Nonparametric posterior learning for emission tomography with multimodal data

Fedor Goncharov, Éric Barat, Thomas Dautremer

Nonparametric posterior learning for emission tomography with multimodal data

Fedor Goncharov, Éric Barat, and Thomas Dautremer

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

September 10, 2021

Abstract

In this work we continue studies of the uncertainty quantification problem in emission tomographies such as PET or SPECT. In particular, we consider a scenario when additional multimodal data (e.g., anatomical MRI images) are available. To solve the aforementioned problem we adapt the recently proposed nonparametric posterior learning technique to the context of Poisson-type data in emission tomography. Using this approach we derive sampling algorithms which are trivially parallelizable, scalable and very easy to implement. In addition, we prove conditional consistency and tightness for the distribution of produced samples in the small noise limit (i.e., when the acquisition time tends to infinity) and derive new geometrical and necessary condition on how MRI images must be used. This condition arises naturally in the context of misspecified generalized Poisson models. We also contrast our approach with bayesian MCMC sampling based on one data augmentation scheme which is very popular in the context of EM-type algorithms for PET or SPECT. We show theoretically and also numerically that such data augmentation significantly increases mixing times for the Markov chain. In view of this, our algorithms seem to give a reasonable trade-off between design complexity, scalability, numerical load and asessement for the uncertainty quantification.

1 Introduction

Emission tomographies (further referred as ET) such as PET (positron emission tomography) or SPECT (single photon emission computed tomography) are functional imaging modalities of nuclear medicine which are used to image activity processes and, in particular, metabolism in soft tissues (e.g., to measure the uptake of a certain biomarker). The level of metabolism provides critical information for diagnostics and treatment of cancers; see e.g., [YCHT04], [Web05], [MMB18] and references therein.

In this work we continue studies on the two following problems:

Problem 1. Quantify the uncertainty of reconstructions in ET.

Problem 2. Use multimodal data (e.g., images from CT or MRI) to regularize the underlying inverse problem.

Problem 1 is not new and several approaches have been established already which in turn can be grouped according to the statistical view of the problem: frequentist [Fes96],

[BWT94], [Li11], bayesian [GM87], [HBJ+97], [Wei97], [BCD+07], [FL07], [Sit12], [BG14], [FBD+18] and bootstrap [HW89], [Dah01], [LAB10]. The list of given references is far from being complete and it should also include references therein.

Problem 2 can be splitted further depending on which type of exterior data are used, CT or MRI. The most common use in the literature of both modalities consists in extracting boundaries of features on side images and embedding them in regularization schemes via special penalties and/or non-invariant filters; see e.g., [FCR92], [CFD13], [HPFT99], [CKF+01], [VAB+11]. Main reasons to use multimodal data in ET are the ill-posedness of corresponding inverse problems (in PET/SPECT forward operators are ill-conditioned; see e.g., [HW16], [Gon19]) and very low signal-to-noise ratio in raw measured data. All this together results in loss of resolution and often in oversmoothing in reconstructed images (e.g., when applying spatially invariant filters for post-smoothing). The modality of PET(SPECT)-CT is well known and first efforts to improve the image quality were made even before the advent of ET-CT coupled scanners. Since then there is a concensus that side information may substantially improve the quality of images in ET but should be used carefully in case of misalignment between features on ET images and side ones; see, e.g., [FCR92], [CKF+01]. Problem 2 for additional MRI data is more recent and it is of particular interest due to appearance of commercially available models of PET-MRI scanners [LVHR13], [JWN+08] which allow simultaneous registrations of both signals, thus significantly reducing motion effects. In the experiment on tumor imaging in [BYH+04] correlations between PET and MRI signals were observed. Therefore, potentially MRI data can be used to regularize accurately the inverse problem; see also [BJT+96], [VAB+11], [FBD+18]. In this work as multimodal data we use series of presegmented anatomical MRI images. In Section 2 we explain in detail how we use MRI data and compare our approach with previous works.

An important feature of this work is that Problems 1, 2 are considered simultaneously. Main conceptual difficulty here is that there is no precise definition of an optimal solution, where the latter can be seen as some numerical algorithm. To be more precise such algorithm should take as an input raw ET data (sinogram or list-mode), acquisition geometry, MRI data, calibration parameters for regularization and parameters which are necessary for uncertainty quantification (e.g., region of interest in the final image, desired statistic (mean, median, quantiles etc.). As an output we expect exactly what the algorithm is designed for given the input parameters. However, the most universal form of an algorithm would be a sampler which at the end generates reconstructed images following some probability distribution. Then, using generated samples one can perform inference on any desired statistic (modulo numerical resources available).

Already the definition of uncertainty for reconstructions in ET is not obvious: during time interval $(0,t)$ raw data Y^t (sinogram) is generated from unknown distribution P^t (typically it is assumed to be a generalized Poisson model with unknown intensity parameter λ_* and known design A, i.e., $P^t = P^t_{A,\lambda} = \mathrm{Po}(t \cdot A\lambda_*)$), so any reconstruction $\widehat{\lambda}^t$ would be also a function of observed data, that is $\widehat{\lambda}^t = \widehat{\lambda}^t(Y^t)$. For example, $\widehat{\lambda}^t$ could be a random variable taking its values in some vector space or a cone which could be finite-dimensional (typically $\mathbb{R}^n$ or $\mathbb{R}^n_+$ for suitable n) or infinite-dimensional (for example, some functional spaces and respective cones, e.g., $L^2(\Omega)$, $H^s_0(\Omega)$, $s>0$, where $\Omega \subset \mathbb{R}^3$ is the imaging domain). In statistics this is known as frequentist approach, and for ET it often leads to estimation of confidence intervals for *the maximum likelihood estimate* or for *maximum penalized log-likelihood* estimate (both are M-estimators [VdV00]); see e.g., [Fes96]. In particular, frequentist approach has an advantage of being relatively robust to *model misspecification* (i.e., when $P^t \neq P^t_{A,\lambda}$ for any A and λ). In this case for large t estimate $\widehat{\lambda}^t$ will tend to a projection of P^t onto family $P^t_{A,\lambda}$ with respect to some chosen distance between probability distributions (e.g., for Kullback-Liebler diver-

gence). Under additional assumptions on P^t even in misspecified case it is still possible to establish asymptotic distribution of $\widehat{\lambda}^t$ (e.g., asymptotic normality), from which, for example, asymptotic confidence intervals can be retrieved; see e.g., [Whi82]. However, use of asymptotic results for practice in ET seems questionable since very little data are available in a single scan.

Bayesian approach is also used for uncertainty quantification in ET. In this case all prior information (e.g., anatomical information from side images, assumptions on support and smoothness) is encoded in some prior measure $\pi_{\mathcal{M}}(\lambda)$ which is used together with family $P^t_{A,\lambda}$ to define posterior distribution via the well-known Bayes formula; see e.g., [BG14]. Sampling from such posteriors is done via MCMC techniques [Wei97], [HBJ+97], [FL07], [FBD+18]. Common bottlenecks here are the following ones: complicated design of the algorithm and its implementation, high numerical load per iteration, lack of scalability and most importantly – poor mixing in constructed Markov chains; see e.g., [DJD18].

Bootstrap is another attractive technique to assess uncertainty which can be also seen as some probabilistic sensitivity analysis in optimization or as approximate sampling from some bayesian posteriors. Nontrivial questions here are the following ones: (1) how to define a bootstrap procedure for Poisson-type raw data in ET and (2) make sure that asymptotic bootstrap intervals coincide with asymptotic bayesian ones. Very common approach to answer question (1) is to use resampling in list-mode data; see e.g., [HW89], [Dah01]. Such approach targets to resample raw data and then propagate the uncertainty in reconstructions for any reconstruction algorithm being used (e.g., FBP, MLEM or MAP). Our approach is somehow similar to bootstrap as it will be explained further.

In view of the above discussion, it is important to note that in ET practice it seems that it is not of great importance which kind of uncertainty model is used – frequentist, bayesian or bootstrap. In the end, most important is to make usable resulting context and algorithms by practitioners.

Being inspired with nonparametric posterior learning proposed in [LWH18], [FLH19], we propose sampling algorithms for ET with and without MRI anatomical data at hand. Main idea of the method is that uncertainty propagates from not knowing true generating distribution P^t and it is encoded in a nonparametric prior on P^t which is permanently updated to nonparametric posterior when observed data Y^t is accessible. Randomness from the nonparametric posterior propagates to reconstructions which are sampled. In particular, for the case of ET without multimodal data our sampling algorithm corresponds to weighting likelihood bootstrap (WLB) from [NR94] being adapted to poisson-type data. With access to multimodal data, in a similar way with [LWH18], we construct a nonparametric prior in form of a mixture of weighted gamma processes (MGP) assuming that P^t is a temporal stationary Poisson point process, so the nonparametric posterior appears also to be an MGP. Important point of our approach is that uncertainty is postulated not in the image space (i.e., in space of reconstructed images), but in observation space which are the measured photon rates along various lines of response. This is done, in particular, using anatomical MRI images. As a byproduct we find that our main calibration parameter θ^t has very simple interpretation which corresponds exactly to the ratio between amount of information in the sinogram and in MRI anatomical images.

All our algorithms are trivially parallelizable, scalable and very easy to implement because they rely on well-known EM-type reconstruction methods from [SV82], [FH95].

Finally, we study theoretically the asymptotic properties of our algorithms when large dataset is available (for ET this is equivalent to $t \to +\infty$). We show that the distribution of produced samples is asymptotically consistent at the "true" intensity map λ_* in the span of the columns of design matrix A if $\theta^t = o(t)$. We also study the asymptotic conditional distribution of produced samples and find two interesting phenomena here. First, the distribution concentrates not at the true point λ_* but around its strongly

consistent estimator $\widehat{\lambda}^t_{sc}$ (modulo $\ker A$) (e.g., $\widehat{\lambda}^t_{sc}$ being the MLE or penalized MLE). Second, because of nonregularity of the statistical model ($\lambda_* \in \partial\mathbb{R}^p_+$, in general) and lack of known results on strongly consistent estimators for the model of ET we fail to obtain a Bernstein von-Mises type theorem for almost any data trajectory Y^t, $t \in (0, +\infty)$. The first phenomenon was already observed in LASSO regression via weighted bootstrap in [NN20] and the second one is completely new to our knowledge.

It appears that using MRI images in the prior on the observation space for ET naturally leads to consideration of misspecification in generalized Poisson models with wrong design. In particular, a breakdown of a certain geometrical condition when using MRI images results in loss of identifiability in models used for the prior, so we say that prior becomes *non-informative* in this case. This condition has geometrical interpretation in terms Radon-type transforms which are commonly used to model the design matrix for ET.

Initially, the main motivation for this work was the problem of poor mixing for the Gibbs-type sampler in [FBD+18] which was designed for PET-MRI context. In this work we give a detailed analysis of this phenomenon and give empirical advice on design of MCMC-samplers for ill-posed inverse problems such as PET or SPECT.

This paper is organized as follows. In Section 2 we give notations and all necessary preliminaries on statistical models of ET and on use of multimodal data. In Section 3 we adapt nonparametric posterior learning for ET context and derive our sampling algorithms. In Section 4 we give a very informative example for the problem of poor mixing for MCMC in ET. In Section 5 we discuss implementations and show numerical tests of our algorithms. In Section 6 we study theoretically the asymptotic properties of our algorithms. In Section 7 we discuss our results and possibilities for future work.

2 Preliminaries

Generic notations. By $\mathbb{N}_0$ we denote the set of non-negative all integers, $\mathbb{R}^n_+$ denotes the nonnegative cone of euclidean space $\mathbb{R}^n$, by $x \succeq y$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$, we denote the property that $x_j \geq y_j$ for all $j = 1, \dots, n$, $x \succ y$ denotes the same but with strict inequalities, $\langle x, y\rangle$ stands for the scalar product $x^T y$ (we will use both notations), $R_+(A)$ denotes the image of positive cone $\mathbb{R}^p_+$ under action of operator $A \in \mathrm{Mat}(d, p)$, by $X \sim F$ we denote the property that random variable X has distribution F, $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with intensity λ, $\lambda \geq 0$, by $\Gamma(\alpha, \beta)$ we denote the gamma distribution with shape parameter α, and scale β ($\xi \sim \Gamma(\alpha, \beta)$, $\mathbb{E}\xi = \alpha\beta^{-1}$, $\mathrm{Var}(\xi) = \alpha\beta^{-2}$), $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in S^n_+$, where the latter denotes the cone of positive definite matrices of size $n \times n$. Let $A \in \mathrm{Mat}(d, p)$, $I \subset \{1, \dots, d\}$, then A_I denotes the submatrix of A with rows indexed by elements in I, $\mathrm{Span}(A^T)$ denotes the span of the rows of A being considered as vectors in $\mathbb{R}^p$. Let Z be a complete separable metric space equipped with metric $\rho_Z(\cdot, \cdot)$ and boundedly finite nonnegative measure dz, $B(Z)$ denotes the sigma algebra of borel sets in Z. By $\mathcal{PP}^t$ we denote a point process on Z defined for each $t \in \mathbb{R}_+$ and $\mathcal{PP}^t_\Lambda$ denotes the Poisson point process on Z with intensity $t\Lambda$, where Λ is the nonnegative function $\Lambda = \Lambda(z)$, $z \in Z$, Λ is integrable with respect to dz. Weighted gamma process on Z is denoted by $GP(\alpha, \beta) = G_{\alpha,\beta}$, where α is the shape measure on Z and β is the scale which is a non-negative function Z and also α-integrable; see, e.g., [Lo82]. Finally, by $\mathcal{KL}(P, Q)$ we denote the Kullback-Leibler divergence between probability distributions P, Q.

Mathematical model for ET. Raw measured data in ET are described by vector $Y^t = (Y^t_1, \dots, Y^t_d) \in (\mathbb{N}_0)^d$ which stands for the photon counts recorded during exposure

time t along lines of responses (LORs) $\{1,\dots,d\}$. It is assumed that

$$\begin{aligned} &Y_i^t \sim \mathrm{Po}(t\Lambda_i),\ \Lambda_i = a_i^T\lambda, \\ &Y_i^t \text{ are mutually independent for } i \in \{1,\dots,d\}, \end{aligned} \tag{2.1}$$

where $\lambda \in \mathbb{R}_+^p$ is the parameter of interest on which we aim to perform inference. In practice, vector λ denotes the spatial emission concentration of the isotope (or tracer uptake) measured in [Bq/mm^3], that is λ_j is the concentration at pixel $j \in \{1,\dots,p\}$.

Vector $\Lambda = (\Lambda_1,\dots,\Lambda_d)$ denotes the observed photon intensities along LORs $\{1,\dots,d\}$, respectively. To separate the LORs with strictly positive intensities from those ones with zeros we use following notations:

$$I_0(\Lambda) = \{i : \Lambda_i = 0\},\ I_1(\Lambda) = \{i : \Lambda_i > 0\},\ I_0 \sqcup I_1 = \{1,\dots,d\}. \tag{2.2}$$

Collection of $a_i \in \mathbb{R}^p$ in (2.1) constitute matrix $A = [a_1^T,\dots,a_d^T]^T$, $A \in \mathrm{Mat}(d,p)$ which is called by projector or system matrix in applied literature on ET and by design or design matrix in statistical literature. Each element a_{ij} in A denotes the probability to observe a pair of photons along LOR $i \in \{1,\dots,d\}$ if both they were emitted from pixel $j \in \{1,\dots,p\}$. In view of such physical interpretation, for design A we assume the following:

$$a_{ij} \geq 0 \text{ for all pairs } (i,j), \tag{2.3}$$

$$A_j = \sum_{i=1}^{d} a_{ij},\ 0 < A_j \leq 1 \text{ for all } j \in \{1,\dots,p\}, \tag{2.4}$$

$$\sum_{j=1}^{p} a_{ij} > 0 \text{ for all } i \in \{1,\dots,d\}. \tag{2.5}$$

If any of formulas (2.4), (2.5) would not be satisfied, then, in practice it would mean that either some pixel is not detectable at all (hence it can be completely removed from the model) or some detector pair is broken and cannot detect any of incoming photons. These scenarios are outside of our scope.

It is well-known that the inverse problems for PET and SPECT are mildly ill-posed (see e.g., [HW16]), so we also assume that

$$\ker A \neq \{0\}. \tag{2.6}$$

Remark 2.1. Numerically A usually represents a discretized version of weighted Radon transform operator R_a for ET with complete data (see e.g., [Nat01], [Gon19]). Since A approximates R_a in strong operator norm (as a discretization of R_a on the uniform grid for full angle acquisition) we can conclude that

$$\sigma_k \asymp k^{-1/2},\ k = 1,\dots,p, \tag{2.7}$$

where σ_k are the singular values of A. In particular, even if matrix A is injective, from numerical point of view it may be not, especially for large dimension p. For p large enough, due to (2.7) it may happen that $\mathrm{cond}(A) > \varepsilon_F^{-1}$, where $\mathrm{cond}(\cdot)$ denotes the condition number, ε_F is the floating-point precision. In the latter case from numerical point of view smallest singular values of A numerically are equivalent to machine zeros which means exactly the existence of a nontrivial kernel for A. □

Likelihood and negative log-likelihood functions for model in (2.1) are given by the formulas:

$$P_{A,\lambda}^t(Y^t) = P(Y^t|A,\lambda,t) = \prod_{i=1}^{d} \frac{(ta_i^T\lambda)^{Y_i^t}}{Y_i^t!} e^{-ta_i^T\lambda},\ \lambda \in \mathbb{R}_+^p,\ t \geq 0, \tag{2.8}$$

$$L(\lambda|Y^t, A, t) = \sum_{i=1}^{d} -Y_i^t \log(t\Lambda_i) + t\Lambda_i,\ \Lambda_i = a_i^T\lambda. \tag{2.9}$$

Note that for A satisfying (2.6) and for any Y^t function $L(Y^t|\lambda, t)$ is not strictly convex even at the point of the global minima since $L(Y^t|\lambda + u,\ t) = L(Y^t|\lambda,\ t)$ for any $\lambda \in \mathbb{R}^p_+$ and $u \in \ker A$. To avoid numerical instabilities due to this phenomenon a convex penalty $\varphi(\lambda)$ is added to $L(Y^t|A, \lambda, t)$, so we also consider the penalized negative log-likelihood:

$$L_p(\lambda|Y^t, A, t, \beta^t) = L(\lambda|Y^t, A, t) + \beta^t\varphi(\lambda),\ \lambda \in \mathbb{R}^p_+, \tag{2.10}$$

where $\beta^t \geq 0$ is the regularization coefficient. Here we assume that β^t may increase with time t at a certain rate, which is important for practice in order to increase the signal-to-noise ratio in reconstructed images.

Convexity of the log-likelihood. Ill-posedness nature of the inverse problem and nontrivial domain for the parameter of interest ($\lambda \in \mathbb{R}^p_+$) require careful review of convexity properties of $L(\lambda|Y^t, A, t)$.

The Hessian for $L(\lambda|Y^t, A, t)$ is given by the formula:

$$\nabla^2_\lambda L(\lambda|Y^t, A, t) = \sum_{i=1}^{d} Y_i^t \frac{a_i a_i^T}{\Lambda_i^2} = A^T D_{\lambda,t} A,\ D_{\lambda,t} = \mathrm{diag}(\dots, Y_i^t/\Lambda_i^2, \dots), \tag{2.11}$$

where $\Lambda_i = a_i^T\lambda$. From (2.11) one can see that $L(\lambda|Y^t, A, t)$ is strongly convex only in directions spanned by vectors a_i, where $Y_i^t > 0$. In case of $Y_i^t = 0$ for some i, function $L(\lambda|Y^t, A, t)$ remains convex in direction spanned by a_i due to linear terms $t\Lambda_i$ in (2.9). Another important conclusion from (2.9), (2.11) is the following one:

$$L(Y^t|A, \lambda, t) \text{ is not strictly convex on } \mathbb{R}^p_+, \text{ in general, even if } A \text{ is injective.} \tag{2.12}$$

Indeed, consider vector $u \in \mathbb{R}^p$ defined by the formulas below

$$a_i^T u = 0 \text{ for all } i \text{ s.t. } Y_i^t > 0,\ \sum_{i:Y_i^t=0} a_i^T u = 0,\ u \neq 0. \tag{2.13}$$

Such u exists, for example, if $\#\{i : Y_i > 0\} < p - 1$, which is realistic in ET practice, for example, when tracer distribution has small spatial support[1]. It is left to note that for u from (2.13) the following property holds:

$$L(\lambda + u|Y^t, A, t) = L(\lambda|Y^t, A, t) \text{ for any } \lambda \in \mathbb{R}^p_+. \tag{2.14}$$

Example in (2.12)-(2.14) is a remainder that positivity constraints in (2.1) are crucial. Though $L(Y^t|A, \lambda, t)$ is not strictly convex in general, for $Y^t \in R_+(A)$ (which is approximately satisfied for large t) it is locally strongly convex in directions of $\mathrm{Span}(A^T)$ at its global minumum on $\mathbb{R}^p_+$. We show this, in particular, in our results on consistency in Section 6.

[1]It is common in ET practice that for many LORs, the registered data are $Y_i^t = 0$. This is mostly due to small acquisition time t and weak sensitivity of the scanner.

Regularization penalty. The role of regularization penalty $\varphi(\lambda)$ in formula (2.10) is to decrease the numerical instability of the underlying inverse problem and to make function $L_p(\lambda|Y^t, A, t, \beta^t)$ more convex, especially in directions close to $\ker A$. In view of the arguments from the previous paragraph we assume that

$$\varphi \text{ is continous and convex on } \mathbb{R}^p, \tag{2.15}$$

$$g_u(w) = \varphi(u + w) \text{ is strictly convex in } w \in \ker A \text{ for any } u \in \mathrm{Span}(A^T). \tag{2.16}$$

Lemma 2.1. *Let $\varphi(\lambda)$ be the function of (2.15), (2.16), A satisfies conditions in (2.3)-(2.5). Let $\lambda \in \mathbb{R}^p_+$ and $U \subset \mathrm{Span}(A^T)$ be a compact such that*

$$\{w : \lambda + u + w \succeq 0,\ w \in \ker A\} \text{ is non-empty for any } u \in U. \tag{2.17}$$

Then, mapping defined by the formula

$$w_{A,\lambda}(u) = \underset{\substack{w:\lambda+u+w\succeq 0,\\ w\in\ker A}}{\arg\min} \varphi(\lambda + u + w),\ u \in U \tag{2.18}$$

is one-to-one. Moreover, $w_{A,\lambda}(u)$ is continuous on U.

In particular, for our numerical tests in Section 5 we choose the well-known in PET imaging log cosh penalty [Gre90] coupled with ℓ_2 convex pairwise difference penalty:

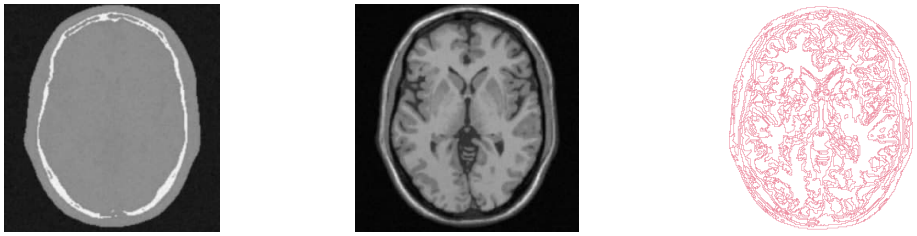
$$\varphi(\lambda) = \sum_{j=1}^{p} \sum_{j' \in \mathcal{N}_j} w_{jj'} \left((1-\nu)\zeta \cdot \log\cosh\left(\frac{\lambda_j - \lambda_{j'}}{\zeta}\right) + \frac{\nu}{2}\left(\lambda_j - \lambda_{j'}\right)^2 \right), \tag{2.19}$$

where $w_{jj'} > 0$, $w_{j'j} = w_{jj'}$ and $\mathcal{N}_j$ the neighborhood of pixel j. In practice, on a square image we consider a 8-adjacent pixels neighborhood with $w_{jj'} = 1$ for horizontal/vertical neighbors and $w_{jj'} = \frac{\sqrt{2}}{2}$ for diagonal ones.

Parameter ζ is chosen to be fixed. Penalty of form (2.19) is attractive since it bridges together Gaussian prior for pairwise interactions ($\zeta \to +\infty$), and for $\zeta = 0$, it corresponds to ℓ^1-penalty on pairwise interactions (Laplace prior). It is easy to check that $\varphi(\lambda)$ in (2.19) is strictly convex except the only direction given by vector $e = \{c(1, \dots, 1),\ c \in \mathbb{R}\}$. From formula (2.5) it follows that $e \notin \ker A$, therefore conditions (2.15), (2.16) are automatically satisfied.

Multimodal data for emission tomography. From the previous paragraph one can see that recorded signal Y^t is essentially the Poisson noise for which its signal-to-noise ratio (SNR) is proportional to $\sqrt{t \cdot \Lambda}$ and is quite low in practice (e.g., because of low injected dose and moderate t in standard medical protocols). In order to increase the SNR in reconstructed images and not to loose a lot in resolution it is proposed to regularize the inverse problem using multimodal data (e.g., anatomical images) from CT or MRI. Between CT and MRI we choose MRI since it provides anatomical information with high contrast in soft tissues in comparison to CT (see Figures 1 (a), (b)).

In this work our exterior data consists of r anatomical (presegmented) MRI images $\mathcal{M} = \{M_1, \dots, M_r\}$ (see Figure 1 (c)). Segmentations of MRI images are precomputed using the ddCRP algorithm from [BF11]. In particular, MRI-guided reconstructions in PET is an active topic of research (see the discussion in [FDC+21] and also references therein) and still a lot of work is needed to describe precisely correlations between ET and MRI signals (especially from biological point of view); see e.g., [BYH+04]. Because of this the current use of MRI data is more image-based: spatially regularizing penalties are constructed using MRI data in [BJT+96], [BYH+04], [VAB+11], model built upon MRI-segmented data for locally-constant tracer distribution are used in [FBD+18] and also in



(a) CT (b) MRI (c) $M \in \mathcal{M}$

Figure 1: Multimodal data for ET of the brain

our work. However, our approach is ideologically different from the one in [VAB+11] because we use $\mathcal{M}$ to construct models of tracer distribution and then sample "pseudo-data" to mix it with real observed data Y^t. That is MRI data are used only in observation space for ET (space of intensities along LORs). This has a practical advantage of interpretability for our main calibration parameter which reflects the ratio between number of real detected photons $N^t = \sum Y^t$ and the number of "pseudo-photons".

3 Posterior learning for emission tomographies

Weighted likelihood/bayesian bootstrap (WLB/WBB) was originally proposed in [NR94], [NPX21] and recently generalized/connected to nonparametric posterior learning originating from [LWH18], [FLH19]. In particular, in latter works it was assumed that observed data are i.i.d. which does not fit directly the observations in emission tomography, where the raw data are described by realizations from spatio-temporal (temporal stationary) Poisson point process; see e.g., [HW16]. In this work we adapt the idea behind nonparametric posterior learning to the case of Poisson-type data and construct bootstrap sampling algorithms to assess the uncertainty.

First, in Subsection 3.1 we explain our adaptation using fully nonparametric model for emission tomographies. In Subsection 3.3, by binning the observations we derive posterior boostrap algorithms for the finite-dimensional model in (2.1). We present two versions of sampling algorithms: in the first one no MRI-data are used and the algorithm is ideologically closer to bootstrap algorithms in [NR94], [NPX21], and in the second one we use MRI images to construct nonparametric priors on spatial temporal point processes, so the algorithm is closer to ones in [LWH18], [FLH19].

3.1 Nonparametric case

In what follows we consider spatio-temporel (Poisson) point processes, in particular, in the context of nonparametric statistical inverse problems. For brevity we do not define precisely all functional spaces and operators involved but explain only the necessary concepts to define nonparametric posterior learning for emission tomographies. Precise definitions and theoretical statements on this topic can be found, for example, in [DVJ05], [DVJ07], [HW16], [JS13]. In the end, only finite-dimensional versions of presented algorithms are used in practice.

Nonparametric model for emission tomographies. Nonparametric statistical framework for emission tomographies can be seen as a classical scanning scenario with a machine having infinite number of infinitely small detectors. Let Z be the manifold of all detector positions available in the acquisition geometry of a scanner, e.g., $Z = \mathbb{R} \times \mathbb{S}^1/\mathbb{Z}_2$ (all non-oriented straight lines in $\mathbb{R}^2$) for full angle acquisition in a single plane slice. For

completeness we also assume that Z is equipped with a boundedly-finite measure dz (which reflects the sensitivity of detectors for various lines) and with a metric $\rho_Z(\cdot)$ describing distances between the lines, e.g., $\rho_Z(\cdot)$ could be a geodesic distance on cylinder $\mathbb{R}\times\mathbb{S}^1/\mathbb{Z}_2$.

For exposure time t the raw data are given by a random measure Z^t generated by a counting point process:

$$Z^t = \sum_{j=1}^{N^t} \delta_{(t_j,z_j)},\ (t_j,z_j)\in\mathbb{R}_+\times Z,\ t_j\le t_{j+1},\ t_j\le t, \tag{3.1}$$

where

$$N^t \text{ is total number of registered photons,} \tag{3.2}$$

$$\{z_j\}_{j=1}^{N^t},\ \{t_j\}_{j=1}^{N^t} \text{ are the LORs and times of photon registrations, respectively.} \tag{3.3}$$

In practical literature on PET/SPECT sample Z^t is known as *list-mode* data, whereas Y^t from (2.1) is called by *sinogram* which is the version of Z^t binned to finite resolution and without information on $\{t_j\}_{j=1}^{N^t}$. From statistical point of view, under the assumption of temporal stationarity Z^t and Y^t contain the same amount of information since Y^t is a sufficient statistic.

For statistical model of Z^t it is natural to take the family of temporal stationary Poisson point processes $\mathcal{PP}^t_{A\lambda}$ on Z, where A, λ stand for the nonparametric versions of the projector and vector denoting the tracer concentration, respectively; see Section 2. For example, in such model the intensity of the poisson flow of photons in LOR $z\in Z$ is given by $\Lambda(z)dz = [A\lambda](z)dz$.

The negative log-likelihood for $\mathcal{PP}^t_{A\lambda}$ and observation Z^t is defined via the following formula (see, e.g., [HW16], Section 2):

$$\begin{aligned} L(Z^t|A,\lambda,t) &= -\sum_{j=1}^{N^t}\log(\Lambda(z_j)) + t\int_Z \Lambda(z)\,dz \\ &= -\int_Z \log(\Lambda) dZ^t + t\int_Z \Lambda(z)\,dz,\ \Lambda = A\lambda. \end{aligned} \tag{3.4}$$

Misspecification and the KL-projection. In reality our model assumption on distribution of Z^t is always incorrect, that is model $\mathcal{PP}^t_{A\lambda}$ is misspecified and $Z^t\sim\mathcal{PP}^t$ for some point process $\mathcal{PP}^t$ for which $\mathcal{PP}^t\neq\mathcal{PP}^t_{A\lambda}$. Since (penalized) maximum log-likelihood estimates are the most popular in ET practice, the best one can hope to reconstruct using model $\mathcal{PP}^t_{A\lambda}$ is the projection of $\mathcal{PP}^t$ onto $\mathcal{PP}^t_{A\lambda}$ in the sense of Kullback-Leibler divergence:

$$\lambda^t_*(\mathcal{PP}^t) = \arg\min_{\lambda\in\mathfrak{L}} \mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}), \tag{3.5}$$

where $\mathfrak{L}$ is some set of admissible solutions for λ (e.g., restrictions on smoothness and support of the tracer). Recall that A is ill-conditioned (see formula (2.6)), so the same property is preserved also in the nonparametric scenario and, in general, λ^t_* in (3.5) may not be defined uniquely even for very natural choices of $\mathfrak{L}$. For this reason we consider the penalized KL-projection defined by the formula:

$$\lambda^t_*(\mathcal{PP}^t) = \arg\min_{\lambda\in\mathfrak{L}} [\mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}) + \beta^t\varphi(\lambda)], \tag{3.6}$$

where β^t is the regularization coefficient and $\varphi(\lambda)$ is a nonparametric extension of penalty from Section 2.

From formula (3.4) and the definition of Kullback-Leibler divergence it follows that

$$\mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}) = -\int_Z \log(\Lambda) \cdot \mathbb{E}_{\mathcal{PP}^t}[dZ^t] + t\int_Z \Lambda(z)dz, \tag{3.7}$$

where $\mathbb{E}_{\mathcal{PP}^t}$ is the expectation with respect to $\mathcal{PP}^t$. Putting together (3.6), (3.7), for the penalized KL-projection we get the following formula:

$$\lambda^t_* = \arg\min_{\lambda\in\mathfrak{L}} \mathbb{L}_p(\lambda|\mathcal{PP}^t,\, A,\, t,\, \beta^t), \text{ where} \tag{3.8}$$

$$\begin{aligned} \mathbb{L}_p(\lambda|\mathcal{PP}^t,\, A,\, t,\, \beta^t) &= -\int_Z \log(\Lambda) \cdot \mathbb{E}_{\mathcal{PP}^t}[dZ^t] + t\int_Z \Lambda(z)dz + \beta^t\varphi(\lambda), \\ \Lambda(z) &= A\lambda(z). \end{aligned} \tag{3.9}$$

Propagation of uncertainty and nonparametric posterior learning. Following the idea from [LWH18], we say that uncertainty on λ propagates from the uncertainty on generating process $\mathcal{PP}^t$ in (3.8), (3.9). Let $\pi(\cdot)$ be a probability measure in which we encode our prior beliefs over a set of possible generating processes, that is $\pi(\cdot)$ is a nonparametric prior on spatio-temporal point processes on $\mathbb{R}_+ \times Z$. In our case prior $\pi(\cdot)$ is constructed using multimodal data $\mathcal{M}$, so for the prior we will write $\pi_{\mathcal{M}}(\cdot)$ instead of $\pi(\cdot)$. Let Z^t be the list-mode data observed in a real experiment, then our prior beliefs on $\mathcal{PP}^t$ can be updated in form of posterior distribution $\pi_{\mathcal{M}}(\cdot|Z^t, t)$. In this case the definition of nonparametric posterior learning for ET with multimodal data is straightforward as shown in Algorithm 1.

Algorithm 1: NPL for ET with multimodal data

Data: list-mode data $Z_t = \sum_{j=1}^{N^t} \delta_{(t_j,z_j)}$, multimodal data $\mathcal{M}$

Input: B – number of samples,
projector A, regularization parameter β^t, penalty $\varphi(\lambda)$

for $b = 1$ **to** B **do**
  Draw point process $\widetilde{\mathcal{PP}}^t \sim \pi_{\mathcal{M}}(\cdot|Z^t, t)$;
  Compute $\widetilde{\lambda}^t_b = \arg\min_{\lambda\in\mathfrak{L}} \mathbb{L}_p(\lambda|\widetilde{\mathcal{PP}}^t, A, t, \beta^t)$ for $\mathbb{L}_p(\cdot)$ defined in (3.9);
end

Output: $\{\widetilde{\lambda}^t_b\}_{b=1}^B$

As it has already been outlined in [LWH18], [FLH19], the above scheme is trivially parallelizable which is a strong advantage in front of MCMC sampling from pure bayesian posteriors. In particular, in Section 4 we show on a very practical example that using MCMC for posterior sampling in ETs can be very challenging already on the level of the sampler design.

Construction of prior $\pi_{\mathcal{M}}(\cdot)$ and of posterior $\pi_{\mathcal{M}}(\cdot|Z^t, t)$. For each t a sample from $\mathcal{PP}^t$ is a purely atomic random measure Z^t on Z which stands for photon registration events along various lines of response durig time interval $(0, t)$. Due to mutual independence of emission events inside the patient, for random measure Z^t it is natural to assume that

for all finite families of mutually disjoint, bounded Borel sets $\{B_1, \dots, B_N\}$,

$$\text{where } B_i \in B(Z), \quad \text{random variables } Z^t(B_i) = \int_{B_i} dZ^t,\ i \in \{1, \dots, N\} \text{ are} \tag{3.10}$$

mutually independent.

Measure Z^t which satisfies property (3.10) is known as *completely random measure*; see [DVJ07], Chapter 10. In particular, under the additional and intuitive assumption that Z^t contains no fixed atoms (i.e., Z^t is purely atomic but locations of registrations may differ from sample to sample) the representation theorem of Kingman (1967) says that $\mathcal{PP}^t$ is characterized uniquely by a Poisson point process with some intensity measure μ^t on $Z \times (0, +\infty)$; see [DVJ07], Section 10.1, Theorem 10.1.III. Therefore, any prior on $\mathcal{PP}^t$ must be a prior on μ^t.

In our case we make an assumption that

$$\begin{gathered}\mathcal{PP}^t = \mathcal{PP}^t_\Lambda, \text{ for some intensity measure } \Lambda \text{ on } Z, \text{ that is} \\ Z^t \sim \mathcal{PP}^t, \text{ i.e., } Z^t(B) \sim \text{Po}(t \cdot \Lambda(B)),\ \Lambda(B) = \int_B \Lambda(z)\, dz, \text{ for any } B \in B(Z)\end{gathered} \tag{3.11}$$

and then we construct a prior on Λ using $\mathcal{M}$.

Remark 3.1. In fact, the choice in (3.11) becomes necessary if process $\mathcal{PP}^t$, $t \in (0, +\infty)$ is temporal stationary, property (3.10) holds and Z^t is *orderly* on Z for each t with respect to metric ρ_Z (see [DVJ05], Section 2.4, Theorem 2.4.V). For construction of more general priors on $\mathcal{PP}^t$ the theory of completely random measures should be used as we noted above. □

For the prior on Λ we use the mixture of gamma processes (further denoted by MGP) which can be written as follows:

$$\Lambda_{\mathcal{M}} \sim P_{\mathcal{M}}(\cdot),\ \Lambda \sim GP(\theta^t \Lambda_{\mathcal{M}}, (\theta^t)^{-1} \cdot \mathbb{1}), \tag{3.12}$$

where

$$\theta^t \text{ is a positive scalar, } \mathbb{1} \text{ is the identity function on } Z, \tag{3.13}$$

$$\Lambda_{\mathcal{M}} \text{ is the mixing parameter, } P_{\mathcal{M}}(\cdot) \text{ is the mixing distribution (hyperprior),} \tag{3.14}$$

$$GP(\alpha, \beta) = G_{\alpha,\beta} \text{ is the weighted gamma process on } Z \text{ with shape } \alpha \text{ and scale } \beta. \tag{3.15}$$

In short, we will use the following notation

$$\mathcal{PP}^t \sim \pi_{\mathcal{M}}(\cdot) = \text{MGP}(P_{\mathcal{M}}, t, \theta^t \Lambda_{\mathcal{M}}, (\theta^t)^{-1}). \tag{3.16}$$

Note that the scale parameter in gamma process in (3.12) is constant for all Z and is equal to $(\theta^t)^{-1}$. Such choice allows to center the gamma process on $\Lambda_{\mathcal{M}}$ in (3.12), so θ^t controls only the spread around $\Lambda_{\mathcal{M}}$ (e.g., choice $\theta^t = 0$ corresponds to improper uniform distribution on Z, $\theta^t = +\infty$ corresponds to prior $\mathcal{PP}^t = \mathcal{PP}^t_{\Lambda_{\mathcal{M}}}$, where $\Lambda_{\mathcal{M}} \sim P_{\mathcal{M}}(\cdot)$).

The key to define posterior for MGP in (3.16) is the following theorem which is an adaptation of Theorem 3.1 from [Lo82].

Theorem 3.1. *Let $Z^t \sim \mathcal{PP}^t_\Lambda$ and $G_{\alpha,\beta}$ be the prior on Λ. Then, the posterior distribution of Λ is given by weighted gamma process $G_{\alpha+Z^t, \frac{\beta}{1+t\beta}}$.*

From the result of Theorem 3.1 it follows that posterior for MGP in (3.16) is also an MGP:

$$\widetilde{\mathcal{P}\mathcal{P}}^t \sim \pi_{\mathcal{M}}(\cdot|Z^t,t) = \mathrm{MGP}(P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Z^t,t), Z^t + \theta^t\Lambda^t_{\mathcal{M}}, (\theta^t+t)^{-1}), \tag{3.17}$$

where $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Z^t,t)$ is posterior for the mixing parameter. From (3.12)-(3.17) one can see that $\Lambda^t_{\mathcal{M}}$ plays the role of an intensity map for LORs in Z. Therefore, mixing takes place in the observation space which reflects the fact that we do not rely completely on particular design A (which is always erroneous in practice).

Detailed constructions of $P_{\mathcal{M}}(\cdot)$, $P_{\mathcal{M}}(\cdot|Z^t,t)$ are given in Subsection 3.4, where finite-dimensional sampling algorithms are considered.

Remark 3.2. MGP prior in (3.16) and the posterior in (3.17) are direct analogs of MDP prior and posterior from [LWH18], respectively. Weighted gamma processes as priors were considered in [Jam03] for various semiparametric intensity models including very elaborate Poisson model for PET (temporal non-stationarity, detector transition kernels). In particular, in [Jam03] a weighted gamma prior was used not in observation space but in the image space (i.e., as a prior on λ) and sampling from posteriors was based on data augmentation schemes similar to the one in Section 4 below. In this work we show that such data augmentation for ill-posed problem of PET leads to very serious mixing problems for MCMC-samplers involved. In addition, in Section 6.4 we show that a badly chosen (even injective) design can lead to loss of idenitifiability for λ in models of type (2.1). In our approach most of complexity is moved to construction of a "good" prior in observation space which should be initially already centered at the true (KL-optimal) intensity map built from MRI data (see also formula (3.12)). □

3.2 Binning to parametric models and algorithms

Each detector has a screen of finite size which detects incoming photons from a family of lines in Z. Let the machine detect photons along d lines (channels). Mathematically it means that Z can be represented as follows:

$$Z = \left(\bigsqcup_{i=1}^{d} Z_i\right) \bigsqcup \overline{Z},\ \{Z_i\}_{i=1}^d \text{ be a disjoint family of sets from } B(Z). \tag{3.18}$$

Each set Z_i corresponds to set of lines which are visible in channel i, $\overline{Z}$ are the lines which are not visible in any channel. For each i we can define binning by the formula (see also [BWP97]):

$$\left(\int_{Z_i} dZ^t, \int_{Z_i} \Lambda(z)\,dz\right) = (Y^t_i, \Lambda_i). \tag{3.19}$$

From (3.11), (3.19) it follows that

$$Y^t_i \text{ are mutually independent and } Y^t_i \sim \mathrm{Po}(t\cdot\Lambda_i), i \in \{1,\dots,d\}. \tag{3.20}$$

Note that data list-mode data Z^t are binned to Y^t which are exactly the sinogram data for model (2.1). Let Y^t be the binning of list-mode data Z^t from the real experiment. Non-parametric weighted gamma prior and its posterior in (3.16), (3.17), penalized negative log-likelihood in (3.9) are also binned in a similar way with (3.19), so a finite-dimensional version of Algorithm 1 can be written as follows.

Algorithm 2: Binned NPL for ET with multimodal data

Data: sinogram data $Y_t = (Y_1^t, \dots, Y_d^t)$, multimodal data $\mathcal{M}$

Input: B – number of samples, parameter θ^t,
projector A, regularization parameter β^t, penalty $\varphi(\lambda)$

1 **for** $b = 1$ **to** B **do**

2 Draw $\Lambda_{\mathcal{M}}^t = (\Lambda_{\mathcal{M},1}^t, \dots, \Lambda_{\mathcal{M},d}^t)$ from $P_{\mathcal{M}}(\Lambda_{\mathcal{M}}^t | Y^t, t)$;

3 Draw $\widetilde{\Lambda}_{b,i}^t \sim \Gamma(Y_i^t + \theta^t \Lambda_{\mathcal{M},i}^t, (\theta^t + t)^{-1})$ independently for each i;

4 Compute $\widetilde{\lambda}_b^t = \underset{\lambda \succeq 0}{\arg\min}\, L_p(\lambda | \, t\widetilde{\Lambda}_b^t, A, t, \beta^t)$ for $L_p(\cdot)$ defined in (2.10);

5 **end**

Output: $\{\widetilde{\lambda}_b^t\}_{b=1}^B$

Remark 3.3. In steps 2, 3 intensities $\widetilde{\Lambda}_{b,i}^t$ are sampled from the binned MGP posterior in (3.17). In step 4 we have used the fact that binned version of $\mathbb{L}_p(\cdot)$ from (3.9) coincides with $L_p(\cdot)$ from (2.10). In addition, from formula (2.10) it follows that

$$L_p(\lambda | t\widetilde{\Lambda}_b^t, A, t, \beta^t) = t \cdot L_p(\lambda | \widetilde{\Lambda}_b^t, A, 1, \beta^t/t) + R, \tag{3.21}$$

where R is a function which is independent of λ. Therefore, minimization problem in step 4 can be directly applied to normalized functional $L_p(\lambda | \widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$. □

Though $P_{\mathcal{M}}(\cdot | Y^t, t)$ we have not yet defined (see Subsection 3.4) one can already see that Algorithm 2 is trivially parallelizable. It is also important to have a numerically efficient and scalable optimization scheme in step 4, which is the case for us in view of the well-known in ET the GEM-type algorithm from [FH95]. In particular, this algorithm is specially designed for Poisson-type log-likelihood $L_p(\cdot)$, where $\varphi(\cdot)$ must be a convex pairwise difference penalty, for example, as one in (2.19).

3.3 WLB for emission tomographies without MRI data

The case when no multimodal data are used corresponds to the choice $\theta^t \equiv 0$. Therefore, Algorithm 2 can be rewritten as follows.

Algorithm 3: WLB for ET without multimodal data

Data: sinogram data Y^t, multimodal data $\mathcal{M}$

Input: B – number of samples, projector A
regularization parameter β^t, penalty $\varphi(\lambda)$

1 **for** $b = 1$ **to** B **do**

2 Draw $\widetilde{\Lambda}_{b,i}^t \sim \Gamma(Y_i^t, t^{-1})$ independently for each $i \in \{1, \dots, d\}$;

3 Compute $\widetilde{\lambda}_b^t = \underset{\lambda \succeq 0}{\arg\min}\, L_p(\lambda | \, \widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$ for $L_p(\cdot)$ defined in (2.10)
using the GEM-type algorithm from [FH95] (see Remark 3.3);

4 **end**

Output: $\{\widetilde{\lambda}_b^t\}_{b=1}^B$

We name Algorithm 3 by weighted likelihood bootstrap since it is a direct analog of the classical sampling algorithm from [NR94], [NPX21] being adapted for ET context; see also [LWH18], [FLH19],[Pom21] for connections between classical WLB/WBB and NPL.

3.4 Binned NPL for emission tomographies with MRI data

First, we construct prior $P_{\mathcal{M}}$, then we proceed with construction of posterior $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t, t)$.

Prior distribution for mixing parameter $\Lambda_{\mathcal{M}}$. Probability distribution $P_{\mathcal{M}}$ is defined via the following sampling scheme:

1. Recall that $\mathcal{M} = \{M_1, \dots, M_r\}$ are the segmented MRI images (see also Section 2), p_k denotes the number of disjoint segments in image $M_k \in \mathcal{M}$. Each segment is a subset of $\{1, \dots, p\}$, collection of segments in image M_k is denoted by $S(M_k) \subset 2^p$.
2. For each image $k \in \{1, \dots, r\}$ and segment $s \in S(M_k)$, we generate $\lambda^k_s \sim \Gamma(1, \infty)$ (uniform (improper) distribution on $\mathbb{R}_+$).
3. Compute random projections

$$\Lambda_{\mathcal{M},i} = \sum_{k=1}^{r} \sum_{s=1}^{p_k} a^k_{is} \lambda^k_s, \text{ for each } i \in \{1, \dots, d\}. \tag{3.22}$$

where

$$a^k_{is} = \sum_{j=1}^{p} a_{ij} \mathbb{1}\{\text{pixel } j \text{ belongs to segment } s \in S(M_k)\}, \; k \in \{1, \dots, r\}. \tag{3.23}$$

Note that $\Lambda_{\mathcal{M},i}$ in (3.22) is defined through the sum of projections over all images in $\mathcal{M}$. This can be seen as concatenating r models with segmentations :

$$A_{\mathcal{M}} = (A_1, \dots, A_r) \in \text{Mat}\,(d, p_{\mathcal{M}})\,, \; A_k = (a^k_{ij}) \in \text{Mat}(d, p_k), \; p_{\mathcal{M}} = \sum_{k=1}^{r} p_k, \tag{3.24}$$

$$\lambda_{\mathcal{M}} = (\lambda^1_1, \dots, \lambda^1_{p_1}, \dots, \lambda^r_1, \dots, \lambda^r_{p_r}), \tag{3.25}$$

Using notations from (3.24), (3.25), formula (3.22) can be rewritten as follows:

$$\Lambda_{\mathcal{M}} = A_{\mathcal{M}} \lambda_{\mathcal{M}}, \; \Lambda_{\mathcal{M}} = (\Lambda_{\mathcal{M},1}, \dots, \Lambda_{\mathcal{M},d}). \tag{3.26}$$

For design matrix $A_{\mathcal{M}}$ we assume that it is injective and well-conditioned, that is

$$\ker A_{\mathcal{M}} = \{0\}, \; \text{cond}(A_{\mathcal{M}}) < c_{\mathcal{M}}. \tag{3.27}$$

The latter assumption reflects our intuition that images in $\mathcal{M}$ consist of low number of large segments. In practice, condition (3.27) can be checked via the singular values of $A^T_{\mathcal{M}} A_{\mathcal{M}}$ which, in turn, can be precomputed due to apriori moderate size of $A_{\mathcal{M}}$.

Posterior distribution for mixing parameter $\Lambda^t_{\mathcal{M}}$. Posterior $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t, t)$ is defined through Bayes' formula for prior $P_{\mathcal{M}}$ and model $P(Y^t|A_{\mathcal{M}}, \lambda_{\mathcal{M}}, t)$ defined in (2.8) for design $A_{\mathcal{M}}$ and parameter $\lambda_{\mathcal{M}}$ from (3.24), (3.25), respectively. In principle, due to moderate size of $A_{\mathcal{M}}$ it is possible to use MCMC-approach (e.g., a Gibbs sampler) to sample from $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t, t)$, however, in order to keep the overall implementation as simple as possible we turn again to bootstrap for approximate posterior sampling.

Algorithm 4: Approximate sampling from $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t, t)$

Data: sinogram data Y^t, multimodal data $\mathcal{M}$

Input: design matrix $A_{\mathcal{M}}$ from (3.23), (3.24)

1 Draw $\Lambda^t_i \sim \Gamma(Y^t_i, t^{-1})$ independently for each $i \in \{1, \dots, d\}$;

2 Compute $\lambda^t_{\mathcal{M}} = \arg\min_{\lambda \succeq 0} L(\lambda|\Lambda^t, A_{\mathcal{M}}, 1)$, where $L(\cdot)$ is defined in (2.9);

3 Compute projections $\Lambda^t_{\mathcal{M}} = A_{\mathcal{M}}\lambda^t_{\mathcal{M}}$;

Output: $\Lambda^t_{\mathcal{M}}$ is sampled approximately from $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t, t)$

Remark 3.4. Since we assume that $A_{\mathcal{M}}$ is injective and is not ill-conditioned (see formula (3.27)), minimizer $\lambda^t_{\mathcal{M}}$ in step 2 of Algorithm 4 can be efficiently computed via the classical MLEM algorithm from [SV82]. □

Final algorithm. Putting together Algorithms 2, 4 and using the argument in Remark 3.3 we get the following sampling algorithm for the problem of ET with multimodal data available.

Algorithm 5: Binned NPL for ET with MRI data

Data: sinogram data Y^t, multimodal data $\mathcal{M}$

Input: B – number of samples, parameter θ^t, projector $A_{\mathcal{M}}$, projector A, regularization parameter β^t, penalty $\varphi(\lambda)$

1 **for** $b = 1$ **to** B **do**

2 Draw $\Lambda^t_{\mathcal{M}} = (\Lambda^t_{\mathcal{M},1}, \dots, \Lambda^t_{\mathcal{M},d})$ from $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t, t)$ via Algorithm 4;

3 Draw $\widetilde{\Lambda}^t_{b,i} \sim \Gamma(Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}, (\theta^t + t)^{-1})$ independently for each i;

4 Compute $\widetilde{\lambda}^t_b = \arg\min_{\lambda \succeq 0} L_p(\lambda|\, \widetilde{\Lambda}^t_b, A, 1, \beta^t/t)$ for $L_p(\cdot)$ defined in (2.10) using the GEM-type algorithm from [FH95];

5 **end**

Output: $\{\widetilde{\lambda}^t_b\}_{b=1}^B$

Remark 3.5. Note that parameter θ^t in Algorithm 5 has the following physical interpretation: it is exactly the rate of creation of "pseudo-photons" in the poisson model constructed from MRI data. For example, by choosing $\theta^t = \rho t$, $\rho \geq 0$ in step 3 we sum up sinograms Y^t and $t \cdot \Lambda^t_{\mathcal{M}}$ in proportions $1/(1+\rho)$ and $\rho/(1+\rho)$, respectively. □

Remark 3.6. A very recent and similar to ours sampling algorithm was proposed in [FDC+21] provided with a very extensive experiment both on synthetic and real data from PET. The algorithm there is also of boostrap-type, based on optimization of a randomized functional (also log-likelihood for the generalized Poisson model) and in fact, it coincides up to minor details with Algorithm 3 (where we do not use MRI data). Instead, MRI data $\mathcal{M}$ are used there to construct very special penalty $\varphi(\lambda) = \varphi_{\mathcal{M}}(\lambda)$ of Bowsher type (see Section 2; paragraph on multimodal data). This penalty satisifes the assumptions in (2.15), (2.16), so theorems 6.2, 6.5 serve as a theoretical foundation also for the algorithms presented there. Here, we note that a nice practical feature of our algorithms is that θ^t has clear physical interpretation of the effect of MRI data on samples (see Remark 3.5), whereas large number of parameters in Bowsher-type penalties have no such easy interpretations making the problem of their calibration cubersome.

Aforementioned minor differences between algorithms consist in the way data Y^t (in [FDC+21]) or intensities Λ_i (in our work) are stochastically perturbed. From the first look this seems to be only a technical question, however, we think that it is not. From the above derivation of Algorithms 3, 5 one can see that initially uncertainty propagates through the KL-projection problem in (3.5) and not concerning at all the problem of limited data. Moreover, we retrieve Algorithm 3 as a particular case of Algorithm 5 when choosing the scale parameter $\theta^t = 0$ in the nonparametric prior in (3.12). This is fully coherent with the derivation of NPL in [LWH18] and nonparametric posterior bootstrap with MDP-prior in [FLH19], where the classical WLB algorithm from [NR94] is retrieved back as a partiuclar case when choosing the concentration parameter $\alpha = 0$ ($c = 0$ in [FLH19]) in the nonparametric Dirichlet process prior. On the other hand, the derivation in [FDC+21] strongly relies on model with finite data and it is claimed that the resulting algorithm is also a version of WLB from [NR94], however, in this case it for us is not clear which nonparametric extension stands behind. □

Numerical tests of Algorithms 3 and 5 are given in Section 5. In the next section we discuss a somewhat negative but very informative example of pure bayesian posterior sampling in ET which motivates us to use the NPL instead.

4 A motivating example for NPL in ET

In recent work [FBD+18] a Gibbs-type sampler was proposed for bayesian inference for PET with exterior MRI data. Despite a number of positive practical features (spatial regularization, use of multimodal data) a problem of slow mixing for the corresponding Markov chain was observed. In fact, slow mixing was observed almost for any values of calibration parameters which, in turn, required a mathematical explanation for this phenomenon.

Below we present an example of a simplified version of a Gibbs sampler for PET which has the same mixing problem. For this example we describe very precisely the asymptotic mixing rate which, in turn, explains completely the observed phenomenon in [FBD+18]. In view of this example, Algorithms 3, 5 seem to give a good trade-off between design complexity of the sampler, numerical load per iteration and provided precision for uncertainty quantification.

In algorithms for emission tomographies (e.g., in PET/SPECT) it is common to introduce latent variables $n^t = \{n^t_{ij}\}$, which are defined as follows:

$$\begin{aligned} &n^t_{ij} \text{ – number of photons emitted from pixel } j \text{ and detected in LOR } i, \\ &n^t_{ij} \sim \mathrm{Po}(t \cdot a_{ij}\lambda_j),\ n^t_{ij} \text{ are mutually independent for all } (i,j). \end{aligned} \tag{4.1}$$

Note that n^t are not observed in the real experiment but only Y^t. For random variable (n^t, Y^t) the following coherence condition must be satisfied:

$$\sum_{j=1}^{p} n^t_{ij} = Y^t_i \text{ for all } i \in \{1, \dots, p\}. \tag{4.2}$$

From (4.2) it follows that Y^t is a function of n^t, so (Y^t, n^t) is a data augmentation of Y^t; see e.g., [SV82].

The point is that n^t greatly simplify the design of samplers (see e.g., [Jam03], [FBD+18], [FBC+11]), because conditional distributions $p(n^t|Y^t, A, \lambda, t)$, $p(\lambda|n^t, A, t)$ admit very simple analytical forms even for nontrivial priors involving multimodal data. For our example

below we use only a simple pixel-wise positivity gamma-prior:

$$\pi(\lambda) = \prod_{i=1}^{p} \pi_j(\lambda_j),\ \pi_j = \Gamma(\alpha, \beta^{-1}),\ \alpha > 0,\ \beta > 0, \tag{4.3}$$

where α, β are some fixed constants. For prior in (4.3) and the poisson model (2.1) conditional distributions $p(n^t|Y^t, A, \lambda, t)$, $p(\lambda|n^t, A, t)$ are as follows:

$$\begin{aligned} p(n_{ij}^t|Y^t, A, \lambda, t) &= \text{Multinomial}(Y_i^t, p_{i1}(\lambda), \dots, p_{ip}(\lambda)), \\ p_{ij}(\lambda) &= \frac{a_{ij}\lambda_j}{\sum_k a_{ik}\lambda_k},\ i \in \{1, \dots, d\}, \end{aligned} \tag{4.4}$$

$$p(\lambda|n^t, Y^t, A, t) = \Gamma\left(\sum_{i=1}^{d} n_{ij}^t + \alpha, (tA_j + \beta)^{-1}\right), \tag{4.5}$$

where A_j is defined in (2.4).

Using (4.4), (4.5) the construction a Gibbs sampler for bayesian posterior sampling from $p(\lambda|Y^t, A, t)$ is straightforward.

Algorithm 6: Gibbs sampler for $p(\lambda|Y^t, A, t)$

Data: sinogram data Y^t

Input: initial point $\lambda_0 \in \mathbb{R}_+^p$, parameters (α, β) for prior $\pi(\lambda_j) \sim \Gamma(\alpha, \beta^{-1})$, design matrix A, N – number of posterior samples

1 **for** $k = 1$ **to** N **do**

2 Sample $n_k^t \sim p(n^t|Y^t, A, \lambda_{k-1}, t)$ using formula (4.4);

3 Sample $\lambda_k^t \sim p(\lambda|n_k^t, Y^t, A, t)$ using formula (4.5);

4 **end**

Output: samples $\{\lambda_k^t\}_{k=1}^N$

Result: For large N distribution of $\{\lambda_k^t\}_{k=1}^N$ approximates posterior $p(\lambda|Y^t, A, t)$ (see e.g., [KS06])

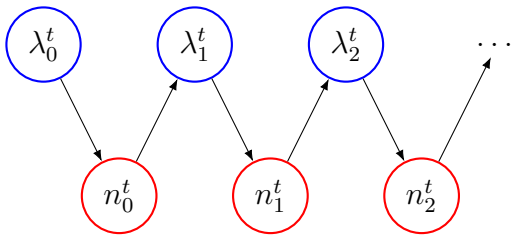


Figure 2: scheme for posterior sampling in Algorithm 6

Remark 4.1. One may argue that prior in (4.3) is a very bad choice from practical point of view, especially in view of ill-posedness of the inverse problem since it does not bring any regularization. We consider the mixing rate for the Markov chain in Algorithm 6 in the small noise limit, i.e., when $t \to +\infty$, and for the latter it is known from the Bernstein von-Mises theorem (see e.g., [BG14]) that asymptotically any prior effect will disappear no matter the choice of $\pi(\lambda)$.

Let $h = h(\lambda)$ be a bounded functional with respect to L^2-norm for integration measure $p(\lambda|Y^t, A, t)$, that is

$$\int\limits_{\mathbb{R}^p_+} |h(\lambda)|^2 p(\lambda|Y^t, A, t)\, d\lambda < +\infty. \tag{4.6}$$

In what follows we choose $h(\lambda)$ to be linear, i.e., $h(\lambda) = h^T\lambda$, for some $h \in \mathbb{R}^p$. In this case condition (4.6) translates as existence of second moments for posterior $p(\lambda|Y^t, A, t)$.

Consider the correlations between values of $h(\lambda)$ for subsequent samples from the Markov chain in Algorithm 6:

$$\gamma^t(h) = \mathrm{corr}(h(\lambda^t_{k+1}), h(\lambda^t_k)|Y^t, t). \tag{4.7}$$

In formula (4.7) we assume that the chain is in stationary state, therefore k can be any.

It is important that Markov chain for the Gibbs sampler in Algorithm 6 coincides with data augmentation schemes from [Liu94], [LWK94], where the latter are exactly Gibbs samplers with only one layer of latent variables.

In bayesian context $\gamma^t(h)$ is known as bayesian fraction of missing information; see [Liu94]. In particular, in [Liu94] authors gave an exact formula for $\gamma^t(h)$ which can be written for our example as follows:

$$\gamma^t(h) = 1 - \frac{\mathbb{E}[\mathrm{Var}(h(\lambda)|n^t, Y^t, t)|Y^t, t]}{\mathrm{Var}(h(\lambda)|Y^t, t)}. \tag{4.8}$$

Exact formulas for the nominator and the denominator in (4.8) for arbitrary t seem difficult (if possible) to be obtained, however, in the asymptotic regime $t \to +\infty$ one can apply the Bernstein von-Mises type theorem from [BG14].

For simplicity assume that

$$\lambda_{*j} > 0 \text{ for all } j \in \{1, \dots, p\} \text{ (see also Remark 4.2).} \tag{4.9}$$

Then, for the asymptotic version of (4.8) one gets the following simple formula:

$$\gamma(h) = \lim_{t\to+\infty} \gamma^t(h) = 1 - \frac{h^T F^{-1}_{aug}(\lambda_*)h}{h^T F^{-1}_{obs}(\lambda_*)h},\ h \in \mathbb{R}^p,\ \text{ a.s. } Y^t, t \in (0, +\infty). \tag{4.10}$$

where

$$\lambda_* \in \mathbb{R}^p_+ \text{ is the true concentration of the tracer,} \tag{4.11}$$

$$F_{obs}(\lambda_*) = \sum_{i=1}^{d} \frac{a_i a_i^T}{\Lambda_i^*} = A^T D^{-1}_{\Lambda^*} A,\ D_{\Lambda^*} = \mathrm{diag}(\dots, \Lambda_i^*, \dots),\ \Lambda_i^* = a_i^T\lambda_*, \tag{4.12}$$

$$F_{aug}(\lambda_*) = \mathrm{diag}(\dots, c_j, \dots),\ c_j = A_j/\lambda_{*j}. \tag{4.13}$$

Note that from (2.5), (4.9) it follows that $\Lambda_i^* > 0$ for all $i \in \{1, \dots, d\}$, therefore division by Λ_i^* in (4.12) is well-defined.

Matrices $F_{obs}(\lambda_*)$, $F_{aug}(\lambda_*)$ are the Fisher information matrices at λ_* for models (2.1), (4.1) with observables Y^t, n^t, respectively. Note also that F_{obs} is not invertible in the usual sense, so in (4.10) its pseudo-inversion in the sense of Moore-Penrose is considered.

Remark 4.2. Assumption in (4.9) is not practical and a precise analytic formula which extends (4.10) for arbitrary $\lambda_* \in \mathbb{R}^p_+$ can be established using the results from [BG14]. The point is that models (2.1), (4.1) are non-regular since the parameter of interest belongs

to a domain with a boundary, so a separate result for Bernstein von-Mises phenomenon is needed in this case. For our purposes it is sufficient to consider this simple case (when the condition in (4.9) is satisfied) since we are mostly interested in mixing times of the Markov chain in areas with positive tracer concentration.

□

Now, let $h_1, \dots, h_p$ be the orthonormal basis of eigenvectors of $F_{obs}(\lambda_*)$ being ordered according to their respective eigenvalues $s_1 \geq s_2 \cdots \geq s_p \geq 0$. From (4.10) it follows that

$$\gamma(h_m) = 1 - s_m h_m^T F_{aug}^{-1} h_m. \tag{4.14}$$

From (4.12) and the ill-conditioning nature of A (see formula (2.6) and Remark 2.1) it follows that $F_{obs}(\lambda_*)$ is also ill-conditioned, moreover, $s_m \approx 0$ for large m. In practice the ill-conditioning of $F_{obs}(\lambda_*)$ is commonly observed in PET/SPECT practice in form of very slow convergence of non-penalized EM-algorithms. At the same time $F_{aug}(\lambda_*)$ is well-conditioned and admits, for example, the following bound:

$$F_{aug}^{-1}(\lambda_*) = \operatorname{diag}(\dots, \frac{\lambda_{*j}}{A_j}, \dots) \Rightarrow h_m^T F_{aug}^{-1}(\lambda_*) h_m \leq \frac{\max_j(\lambda_{*j})}{\min_j(A_j)}. \tag{4.15}$$

Regular behavior of F_{aug}^{-1} in (4.15) is not surprising because this is the Fisher information matrix for latent variables n^t for which the inverse problem is not ill-posed at all (photon counts n_{ij}^t are observed separately for each pixel j for any LOR i).

From (4.14), (4.15) we conclude that

$$\gamma(h_m) \approx 1 \text{ for large } m. \tag{4.16}$$

Formulas (4.7), (4.16) constitute a clear evidence of poor mixing in the Markov chain in Algorithm 6. Though formulas (4.10)-(4.16) were derived for $t \to +\infty$, they reflect well the behavior for moderate t which is seen from the numerical experiment below.

λ_* – image of size 64×64 (see figure 3),
A – Radon transform matrix of size 4096×4096,
prior $\pi_j = \Gamma(1, 1)$,
time $t = 10^4, 10^{10}$ ($\sim$ photons per LOR),
initial point: λ_*,
burn-in samples: 1000,
number of samples for the output: 2000

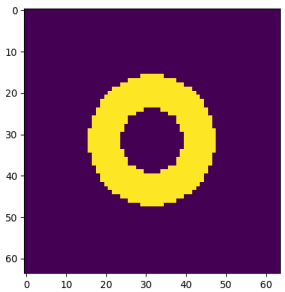


Figure 3: true distribution λ_*

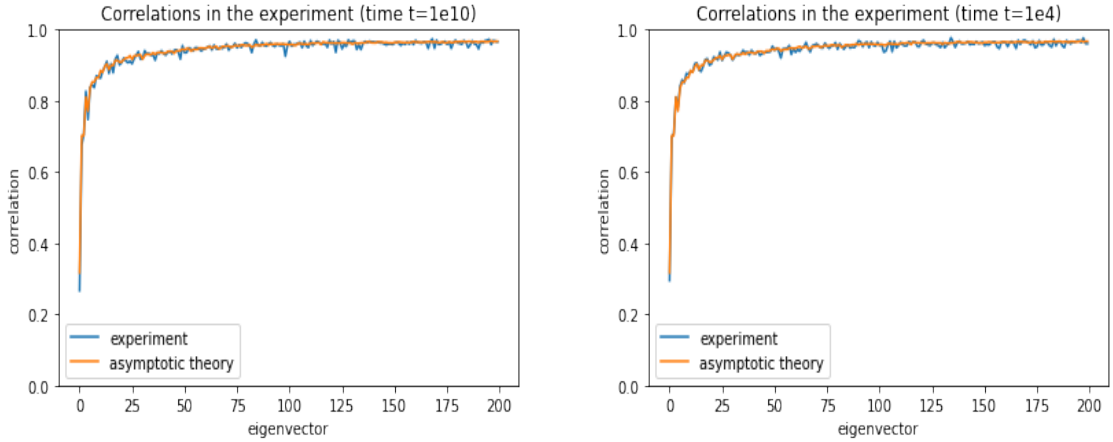
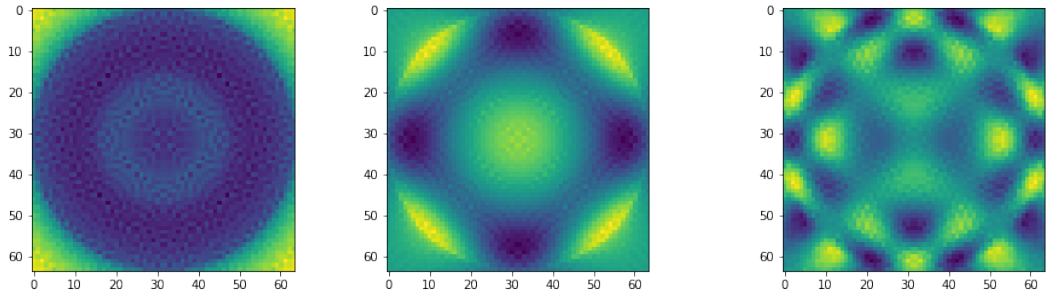


Figure 4: values for $\mathrm{corr}(h^T\lambda_k^t, h^T\lambda_{k+1}^t|Y^t)$ for $t = 10^4, 10^{10}$ for $h = h_m$; blue curve – empirical correlations computed from 2000 samples, orange curve – values for $\gamma(h_m)$ for $m = 1, \ldots, 200$.



(a) h_1 (b) h_2 (c) h_{50}

Figure 5: eigenvectors h_m for $F_{obs}(\lambda_*)$, where λ_* is from Figure 3

We end this section by giving a practical interpretation of obtained results.

First, in Figure 4 one can see that correlations $\mathrm{corr}(h_m(\lambda_k^t), h_m(\lambda_{k+1}^t)|Y^t, t)$ increase fast with index m and that values by formula (4.14) are in full correspondence with our numerical results. Vectors $h_1, \ldots, h_p$ constitute a basis in space of reconstructed images and higher indices m correspond to higher frequencies of images (see Figure 5). From (4.16) and from the plot in Figure 4 one concludes that mixing is much slower for high frequency parts of images.

Example 1. In practice produced samples are used to compute credible intervals for weighted means in certain subregions of reconstructed images. Let $h \in \mathbb{R}^p$ be a weighting mask which corresponds to subregion $\Omega \subset \{1, \ldots, p\}$. For example, if $h_j = \frac{1}{\#\Omega}$ for pixel $j \in \Omega$ and $h_j = 0$ otherwise, then $h^T\lambda$ gives the average tracer concentration in subregion Ω. Let N be the number of generated samples which we denote by $\{\lambda_k^t\}_{k=1}^N$. Then, the posterior mean of $h^T\lambda$ can be approximated by the following expression:

$$\widehat{f}_{h,N}^t = \frac{1}{N}\sum_{k=1}^{N} h^T\lambda_k^t, \tag{4.17}$$

An important question here would be:

$$\text{How many samples } \lambda_k^t \text{ are needed to estimate } h^T\lambda \text{ reliably?} \tag{4.18}$$

The variance of estimator $\widehat{f}^t_{h,N}$ can be approximated as follows:

$$\begin{aligned}\mathrm{Var}(\widehat{f}^t_{h,N}|Y^t,t) &= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{s=1}^{N}\mathrm{cov}(h(\lambda^t_k),h(\lambda^t_s)|Y^t,t)\\ &\asymp \frac{\sigma^2}{N}(1+2\sum_{k=1}^{\infty}\rho^t_k(h)),\end{aligned} \tag{4.19}$$

where

$$\rho^t_k(h) = \mathrm{corr}(h^T\lambda^t_1, h^T\lambda^t_{k+1}|Y^t,t),\ \sigma^2 = \mathrm{Var}(h^T\lambda). \tag{4.20}$$

In [LWK94] it was shown, in particular, that $\rho^t_k(h) \asymp \gamma^t(h)^k$, so from this and the above formula we get the following expression for the variance of $\widehat{f}^t_{h,N}$ (modulo a universal multiplicative factor):

$$\mathrm{Var}(\widehat{f}_{h,N}|Y^t,t) \asymp \frac{\sigma^2}{N}\left(\frac{1+\gamma^t(h)}{1-\gamma^t(h)}\right) \approx \frac{\sigma^2}{N}\left(\frac{1+\gamma(h)}{1-\gamma(h)}\right) \tag{4.21}$$

where $\gamma^t(h)$, $\gamma(h)$ are defined in (4.7), (4.10), respectively. The rule of thumb in [AG91] tells to choose N such that empirical variance of $\widehat{f}_{h,N}$ does not exceed 1% of σ^2, which is then translated to the following rule:

$$\frac{\mathrm{Var}(\widehat{f}_{h,N}|Y^t,t)}{\sigma^2} < 0.01 \Rightarrow N \gtrsim 100\cdot\left(\frac{1+\gamma(h)}{1-\gamma(h)}\right) \to +\infty \text{ for } h = h_m,\ m \gg 1. \tag{4.22}$$

Therefore, to estimate reliably the average signal using mask $h \in \mathbb{R}^p$, one needs almost infinite number of samples if h contains a high-frequency component in terms of basis $\{h_k\}_{k=1}^p$. This also can be seen as a recommendation for choosing mask h in practice: h should belong to $\mathrm{Span}(A^T)$ and projections $h^T h_m$ should be as small as possible for large m. □

It is important that such behavior of the sampler is not due to the choice of pixel-wise prior but due to the decision to use latent variables n^t which correspond to observations for the well-posed inverse problem for PET. In this situation a practical advice would be to avoid to use n^t is the design of the sampler or to use a strong smoothing prior (for example by greatly increasing regularization coefficients so that asymptotic arguments in (4.10) will no longer hold but the posterior consistency is still preserved). The latter approach will accelerate mixing at cost of oversmoothing in sampled images.

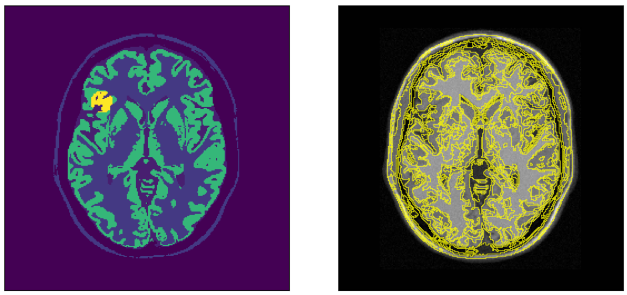
By this negative but informative example we support the message in [VDM01] that design of a data agumentation scheme while preserving good mixing in the Markov chain is an "art", especially, in the case of ill-posed inverse problems. In view of complexity of the design and implementation, high numerical load while using MCMC-sampling [Wei97], [HBJ+97], [FL07], [FBD+18], Algorithms 3, 5 seem to be a good practical relaxation of exact posterior sampling for the problem of ETs.

5 Numerical experiment

5.1 Experiment design

We illustrate our algorithm on synthetic PET data based on a realistic phantom from the BrainWeb database [VAB+11]. Typical activity concentrations have been assigned to annotated tissues (gray matter, white matter, skin, *etc.*) and we delineated a tumor

lesion area, not present in the initial phantom with an uptake of 50% compared to the gray matter activity. The anatomical MRI (T1) phantom does not contain any information relative to the lesion. Nonparametric bayesian over-segmentation of side images is amenable with DP-Potts [XYCD16] or ddCRP [BF11], [GUSB11] MCMC algorithms. In our experiment, we used ddCRP with a concentration parameter fixed to 10^{-5}, leading to a few hundreds of random superpixels for a 2D brain slice during sampling. Though several random segmentations might be considered, for seek of simplicity, we selected a single sample among few ones which maximized the corresonding MRI log-likelihood. In Figure 6 are shown the 2D emission map used for data generation and the ddCRP over-segmentation overlayed to MRI.



(a) λ_* (b) $M_1 \in \mathcal{M}$

Figure 6: emission map with lesion hot spot at (a) and segmented MRI at (b)

The reconstruction grid was taken 256×256 pixels ($p = 2^{16}$) being identical to the phantom's one. The binned observation space consists from LORs derived from a ring of 512 detectors spaced uniformly on a circle. Design A was generated using the Siddon's algorithm [Sid85] and $A_{\mathcal{M}}$ was computed from A and segmented image $M_1 \in \mathcal{M}$ using formulas (3.23), (3.24). The intensity was set so that $\sum_{i=1}^{p} \lambda_{*k} = 5 \cdot 10^5$ and for the experiment with mild t time was set to $t_1 = 1$; for large t (when asymptotic approximation is better) we set $t_2 = 100$. Sinograms for t_1, t_2 were generated via formula (2.1).

Note that the misspecification in using design $A_{\mathcal{M}}$ in the prior is mainly due here to the fact that the lesion signature is not reflected in $\mathcal{M}$ and, more generally, to the mismatch between the actual emission map λ_* and the segmentation in $\mathcal{M}$.

5.2 GEM-type algorithm

The attractiveness of Algorithms 3, 5 depends solely on having an efficient procedure for minimizing $L_p(\lambda|\,\widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$ (or equivalently $L(\lambda|\Lambda_b^t, A_{\mathcal{M}}, 1)$). For integer-valued data $Y^t \in \mathbb{N}_0^d$ (e.g., when the assumption in (2.1) holds) the $L_p(\lambda|Y^t, \cdot)$ coincides with the penalized negative log-likelihood for Poisson-type sample. In this situation, provided penalty $\varphi(\lambda)$ satisfies elementary conditions (convex, C^2 – smooth), fast monotonic GEM algorithms [FH95], [WQ15] can be used.

In our setting intensities $\widetilde{\Lambda}_b^t$ are not integer-valued anymore, hence the GEM derivation machinery must be re-verified. We claim that the same so-called "GEM-type" iterative algorithms can be derived outside the context of a Poisson model and missing data. First, notice that EM belongs to the class of optimization transfer algorithms [LHY00] also denoted as MM (Majoration Minimization). In this context, the E-step is interpreted as the construction of a *majorizing surrogate* for the objective function, M-step corresponds to its consequent minimization (negative log-likelihood). Using the convexity argument from [DP93] we construct the same majoring surrogate for $L(\lambda|\,\widetilde{\Lambda}_b^t, A, 1)$ as in [FH95] in a completely algebraic way but now for arbitrary nonnegative data term $\widetilde{\Lambda}_b^t$. Further

extension to $L_p(\lambda|\,\widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$ is straightforward by considering a separate surrogate for $\varphi(\lambda)$. Details are given in Appendix C.

An immediate and substantial consequence for practitioners is that all celebrated GEM algorithms for MLE and MAP reconstructions can be used in the bootstrap context by simply replacing Poisson data term by $\widetilde{\Lambda}_b^t$.

5.3 Algorithm settings

For $\varphi(\lambda)$ we use the function from (2.19), where parameters are chosen as follows: $\zeta = 0.05$, $\nu = 0.15$, $\beta^t = 2 \cdot 10^{-3}$. For $t_1 = 1$, we present results for $\rho = \theta^t/t \in \{0, 0.25, 0.5, 1, 2\}$ (see Remark 3.5). For $t_2 = 100$ we choose only one value $\rho = 0.1$. For each combination of t, ρ, Algorithm 5 was generating $B = 1000$ bootstrap draws from which further statistics were computed (empricial mean, standard devation, etc.).

5.4 Results

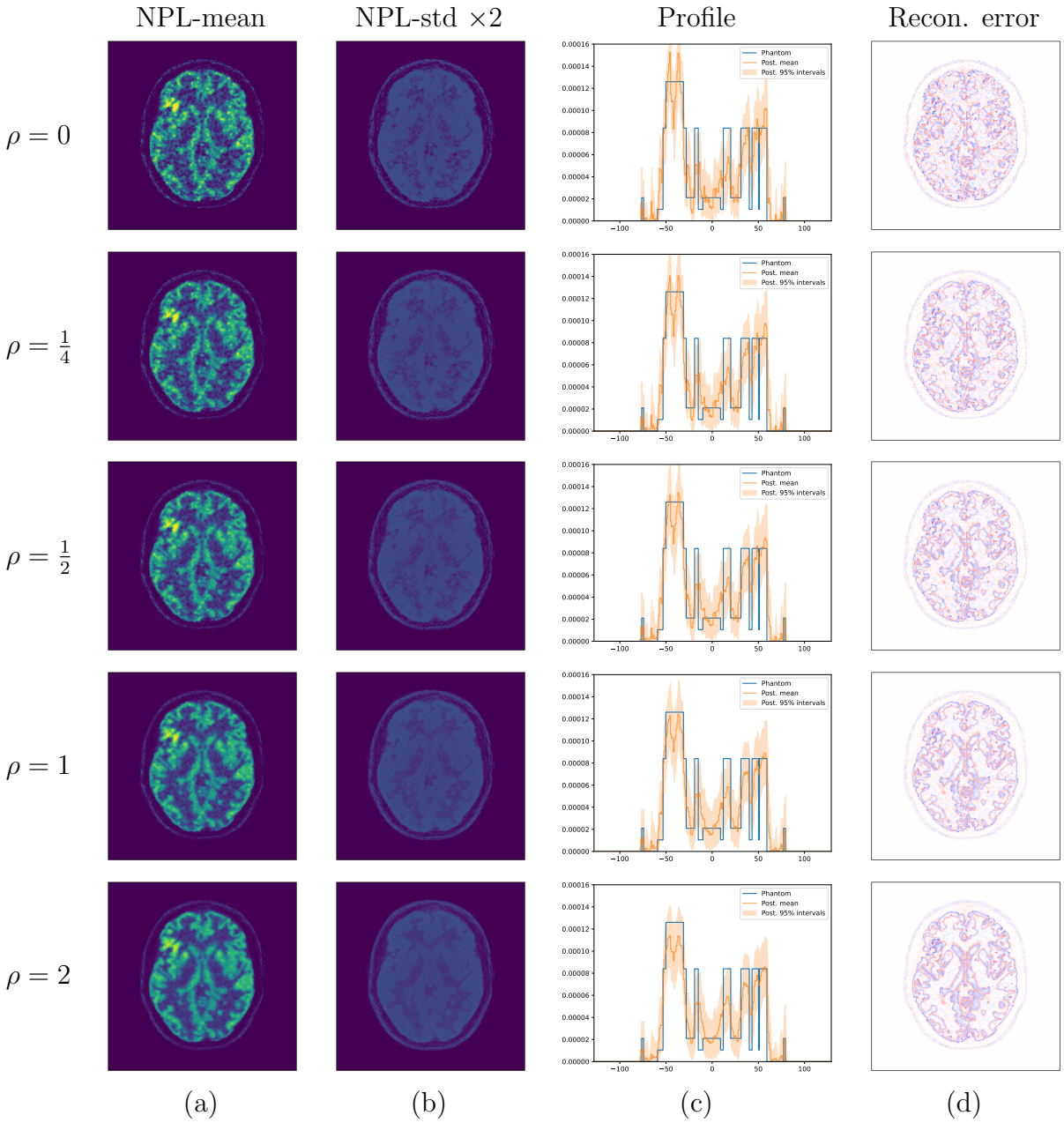


Figure 7: Columns represent respectively the posterior mean (a), twice the posterior standard deviation with same color scale as mean (b), posterior 95% band on an horizontal profile through the lesion (c), the absolute reconstruction error (d).

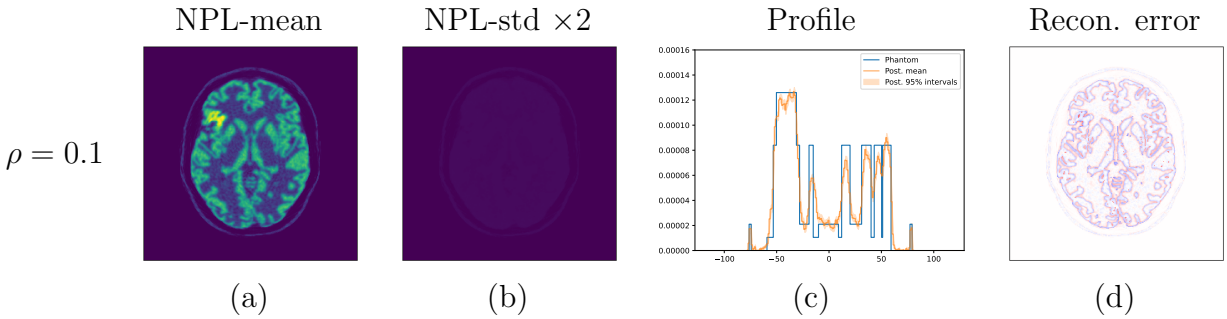


Figure 8: NPL with $\rho = 0.1$ for $t = 100$

As expected, higher ρ values reinforce the effect of the MRI on reconstructions. Posterior variance decreases slowly with ρ while bias in the lesion area increases. As a rule of thumb, it seems reasonable not to exceed $\rho = 1$ – the weight of pseudo-data from the misspecified model of the prior should not exceed the weight of observed data. We checked that NPL posterior mean with $\rho = 0$ (no MRI information) is indistinguishable from the MAP reconstruction with the same penalty tuning (see Appendix D). This supports our theoretical finding in Subsection 6.3, namely Theorem 6.5, saying that the asymptotic distribution is concentrated not around λ_* but a strongly consistent estimator for which we conjecture to coincide with MAP (or MLE).

To give an intuition of asymptotic behavior of NPL reconstruction, we made an additional experiment for $t = 100$ for the same phantom intensity map (Figure 8). The posterior 95% band is not able to capture the phantom's high frequencies (sharp edges), however, this is not surprising in view of result of Theorem 6.1 which says that reconstructions asymptotically converge to the true image up to its projection onto $\ker A$ which mainly contains high frequencies. In addition, in view of ill-conditionality of A (see Remark 2.1) and access to limited data set Y^t not only components from $\ker A$ are not captured, but also components which are close to $\ker A$. Moreover, again the result of Theorem 6.5 says, that essentially, the posterior mean is controlled by the strongly consistent estimator, so if MAP estimate (or any other frequentist estimator) does not capture higher frequencies, then it is not likely that bootstrap samples will.

6 Asymptotic analysis of the new algorithm

Statistical model (2.1) is non-regular because the domain for parameter λ is not open, contains boundary $\partial\mathbb{R}^p_+ = \{\lambda \in \mathbb{R}^p_+ : \exists j \text{ s.t. } \lambda_j = 0\}$ and it is possible, in general, that $\lambda_* \in \partial\mathbb{R}^p_+$. This model was investigated in the small noise limit (i.e., when $t \to +\infty$) in pure bayesian framework in [BG14] for large class of priors for the well-specified case (i.e., when $Y^t \sim P^t_{A,\lambda_*}$ for some $\lambda_* \in \mathbb{R}^p_+$) and for design A of the full rank though also ill-conditioned. It was shown that the posterior is consistent at λ_*, the asymptotic distribution is centered around the MLE estimate for the quadratic approximation of $L(\lambda|Y^t, A, t)$ and the nonregularity results in splitting of the asymptotic posterior in three parts: multivariate exponential (for coordinates which are related to pixels intersected by LORs with zero intensities) contracting to zero with the fastest rate $\sim t$, gaussian (for coordinates where λ_* is in the interior of $\mathbb{R}^p_+$) and half-gaussian (for coordinates with $\lambda_{*,j} = 0$ and pixels being intersected only by LORs with positive intensities) with rate $\sim \sqrt{t}$.

Our results for consistency and characterization of asymptotic conditional distribution of samples are similar to ones from [BG14], however, there are several major and minor differences. Intuitively, the asymptotic distribution should be similar to the distribution of MAP estimates for bayesian posterior from [BG14] (but with perturbed data Y^t): atom at zero for the exponential part, gaussian – for the gaussian part, and sum of atom at zero and half-gaussian for the half-gaussian part (see [Gey94]).

Asymptotic consistency at λ_* and the aforementioned splitting are also present in posterior learning, and the asymptotic distribution is tight around a strongly consistent estimator $\widehat{\lambda}^t_{sc}$ satisfying additional properties in observation space for which the MAP estimate from [BG14] seems to be the right candidate. Interestingly, the splitting of the distribution into different modes depends not on λ_* (as it was in [BG14]) but again on $\widehat{\lambda}^t_{sc}$ because of which we fail to demonstrate the asymptotic normality since it requires additional results on behavior of strongly consistent estimators with constraints for the problem of ET. We address this investigation for future and conjecture that classical MLE or penalized MLE (i.e., MAP) are the right candidates for $\widehat{\lambda}^t_{sc}$. The assumptions we put

on $\widehat{\lambda}^t_{sc}$ for conditional tightness seem very natural and we discuss them thoroughly in the text.

A minor remark would be that, in pure bayesian framework there is only one free parameter that is controlled by a specialist – the prior distribution, whereas in Algorithm 5 we have several free parameters: θ^t, β^t, $A_{\mathcal{M}}$. Therefore, our theoretical results are also different from ones in [BG14] that they contain restrictions on the above parameters. At the end, we address the problem of model misspecification for the generalized Poisson model which arises twice our setting: first, in Algorithm 4 when sampling $\Lambda^t_{\mathcal{M}}$ (because we use data Y^t in model (2.1) with incorrect design $A_{\mathcal{M}}$) and, second, when assume that model (2.1) is wrong, in general.

This section is organized as follows. In Subsection 6.1 we define convergence of conditional probabilities which is necessary for statment of our theoretical results. In Subsection 6.2 we show that if model (2.1) is well specified, then posterior distribution is consistent at the true point λ_* up to its projection onto $\ker A$. In Subsection 6.3 we characterize the asymptotic distribution of $\widetilde{\lambda}^t_b$ in terms of conditional tightness. For this we study accurately posterior $P_{\mathcal{M}}(\Lambda^t_{\mathcal{M}}|Y^t,t)$ (see Algorithm 4) and show its concentration near the KL-minimizer of $\mathcal{KL}(P^t_{A,\lambda_*}, P^t_{A_{\mathcal{M}},\lambda_{\mathcal{M}}})$ with respect to $\lambda_{\mathcal{M}}$. Next, we characterize the concentration rate under a new identifiability condition for $\lambda_{\mathcal{M}}$ on A, $A_{\mathcal{M}}$ and λ_*. Using the model where designs A, $A_{\mathcal{M}}$ represent Radon-type transforms along straight lines we find a simple geometrical interpretation of this condition and propose to name it by *the non-expansiveness condition* or alternatively by *the mask condition*. In Subsection 6.4 we relax the initial assumption of correctness of the model and analyze the effect of wrong design in (2.1) in greater detail. We show that if the non-expansiveness condition fails, then minimizer of $\mathcal{KL}(P^t, P^t_{A,\lambda})$ with respect to λ may not be uniquely defined even if A is injective and satisfies (2.3)-(2.5). That is the identification problem for generalized poisson models in ET has a negative answer, in general. Finally, we propose a generalized version of the non-expansiveness condition and show that for it being satisfied the identification property holds.

6.1 Convergence for conditional probabilities.

In our theoretical considerations there are two levels of randomness: first one, which is closer to a practitioner, contains conditional distributions of $\widetilde{\lambda}^t_b$, $\Lambda^t_{\mathcal{M}}$ given data Y^t, and second is where the aforementioned distributions are considered to be random themselves due to randomness in Y^t. Theoretical validation of proposed algorithms consists in proving that conditional distributions in Algorithms 3, 5 will concentrate near the "true" point λ_* when $t \to +\infty$ almost for any trajectory Y^t, $t \in (0, +\infty)$. At the same time it is also important to characterize the rate of this concentration, which equivalent to characterization of the asymptotic conditional distribution of $\widetilde{\lambda}^t_b$ or, at least, its conditional tightness.

Let $(\Omega, \mathcal{F}, P)$ be the common probability space on which process Y^t, $t \in (0, +\infty)$ and MGP prior in (3.16) are defined (see Appendix A). Let

$$\mathcal{F}^t = \sigma(Y^\tau,\, \tau \in (0,t)) \subset \mathcal{F}, \tag{6.1}$$

where $\sigma(\cdot)$ denotes the sigma-algebra generated by a family of random variables.

Definition 6.1. We say that U^t converges in conditional probability to U almost surely Y^t, $t \in (0, +\infty)$ if for every $\varepsilon > 0$ the following holds:

$$P(\|U^t - U\| > \varepsilon \,|\, \mathcal{F}^t) \to 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{6.2}$$

This type of convergence will be denoted as follows:

$$U^t \xrightarrow{c.p.} U \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{6.3}$$

In the proofs for $U^t \xrightarrow{c.p.} 0$ we also write

$$U^t = o_{cp}(1). \tag{6.4}$$

□

Definition 6.2. We say that U^t is conditionally tight almost surely Y^t, $t \in (0, +\infty)$ if for any $\varepsilon > 0$ and almost any trajectory Y^t, $t \in (0, +\infty)$ there exists $M = M(\varepsilon, \{Y^t\}_{t\in(0,+\infty)})$ such that

$$\sup_{t\in(0,+\infty)} P(\|U^t\| > M \,|\, \mathcal{F}^t) < \varepsilon. \tag{6.5}$$

□

Definition 6.3. We say that U^t converges in conditional distribution to V almost surely Y^t, $t \in (0, +\infty)$ if for every Borel set $A \in B(\mathbb{R}^n)$ the following holds:

$$P(U^t \in A \,|\, \mathcal{F}^t) \to P(V \in A) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{6.6}$$

This type of convergence will be denoted as follows:

$$U^t \xrightarrow{c.d.} U \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{6.7}$$

□

In short, in defenitions above *almost surely* Y^t, $t \in (0, +\infty)$ means that limits in (6.2), (6.6) hold for *almost every trajectory* Y^t, $t \in (0, +\infty)$.

6.2 Consistency

Assumption 1. Model (2.1) is well-specified, that is

$$Y^t \sim P^t_{A,\lambda_*}, \text{ for some } \lambda_* \in \mathbb{R}^p_+ \text{ and all } t \in (0, +\infty), \tag{6.8}$$

where A satisfies (2.3)-(2.6), $P^t_{A,\lambda}$ is defined in (2.8).

Theorem 6.1. *Let Assumption 1 and conditions (2.15), (2.16) for $\varphi(\lambda)$ be satisfied and parameters β^t, θ^t be such that*

$$\beta^t/t \to 0,\ \theta^t/t \to 0 \ \text{ when } t \to +\infty. \tag{6.9}$$

Let $\widetilde{\lambda}^t_b$ be defined as in Algorithm 5. Then,

$$\widetilde{\lambda}^t_b - \lambda_* \xrightarrow{c.p.} w_{A,\lambda_*}(0) \ \text{ when } t \to +\infty, \ \text{ a.s. } Y^t, \ t \in (0, +\infty). \tag{6.10}$$

where $w_{A,\lambda}(\cdot)$ is defined in (2.18).

Note that the result of Theorem 6.1 automatically implies consistency for Algorithm 3.

The interpretation of formula (6.10) is straightforward: conditional distribution of $\widetilde{\lambda}^t_b$ asymptotically concentrates near λ_* in the subspace $\mathrm{Span}(A^T)$, where parameter λ is identifiable through design A and also regarding the positivity constraints. On the other hand, projection of λ_* onto $\ker A$ is not identifiable in model (2.1) and it is defined solely by penalty $\varphi(\lambda)$, the orthogonal projection of λ_* onto $\mathrm{Span}(A^T)$ and by $\ker A$.

There is also a natural generalization of the above theorem for any generic bootstrap type procedure provided that perturbation of data Y^t asymptotically is not too excessive.

Theorem 6.2. *Let conditions of Theorem 6.1 be satisified but Assumption 1. Assume also that*

$$\begin{aligned} &\widetilde{\Lambda}^t_{b,i} \xrightarrow{c.p.} \Lambda^*_i = a_i^T \lambda_*,\ i = 1, \dots, d,\ \textit{for some } \lambda_* \in \mathbb{R}^p_+ \\ &\textit{when } t \to +\infty,\ \textit{a.s. } Y^t,\ t \in (0, +\infty). \end{aligned} \tag{6.11}$$

Then, formula (6.10) remains valid.

6.3 Asymptotic distribution

Assumption 2. Let $A_{\mathcal{M}}$ be the design matrix defined by formulas (3.23)-(3.26). For $A_{\mathcal{M}}$ the property in (3.27) (injectivity) holds.

Assumption 3 (non-expansiveness condition)**.** Let $\Lambda^* \in \mathbb{R}^d_+$, $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$, $A_{\mathcal{M}}$ has only positive entries and the property in (2.4) holds. Consider $\lambda_{\mathcal{M},*}$ which is defined by the formula:

$$\lambda_{\mathcal{M},*} = \arg\min_{\lambda_{\mathcal{M}} \succeq 0} L(\lambda_{\mathcal{M}} | \Lambda^*, A_{\mathcal{M}}, 1), \tag{6.12}$$

where $L(\lambda_{\mathcal{M}} | \Lambda^*, A_{\mathcal{M}}, 1)$ is defined in (2.8). There is at least one $\lambda_{\mathcal{M},*}$ for which the following holds:

$$I_0(\Lambda^*_{\mathcal{M}}) = I_0(\Lambda^*),\ \Lambda^*_{\mathcal{M}} = A_{\mathcal{M}} \lambda_{\mathcal{M},*}, \tag{6.13}$$

where $I_0(\cdot)$ is defined in (2.2).

Remark 6.1. Note that set of minimizers in (6.12) is always nonempty. From the Karush-Kuhn-Tucker optimality conditions (see e.g., [Ber97], Section 3.3) it follows that

$$\exists (\lambda_{\mathcal{M},*}, \mu_{\mathcal{M},*}) \in \mathbb{R}^p_+ \times \mathbb{R}^p_+ \text{ such that}$$

$$\sum_{i \in I_1(\Lambda^*)} -\Lambda^*_i \frac{a_{\mathcal{M},ij}}{\Lambda^*_{\mathcal{M},i}} + \sum_{i=1}^d a_{\mathcal{M},ij} - \mu_{\mathcal{M},*,j} = 0, \tag{6.14}$$

$$\mu_{\mathcal{M},*,j} \lambda_{\mathcal{M},*,j} \equiv 0,\ \text{for all } j \in \{1, \dots, p_{\mathcal{M}}\}. \tag{6.15}$$

By multiplying both sides of (6.14) on $\lambda_{\mathcal{M},*,j}$, summing all equations with respect to j and using (6.15) we obtain the following necessary optimality condition:

$$\begin{aligned} &\langle \sum_{i=1}^d a_{\mathcal{M},i}, \lambda_{\mathcal{M},*} \rangle = \sum_{j=1}^{p_{\mathcal{M}}} A_{\mathcal{M},j} \lambda_{\mathcal{M},*,j} = \sum_{i=1}^d \Lambda^*_i, \\ &A_{\mathcal{M},j} = \sum_{i=1}^d a_{\mathcal{M},ij}. \end{aligned} \tag{6.16}$$

From (2.4), (3.23) one can see that $A_{\mathcal{M},j} > 0$ for all $j \in \{1, \dots, p_{\mathcal{M}}\}$, hence, the set of constraints in (6.16) is a $(p-1)$ – dimensional simplex which is a convex compact in $\mathbb{R}^p$. This constraint can be added to the set of constraints in (6.12) without any effect since it is necessary. Because the minimized functional in (6.12) is convex and the domain of constraints is now a convex compact there always exists at least one minimizer. □

Remark 6.2. It is always true that

$$I_1(\Lambda^*) \subset I_1(\Lambda^*_{\mathcal{M}}) \text{ or equivalently } I_0(\Lambda^*_{\mathcal{M}}) \subset I_0(\Lambda^*). \tag{6.17}$$

Indeed, if for some i we have $\Lambda^*_i > 0$, then necessarily $\Lambda^*_{\mathcal{M},i} > 0$, otherwise the value of the target functional becomes $+\infty$. □

Remark 6.3. Assumption 3 is named as the non-expansiveness condition because in (6.13) it directly forbids to expand $I_1(\Lambda^*)$ when projecting P^t_{A,λ_*} onto $P^t_{A_\mathcal{M},\lambda_\mathcal{M}}$ in the sense of Kullback-Leibler divergence; see Remark 6.2. In the following paragraph we interpret this condition geometrically and propose an alternative name for it – *the mask condition.* □

MRI data and the mask condition. Below we consider a geometrical interpretation of the non-expansiveness condition based on representation of designs A, $A_\mathcal{M}$ as weighted Radon transforms over the space of discrete images. We show that failure of this condition implies presence of a segment in some $M \in \mathcal{M}$ which is badly aligned with respect to the convex hull of the tracer support. To avoid such situations we propose to preprocess MRI images before using them in the context of ET.

For simplicity, let $k = 1$, i.e., MRI data consists of one segmented image $\mathcal{M} = \{M\}$, and let

$$\Gamma = \{\gamma_i\}_{i=1}^d \text{ be the set of rays available in the acquisition geometry.} \tag{6.18}$$

Assume that $A = (a_{ij})$ is a discretized version of some weighted Radon transform on set of rays Γ with positive weight W. That is

$$a_{ij} = \int_{\gamma_i} W(x, \gamma_i)\, \mathbb{1}_j(x)\, dx,\ \gamma_i \in \Gamma,\ j \in \{1, \dots, p\}, \tag{6.19}$$

$$W = W(x, \gamma),\ (x, \gamma) \in \mathbb{R}^2 \times T\mathbb{S}^1,\ 0 < c \le W \le C, \tag{6.20}$$

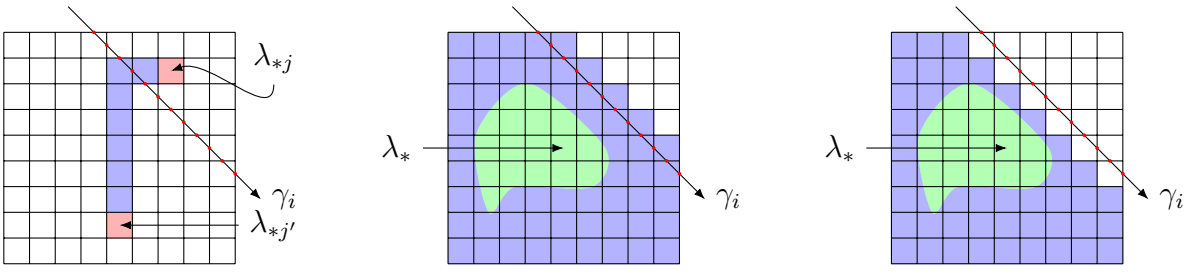
where dx denotes the standard Lebesgue measure on ray γ_i, $\mathbb{1}_j(x)$ is the indicator function of pixel j on the image. Weight $W(x, \gamma)$ is some known sufficiently regular function of spatial coordinates and oriented rays in $\mathbb{R}^2$ which are parameterized by $T\mathbb{S}^1$ (tangent bundle of the unit sphere, see e.g., [Nat01]). Projectors defined by the formulas of type (6.19), (6.20) are common in CT and ET practice; see e.g., [Sid85], [HLY99]. For example, in PET and SPECT weight W is used to model attenuation and nonuniform sensitivity of detectors; see e.g., [Qui83], [Nov19], [Gon19].

From (3.23), (6.19) it follows that

$$A_\mathcal{M} = (a_{M,is}),\ a_{M,is} = \int_{\gamma_i} W(x, \gamma_i) \mathbb{1}_{M,s}(x)\, dx,\ s \in S(M), \tag{6.21}$$

where $\mathbb{1}_{M,s}(x)$ is the indicator function of segment s in image M.

Let $\lambda_* \in \mathbb{R}^p_+$ be the discretized version of the real spatial distribution of the tracer and assume that $\lambda_* \in \mathbb{R}^p_+$ is pixel-wise connected (i.e., between two arbitrary pixels with positive tracer uptake there is a path of pixels preserving the positivity; two pixels are neighboors if they share an edge (see Figure 9)). This assumption is natural, for example, in the context of brain imaging when the tracer is distributed in the whole volume inside the cranium and only relative spatial variations are of practical interest.



(a) (b) $\mathrm{DH}(\lambda_*; \gamma_i)$ (c) $\mathring{\mathrm{DH}}(\lambda_*; \gamma_i)$

Figure 9

Definition 6.4. Let Γ be the finite family of oriented rays in $\mathbb{R}^2$, A be the projector defined by formulas (6.19), (6.20), $\lambda_* \in \mathbb{R}^p_+$, $\lambda_* \neq 0$ and λ_* is pixel-wise connected. Consider $\gamma_i \in \Gamma$ and assume that $i \in I_0(A\lambda_*)$. Then, support of λ_* lies completely in one of the closed half-spaces in $\mathbb{R}^2$ separated from each other with ray γ_i. Let $H(\lambda_*, \gamma_i)$ be such a closed half-space. Consider the discrete version of $H(\lambda_*, \gamma_i)$ defined by the formula

$$\mathrm{DH}(\lambda_*; \gamma_i) = \{j \in \{1, \ldots, p\} | \text{ intersection between pixel } j \text{ and } H(\lambda_*, \gamma_i) \text{ is of non-zero Lebesgue measure on } \mathbb{R}^2\}. \tag{6.22}$$

Consider

$$\mathring{\mathrm{DH}}(\lambda_*; \gamma_i) = \{j \in \mathrm{DH}(\lambda_*, \gamma_i) | \text{ intersection between pixel } j \text{ and ray } \gamma_i \text{ is of length zero}\}. \tag{6.23}$$

Discrete convex hull of λ_* for family Γ is defined by the formula

$$\mathrm{DConv}(\lambda_*; \Gamma, A\lambda_*) = \bigcap_{\substack{\gamma_i \in \Gamma, \\ i \in I_0(\Lambda_*)}} D\mathring{H}(\lambda_*; \gamma_i). \tag{6.24}$$

□

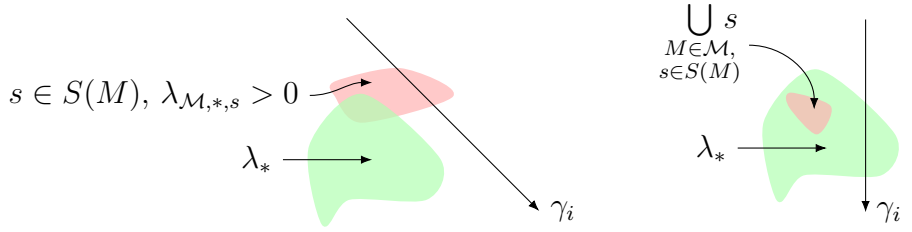
For the geometrical intuition behind definitions $\mathrm{DH}(\cdot)$, $\mathring{\mathrm{DH}}(\cdot)$, $\mathrm{DConv}(\cdot)$, see examples in Figure 9.

Now assume that the non-expansiveness condition fails in the following sense:

$$\text{there exists } i \in I_0(\Lambda^*) \text{ such that } \Lambda^*_{\mathcal{M},i} > 0, \tag{6.25}$$

where $\Lambda^*_{\mathcal{M}}$ is defined in (6.13). From (6.18)-(6.21) and Definition 6.4 it follows that in the image for $\lambda_{\mathcal{M},*}$ there is a segment $s \in S(M)$ which interesected by $\gamma_i \in \Gamma$ and such that $\lambda_{\mathcal{M},*,s} > 0$ (see Figure 10(a)), that is

$$\bigcup_{\substack{M \in \mathcal{M}, \\ s \in S(M), \\ \lambda_{\mathcal{M},*,s} > 0}} s \not\subset \mathrm{DConv}(\lambda_*; \Gamma, \Lambda^*) \tag{6.26}$$



(a) $\Lambda^*_i = 0, \Lambda^*_{\mathcal{M},i} > 0$ (b) $\Lambda^*_i > 0$, $\Lambda^*_{\mathcal{M},i} = 0$

Figure 10

If we assume that $\lambda_{\mathcal{M},*}$ is also pixel-wise connected, then from (6.26) it follows that

$$\mathrm{DConv}(\lambda_{\mathcal{M},*}; \Gamma, A_{\mathcal{M}}\lambda_{\mathcal{M},*}) \not\subset \mathrm{DConv}(\lambda_*; \Gamma, A\lambda_*). \tag{6.27}$$

In total, we have demonstrated the following

Proposition 6.1. *Let $\lambda_* \in \mathbb{R}^p_+$, $\lambda_* \neq 0$, λ_* is pixel-wise connected and designs A, $A_{\mathcal{M}}$ be of type (6.18)-(6.21). Let $\lambda_{\mathcal{M},*}$ be a solution of the minimization problem in (6.12) and $\lambda_{\mathcal{M},*}$ be also pixel-wise connected. Assume that the non-expansiveness condition (Assumption 3) fails in the sense of (6.25). Then, formula (6.27) holds.*

To avoid the situation in Proposition 6.1 one may propose to use a significantly smaller segmentation area, for example, such that

$$\bigcup_{\substack{M\in\mathcal{M},\\ s\in S(M)}} s \subsetneq \mathrm{DConv}(\lambda_*;\Gamma,\Lambda^*), \tag{6.28}$$

where $A \subsetneq B$ denotes the strict inclusion of sets. In this case even a small misalignment may lead to a situation when $\mathcal{KL}(P^t_{A,\lambda_*}, P^t_{A_{\mathcal{M},\lambda_{\mathcal{M}}}}) = +\infty$, so the KL-projection of P^t_{A,λ_*} onto MRI-based model $P^t_{A_{\mathcal{M}},\lambda_{\mathcal{M}}}$ is impossible; see Figure 10(b).

In view of the latter a "good" choice for $S(M)$ would be such that

$$\mathrm{DConv}(\lambda_{\mathcal{M},*};\Gamma,A_{\mathcal{M}}\lambda_{\mathcal{M},*}) = \mathrm{DConv}(\lambda_*;\Gamma,A\lambda_*). \tag{6.29}$$

Note that the above arguments are can be easily extended to the case of $k>1$ by simply checking the alignments for all images in $\mathcal{M}$.

We conclude with a proposition to use the following pipeline for preprocessing anatomical MRI-images:

1. Estimate $\mathrm{DConv}(\lambda_*;\Gamma,A\lambda^*)$ using any well-suited and fast algorithm. Let D be such an estimate.
2. In all MRI-images remove pixels lying outside of D and perform segmentations only on those which are left inside of D.

In view of step 2 we propose an alternative name for Assumption 3 – *the mask condition*.

Theorem 6.3 (identifiability in the prior model)**.** *Let Assumptions 2-3 be satisfied. Then, $\lambda_{\mathcal{M},*}$ defined in (6.12) is unique and the following formula holds:*

$$\begin{aligned} L(\lambda_{\mathcal{M}}|\Lambda^*,A_{\mathcal{M}},1) - L(\lambda_{\mathcal{M},*}|\Lambda^*,A_{\mathcal{M}},1) = \mu^T_{\mathcal{M},*}\lambda_{\mathcal{M}} + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)} \Lambda^*_i \frac{(\Lambda_{\mathcal{M},i}-\Lambda^*_{\mathcal{M},i})^2}{(\Lambda^*_{\mathcal{M},i})^2} \\ + o(\|\Pi_{A^T_{\mathcal{M},I_1(\Lambda^*)}}(\lambda_{\mathcal{M}}-\lambda_{\mathcal{M},*})\|^2), \end{aligned} \tag{6.30}$$

where $\Pi_{A^T_{\mathcal{M},I_1(\Lambda^*)}}$ denotes the orthogonal projector onto $\mathrm{Span}(A^T_{\mathcal{M},I_1(\Lambda^*)})$,

$$\begin{aligned} &\mu_{\mathcal{M},*} = \sum_{i\in I_1(\Lambda^*)} -\Lambda^*_i \frac{a_{\mathcal{M},i}}{\Lambda^*_{\mathcal{M},i}} + \sum_{i=1}^{d} a_{\mathcal{M},i}, \\ &\mu_{\mathcal{M},*} \succeq 0,\ \mu_{\mathcal{M},*,j}\lambda_{\mathcal{M},*,j} = 0 \text{ for all } j\in\{1,\dots,p_{\mathcal{M}}\}. \end{aligned} \tag{6.31}$$

In particular, there exists an open ball $B_* = B(\lambda_{\mathcal{M},*},\delta_*)$, $\delta_* = \delta_*(A_{\mathcal{M}},\Lambda_*) > 0$ and constant $C_* = C_*(A_{\mathcal{M}},\Lambda_*) > 0$ such that

$$L(\lambda_{\mathcal{M}}|\Lambda^*,A_{\mathcal{M}},1) - L(\lambda_{\mathcal{M},*}|\Lambda^*,A_{\mathcal{M}},1) \geq C_*\|\lambda_{\mathcal{M}}-\lambda_{\mathcal{M},*}\|^2 \text{ for any } \lambda\in B_*\cap\mathbb{R}^{p_{\mathcal{M}}}_+. \tag{6.32}$$

Result of Theorem 6.3 is a positive answer to the identification problem when model (2.1) is misspecified in the sense of wrong design. The non-expansiveness condition is essential and counterexamples are possible if it is removed. One such example is constructed in the proof of Theorem 6.6 in Subsection 6.4.

Theorem 6.4 (concentration rate for the mixing parameter)**.** *Let Assumptions 1-3 be satisfied. Let $\lambda^t_{\mathcal{M}}$ be sampled as in Algorithm 4 and $r(t) = o(\sqrt{t/\log\log t})$. Then,*

$$r(t)(\lambda^t_{\mathcal{M}} - \lambda_{\mathcal{M},*}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \quad a.s.\ Y^t, t \in (0, +\infty), \tag{6.33}$$

where $\lambda_{\mathcal{M},*}$ is from Theorem 6.3. In particular, for parameter $\Lambda^t_{\mathcal{M}} = A_{\mathcal{M}}\lambda^t_{\mathcal{M}}$ from Algorithm 4 formula (6.33) implies that

$$r(t)(\Lambda^t_{\mathcal{M}} - \Lambda^*_{\mathcal{M}}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \quad a.s.\ Y^t, t \in (0, +\infty), \tag{6.34}$$

where $\Lambda^*_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M},*}$.

Remark 6.4. The log-factor for $r(t)$ in Theorem 6.4 is necessary for the "almost sure" character of formula (6.34) and, in particular, it is due to the Law of the Iterated Logarithm for trajectory Y^t (see Appendix B). For our purposes it is sufficient to have the result for rate $r(t) = o(\sqrt{t/\log\log t})$ because $\Lambda^t_{\mathcal{M}}$ is used in the prior whose effect disappears in view of the well-known Bernstein von-Mises phenomenon for bayesian posteriors; see, e.g. Section 10.2 in [VdV00] and [KVdV12]. □

Let $\{e_j\}_{j=1}^p$ be the standard basis in $\mathbb{R}^p$ and define the following spaces:

$$\mathcal{V} = \mathrm{Span}\{e_j \,|\, \exists\, i \in I_0(\Lambda^*) \text{ s.t. } a_{ij} > 0\}, \tag{6.35}$$

$$\mathcal{U} = \mathcal{V}^\perp \cap \mathrm{Span}\{A^T_{I_1(\Lambda^*)}\}, \tag{6.36}$$

$$\mathcal{W} = (\mathcal{V} \oplus \mathcal{U})^\perp \cap \ker A. \tag{6.37}$$

Subspace $\mathcal{V}$ is the span of pixels which correspond to LORs with zero intensities for the true value of the parameter. Note that positivity constraint on the parameter of interest gives us additional information that all such pixels must have zero tracer uptake inside and one could expect the fastest estimation rate for λ on $\mathcal{V}$. Subspace $\mathcal{U}$ corresponds to pixels intersected by LORs with positive intensities and which signal can be recovered, in general, using design matrix A (after having removed pixels from $\mathcal{V}$ where the estimation rates are faster). Finally, subspace $\mathcal{W}$ corresponds to parts of λ which cannot be reconstructed neither by positivity constraints nor using design A (note that $\mathcal{W} \subset \ker A$) and which, in turn, are defined completely by the regularization penalty $\varphi(\cdot)$ at λ_* and by $\ker A$.

Let

$$\Pi_{\mathcal{V}}, \Pi_{\mathcal{V}}, \Pi_{\mathcal{W}} \text{ be the orthogonal projectors on } \mathcal{V}, \mathcal{V}, \mathcal{W}, \text{ respectively.} \tag{6.38}$$

Theorem 6.5 (tightness of the asymptotic distribution)**.** *Let assumptions 1-3 be satisfied and assume also that*

$$\varphi \text{ satisfies (2.15), (2.16) and } \varphi \text{ is locally Lipschitz continous.} \tag{6.39}$$

Let $\widetilde{\lambda}^t_b$ be defined as in Algorithm 5 and $\theta^t = o(\sqrt{t/\log\log t})$, $\beta^t = o(\sqrt{t})$ and assume that there exists a strongly consistent estimator $\widehat{\lambda}^t_{sc}$ of λ_* on $\mathcal{V}\oplus\mathcal{U}$ (i.e., $\Pi_{\mathcal{U}\oplus\mathcal{V}}\widehat{\lambda}^t_{sc} \xrightarrow{a.s.} \Pi_{\mathcal{U}\oplus\mathcal{V}}\lambda_*$) such that

$$\widehat{\lambda}^t_{sc} \succeq 0, \tag{6.40}$$

$$\limsup_{t\to+\infty}\left|\sum_{i\in I_1(\Lambda^*)}\sqrt{t}\cdot\frac{Y_i^t/t-\widehat{\Lambda}_{sc,i}^t}{\widehat{\Lambda}_{sc,i}^t}\cdot a_i\right|<+\infty\ \ a.s.\ Y^t,\ t\in(0,+\infty),\tag{6.41}$$

$$t\widehat{\Lambda}_{sc,i}^t\to 0\ \ when\ t\to+\infty,\ \ a.s.\ Y^t,\ t\in(0,+\infty)\ \ for\ i\in I_0(\Lambda^*),\tag{6.42}$$

where $\widehat{\Lambda}_{sc}^t=A\widehat{\lambda}_{sc}^t$. *Then,*

i)

$$t\cdot\Pi_{\mathcal{V}}(\widetilde{\lambda}_b^t-\widehat{\lambda}_{sc}^t)\xrightarrow{c.p.}0\ \ when\ t\to+\infty,\ \ a.s.\ Y^t,\ t\in(0,+\infty).\tag{6.43}$$

ii) Vector $\sqrt{t}\cdot\Pi_{\mathcal{U}}(\widetilde{\lambda}_b^t-\widehat{\lambda}_{sc}^t)$ *is conditinally tight a.s.* Y^t, $t\in(0,+\infty)$.

Statement in (i) claims that in pixels which are interested by LORs with zero intensities the posterior distribution contracts to zero with faster rate than for the ones intersected by LORs with positive intensities. Similar splitting for bayesian posteriors was already investigated in [BG14], where it was established that $t\cdot\Pi_{\mathcal{V}}\lambda|Y^t,t$ asymptotically has exponential distribution. In view of this, the result in (6.43) is not surprising since samples $\widetilde{\lambda}_b^t$ are essentially MAP-estimates (for perturbed data) and the maximizer of exponential distribution is exactly zero. Statement in (ii) claims that, in general, the posterior concentrates around $\widehat{\lambda}_{sc}^t$ in subspace $\mathcal{U}$ with standard rate $\sim\sqrt{t}$.

For $\widehat{\lambda}_{sc}^t$ we propose to take the penalized MLE-estimate which is defined by the formula:

$$\widehat{\lambda}_{pMLE}^t=\arg\min_{\lambda\succeq 0}L_p(\lambda|Y^t,A,t,\beta^t),\tag{6.44}$$

where $L_p(\cdot)$ is defined in (2.10). Note that $\widehat{\lambda}_{pMLE}^t$ can be efficiently computed using the GEM-type algorithm from [FH95], which we also use in Algorithms 3-5.

Conjecture 1. *Let assumptions of Theorem 6.5 be satisfied and* $\widehat{\lambda}_{sc}^t=\widehat{\lambda}_{pMLE}^t$, *where the latter is defined by formula* (6.44). *Then,* $\widehat{\lambda}_{sc}^t$ *is a strongly consistent estimator of* λ_* *on* $\mathcal{V}\oplus\mathcal{V}$ *and formulas* (6.40)-(6.42) *hold.*

The requirement for existence of a strongly consistent estimator for characterization of weighted bootstrap is not new and already appears in [NN20]. However, in that case the sampling is performed via unconstrained optimization of quadratic functionals with ℓ_1-penalties for which existence of such estimators is trivial by taking the standard OLS estimator or LASSO estimator; see the discussion after Theorem 3.3 in [NN20]. Conditions (6.40)-(6.42) are somehow analogous to the ones in the aforementioned work. Note also that according to Kolmogorov's 0-1 Law the inequality in (6.41) either holds with probability one (i.e., almost surely Y^t, $t\in(0,+\infty)$) or zero, and the case of zero probability would mean a very exotic and unexpected behavior of the constrained MLE estimate for such model because conditions (6.40)-(6.42) are trivially satisfied, for example, if A is a diagonal matrix.

Proving Conjecture 1 and further investigation of possible $\widehat{\lambda}_{sc}^t$ are outside of the scope of this work and will be given in future. To our knowledge this is a completely new open problem and such result is necessary for further investigation of bootstrap procedures for the model of ET.

Remark 6.5. Centering the distribution of $\widetilde{\lambda}_b^t$ at the true parameter λ_* in (ii) does not allow to achieve conditional tightness almost surely Y^t, $t\in(0,+\infty)$ which we briefly explain below.

As a part of the proof of Theorem 6.5 (see also lemmas 8.8, 8.9) we show that

$$\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_b - \widehat{\lambda}^t_{sc}) - u^t(\widetilde{\xi}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty), \tag{6.45}$$

where

$$u^t(\xi) = \underset{\substack{u:(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0,\\ u\in\mathcal{U},\, w\in\mathcal{W}}}{\arg\min} -u^T(A_{I_1(\Lambda^*)})^T(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\xi + \frac{1}{2}u^T\widehat{F}^t_{I_1(\Lambda^*)}u, \tag{6.46}$$

$$\widehat{D}^t_{I_1(\Lambda^*)} = \operatorname{diag}(\dots, \widehat{\Lambda}^t_{sc,i}, \dots),\ i \in I_1(\Lambda^*), \tag{6.47}$$

$$\widehat{F}^t_{I_1(\Lambda^*)} = \sum_{i\in I_1(\Lambda^*)} \frac{a_i a_i^T}{\widehat{\Lambda}^t_{sc,i}} = (A_{I_1(\Lambda^*)})^T(\widehat{D}^t_{I_1(\Lambda^*)})^{-1}A_{I_1(\Lambda^*)}, \tag{6.48}$$

$$\xi \in \mathbb{R}^{\#I_1(\Lambda^*)},\ \widetilde{\xi}^t = (\dots, \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}}{\sqrt{\widehat{\Lambda}^t_{sc,i}}}, \dots),\ i \in I_1(\Lambda^*). \tag{6.49}$$

That is the conditional tightness (and also asymptotic distribution) of $\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_b - \widehat{\lambda}^t_{sc})$ asymptotically coincides with the one of $u^t(\widetilde{\xi}^t)$, where the latter is the minimizer of a quadratic function on a polyhedral set depending on $\widehat{\lambda}^t_{sc}$. In the proof we show that conditional tightness is implied by tightness of $\widetilde{\xi}^t$ (this is especially obvious if the constraints in (6.46) are not active for large t, e.g., when $\lambda_* \succ 0$) and that under the assumptions of the theorem it holds that

$$(\dots, \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,i} - \frac{Y_i}{t}}{\sqrt{\widehat{\Lambda}^t_{sc,i}}}, \dots) \xrightarrow{c.d.} \mathcal{N}(0, I) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty),$$
$$I - \text{identity matrix of size } \#I_1(\Lambda^*) \times \#I_1(\Lambda^*). \tag{6.50}$$

From (6.46)-(6.50) and the Prohorov's theorem on tightness of weakly convergent sequences or r.v.s, the asymptotic behavior (tightness, distribution) of $u^t(\widetilde{\xi}^t)$ is essentially depends on the term $(\dots, \sqrt{t}\cdot\frac{\widehat{\Lambda}^t_{sc,i} - \frac{Y_i}{t}}{\sqrt{\widehat{\Lambda}^t_{sc,i}}}, \dots)$, $i \in I_1(\Lambda^*)$. For tightness this term needs to be asymptotically bounded for almost any trajectory Y^t, $t \in (0, +\infty)$, which is exactly asked in (6.41) (in a slightly weakened form).

Now, if we center $\widetilde{\lambda}^t_b$ on λ_* one finds that $\widehat{\lambda}^t_{sc}$ must be replaced everywhere with λ_* in formulas (6.46)-(6.50) and, most importantly, the latter term is equal $(\dots, \frac{Y^t_i - t\Lambda^*_i}{\sqrt{t\Lambda^*_i}}, \dots)$ which is asymptotically standard normal (by CLT; see Section B in Appendix). Therefore, the mean of the asymptotic distribution of $\sqrt{t}\cdot\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_b - \lambda_*)$ depends on the trajectory of $(Y^t_i - t\Lambda^*_i)/\sqrt{t\Lambda^*_i}$, $i \in I_1(\Lambda^*)$, which is almost surely unbounded infinitely often on $t \in (0, +\infty)$ in view of the Law of Iterated Logarithm for Y^t (see formula (B.4) in Appendix). So the tightness for $\sqrt{t}\cdot\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_b - \lambda_*)$ almost surely for any trajectory Y^t, $t \in (0, +\infty)$ is impossible. A very similar behavior for centering of the posterior distribution for weighted bootstrap was also ovserved in Theorem 3.3 from [NN20].

We also note that (6.45)-(6.50) is a key to establish further asymptotic normality of the posterior, however, for this one needs a separate investigation of behavior of $\widehat{\lambda}^t_{sc}$ near $\partial\mathbb{R}^p_+$.

6.4 Misspecification in design and identifiability

Assumption 1 in Subsection 6.2 reflects our belief that model (2.1) is not misspecified (at least asymptotically): observed data Y^t has distribution $P^t_{A,\lambda}$ for some parameter $\lambda = \lambda_*$

and known design A. At the same time, for any practitioner in ET it is known that such model is by far approximate. For example, the tracer inside the human body surely does not respect locally constant behavior in each pixel on which our discretized model is based. And even if we assume a situation when the discretization is correct, in practice, matrix A is known only approximately, with non-negligible errors, since it contains patient's attenuation map which reconstructed via a separate MRI or CT scan; see e.g., [SC13]. There also are many other practical issues which are not included in (2.1), e.g., non-stationarity of the process due to kinetics for the tracer, scattered photons, electronic noise in detectors, errors from multiple events etc.; see e.g., [LDH95], [RTZ09] [Lew10].

Assuming temporal stationarity of the process we consider the following scenario for ET[2]:

$$Y^t \sim P^t,\ Y^t \in (\mathbb{N}_0)^d, \tag{6.51}$$

$$\mathbb{E}_{P^t}[Y^t] = \mathrm{Var}_{P^t}[Y] = t\Lambda^* \text{ for some } \Lambda^* = (\Lambda^*_1, \dots, \Lambda^*_d) \in \mathbb{R}^d_+,\ t \in \mathbb{R}_+, \tag{6.52}$$

Formulas (6.51), (6.52) reflect our belief that Y^t has Poisson-type behavior at least for its two first moments which is not far from truth in practice [SC15]. Most importantly, we do not assume that $\Lambda^* \in R_+(A)$.

The main question now is the identifiability of λ which translates to the problem of uniqueness in the following minimization problem:

$$\lambda_{A,*}(P^t) = \arg\min_{\lambda \succeq 0} \mathcal{KL}(P^t, P^t_{A,\lambda}), \tag{6.53}$$

where $P^t_{A,\lambda}$ is defined in (2.8).

Using formulas (2.9), (6.51), (6.52), formula (6.53) can be rewritten as follows:

$$\lambda_{A,*}(P^t) = \arg\min_{\lambda \succeq 0} L(\lambda|\Lambda^*, A, 1). \tag{6.54}$$

The following result gives the negative answer to the identifiability problem for general $\Lambda^* \in \mathbb{R}^d_+$, even if design A is satisfies (2.3)-(2.5), it is stochastic column-wise and injective.

Theorem 6.6. *There exist $\Lambda^* = (\Lambda^*_1, \dots, \Lambda^*_d) \in \mathbb{R}^d_+$, $\Lambda^* \neq 0$, $A \in \mathrm{Mat}(d, p)$ which has only nonnegative entries, it is stochastic column-wise (that is $\sum_{i=1}^d a_{ij} = 1$ for all j) and injective such that solutions of the optimization problem (6.54) constitute a non-empty affine subset of the $(p-1)$-simplex $\Delta_p(\Lambda^*)$ defined by the formula:*

$$\Delta_p(\Lambda^*) = \left\{ \lambda \in \mathbb{R}^p_+ : \sum_{j=1}^p \lambda_j = \sum_{i=1}^d \Lambda^*_i \right\}. \tag{6.55}$$

Proof of Theorem 6.6. We construct Λ_* and A for $p = 4$, $d = 6$.

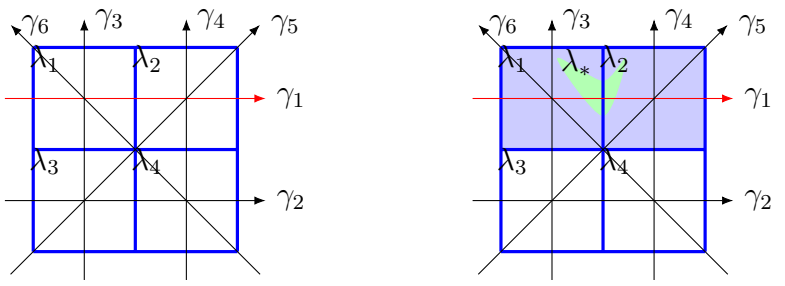


Figure 11

[2]Temporal stationarity of P^t is equivalent to neglecting the kinetic evolution of the tracer which is always assumed in classical ET scenario; see e.g., [SV82].

Let $\mathcal{I}$ be the image consisting of four square pixels each with side length equal to 1 as shown in Figure 11(a), i.e., $\lambda = (\lambda_1, \dots, \lambda_4) \in \mathbb{R}^4_+$. Let $\Gamma = \{\gamma_1, \dots, \gamma_6\}$ be the family of rays as it is shown in the Figure 11(a) and matrix A' corresponds to the the classical Radon transform on $\mathcal{I}$:

$$a'_{ij} = \text{length of intersection of ray } \gamma_i \in \Gamma \text{ with pixel } j, \tag{6.56}$$

$$A' = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \end{pmatrix}. \tag{6.57}$$

Moreover,

$$\det(A'^T A') = 128 \neq 0, \tag{6.58}$$

so A' is injective. Let A be a normalization of A' with respect to columns such that A is stochastic column-wise , i.e., $a_{ij} = a'_{ij}/(\sum_i a'_{ij})$. Such normalization obviously does not break the injectivity of A'.

Let $\Lambda^* = (1, 0, 0, 0, 0, 0)$. Then, the optimization problem in (6.54) has the following form:

$$\lambda_{A,*} = \arg\min_{\lambda \succeq 0} -\log\left(\frac{\lambda_1 + \lambda_2}{2 + \sqrt{2}}\right) + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4. \tag{6.59}$$

Note that in (6.59) we have used the fact that $\sum_i a_{ij} = 1$ for all $j \in \{1, \dots, 4\}$. It is obvious that the set of minimizers in (6.59) is an affine set of the following form:

$$\lambda_{A,*3} = \lambda_{A,*4} = 0,\ \lambda_{A,*1} + \lambda_{A,*2} = 1. \tag{6.60}$$

Note that the property that $\lambda_* \in \Delta_p(\Lambda^*)$ follows from KKT optimality conditions for the problem in (6.54); see Remark 6.1.

Theorem 6.6 is proved. □

Remark 6.6. Constructed pair (Λ^*, A) in the proof of Theorem 6.6 is meaningful even from the physical point of view. Indeed, intensity vector $\Lambda^* = (1, 0, 0, 0, 0, 0)$ can be observed, for example, if in pixel $j = 1$ there is a region of subpixel size containing the isotope and is being intersected only by γ_1; see Figure 11(b). Note also that for the constructed family of minimizers the non-expansiveness condition is not satisfied: indeed, any choice of $\lambda_{A,*1}, \lambda_{A,*2}, \lambda_{A,*3}, \lambda_{A,*4}$ in (6.60) results in at least two rays with positive projected intensities. □

As it was shown in Subsection 6.3, for the well-specified case the non-expansiveness condition has a meaningful geometrical interpretation. But when the distribution of Y^t is completely unknown, a similar geometrical interpretation of this assumption is not straightforward – at least the non-expansiveness condition needs to be investigated further if it can be relaxed.

Definition 6.5 (generic non-expansiveness condition)**.** We say that $\Lambda^* \in \mathbb{R}^d_+$ is *non-expansive* for design $A \in \mathrm{Mat}(d, p)$ having nonnegative entries, if there is at least one minimizer $\lambda_{A,*}$ in (6.54) for which the following holds:

$$I_0(\Lambda^*) = I_0(\Lambda^*_A),\ \Lambda^*_A = A\lambda_{A,*}, \tag{6.61}$$

where $I_0(\cdot)$ is defined in (2.2).

The following result states that identification problem in (6.54) has positive answer for injective design if the generic non-expansiveness condition holds.

Theorem 6.7. *Let $\Lambda^* \in \mathbb{R}^d_+$ be non-expansive for design $A \in \mathrm{Mat}(d,p)$ which has non-negative entries and is also injective. Then, $\lambda_{A,*}$ is a unique minimizer in (6.54) and the following formula holds:*

$$L(\lambda|\Lambda^*,A,1) - L(\lambda_{A,*}|\Lambda^*,A,1) = \mu_{A,*}^T\lambda + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)}\Lambda_i^*\frac{(\Lambda_i-\Lambda^*_{A,i})^2}{(\Lambda^*_{A,i})^2} + o(\|\Pi_{A^T_{I_1(\Lambda^*)}}(\lambda-\lambda_{A,*})\|^2), \tag{6.62}$$

where $\Pi_{A^T_{I_1(\Lambda^*)}}$ denotes the orthogonal projector onto $\mathrm{Span}\{a_i|i\in I_1(\Lambda^*)\}$,

$$\mu_{A,*} = \sum_{i\in I_1(\Lambda^*)} -\Lambda_i^*\frac{a_i}{\Lambda^*_{A,i}} + \sum_{i=1}^d a_i, \quad \mu_{A,*} \succeq 0,\ \mu_{A,*,j}\lambda_{A,*,j} = 0 \text{ for all } j\in\{1,\dots,p\}. \tag{6.63}$$

In particular, there exists an open ball $B_* = B(\lambda_{A,*},\delta_*)$, $\delta_* = \delta_*(A,\Lambda_*) > 0$ and positive constant $C_* = (A,\Lambda_*)$ such that

$$L(\lambda|\Lambda^*,A,1) - L(\lambda_{A,*}|\Lambda^*,A,1) \geq C_*\|\lambda-\lambda_{A,*}\|^2 \text{ for any } \lambda\in B_*\cap\mathbb{R}^p_+. \tag{6.64}$$

If the claim by Theorem 6.7 holds, and in particular (6.64), then it is straightforward to extend theorems 6.1, 6.2 for the misspecified scenario given by (6.51), (6.52) and also when

$$\frac{Y_i^t}{t} \xrightarrow{a.s.\ P^t} \Lambda_i^*, \text{ for } t\to+\infty, \text{ and } Y^t\equiv 0 \text{ if } \Lambda_i^* = 0. \tag{6.65}$$

Requiring the injectivity for A may seem to restrictive and in fact is not necessary to obtain such extensions since the result of Theorem 6.7 can be refined for noninjective A by demanding only uniqueness and strong convexity in directions from $\mathrm{Span}(A^T)$. This work will be done elsewhere.

However, the most interesting thing would be to extend the result of Theorem 6.5 and also demonstrate asymptotic normality of the posterior projected on $\mathcal{U}$ for the fully misspecified case. From the result of [Pom21] we know that for $\theta^t = 0$ (i.e., Algorithm 3) our method coincides with weighted likelihood bootstrap with prior information induced by penalty $\varphi(\cdot)$ which was proven to provide robust estimates of credible intervals in regular statistical models under misspecification. Though our model is nonregular this seems not to be a fundamental issue for the robustness to hold also in our case. At the same time in [LHW19], for regular models and using nonparametric Dirichlet process priors, it was shown that in Algorithm 5 for $\beta^t = 0$ the robustness for credible intervals also holds under misspecification. So the intuitive conclusion is that Algorithms 3, 5 will also correct the misspecification (at least up to the first order of Edgeworth's expansion) in asymptotic credible intervals in comparison with the classical bayesian approach which gives suboptimal estimates [KVdV12]. Explicit extension of our results following [Pom21], [LHW19] to the nonregular case of ET is left for future work.

7 Discussion

We have proposed an extension of the nonparametric posterior learning to generalized Poisson models which are used for PET/SPECT. The main idea of the extension is that

the uncertainty on the parameter of interest (tracer concentration) propagates from the uncertainty on the true point process that generates the observed data (photon counts). Then, the sampling procedure is defined via minimization of the constrained penalized Kullback-Leibler divergence between the candidate point process from the nonparametric posterior and the chosen model for observed data (spatial temporal stationary Poisson point processes with fixed design).

By using additional MRI images for constructing nonparametric prior on the set of spatial temporal stationary Poisson point processes we regularize the inverse problem of ET. In particular, used MRI-images are presegmented and in the prior it is assumed that tracer concentration has locally constant values in each segment. To sample photon counts from the prior model we construct new design matrix from the original high-dimensional one by collapsing columns that correspond to pixels sharing same segments. Low number of large segments on MRI-images while keeping the same acquisition geometry makes the new design matrix injective and not ill-conditioned as in the original problem. This is particularly important for regularization when we add the effect of MRI data on our samples: there is unique correspondance between sinogram from pseudo-photons and the emission map in the prior. Overall, the new MRI-based model is used further to sample pseudo-sinograms and mix them with real observations.

It is assumed that the unknown generating process belongs to a family of temporal stationary Poisson point processes on the manifold of LORs, so the family is parameterized by function assigning intensity of the photon flow along each LOR in the acquisition geometry. Hence, the nonparametric prior is constructed on the set of these functions. Defining prior on observed intensities in the LOR-space but not on the image-space allows to address the problem of misspecification with the wrong design and circumvent the use of data-augmentation schemes which, as we have clearly shown, lead to high correlations between samples (i.e., mixing problems for MCMC). The cost to be payed is that the prior is defined now on a larger space and no structural information is used here (e.g. that observed intensities belong to the image space of the system matrix of the scanner which is not too far from the chosen design). Therefore, more data may be needed to contract the posterior on the image of the forward operator. The only help here is that prior, as explained further, is centered on intensities that would be observed from the MRI-model with reduced design matrix.

In particular, for the nonparametric prior on intensity function for LORs we choose a mixture of gamma processes. In view of the conjugacy between Poisson and gamma processes, the resulting posterior is again a mixture of gamma processes but with updated parameters and, in particular, with only one free parameter ρ, $\rho \geq 0$, which pleasantly appears to be interpretable as the ratio between the total number of detected photons and the number of pseudo-photons generated from the MRI-based model ($\rho/(1+\rho)$ for photons are taken from the prior, $1/(1+\rho)$ for photons from the sinogam). For example, for $\rho = 0$ no MRI data are used in sampling and the algorithm corresponds to classical weighted likelihood bootstrap (WLB), whereas $\rho > 0$ corresponds to NPL for ET with side information from MRI.

An important and final remark on the design of the new sampler would be that one may think of more involved tracer models constructed from MRI data, e.g. using side knowledge of correlations between PET and MRI signals or, at least, learning them (e.g., via Neural Networks or other machine learning techniques). Also, for the gamma processes one could use some additional structural information for the scale parameter (e.g., coming from the knowledge of design A), so the posterior will concentrate faster on the image space of the real system matrix of the scanner. In our algorithm, the scale parameter was chosen to be constant on the whole manifold of LORs (equals $(\theta^t)^{-1} \cdot 1$; see formula (3.12)) which is clearly not optimal, and an obvious extension would be to choose θ^t not being a

scalar but vector $\theta^t = (\theta_1^t, \dots, \theta_d^t)$, where θ_i^t controls the fraction of photons retrieved from each detector channel $i \in \{1, \dots, d\}$. Of course, more involved extensions are possible for the list-mode data. To conclude, the important requirement in all the above steps is that in the end one must to be able to sample efficiently from the resulting nonparametric posterior since problems of high-dimensionality and ill-posedness are crucial.

New obtained sampling algorithms are scalable, trivially parallelizable and very easy to implement since they rely only on well-known EM-type algorithms in ET. Important advantage of NPL in ET is the circumvent of poor mixing in bayesian MCMC when using one very popular data augmentation scheme. In particular, the problem of poor mixing in MCMC for bayesian posteriors was the main motivation for this work and we have demonstrated and explained the phenomenon theoretically and numerically on one very practical example of Gibbs sampling for PET.

Theoretical studies of new algorithms demonstrate consistency of the posterior in the span of the design matrix and regarding the positivity constraints if the model is correctly specified. Analysis of the asymptotic distribution is more involved and we have several new findings here.

First, it is necessary that prior constructed using MRI data and the consequent posterior do not produce excessive amount of pseudo-photons (or, equivalently, assign high intensities) in LORs where the true intensity is zero (and consequently no detected photons). This is translated to the non-expansiveness condition which can be interpreted in terms of applying masks to segments of MRI images before to use it (see Section 6.3 for details). From practical point of view, here one needs to have a good algorithm for estimation of the convex support of the true Poisson point process (also with constraints); see e.g., [BR16]. From statistical point of view this condition (or the generalized non-expansiveness condition in Section 6.4) is a sufficient one for the identification to hold in the misspecified generalized Poisson models with wrong design. To check it in practice one would need side information on the convex hull of the support of the tracer, which could be obtained, for example from some medical expertise. In view of this, it is also of independent interest to consider the following relaxation of the classical inverse problem in ET:

Problem for ET with side information on support. Let $Y^t \sim P^t_{A,\lambda_*}$, where A is known and λ_* is not and needs to be estimated. Assume that $\mathrm{DConv}(\lambda_*; \Gamma, \Lambda^*)$ is known for given Γ (shortly, convex hull of the support of the tracer; see Definition 6.4). How this additional information decreases the ill-posedness of the inverse problem in ET?

Second, the asymptotic posterior concentrates around a strongly consistent estimator $\widehat{\lambda}^t_{sc}$ in the subspace $\mathcal{V} \oplus \mathcal{U}$ satisfying additional constraints (see formulas (6.35), (6.36) and Remark 3) which is also the case for bayesian posteriors (see [BG14]; Remark 6.5). Such behavior is known for weighted bootstrap for regular models and their slight extensions with non-smooth penalties; see [NR94], [NN20].

Third, the nonregularity of the model results in splitting of the posterior distribution in three different modes and the splitting is essentially defined by the behavior of $\widehat{\lambda}^t_{sc}$ near $\partial\mathbb{R}^p_+$. Because of the lack of results on $\widehat{\lambda}^t_{sc}$ (and on penalized MLE, in particular) we fail to demonstrate the asymptotic normality almost for any trajectory Y^t, $t \in (0, +\infty)$.

The general conclusion here is that, surprisingly, little work (if any) has been done to study strongly consistent estimators under constraints (such as MLE or penalized MLE) in Poisson models for ET. It appears that such results are necessary for analysis of asymptotic distributions for NPL or weighted bootstrap and in general, and this is a completely open problem to our knowledge. For example, extension of the results from [Gey94] for bootstrap-type procedures with constraints is very needed in view of their very recent and

active studies (for the case of regular models see [NR20], [NN20], [Pom21]).

The problem of identifiability in the misspecified generalized Poisson model with wrong design is the last being considered here, where we show that the (generalized) non-expansiveness condition is essential and counterexamples are possible if it is removed. Nevertheless, this condition is only sufficient one and does not give precise restrictions on Λ^* for fixed design A. Therefore, it is of interest to find a necessary condition which we adress as the following problem:

Necessary condition for identifiability. Let $A \in \mathrm{Mat}(d, p)$ be injective, with non-negative entries and stochastic column-wise (e.g. discretized Radon-type transform). Under which restrictions on $\Lambda^* = (\Lambda_1^*, \ldots, \Lambda_d^*) \in \mathbb{R}_+^d$, the minimization problem in (6.54) has a unique solution $\lambda_* \in \mathbb{R}_+^p$ and functional $L(\lambda; \Lambda^*, A, 1)$ is locally strongly convex at its global minimizer?

There is also a question of extending the obtained results in Subsections 6.2, 6.3 for fully misspecified scenario when $P^t \neq P^t_{A,\lambda}$. An extensive discussion was given in the end of Subsection 6.4.

Finally, numerical tests of our algorithms were performed only on synthetic data. Future work will be focused to perform tests on real patient data.

8 Proofs

8.1 Proof of Lemma 2.1

Proof. For the proof we use the two following lemmas.

Lemma 8.1. *Let $\lambda \in \mathbb{R}_+^p$ and A satisfies (2.3), (2.4). Then, for any compact $U \subset \mathrm{Span}(A^T)$ it holds that*

$$S_{A,\lambda}(U) = (\lambda + U + \ker A) \cap \mathbb{R}_+^p \text{ is convex and compact,} \tag{8.1}$$

where the summation sign denotes the Minkowski sum

$$A + B = \{w = u + v \subset \mathbb{R}^p : u \in A,\, v \in B\},\ A \subset \mathbb{R}^p,\ B \subset \mathbb{R}^p. \tag{8.2}$$

Lemma 8.2. *Let assumptions of Lemma 8.1 be satisfied and $d_H(A, B)$ denote the Hausdorff distance between compact sets $A, B \subset \mathbb{R}^p$*

$$d_H(A, B) = \max\left(\sup_{x \in A} \inf_{y \in B} \|x - y\|, \sup_{x \in B} \inf_{y \in A} \|x - y\|\right). \tag{8.3}$$

Then,

$$d_H(S_{A,\lambda_*}(\{u_0\}), S_{A,\lambda_*}(\{u\})) \to 0 \text{ for } u \to u_0,\ u, u_0 \in U, \tag{8.4}$$

where $S_{A,\lambda_*}(\cdot)$ is defined in (8.1).

From the result of Lemma 8.1 and from assumption in (2.16) it follows that for each $u \in U$ the following optimization problem

$$\begin{aligned} &\text{minimize } \varphi(\lambda + u + w) \text{ w.r.t } w, \\ &\text{subject to: } \lambda + u + w \succeq 0,\ w \in \ker A. \end{aligned} \tag{8.5}$$

is a minimization problem of a strictly convex function in w on a convex compact domain. Therefore, there is always exists a unique minimizer $w(u) \in \ker A$. This proves the first assertion of the lemma.

Now, we prove the continuity of $w(u)$ on its domain of definition. Let u_k be a sequence in U such that $u_k \to u_0$ for some $u_0 \in U_0$. Let $w_k = w(u_k)$, where the latter are minimizers in (8.5) for $u = u_k$, and $w_0 = w(u_0)$. We know that $\lambda_k = \lambda + u_k + w(u_k) \in S_{A,\lambda}(U)$, where the latter is a compact by Lemma 8.1. Then all w_k belong to compact $W_{A,\lambda}(U)$ which is the orthogonal projection of $(S_{A,\lambda}(U) - \lambda)$ onto $\ker A$ (image of a compact by continuous function is compact). From compactness of $W_{A,\lambda}(U)$ it follows that w_k contains a converging subsequence $w_m \to w_*$, $w_* \in W_{A,\lambda}(U)$, where $w_m = w(u_m)$.

Since w_m are minimizers in (8.5) we know that

$$\begin{aligned} &\varphi(\lambda + u_m + w_m) \le \varphi(\lambda + u_m + w), \\ &\text{for all } w \in \ker A, \text{ s.t. } \lambda + u_m + w \succeq 0. \end{aligned} \tag{8.6}$$

Taking the limit $m \to +\infty$, $u_m \to u_0$, $w_m \to w_*$ we aim to show that

$$\begin{aligned} &\varphi(\lambda + u_0 + w_*) \le \varphi(\lambda + u_0 + w), \\ &\text{for all } w \in \ker A, \text{ s.t. } \lambda + u_0 + w \succeq 0. \end{aligned} \tag{8.7}$$

Therefore, $w_* = w(u_0)$ which is unique (by the strict convexity of φ along $\ker A$). The fact that any sequence has a convergent subsequence having the same limit $w(u_0)$ implies that the whole sequence $w_k = w(u_k)$ also converges to $w(u_0)$.

However, taking the limit $m \to +\infty$ in (8.6) may not preserve the positivity constraint. To show (8.7) we find a sequence $\{w'_m\}$ such that

$$\lambda + u_m + w'_m \succeq 0,\ w'_m \to w \text{ for } m \to +\infty. \tag{8.8}$$

In this case we can replace w with w'_m in (8.6) and take the limit $m \to \infty$ to obtain (8.7), so the previous argument on contiunuity applies.

It is left how to choose w'_m so that (8.8) holds. Let w'_m be a minimizer in the following minimization problem

$$\begin{aligned} &\text{minimize } \|(\lambda + u_0 + w) - (\lambda + u_m + w'_m)\| \text{ w.r.t } w'_m, \\ &\text{subject to: } w'_m \in \ker A,\ \lambda + u_m + w'_m \succeq 0. \end{aligned} \tag{8.9}$$

Minimizer w'_m in (8.9) corresponds to the euclidean projection of $\lambda+u_0+w$ onto $S_{A,\lambda}(\{u_m\})$ in the sense of Euclidean norm, that is

$$w'_m = \Pi_{\ker A} \cdot [\mathrm{Proj}(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\})) - \lambda], \tag{8.10}$$

where $\Pi_{\ker A}$ is the orthogonal projector onto $\ker A$, $\mathrm{Proj}(x, X)$ denotes the euclidean projection of point x onto convex set X.

From (8.10) and the fact that $\lambda + u_0 + w \in S_{A,\lambda}(\{u_0\})$ it follows that

$$w'_m - w = \Pi_{\ker A}[\mathrm{Proj}(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\})) - \mathrm{Proj}(\lambda + u_0 + w, S_{A,\lambda}(\{u_0\}))]. \tag{8.11}$$

Using (8.11) and Proposition in 5.3 from [AW93] we get the following estimate:

$$\|w'_m - w\| \le \rho_m^{1/2} \cdot d_{H,\rho_m}(S_{A,\lambda}(\{u_0\}), S_{A,\lambda}(\{u_m\}))^{1/2}, \tag{8.12}$$

where $\rho_m = \|\lambda + u_0 + w\| + d(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\}))$, $d(x, y)$ denotes the standard euclidean distance between x, y (for set X, $d(x, X) = \inf_{x' \in X} d(x, x')$), $d_{H,\rho}(\cdot, \cdot)$ is the

bounded Hausdorff distance (Section 3 in [AW93]). In particular, for $d_{H,\rho}$ the following bound holds:

$$d_{H,\rho}(A,B) \le d_H(A,B), \tag{8.13}$$

for any sets A, B.

Note that $\sup_m \rho_m$ is finite. Indeed, this follows from the fact that $u_m \to u_0$ and following estimates:

$$\begin{aligned} d(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\})) &\le d(\lambda + u_0 + w, 0) + d(0, S_{A,\lambda}(\{u_m\})) \\ &\le \|\lambda + u_0 + w\| + d(0, S_{A,\lambda}(\{u_m\})), \end{aligned} \tag{8.14}$$

$$d(0, S_{A,\lambda_*}(\{u_m\})) \le \max_{j\in\{1,\dots,p\}} \left(\sum_{i=1}^{d} a_i^T(\lambda + u_m) \right) / A_j,\ A_j = \sum_{i=1}^{d} a_{ij}. \tag{8.15}$$

The estimate in (8.15) follows from the fact that $S_{A,\lambda}(\{u\})$ is the affine subset of $(p-1)$ – simplex:

$$\Delta^p_{A,\lambda}(u) = \{\lambda' \in \mathbb{R}^p_+ : \sum_{j=1}^{p} \lambda'_j A_j = \sum_{i=1}^{d} a_i^T(\lambda + u)\},\ A_j = \sum_{i=1}^{d} a_{ij}. \tag{8.16}$$

Note that $A_j > 0$ for all $j \in \{1,\dots,p\}$ (see formula (2.4)).

From (8.12), the fact that $\sup_m \rho_m < +\infty$ and the result of Lemma 8.2 it follows that $w'_m \to w$, where $\lambda + u_m + w_m \succeq 0$. So conditions in (8.8) are satisfied, which, in turn, proves (8.7) and the second claim of the lemma.

Lemma is proved. □

8.2 Proof of Lemma 8.1

Proof. Closedness and convexity of $S_{A,\lambda}(U)$ follow directly from the fact that $(\lambda + U + \ker A)$, $\mathbb{R}^p_+$ are both closed and convex and the intersection preserves these properties.

We prove boundedness of $S_{A,\lambda}(U)$ by the contradiction argument.

Assume that $S_{A,\lambda}(U)$ is not bounded, then there exists a sequence $\{(u_k, w_k)\}_{k=1}^{\infty}$, $u_k \in U$, $w_k \in \ker A$, such that

$$\lambda + u_k + w_k \in \mathbb{R}^p_+,\ \|\lambda + u_k + w_k\| \to \infty. \tag{8.17}$$

From (8.17) and compactness of U it follows, in particular, that

$$w_k \text{ in } \ker A, \|w_k\| \to +\infty. \tag{8.18}$$

Also there exists a converging subsequence $\{u_{k_n}\}_{n=1}^{\infty}$ such that

$$u_{k_n} \to u_0 \in U \text{ for some } u_0, \text{ as } n \to +\infty. \tag{8.19}$$

Consider the corresponding subsequence $\{w_{k_n}\}_{n=1}^{\infty}$ for which we know that

$$w_{k_n} \in \ker A,\ \|w_{k_n}\| \to +\infty \text{ for } n \to +\infty. \tag{8.20}$$

Let

$$\theta_n = \frac{w_{k_n}}{\|w_{k_n}\|},\ \theta_n \in \mathbb{S}^{p-1} \cap \ker A. \tag{8.21}$$

Since $\mathbb{S}^{p-1} \cap \ker A$ is compact, $\{\theta_n\}_{n=1}^{\infty}$ has a converging subsequence $\{\theta_m\}_{m=1}^{\infty}$ such that

$$\theta_m \to \theta_0,\ \theta_0 \in \mathbb{S}^{p-1} \cap \ker A. \tag{8.22}$$

Let $\{u_m\}_{m=1}^{\infty}$ be the corresponding subsequence of $\{u_{k_n}\}_{n=1}^{\infty}$ for index m in formula (8.22). From (8.17)-(8.22) it follows that we have constructed a sequence $\{(u_m, w_m)\}_{m=1}^{\infty}$ such that

$$\lambda + u_m + w_m \in \mathbb{R}^p_+, u_m \in U,\ w_m \in \ker A, \tag{8.23}$$

$$u_m \to u_0,\ \|w_m\| \to +\infty, \tag{8.24}$$

$$\theta_m = \frac{w_m}{\|w_m\|} \to \theta_0 \in \mathbb{S}^{p-1} \cap \ker A. \tag{8.25}$$

Now we show that under our initial assumption we arrive to the fact that

$$\lambda + s\theta_0 \in \mathbb{R}^p_+ \text{ for any } s > 0, \tag{8.26}$$

where θ_0 is defined in (8.25).

Indeed, from the fact that $\lambda \in \mathbb{R}^p_+$ and that $\mathbb{R}^p_+$ is convex it follows that

$$\lambda + t(u_m + w_m) = \lambda + t(u_m + \|w_m\|\theta_m) \in \mathbb{R}^p_+ \text{ for any } t \in [0, 1]. \tag{8.27}$$

Let $s > 0$. By choosing $t = t_m(s) = s/\|w_m\|$ in (8.27) ($t_m(s) \in [0, 1]$ for large m; see (8.24)) and using formulas (8.23)-(8.25) we obtain

$$\begin{aligned}(\lambda + s\theta_0) - (\lambda + t_m(s)u_m + t_m(s)\|w_m\|\theta_m) \\ = s(\theta_0 - \theta_m) - s\frac{u_m}{\|w_m\|} \to 0 \text{ for } m \to +\infty.\end{aligned} \tag{8.28}$$

From (8.28) it follows that $\lambda + s\theta_0$ is a limiting point in $\mathbb{R}^p_+$, and due to its closedness it follows that $\lambda + s\theta_0 \in \mathbb{R}^p_+$, $s \geq 0$.

The statement in (8.26) cannot hold, because from (2.5) it follows that

$$\text{for any } \theta \in \ker A,\ \theta \neq 0\, \exists j \in \{1, \dots, p\} \text{ s.t. } \theta_j < 0. \tag{8.29}$$

Since $\theta_0 \in \ker A$, by taking $s > 0$ large enough in formula (8.26), we will arrive to the case when $\lambda + s\theta_0 \notin \mathbb{R}^p_+$, which gives the desired contradiction.

Lemma is proved. □

8.3 Proof of Lemma 8.2

Proof. The claim of the lemma is a part of Theorem 1 from [WW69] which, informally, says that a closed convex set K is a polyhedra iff the Hausdorff distance on the space sections by any family of shifted linear subspaces and parameterized by the shift is Lipschitz continuous.

Using notations from [WW69] we define the following affine mapping

$$\tau_{A,\lambda}(u) = A\lambda + Au,\ u \in \mathbb{R}^p, \tag{8.30}$$

where λ is a parameter, $A \in \mathrm{Mat}(d, p)$ is the design matrix satisfying (2.3), (2.4).

Let $K = \mathbb{R}^p_+$. Define family of sections of K by the formula

$$k(\Lambda) = \tau^{-1}_{A,\lambda}(\Lambda) \cap K,\ \Lambda \in \mathbb{R}^d. \tag{8.31}$$

Essentially, $k(\Lambda)$ is an intersection $\ker A$ being shifted on u with K (in some cases $k(\Lambda)$ can be an empty set).

In particular, if $\Lambda(u) = A\lambda + Au$ for some $u \in \mathbb{R}^p$, then it is easy to see that

$$k(\Lambda(u)) = (\lambda + u + \ker A) \cap K = (\lambda + u + \ker A) \cap \mathbb{R}^p_+ = S_{A,\lambda}(\{u\}), \tag{8.32}$$

where $S_{A,\lambda}$ is defined in (8.1).

The Theorem 1 from [WW69] says, in particular, that

$$d_H(k(\Lambda), k(\Lambda')) \leq C\|\Lambda - \Lambda'\|, \tag{8.33}$$

where C is some constant (depending on K and A), $d_H(\cdot,\cdot)$ is the standard Hausdorff distance being properly extended for empty sets. This extension is not needed for our case since we always consider parameters $\Lambda(u)$ for u from some U which corresponds to apriori non-empty sets $S_{A,\lambda}(\{u\})$.

From formulas (8.32), (8.33) it follows that

$$d_H(S_{A,\lambda}(\{u\}), S_{A,\lambda}(\{u'\})) \leq C\|A(u-u')\|, \tag{8.34}$$

which directly implies (8.4).

Lemma is proved. □

8.4 Proof of Theorem 3.1

Proof. Claim follows directly from the result of Theorem 3.1 from [Lo82]. Indeed, having sample $N_1, \ldots, N_n$ of size n from a Poisson point process with intensity ν is equivalent having sample $N_1 + \ldots N_n$ of size 1 for intensity $n \cdot \nu$. Therefore, parameter n is a direct analog of t in our considerations. Moreover, it is trivial to check that all results from Section 3 of [Lo82] hold for n being replaced with t.

Theorem is proved. □

8.5 Proofs of theorems 6.1 and 6.2

First we prove Theorem 6.2. Then we will show that under (6.9) the conditions in (6.11) for Theorem 6.2 are satisfied which automatically proves Theorem 6.1.

Proof of Theorem 6.2. Minimization problem in step 4 in Algorithm 5 can be rewritten as as follows:

$$\begin{aligned}\widetilde{\lambda}_b^t &= \arg\min_{\lambda \succeq 0} L_p(\lambda|\widetilde{\Lambda}_b^t, A, 1, \beta^t/t) \\ &= \arg\min_{\lambda \succeq 0} [L_p(\lambda|\widetilde{\Lambda}_b^t, A, 1, \beta^t/t) - L_p(\lambda_*|\widetilde{\Lambda}_b^t, A, 1, \beta^t/t)].\end{aligned} \tag{8.35}$$

From (2.9), (2.10) it follows that

$$\begin{aligned}\mathcal{L}^t(\lambda) &= L_p(\lambda|\widetilde{\Lambda}_b^t, A, 1, \beta^t/t) - L_p(\lambda_*|\widetilde{\Lambda}_b^t, A, 1, \beta^t/t) \\ &= \sum_{i \in I_1(\Lambda^*)} (-\widetilde{\Lambda}_{b,i}^t + \Lambda_i^*) \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) \\ &+ \sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) + (\Lambda_i - \Lambda_i^*) \\ &+ \sum_{i \in I_0(\Lambda^*)} -\widetilde{\Lambda}_{b,i}^t \log(\Lambda_i) + \Lambda_i + \frac{\beta_t}{t}(\varphi(\lambda) - \varphi(\lambda_*)),\end{aligned} \tag{8.36}$$

where $I_0(\cdot)$, $I_1(\cdot)$ are defined in (2.2) and

$$\Lambda^* = A\lambda_*. \tag{8.37}$$

We will use the following lemma.

Lemma 8.3. *Let $\mathcal{L}^t(\lambda)$ be defined in (8.36) and conditions of Theorem 6.2 be satisfied. Then,*

i) there exists $\delta_0 = \delta_0(A, \lambda_*) > 0$ such that for any $\delta < \delta_0$ it holds that

$$\inf_{\lambda \in C_{A,\delta}(\lambda_*)} \mathcal{L}^t(\lambda) \geq C\delta^2 + o_{cp}(1) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.38}$$

where C is a positive constant independent of δ and

$$C_{A,\delta}(\lambda_*) = \{\lambda \in \mathbb{R}^p_+ | \lambda = \lambda_* + \delta u + w, \, (u, w) \in \mathrm{Span}(A^T) \times \ker A, \|u\| = 1\}. \tag{8.39}$$

ii) there is a family $\widetilde{\lambda}^t \in \mathbb{R}^p_+$, $t \in (0, +\infty)$, such that

$$\widetilde{\lambda}^t \xrightarrow{c.p.} \lambda_* \text{ and } \mathcal{L}^t(\widetilde{\lambda}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, \, t \in (0, +\infty). \tag{8.40}$$

From Lemma 8.3(i) it follows that for all λ at distance δ from λ_* in the $\mathrm{Span}(A^T)$ values of $\mathcal{L}^t(\lambda)$ are greater or equal than $C\delta^2$ with conditional probability tending to one. At the same time, from Lemma 8.3(ii) it follows that there is a point $\lambda^t \in \mathbb{R}^p_+$ which is arbitrarily close to λ_* and $\mathcal{L}^t(\lambda^t)$ is converges to zero for $t \to +\infty$ with conditional probability also tending to one. Note that function $\mathcal{L}^t(\lambda)$ is convex on $\mathbb{R}^p_+$. The above arguments, convexity of $\mathcal{L}^t(\lambda)$ and the fact that $\widetilde{\lambda}^t_b$ is a minimizer of $\mathcal{L}^t(\lambda)$ in (8.35) imply that

$$P(\, \|\Pi_{A^T}(\widetilde{\lambda}^t_b - \lambda_*)\| < \delta | Y^t, t) \to 1 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.41}$$

where Π_{A^T} is the orthogonal projector onto $\mathrm{Span}(A^T)$.

Since δ is chosen arbitrarily in Lemma 8.3 and using (8.41) we find that

$$\Pi_{A^T}(\widetilde{\lambda}^t_b - \lambda_*) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.42}$$

Let

$$\widetilde{\lambda}^t_b = \lambda_* + \widetilde{u}^t_b + \widetilde{w}^t_b, \text{ where } (\widetilde{u}^t_b, \widetilde{w}^t_b) \in \mathrm{Span}(A^T) \times \ker A. \tag{8.43}$$

From formulas (8.35), (8.36), (8.43) it follows that

$$\widetilde{w}^t_b = \underset{\substack{w: \lambda_* + \widetilde{u}^t_b + w \geq 0, \\ w \in \ker A}}{\arg\min} \varphi(\lambda_* + \widetilde{u}^t_b + w) = w_{A,\lambda_*}(\widetilde{u}^t_b), \tag{8.44}$$

where $w_{A,\lambda}(\cdot)$ is defined in (2.18).

From (8.44), the fact that $\widetilde{u}^t_b \xrightarrow{c.p.} 0$ (see formula (8.42)), the result of Lemma 2.1 and the Continuous Mapping Theorem (see, e.g. [VdV00], Theorem 2.3, p. 7) it follows that

$$\widetilde{w}^t_b \xrightarrow{c.p.} w_{A,\lambda_*}(0) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.45}$$

Formula (6.10) follows from (8.42), (8.43), (8.45).

Theorem is proved. □

Proof of Theorem 6.1. To prove the theorem we use the following lemma.

Lemma 8.4. *Let $\widetilde{\Lambda}^t_b$ be defined as in Algorithm 5 and let $\theta^t/t \to 0$ when $t \to +\infty$. Then,*

$$\widetilde{\Lambda}^t_b \xrightarrow{c.p.} \Lambda^*_i = a_i^T \lambda_* \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.46}$$

In view of (8.46) in Lemma 8.4 we find that all assumptions for Theorem 6.2 are satisfied, therefore formula (6.10) holds.

Theorem is proved. □

8.6 Proof of Lemma 8.3

Proof. First we prove (i), then for (ii) we will give an explicit formula for $\widetilde{\lambda}^t$ and show that (8.40) holds.

First, in formula (8.36) one can see that

$$\inf_{\lambda\in C_{A,\delta}(\lambda_*)}\sum_{i\in I_1(\Lambda^*)}\left(-\widetilde{\Lambda}^t_{b,i}+\Lambda^*_i\right)\log\left(\frac{\Lambda_i}{\Lambda^*_i}\right)\xrightarrow{c.p.}0,\ \text{ when } t\to+\infty,\ \text{ a.s. } Y^t, t\in(0,+\infty). \tag{8.47}$$

The above formula follows from the assumption that $\widetilde{\Lambda}^t_{b,i}\xrightarrow{c.p.}\Lambda^*_i$ and that $\log(\Lambda_i/\Lambda^*_i)=\log(1+\delta a_i^T u/\Lambda^*_i)$, $\lambda\in C_{A,\delta}(\lambda_*)$ is uniformly bounded from above and below for δ small enough. For example, in (8.47) we choose δ such that

$$0<\delta<\min_{i\in I_1(\Lambda^*)}\left(\Lambda^*_i\cdot\|a_i\|^{-1}\right). \tag{8.48}$$

Since $\varphi(\lambda)$ satisfies (2.15), (2.16), there exists a constant $M=M(\lambda_*,\delta,A)$ such that

$$\inf_{\lambda\in C_{A,\delta}(\lambda_*)}\varphi(\lambda)\geq M. \tag{8.49}$$

From (6.9), (8.49) it follows that

$$\frac{\beta^t}{t}\cdot\inf_{C_{A,\delta}(\lambda_*)}(\varphi(\lambda)-\varphi(\lambda_*))\geq o(1),\ \text{ when } t\to+\infty. \tag{8.50}$$

From formulas (8.36), (8.47), (8.50) it follows that

$$\begin{aligned}\inf_{\lambda\in C_{A,\delta}(\Lambda_*)}\mathcal{L}^t(\lambda)\geq&\inf_{\lambda\in C_{A,\delta}(\lambda_*)}\sum_{i\in I_1(\Lambda^*)}-\Lambda^*_i\log\left(\frac{\Lambda_i}{\Lambda^*_i}\right)+(\Lambda_i-\Lambda^*_i)\\&+\sum_{i\in I_0(\Lambda^*)}-\widetilde{\Lambda}^t_{b,i}\log(\Lambda_i)+\Lambda_i+o_{cp}(1).\end{aligned} \tag{8.51}$$

Note that

$$-\widetilde{\Lambda}^t_{b,i}\log(\Lambda_i)\geq 0 \text{ for } \Lambda_i\leq 1. \tag{8.52}$$

From (2.2), (8.37), (8.39) it follows that we can choose δ sufficiently small so that

$$\Lambda_i\leq 1 \text{ for } \lambda\in C_{A,\delta}(\lambda_*), i\in I_0(\Lambda^*). \tag{8.53}$$

For example, to have (8.53) it suffices to choose δ as follows:

$$0<\delta\leq\min_{i\in\{1,\dots,d\}}(\|a_i\|^{-1}). \tag{8.54}$$

Using formulas (8.48), (8.51), (8.52)-(8.54) we obtain

$$\begin{aligned}\inf_{\lambda\in C_{A,\delta}(\Lambda_*)}\mathcal{L}^t(\lambda)\geq&\inf_{\lambda\in C_{A,\delta}(\lambda_*)}\sum_{i\in I_1(\Lambda^*)}-\Lambda^*_i\log\left(\frac{\Lambda_i}{\Lambda^*_i}\right)+(\Lambda_i-\Lambda^*_i)\\&+\sum_{i\in I_0(\Lambda^*)}\Lambda_i+o_{cp}(1).\end{aligned} \tag{8.55}$$

Consider the following function:

$$\Phi_{s^*}(s)=-s^*\log(s)+s,\ s>0,\ s^*>0. \tag{8.56}$$

Function $F_{s^*}(s)$ is convex and at $s = s^*$ it has its minimum. Therefore, for $\varepsilon > 0$ small enough (for example, for $\varepsilon < s^*$) it holds that

$$\Phi_{s^*}(s) - \Phi_{s^*}(s^*) \geq C(\varepsilon, s^*)(s - s^*)^2 \text{ for } |s - s^*| < \varepsilon, \tag{8.57}$$

where $C(\varepsilon, s^*)$ is some positive constant.

From (8.57) it follows that one can choose $\delta_0 > 0$ such that

$$\begin{aligned} &\sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) + (\Lambda_i - \Lambda_i^*) \\ &+ \sum_{i \in I_0(\Lambda^*)} \Lambda_i \geq C(\delta_0, \Lambda^*) \sum_{i \in I_1(\Lambda^*)} (\Lambda_i - \Lambda_i^*)^2 + \sum_{i \in I_0(\Lambda^*)} \Lambda_i \\ &\text{for } |\Lambda_i - \Lambda_i^*| < \delta_0,\ i \in I_1(\Lambda^*). \end{aligned} \tag{8.58}$$

Value for δ_0 will be precised later. Let $\lambda \in C_{A,\delta}(\lambda_*)$ and $\delta < \delta_0$, that is $\lambda = \lambda_* + \delta u + w$, where $u \in \mathrm{Span}(A^T)$, $\|u\| = 1$, $w \in \ker A$. Then, for δ satisfying (8.54) we get the following estimate:

$$\Lambda_i = a_i^T \lambda = \delta a_i^T u \geq \delta^2 (a_i^T u)^2 \geq 0 \text{ for } i \in I_0(\Lambda^*). \tag{8.59}$$

Note that in formula (8.59) we used the fact that $\Lambda_i^* = a_i^T \lambda_* = 0$, $i \in I_0(\Lambda^*)$.

From (8.58), (8.59) it follows that

$$\begin{aligned} \inf_{\lambda \in C_{A,\delta}(\lambda_*)} &\sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) + (\Lambda_i - \Lambda_i^*) + \sum_{i \in I_1(\Lambda^*)} \Lambda_i \\ &\geq \min(C(\delta_0, \Lambda^*), 1)\delta^2 \sum_{i=1}^{d} (a_i^T u)^2 \\ &\geq \min(C(\delta_0, \Lambda^*), 1)\delta^2 \sigma_{min}^+(A^T A), \end{aligned} \tag{8.60}$$

where $\sigma_{min}^+(A^T A)$ is the smallest non-zero eigenvalue of $A^T A$. In particular, in formula (8.60) we have used the property that $u \in \mathrm{Span}(A^T)$ which guarantees that

$$u^T A^T A u \geq \sigma_{min}^+(A^T A) > 0 \text{ for } \|u\| = 1. \tag{8.61}$$

Formula (8.38) follows from formulas (8.55), (8.60).

Finally, we choose δ_0 such that conditions (8.48), (8.54) are simultaneously satisifed

$$\delta_0 = \frac{1}{2} \min\left[\min_{i \in \{1,\dots,d\}} (\|a_i\|^{-1}), \min_{i \in I_1(\Lambda^*)} (\Lambda_i^* \cdot \|a_i\|^{-1})\right]. \tag{8.62}$$

Part (i) of Lemma 8.3 is proved.

Now we prove part (ii) of the lemma. Let

$$\widetilde{\lambda}^t = \lambda_* + \sum_{i \in I_0(\Lambda^*)} \widetilde{\Lambda}_{b,i}^t \frac{a_i}{\|a_i\|^2}. \tag{8.63}$$

Note that $\widetilde{\lambda}^t \in \mathbb{R}_+^p$ because $a_i \in \mathbb{R}_+^p$ and $\widetilde{\Lambda}_{b,i}^t \geq 0$. Since $\widetilde{\Lambda}_{b,i}^t \xrightarrow{c.p.} 0$ for $i \in I_0(\Lambda^*)$ (see the proof of Lemma 8.7) we have that

$$\widetilde{\lambda}^t \xrightarrow{c.p.} \lambda_* \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.64}$$

Note that in (8.36) for $\mathcal{L}^t(\lambda)$ all summands are continuous at λ_* and equal to zero except the logarithmic part which is given by the formula

$$g(\lambda) = \sum_{i \in I_0(\Lambda^*)} -\widetilde{\Lambda}^t_{b,i} \log(\Lambda_i),\ \Lambda_i = a_i^T \lambda. \tag{8.65}$$

From the fact that $a_i \in \mathbb{R}^p_+$ (see formula (2.3)) it follows that $a_i^T a_{i'} \geq 0$ for all i, i'. Using this property and monotonicity of the logarithm it follows that

$$\begin{aligned} g(\widetilde{\lambda}^t) &= \sum_{i \in I_0(\Lambda^*)} -\widetilde{\Lambda}^t_{b,i} \log\left(a_i^T \widetilde{\lambda}^t\right) \\ &\leq \sum_{i \in I_0(\Lambda^*)} -\widetilde{\Lambda}^t_{b,i} \log\left(\widetilde{\Lambda}^t_{b,i}\right) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty \text{ a.s. } Y^t,\ t \in (0, +\infty). \end{aligned} \tag{8.66}$$

Formula (8.66) gives an asymptotic upper bound on $g(\widetilde{\lambda}^t)$ which is equal to zero. For the lower bound we use formulas (8.52), (8.64) and the fact that $a_i^T \widetilde{\lambda}^t \xrightarrow{c.p.} 0$ for $i \in I_0(\Lambda^*)$ from which it follows that

$$\begin{aligned} &g(\widetilde{\lambda}^t) \geq 0 \text{ with conditional probability tending to one when } t \to +\infty, \\ &\text{a.s. } Y^t, t \in (0, +\infty). \end{aligned} \tag{8.67}$$

From (8.66), (8.67) it follows that

$$g(\widetilde{\lambda}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty,\ \text{a.s. } Y^t, t \in (0, +\infty). \tag{8.68}$$

From (8.36), (8.63), (8.65), (8.68) it follows that

$$\mathcal{L}^t(\widetilde{\lambda}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty,\ \text{a.s. } Y^t, t \in (0, +\infty). \tag{8.69}$$

This proves part (ii) of the lemma.

Lemma is proved. □

8.7 Proof of Lemma 8.4

Proof. Recall that

$$\widetilde{\Lambda}^t_{b,i} | Y^t, \Lambda^t_{\mathcal{M}}, t \sim \Gamma(Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}, (\theta^t + t)^{-1}),\ i \in \{1, \ldots, d\}, \tag{8.70}$$

where $\Lambda^t_{\mathcal{M}} | Y^t, t$ is sampled in Algorithm 4.

From the definition of Λ^t in step 1 of Algorithm 4 and necessary optimality conditions in step 2 (see also formula (6.16) in Remark 6.1) it follows that

$$\sum_{i=1}^d \Lambda^t_{\mathcal{M},i} = \sum_{i=1}^d \Lambda^t_i, \tag{8.71}$$

$$\Lambda^t_{\mathcal{M}} \succeq 0,\ \Lambda^t \succeq 0,\ \mathbb{E}[\Lambda^t_i | Y^t, t] = Y^t_i / t,\ i \in 1, \ldots, d. \tag{8.72}$$

Using (8.71), (8.72) we get the following estimate:

$$\mathbb{E}[\Lambda^t_{\mathcal{M},i} | Y^t, t] \leq \sum_{i=1}^d \frac{Y^t_i}{t},\ i \in \{1, \ldots, d\}. \tag{8.73}$$

Let $\varepsilon > 0$. Using the Markov inequality we obtain

$$P(|\widetilde{\Lambda}^t_{b,i} - \Lambda^*_i| > \varepsilon |\, Y^t, t) \le \frac{\mathbb{E}[|\widetilde{\Lambda}^t_{b,i} - \Lambda^*_i||Y^t, t]}{\varepsilon} \le \frac{\mathbb{E}[|\widetilde{\Lambda}^t_{b,i} - \frac{Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}}{\theta^t + t}||Y^t, t]}{\varepsilon} + \frac{\mathbb{E}[|\frac{Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}}{\theta^t + t} - \Lambda^*_i||Y^t, t]}{\varepsilon}. \tag{8.74}$$

Using the Jensen's inequality $\mathbb{E}|X|^2 \ge (\mathbb{E}|X|)^2$, formulas (8.70), (8.73), the Strong Law of Large Numbers for Y^t (see Appendix B) and the fact that $\theta^t/t \to 0$, we get the following:

$$\begin{aligned}
\mathbb{E}[|\widetilde{\Lambda}^t_{b,i} - \frac{Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}}{\theta^t + t}||Y^t, t] &\le \left(\mathbb{E}[|\widetilde{\Lambda}^t_{b,i} - \frac{Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}}{\theta^t + t}|^2|Y^t, t]\right)^{1/2} \\
&= \left(\mathbb{E}[\mathrm{Var}[(\widetilde{\Lambda}^t_{b,i})|Y^t, \Lambda^t_{\mathcal{M}}, t]|Y^t, t]\right)^{1/2} \\
&= \left(\frac{Y^t_i + \theta^t \mathbb{E}[\Lambda^t_{\mathcal{M},i}|Y^t, t]}{(t + \theta^t)^2}\right)^{1/2} \\
&\le \left(\frac{Y^t_i + (\theta^t/t)\sum_{i=1}^d Y^t_i}{(t + \theta^t)^2}\right)^{1/2} \to 0 \text{ a.s. } Y^t,\ t \in (0, +\infty).
\end{aligned} \tag{8.75}$$

For the second term in (8.74) we use formula (8.73), the triangle inequality and the property that $\theta^t/t \to 0$ to get the following:

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{Y^t_i + \theta^t \Lambda^t_{\mathcal{M},i}}{\theta^t + t} - \Lambda^*_i\right| |Y^t, t\right] &\le \left|\frac{Y^t_i}{\theta^t + t} - \Lambda^*_i\right| + \mathbb{E}\left[\frac{\theta^t \Lambda^t_{\mathcal{M}}}{\theta^t + t}|Y^t, t\right] \\
&\le \left|\frac{Y^t_i}{\theta^t + t} - \Lambda^*_i\right| + \frac{\theta^t}{\theta^t + t}\sum_{i=1}^d \frac{Y^t_i}{t} \to 0 \text{ a.s. } Y^t, t \in (0, +\infty).
\end{aligned} \tag{8.76}$$

Formula (8.46) follows from formulas (8.74)-(8.76).

Lemma is proved. □

8.8 Proof of Theorem 6.3

Proof. We prove directly (6.32) which automatically implies uniqueness of the minimizer.

Let $\lambda_{\mathcal{M},*} \in \mathbb{R}^{p_{\mathcal{M}}}_+$ be a minimizer in (6.12) (see also Remark 6.1).

Let

$$\lambda_{\mathcal{M}} = \lambda_{\mathcal{M},*} + u_{\mathcal{M}},\ \lambda_{\mathcal{M}} \in \mathbb{R}^{p_{\mathcal{M}}}_+. \tag{8.77}$$

Consider the second order Taylor expansion of $L(\lambda|\Lambda^*, A_{\mathcal{M}})$ in (6.12) in a vicinity of $\lambda_{\mathcal{M},*}$:

$$\begin{aligned}
&L(\lambda_{\mathcal{M}}|\Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}, 1) \\
&\quad = u^T_{\mathcal{M}} \nabla L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}, 1) + \frac{1}{2}\sum_{i \in I_1(\Lambda^*)} \Lambda^*_i \frac{(u^T_{\mathcal{M}} a_{\mathcal{M},i})^2}{(\Lambda^*_{\mathcal{M},i})^2} \\
&\quad + o(\|\Pi_{A^T_{\mathcal{M},I_1(\Lambda^*)}} u_{\mathcal{M}}\|^2),
\end{aligned} \tag{8.78}$$

where $\Lambda^*_{\mathcal{M}} = A_{\mathcal{M}} \lambda_{\mathcal{M},*}$ and

$$\nabla L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}, 1) = \sum_{i \in I_1(\Lambda^*)} -\Lambda^*_i \frac{a_{\mathcal{M},i}}{\Lambda^*_{\mathcal{M},i}} + \sum_{i=1}^d a_{\mathcal{M},i}. \tag{8.79}$$

From formulas (6.14), (6.15) of Remark 6.1 and (8.77), (8.79) it follows that

$$\begin{aligned} u_{\mathcal{M}}^T \nabla L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}) = u_{\mathcal{M}}^T \mu_{\mathcal{M},*} = (\lambda_{\mathcal{M}} - \lambda_{\mathcal{M},*})^T \mu_{\mathcal{M},*} \\ = \lambda_{\mathcal{M}}^T \mu_{\mathcal{M},*} \geq 0,\ \mu_{\mathcal{M},*} \succeq 0, \end{aligned} \tag{8.80}$$

where $\mu_{\mathcal{M},*}$ is the optimal lagrangian multiplier for the problem in (6.12). Formulas (6.30), (6.31) follow from (8.78)-(8.80). Next, we prove that (6.32) holds.

Using (8.80) we obtain the following estimate:

$$u_{\mathcal{M}}^T \nabla L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}) = u_{\mathcal{M}}^T \mu_{\mathcal{M},*} \geq (u_{\mathcal{M}}^T \mu_{\mathcal{M},*})^2 \text{ if } \|u_{\mathcal{M}}\| \leq \|\mu_{\mathcal{M},*}\|^{-1}. \tag{8.81}$$

From (8.78), (8.80) it follows that

$$\begin{aligned} L(\lambda_{\mathcal{M}}|\Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}, 1) \\ \geq u_{\mathcal{M}}^T C_{\mathcal{M},*} u_{\mathcal{M}} + o(\|u_{\mathcal{M}}\|^2), \\ \text{for } \|u_{\mathcal{M}}\| \leq \|\mu_{\mathcal{M},*}\|^{-1}, \end{aligned} \tag{8.82}$$

where

$$C_{\mathcal{M},*} = \mu_{\mathcal{M},*} \mu_{\mathcal{M},*}^T + \frac{1}{2} \sum_{i \in I_1(\Lambda^*)} \Lambda_i^* \frac{a_{\mathcal{M},i} a_{\mathcal{M},i}^T}{(\Lambda_{\mathcal{M},i}^*)^2}. \tag{8.83}$$

To prove the claim we use two following lemmas.

Lemma 8.5. *Let*

$$C_\delta = \inf_{\substack{u_{\mathcal{M}}: \lambda_{\mathcal{M},*} + u_{\mathcal{M}} \succeq 0, \\ \|u_{\mathcal{M}}\| = \delta}} u_{\mathcal{M}}^T C_{\mathcal{M},*} u_{\mathcal{M}}. \tag{8.84}$$

Let assumptions of Theorem 6.3 be satisfied. Then,

$$C_\delta > 0 \text{ for any } \delta > 0. \tag{8.85}$$

Lemma 8.6. *Let $\lambda_{\mathcal{M},*} \in \mathbb{R}_+^{p_{\mathcal{M}}}$. There exists $\delta_* > 0$ such that for any $u_{\mathcal{M}} \in \mathbb{R}^{p_{\mathcal{M}}}$, $0 < \|u_{\mathcal{M}}\| \leq \delta_*$, $\lambda_{\mathcal{M},*} + u_{\mathcal{M}} \succeq 0$ it also holds that*

$$\lambda_{\mathcal{M},*} + \delta_* \frac{u_{\mathcal{M}}}{\|u_{\mathcal{M}}\|} \succeq 0. \tag{8.86}$$

Let δ_* be the one of Lemma 8.6. From formula (8.82) and the result of Lemmas 8.5, 8.6 it follows that

$$\begin{aligned} L(\lambda_{\mathcal{M}}|\Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}, 1) \\ \geq \frac{\delta_* u_{\mathcal{M}}^T}{\|u_{\mathcal{M}}\|} C_{\mathcal{M},*} \frac{\delta_* u_{\mathcal{M}}}{\|u_{\mathcal{M}}\|} \cdot \frac{\|u_{\mathcal{M}}\|^2}{\delta_*^2} + o(\|u_{\mathcal{M}}\|^2) \\ \geq C_{\delta_*} \frac{\|u_{\mathcal{M}}\|^2}{\delta_*^2} + o(\|u_{\mathcal{M}}\|^2),\ C_{\delta_*} > 0, \\ \text{for } \lambda_{\mathcal{M},*} + u_{\mathcal{M}} \succeq 0,\ \|u_{\mathcal{M}}\| \leq \min(\delta_*, \|\mu_{\mathcal{M},*}\|^{-1}) \end{aligned} \tag{8.87}$$

This proves the claim in (6.32).

Theorem is proved. □

Proof of Lemma 8.5. We prove Lemma 8.5 by contradiction argument. Assume that it exists $\delta > 0$ such that $C_\delta = 0$, where C_δ is defined in (8.84). Since the infimum in (8.84) is taken over a compact set, there should exist $u_{\mathcal{M}}$, $\|u_{\mathcal{M}}\| = \delta$, $\lambda_{\mathcal{M},*} + u_{\mathcal{M}} \succeq 0$ such that

$$u_{\mathcal{M}}^T C_{\mathcal{M},*} u_{\mathcal{M}} = 0. \tag{8.88}$$

From (8.83) it follows that

$$u_{\mathcal{M}}^T a_{\mathcal{M},i} = 0,\ i \in I_1(\Lambda^*),\ u_{\mathcal{M}}^T \mu_{\mathcal{M},*} = 0. \tag{8.89}$$

Using formulas (6.13), (6.14), (8.89) we obtain the following formula:

$$\begin{aligned} u_{\mathcal{M}}^T \mu_{\mathcal{M},*} &= \sum_{i \in I_0(\Lambda^*)} u_{\mathcal{M}}^T a_{\mathcal{M},*,i} = \sum_{i \in I_0(\Lambda^*)} (\lambda_{\mathcal{M},i} - \lambda_{\mathcal{M},*,i})^T a_{\mathcal{M},*,i} \\ &= \sum_{i \in I_0(\Lambda^*)} (\Lambda_{\mathcal{M},i} - \Lambda_{\mathcal{M},*,i}) = \sum_{i \in I_0(\Lambda^*)} \Lambda_{\mathcal{M},i} = 0,\ \Lambda_{\mathcal{M},i} = \lambda_{\mathcal{M}}^T a_{\mathcal{M},i} \end{aligned} \tag{8.90}$$

From (8.90) and the fact that $\Lambda_M \succeq 0$ it follows that

$$\Lambda_{\mathcal{M},i} = u_{\mathcal{M}}^T a_{\mathcal{M},i} = 0,\ i \in I_0(\Lambda^*). \tag{8.91}$$

Putting formulas (8.89), (8.91) we arrive to the following:

$$u_{\mathcal{M}}^T a_{\mathcal{M},i} = 0 \text{ for } i \in \{1, \dots, d\}. \tag{8.92}$$

The injectivity of $A_{\mathcal{M}}$ and (8.92) imply that $u_{\mathcal{M}} = 0$ which contradicts the initial assumption that $\|u_{\mathcal{M}}\| = \delta > 0$.

Lemma is proved. □

Proof of Lemma 8.6. We prove the claim by contradiction.

The claim is obvious for $\lambda_{\mathcal{M},*} = 0$.

Let $\lambda_{\mathcal{M},*} \neq 0$ and

$$\delta_* = \frac{1}{2} \min\{\lambda_{\mathcal{M},*,j} | \lambda_{\mathcal{M},*,j} > 0\},\ \delta_* > 0. \tag{8.93}$$

Let $u_{\mathcal{M}}$ be such that

$$0 < \|u_{\mathcal{M}}\| \leq \delta_*,\ \lambda_{\mathcal{M},*} + u_{\mathcal{M}} \succeq 0 \tag{8.94}$$

and assume that

$$\lambda_{\mathcal{M},*} + \delta_* \frac{u_{\mathcal{M}}}{\|u_{\mathcal{M}}\|} \not\succeq 0 \Leftrightarrow \exists j \in \{1, \dots, p_{\mathcal{M}}\} \text{ s.t. } \lambda_{\mathcal{M},*,j} + \delta_* \frac{u_{\mathcal{M},j}}{\|u_{\mathcal{M}}\|} < 0. \tag{8.95}$$

From the fact that $\lambda_{\mathcal{M},*} \succeq 0$ and (8.94), (8.95) it follows that

$$\text{for } j \text{ from (8.95) it holds that } \lambda_{\mathcal{M},*,j} > 0,\ u_{\mathcal{M},j} < 0. \tag{8.96}$$

Using (8.93), (8.95), (8.96) we get the following inequality:

$$\frac{\delta_*}{\|u_{\mathcal{M}}\|}(-u_{\mathcal{M},j}) > \lambda_{\mathcal{M},*,j} \geq 2\delta_* \Rightarrow (-u_{\mathcal{M},j}) > 2\|u_{\mathcal{M}}\|. \tag{8.97}$$

The inequality in the right hand-side of (8.97) gives the desired contradiction.

Lemma is proved. □

8.9 Proof of Theorem 6.4

Proof. Claim in (6.34) directly follows from (6.33) by the Continuous Mapping Theorem, so we prove only (6.33).

Step 2 in Algorithm 4 can be rewritten as follows:

$$\lambda_{\mathcal{M}}^t = \arg\min_{\lambda_{\mathcal{M}} \succeq 0} L_{\mathcal{M}}^t(\lambda_{\mathcal{M}}), \tag{8.98}$$

$$L^t_{\mathcal{M}}(\lambda_{\mathcal{M}}) = \sum_{i\in I_1(\Lambda^*)} -\log\left(\frac{\Lambda_{\mathcal{M},i}}{\Lambda^*_{\mathcal{M},i}}\right)(\Lambda^t_i - \Lambda^*_i) + L(\lambda_{\mathcal{M}}|\Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*}|\Lambda^*, A_{\mathcal{M}}, 1), \tag{8.99}$$

where $\lambda_{\mathcal{M},*}$ is the point from Theorem 6.3, $\Lambda^*_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M},*}$, and

$$\begin{aligned}&\Lambda^t_i \sim \Gamma(Y^t_i, t^{-1}) \text{ are mutually independent,}\\ &\mathbb{E}[\Lambda^t_i|Y^t, t] = Y^t_i/t,\ \mathrm{Var}[\Lambda^t_i|Y^t, t] = Y^t_i/t^2,\ i \in \{1,\dots,d\}.\end{aligned} \tag{8.100}$$

Note that

$$L^t_{\mathcal{M}}(\lambda_{\mathcal{M}}) \text{ is convex on } \mathbb{R}^{p_{\mathcal{M}}}_+,\ L^t_{\mathcal{M}}(\lambda_{\mathcal{M},*}) = 0. \tag{8.101}$$

Consider the following parametrization:

$$\lambda_{\mathcal{M}} = \lambda_{\mathcal{M},*} + \frac{u_{\mathcal{M}}}{r(t)},\ \lambda_{\mathcal{M}} \in \mathbb{R}^{p_{\mathcal{M}}}_+,\ r(t) = o(\sqrt{t/\log\log t}). \tag{8.102}$$

Let $\delta > 0$. In view of (8.98), (8.101), (8.102) the following implication holds:

$$\inf_{\substack{\lambda_{\mathcal{M}}:\|u_{\mathcal{M}}\|=\delta,\\ \lambda_{\mathcal{M}}\succeq 0}} L^t_{\mathcal{M}}(\lambda_{\mathcal{M}}) > 0 \Rightarrow r(t)\|\lambda^t_{\mathcal{M}} - \lambda_{\mathcal{M},*}\| < \delta \tag{8.103}$$

Therefore, to prove (6.33) it is sufficient to show that for any small $\delta > 0$ the conditional probability of the event left hand-side in (8.103) tends to one a.s. Y^t, $t \in (0, +\infty)$.

Let C_*, δ_* be the values of (6.32) from Theorem 6.3 and $\|u_{\mathcal{M}}\| = \delta$, $\delta < \delta_*$.

Using (6.32) and formulas (8.99), (8.102) we get the following estimate:

$$\begin{aligned}L^t(\lambda_{\mathcal{M}}) &\geq \sum_{i\in I_1(\Lambda^*)} -\log\left(1 + \frac{u^T_{\mathcal{M}} a_{\mathcal{M},i}}{r(t)\Lambda^*_{\mathcal{M},i}}\right)(\Lambda^t_i - \Lambda^*_i) + C_*\delta^2/r^2(t)\\ &\geq C_*\delta^2/r^2(t) - \sum_{i\in I_1(\Lambda^*)} \frac{|u^T_{\mathcal{M}} a_{\mathcal{M},i}|}{r(t)\Lambda^*_{\mathcal{M},i}} \cdot |\Lambda^t_i - \Lambda^*_i|\\ &= r^{-2}(t)\left(C_*\delta^2 - \sum_{i\in I_1(\Lambda^*)} \frac{|u^T_{\mathcal{M}} a_{\mathcal{M},i}|}{\Lambda^*_{\mathcal{M},i}} \cdot r(t)|\Lambda^t_i - \Lambda^*_i|\right)\\ &\geq r^{-2}(t)\left(C_*\delta^2 - \sum_{i\in I_1(\Lambda^*)} \frac{\delta\|a_{\mathcal{M},i}\|}{\Lambda^*_{\mathcal{M},i}} \cdot r(t)|\Lambda^t_i - \Lambda^*_i|\right).\end{aligned} \tag{8.104}$$

Also in (8.104) we have used the property that $\log(1+x) \leq x$, $x \in (-1, +\infty)$.

Estimate in (8.104) implies the left hand-side of (8.103), for example, if

$$r(t)|\Lambda^t_i - \Lambda^*_i| \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty),\ i \in I_1(\Lambda^*). \tag{8.105}$$

To prove (8.105) we use (8.100) and the Markov inequality as follows:

$$\begin{aligned}P(r(t)|\Lambda^t_i - \Lambda^*_i| > \varepsilon|Y^t, t) &\leq \frac{r^2(t)\mathbb{E}(|\Lambda^t_i - \Lambda^*_i|^2|Y^t, t)}{\varepsilon^2}\\ &\leq \frac{2r^2(t)\mathbb{E}(|\Lambda^t_i - Y^t_i/t|^2|Y^t, t) + 2r^2(t)|Y^t_i - \Lambda^*_i|^2}{\varepsilon^2}\\ &= \frac{2r^2(t)/t^2 + 2|r(t)(Y^t_i/t - \Lambda^*_i)|^2}{\varepsilon^2},\end{aligned} \tag{8.106}$$

where $\varepsilon > 0$ is arbitrary. For $r(t) = o(\sqrt{t/\log\log t})$ the following holds (see Appendix B):

$$r^2(t)/t^2 \to 0 \text{ and } r(t)(Y_i^t/t - \Lambda_i^*) \to 0 \text{ a.s. } Y^t, t \in (0, +\infty). \tag{8.107}$$

From (8.106), (8.107) it follows that formula (8.105) holds which together with (8.104) imply (8.103).

Theorem is proved.

□

8.10 Proof of Theorem 6.5

Proof. The formula for $\widetilde{\lambda}_b^t$ in step 4 of Algorithm 5 can be rewritten as follows:

$$\widetilde{\lambda}_b^t = \arg\min_{\lambda \succeq 0} A^t(\lambda), \tag{8.108}$$

$$\begin{aligned} A^t(\lambda) &= L_p(\lambda | t\widetilde{\Lambda}_b^t, A, t, \beta^t) - L_p(\widehat{\lambda}_{sc}^t | t\widehat{\Lambda}_{sc}^t, A, t, \beta^t) \\ &= \sum_{i \in I_1(\Lambda^*)} -t(\widetilde{\Lambda}_{b,i}^t - \widehat{\Lambda}_{sc,i}^t) \log\left(\frac{\Lambda_i}{\widehat{\Lambda}_{sc,i}^t}\right) \\ &+ \sum_{i \in I_1(\Lambda^*)} -t\widehat{\Lambda}_{sc,i}^t \log\left(\frac{\Lambda_i}{\widehat{\Lambda}_{sc,i}^t}\right) + t(\Lambda_i - \widehat{\Lambda}_{sc,i}^t) \\ &+ \sum_{i \in I_0(\Lambda^*)} -t\widetilde{\Lambda}_{b,i}^t \log(t\Lambda_i) + t\Lambda_i \\ &- \left(\sum_{i \in I_0(\Lambda^*)} -t\widehat{\Lambda}_{sc,i}^t \log(t\widehat{\Lambda}_{sc,i}^t) + t\widehat{\Lambda}_{sc,i}^t \right) \\ &+ \beta^t(\varphi(\lambda) - \varphi(\widehat{\lambda}_{sc}^t)), \end{aligned} \tag{8.109}$$

where $\widehat{\lambda}_{sc}^t$ is the strongly consistent estimator from (6.40)-(6.42).

To prove the claim, first, we approximate $A^t(\lambda)$ with quadratic process $B^t(\lambda)$ for which its minimizers have the same asymptotic distribution in the $\mathrm{Span}(A^T) \cap \mathbb{R}_+^p$ as for $A^t(\lambda)$. Second, using this approximation we establish the statements in (i), (ii), but for minimizers of $B^t(\lambda)$ which together with the previous argument completes the proof.

Approximations $B^t(\lambda)$, $\widetilde{\lambda}_{b,app}^t$ of $A^t(\lambda)$, $\widetilde{\lambda}_b^t$ are defined by the formulas:

$$\begin{aligned} B^t(\lambda) = &\sum_{i \in I_1(\Lambda^*)} -t(\widetilde{\Lambda}_{b,i}^t - \widehat{\Lambda}_{sc,i}^t) \frac{\Lambda_i - \widehat{\Lambda}_{sc,i}}{\widehat{\Lambda}_{sc,i}} + t \cdot \frac{(\Lambda_i - \widehat{\Lambda}_{sc,i})^2}{2\widehat{\Lambda}_{sc,i}} \\ &+ \sum_{i \in I_0(\Lambda^*)} t\Lambda_i,\ \Lambda_i = a_i^T \lambda. \end{aligned} \tag{8.110}$$

$$\widetilde{\lambda}_{b,app}^t = \arg\min_{\lambda \succeq 0} B^t(\lambda). \tag{8.111}$$

Note that $B^t(\lambda)$ is flat in directions from $\ker A$, therefore, though $\widetilde{\lambda}_{b,app}^t$ in (8.111) always exists, it may not be unique, and, in general, $\widetilde{\lambda}_{b,app}^t$ is set-valued. In what follows, if not said otherwise, for $\widetilde{\lambda}_{b,app}^t$ one chooses any point from the set of minimizers.

For large t it may happen that $a_i^T \widetilde{\lambda}_{b,app}^t = 0$ for some $i \in I_0(\Lambda^*)$, so $A^t(\widetilde{\lambda}_{b,app}^t)$ may not be defined due to the presence of logarithmic terms in (8.109). For this reason we approximate $\widetilde{\lambda}_{b,app}^t$ with an auxiliary point $\widetilde{\lambda}_{app}^t$ defined by the formula:

$$\widetilde{\lambda}^t_{app} = \widetilde{\lambda}^t_{b,app} + \sum_{i \in I_0(\Lambda^*)} \widetilde{\Lambda}^t_{b,i} \frac{a_i}{\|a_i\|^2}, \tag{8.112}$$

where $\widetilde{\Lambda}^t_b$ is from step 3 of Algorithm 5. It is easy to check that value $A^t(\widetilde{\lambda}^t_{app})$ is always well-defined.

Let $\mathcal{V}$, $\mathcal{U}$ be the subspaced defined in (6.35), (6.36), respectively. From (8.112) and the definiton of $\mathcal{V}, \mathcal{U}$ it follows that

$$\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{app} - \widetilde{\lambda}^t_{b,app}) \equiv 0, \tag{8.113}$$

where $\Pi_{\mathcal{U}}$ is defined in (6.36). For the approximation on $\mathcal{V}$ the following result holds.

Lemma 8.7. *Let $\mathcal{V}$ be the subspace defined in (6.35), $\Pi_{\mathcal{V}}$ be defined in (6.38). Then,*

$$t \cdot \Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \widetilde{\lambda}^t_{app}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{8.114}$$

Let $\delta > 0$. Consider the two following sets:

$$D^t_{A,\delta}(\lambda) = \{\lambda' \in \mathbb{R}^p_+ : \lambda' = \lambda + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W},\ \|u\|_2 + \|v\|_1 \le \delta\}, \tag{8.115}$$

$$C^t_{A,\delta}(\lambda) = \{\lambda' \in \mathbb{R}^p_+ : \lambda' = \lambda + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W},\ \|u\|_2 + \|v\|_1 = \delta\}, \tag{8.116}$$

where subspaces $\mathcal{V}, \mathcal{U}, \mathcal{W}$ are defined in (6.35)-(6.37), respectively and $\|\cdot\|_2, \|\cdot\|_1$ denote the standard ℓ_2 and ℓ_1-norms in $\mathbb{R}^p$.

Main idea behind the approximation is the convexity argument for $A^t(\lambda)$ which in our case boils down to the following implication:

$$\inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})) > 0 \Rightarrow \widetilde{\lambda}^t_b \in D^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{8.117}$$

In view of this, for the approximation it suffices to establish the following result.

Lemma 8.8. *Let $A^t(\lambda)$, $B^t(\lambda)$, $\widetilde{\lambda}^t_b$, $\widetilde{\lambda}^t_{b,app}$, $\widetilde{\lambda}^t_{,app}$ be defined in (8.109), (8.110), (8.108), (8.111), (8.112), respectively. Then, for any $\delta > 0$ the following formula holds:*

$$P\left(\inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] > 0 \,|Y^t, t\right) \to 1 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{8.118}$$

In particular, from (8.117), (8.118) it follows that

$$\sqrt{t} \cdot \Pi_{\mathcal{U}}(\widetilde{\lambda}^t_b - \widetilde{\lambda}^t_{b,app}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty), \tag{8.119}$$

$$t \cdot \Pi_{\mathcal{V}}(\widetilde{\lambda}^t_b - \widetilde{\lambda}^t_{b,app}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{8.120}$$

Let

$$\lambda = \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W}. \tag{8.121}$$

Process $B^t(\cdot)$ defined in (8.110) has the following form in terms of variables u, v (note that $B^t(\cdot)$ is independent of $w \in \mathcal{W}$):

$$B^t(u, v) = \widetilde{B}^t(u, v) + \widetilde{R}^t(u, v), \tag{8.122}$$

$$\widetilde{B}^t(u,v) = \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{a_i^T u}{\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} a_i^T v, \tag{8.123}$$

$$\begin{aligned}\widetilde{R}^t(u,v) = &\sum_{i\in I_1(\Lambda^*)} -(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})a_i^T v + \frac{(a_i^T v)^2}{2\widehat{\Lambda}^t_{sc,i}t} + \frac{(a_i^T u)(a_i^T v)}{\sqrt{t}\widehat{\Lambda}^t_{sc,i}} \\ &+ \sum_{i\in I_0(\Lambda^*)} t\widehat{\Lambda}^t_{sc,i}.\end{aligned} \tag{8.124}$$

Let

$$(\widetilde{u}^t, \widetilde{v}^t) = \underset{\substack{(u,v):\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+\frac{v}{t}+w\succeq 0\\ u\in\mathcal{U},\, v\in\mathcal{V},\, w\in\mathcal{W}}}{\arg\min}\ \widetilde{B}^t(u,v) \tag{8.125}$$

In particular, from the definition of $\mathcal{V}$ in (6.35) and (8.121), (8.123), (8.125) it follows that

$$\frac{\widetilde{v}^t_j}{t} = -\widehat{\lambda}^t_{sc,j} \text{ for } j \text{ s.t. } \exists a_{ij} > 0,\ i \in I_0(\Lambda^*) \Leftrightarrow \Pi_{\mathcal{V}}(\widehat{\lambda}^t_{sc} + \frac{\widetilde{v}^t}{t}) = 0. \tag{8.126}$$

Indeed, formulas (6.35), (6.38), (8.123) imply that the choice in (8.126) satisfies the positivity constraint in (8.125) and at the same time minimizes the linear term $\sum_{i\in I_0(\Lambda^*)} a_i^T v$ since all entries a_{ij} are non-negative.

Lemma 8.9. *Let $\widetilde{u}^t_{b,app}$, $\widetilde{v}^t_{b,app}$ be defined by (8.111) via parametrization in (8.121) and $\widetilde{u}^t$, $\widetilde{v}^t$ be defined by (8.125), respectively. Then,*

$$\widetilde{u}^t - \widetilde{u}^t_{b,app} \xrightarrow{c.p.} 0 \text{ when } t\to+\infty,\ \text{a.s. } Y^t, t\in(0,+\infty), \tag{8.127}$$

$$\widetilde{v}^t - \widetilde{v}^t_{b,app} \xrightarrow{c.p.} 0 \text{ when } t\to+\infty,\ \text{a.s. } Y^t, t\in(0,+\infty). \tag{8.128}$$

In view of Lemma 8.9 it sufficies to prove conditional tightness of $(\widetilde{u}^t, \widetilde{v}^t)$.

Statement in (i) (i.e., formula (6.43)) follows from (8.120), (8.126), (8.128) and the assumption in (6.42).

From (8.125), (8.126) it follows that

$$\widetilde{u}^t = \underset{\substack{u:(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0\\ u\in\mathcal{U},\, w\in\mathcal{W}}}{\arg\min} \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{a_i^T u}{\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}}. \tag{8.129}$$

Since the minmized functional in (8.129) is strongly convex in $u\in\mathcal{U}$ and the set of constraints is also convex, the following mapping is well-defined:

$$\widetilde{u}^t(\xi) = \widetilde{u}(\xi,t)\in\mathcal{U},\ \xi\in\mathbb{R}^{\#I_1(\Lambda^*)},\ t\in(0,+\infty), \tag{8.130}$$

$$\widetilde{u}(\xi,t) = \underset{\substack{u:(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0\\ u\in\mathcal{U},\, w\in\mathcal{W}}}{\arg\min} -\xi^T(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}A_{I_1(\Lambda^*)}u + \frac{1}{2}u^T\widehat{F}^t_{I_1(\Lambda^*)}u, \tag{8.131}$$

where

$$\widehat{D}^t_{I_1(\Lambda^*)} = \mathrm{diag}(\dots, \widehat{\Lambda}^t_{sc,i}, \dots), \tag{8.132}$$

$$\widehat{F}^t_{I_1(\Lambda^*)} = \sum_{i\in I_1(\Lambda^*)} \frac{a_i a_i^T}{\widehat{\Lambda}^t_{sc,i}} = A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1}A_{I_1(\Lambda^*)}. \tag{8.133}$$

Note that the minimized functional in (8.131) does not depend on $w\in\mathcal{W}$ which in turn affects only the set of constraints.

Lemma 8.10. *Let $\widetilde{u}^t(\xi)$ be the mapping defined in (8.130)-(8.133). Then,*

$$\|\widetilde{u}^t(\xi)\| \leq c_t \|A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\xi\|,\ \xi \in \mathbb{R}^{\#I_1(\Lambda^*)}, \tag{8.134}$$

$$c_t = \|(\widehat{F}^t_{I_1(\Lambda^*)})^{-1}\|_{\mathcal{U}} + 2\|(\widehat{F}^t_{I_1(\Lambda^*)})^{-1}\|_{\mathcal{U}} \cdot \left(\max_{\sigma \in \sigma_{\mathcal{U}}(\widehat{F}^t_{I_1(\Lambda^*)})} \sigma^{-1/2} \cdot \|(\widehat{F}^t_{I(\Lambda^*)})^{-1/2}\|_{\mathcal{U}} \right), \tag{8.135}$$

where $\|\cdot\|_{\mathcal{U}}$ denotes the operator norm being reduced to subspace $\mathcal{U}$.

Lemma 8.11. *Let*

$$\widetilde{\xi}^t = (\dots, \sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})/\sqrt{\widehat{\Lambda}^t_{sc,i}}, \dots),\ i \in I_1(\Lambda^*),\ \widetilde{\xi}^t \in \mathbb{R}^{\#I_1(\Lambda^*)}. \tag{8.136}$$

Then, under the assumptions of Theorem 6.5

(i) $A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\widetilde{\xi}^t$ is conditionally tight a.s. Y^t, $t \in (0, +\infty)$.

(ii) $\widetilde{u}^t(\widetilde{\xi}^t)$ is conditionally uniformly tight almost surely Y^t, $t \in (0, +\infty)$.

Statement (ii) of the lemma follows directly from the results of lemmas 8.8-8.11. Theorem is proved.

□

8.11 Proof of Lemma 8.7

Proof. To prove the claim is suffices to show that

$$t\widetilde{\Lambda}^t_{b,i} \xrightarrow{c.p.} 0 \text{ for } i \in I_0(\Lambda^*) \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{8.137}$$

Let $\delta > 0$. Using step 3 in Algorithm 5 and Assumption 1 we obtain

$$\begin{aligned} P(t\widetilde{\Lambda}^t_{b,i} > \delta | Y^t, t) &= \int_0^{+\infty} P(t\widetilde{\Lambda}^t_{b,i} > \delta | \Lambda^t_{\mathcal{M},i} = \Lambda, Y^t, t)\, P(\Lambda^t_{\mathcal{M},i} = \Lambda | Y^t, t) d\Lambda \\ &\leq \int_0^{+\infty} \min\left(\frac{t\theta^t\Lambda}{(\theta^t + t)\delta}, 1\right) P(\Lambda^t_{\mathcal{M},i} = \Lambda | Y^t, t) d\Lambda \\ &\leq \int_0^{\frac{(\theta^t+t)\delta}{t\theta^t}} \frac{t\theta^t\Lambda}{(\theta^t + t)\delta} P(\Lambda^t_{\mathcal{M},i} = \Lambda | Y^t, t) d\Lambda + P\left(\frac{t\theta^t\Lambda^t_{\mathcal{M},i}}{\theta^t + t} > \delta | Y^t, t\right). \end{aligned} \tag{8.138}$$

Note that in (8.138) we have used the Markov inequality for $\Lambda^t_b | Y^t, t, \Lambda^t_{\mathcal{M}}$, $i \in I_0(\Lambda^*)$ for which it is known that $\Lambda^t_{b,i} | Y^t, t, \Lambda^t_{\mathcal{M},i} \sim \Gamma(\theta^t\Lambda_{\mathcal{M},i}, (t + \theta^t)^{-1})$.

The last term in (8.138) tends to zero a.s. Y^t, $t \in (0, +\infty)$ due to (6.34) from Theorem 6.4.

Next, we show that the first integral in (8.138) it is arbitrarily small a.s. Y^t, $t \in (0, +\infty)$ and, hence, tends to zero a.s. Y^t, $t \in (0, +\infty)$. The integral in (8.138) is rewritten as follows:

$$\begin{aligned} &\int_0^{\frac{(\theta^t+t)\delta}{t\theta^t}} \frac{t\theta^t\Lambda}{(\theta^t + t)\delta} P(\Lambda^t_{\mathcal{M},i} = \Lambda | Y^t, t) d\Lambda = \\ &= \frac{\delta(\theta^t + t)}{t\theta^t} \int_0^1 s P(\theta^t\Lambda^t_{\mathcal{M},i} = s\delta(t + \theta^t)/t \,|\, Y^t, t)\, ds. \end{aligned} \tag{8.139}$$

From the choise of $\theta^t = o(\sqrt{t/\log\log t})$, multiplier $\delta(\theta^t + t)/t\theta^t$ in (8.139) is uniformly bounded and does not affect the estimate.

Let $0 < \varepsilon < 1$. Then, by splitting the integral in (8.139) we obtain the following estimate:

$$\begin{aligned}\int\limits_0^1 sP(\theta^t\Lambda^t_{\mathcal{M},i} = s\delta(t+\theta^t)/t \,|\, Y^t, t)\, ds &= \int\limits_0^\varepsilon \dots\, ds + \int\limits_\varepsilon^1 \dots\, ds \\ &\le \varepsilon + P(\theta^t\Lambda^t_{\mathcal{M},i} > \varepsilon\delta(t+\theta^t)/t \,|\, Y^t, t).\end{aligned} \tag{8.140}$$

The second term in (8.140) tends to zero a.s. Y^t, $t \in (0, +\infty)$ again due to (6.34) from Theorem 6.4. Since ε is arbitrary, it follows that the integral in (8.140) is arbitrarily small when $t \to +\infty$, a.s. Y^t, $t \in (0, +\infty)$, and hence, the integral in (8.139) also converges to zero when $t \to +\infty$, a.s. Y^t, $t \in (0, +\infty)$.

Since parameter δ was chosen arbitrarily, this proves the convergence in (8.137).

Lemma is proved. □

8.12 Proof of Lemma 8.8

Proof. Let $\delta > 0$. The left hand-side of (8.117) can be estimated as follows:

$$\begin{aligned}\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] \ge &\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - B^t(\lambda)] \\ &+ \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app})] \\ &+ [B^t(\widetilde{\lambda}^t_{b,app}) - B^t(\widetilde{\lambda}^t_{app})] \\ &+ [B^t(\widetilde{\lambda}^t_{app}) - A^t(\widetilde{\lambda}^t_{app})].\end{aligned} \tag{8.141}$$

We will show that under the assumptions of Theorem 6.5 the following holds:

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] \ge \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app})] + o_{cp}(1). \tag{8.142}$$

Note that the term in right hand-side of (8.142) is expected to be positively separated from zero in view of (8.111), (8.116), and gives the main contribution for (8.117) to hold. This is described precisely by the following lemma.

Lemma 8.12. *Let $B^t(\lambda)$, $\widetilde{\lambda}^t_{b,app}$ be defined in (8.110), (8.111), respectively. Then, the following formulas hold:*

$$\begin{aligned}&B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app}) = t\cdot\left(\sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i})^2}{2\widehat{\Lambda}^t_{sc,i}} + \langle\widetilde{\mu}^t_{b,app}, \lambda\rangle\right), \\ &\lambda \in \mathbb{R}^p_+,\ \widetilde{\Lambda}^t_{b,app} = A\widetilde{\lambda}^t_{b,app},\end{aligned} \tag{8.143}$$

where

$$\widetilde{\mu}^t_{b,app} = \sum_{i\in I_1(\Lambda^*)} \frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} a_i + \sum_{i\in I_0(\Lambda^*)} a_i, \tag{8.144}$$

$$\widetilde{\mu}^t_{b,app} \in \mathbb{R}^p_+,\ \widetilde{\mu}^t_{b,app,j}\cdot\widetilde{\lambda}^t_{b,app,j} = 0 \text{ for all } j \in \{1,\dots,p\}. \tag{8.145}$$

Now, we show that (8.142) and the result of Lemma 8.12 imply the statement in (8.118). Let

$$\lambda(u,v,w) = \widetilde{\lambda}^t_{b,app} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W},\ \lambda(u,v,w) \in \mathbb{R}^p_+. \tag{8.146}$$

Using the parametrization from (8.146) and formulas (8.116), (8.143)-(8.145) we obtain

$$B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app}) = K^t(u,v,w) + R^t(u,v,w),\ \lambda = \lambda(u,v,w), \tag{8.147}$$

$$\begin{aligned} K^t(u,v,w) &= \sum_{i \in I_1(\Lambda^*)} \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}} + t\langle \widetilde{\mu}^t_{b,app},\ \lambda(u,v,w)\rangle \\ &= \sum_{i \in I_1(\Lambda^*)} \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}} + t\langle \widetilde{\mu}^t_{b,app},\ \lambda(u,v,w) - \widetilde{\lambda}^t_{b,app}\rangle \\ &= \sum_{i \in I_1(\Lambda^*)} \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}} + t\langle \widetilde{\mu}^t_{b,app},\ \frac{u}{\sqrt{t}} + \frac{v}{t}\rangle \\ &= \sum_{i \in I_1(\Lambda^*)} \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}} + \sum_{i \in I_1(\Lambda^*)} \sqrt{t} \cdot \frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} a_i^T u \\ &+ \sum_{i \in I_1(\Lambda^*)} \frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} \cdot a_i^T v + \sum_{i \in I_0(\Lambda^*)} a_i^T v, \end{aligned} \tag{8.148}$$

$$R^t(u,v,w) = \sum_{i \in I_1(\Lambda^*)} \frac{(a_i^T u)(a_i^T v)}{\sqrt{t}\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T v)^2}{2t\widehat{\Lambda}^t_{sc,i}}. \tag{8.149}$$

In particular, from the fact that $\widehat{\Lambda}^t_{sc,i} \to \Lambda^*_i$ a.s. $Y^t, t \in (0,+\infty)$, the definition of $C^t_{A,\delta}(\cdot)$ in (8.116) and (8.149) it follows that

$$\sup_{\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} |R^t(u,v,w)| = o_{cp}(1). \tag{8.150}$$

In view of formulas (8.121), (8.126), the results of lemmas 8.9, 8.11 we also find that

$$\frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} = \frac{\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} + \frac{\widehat{\Lambda}^t_{sc,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} = o_{cp}(1),\ i \in I_1(\Lambda^*). \tag{8.151}$$

Formulas (8.147)-(8.151) imply that

$$\inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app})] \geq \inf_{\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} K^t(u,v,w) + o_{cp}(1). \tag{8.152}$$

Using formulas (6.42), (8.144), (8.145), (8.148), (8.151) and the result of Lemma 8.12, we obtain the following lower bound

$$\begin{aligned} &K^t(u,v) \geq \begin{cases} c(\delta - \|v\|_1)^2, & \text{for } \|v\|_1 < \delta, \\ \sum\limits_{i \in I_0(\Lambda^*)} a_i^T v + o_{cp}(1), & \text{for } \|v\|_1 = \delta \text{ (i.e., for } \|u\|_2 = 0), \end{cases} \\ &\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}), \end{aligned} \tag{8.153}$$

where constant c is strictly positive and depends only on design A.

In addition, the Karush-Kuhn-Tucker optimality conditions in (8.144), (8.145), formula (8.151) and the definition of space $\mathcal{V}$ in (6.35) imply that

$$P(\Pi_{\mathcal{V}} \widetilde{\lambda}^t_{b,app} = 0 \,|\, Y^t, t) \to 1 \text{ when } t \to +\infty,\ \text{a.s. } Y^t, t \in (0,+\infty). \tag{8.154}$$

In view of formulas (8.153), (8.154) and the fact that $\sum_{i\in I_0(\Lambda^*)} a_i^T v \geq c\|v\|_1$ for some constant c depending on A and for all $v \in \mathcal{V}$, $\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$ when $\Pi_{\mathcal{V}}\widetilde{\lambda}^t_{b,app} = 0$ ($v \succeq 0$, when $\Pi_{\mathcal{V}}\widetilde{\lambda}^t_{b,app} = 0$), one concludes that with conditional probability tending to one a.s. Y^t, $t \in (0,+\infty)$ the following estimate holds:

$$K^t(u,v,w) \geq \begin{cases} c(\delta - \|v\|_1)^2, \ \text{ for } \|v\|_1 < \delta, \\ c\delta, \ \text{ for } \|v\|_1 = \delta \ (\text{i.e., for } \|u\|_2 = 0), \end{cases} \tag{8.155}$$
$$\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}).$$

Yet we have shown that $K^t(u,v,w)$ is positive up to errors of order $o_{cp}(1)$ (see formulas (8.153), (8.155)) which is not yet sufficient to prove (8.118) since $K^t(u,v,w)$ is not separated from zero for $\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$ (see also formulas (8.142), (8.152), (8.155)).

Next, we show that $K^t(u,v,w)$ can be bounded uniformly positively from below on $C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$ with conditional probability arbitrarily close to one, so that the bound depends, in particular, only on δ and on sequence $(\dots, \sqrt{t}(\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i})/\widehat{\Lambda}^t_{sc,i}, \dots)$, $i \in I_1(\Lambda^*)$, which is conditionally uniformly tight in view of result of Lemma 8.11 and the assumption in (6.41). After that we will show that such bound is sufficient to demonstrate (8.118).

Consider $K^t(u,v,w)$, $\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$ in the vicinity of $u = 0$.

Recall that

$$\|u\|_2 + \|v\|_1 = \delta \text{ for } \lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}),\ u \in \mathcal{U},\ v \in \mathcal{V}. \tag{8.156}$$

Let

$$\alpha = \|u\|_2 = \delta - \|v\|_1,\ \alpha \in [0,\delta],\ \text{ for } \lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{8.157}$$

So the vicinity of $u = 0$ corresponds to the vicinity of zero for parameter α in (8.157).

From (8.148), (8.151), (8.154), (8.157) and the Cauchy-Schwarz inequality it follows that

$$\begin{aligned} K^t(u,v,w) &\geq c_1\|u\|_2^2 + \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}\cdot a_i^T u + c_2\|v\|_1 + o_{cp}(1) \\ &= c_1\alpha^2 + \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}\cdot a_i^T u + c_2(\delta - \alpha) + o_{cp}(1) \\ &\geq c_2\delta - c_2\alpha - \left\| \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} a_i \right\| \cdot \alpha + o_{cp}(1), \\ &\text{for } \lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}), \end{aligned} \tag{8.158}$$

where c_1, c_2 are some positive constands depending only on A.

Let

$$\widetilde{\zeta}^t = \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}} a_i, \tag{8.159}$$

and assume that

$$\|\widetilde{\zeta}^t\| \leq r, \ \text{ for some } r \in (0,+\infty). \tag{8.160}$$

From (8.158)-(8.160) it follows that

$$K^t(u,v,w) \geq c_2\delta/2 + o_{cp}(1) \text{ for } \alpha \leq \frac{c_2\delta}{2(c_2+r)}. \tag{8.161}$$

Using (8.155), (8.157) and (8.161) we obtain the following estimate:

$$\inf_{\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} K^t(u,v,w) \geq \left(\frac{c_2\delta}{2(c_2+r)}\right)^2 + o_{cp}(1) \text{ if (8.160) holds with } r \in (0,+\infty). \tag{8.162}$$

Note that (8.162) gives a uniform lower bound on $K^t(u,v,w)$ which is now separated from zero (compare also with formula (8.155)) but depends on parameter r from the assumption in (8.160).

From (8.142), (8.152), (8.162) and the fact that parameter r is fixed, it follows that

$$P\left(\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] > 0 \,|\, Y^t, t\right) \geq P(\|\widetilde{\zeta}^t\| \leq r \,|\, Y^t, t) + o_{as}(1), \tag{8.163}$$

where $o_{as}(1)$ denotes the random value which is independent of r and converges to zero a.s. Y^t, $t \in (0,+\infty)$.

Lemma 8.13. *Let $\widetilde{\zeta}^t$ be definded by (8.159). Under the assumptions of Theorem 6.5 sequence $\widetilde{\zeta}^t$ is conditionally tight almost surely Y^t, $t \in (0,+\infty)$.*

Note that parameter r is arbitrary in (8.160) and, hence, it is also arbitrary in (8.163).

Let $\varepsilon > 0$. In view of formula (8.163) and the result of Lemma 8.13, choice $r > M(\varepsilon, \{Y^t\}_{t\in(0,+\infty)})$ results in the following estimate:

$$\liminf_{t\to+\infty} P\left(\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] > 0 \,|\, Y^t, t\right) \geq 1-\varepsilon. \tag{8.164}$$

Formula (8.164) and the fact that ε is arbitrary positive imply formula (8.118).

Now it is left to demonstrate the statement in (8.142).

Consider the first term in the left hand-side of (8.141).

Using (8.109), (8.110), the definitions in (8.111), (8.116) and the facts that $\Pi_{\mathcal{V}\oplus\mathcal{U}}(\widehat{\lambda}^t_{sc} - \lambda_*) \xrightarrow{a.s.} 0$, $\Pi_{\mathcal{V}\oplus\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \lambda_*) \xrightarrow{c.p.} 0$, one can use the Taylor expansion at $\widehat{\lambda}^t_{sc}$ up to the second order of $A(\lambda)$ to get the following estimates

$$\begin{aligned} A^t(\lambda) - B^t(\lambda) \geq & \sum_{i\in I_1(\Lambda^*)} -tC_1|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \cdot \frac{|\Lambda_i - \widehat{\Lambda}^t_{sc,i}|^2}{|\widehat{\Lambda}^t_{sc,i}|^2} + \sum_{i\in I_1(\Lambda^*)} -tC_2|\Lambda_i - \widehat{\Lambda}^t_{sc,i}|^3 \\ & + \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) + \sum_{i\in I_0(\Lambda^*)} t\widehat{\Lambda}^t_{sc,i}\log(t\widehat{\Lambda}^t_{sc,i}) - t\widehat{\Lambda}^t_{sc,i} \\ & + \beta^t(\varphi(\lambda) - \varphi(\widehat{\lambda}^t_{sc})),\ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})). \end{aligned} \tag{8.165}$$

where C_1, C_2 are some positive constants which depend only design A. More precisely, the above estimate holds with conditional probability tending to one when $t \to +\infty$ a.s. Y^t, $t \in (0,+\infty)$.

In particular, in (8.165) to bound uniformly the error-terms in the expansion we have used the following estimates

$$\sup_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} |\Lambda_i - \widehat{\Lambda}^t_{sc,i}|/|\widehat{\Lambda}^t_{sc,i}| = o_{cp}(1),\ i \in I_1(\Lambda^*), \tag{8.166}$$

$$|\log(1+x)-x| \le C_1 \cdot |x|^2, \text{ for some } C_1 > 0 \text{ for } |x| \le 1/2, \tag{8.167}$$

$$|-\widehat{s}\log(s/\widehat{s}) + (s-\widehat{s}) - \frac{s^2}{2\widehat{s}}| \le C_2 |s-\widehat{s}|^3, \tag{8.168}$$

for some $C_2 = C_2(s_*, \varepsilon) > 0$ and $|s-\widehat{s}| < \widehat{s}/2$, $|\widehat{s} - s_*| < \varepsilon$ for some fixed ε, $s_* > 0$.

Formulas (8.167), (8.168) describe the standard second order Taylor expansions of the logarithm in vicinity of $x = 0$ and $\widehat{s} = s_*$, respectively.

Formula (8.166) can be proved via the following triangle-type inequality:

$$|\Lambda_i - \widehat{\Lambda}^t_{sc,i}| \le |\Lambda_i - \widetilde{\Lambda}^t_{b,app,i}| + |\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}| + |\widetilde{\Lambda}^t_{b,i} - \Lambda^*_i| + |\Lambda^*_i + \widehat{\Lambda}^t_{sc,i}|, \ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{8.169}$$

The first term in the right hand-side of (8.169) is of order $o_{cp}(1)$ in view of the definition in (8.116) and the fact that $\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$. The last two terms are also $o_{cp}(1)$ in view of Lemma 8.4 and the fact that $\widehat{\Lambda}^t_{sc,i} \to \Lambda^*_i$ a.s. $Y^t, t \in (0,+\infty)$. Finally, from (8.151) and again the fact that $\widehat{\Lambda}^t_{sc,i} \to \Lambda^*_i$ a.s. Y^t, $t \in (0,+\infty)$, it follows that the second term in (8.169) is also of order $o_{cp}(1)$. This completes the proof of (8.166).

Using the restriction that $\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app,t})$ two first sums in (8.165) can be estimated as follows:

$$\begin{aligned}
\sum_{i\in I_1(\Lambda^*)} -tC_1|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \cdot \frac{|\Lambda_i - \widehat{\Lambda}^t_{sc,i}|^2}{(\widehat{\Lambda}^t_{sc,i})^2} &\ge \sum_{i\in I_1(\Lambda^*)} -tC_1|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \\
&\times \left(\frac{2|\Lambda_i - \widetilde{\Lambda}^t_{b,app,i}|^2}{|\widehat{\Lambda}^t_{sc,i}|^2} + \frac{2|\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}|^2}{|\widehat{\Lambda}^t_{sc,i}|^2} \right) \\
&\ge \sum_{i\in I_1(\Lambda^*)} -tC_1|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \left(\frac{c\delta^2}{t|\widehat{\Lambda}^t_{sc,i}|^2} + \frac{2|\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}|^2}{|\widehat{\Lambda}^t_{sc,i}|^2} \right),
\end{aligned} \tag{8.170}$$

where c depends only A.

Using same argument for the second sum in (8.165) we obtain the following:

$$\begin{aligned}
\sum_{i\in I_1(\Lambda^*)} -tC_2|\Lambda_i - \widehat{\Lambda}^t_{sc,i}|^3 &\ge \sum_{i\in I_1(\Lambda^*)} -8tC_2 \left(|\Lambda_i - \widetilde{\Lambda}^t_{b,app,i}|^3 + |\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}|^3 \right) \\
&\ge \sum_{i\in I_1(\Lambda^*)} -8tC_2 \left(\frac{c\delta^3}{t^{3/2}} + |\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}|^3 \right),
\end{aligned} \tag{8.171}$$

for $\lambda \in C^t_{A,\delta}(\widetilde{\Lambda}^t_{b,app,t})$, where c depends only on A.

From formulas (6.42), (8.121), (8.126), the results of Lemma 8.9 and Lemma 8.11, in particular, it follows that

$$t|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \cdot |\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}|^2 = o_{cp}(1), \tag{8.172}$$

$$t|\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}|^3 = o_{cp}(1). \tag{8.173}$$

The above formulas imply that sums in (8.170), (8.171) are bounded from below of order $o_{cp}(1)$.

The logarithmic term in (8.165) can be estimated as follows:

$$\begin{aligned}
\sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) &= \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i}) + t\widetilde{\Lambda}^t_{b,app,i}) \\
&\ge \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t|\Lambda_i - \widetilde{\Lambda}^t_{b,app,i}| + t\widetilde{\Lambda}^t_{b,app,i}) \\
&\ge \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(c\delta + t\widetilde{\Lambda}^t_{b,app,i}), \ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}).
\end{aligned} \tag{8.174}$$

where c is some positive constant depending on A.

From (8.137) in the proof of Lemma 8.7 it follows that $t\widetilde{\Lambda}^t_{b,i} = o_{cp}(1)$ for $i \in I_0(\Lambda^*)$, which together with (8.174) imply that

$$\sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) \geq \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(c\delta + o_{cp}(1)). \tag{8.175}$$

By choosing δ smaller than some fixed constant ($\delta < c/2$) in (8.175) we find that the right hand-side in (8.175) becomes negative with conditional probability tending to one a.s. Y^t, $t \in (0, +\infty)$. This, together with the above formula imply that

$$\sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) \geq o_{cp}(1),\ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}) \text{ for } \delta < c/2. \tag{8.176}$$

In addition, from (6.42) it directly follows that

$$\sum_{i\in I_0(\Lambda^*)} t\widehat{\Lambda}^t_{sc,i}\log(t\widehat{\Lambda}^t_{sc,i}) - t\widehat{\Lambda}^t_{sc,i} = o_{cp}(1). \tag{8.177}$$

Using formulas (8.165), (8.170)-(8.173), (8.176), (8.177) we obtain the following estimate:

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - B^t(\lambda)] \geq o_{cp}(1) + \beta^t \cdot \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (\varphi(\lambda) - \varphi(\widehat{\lambda}^t_{sc})). \tag{8.178}$$

Now, consider the third term in the left-hand side of (8.141).

Using formulas (8.122)-(8.124) we rewrite it as follows:

$$\begin{aligned} B^t(\widetilde{\lambda}^t_{b,app}) - B^t(\widetilde{\lambda}^t_{app}) = \widetilde{B}^t(\widetilde{\lambda}^t_{b,app}) - \widetilde{B}^t(\widetilde{\lambda}^t_{app}) \\ + \widetilde{R}^t(\widetilde{\lambda}^t_{b,app}) - \widetilde{R}^t(\widetilde{\lambda}^t_{app}). \end{aligned} \tag{8.179}$$

From (8.113), the result of Lemma 8.7, (8.122)-(8.124), (8.126), the result of lemmas 8.9, 8.11 and formula (8.179) it follows directly that

$$B^t(\widetilde{\lambda}^t_{b,app}) - B^t(\widetilde{\lambda}^t_{app}) = o_{cp}(1). \tag{8.180}$$

Now we estimate the last term in the left-hand side of (8.141). Using the same argument as in (8.165)-(8.177) one gets the following estimate:

$$\begin{aligned} B^t(\widetilde{\lambda}^t_{app}) - A^t(\widetilde{\lambda}^t_{app}) \geq & \sum_{i\in I_1(\Lambda^*)} -tC_1|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \cdot \frac{|\widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i}|^2}{|\widehat{\Lambda}^t_{sc,i}|^2} \\ & + \sum_{i\in I_1(\Lambda^*)} -tC_2|\widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i}|^3 \\ & + \sum_{i\in I_0(\Lambda^*)} t\widetilde{\Lambda}^t_{b,i}\log(\widetilde{\Lambda}^t_{app,i}) \\ & + \sum_{i\in I_0(\Lambda^*)} -t\widehat{\Lambda}^t_{sc,i}\log(t\widehat{\Lambda}^t_{sc,i}) + t\widehat{\Lambda}^t_{sc,i} \\ & - \beta^t(\varphi(\widetilde{\lambda}^t_{app}) - \varphi(\widehat{\lambda}^t_{sc})). \end{aligned} \tag{8.181}$$

$$
\begin{aligned}
&\geq \sum_{i\in I_1(\Lambda^*)} -tC_1|\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \cdot \frac{|\widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i}|^2}{|\widehat{\Lambda}^t_{sc,i}|^2} \\
&+ \sum_{i\in I_1(\Lambda^*)} -tC_2|\widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i}|^3 \\
&+ \sum_{i\in I_0(\Lambda^*)} t\widetilde{\Lambda}^t_{b,i}\log(t\widetilde{\Lambda}^t_{b,i}) \\
&+ \sum_{i\in I_0(\Lambda^*)} -t\widehat{\Lambda}^t_{sc,i}\log(t\widehat{\Lambda}^t_{sc,i}) + t\widehat{\Lambda}^t_{sc,i} \\
&- \beta^t(\varphi(\widetilde{\lambda}^t_{app}) - \varphi(\widehat{\lambda}^t_{sc,i})),
\end{aligned} \tag{8.182}
$$

where constants C_1, C_2 depend only on A. Note that to pass from formula (8.181) to (8.182) we have used the monotonicity of the logarithm, i.e., $\log(x+y) \geq \log(x)$, for any $y > 0$. The above estimate holds with conditional probability tending to one a.s. Y^t, $t \in (0, +\infty)$.

From formulas (8.112), (8.114), (8.121), (8.126), the results of Lemma 8.9 and of Lemma 8.11 it follows that

$$
t \cdot |\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}| \cdot |\widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i}|^2 = o_{cp}(1), \tag{8.183}
$$

$$
t \cdot |\widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i}|^3 = o_{cp}(1). \tag{8.184}
$$

In addition, using (8.137) in the proof of Lemma 8.7 we find that

$$
\sum_{i\in I_0(\Lambda^*)} t\widetilde{\Lambda}^t_{b,i}\log(t\widetilde{\Lambda}^t_{b,i}) = o_{cp}(1). \tag{8.185}
$$

Putting together (8.182)-(8.185) and using again (8.177) we obtain

$$
B^t(\widetilde{\lambda}^t_{app}) - A^t(\widetilde{\lambda}^t_{app}) \geq o_{cp}(1) - \beta^t \cdot (\varphi(\widetilde{\lambda}^t_{app}) - \varphi(\widehat{\lambda}^t_{sc})). \tag{8.186}
$$

Formulas (8.141), (8.178), (8.180) (8.186) imply that

$$
\begin{aligned}
\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] = &\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app})] \\
&+ \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} \beta^t(\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{app})) \\
&+ o_{cp}(1).
\end{aligned} \tag{8.187}
$$

Lemma 8.14. *Let β^t, $\varphi(\cdot)$ satisfy the assumptions of Theorem 6.5 and $\widetilde{\lambda}^t_{b,app}$, $\widetilde{\lambda}^t_{app}$ be defiend in (8.111), (8.112), respectively. Then,*

$$
\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} \beta^t \cdot (\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{app})) = o_{cp}(1). \tag{8.188}
$$

Formula (8.142) directly follows from (8.187) and the result of Lemma 8.14.
Lemma is proved. □

8.13 Proof of Lemma 8.9

Proof. To prove the claim we use essentially the same convexity argument as one in (8.117) and by Lemma 8.8.

Let $\delta > 0$.

Let also

$$\widetilde{\lambda}^t = \widehat{\lambda}^t_{sc} + \frac{\widetilde{u}^t}{\sqrt{t}} + \frac{\widetilde{v}^t}{t} + \widetilde{w}^t, \tag{8.189}$$

$$\lambda(u,v,w) = \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ (u,v,w) \in \mathcal{U}\times\mathcal{V}\times\mathcal{W}, \tag{8.190}$$

where $(\widetilde{u}^t, \widetilde{v}^t)$ is defined in (8.125) and $\widetilde{w}^t$ is any vector from $\mathcal{W}$ such that $\widetilde{\lambda}^t \succeq 0$. Restrictions on (u,v,w) in (8.190) are such that $\lambda(u,v,w) \succeq 0$.

Recall that

$$\|u - \widetilde{u}^t\|_2 + \|v - \widetilde{v}^t\|_1 = \delta \text{ for } \lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t). \tag{8.191}$$

where $C^t_{A,\delta}(\cdot)$ is defined in (8.116).

Next we show that

$$P(\inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} [B^t(u,v) - B^t(\widetilde{u}^t,\widetilde{v}^t)] > 0 \,|\, Y^t, t) \to 1 \text{ when } t\to+\infty, \text{ a.s. } Y^t, t \in (0,+\infty) \tag{8.192}$$

which together with the fact that δ is arbitrary and convexity of $B^t(u,v)$ in (u,v) implies the claim of the lemma.

Using formulas (8.122)-(8.124) we obtain

$$\begin{aligned} B^t(u,v) - B^t(\widetilde{u}^t,\widetilde{v}^t) = [\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t)] \\ + [\widetilde{R}^t(u,v) - \widetilde{R}^t(\widetilde{u}^t,\widetilde{v}^t)], \\ (u,v) \text{ s.t. } \exists w\in\mathcal{W},\ \lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t). \end{aligned} \tag{8.193}$$

From the facts that $\widetilde{\Lambda}^t_{b,i} \xrightarrow{c.p.} \Lambda^*_i$ (by Lemma 8.4), $\widehat{\Lambda}^t_{sc,i} \xrightarrow{a.s.} \Lambda^*_i$ for $i\in\{1,\dots,d\}$ (see (6.41), (6.42) and (B.2) in Appendix B), the conditional tightness of $\widetilde{u}^t$ (by Lemma 8.11) and formulas (6.42), (8.126), (8.191) it follows that

$$\sup_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} |\widetilde{R}^t(u,v) - \widetilde{R}^t(\widetilde{u}^t,\widetilde{v}^t)| = o_{cp}(1), \tag{8.194}$$

where $\widetilde{R}^t(\cdot)$ is defined in (8.124).

Formulas (8.193), (8.194) imply that

$$\inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} [B^t(u,v) - B^t(\widetilde{u}^t,\widetilde{v}^t)] = \inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} [\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t)] + o_{cp}(1). \tag{8.195}$$

Now, note that the minimized function in (8.129) coincides with $\widetilde{B}^t(u,v)$ up to a linear term depending on v. Therefore, the difference in the right hand-side of (8.195) for terms depending on $u\in\mathcal{U}$ can be estimated through optimality conditions for the problem in (8.129).

Since the positivity constraints in (8.129) include restrictions on $u\in\mathcal{U}$ and also depend on $w\in\mathcal{W}$, for simplicity, we include w in the minimization problem as an independent variable

$$(\widetilde{u}^t,\widetilde{w}^t) = \underset{\substack{(u,w):(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0 \\ u\in\mathcal{U},\, w\in\mathcal{W}}}{\arg\min} \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{a_i^T u}{\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}}. \tag{8.196}$$

Clearly, minimizer $\widetilde{u}^t$ in (8.196) coincides with the original solution from (8.129). The problem in (8.196) is convex and the strong duality is satisfied (e.g., by Slater's condition).

From the Karush-Kuhn-Tucker necessary optimality conditions (see e.g., [Ber97], Section 3.3) for the optimization problem in (8.196) and the strong duality it follows that

$$\exists\, \widetilde{\mu}^t \succeq 0,\ \widetilde{\mu}^t \in \mathcal{W}^\perp, \tag{8.197}$$

$$\sum_{i\in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}) \cdot \frac{\Pi_{\mathcal{U}} a_i}{\widehat{\Lambda}^t_{sc,i}} + \frac{\Pi_{\mathcal{U}} a_i a_i^T \widetilde{u}^t}{\widehat{\Lambda}^t_{sc,i}} = \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}},\ \widetilde{\mu}^t_{\mathcal{U}} = \Pi_{\mathcal{U}} \widetilde{\mu}^t, \tag{8.198}$$

$$\widetilde{\mu}^t_j \left([(I - \Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}]_j + \frac{\widetilde{u}^t_j}{\sqrt{t}} + \widetilde{w}^t_j \right) = 0,\ j \in \{1, \dots, p\}, \tag{8.199}$$

where $(\widetilde{u}^t, \widetilde{w}^t)$ are defined in (8.196). Strong duality implies, in particular, that $\widetilde{\mu}^t$ is a solution for the dual problem and $\widetilde{\mu}^t \in \mathcal{W}^\perp$ (dual functional equals $-\infty$ for $\widetilde{\mu}^t \notin \mathcal{W}^\perp$). Note also that the optimized functional in (8.196) is strongly convex in u, so $\widetilde{u}^t$ is always unique, whereas at least one $\widetilde{w}^t$ always exists may not be unique. The latter fact does not pose any problem since the target functional is flat for $w \in \mathcal{W}$, so if not said otherwise, we choose any solution $\widetilde{w}^t$ in (8.196) so that positivity constraints are satisfied.

From (8.122), (8.197)-(8.199) it follows that

$$\begin{aligned}
\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t, \widetilde{v}^t) &= \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}\frac{\widetilde{\Lambda}_{b,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i^T(u - \widetilde{u}^t) + \frac{1}{2}\frac{(a_i^T u)^2 - (a_i^T \widetilde{u}^t)^2}{\widehat{\Lambda}^t_{sc,i}} \\
&+ \sum_{i\in I_0(\Lambda^*)} a_i^T(v - \widetilde{v}^t) \\
&= \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}\frac{\widetilde{\Lambda}_{b,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i^T(u - \widetilde{u}^t) + \frac{1}{2}\frac{(a_i^T(u - \widetilde{u}^t))^2}{\widehat{\Lambda}^t_{sc,i}} \\
&+ \frac{(\widetilde{u}^t)^T a_i a_i^T(u - \widetilde{u}^t)}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} a_i^T(v - \widetilde{v}^t) \\
&= \langle \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}}, u - \widetilde{u}^t \rangle + \frac{1}{2} \sum_{i\in I_1(\Lambda^*)} \frac{|a_i^T(u - \widetilde{u}^t)|^2}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} a_i^T(v - \widetilde{v}^t).
\end{aligned} \tag{8.200}$$

Note that

$$\langle \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}}, u - \widetilde{u}^t \rangle \geq 0, \tag{8.201}$$

$$v - \widetilde{v}^t \succeq 0, \tag{8.202}$$

$$\text{for } (u,v) \in \mathcal{U} \times \mathcal{V} \text{ s.t. } \lambda(u,v,w) = \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w \succeq 0 \text{ for some } w \in \mathcal{W}.$$

Indeed, in view of (8.197), (8.199) the left hand-side in (8.201) can be rewritten as follows:

$$\begin{aligned}
\langle \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}}, u - \widetilde{u}^t \rangle &= \langle \widetilde{\mu}^t_{\mathcal{U}}, \frac{u}{\sqrt{t}} - \frac{\widetilde{u}^t}{\sqrt{t}} \rangle \\
&= \langle (I - \Pi_{\mathcal{V}})\widetilde{\mu}^t, (I - \Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} \rangle \\
&= \langle (I - \Pi_{\mathcal{V}})\widetilde{\mu}^t, \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w \rangle.
\end{aligned} \tag{8.203}$$

Note also that from (8.197) and the definition of $\mathcal{V}$ in (6.35) it follows that

$$\mu^t_{\mathcal{U}} = (I - \Pi_{\mathcal{V}})\mu^t \succeq 0. \tag{8.204}$$

Formula (8.201) follows directly from (8.203), (8.204).

In turn, formula (8.202) follows from (8.126).

Formulas (8.191), (8.200)- (8.202) and the fact that $\widehat{\Lambda}_{sc,i} \to \Lambda^*_i$ for $i \in \{1,\dots,d\}$ a.s. Y^t, $t \in (0,+\infty)$ (as a strongly consistent estimator), imply that with conditional probability tending to one a.s. Y^t, $t \in (0,+\infty)$ the following estimate holds:

$$\inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widehat{\lambda}^t)} [\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t)] \geq c\delta^2, \tag{8.205}$$

where c is some fixed positive constant depending only on Λ^* and A.

Formula (8.192) follows directly from (8.194), (8.205).

Lemma is proved. □

8.14 Proof of Lemma 8.10

Let $\xi \in \mathbb{R}^{\#I_1(\Lambda^*)}$ be a parameter and consider $\widetilde{u}^t(\xi)$ defined in (8.131).

Since the positivity constraints in (8.131) include restrictions on $u \in \mathcal{U}$ and $w \in \mathcal{W}$, for simplicity, we include w in the minimization problem as an independent variable

$$(\widetilde{u}^t,\widetilde{w}^t) = \underset{\substack{(u,w):(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0\\ u\in\mathcal{U},w\in\mathcal{W}}}{\arg\min} -\xi^T C^t u + \frac{1}{2}u^T F^t u, \tag{8.206}$$

where

$$\begin{aligned} &C^t = (\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2} A_{I_1(\Lambda^*)},\ F^t = \widehat{F}^t_{I_1(\Lambda^*)},\\ &\widehat{D}^t_{I_1(\Lambda^*)},\ \widehat{F}^t_{I_1(\Lambda^*)} \text{ are defined in (8.132), (8.133).} \end{aligned} \tag{8.207}$$

The Lagrangian function for the primal problem in (8.206) is defined by the formula:

$$\mathcal{L}^t(u,w;\mu) = -\xi^T C^t u + \frac{1}{2}u^T F^t u - \mu^T((1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + w), \tag{8.208}$$

$$u \in \mathcal{U},\ w \in \mathcal{W},\ \mu \succeq 0. \tag{8.209}$$

The dual function for $G^t(\mu)$ and solution μ^t for the dual problem are defined by the formulas:

$$G^t(\mu) = \inf_{u\in\mathcal{U},\,w\in\mathcal{W}} \mathcal{L}^t(u,w;\mu),\ \mu^t = \arg\max_{\mu\succeq 0} G^t(\mu). \tag{8.210}$$

From the Karush-Kuhn-Tucker necessary optimality conditions (see e.g., [Ber97], Section 3.3), the fact that the primal problem is strongly convex in $u \in \mathcal{U}$ and the strong duality it follows that

$$\exists (u^t,w^t) \in \mathcal{U}\times\mathcal{W},\ \mu^t \succeq 0,\ \mu^t \in \mathcal{W}^{\perp} \text{ s.t.} \tag{8.211}$$

$$(u^t,w^t) \text{ is a solution for the primal problem in (8.206),} \tag{8.212}$$

$$\mu^t = \mu^t(\xi) \text{ is a solution for the dual problem in (8.210),} \tag{8.213}$$

$$\nabla_{u,w}\mathcal{L}^t(u^t,w^t;\mu^t) = 0, \tag{8.214}$$

$$((1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc,j} + \frac{u^t_j}{\sqrt{t}} + w^t_j)\cdot\mu^t_j = 0,\ j \in \{1,\dots,p\}. \tag{8.215}$$

Using formulas (8.208), (8.214) we obtain the following:

$$-\Pi_{\mathcal{U}}(C^t)^T\xi + (\Pi_{\mathcal{U}}F^t\Pi_{\mathcal{U}})u^t - \frac{\Pi_{\mathcal{U}}\mu^t(\xi)}{\sqrt{t}} = 0, \tag{8.216}$$

$$\Pi_{\mathcal{W}}\mu^t = 0, \tag{8.217}$$

where $\Pi_{\mathcal{U}}$, $\Pi_{\mathcal{W}}$ are defined in (6.38).

Let

$$C^t_{\mathcal{U}} = C^t\Pi_{\mathcal{U}},\ F^t_{\mathcal{U}} = (\Pi_{\mathcal{U}}F^t\Pi_{\mathcal{U}}),\ \mu^t_{\mathcal{U}} = \Pi_{\mathcal{U}}\mu^t. \tag{8.218}$$

Note that formulas (6.40)-(6.42), Assumption 1 and the Continuous Mapping Theorem imply that

$$C^t_{\mathcal{U}} \to C^*_{\mathcal{U}},\ F^t_{\mathcal{U}} \to F^*_{\mathcal{U}} \text{ when } t \to +\infty,\ \text{ a.s. } Y^t,\ t \in (0, +\infty), \tag{8.219}$$

where

$$C^*_{\mathcal{U}} = \Pi_{\mathcal{U}}C^*,\ F^*_{\mathcal{U}} = \Pi_{\mathcal{U}}F^*\Pi_{\mathcal{U}}, \tag{8.220}$$

$$C^* = (D_{I_1(\Lambda^*)})^{-1/2}A_{I_1(\Lambda^*)},\ D_{I_1(\Lambda^*)} = \mathrm{diag}(\dots, \Lambda^*_i, \dots),\ i \in I_1(\Lambda^*), \tag{8.221}$$

$$F^* = \sum_{i \in I_1(\Lambda^*)} \frac{a_i a_i^T}{\Lambda^*_i} = A_{I_1(\Lambda^*)})^T D^{-1}_{I_1(\Lambda^*)} A_{I_1(\Lambda^*)}. \tag{8.222}$$

Using notations from (8.218) formula (8.216) can be rewritten as follows:

$$u^t(\xi) = (F^t_{\mathcal{U}})^{-1}(C^t_{\mathcal{U}})^T\xi + (F^t_{\mathcal{U}})^{-1}\frac{\mu^t_{\mathcal{U}}(\xi)}{\sqrt{t}}. \tag{8.223}$$

Next, we show that the following estimate always holds:

$$\|\frac{\mu^t_{\mathcal{U}}(\xi)}{\sqrt{t}}\| \le 2 \max_{\sigma \in \sigma_{\mathcal{U}}(F^t_{\mathcal{U}})} \sigma^{-1/2} \cdot \|(F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\|_{\mathbb{R}^{\#I_1(\Lambda^*)} \to \mathcal{U}} \cdot \|\xi\|, \tag{8.224}$$

where $\sigma_{\mathcal{U}}(F^t_{\mathcal{U}})$ denotes the spectrum of $F^t_{\mathcal{U}}$ on $\mathcal{U}$ (which in view of (8.219), (8.222) contains only non-zero positive elements starting from some $t \ge t_0$).

We begin with characterzation of mapping $\mu^t_{\mathcal{U}}(\xi)$.

First, note that from (8.208), (8.210) it follows that

$$G^t(\mu) = -\infty \text{ if } \mu \notin \mathcal{W}^\perp. \tag{8.225}$$

That is for $\mu \notin \mathcal{W}^\perp$ the dual problem is unfeasible. In view of this and the strong duality, formulas in (8.210) can be rewritten as follows:

$$G^t(\mu) = \inf_{u \in \mathcal{U}} \mathcal{L}^t(u, 0; \mu),\ \mu \succeq 0,\ \mu \in \mathcal{W}^\perp, \tag{8.226}$$

$$\mu^t = \operatorname*{arg\,max}_{\mu \succeq 0,\, \mu \in \mathcal{W}^\perp} G^t(\mu). \tag{8.227}$$

Using (8.208), (8.218) the first order optimality condition in (8.226) has the following form:

$$\begin{aligned} &u^t_{min}(\mu) = (F^t_{\mathcal{U}})^{-1}(C^t_{\mathcal{U}})^T\xi + (F^t_{\mathcal{U}})^{-1}\frac{\mu_{\mathcal{U}}}{\sqrt{t}}, \\ &\mu_{\mathcal{U}} = \Pi_{\mathcal{U}}\mu,\ \mu \succeq 0,\ \mu \in \mathcal{W}^\perp. \end{aligned} \tag{8.228}$$

From (8.208), (8.210), (8.226), (8.228) it follows that

$$\begin{aligned} G^t(\mu) = \mathcal{L}^t(u^t_{min}(\mu), 0; \mu) = -\xi^T C^t_{\mathcal{U}} u^t_{min}(\mu) + \frac{1}{2}[u^t_{min}(\mu)]^T F^t_{\mathcal{U}} u^t_{min}(\mu) \\ - \mu^T((1 - \Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u^t_{min}(\mu)}{\sqrt{t}}), \\ \mu_{\mathcal{U}} = \Pi_{\mathcal{U}}\mu,\ \mu \succeq 0,\ \mu \in \mathcal{W}^\perp. \end{aligned} \tag{8.229}$$

Formulas (8.228), (8.229) imply that

$$G^t(\mu) = -\frac{1}{2}\frac{\mu_{\mathcal{U}}^T}{\sqrt{t}}(F_{\mathcal{U}}^t)^{-1}\frac{\mu_{\mathcal{U}}}{\sqrt{t}} - \xi^T C_{\mathcal{U}}^t (F_{\mathcal{U}}^t)^{-1}\frac{\mu_{\mathcal{U}}}{\sqrt{t}} - \mu^T(I-\Pi_{\mathcal{V}})\widehat{\lambda}_{sc}^t, \\ \mu \succeq 0,\ \mu \in \mathcal{W}^{\perp}. \tag{8.230}$$

Note that from the facts that $\mu \in \mathcal{W}^{\perp}$, $\mu \succeq 0$ and the definition of $\mathcal{V}$ in (6.35) it follows that

$$\mu_{\mathcal{U}} = (I-\Pi_{\mathcal{V}})\mu = \begin{cases} \mu_j, \text{ if } \sum\limits_{i\in I_0(\Lambda^*)} a_{ij} = 0, \\ 0, \text{ otherwise,} \end{cases} \Rightarrow \mu_{\mathcal{U}} = (I-\Pi_{\mathcal{V}})\mu \succeq 0. \tag{8.231}$$

From (8.231) and the fact that $\widehat{\lambda}_{sc}^t \succeq 0$ it follows that

$$\mu^T(I-\Pi_{\mathcal{V}})\widehat{\lambda}_{sc}^t = [(I-\Pi_{\mathcal{V}})\mu]^T\widehat{\lambda}_{sc}^t = \mu_{\mathcal{U}}^T\widehat{\lambda}_{sc}^t \geq 0. \tag{8.232}$$

From (8.230) one can see that solution μ^t in (8.227) may not be unique, however, its projection $\mu_{\mathcal{U}}^t$ is unique since functional $G^t(\mu)$ is strongly convex in $\mu_{\mathcal{U}}$. At the same time, from (8.223) it follows that only $\mu_{\mathcal{U}}^t$ is essential. In view of (8.223), (8.230), the optimization problem in (8.227) can be rewritten as follows:

$$\frac{\mu_{\mathcal{U}}^t}{\sqrt{t}} = \widetilde{\mu}_{\mathcal{U}}^t = \underset{\mu_{\mathcal{U}}\in\Pi_{\mathcal{U}}(\mathbb{R}_+^p\cap\mathcal{W}^{\perp})}{\arg\min} \frac{1}{2}\|(F_{\mathcal{U}}^t)^{-1/2}\mu_{\mathcal{U}} + (F_{\mathcal{U}}^t)^{-1/2}(C_{\mathcal{U}}^t)^T\xi\|^2 + \sqrt{t}\mu_{\mathcal{U}}^T\widehat{\lambda}_{sc}^t. \tag{8.233}$$

From (8.233) and the fact that $0 \in \Pi_{\mathcal{U}}(\mathbb{R}_+^p \cap \mathcal{W}^{\perp})$ it follows that

$$\frac{1}{2}\|(F_{\mathcal{U}}^t)^{-1/2}\widetilde{\mu}_{\mathcal{U}}^t + (F_{\mathcal{U}}^t)^{-1/2}(C_{\mathcal{U}}^t)^T\xi\|^2 + \sqrt{t}\mu_{\mathcal{U}}^t\widehat{\lambda}_{sc}^t \leq \|(F_{\mathcal{U}}^t)^{-1/2}(C_{\mathcal{U}}^t)^T\xi\|^2. \tag{8.234}$$

Formulas (8.232), (8.234) imply that

$$\|(F_{\mathcal{U}}^t)^{-1/2}\widetilde{\mu}_{\mathcal{U}}^t + (F_{\mathcal{U}}^t)^{-1/2}(C_{\mathcal{U}}^t)^T\xi\| \leq \|(F_{\mathcal{U}}^t)^{-1/2}(C_{\mathcal{U}}^t)^T\xi\|. \tag{8.235}$$

which together with inequality $\|a+b\| \geq \|a\| - \|b\|$ (a, b any vectors in $\mathbb{R}^p$) imply the following estimate

$$\|(F_{\mathcal{U}}^t)^{-1/2}\widetilde{\mu}_{\mathcal{U}}^t\| \leq 2\|(F_{\mathcal{U}}^t)^{-1/2}(C_{\mathcal{U}}^t)^T\xi\|. \tag{8.236}$$

From (6.36), (8.218), (8.219), (8.222) it follows that $F_{\mathcal{U}}^t$ is of full rank on $\mathcal{U}$ (starting from some $t \geq t_0$), therefore, for large t matrix $(F_{\mathcal{U}}^t)^{-1/2}$ is positive definite, injective on $\mathcal{U}$ and, hence, $\|(F_{\mathcal{U}}^t)^{-1/2}\widetilde{\mu}_{\mathcal{U}}^t\| \geq \min_{\sigma\in\sigma_{\mathcal{U}}(F_{\mathcal{U}}^t)}\sigma^{1/2}\cdot\|\widetilde{\mu}_{\mathcal{U}}^t\|$, where $\sigma_{\mathcal{U}}(F_{\mathcal{U}}^t)$ denotes the spectrum.

The above argument with formula (8.236) directly imply (8.224).

Formulas (8.134), (8.135) follows from (8.218), (8.223), (8.224).

Lemma is proved.

8.15 Proof of Lemma 8.11

Proof. In view of step 3 in Algorithm 5 intensities $\widetilde{\Lambda}_{b,i}^t$ can be represented as follows:

$$\widetilde{\Lambda}_{b,i}^t = \frac{1}{\theta^t + t}\sum_{k=1}^{Y_i^t} w_{ik} + r_{b,\mathcal{M},i}^t,\ i \in I_1(\Lambda^*), \tag{8.237}$$

$$\{w_{ik}\}_{k=1,\,i=1}^{\infty,\,d} \text{ are mutually independent, } w_{ik} \sim \Gamma(1,1), \tag{8.238}$$

where

$$r^t_{b,\mathcal{M},i}|\Lambda^t_{\mathcal{M},i}, Y^t, t \sim \Gamma(\theta^t\Lambda^t_{\mathcal{M},i}, (\theta^t+t)^{-1}),$$
$$\Lambda^t_{\mathcal{M},i} \text{ are sampled in Algorithm 4.} \tag{8.239}$$

In particular,

$$\sqrt{t}\cdot r^t_{b,\mathcal{M},i} = o_{cp}(1). \tag{8.240}$$

Indeed, from (8.73), (8.239) and the Markov inequality it holds that

$$\begin{aligned} P(\sqrt{t}\cdot r^t_{b,\mathcal{M},i} > \delta \,|\, Y^t, t) &\le \frac{\sqrt{t}\cdot\theta^t}{\delta(\theta^t+t)}\mathbb{E}[\Lambda^t_{\mathcal{M},i}|Y^t,t] \\ &\le \frac{\sqrt{t}\cdot\theta^t}{\delta(\theta^t+t)}\sum_{i\in I_1(\Lambda^*)}\frac{Y^t_i}{t} \to 0 \text{ when } t\to+\infty, \text{ a.s. } Y^t, t\in(0,+\infty), \end{aligned} \tag{8.241}$$

where δ is arbitrary positive value.

Using the Central Limit Theorem for sums of w_{ik} in (8.237), (8.238) and the Strong Law of Large Numbers for Y^t (see Theorem B.1, formula (B.2) in Appendix) we obtain:

$$\frac{\sqrt{t}}{(\theta^t+t)\sqrt{Y^t_i/t}}\sum_{k=1}^{Y^t_i}(w_{ik}-1) \xrightarrow{c.d.} \mathcal{N}(0,1) \text{ when } t\to+\infty, \text{ a.s. } Y^t, t\in(0,+\infty). \tag{8.242}$$

Due to mutual independence between w_{ik}, the above convergence holds for all $i\in I_1(\Lambda^*)$ simultaneously as for the vector in $\mathbb{R}^{\#I_1(\Lambda^*)}$.

Using formula (8.136) we obtain:

$$\begin{aligned} A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\widetilde{\xi}^t &= \sum_{i\in I_1(\Lambda^*)}\sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i \\ &= \sum_{i\in I_1(\Lambda^*)}\sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,i}-Y^t_i/t}{\widehat{\Lambda}^t_{sc,i}}a_i + \sum_{i\in I_1(\Lambda^*)}\sqrt{t}\cdot\frac{Y^t_i/t-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i. \end{aligned} \tag{8.243}$$

The first sum is conditionally tight in view of the Prokhorov theorem on tightness of weakly convergence sequences and the result in (8.242). Due to (6.41) the second sum is simply bounded for large t for almost any trajectory Y^t, $t\in(0,+\infty)$. These arguments directly imply conditional tightness of $A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\widetilde{\xi}^t$ for almost any trajectory Y^t, $t\in(0,+\infty)$. Statement (i) of the lemma is proved.

Statement (ii) follows directly from (i) and the result of Lemma 8.10. Indeed, this follows, in particular, from the fact that $\Lambda_{sc}\xrightarrow{a.s.}\Lambda_*$, the coefficient c_t in (8.135) is separated from zero for large t and has a limit $c_*>0$ which coincides with c_t where $\widehat{F}^t_{I_1(\Lambda^*)}$ is replaced with its limit $F^*_{I_1(\Lambda^*)} = \sum\limits_{i\in I_1(\Lambda^*)}\dfrac{a_ia_i^T}{\Lambda^*_i}$.

Lemma is proved. □

8.16 Proof of Lemma 8.12

Proof. Since $B^t(\lambda)$ is proportional to t in (8.110), it suffices to prove formula (8.143) for normalized process $B^t(\lambda)/t$ which we denote here by $G^t(\lambda)$, that is

$$\begin{aligned} G^t(\lambda) = &\sum_{i\in I_1(\Lambda^*)} -(\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i})\frac{\Lambda_i-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)}\frac{(\Lambda_i-\widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} \\ &+ \sum_{i\in I_0(\Lambda^*)}\Lambda_i,\ \Lambda_i = a_i^T\lambda,\ i\in\{1,\dots,d\}. \end{aligned} \tag{8.244}$$

Note also that minimizers of B^t and of G^t coincide.

From the necessary Karush-Kuhn-Tucker optimality conditions in (8.111) (see e.g., [Ber97], Section 3.3) it follows that

$$\exists \widetilde{\lambda}^t_{b,app}, \widetilde{\mu}^t_{b,app} \in \mathbb{R}^p_+ \text{ such that}$$

$$-\sum_{i\in I_1(\Lambda^*)} \frac{\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i + \sum_{i\in I_1(\Lambda^*)} \frac{\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i + \sum_{i\in I_0(\Lambda^*)} a_i - \widetilde{\mu}^t_{b,app} = 0, \tag{8.245}$$

$$\widetilde{\Lambda}^t_{b,app} = A\widetilde{\lambda}^t_{b,app},$$

$$\widetilde{\mu}^t_{b,app,j} \cdot \widetilde{\lambda}^t_{b,app,j} = 0 \text{ for all } j \in \{1,\dots,p\}. \tag{8.246}$$

Multiplying both sides of (8.245) on $(\widetilde{\lambda}^t_{b,app} - \widehat{\lambda}^t_{sc})$ and using formula (8.246) we obtain following formulas:

$$-\langle \widetilde{\mu}^t_{b,app}, \widehat{\lambda}^t_{sc}\rangle = -\sum_{i\in I_1(\Lambda^*)} \frac{(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})(\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_1(\Lambda^*)} \frac{(\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}}$$
$$+ \sum_{i\in I_0(\Lambda^*)} \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}, \tag{8.247}$$

$$-\langle \widetilde{\mu}^t_{b,app}, \widehat{\lambda}^t_{sc}\rangle = \sum_{i\in I_1(\Lambda^*)} \widetilde{\Lambda}^t_{b,i} - \widetilde{\Lambda}^t_{b,app,i} - \sum_{i\in I_0(\Lambda^*)} \widehat{\Lambda}^t_{sc,i}. \tag{8.248}$$

From formulas (8.244), (8.245), (8.247) it follows that

$$G^t(\widetilde{\lambda}^t_{b,app}) = -\langle \widetilde{\mu}^t_{b,app}, \widehat{\lambda}^t_{sc}\rangle - \frac{1}{2}\sum_{i\in I_1(\Lambda^*)} \frac{(\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} \widehat{\Lambda}^t_{sc,i}. \tag{8.249}$$

Using (8.244)-(8.249) we get the following identity:

$$\begin{aligned}
G^t(\lambda) - G^t(\widetilde{\lambda}^t_{b,app}) &= \sum_{i\in I_1(\Lambda^*)} -(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{\Lambda_i - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} \Lambda_i - \widehat{\Lambda}^t_{sc,i} \\
&+ \frac{1}{2}\sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widehat{\Lambda}^t_{sc,i})^2 + (\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} + \langle \widetilde{\mu}^t_{b,app}, \widehat{\lambda}^t_{sc}\rangle \\
&= \sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i})^2}{2\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_1(\Lambda^*)} \frac{(\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})(\Lambda_i - \widehat{\Lambda}^t_{sc,i})}{\widehat{\Lambda}^t_{sc,i}} \\
&+ \sum_{i\in I_0(\Lambda^*)} \Lambda_i - \widehat{\Lambda}^t_{sc,i} + \sum_{i\in I_1(\Lambda^*)} \widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i} + \sum_{i\in I_0(\Lambda^*)} \widehat{\Lambda}^t_{sc,i} \\
&- \sum_{i\in I_1(\Lambda^*)} (\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{\Lambda_i - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} \\
&= \sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i})^2}{2\Lambda^*_i} + \sum_{i\in I_0(\Lambda^*)} \Lambda_i + \sum_{i\in I_1(\Lambda^*)} \frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}\Lambda_i.
\end{aligned} \tag{8.250}$$

Formulas (8.143)-(8.145) follow from (8.244) (8.245), (8.246), (8.250).

Lemma is proved. □

8.17 Proof of Lemma 8.13

Proof. Formula (8.159) can be rewritten as follows:

$$\widetilde{\zeta}^t = \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}a_i = \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i + \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widehat{\Lambda}^t_{sc,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}a_i. \tag{8.251}$$

The second sum is conditionally uniformly tight a.s. Y^t, $t\in(0,+\infty)$ in view of the result of Lemma 8.11(i).

Using formula (8.121) and the result of Lemma 8.11(ii) we find that

$$\sum_{i\in I_1(\Lambda^*)} \sqrt{t}\cdot\frac{\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} = \sum_{i\in I_1(\Lambda^*)} \frac{a_i^T\widetilde{u}^t(\widetilde{\xi}^t)}{\widehat{\Lambda}^t_{sc,i}}, \tag{8.252}$$

where $\widetilde{u}^t$ is defined in (8.129), $\widetilde{\xi}^t = (\dots,\sqrt{t}(\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i})/\sqrt{\widehat{\Lambda}^t_{sc,i}},\dots)$, $i\in I_1(\Lambda^*)$. By the result of Lemma 8.11(ii) the expression in (8.252) is also conditionally tight a.s. Y^t, $t\in(0,+\infty)$.

Conditional tightness of ζ^t a.s. Y^t, $t\in(0,+\infty)$ the from (8.251), (8.252) and the above arguments.

Lemma is proved. □

8.18 Proof of Lemma 8.14

Proof. Consider the following formula

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}[\varphi(\lambda)-\varphi(\widetilde{\lambda}^t_{app})] = \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}[\varphi(\lambda-\varphi(\widetilde{\lambda}^t_{b,app})] + [\varphi(\widetilde{\lambda}^t_{b,app})-\varphi(\widetilde{\lambda}^t_{app})]. \tag{8.253}$$

Recall that $\widetilde{\lambda}^t_{b,app}$ may not be chosen uniquely since the functional $B^t(\lambda)$ is strongly convex only in directions from $\mathrm{Span}\{a_i : i\in I_1(\Lambda^*)\}$ (see formula (8.110)) and it is flat in directions from $\ker A$. From the strong convexity of $B^t(\lambda)$ on $\mathrm{Span}\{a_i : i\in I_1(\Lambda^*)\}$ and formulas (8.110), (8.111), (8.121) it follows that

$$\widetilde{u}^t_{b,app} = \sqrt{t}\cdot\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app}-\widehat{\lambda}^t_{sc}) \text{ is unique.} \tag{8.254}$$

At the same time, from (6.42), (8.126) and the result of Lemma 8.9 it follows that

$$\widetilde{v}^t_{b,app} = t\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app}-\widehat{\lambda}^t_{sc}) = o_{cp}(1), \tag{8.255}$$

where the above formula is understood as a uniform bound on the set of all possible minimizers $\widetilde{\lambda}^t_{b,app}$. We may assume that for each t there is some unique $\widetilde{v}^t_{b,app}$. Then, to choose uniquely $\widetilde{\lambda}^t_{b,app}$ one has to fix its projection onto $\mathcal{W}$ regarding the positivity constraints.

Consider the following mapping

$$\begin{aligned} w(u,v) &= \operatorname*{arg\,min}_{\substack{w:\lambda_*+u+v+w\geq 0\\ w\in\mathcal{W}}} \varphi(\lambda_*+u+v+w),\\ u\in\mathcal{U},\ v\in\mathcal{V} &: (\lambda_*+u+v+\mathcal{W})\cap\mathbb{R}^p_+\neq\emptyset. \end{aligned} \tag{8.256}$$

From the strict convexity of $\varphi(\cdot)$ along $\ker A$ (by the assumption in (2.16)), the definition of $\mathcal{W}$ in (6.37) and the result of Lemma 2.1 it follows that mapping $w(u,v)$ is one-to-one and continuous in (u,v) on its domain of definition.

Note also that

$$w(0,0) = w_*,\ w_* = \underset{\substack{w:\lambda_*+w\succeq 0\\ w\in\ker A}}{\arg\min}\ \varphi(\lambda_* + w), \tag{8.257}$$

where $w(\cdot,\cdot)$ is defined in (8.256), w_* appears in Theorems 6.1, 6.2.

The fact that $w_* \in \mathcal{W}$ can be proved by the contradiction argument. Assume that $w_* \in \ker A$ but $w_* \notin \mathcal{W}$, $w_* \neq 0$. Then, from the defintion of $\mathcal{V}$, $\mathcal{U}$, $\mathcal{W}$ it follows that

$$\exists i \in I_0(\Lambda^*),\ j \in \{1,\dots,p\} : a_{ij} > 0,\ w_{*j} > 0. \tag{8.258}$$

At the same time from the fact that $w_* \in \ker A$ it follows that

$$0 = \sum_{i\in I_0(\Lambda^*)} a_i^T w_* = \sum_{j=1}^{p} \left(\sum_{i\in I_0(\Lambda^*)} a_{ij} \right) w_{*j} \tag{8.259}$$

Formulas (8.258), (8.259) imply that

$$\exists i' \in I_0(\Lambda^*),\ j' \in \{1,\dots,p\} : a_{i'j'} > 0,\ w_{*j'} < 0. \tag{8.260}$$

At the same time, from the definition of $I_0(\Lambda^*)$ in (2.2) it follows that $\lambda_{*j'} = 0$ which together with the results from (8.260) contradicts the positivity constraint in (8.257). Thus, $w_* \in \mathcal{W}$.

Let

$$\widetilde{w}^t_{b,app} = w\left(\Pi_{\mathcal{U}}(\widehat{\lambda}^t_{sc} - \lambda_*) + \frac{\widetilde{u}^t_{b,app}}{\sqrt{t}}, \Pi_{\mathcal{V}}(\widehat{\lambda}^t_{sc} - \lambda_*) + \frac{\widetilde{v}^t_{b,app}}{t} \right), \tag{8.261}$$

where $\widetilde{u}^t_{b,app}$, $\widetilde{v}^t_{b,app}$ are defined in (8.111), (8.121), w is the mapping from (8.256).

Then $\widetilde{\lambda}^t_{b,app}$ in (8.111) is defined as follows

$$\widetilde{\lambda}^t_{b,app} = \widehat{\lambda}^t_{sc} + \frac{\widetilde{u}^t_{b,app}}{\sqrt{t}} + \frac{\widetilde{v}^t_{b,app}}{t} + \widetilde{w}^t_{b,app}, \tag{8.262}$$

For $\widetilde{\lambda}^t_{b,app}$ from (8.262) it holds that

$$\widetilde{\lambda}^t_{b,app} \xrightarrow{c.p.} \lambda_* + w_* \text{ when } t \to +\infty,\ \text{a.s. } Y^t,\ t \in (0,+\infty), \tag{8.263}$$

where w_* is defined in (8.257).

Indeed, formula (8.263) follows from the fact that $\Pi_{\mathcal{U}\oplus\mathcal{V}}\widehat{\lambda}^t_{sc} \xrightarrow{c.p.} \Pi_{\mathcal{U}\oplus\mathcal{V}}\lambda_*$, the fact that $\widetilde{u}^t_{b,app}/\sqrt{t} = o_{cp}(1)$, $\widetilde{v}^t_{b,app}/t = o_{cp}(1)$ (see formula (8.126) and results of Lemma 8.11) and the continuity of mapping w.

From the local Lipschitz continuity of φ and (8.113), (8.114), (8.263) it follows that there exits some constant $L > 0$ such that with conditional probability tending to one a.s. Y^t, $t \in (0,+\infty)$ it holds that:

$$\varphi(\widetilde{\lambda}^t_{b,app}) - \varphi(\widetilde{\lambda}^t_{app}) \leq L\|\widetilde{\lambda}^t_{b,app} - \widetilde{\lambda}^t_{app}\|. \tag{8.264}$$

In particular, from (8.113), (8.114), (8.264) it follows that

$$\beta^t \cdot (\varphi(\widetilde{\lambda}^t_{b,app}) - \varphi(\widetilde{\lambda}^t_{app})) = o_{cp}(1). \tag{8.265}$$

To prove that the first term in (8.253) is also of order $o_{cp}(1)$ we will use extensively results from [Wet03].

The first term in (8.253) can be rewritten as a double inf-operation:

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}(\varphi(\lambda)-\varphi(\widetilde{\lambda}^t_{b,app}))=\inf_{\substack{(u,v)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})\\(u,v)\in\mathcal{U}\times\mathcal{V}}}[\varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app}-\lambda_*)+\frac{u}{\sqrt{t}},\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app}-\lambda_*)+\frac{v}{t})$$
$$-\varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app}-\lambda_*),\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app}-\lambda_*))], \tag{8.266}$$

where

$$\varphi_*(u,v)=\inf_{\substack{w:\lambda_*+u+v+w\succeq 0,\\ w\in\mathcal{W}}}\varphi(\lambda_*+u+v+w),$$
$$u\in\mathcal{U},\ v\in\mathcal{V}:(\lambda_*+u+v+\mathcal{W})\cap\mathbb{R}^p_+\neq\emptyset. \tag{8.267}$$

The expression in the square brackets in (8.266) is essentially the variation of the inf-projection for $\varphi_*(u,v)$ with parameters $u\in\mathcal{U}$ and $v\in\mathcal{V}$ in the vicinities of their respective zeros. Indeed, this follows from the facts that $\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app}-\lambda_*)$ and $\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app}-\lambda_*)$ are both of order $o_{cp}(1)$ and $u/\sqrt{t}$, v/t are also $o_{cp}(1)$ in view of the fact that $(u,v)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$.

Using Theorem 3.4 and examples in Section 4 (pp. 278-282) of [Wet03] we find that

$$\varphi_*(u,v)\text{ is locally Lipschitz continuous.} \tag{8.268}$$

Indeed, consider the optimization problem in (8.267), where $(u,v)\in\mathcal{U}\times\mathcal{V}$ is a parameter. Then, the problem can be rewritten as follows:

$$\inf_w\varphi_0((u,v);w),\ \varphi_0:(\mathcal{U}\times\mathcal{V})\times\mathcal{W}\to\overline{\mathbb{R}}, \tag{8.269}$$

$$\varphi_0((u,v);w)=\begin{cases}\varphi(\lambda_*+u+v+w),&\text{if }\lambda_*+u+v+w\succeq 0,\\ +\infty,&\text{otherwise,}\end{cases} \tag{8.270}$$

where $\overline{\mathbb{R}}$ denotes the extended real line. From the fact that $\varphi(\cdot)$ is locally Lipschitz continuous it is easy to see that φ_0 is locally Lipschitz continuous on $D=\{(u,v,w)\in\mathcal{U}\times\mathcal{V}\times\mathcal{W}:\lambda_*+u+v+w\succeq 0\}$, where the latter is a polyhedral subset of $\mathcal{U}\times\mathcal{V}\times\mathcal{W}$.

Consider the feasibility mapping

$$S:\mathcal{U}\times\mathcal{V}\rightrightarrows\mathcal{W}\text{ with }S(u,v)=\{w\in\mathcal{W}:\lambda_*+u+v+w\succeq 0\}, \tag{8.271}$$

where $\rightrightarrows$ denotes the property to be a set-valued mapping. From (8.271) one can see that $\operatorname{gph}S=D$ (gph denotes the graph of a mapping). Therefore, $\operatorname{gph}S$ is polyhedral and Proposition 4.1 from [Wet03] applies to our case (see also Example 9.35 in [RW09]), so mapping S in (8.271) Lipschitz continuous on $\operatorname{dom}S$ (as set-valued mapping). At the same time, the result of Lemma 8.1 implies that feasibility mapping S is locally bounded which yeilds level boundeness in w locally uniformly in (u,v) of $\varphi_0(\cdot,\cdot)$. The above properties are exactly the same is in Section 4 of [Wet03], so Theorem 3.4 therein applies to the case of φ_0 from (8.269) and $\varphi_*(u,v)=\inf_w\varphi((u,v);w)$ is locally Lipschitz continous. This proves the claim in (8.268) is proved.

From (8.268) it follows that there exists a constant $L>0$ such that with conditional probability tending to one a.s. Y^t, $t\in(0,+\infty)$ the following holds

$$|\varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app}-\lambda_*)+\frac{u}{\sqrt{t}},\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app}-\lambda_*)+\frac{v}{t})-\varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app}-\lambda_*),\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app}-\lambda_*))|$$
$$\leq L\left(\frac{\|u\|}{\sqrt{t}}+\frac{\|v\|}{t}\right)\leq L\left(\frac{\delta}{\sqrt{t}}+c\frac{\delta}{t}\right)\text{ for any }(u,v)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}), \tag{8.272}$$

where c is a positive constant depending only on dimension p.

Using formulas (8.266), (8.272) and the assumption that $\beta^t = o(\sqrt{t})$ we obtain

$$\beta^t \cdot \inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{b,app})) = o_{cp}(1). \tag{8.273}$$

Formula (8.188) directly follows from (8.265), (8.273).

Lemma is proved. □

8.19 Proof of Theorem 6.7

Proof. Proof of the theorem is essentially the same as the proof of Theorem 6.3. Indeed, the assumptions for Theorem 6.3 contain restrictions only on Λ^* which coincide with ones in the generalized non-expansiveness condition which is assumed to hold. All steps in the proof of Theorem 6.3 remain the same with $A_{\mathcal{M}}$ being replaced with A.

Theorem is proved. □

Acknowledgments

We are grateful to Zacharie Naulet from Université d'Orsay for many valuable comments on statistical side of the paper. We are also grateful to our colleagues from *Service Hospitalier Frédéric Joliot* (SHFJ) – Marina Filipović, Claude Comtat and Simon Stute for many practical insights on the topic of PET-MRI reconstructions.

References

[AG91] R. G. Aykroyd and P. J. Green. Global and local priors, and the location of lesions using gamma-camera imagery. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):323–342, 1991.

[AW93] Hedy Attouch and Roger J.-B. Wets. Quantitative stability of variational systems. ii. a framework for nonlinear conditioning. *SIAM Journal on Optimization*, 3(2):359–381, 1993.

[BCD+07] É. Barat, C. Comtat, T. Dautremer, T. Montagu, and R. Trebossen. A non-parametric bayesian approach for pet reconstruction. In *IEEE. Nucl. Sci. Symp. Conf. Rec.*, pages 4155–4162, 2007.

[Ber97] D. P Bertsekas. Nonlinear programming. *Journal of Operational Research Society*, 48(3):334, 1997.

[BF11] D.M. Blei and P.I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(8), 2011.

[BG14] N. A. Bochkina and P. J. Green. The bernstein–von mises theorem and non-regular models. *The Annals of Statistics*, 42(5):1850–1878, 2014.

[BJT+96] J.E. Bowsher, V.E. Johnson, T.G. Turkington, R.J. Jaszczak, C.E. Floyd, and R.E. Coleman. Bayesian reconstruction and use of anatomical a priori information for emission tomography. *IEEE Transactions on Medical Imaging*, 15(5):673–686, 1996.

[BR16] N. Baldin and M Reiß. Unbiased estimation of the volume of a convex body. *Stochastic Processes and their Applications*, 126(12):3716–3732, 2016.

[BWP97] Harrison H Barrett, Timothy White, and Lucas C Parra. List-mode likelihood. *JOSA A*, 14(11):2914–2923, 1997.

[BWT94] H. H. Barrett, D. W. Wilson, and B. M. Tsui. Noise properties of the em-algorithm. i. theory. *Phys. Med. Biol.*, 39(5):833, 1994.

[BYH+04] J.E. Bowsher, Hong Yuan, L.W. Hedlund, T.G. Turkington, G. Akabani, A. Badea, W.C. Kurylo, C.T. Wheeler, G.P. Cofer, M.W. Dewhirst, and G.A. Johnson. Utilizing mri information to estimate f18-fdg distributions in rat flank tumors. In *IEEE Symposium Conference Record Nuclear Science*, volume 4. IEEE, 2004.

[CFD13] Se Young Chun, Jeffrey A Fessler, and Yuni K Dewaraja. Post-reconstruction non-local means filtering methods using ct side information for quantitative spect. *Physics in Medicine & Biology*, 58(17):6225, 2013.

[CKF+01] Claude Comtat, Paul E Kinahan, Jeffrey A Fessler, Thomas Beyer, David W Townsend, Michel Defrise, and Christian Michel. Clinically feasible reconstruction of 3d whole-body pet/ct data using blurred anatomical labels. *Physics in Medicine & Biology*, 47(1):1, 2001.

[Dah01] M. Dahlbom. Estimation of image noise in pet using the bootstrap method. In *IEEE Nuclear Science Symposium Conference Record*, volume 4. IEEE, 2001.

[DJD18] L. L. Duan, J. E. Johndrow, and D. B. Dunson. Scaling up data augmentation mcmc via calibration. *Journal of Machine Learning Research*, 19(1):2575–2608, 2018.

[DP93] A.R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Transactions on Medical Imaging*, 12(2):328–333, 1993.

[DVJ05] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer Science & Business Media, 2005.

[DVJ07] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[EF99] H. Erdogan and J.A. Fessler. Monotonic algorithms for transmission tomography. *IEEE Transactions on Medical Imaging*, 18(9):801–814, 1999.

[FBC+11] Mame Diarra Fall, Éric Barat, Claude Comtat, Thomas Dautremer, Thierry Montagu, and Ali Mohammad-Djafari. A discrete-continuous Bayesian model for emission tomography. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 1373–1376, Brussel,, Belgium, 2011.

[FBD+18] Marina Filipović, Éric Barat, Thomas Dautremer, Claude Comtat, and Simon Stute. Pet reconstruction of the posterior image probability, including multimodal images. *IEEE transactions on medical imaging*, 38(7):1643–1654, 2018.

[FCR92] Jeffrey A Fessler, Neal H Clinthorne, and W Leslie Rogers. Regularized emission image reconstruction using imperfect side information. *IEEE Transactions on Nuclear Science*, 39(5):1464–1471, 1992.

[FDC+21] Marina Filipovic, Thomas Dautremer, Claude Comtat, Simon Stute, and Eric Barat. Reconstruction, analysis and interpretation of posterior probability distributions of pet images, using the posterior bootstrap. *Physics in Medicine & Biology*, 2021.

[Fes96] J. A. Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Transactions on Image Processing*, 5(3):493–506, 1996.

[FH95] J.A. Fessler and A.O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Transactions on Image Processing*, 4(10):1417–1429, 1995.

[FL07] A. R. Ferreira and K. H. Lee. Single photon emission computed tomography example. In *Multiscale Modeling*. Springer Series in Statistics, 2007.

[FLH19] Edwin Fong, Simon Lyddon, and Chris Holmes. Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1952–1962. PMLR, 09–15 Jun 2019.

[Gey94] Charles J. Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, 22(4):1993–2010, 1994.

[GM87] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstructions. *Bull. Int. Stat. Inst.*, LII(4):5–21, 1987.

[Gon19] F. Goncharov. *Weighted Radon transforms and their applications*. PhD thesis, Université Paris Saclay (COmUE), 2019.

[Gre90] P. J. Green. Bayesian reconstructions from emission tomography data using a modified em algorithm. *IEEE Trans. Med. Imag.*, 9:84–93, 1990.

[GUSB11] Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei. Spatial distance dependent chinese restaurant processes for image segmentation. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1476–1484. Curran Associates, Inc., 2011.

[Gut13] Allan Gut. *Probability: a graduate course*. New York, NY: Springer, 2013.

[HBJ+97] D.M. Higdon, J.E. Bowsher, V.E. Johnson, T.G. Turkington, D.R. Gilland, and R.J. Jaszczak. Fully bayesian estimation of gibbs hyperparameters for emission computed tomography data. *IEEE Transactions on Medical Imaging*, 16:516, 1997.

[HLY99] G. Han, Z. Liang, and J. You. A fast ray-tracing technique for tct and ect studies. In *1999 IEEE Nuclear Science Symposium. Conference Record. 1999 Nuclear Science Symposium and Medical Imaging Conference (Cat. No.99CH37019)*, volume 3, pages 1515–1518 vol.3, 1999.

[HPFT99] Alfred O Hero, Robinson Piramuthu, Jeffrey A Fessler, and Steven R. Titus. Minimax emission computed tomography using high-resolution anatomical side information and b-spline models. *IEEE Transactions on Information Theory*, 45(3):920–938, 1999.

[HW89] D. R. Haynor and S. D. Woods. Resampling estimates of precision in emission tomography. *IEEE Transactions on Medical Imaging*, 8(4), 1989.

[HW16] T. Hohage and F. Werner. Inverse problems with poisson data: statistical regularization theory, applications and algorithms. *Inverse Problems*, 32(9):093001, 2016.

[Jam03] L. F. James. Bayesian calculus for gamma processes with applications to semiparametric intensity models. *Sankhyā: The Indian Journal of Statistics*, pages 179–206, 2003.

[JS13] J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes.* Springer Science & Business Media, 2013.

[JWN+08] M. Judenhofer, H. Wehrl, D. Newport, C. Catana, S. Siegel, Markus Becker, A. Thielscher, M. Kneilling, M. Lichy, M. Eichner, K. Klingel, G. Reischl, S. Widmaier, M. Röcken, R. Nutt, H. Machulla, K. Uluda, S. Cherry, C. Claussen, and B. Pichler. Simultaneous pet-mri: a new approach for functional and morphological imaging. *Nature medicine*, 14(4):459–465, 2008.

[KS06] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[KVdV12] B. J. K. Kleijn and A. W. Van der Vaart. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[LAB10] C. Lartizien, J.-B. Aubin, and I. Buvat. Comparison of bootstrap resampling methods for 3-d pet imaging. *IEEE Transactions on Medical Imaging*, 29(7):1442–1454, 2010.

[LDH95] C. S. Levin, M. Dahlbom, and E. J. Hoffman. A monte carlo correction for the effect of compton scattering in 3-d pet brain imaging. *IEEE Transactions on Nuclear Science*, 42(4):1181–1185, 1995.

[Lew10] T. K. Lewellen. The challenge of detector designs for pet. *American Journal of Roentgenology*, 195(2):301–309, 2010.

[LHW19] SP Lyddon, CC Holmes, and SG Walker. General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

[LHY00] Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[Li11] Y. Li. Noise propagation for iterative penalized-likelihood image reconstruction based of fisher information. *Phys. Med. Biol.*, 56(4):1083, 2011.

[Liu94] J. S. Liu. The fraction of missing information and convergence rate for data augmentation. *Computing Science and Statistics*, pages 490–490, 1994.

[Lo82] A. Y. Lo. Bayesian nonparametric statistical inference for poisson point processes. *Zeitschrift fur Wahrscheinlichkeitsteorie und verwandte Gebiete*, 59(1):55–66, 1982.

[LVHR13] A. Luna, J. C. Vilanova, L. C. Hygino da Cruz Jr, and S. E. Rossi. *Functional imaging in oncology: biophysical basis and technical approaches - Vol. 1.* Springer Science & Business Media, 2013.

[LWH18] S. Lyddon, S. Walker, and C. Holmes. Nonparametric learning from bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems*, 2018.

[LWK94] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1), 1994.

[MMB18] L. G. Marcu, L. Moghaddasi, and E. Bezak. 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, 2018.

[Nat01] F. Natterer. *The mathematics of computerized tomography.* Society for Industrial and Applied Mathematics, 2001.

[NN20] Tun Lee Ng and Michael A Newton. Random weighting in lasso regression. *arXiv preprint arXiv:2002.02629*, 2020.

[Nov19] RG Novikov. 5. non-abelian radon transform and its applications. pages 115–128, 2019.

[NPX21] Michael A. Newton, Nicholas G. Polson, and Jianeng Xu. Weighted bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*, 49(2):421–437, 2021.

[NR94] M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.

[NR20] Lizhen Nie and Veronika Ročková. Bayesian bootstrap spike-and-slab lasso. *arXiv preprint arXiv:2011.14279*, 2020.

[Pom21] Emilia Pompe. Introducing prior information in weighted likelihood bootstrap with applications to model misspecification. *arXiv preprint arXiv:2103.14445*, 2021.

[Qui83] E. T. Quinto. The invertibility of rotation invariant radon transforms. *Journal of Mathematical Analysis and Applications*, 91(2):510–522, 1983.

[RTZ09] Arman Rahmim, Jing Tang, and Habib Zaidi. Four-dimensional (4d) image reconstruction strategies in dynamic pet: Beyond conventional independent frame reconstruction. *Medical Physics*, 36(8):3654–3670, 2009.

[RW09] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[SC13] S. Stute and C. Comtat. Practical considerations for image-based psf and blobs reconstruction in pet. *Physics in Medicine & Biology*, 58(11):3849, 2013.

[SC15] A. Sitek and M. A. Celler. Limitations of poisson statistics in describing radioactive decay. *Physica Medica*, 31(8):1105–1107, 2015.

[Sid85] R. Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical physics*, 12 2:252–5, 1985.

[Sit12] A. Sitek. Data analysis in emission tomography using emission count posteriors. *Physics in Medicine & Biology*, 52(21):6779, 2012.

[SV82] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE transactions on medical imaging*, 1(2):113–122, 1982.

[VAB+11] Kathleen Vunckx, Ameya Atre, Kristof Baete, Anthonin Reilhac, Christophe M. Deroose, Koen Van Laere, and Johan Nuyts. Evaluation of three mri-based anatomical priors for quantitative pet brain imaging. *IEEE transactions on medical imaging*, 31(3):599–612, 2011.

[VDM01] D. A. Van Dyk and X.-L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[VdV00] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Web05] W. A. Weber. Use of pet for monitoring cancer therapy and for predicting outcome. *Journal of Nuclear Medicine*, 46(6):983–995, 2005.

[Wei97] I. S. Weir. Fully bayesian reconstructions from single-photon emission computed tomography data. *Journal of the American Statistical Association*, 92(437):49–60, 1997.

[Wet03] Roger J.-B. Wets. Lipschitz continuity of inf-projections. *Computational Optimization and Applications*, 25(1-3):269–282, 2003.

[Whi82] H. White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1982.

[WQ15] Guobao Wang and Jinyi Qi. Edge-preserving pet image reconstruction using trust optimization transfer. *IEEE Transactions on Medical Imaging*, 34(4):930–939, 2015.

[WW69] David W Walkup and Roger J-B Wets. A lipschitzian characterization of convex polyhedra. *Proceedings of the American Mathematical Society*, pages 167–173, 1969.

[XYCD16] Da Xu, Richard Yi, Francois Caron, and Arnaud Doucet. Bayesian nonparametric image segmentation using a generalized swendsen-wang algorithm. *arXiv preprint arXiv:1602.03048*, 2016.

[YCHT04] Jeffrey T Yap, Jonathan P J Carney, Nathan C Hall, and David W Townsend. Image-guided cancer therapy using pet/ct. *The Cancer Journal*, 10(4):221–233, 2004.

A Construction of the common probability space.

Let $(\Omega', \mathcal{F}', P')$ be the probability space on which the stationary spatio-temporel Poisson point process Z^t is defined (Z^t has values in $Z \times (0, +\infty)$; recall that Z is the space of LORs). Sinogram data Y^t is obtained from binning Z^t to detector elements (see Section 3.2), therefore process Y^t is a well-defined random variable on $(\Omega', \mathcal{F}', P')$. Measure-theoretic construction of Z^t and $(\Omega', \mathcal{F}', P')$ can be found, for example, in [DVJ07], Section 9.2, Example 9.2(b).

Algorithms 4, 5 rely on perturbed intensities $\Lambda^t_{\mathcal{M}}$ and $\widetilde{\Lambda}^t_b$ for which we show that they can be expressed as functions of random weighting of the list-mode data. Let

$$G^t = \{\delta_{(k,i)} : (k,i) - k^{\text{th}} \text{ photon was detected at detector } i\},$$

where G^t corresponds to observed data Y^t. Indeed, from steps 1, 2 in Algorithm 4 we can see that $\Lambda^t_{\mathcal{M}}$ is a function of Λ^t for which the following representation holds:

$$\Lambda^t_i = t^{-1} \sum_{k=1}^{N^t} \delta_{(k,i)} \widetilde{w}_k, \; i \in \{1, \dots, d\}, \tag{A.1}$$

$$\{\widetilde{w}_k\}_{k=1}^{N^t} \overset{iid}{\sim} \Gamma(1,1), \tag{A.2}$$

where N^t is the total number of photons.

For $\widetilde{\Lambda}^t_b$ in step 3 of Algorithm 5 we have the following representation:

$$\widetilde{\Lambda}^t_{b,i} = (\theta^t + t)^{-1} \left(\sum_{k=1}^{N^t} \delta_{(k,i)} w_k + w_p \theta^t \Lambda^t_{\mathcal{M},i} \right), \; i \in \{1, \dots, d\}, \tag{A.3}$$

$$\{w_k\}_{k=1}^{N^t}, w_p \overset{iid}{\sim} \Gamma(1,1). \tag{A.4}$$

From formulas (A.1)-(A.4) one can see that perturbations $\Lambda^t_{\mathcal{M}}$ and $\widetilde{\Lambda}^t_b$ depend on data Y^t and on family of random mutually independent weights $(\{(w_k, \tilde{w}_k)\}_{k=1}^{N^t}, w_p)$ and also independent of Y^t. Therefore, the common probability space can be defined as follows:

$$(\Omega', \mathcal{F}', P') = (\Omega' \times \Omega_w \times \Omega_{\widetilde{w}} \times \Omega_{w_p}, \mathcal{F}' \times \mathcal{F}_w \times \mathcal{F}_{\widetilde{w}} \times \mathcal{F}_{w_p}, P' \times P_w \times P_{\widetilde{w}} \times P_{w_p}), \tag{A.5}$$

where $(\Omega_w, \mathcal{F}_w, P_w)$, $(\Omega_{\widetilde{w}}, \mathcal{F}_{\widetilde{w}}, P_{\widetilde{w}})$, $(\Omega_{w_p}, \mathcal{F}_{w_p}, P_{w_p})$ are the probability spaces for infinite sequences of i.i.d r.v.s $\{w_k\}_{k=1}^{\infty}$, $\{\widetilde{w}_k\}_{k=1}^{\infty}$, $w_k \sim \Gamma(1,1)$, $\widetilde{w}_k \sim \Gamma(1,1)$ and for $w_p \sim \Gamma(1,1)$, respectively. This construction is not new and it originates to [NR94]; similar ones have been recently used in [NN20], [Pom21].

B Limit theorems for stationary Poisson processes.

Let

$$Y^t \sim \text{Po}(\Lambda \cdot t), \; \Lambda > 0, \; t \in [0, +\infty). \tag{B.1}$$

The following result is a composition of theorems 9.3, 4.1 and 7.5 (pp. 306, 350, 417, respectively) from [Gut13].

Theorem B.1. *Let Y^t be the Poisson process defined in (B.1). Then,*

i)

$$\frac{Y^t}{t} \xrightarrow{a.s.} \Lambda \text{ as } t \to +\infty. \tag{B.2}$$

ii)

$$\frac{Y^t - \Lambda t}{\sqrt{\Lambda t}} \xrightarrow{d} \mathcal{N}(0,1) \ \textit{as} \ t \to +\infty. \tag{B.3}$$

iii)

$$\liminf_{t\to+\infty}(\limsup_{t\to+\infty}) \frac{Y^t - \Lambda t}{\sqrt{\Lambda t \log\log t}} = \sqrt{2}\,(-\sqrt{2}) \ \textit{a.s.}, \tag{B.4}$$

where $\xrightarrow{a.s.}$, $\xrightarrow{d}$ *denote the convergence almost surely and in distribution, respectively, a.s. denotes that a statement holds for almost any trajectory* Y^t, $t \in (0, +\infty)$.

C GEM-type algorithm derivation

We mainly follow [WQ15] for the derivation of the minimization algorithm based on optimization transfer. Our aim is to build a majoring surrogate of $L_p(\lambda|\,\widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$. Using the fact that $L_p(\lambda|\,\widetilde{\Lambda}_b^t, A, 1, \beta^t/t) = L(\lambda|\,\widetilde{\Lambda}_b^t, A, 1) + \frac{\beta^t}{t}\varphi(\lambda)$, we proceed by finding a surrogate for each of both terms in the right hand-side.

C.1 Majoring surrogate of $L(\lambda|\,\widetilde{\Lambda}_b^t, A, 1)$

In [DP93] authors propose a purely algebraic derivation of the surrogate outside the context of latent variables and evidence lower bound (ELBO) computation.

Let $f_i(x) \triangleq x - \widetilde{\Lambda}_{b,i}^t \log(x)$, $\lambda_j^{(r)} \succeq 0$, $j = 1, \ldots, p$, be the r^{th} iterate of the optimization algorithm minimizing $L(\lambda|\widetilde{\Lambda}_b^t, A, 1)$, and denote also $\Lambda_i^{(r)} = a_i^T \lambda^{(r)}$.

Consider the formula

$$\begin{aligned} L(\lambda|\,\widetilde{\Lambda}_b^t, A, 1) &= \sum_{i=1}^{d} f_i(\Lambda_i) \\ &= \sum_{i=1}^{d} f_i\left(\sum_{j=1}^{p} a_{ij}\lambda_j\right) \\ &= \sum_{i=1}^{d} f_i\left(\sum_{j=1}^{p} \left[\frac{a_{ij}\lambda_j^{(r)}}{\Lambda_i^{(r)}}\right]\left[\frac{\lambda_j}{\lambda_j^{(r)}}\Lambda_i^{(r)}\right]\right) \end{aligned}$$

Since f_i is convex for $\widetilde{\Lambda}_{b,i}^t \geq 0$ and using the fact that $\sum_{j=1}^{p} \frac{a_{ij}\lambda_j^{(r)}}{\Lambda_i^{(r)}} = 1$ together with the Jensen's inequality we obtain

$$L(\lambda|\,\widetilde{\Lambda}_b^t, A, 1) \leq Q(\lambda, \lambda^{(r)})$$

where

$$Q(\lambda, \lambda^{(r)}) = \sum_{i=1}^{d}\sum_{j=1}^{p} \left[\frac{a_{ij}\lambda_j^{(r)}}{\Lambda_i^{(r)}}\right] f_i\left(\frac{\lambda_j}{\lambda_j^{(r)}}\Lambda_i^{(r)}\right)$$

Note also that $Q(\lambda^{(r)}, \lambda^{(r)}) = L(\lambda^{(r)}|\,\widetilde{\Lambda}_b^t, A, 1)$. Using the definition of f_i we find that

$$Q(\lambda, \lambda^{(r)}) = \sum_{i=1}^{d}\sum_{j=1}^{p} \left[a_{ij}\lambda_j - \frac{a_{ij}\lambda_j^{(r)}}{\Lambda_i^{(r)}}\widetilde{\Lambda}_{b,i}^t \log\left(\frac{\lambda_j}{\lambda_j^{(r)}}\Lambda_i^{(r)}\right)\right]$$

$$= \sum_{j=1}^{p} A_j \left[\lambda_j - \left(\frac{\lambda_j^{(r)}}{A_j} \sum_{i=1}^{d} \frac{a_{ij} \widetilde{\Lambda}_{b,i}^{t}}{\Lambda_i^{(r)}} \right) \log \lambda_j \right] + const.$$

where R denotes terms independent of λ.

Function $Q(\lambda, \lambda^{(r)})$ can be rewritten as follows:

$$Q(\lambda, \lambda^{(r)}) \triangleq \sum_{j=1}^{p} A_j \left(\lambda_j - \lambda_{j,L}^{(r+1)} \log \lambda_j \right) \tag{C.1}$$

with

$$\lambda_j^{(r+1),L} \triangleq \frac{\lambda_j^{(r)}}{A_j} \sum_{i=1}^{d} \frac{a_{ij} \widetilde{\Lambda}_{b,i}^{t}}{\Lambda_i^{(r)}} \tag{C.2}$$

C.2 Majoring surrogate for $\varphi(\lambda)$

Let

$$\varphi(\lambda) = \sum_{j=1}^{p} \sum_{k \in \mathcal{N}_j} w_{jk} \, \psi(\lambda_j - \lambda_k)$$

with $w_{jk} > 0$, $w_{kj} = w_{jk}$ are the weigths and $\mathcal{N}_j$ is the neighborhood of pixel j.

From [EF99], any potential function ψ satisfying the conditions

i. ψ is symmetric.

ii. ψ is continuous and differentiable everywhere.

iii. ψ is convex.

iv. $\omega_\psi(u) \triangleq \frac{1}{u} \frac{\mathrm{d}\psi(u)}{\mathrm{d}u}$ is non-increasing for $u \geqslant 0$.

v. $\lim_{u \to 0} \omega_\psi(u)$ is finite and positive.

can be majorized by a parabolic curve.

With these requirements satisfied, $\varphi(\lambda)$ is majorized by a separable quadratic penalty given below (see [WQ15] and references therein):

$$\varphi(\lambda) \leq Q_\varphi(\lambda; \lambda^{(r)})$$

where

$$Q_\varphi(\lambda; \lambda^{(r)}) = \frac{1}{2} \sum_{j=1}^{p} p_{j,\varphi}^{(r+1)} (\lambda_j - \lambda_{j,\varphi}^{(r+1)})^2, \tag{C.3}$$

$$p_{j,\varphi}^{(r+1)} = 4 \sum_{k \in \mathcal{N}_j} w_{jk} \, \omega_\psi(\lambda_j^{(r)} - \lambda_k^{(r)}), \tag{C.4}$$

$$\lambda_{j,\varphi}^{(r+1)} = \frac{2}{p_{j,\varphi}^{(r+1)}} \sum_{k \in \mathcal{N}_j} w_{jk} \, \omega_\psi(\lambda_j^{(r)} - \lambda_k^{(r)})(\lambda_j^{(r)} + \lambda_k^{(r)}). \tag{C.5}$$

C.3 Global surrogate minimization

At iteration $(r+1)$, solving the Karush-Kuhn-Tucker condition for minimizing the combined surrogate, we get

$$\lambda^{(r+1)} = \underset{\lambda \succeq 0}{\arg\min} \; Q_L(\lambda, \lambda^{(r)}) + \frac{\beta^t}{t} Q_\varphi(\lambda, \lambda^{(r)})$$

which gives a unique analytical solution

$$\lambda_j^{(r+1)} = \frac{2\lambda_{j,L}^{(r+1)}}{\sqrt{(b_j^{(r+1)})^2 + 4\beta_j^{(r+1)}\lambda_{j,L}^{(r+1)}} + b_j^{(r+1)}} \tag{C.6}$$

with $\beta_j^{(r+1)} = \frac{\beta^t}{t\, A_j} p_{j,\varphi}^{(r+1)}$ and $b_j^{(r+1)} = 1 - \beta_j^{(n+1)}\lambda_{j,\varphi}^{(r+1)}$.

The GEM-type algorithm is summarized in Algorithm 7.

Algorithm 7: $\arg\min_{\lambda \succeq 0} L_p(\lambda|\, \widetilde{\Lambda}_b^t, A, 1, \frac{\beta^t}{t})$ by optimization transfer

Data: Stochastic intensities $\widetilde{\Lambda}_b^t$;
Input: Initial image $\lambda^{(0)}$, number max. of iterations R, projector A, regularization parameter β^t, penalty $\varphi(\lambda)$

```
for r = 0 to R − 1 do
    for j = 1 to p do
        compute λ_{j,L}^{(r+1)} using formula (C.2);
        compute λ_{j,φ}^{(r+1)} using formula (C.5);
        compute λ_j^{(r+1)} using formula (C.6);
    end
end
```

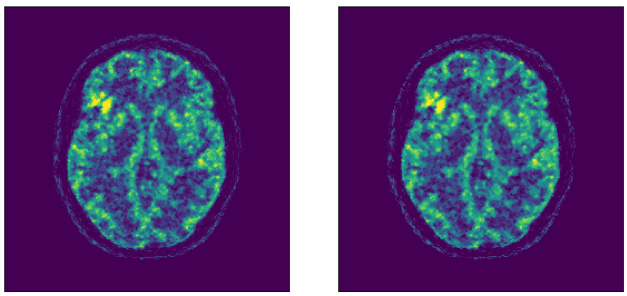
Output: Minimizing intensity map $\lambda^{(R)}$

Remark C.1. By setting $\frac{\beta^t}{t} \to 0$ in (C.6), we immediately check that $\lambda^{(r+1)} \to \lambda_L^{(r+1)}$.

Remark C.2. Parameter $\lambda_{\mathcal{M}}^t$ in Algorithm 4 is easily obtained by iterating formula (C.2) with projector $A_{\mathcal{M}}$ and random intensities Λ^t

$$\lambda_{\mathcal{M},s}^{(r+1)} = \frac{\lambda_{\mathcal{M},s}^{(r)}}{A_s^{\mathcal{M}}} \sum_{i=1}^{d} \frac{a_{is}^{\mathcal{M}} \Lambda_i^t}{\Lambda_{\mathcal{M},i}^{(r)}} \tag{C.7}$$

D Visual comparison between the NPL mean without MRI and the MAP reconstructions



NPL without MRI MAP

Figure 12: Comparison of NPL mean without MRI (left) and MAP (right) reconstructions