual noise schedule introduced earlier when no stationary distribution exists or a constant sequence $\sigma_t = \sigma$ for all $t \geq 0$ when a stationary distribution exists, we could simulate a process $\hat{X}_t^{B,\theta}$ solving the following estimated backward diffusion SDE

$$\mathrm{d}\hat{X}_t^{B,\theta} = \Big(- f(\hat{X}_t^{B,\theta}, t) + g(t)^2 \hat{s}_t^\theta(\hat{X}_t^{B,\theta})\Big)\mathrm{d}t + g(t)\mathrm{d}\hat{W}_t^B, \quad \hat{X}_0^{B,\theta} \sim \mathcal{N}\big(0, \sigma_T^2 I_p\big), \tag{33}$$

where $(\hat{W}_t^B)_{t\geq 0}$ is another Wiener process and the distribution $\hat{\rho}_0^\theta$ of $\hat{X}_0^{B,\theta}$ approximates ρ_0. Does $\hat{\rho}_0^\theta$ really approximate ρ_0 and how can we measure "how close" it does it? To answer this question we need to leverage variational inference tools.

4.2.3 A variational perspective on diffusion models

Suppose the estimated score $\hat{s}_t^{\theta^*}$ is obtained by minimisation of the following regression loss :

$$L_{SGM}(\theta) = \frac{1}{2}\int_0^T \psi(t)\mathbb{E}||\hat{s}_t^\theta(X_t) - s_t(X_t)||^2 dt \tag{34}$$

where $\psi(t)$ is a positive time-weighting function. Under mild regularity assumptions on the data, ρ_t, s_t, $\hat{s}_t^\theta$, and selecting $\psi(t) = g(t)^2$ we can bound the Kullback-Leibler divergence between ρ_0 and $\hat{\rho}_0^\theta$ [Song et al., 2021a]

$$\mathrm{KL}\big(\rho_0 \;||\; \hat{\rho}_0^\theta\big) \leq L_{SGM}(\theta) + \mathrm{KL}\big(\rho_T \;||\; \hat{\rho}_T\big) \tag{35}$$

Since distribution $\hat{\rho}_T^\theta = \hat{\rho}_T = \mathcal{N}\big(\;\cdot\;\; 0, I_p\big)$ is given - i.e., is constant for any θ - $\mathrm{KL}\big(\rho_T \;||\; \hat{\rho}_T\big)$ is independent of θ. Furthermore, recalling bound (30) for diffusion processes where no stationary distribution exists and observing that when such a stationary distribution exists for a diffusion process X_t [Franzese et al., 2023]

$$\mathrm{KL}\left[\rho_t \;\Big|\Big|\; \mathcal{N}\big(\,\cdot\,; 0, \sigma^2 I_p\big)\right] \xrightarrow[t\to+\infty]{} 0$$

then, for any diffusion process X_t solution the forward diffusion SDE 29, we have

$$\mathrm{KL}\big(\rho_T \mid\mid \hat{\rho}_T\big) \xrightarrow[t\to+\infty]{} 0$$

This means that for a rich parametric function space $\theta \in \Theta$ - i.e. a complex neural network - and for T large enough, the Kullback-Leibler divergence between ρ_0 and $\hat{\rho}_0$ gets close to 0 for some optimal θ^*. Therefore, minimizing the Kullback-Leibler divergence between ρ_0 and $\hat{\rho}_0$ is approximately equivalent to minimizing L_{SGM} for T large enough.

Finally, let's link the likelihood of the diffusion models to the Kullback-Leibler divergence between ρ_0 and $\hat{\rho}_0$. The negative log-likelihood of the model is written as :

$$\mathbb{E}_{X_0\sim\rho_0}\Big[-\log\hat{\rho}_0^\theta(X_0)\Big] = \int_{\mathbb{R}^P} \log\hat{\rho}_0^\theta(X)\rho_0(x)dx = \underbrace{\int_{\mathbb{R}^P} \log\frac{\hat{\rho}_0^\theta(x)}{\rho_0(x)}\rho_0(x)dx}_{=\mathrm{KL}\left(\rho_0 \ \mid\mid\hat{\rho}_0^\theta\right)} \quad \underbrace{-\int_{\mathbb{R}^P} \log\rho_0(x)\rho_0(x)}_{:=H(\rho_0)} \ dx \tag{36}$$

Where $H(\rho_0)$ is the entropy of ρ_0 which is always positive and is independent of θ. Thus, maximizing the likelihood of the model is equivalent to minimizing the Kullback-Leibler divergence between ρ_0 and $\hat{\rho}_0$.

4.2.4 Score matching objective

In order to compute L_{SGM} we need an analytic expression for the score. However, by definition of a diffusion model, its expression is dependent on the data distribution, which is unknown. To overcome this issue, the trick is to regress the estimator on the data conditional score, which is generally computable. Indeed, for VP and VE processes, X_t conditioned on X_0 is Gaussian. In such case, ([Vincent, 2010]) proved the following equivalence

$$L_{SGM}(\theta) \equiv \frac{1}{2}\int_0^T \psi(t)\mathbb{E}||\hat{s}_t^\theta(X_t) - s_{t|0}(X_t|X_0)||^2 dt \tag{37}$$

where the expectation is taken with respect to the joint law $\rho_{0,t}(x_0, x) = \rho_{t|0}(x \mid x_0)\rho_0(x_0)$ of (X_0, X_t) and the equivalence here means equality up to additive constant - i.e. independent of θ. The equivalence can be extended to more general classes of diffusion processes - and other generative processes.

4.2.5 From training to sampling

We now have everything we need to build our diffusion model. All that's left is to put it all together.

There are two phases to distinguish: training and sampling. The training phase is resumed in Algorithm 2 while sampling in 3. The score estimator is generally a neural network parameterized by θ and initiated at random. We use stochastic gradient descent to optimize the network's parameters and every optimizer step is computed over a mini-batch. More details on the implementation are given in section 5 Once the model has converged, we feed the estimated score to the sampler which solves the SDE 33 using any numerical integrator - more details on the numerical solvers in subsection 5.2.

Algorithm 2 Score-Matching Diffusion Training

```
Require: $\rho_0$, $\rho_{t|0}$, $s_{t|0}$, $\hat{s}_t^\theta$, $\psi(t)$.
  while Training do
    Sample $x_0 \sim \rho_0(x_0)$, $x \sim \rho_{t|0}(x \mid x_0)$, $t \sim \mathcal{U}(0, T)$
    $L_{SGM}(\theta) \leftarrow \frac{1}{2}\psi(t)\|\hat{s}_t^\theta(x) - s_{t|0}(x|x_0)\|^2$
    $\theta \leftarrow \mathrm{Update}(\theta, \nabla_\theta L_{SGM}(\theta))$
  end while
  return $\hat{s}_t^\theta$
```

Algorithm 3 Score-Matching Diffusion Sampling

```
Require: $\hat{\rho}_T$, $\hat{s}^\theta_t$.
    Sample $x \sim \hat{\rho}^\theta_T$
    Solve SDE $\mathrm{d}\hat{X}^B_t = \left(f(\hat{X}^B_t, t) - g(t)^2 \hat{s}^\theta_t(\hat{X}^B_t)\right)\mathrm{d}t + g(t)\mathrm{d}W_t \quad \hat{X}^B_0 = x$
3: return $\hat{X}^B_T$.
```

4.2.6 An alternative deterministic sampler for Diffusion model

One may observe, recalling identity $\nabla \log \rho_t = \frac{\nabla \rho_t}{\rho_t}$ that the forward diffusion FPE (31) factorizes into

$$\partial_t \rho_t(x) = -\nabla \cdot \left(\left(f(x,t) - \frac{g(t)^2}{2}\nabla \log \rho_t(x)\right)\rho_t(x)\right)$$

This is exactly the Continuity Equation (CE) defined in (25) with velocity field $u_t = f(x,t) - \frac{g(t)^2}{2}\nabla \log \rho_t(x)$. Hence, using the duality between CE and probability flow ODE, X_t solves the equation (26). It means that for initial condition $X_0 \sim \rho_0$, solutions to the probability flow ODE have the same marginal distribution ρ_t as solutions to the forward diffusion SDE for all $t \in [0, +\infty)$. If we now define an estimator of the velocity field $\hat{u}^\theta_t$ in terms of the estimated score $\hat{s}^\theta_t$, that is

$$\hat{u}^\theta_t = f(x,t) - \frac{g(t)^2}{2}\hat{s}^\theta_t(x)$$

keeping the training procedure as in Algorithm 2 and picking an upper time limit T large enough, we can obtain an approximation of the data distribution as the solution generated at time t=0 by the flow matching sampler as described in Algorithm 1, where we use $\pi = \hat{\rho}_T$. An illustration of the difference between trajectories generated by the probability flow ODE and the diffusion SDE is presented in figure 6. This sampling procedure is generally faster as generated trajectories are straighter. However, stochastic samplers generally produce higher-quality samples [Albergo et al., 2023a, Karras et al., 2022].

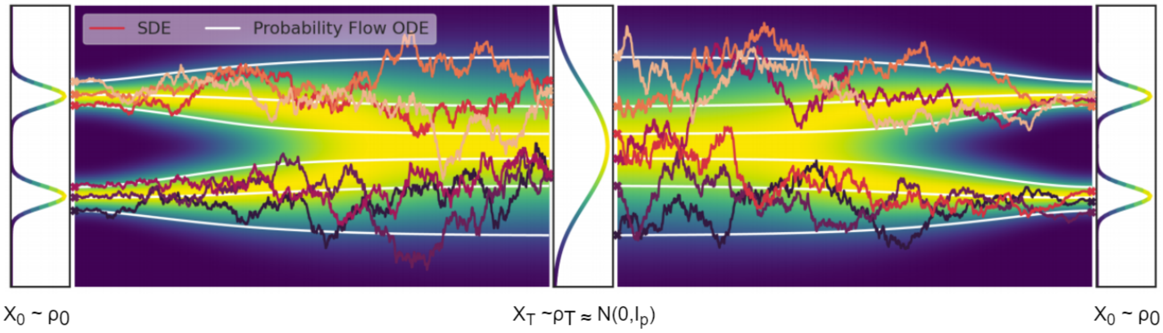


Figure 6: Illustration of the difference between trajectories generated by the probability flow ODE and the diffusion SDE in diffusion models

4.3 Conditional Diffusion models

In this section, we will note $Y \sim \pi$ a reconstructed brain slice obtained from fixed exposure time T_1, such that $T_1 < T_2$, where we recall that T_1 is the time exposure of reconstructed brain slice X_0. We aim to extend the unconditional diffusion model presented in the previous subsection to approximate the posterior distribution $q_0(\cdot|y)$ of X conditioned on $Y = y$ [Heng et al., 2023].

Let $(X_t)_{t\geq 0}$ be the usual process solution to the forward diffusion SDE given by equation (29), however, this time, for fixed $y \in \mathbb{R}^p$, let the initial sampling condition be given by $X_0 \sim q_0(\cdot \mid y)$. Let $q_t(\cdot|y)$ be the distribution of X_t and let $\rho_{t|0}(\cdot|x_0)$ be the transition kernel of the forward diffusion SDE, that is the distribution of X_t conditioned on $X_0 = x_0$. Then applying Bayes' rule we get

$$q_t(x|y) = \int_{\mathbb{R}^p} \rho_{t|0}(x|x_0) q_0(x_0|y) dx_0$$

Under similar assumptions as stated in [Cattiaux et al., 2022, Haussmann and Pardoux, 1986], $X_t^B = X_{T-t}$ is also a solution of the backward diffusion SDE given by equation (32), but where the time-reversed score in the right-hand side is replaced by $s_{T-t}(\cdot, y) = \nabla \log q_{T-t}(\cdot \mid y)$ and the initial sampling condition is given by $X_0^B \sim q_T(\cdot \mid y)$.

For T large enough and for every $y \in \mathbb{R}^p$ we define an estimator $\hat{q}_t^\theta(\cdot, y)$ such that $\hat{q}_T^\theta(\cdot, y) = \hat{q}_T(\cdot, y) = \mathcal{N}\big(\ \cdot\ ; x_0, I_p\big)$. Applying the same variational decomposition to $\hat{q}_t^\theta(\cdot, Y)$, as done in, (36), and taking the expectation on both sides of the equality we obtain :

$$\mathbb{E}_{Y\sim\pi, X_0\sim q_0(\cdot|y)}\Big[-\log \hat{q}_0^\theta(X_0, Y)\Big] = \mathbb{E}_{Y\sim\pi}\Big[H(q_0(\cdot \mid Y)\Big] + \mathbb{E}_{Y\sim\pi}\Big[\text{KL}\left(q_0(\cdot \mid Y) \mid\mid \hat{q}_0^\theta(\cdot \mid Y)\right)\Big]$$

Thus, the expectation of the entropy is independent of θ. Assuming similar hypotheses as in the unconditional case, we apply inequality (35) to the conditional diffusion model. Both sides of the inequality are positive, by growth property of the expectation we get :

$$\mathbb{E}_{Y\sim\pi}\Big[\text{KL}\big(q_0(\cdot|y) \mid\mid \hat{q}_0^\theta(\cdot \mid Y)\big)\Big] \le L_{CSGM}(\theta) + \mathbb{E}_{Y\sim\pi}\Big[\text{KL}\big(q_T(\cdot|y) \mid\mid \hat{q}_T(\cdot \mid Y)\big)\Big]$$

where the Conditional Score-Based Generative Model loss is given by

$$L_{CSGM}(\theta) = \frac{1}{2}\int_0^T \mathbb{E}_{Y\sim\pi,\ X_t\sim q_t(\cdot|Y)} ||\hat{s}_t^\theta(X_t, Y) - s_t(X_t)||^2 dt \tag{38}$$

$$\equiv \frac{1}{2}\int_0^T \mathbb{E}_{Y\sim\pi,\ X_0\sim q_0(\cdot|Y),\ X_t\sim\rho_{t|0}(\cdot|X_0)} ||\hat{s}_t^\theta(X_t, Y) - s_{t|0}(X_t|X_0)||^2 dt \tag{39}$$

where for any $y \in \mathbb{R}^p$, $\hat{s}_t^\theta(\cdot, y)$ is the score of $\hat{q}_t^\theta(\cdot, y)$ and the last equivalence is obtained from [Vincent, 2010] and Fubini theorem. Using similar arguments as for the unconditional case, selecting an upper time limit T large enough, maximizing the likelihood is approximately equivalent to minimizing L_{CSGM}.

4.4 Stochastic Interpolants

In an attempt to present more clearly, in this section we will rename variable Y with X_1 and note ρ_0 the density of X_0 and ρ_1 the density of X_1.

Until now, we have been constructing transport maps from the data X_0 to a well-known Gaussian distribution. This is still the case for the conditional diffusion model we introduced in the previous section. However, intuition would suggest that the "distance" separating a high-dose and low-dose reconstructed images X_0 and X_1, is significantly smaller than the distance separating a high-dose reconstructed image conditioned on the low-dose one $X_0|Y$ to the Gaussian. The goal of this section is to present a novel class of generative models enabling the construction of such transport maps, these are called *stochastic interpolants* [Albergo et al., 2023a]. Specifically, we will be dealing with the most recent version of the model, which has been extended to deal with data-dependent couplings [Albergo et al., 2023b].

The model can handle very intricate data dependencies. We introduce an additional conditioning variable $\xi \sim \eta$, which we can interpret as some common label to both X_0 and X_1. For our purpose, this will correspond to a continuous random variable with support in $[0, 1]$, describing the relative height of the brain slice associated with the pair of high-dose and low-dose reconstructed images (X_0, X_1). In case we were not interested in modeling additional dependencies between X_0 and X_1, consider ξ to be independent of the couple (X_0, X_1) in what will follow. Therefore, we will consider for this section that the distribution of X_0 is $\rho_0(\cdot \mid \xi)$, the distribution of X_1 by $\rho_1(\cdot \mid \xi)$ and the joint distribution of (X_0, X_1) by $\rho_{0,1}(x_0, x_1 \mid \xi) = \rho_{0|1}(x_0 \mid x_1, \xi)\rho_1(x_1 \mid \xi)$, where by abuse of notation ξ is both the random variable and a realization. We recover the marginal probability law through the following integrals :

$$\int_{\mathbb{R}^p} \rho_{0,1}\left(x_0, x_1 \mid \xi\right) dy = \rho_0\left(x_0 \mid \xi\right), \quad \int_{\mathbb{R}^p} \rho_{0,1}\left(x_0, x_1 \mid \xi\right) dx_0 = \rho_1\left(x_1 \mid \xi\right)$$

The process $(X_t)_{t\in[0,1]}$ is said to be a stochastic interpolant if it meets the following conditions :

- There exists $\alpha(t), \beta(t)$, and $\gamma^2(t)$ differentiable functions of time such that $\alpha(0) = \beta(1) = 1$, $\alpha(1) = \beta(0) = \gamma(0) = \gamma(1) = 0$, $\alpha^2(t) + \beta^2(t) + \gamma^2(t) > 0$ for all $t \in [0, 1]$ and

$$X_t = \alpha(t)X_0 + \beta(t)X_1 + \gamma(t)W \quad t \in [0, 1]$$

- $W \sim \mathcal{N}(0, I_p)$, independent of (x_0, x_1, ξ).

Take for example $\alpha(t) = 1 - t, \beta(t) = t$, and $\gamma(t) = \sqrt{2t(1-t)}$. We will note $\rho_t(\cdot \mid \xi)$ the distribution of X_t.

Under these conditions [Albergo et al., 2023b] demonstrated that the velocity field defined as

$$u_t(x \mid \xi) = \dot{\alpha}(t)g_0(t, x, \xi) + \dot{\beta}(t)g_1(t, x, \xi) + \dot{\gamma}(t)g_w(t, x, \xi) \tag{40}$$

where

$$g_0(t, x, \xi) = \mathbb{E}\left(X_0 \mid X_t = x\right), \quad g_1(t, x, \xi) = \mathbb{E}\left(X_1 \mid X_t = x\right), \quad g_w(t, x, \xi) = \mathbb{E}\left(W \mid x_t = x\right)$$

generates the density $\rho_t(\cdot \mid \xi)$; that is $\rho_t(\cdot \mid \xi)$ solves the Continuity Equation (CE) defined in (25) where u_t in the expression is replaced by the velocity field $u_t(\cdot \mid \xi)$ and initial condition is either taken with respect to $\rho_0(\cdot \mid \xi)$ and defined as the density of X_0 or with respect to $\rho_1(\cdot \mid \xi)$ and defined as the density of X_1, both conditions results in the same solution by definition of a stochastic interpolant. By duality between CE and probability flow equation, X_t solves the probability flow ODE defined in (26) with velocity field $u_t(\cdot \mid \xi)$.

Furthermore, the functions g_0, g_1, and g_w are the unique minimizers of the objectives [Albergo et al., 2023b]

$$L_0\left(\hat{g}_0\right) = \int_0^1 \mathbb{E}\left[\left|\hat{g}_0\left(t, X_t, \xi\right)\right|^2 - 2X_0 \cdot \hat{g}_0\left(t, X_t, \xi\right)\right] dt \tag{41}$$

$$L_1\left(\hat{g}_1\right) = \int_0^1 \mathbb{E}\left[\left|\hat{g}_1\left(t, X_t, \xi\right)\right|^2 - 2X_1 \cdot \hat{g}_1\left(t, X_t, \xi\right)\right] dt \tag{42}$$

$$L_w\left(\hat{g}_w\right) = \int_0^1 \mathbb{E}\left[\left|\hat{g}_w\left(t, X_t, \xi\right)\right|^2 - 2W \cdot \hat{g}_w\left(t, X_t, \xi\right)\right] dt \tag{43}$$

where the expectation is taken over $(X_0, X_1) \sim \rho\left(x_0, x_1 \mid \xi\right), \xi \sim \eta$, and $W \sim \mathcal{N}(0, I_p)$. Notice, that we can always recover the third g given 2 of the g's since by definition of the conditional expectation

$$\alpha(t)g_0(t, x, \xi) + \beta(t)g_1(t, x, \xi) + \gamma(t)g_z(t, x, \xi) = x \tag{44}$$

The transport map is finally defined as the following steps :

- **Training**: Train two neural networks $\hat{g}_0^\theta$, $\hat{g}_1^\vartheta$ over two of the three objectives given in (41).
- **Sampling**: Recover the third estimator using identity (44). Compute the estimated velocity field $u_t^{\theta,\vartheta}(\cdot \mid \xi)$ plugging the estimators in expression (40). Sample first $\xi \sim \eta$ and then use the Flow Matching sampling procedure by feeding the estimated velocity field $u_t^{\theta,\vartheta}(\cdot \mid \xi)$, $T = 1$ and initial sampling density $\pi = \rho_1(\cdot \mid \xi)$ as an input to Algorithm 1.

[Albergo et al., 2023b] also proved that for any $\epsilon(t) > 0$, the backward process $X_t^B = X_{1-t}$ is also solution the following SDE

$$dX_t^B = b\left(t, X_t^B, \xi\right) dt + \epsilon(t)\gamma^{-1}(t)g_w\left(t, X_t^B, \xi\right) dt + \sqrt{2\epsilon(t)}dW_t \quad X_{t=0}^B \sim \rho_1\left(\cdot \mid \xi\right)$$

and enjoy the property that $X^B_{t=1} \sim \rho_0\left(\cdot \mid \xi\right)$. The score of the model is given by the expression

$$\nabla \log \rho_t(x \mid \xi) = -\gamma^{-1}(t) g_z(t, x, \xi)$$

This gives a stochastic sampling alternative to the deterministic sampling procedure described earlier using duality between the Fokker Planck Equation expressed above and diffusion SDEs.

5 Model implementation

This section presents the general workflow implementation for the unconditional and conditional diffusion model. Even though we decided only to tackle the dose reduction problem, we wanted to assert that we were able to generate unconditional samples before extending the model to the conditional setting, as the model implementation is actually very similar. That is why we first trained an unconditional diffusion model on images constructed as the concatenation of a low and high dose reconstruction pair along the pixel height axis dimension. Once asserted that the unconditional model worked we trained the conditional diffusion model on pairs of low (the conditioning image) and high dose reconstruction (the target) images.

To implement the models, we first construct a dataset that stores fixed data points. Data points are then selected at random by the dataset sampler which then transforms them with stochastic operations. The output of the dataset sampler is then fed to the neural networks during training, which parametrizes the estimator of the model. Once trained we follow the model sampling procedure to generate data.

We mention here that the implementation of the stochastic interpolants is something we are working on but we assume that it follows closely the implementations of the diffusion models. In fact, we will use the same neural network in all cases.

5.1 Dataset construction

The training dataset is synthetically constructed from twenty 3D brain phantoms extracted from the BrainWeb dataset [Aubert-Broche, 2006]. These are synthetic MRIs, where different brain tissue types are segmented into O classes. For each of these tissue classes, an expert from CEA assigned a specific concentration of the isotope. Each phantom $m \in \{0, ..., M = 20\}$ is sliced along the vertical axis into K 2D images. The skull dimension may vary from one phantom to another, but the first slice starting from the neck always starts at the same location. We select the last slice index $K^m \leq K$ such that the slice is still located inside the skull. For every phantom $m \in \{1, ..., M = 20\}$ we keep slices $l \in \{1, ..., K^m\}$ and discard the rest of the slices.

We introduce the index function $\kappa(m, l) = l + \sum_{j=1}^{m-1} K^j$ and define for every slice $n \in \{1, ..., \kappa(M, K^M)\}$, the label variable $\xi^n = \dfrac{l}{K^m}$ - notice that there exists $m \in \{0, ..., M\}$ and $l \in \{1, ..., K^m\}$ such that $n = \kappa(m, l)$. Assuming that the distribution of brain region proportions along the vertical axis remains relatively constant from one individual to another, ξ^n represents the brain regions associated with the relative height of the slice. This label variable can be used as a conditional random variable. For every slice n, we also pre-compute the projection matrix A^n of that slice (the matrix is inferred from the MRI data).

We now have for every slice $n \in \{1, ..., \kappa(M, K^M)\}$ a triplet (ξ^n, λ^n, A^n), where $\lambda^n \in \mathbb{R}^p$ represents the isotope density of slice n. For each slice n, there exists a partition $\{I^n_1, ..., I^n_O\}$ of the pixel grid $\mathbb{R}^p$ which maps each pixel in the brain slice belonging to its unique tissue class $o \in \{1, ..., O\}$. Since we attributed a constant value across all phantoms for each tissue class, for each brain slice n there exists a bijection between the isotope concentration λ^n and the partition $\{I^n_1, ..., I^n_O\}$. Furthermore, we only care about the ratios of the isotope concentrations among the different classes and not about their actual values. Hence, if $c = (c_0, ..., c_O)$ are the isotope concentrations values per class tissue, we can assume that c belongs to the simplex $\mathbb{S}^{O-1} = \{c \in \mathbb{R}^O \ : \ \sum_{i=1}^{O} c_i = 1\}$.

The isotope concentrations per class tissue could be considered as the average value we find in the general population, as we may assume that the isotope concentration ratios slightly vary from one individual to another. To model some variation, we will construct a map that will perturb the isotope concentrations per class. Given a slice n and parameter $\alpha \in \mathbb{R}_{>0}$, the transformation first samples a random variable D from a Dirichlet distribution of parameter αc, which takes support in the simplex $\mathbb{S}^{O-1}$. The expected value of D is equal to c and the variance of D is a decreasing function of α. We choose parameter α big enough to model the slight variations of isotope concentrations per tissue encountered in the general population. Finally we assign to the partition $\{I_1^n, ..., I_O^n\}$ isotope concentration values $D = (D_1, ..., D_O)$

5.2 Dataset sampler

The function of the sampler is to pick a random data point from the dataset and to output a pair of low-quality and high-quality reconstructed images (corresponding to exposure time $T_1 = 5$ min and $T_2 = 20$ min respectively), defined on the same probability event. We will now explicit the steps the sampler applies to a single data point, however, these steps are in reality applied in parallel to a mini-batch sample.

1. Pick a random sample (ξ, λ, A) from the dataset constructed above.
2. Sample $D \sim \mathcal{D}ir\big(\alpha c\big)$ and assign D to partition $\{I_1, ..., I_O\}$. We obtain a perturbed isotope concentration $\tilde{\lambda} \in \mathbb{R}^p$
3. Construct the high-quality sinogram by sampling $\overline{Y} \sim \bigotimes_{j \in \{1,...,d\}} \mathcal{P}oi\big(T_2 A \tilde{\lambda}\big)$.
4. Construct the low-quality sinogram by sampling $\underline{Y} \sim \bigotimes_{j \in \{1,...,d\}} \mathcal{B}in\big(\overline{Y}, \frac{T_2}{T_1}\big)$.
5. Output a pair of low-quality and high-quality reconstructed images $(\underline{X}, \overline{X}) = (\text{MLEM}(\underline{Y}, \epsilon), \text{MLEM}(\overline{Y}, \epsilon)$

where $\text{MLEM}(\cdot, \epsilon)$ is the maximum likelihood estimation operator defined by iteration (16) which solve the convex problem with tolerance ϵ.

To diversify the dataset, we create an augmentation pipeline that applies the same geometric transformations to the pair of images, such as rotation, translation, dilation, or combinations of these. We have a pool totaling nine geometric transformations and for every sample we apply each of these transformations sequentially with a probability $p_{\text{aug}} \in [0, 1]$ defined as a hyperparameter.

However, we want to apply the augmentation only to training samples and would like to avoid leaking during the sampling phase. We generate an augmentation parameter for each of the pairs of images and set it to zero when we aren't applying any transformation. We then feed it to the neural network as a conditioning variable.

5.3 Network architecture

To implement the different estimators we adopted a reimplementation of the Dhariwal UNet [Dhariwal and Nichol, 2021]. We chose this class of neural networks for its capacity to represent a large functional space. A diagram overview of the network is drawn in figure 7.

The network is divided into the following components :

- An initial convolutional layer embeds the image into the model channel dimension (in the diagram we fixed $C = 128$ channels)

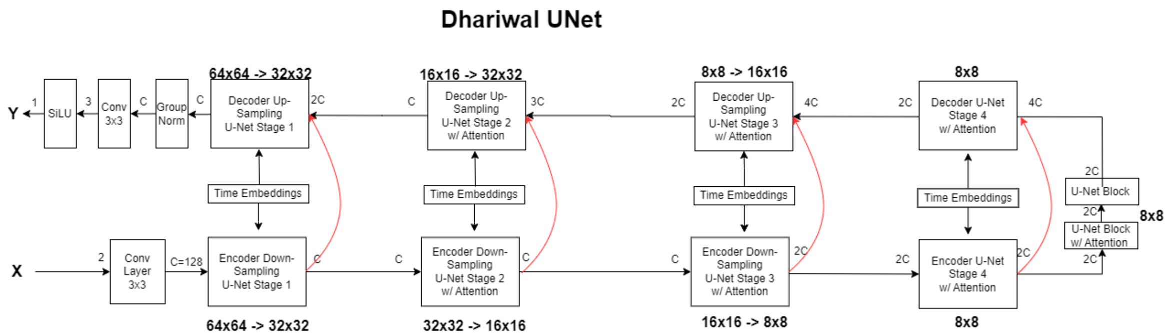

Dhariwal UNet
64x64 -> 32x32
16x16 -> 32x32
8x8 -> 16x16
8x8
Y
1
SiLU
3
Conv 3x3
C
Group Norm
C
Decoder Up-Sampling U-Net Stage 1
2C
C
Decoder Up-Sampling U-Net Stage 2 w/ Attention
3C
2C
Decoder Up-Sampling U-Net Stage 3 w/ Attention
4C
2C
Decoder U-Net Stage 4 w/ Attention
4C
2C
U-Net Block
2C
U-Net Block w/ Attention
2C
8x8
Time Embeddings
Time Embeddings
Time Embeddings
Time Embeddings
X
2
Conv Layer 3x3
C=128
Encoder Down-Sampling U-Net Stage 1
C
C
Encoder Down-Sampling U-Net Stage 2 w/ Attention
C
C
Encoder Down-Sampling U-Net Stage 3 w/ Attention
2C
2C
Encoder U-Net Stage 4 w/ Attention
2C
64x64 -> 32x32
32x32 -> 16x16
16x16 -> 8x8
8x8

Figure 7: Dhariwal UNet - this specific instantiation represents the *conditional* diffusion model implementation.

- Four encoder and decoder stages interconnected with skip connections (see figure 8b). Skip connections concatenate output from the encoder modules to the input of the decoder modules along the channel dimension.
- Each of these stages performs convolutional and attention operations at a specific resolution. Resolution gets divided by two along increasing encoder stages and multiplied by two along decreasing decoder stages. In the diagram, the input image has a resolution of 64 by 64 pixels. The division of the resolution is applied by a convolutional layer located inside a UNet block (see figure 8a) and is appended at the end of an encoder stage. The multiplication of the resolution is performed by a transposed convolutional layer and is placed at the beginning of a decoder stage.
- Stages also perform operations at a specified channel dimension, which is a factor of the model channel base dimension. Here the channel multiplier is one for stage 1 and two for the other stages.
- An intermediate stage is placed between the encoder and the decoder stages and composed of two UNet blocks, one with attention operation appended at the end of it.
- A final decoder stage which outputs the image to the appropriate format.
- A time embedding is used as a conditional variable.

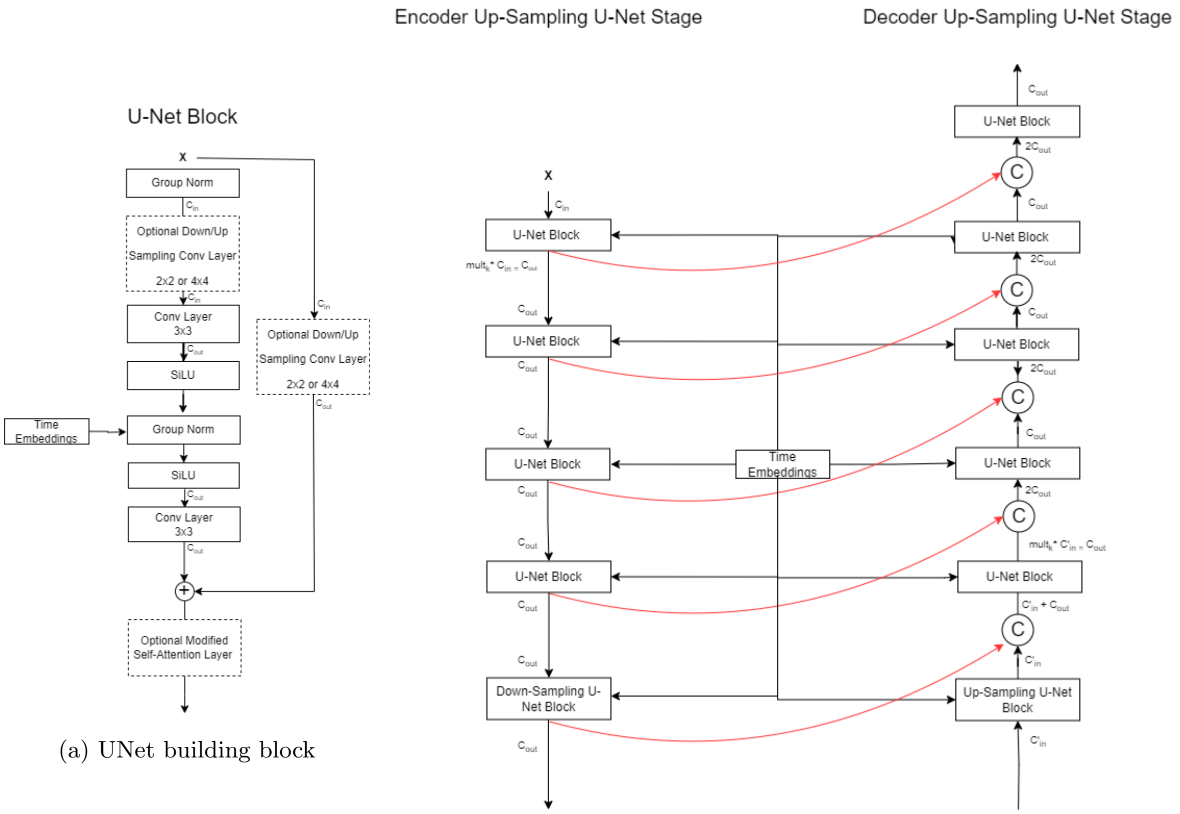
The time-embedding module first creates a vector of frequencies of dimensions d_{freq}, multiplies each of the components of that vector with time t, takes the cosine and sine of that vector and concatenates these two vectors to obtain a vector of dimensions $2d_{\text{freq}}$. In parallel, we linearly project the augmentation parameter a onto $\mathbb{R}^{2d_{\text{freq}}}$ with a matrix of learnable parameters. We then add together the embedded frequency vector with the projected augmentation parameter. We pass them into a non-linear layer such as a SiLU operator. Finally, we linearly project the obtained vector onto $\mathbb{R}^{d_{\text{emb}}}$ with another matrix of learnable parameters. This whole sequence of operation is what we name the time-embedding and note it $\text{emb}(t, a)$.

Time and augmentation conditioning is applied as an adaptive scale normalizing operation placed before every UNet block. Precisely, given input $X \in \mathbb{R}^{C\times H,\times W}$ and two learnable projection matrix $P_1, P_2 \in \mathbb{R}^{C\times d_{emb}}$, we apply transformation

$$Y = P_2\,\text{emb}(t, a) + (P_1\,\text{emb}(t, a) + 1) \times \text{GroupNorm}(X)$$

where $\text{GroupNorm}(\cdot)$ is a type of normalizing operator that standardizes the input. To be precise, given a divisor D of C, such that there exists a $M \in \mathbb{N}^*$ and $C = MD$, then let $X = (X_1, ..., X_M) \in \left(\mathbb{R}^{D\times H\times W}\right)^M$ and given two learnable parameters $\gamma = (\gamma_1, ..., \gamma_M) \in \left(\mathbb{R}^D\right)^M$, $\beta = (\beta_1, ..., \beta_M) \in \left(\mathbb{R}^D\right)^M$ we have

$$\text{GroupNorm}(X) = \left(\frac{X_1 - \hat{\mathbb{E}}\left[X_1\right]}{\hat{\text{Var}}\left[(X_1)\right]}, ..., \frac{X_1 - \hat{\mathbb{E}}\left[X_M\right]}{\hat{\text{Var}}\left[X_M\right]}\right)$$


Encoder Up-Sampling U-Net Stage
Decoder Up-Sampling U-Net Stage
U-Net Block
x
Group Norm
C_{in}
Optional Down/Up Sampling Conv Layer 2x2 or 4x4
Conv Layer 3x3
C_{out}
SiLU
Time Embeddings
Group Norm
SiLU
Conv Layer 3x3
Optional Down/Up Sampling Conv Layer 2x2 or 4x4
Optional Modified Self-Attention Layer
U-Net Block
$mult_k * C_{in} = C_{out}$
Down-Sampling U-Net Block
Time Embeddings
C
$2C_{out}$
$mult_k * C'_{in} = C_{out}$
$C'_{in} + C_{out}$
C'_{in}
Up-Sampling U-Net Block

(a) UNet building block

(b) UNet encoder stage (left) and decoder stage (right)

Figure 8: UNet components

and

$$\hat{\mathbb{E}}[X_i] = \frac{1}{D \times H \times W} \sum_{(d,h,w) \in \{1,\dots,D\} \times \{1,\dots,H\} \times \{1,\dots,W\}} (X_i)_{d,h,w} \tag{45}$$

$$\hat{\text{Var}}[X_i] = \hat{\mathbb{E}}[X_i^2] - \hat{\mathbb{E}}[X_i]^2 \tag{46}$$

A UNet block (figure 8a) is the building block of the encoder and decoder stages. One stage is composed of four UNet blocks and one down or upsampling UNet block. UNet blocks are composed of GroupNorm operators, convolutional layers and an optional attention layer. The down and upsampling UNet block has an additional convolutional and or transposed convolutional layer. The block also possesses a residual connection.

The input and output of the neural network varies for the different models. In the unconditional diffusion model, the input is the concatenation of the noise-corrupted pair $X_t = (\underline{X}, \overline{X}) + t(\underline{W}, \overline{W})$ along the height dimension, that is $X_t \in \mathbb{R}^{C \times 2H \times W}$, where $\underline{W}$ and $\overline{W}$ are two independent standard Gaussians. The output has the same dimensions. In the conditional diffusion model, the input is the concatenation of the pair $(X_t, Y) = (\overline{X} + t\overline{W}, \underline{X})$ along the channel dimension, that is $(X_t, Y) \in \mathbb{R}^{2C \times H \times W}$. The output, however, is a single reconstructed image and is of dimensions $C \times H \times W$, as Y serves as a conditional variable.

5.4 Training

To select the specific diffusion process we followed [Karras et al., 2022], which extensively explored the design space for diffusion models. We chose a VE process with a noise schedule equal to the identity function. That is, we model the process

$$\forall t \in [0, +\infty) \quad X_t = X_0 + tW$$

where $X_0 \sim \rho_0$, $W \sim \mathcal{N}(\,\cdot\,; 0, I_p)$ and is independent of X_0 for all $t \geq 0$. We enforce notations employed in the unconditional case, however, ρ_0 should be replaced by $q_0(\cdot \mid y)$, where $Y = y$ is sampled from π, the low-quality reconstructed images for the unconditional case. This process is a solution of the forward diffusion SDE

$$dX_t = \sqrt{2t}dW_t \quad X_0 \sim \rho_0$$

and of the probability flow ODE :

$$X_t = -t\nabla \log \rho_t(X_t)dt \quad X_0 \sim \rho_0$$

where W_t is a Wiener process and $s_t(x) = \nabla \log \rho_t(x)$ is the score of the model. Similarly as for the previous remark, read here $q_t(\cdot \mid y)$ instead of ρ_t for the conditional case. For some upper time limit T, $(X_t)_{t\in[0,T]}$ is also a solution to the probability flow ODE with initial condition $X_T \sim \rho_T$ and solved backward in time, where ρ_T the distribution of the solution to one of the two equations written above, evaluated at time $t = T$. The associated backward diffusion process $(X_t^B)_{T-t\in[0,T]}$ is solution to the backward diffusion SDE :

$$dX_t^B = 2ts_{T-t}(X_t^B)dt + \sqrt{2t}dW_t^B \quad X_0^B \sim \rho_T$$

where W_t^B is another Wiener process. The data-conditioned score of the model is given by the expression

$$\forall (x, x_0) \in \left(\mathbb{R}^p\right)^2, \quad s_{t|0}(x \mid x_0) = \frac{x_0 - x}{t^2}$$

Instead of directly regressing the score estimator $\hat{s}_t^\theta(x)$ against the data-conditioned score $s_{t|0}(x|x_0)$ ($\hat{s}_t^\theta(x, y)$ against $s_{t|0}(x|x_0)$ for the conditional diffusion model and remember $X_t = x \sim q_t(\cdot \mid y)$ and in particular $X_0 = x_0 \sim q_0(\cdot \mid y)$), it is equivalent to consider a denoiser function $\hat{D}_t^\theta(x)$ (read here $\hat{D}_t^\theta(x \mid y)$ for the conditional diffusion model) that minimizes the expected L_2 denoising error for samples drawn ρ_0 [Karras et al., 2022]. Specifically, we define the estimator

$$\forall (x, t) \in \mathbb{R}^p \times [0, +\infty) \quad \hat{D}_t^\theta(x) = t^2\hat{s}_t^\theta(x) + x$$

and sought to solve the following equivalent objective

$$L_{SGM}(\theta) \equiv \int_0^T \psi(t)\mathbb{E}||\hat{D}_t^\theta(X_t) - X_0||^2 dt \tag{47}$$

$$\equiv \int_0^T \psi(t)\mathbb{E}||\hat{D}_t^\theta(X_0 + tW) - X_0||^2 dt \tag{48}$$

A good practice in deep learning is to keep the input and output signal magnitude fixed to unit variance and avoid large variations of signal magnitude. Training a neural network D_t^θ directly is therefore far from optimal as noise-corrupted data may vary immensely depending on noise level. Predicting the noise component of the noise-corrupted data, i.e. defining a noise estimator $\hat{F}_t^\theta$ such that $\hat{D}_t^\theta(x) = x - t\hat{F}_t^\theta(x)$, results in error amplification by a factor t which is problematic for high noise corruption levels. As a result, we choose to adaptively mix clean data signal and noise. We define an estimator $\hat{F}^\theta(t, x)$, i.e. a neural network and noise-dependent weighting functions $c_{\text{skip}}(t)$, $c_{\text{in}}(t)$, $c_{\text{out}}(t)$ and $c_{\text{noise}}(t)$ such that :

$$\hat{D}_t^\theta(x) = c_{\text{skip}}(t)x + c_{\text{out}}(t)\hat{F}^\theta(c_{\text{noise}}(t), c_{\text{in}}(t)x)$$

so that we can reformulate the loss function as

$$\int_0^T \underbrace{\psi(t)c_{\text{out}}(t)^2}_{\text{effective noise weight}} \; \mathbb{E}\left[\left|\left| \underbrace{\hat{F}^\theta\big(c_{\text{noise}}(t), c_{\text{in}}(t)(X_0 + W)\big)}_{\text{network output}} - \underbrace{\frac{1}{c_{\text{out}}(t)}\Big(X_0 - c_{\text{skip}}(t)(X_0 + W)\Big)}_{\text{effective training target}} \right|\right|^2\right] dt \tag{49}$$

where $c_{\text{in}}(t)$ is defined such that weighted input $c_{\text{in}}(t)X_t$ has unit variance, $c_{\text{out}}(t)$ is defined as a function of $c_{\text{skip}}(t)$ such that effective training target has unit variance, $c_{\text{skip}}(t)$ as the minimizer of the $c_{\text{out}}(t)$ and c_{noise} is chosen empirically following [Karras et al., 2022].

Finally, we define $\psi(t) = \dfrac{\tilde{\psi}(t)}{c_{\text{out}}(t)^2}$ such that $\tilde{\psi}(t)$ is a probability density with respect to Lebesgue's measure on $[0, T]$. If we define a uniform probability law on $[0, T]$, then the effective noise weight is constant for all noise levels. In such a case, [Karras et al., 2022] found that the loss magnitude was higher for intermediate noise levels. To target the training efforts to the relevant range, we define $\tilde{\psi}(t)$ as the density of a log-normal, that is we set $\ln(t) \sim \mathcal{N}\big(\,\cdot\,; P_{\text{mean}}, P_{\text{std}}^2 I_p\big)$ - where t take support on $(0, +\infty)$ in that case, that is, take the limit of $T \to +\infty$ in loss (49). This means that we can replace the uniform time-sampling step in Algorithm 3 by directly sampling t as a log-normal random variable. We resume the *unconditional* diffusion training procedure in Algorithm 4. A very similar preconditioning procedure is applied to the *conditional* diffusion model.

Algorithm 4 Unconditional Diffusion Training

```
Require: $\rho_0$, $\hat{F}^\theta$
  while Training do
    Sample $X_0 \sim \rho_0$, $W \sim \mathcal{N}\big(\,\cdot\,; 0, I_p\big)$, $\ln(t) \sim \mathcal{N}\big(\,\cdot\,; P_{\text{mean}}, P_{\text{std}}^2 I_p\big)$
    $L_{SGM}(\theta) \leftarrow \left|\left|\hat{F}^\theta\big(c_{\text{noise}}(t), c_{\text{in}}(t)(X_0 + W)\big) - \frac{1}{c_{\text{out}}(t)}\Big(X_0 - c_{\text{skip}}(t)(X_0 + W)\Big)\right|\right|^2$
    $\theta \leftarrow \text{Update}(\theta, \nabla_\theta L_{SGM}(\theta))$
  end while
  return $\hat{F}^\theta$
```

5.5 Sampling

Given some upper time limit T, we follow the sampling procedure proposed in [Karras et al., 2022] which employs Heun's second-order method - an ODE solver with trapezoid 2nd-order correction - to solve ODE

$$d\hat{X}_t = -t\hat{s}_t^\theta(X_t)dt \quad X_T \sim \mathcal{N}\big(\,\cdot\,; 0, T^2 I_p\big)$$

backward in time or a modified stochastic version that solves the SDE

$$d\hat{X}_t^{B,\theta} = 2t\hat{s}_{T-t}^\theta(\hat{X}_t^{B,\theta})dt + \sqrt{2t}d\hat{X}_t^B \quad \hat{X}_t^B \sim \mathcal{N}\big(\,\cdot\,; 0, T^2 I_p\big)$$

6 Results

The unconditional diffusion ran successfully for image resolution 32x32pixels. In figure 9 we observe a couple of generated pairs of low-quality and high-quality reconstructed images (right) and compare them to a random couple of pairs sampled from the original dataset (left). Validating this step was essential to test the soundness of our diffusion model in the context of PET/SPECT medical imaging.

We next trained the unconditional diffusion model for the same resolution and at inference obtained excellent results as shown in the set of images 10.

7 Conclusion and future perspectives

To conclude we review what has been done during the internship.

1. Familiarization with PET/SPECT inverse problem and enunciation of a statistical model.

2. Identification of two modalities of improvement : (**prior regularization** and **dose reduction**). Reviewing the different problematics circumventing the former approach, we finally decided to only tackle the **dose reduction** problem.

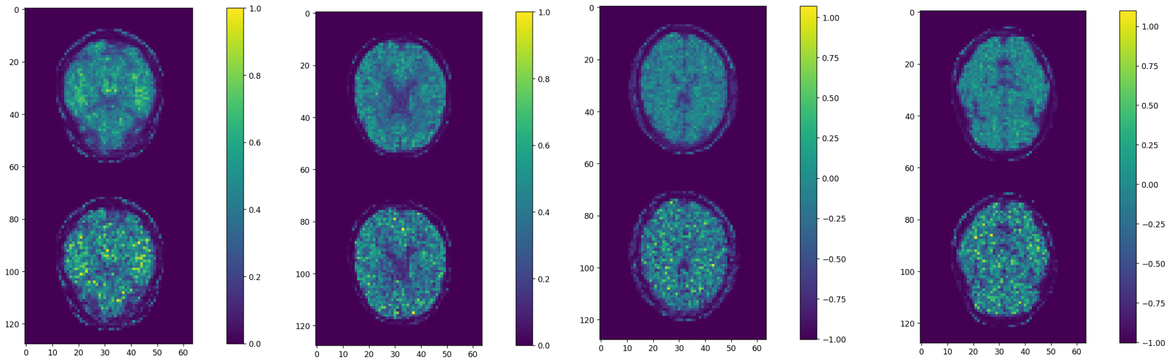


Figure 9: Pairs of low-quality and high-quality images for the *unconditional* model. The first on the left is sampled from the dataset, the other three are generated from the unconditional diffusion model (right)

3. Review of the literature on generative models and particularly on diffusion models.
4. Careful design of a synthetic dataset and associated sampler adapted to the dose reduction problem.
5. Implementation of an unconditional and conditional diffusion model applied to the PET/SPECT inverse problem.

Future perspectives includes

1. Extension of the synthetic dataset. For example, there exists a very large MRI database. If we could find a map that transforms the MRIs into phantoms we could significantly enrich the model.
2. Fine-tuning the model on real patient data, by first learning on a synthetic dataset.
3. Modelization of a transport map in the sinogram space. This would require a separate study of the geometrical properties of sinograms.
4. Derive further theoretical insurances on the adeptness of stochastic interpolants. Specifically, find an asymptotic vanishing bound on the Kullback-Leibler divergence between $\rho_{0|1}$ and its estimator similarly as for diffusion models.
5. Select a specific process for stochastic interpolants and tailor a preconditioning procedure to train the model as done for diffusion models.
6. Compare the different performances between the conditional diffusion model and the stochastic interpolants.
7. Develop a metric for generative models adapted to medical imaging.

Let's further detail the last point. To measure the quality of the generated samples, literature on generative modeling generally adopts Frechet Inception Distance (FID). Given two probability distributions γ, υ on $\mathbb{R}^p$, the FID between these two measures is defined as the 2-Wasserstein distance

$$W_2(\gamma, \upsilon) = \min_{\Pi \in \Gamma(\gamma, \upsilon)} \left(\int_{\mathbb{R}^p \times \mathbb{R}^p} ||x - y||^2 d\Pi(x, y) \right)^{\frac{1}{2}}$$

where $\Gamma(\gamma, \upsilon)$ is the set of all possible joint distributions between γ and υ. For two multidimensional Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$, we get the closed form

$$W_2\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')\right)^2 = \|\mu - \mu'\|_2^2 + \operatorname{tr}\left(\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}} \cdot \Sigma' \cdot \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$$

We define the FID computation for general generative models in pseudocode form:

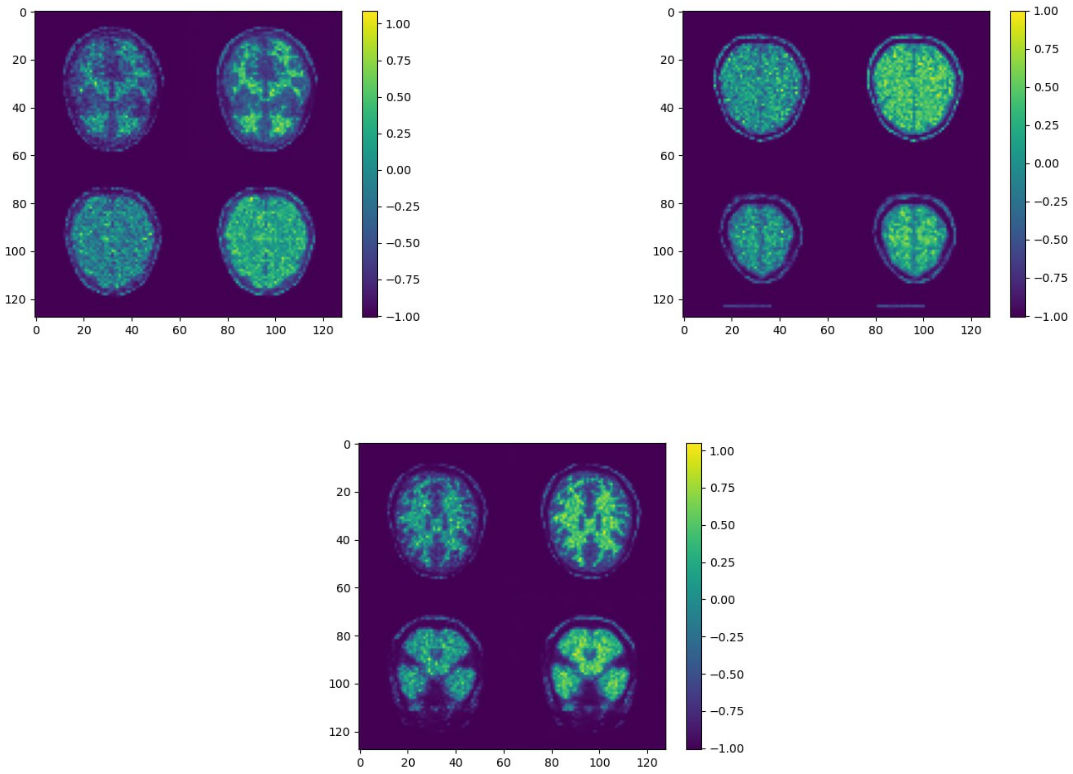


Figure 10: Pairs of low-quality and high-quality images generated from the trained *conditional* diffusion model. The image on the left is the conditioning image and the one the right is the denoised version.

1. Input a function $f : \mathbb{R}^p \rightarrow \mathbb{R}^{p'}$.
2. Input two datasets $S, S' \subset \mathbb{R}^p$, where S is the training dataset and S' is a generated datset from the trained generative model.
3. Compute $f(S), f\left(S'\right) \subset \mathbb{R}^{p'}$.
4. Fit two gaussian distributions $\mathcal{N}(\mu, \Sigma), \mathcal{N}\left(\mu', \Sigma'\right)$, respectively for $f(S), f\left(S'\right)$.
5. Return $W_2\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}\left(\mu', \Sigma'\right)\right)^2$

Here f is an Inception v3 model trained on the ImageNet and amputated from its classifier head. ImageNet is a huge dataset of images labeled with over 22 thousand classes ranging from object descriptions, face features, and other object labels we encounter in everyday life. However, these classes aren't adapted to medical imaging. Indeed, they do not include feature descriptions of PET/SPECT scans and therefore cannot discriminate between the different brain images. The output $f(S)$ and $f(S')$ will both be concentrated in a very narrow region of the $\mathbb{R}^{p'}$. That is why we are currently working on a metric-adapted to the dose enhancement task.

A last word on point 3 of the future perspectives. As generally $p >> d$, that is we record a much higher number of lines of responses than there are pixels in the reconstructed images, the image of the projection matrix A is a manifold of dimension d embedded in $\mathbb{R}^p$. More precisely, as the isotope concentration $\lambda \geq 0$, the image forms a cone. Then, invoking tools from information geometry, the space of sinograms is seen as a doubly flat statistical manifold. This approach could open new perspectives to solve PET/SPECT inverse problem.

References

[Albergo et al., 2023a] Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. (2023a). Stochastic interpolants: A unifying framework for flows and diffusions.

[Albergo et al., 2023b] Albergo, M. S., Goldstein, M., Boffi, N. M., Ranganath, R., and Vanden-Eijnden, E. (2023b). Stochastic interpolants with data-dependent couplings.

[Aubert-Broche, 2006] Aubert-Broche, B., E. A. C. . C. L. (2006). A new improved version of the realistic digital brain phantom. neuroimage.

[Bowsher et al., 1996] Bowsher, J., Johnson, V., Turkington, T., Jaszczak, R., Floyd, C., and Coleman, R. (1996). Bayesian reconstruction and use of anatomical a priori information for emission tomography. *IEEE Transactions on Medical Imaging*, 15(5):673–686.

[Bowsher et al., 2004] Bowsher, J., Yuan, H., Hedlund, L., Turkington, T., Akabani, G., Badea, A., Kurylo, W., Wheeler, C., Cofer, G., Dewhirst, M., and Johnson, G. (2004). Utilizing mri information to estimate f18-fdg distributions in rat flank tumors. In *IEEE Symposium Conference Record Nuclear Science*, volume 4. IEEE.

[Cardoso et al., 2023] Cardoso, G., Idrissi, Y. J. E., Corff, S. L., and Moulines, E. (2023). Monte carlo guided diffusion for bayesian linear inverse problems. *arXiv preprint arXiv:2308.07983*.

[Cattiaux et al., 2022] Cattiaux, P., Conforti, G., Gentil, I., and Léonard, C. (2022). Time reversal of diffusion processes under a finite entropy condition.

[Chung et al., 2022] Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. (2022). Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.

[Chung et al., 2023] Chung, H., Ryu, D., McCann, M. T., Klasky, M. L., and Ye, J. C. (2023). Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22542–22551.

[De Bortoli, 2022] De Bortoli, V. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*.

[De Bortoli et al., 2021] De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709.

[Dhariwal and Nichol, 2021] Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis.

[Fishman et al., 2023] Fishman, N., Klarner, L., De Bortoli, V., Mathieu, E., and Hutchinson, M. (2023). Diffusion models for constrained domains. *arXiv preprint arXiv:2304.05364*.

[Franzese et al., 2023] Franzese, G., Rossi, S., Yang, L., Finamore, A., Rossi, D., Filippone, M., and Michiardi, P. (2023). How much is enough? a study on diffusion times in score-based generative models. *Entropy*, 25(4):633.

[Goncharov, 2019] Goncharov, F. (2019). *Weighted Radon transforms and their applications.* PhD thesis, Université Paris Saclay (COmUE).

[Goncharov et al., 2023] Goncharov, F., Barat, E., and Dautremer, T. (2023). Nonparametric posterior learning for emission tomography. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):452–479.

[Gong et al., 2023] Gong, K., Johnson, K., El Fakhri, G., Li, Q., and Pan, T. (2023). Pet image denoising based on denoising diffusion probabilistic model. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–11.

[Haussmann and Pardoux, 1986] Haussmann, U. G. and Pardoux, E. (1986). Time Reversal of Diffusions. *The Annals of Probability*, 14(4):1188 – 1205.

[Heng et al., 2023] Heng, J., Bortoli, V. D., and Doucet, A. (2023). Diffusion schrödinger bridges for bayesian computation.

[Ho et al., 2020] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

[Jiang et al., 2023] Jiang, C., Pan, Y., Liu, M., Ma, L., Zhang, X., Liu, J., Xiong, X., and Shen, D. (2023). Pet-diffusion: Unsupervised pet enhancement based on the latent diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer.

[Karras et al., 2022] Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models.

[Kazerouni et al., 2023] Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846.

[Krishnan and Quinto, 2015] Krishnan, V. P. and Quinto, E. T. (2015). Microlocal analysis in tomography. *Handbook of mathematical methods in imaging*, 1:3.

[Lipman et al., 2023] Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling.

[Liu et al., 2023] Liu, G.-H., Vahdat, A., Huang, D.-A., Theodorou, E. A., Nie, W., and Anandkumar, A. (2023). I ˆ2 sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*.

[Marcu et al., 2018] Marcu, L. G., Moghaddasi, L., and Bezak, E. (2018). Imaging of tumor characteristics and molecular pathways with pet: developments over the last decade toward personalized cancer therapy. *International Journal of Radiation Oncology Biology Physics*, 102(4):1165–1182.

[Mardani et al., 2023] Mardani, M., Song, J., Kautz, J., and Vahdat, A. (2023). A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*.

[Müller-Franzes et al., 2022] Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nebelung, S., Kather, J. N., et al. (2022). Diffusion probabilistic models beat gans on medical images. *arXiv preprint arXiv:2212.07501*.

[Natterer, 2001] Natterer, F. (2001). *The mathematics of computerized tomography*. SIAM.

[Natterer and Hadeler, 1980] Natterer, F. and Hadeler, K. (1980). Efficient implementation of 'optimal'algorithms in computerized tomography. *Mathematical Methods in the Applied Sciences*, 2(4):545–555.

[NIH, 2016] NIH (2016). ://nibib.nih.gov/science-education/science-topics/nuclear-medicine Cited 30.10.2023.

[Øksendal, 2014] Øksendal, B. (2014). *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edition.

[Reader et al., 2020] Reader, A. J., Corda, G., Mehranian, A., da Costa-Luis, C., Ellis, S., and Schnabel, J. A. (2020). Deep learning for pet image reconstruction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(1):1–25.

[Romano et al., 2017] Romano, Y., Elad, M., and Milanfar, P. (2017). The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844.

[Shepp and Vardi, 1982] Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE transactions on medical imaging*, 1(2):113–122.

[Siddon, 1985] Siddon, R. L. (1985). Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical physics*, 12(2):252–255.

[Song et al., 2022] Song, J., Vahdat, A., Mardani, M., and Kautz, J. (2022). Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*.

[Song et al., 2021a] Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models.

[Song et al., 2021b] Song, Y., Shen, L., Xing, L., and Ermon, S. (2021b). Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*.

[Song et al., 2020] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

[Toft, 1996] Toft, P. (1996). The radon transform. *Theory and Implementation (Ph. D. Dissertation)(Copenhagen: Technical University of Denmark)*.

[Tong et al., 2023] Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. (2023). Conditional flow matching: Simulation-free dynamic optimal transport.

[Villani, 2009] Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

[Vincent, 2010] Vincent, P. (2010). A connection between score matching and denoising autoencoders. *Technical Report 1358, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal*. Preprint version.

[Vunckx et al., 2011] Vunckx, K., Atre, A., Baete, K., Reilhac, A., Deroose, C. M., Van Laere, K., and Nuyts, J. (2011). Evaluation of three mri-based anatomical priors for quantitative pet brain imaging. *IEEE transactions on medical imaging*, 31(3):599–612.

[Wang et al., 2021] Wang, G., Jiao, Y., Xu, Q., Wang, Y., and Yang, C. (2021). Deep generative learning via schrödinger bridge. In *International Conference on Machine Learning*, pages 10794–10804. PMLR.

[Zhu et al., 2020] Zhu, R., Li, X., Zhang, X., and Ma, M. (2020). Mri and ct medical image fusion based on synchronized-anisotropic diffusion model. *IEEE Access*, 8:91336–91350.