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[1], **Éric Barat**[1] **and Thomas Dautremer**[1]

[1] *Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France.*
fedor.goncharov@cea.fr, eric.barat@cea.fr, thomas.dautremer@cea.fr

Abstract: We continue studies of the uncertainty quantification problem in emission tomographies such as PET or SPECT 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 identifiability problem for misspecified generalized Poisson models. We also contrast our approach with Bayesian Markov Chain Monte Carlo sampling based on one data augmentation scheme which is very popular in the context of Expectation-Maximization 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 assessment for the uncertainty.

1. Introduction

Emission tomographies (further referred as ET) such as Positron Emission Tomography (PET) or Single Photon Emission Computed Tomography (SPECT) are functional imaging modalities of nuclear medicine which are used to image activity processes and, in particular, metabolism in soft tissues via the uptake of certain injected biomarkers. The level of metabolism provides critical information for diagnostics and treatment of cancers; see e.g., Weber (2005), Marcu et al. (2018) 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. *Regularize the inverse problem using the multimodal data (e.g., images from CT or MRI).*

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 (Fessler (1996), Barrett et al. (1994), Li (2011)), Bayesian (Higdon et al. (1997), Weir (1997), Ferreira and Lee (2007), Sitek (2012), Bochkina and Green (2014), Filipović et al. (2018)) and bootstrap (Haynor and Woods (1989), Dahlbom (2001), Lartizien et al. (2010), Filipović et al. (2021)). 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 of both modalities consists in extracting boundaries of anatomical features on side images and embedding them into regularization schemes via special penalties and/or non-invariant filters; see e.g., Fessler et al. (1992), Chun et al. (2013), Hero et al. (1999), Comtat et al. (2001), Vunckx et al. (2011). 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., Hohage and Werner (2016)) and very low signal-to-noise ratio in the raw measured data. All this together results in loss of resolution in reconstructed images and consequently in oversmoothing, e.g., when applying spatially invariant filters for post-smoothing. In our work as multimodal data we use series of presegmented anatomical MRI images. Problem 2 for additional MRI data is now of particular interest due to appearance of commercially available models of PET-MRI scanners Luna et al. (2013), Judenhofer et al. (2008) which allow simultaneous registrations of both signals, thus significantly reducing motion effects. Moreover, in the experiment on tumor imaging in Bowsher et al. (2004) correlations between PET and MRI signals were observed, therefore, potentially MRI data can

be used to regularize accurately the inverse problem. In Section 2 we explain in detail how we use MRI data and compare our approach with previous works.

For Problem 1 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 from the generalized Poisson model with unknown intensity parameter λ_* and known design A, i.e., $P^t = P^t_{A,\lambda} = \mathrm{Po}(tA\lambda_*)$), so any reconstruction $\widehat{\lambda}^t$ would be also a function of observed data, that is $\widehat{\lambda}^t = \widehat{\lambda}^t(Y^t)$ and uncertainty propagates directly from Y^t. This is known as frequentist approach, and for ET it often leads to estimation of confidence intervals for the maximum likelihood estimator (MLEM) or for penalized maximum log-likelihood estimator (pMLEM or MAP) (both are M-estimators Van der Vaart (2000)); see e.g., Fessler (1996). 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 $P^t_{A,\lambda}$ with respect to some chosen distance between probability distributions (e.g., for Kullback-Liebler divergence). Under additional assumptions on P^t even in misspecified case it is still possible to establish asymptotic distribution of $\widehat{\lambda}^t$ (e.g., via asymptotic normality), from which, for example, the asymptotic confidence intervals can be retrieved. However, use of asymptotic results for ET practice seems doubtful since very little data are available in a single scan.

Bayesian approach is also used for uncertainty quantification in ET. In this case the initial uncertainty on the parameter of interest (e.g., anatomical information from side images, assumptions on support and smoothness) is encoded in some prior measure $\pi_{\mathcal{M}}(\lambda)$ which is updated using model family $P^t_{A,\lambda}$ and data Y^t to define posterior distribution via the well-known Bayes' formula; see e.g., Bochkina and Green (2014). Sampling from such posteriors is done via Markov Chain Monte Carlo (MCMC) techniques Weir (1997), Higdon et al. (1997), Ferreira and Lee (2007), Filipović et al. (2018). Common bottlenecks here are: complicated design of the algorithm and its implementation, high numerical load per iteration, lack of scalability and most importantly – poor mixing in constructed chains; see e.g., Van Dyk and Meng (2001), Duan et al. (2018). Additional issue is the misspecification of the model which cannot be included in the classical Bayesian framework and for robust inference it leads to the recently proposed general Bayesian updating and bootstrap-type sampling; see Pompe (2021), Section 1.

As noted above bootstrap is another attractive technique to assess the uncertainty which can be also seen as some probabilistic sensitivity analysis or as approximate/exact sampling via (nonparametric) Bayesian posteriors; see e.g., Newton and Raftery (1994), Lyddon et al. (2018), Fong et al. (2019). Nontrivial questions for ET are the following ones: (1) how to define a bootstrap procedure for Poisson-type raw data in ET and also include side information (multimodal images) (2) provide theoretical guarantees on the coverage by asymptotic credible intervals. A common approach to answer question (1) is to use resampling in list-mode data; see e.g., Haynor and Woods (1989), Dahlbom (2001). Such approach targets to resample photon counts and then propagate the uncertainty by using some reconstruction algorithm (e.g., FBP (Filtered backprojection), MLEM or MAP (maximum a posteriori)). In this sense our approach is similar to bootstrap as it will be explained further. Question (2) is often resolved by demonstrating asymptotic equivalence between bootstrap, Bayesian and frequentist approaches via Bernstein von-Mises type theorems; see e.g., Van der Vaart (2000), Lyddon et al. (2018), Ng and Newton (2020) or equivalence of Edgeworth's expansions for higher orders; see Pompe (2021).

In view of the above discussion, we note that for practice it seems that it is not of great importance which kind of uncertainty model is used – frequentist, Bayesian or bootstrap. Most important is to make usable the resulting framework and algorithms by practitioners, hence, it should be simple, tractable and numerically feasible.

Being inspired with nonparametric posterior learning (further referred as NPL) originating from Lyddon et al. (2018), Fong et al. (2019), we propose sampling algorithms for ET with and without MRI data at hand. Therefore, our main contribution is that we extend the NPL originally proposed for regular statistical models and i.i.d data to the non-regular generalized Poisson model of ET (see Bochkina and Green (2014)), where the raw data are not i.i.d but a realization from a point process. The initial motivation for this work was the problem of poor mixing for the Gibbs-type sampler in Filipović et al. (2018) which was designed for posterior sampling in the PET-MRI context. Below we give a detailed analysis of this phenomenon and give few empirical advises on design of MCMC-samplers for ill-posed inverse problems such as PET or SPECT. Our new algorithms solve the above

problem since sampled images are automatically i.i.d, moreover, the scheme is trivially parallelizable, scalable and very easy to implement because it relies on the well-known EM-type reconstruction methods from Shepp and Vardi (1982), Fessler and Hero (1995). Because of the aforementioned non-regularity of the model we conduct a separate theoretical study of our algorithms for when large dataset is available (for ET this is equivalent to $t \to +\infty$) and establish consistency and tightness of the posterior for almost any trajectory Y^t, $t \in (0, +\infty)$. Establishing further the asymptotic normality requires existence of a strongly consistent estimator which has specific contraction rates in the span of the design and for components activated by positivity constraints. Existence of such estimator is left conjectured, however, we propose one candidate and explain the intuition behind which makes the requirement quite natural.

Though our main theoretical results rely on the assumption of well-specified model, at the end we study the identification problem for the KL-criterion in the misspecified case with wrong design. If a certain geometrical condition on design matrix and observed asymptotic sinogram are satisfied, then the identification problem has positive answer and negative otherwise. In particular, the latter result gives a clue to extend our theoretical results to fully misspecified scenario for the model of ET when design matrix is incorrect. The latter case is meaningful in practice since the design in ET is always computed very approximately and it does not reflect very complicated photon-matter interactions inside the human body.

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 give a very informative example for the problem of poor mixing for MCMC in ET. In Section 4 we adapt nonparametric posterior learning for ET context and derive our sampling algorithms. In Section 5 we study theoretically the asymptotic properties of our algorithms. In Section 6 we discuss our results and possibilities for future work.

2. Preliminaries

2.1. Notations

By $\mathbb{N}_0$ we denote the set of non-negative all integers, $\mathbb{R}^n_+$ denotes the nonnegative cone of $\mathbb{R}^n$, 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 ($\xi \sim \Gamma(\alpha, \beta)$, $\mathbb{E}\xi = \alpha\beta^{-1}$, $\mathrm{var}(\xi) = \alpha\beta^{-2}$). Let $A \in \mathrm{Mat}(d, p)$, $I \subset \{1, \dots, d\}$, then $\mathrm{cond}(A)$ denotes the condition number of A, 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 non-negative 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., Lo (1982). Finally, by $\mathcal{KL}(P, Q)$ we denote the standard Kullback-Leibler divergence between probability distributions P, Q.

2.2. Mathematical model for ET

Raw data in ET are described by vector $Y^t = (Y^t_1, \dots, Y^t_d) \in (\mathbb{N}_0)^d$ called sinogram which stands for the photon counts recorded during exposure time t along d lines of response (LORs). It is assumed that

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

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 introduce the 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 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., Hohage and Werner (2016), Natterer (2001)), which in practice means that

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

Remark 1. *Numerically A represents a discretized version of weighted Radon transform operator R_a for ET with complete data (see e.g., Natterer (2001)). Since A approximates R_a in strong operator norm (e.g., for $R_a : L_0^2(D) \to L_0^2([-1, 1] \times \mathbb{S}^1)$, D is the centered unit ball in $\mathbb{R}^2$) we know 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 A is injective for p large enough, due to (2.7), it may happen that $\mathrm{cond}(A) > \varepsilon_F^{-1}$, where ε_F is the floating-point precision. In the latter case, due to the cancelling effect singular values of A numerically will be equivalent to machine zeros which means then 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:

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

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

For A satisfying (2.6) and for any Y^t function $L(\lambda \mid Y^t, A, t)$ is not strictly convex even at the point of the global minima since $L(\lambda + u \mid Y^t, A, t) = L(\lambda \mid Y^t, A, t)$ 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 $L(\lambda \mid Y^t, A, t)$, so we also consider the penalized negative log-likelihood:

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

where $\beta^t \geq 0$ is the regularization coefficient. 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.

2.3. Regularization penalty

The role of regularization penalty $\varphi(\lambda)$ in (2.10) is to decrease the numerical instability in the underlying inverse problem and to make function $L_p(\lambda \mid Y^t, A, t, \beta^t)$ more convex, especially in directions close to $\ker A$.

In view of this we assume that

$$\varphi \text{ is continous and convex on } \mathbb{R}^p, \tag{2.11}$$

$$g_u(w) = \varphi(u + w) \text{ is strictly convex in } w \in \ker A \text{ for any } u \in \mathrm{Span}(A^T). \tag{2.12}$$

In Subsection 5.2 and in our proofs we use extensively the following technical result.

Lemma 2.1. *Let $\varphi(\lambda)$ be the function satisfying (2.11), (2.12), 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*

$$\{w : \lambda + u + w \succeq 0,\ w \in \ker A\} \text{ is non-empty for any } u \in U. \tag{2.13}$$

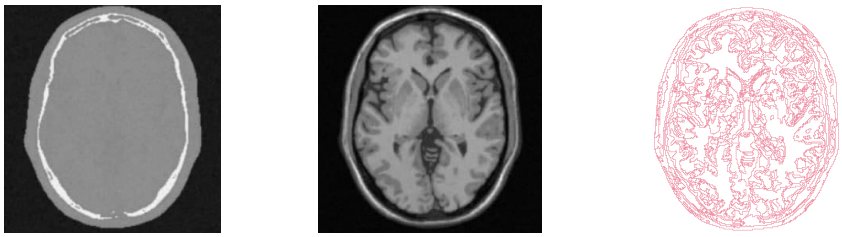
Then, mapping defined by the formula

$$w_{A,\lambda}(u) = \underset{\substack{w:\lambda+u+w\succeq 0,\\ w\in\ker A}}{\arg\min}\ \varphi(\lambda + u + w),\ u \in U \tag{2.14}$$

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

2.4. Multimodal data for ET

From (2.1) one can see that recorded signal Y^t is essentially the Poisson noise for which its signal-to-noise ratio (SNR) is proportional to $\sqrt{t\Lambda}$ 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 – images from CT or MRI. We choose MRI since it provides anatomical information with high contrast in soft tissues in comparison to CT (see Figures 1 (a), (b)).



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

Fig 1: Multimodal data for ET of the brain

We assume that our exterior data consists of r presegmented MRI images $\mathcal{M} = \{M_1, \ldots, M_r\}$ (see Figure 1 (c)) (segmentations of MRI images are precomputed using the ddCRP algorithm from Ghosh et al. (2011)). In fact, MRI-guided reconstructions in PET is an active topic of research (see the discussion in Filipović et al. (2021) 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., Bowsher et al. (2004). Current use of MRI data is purely image-based: spatially regularizing penalties are constructed using MRI data in Bowsher et al. (1996), Bowsher et al. (2004), Vunckx et al. (2011), models built upon MRI-segmented data for locally-constant tracer distribution are used in Filipović et al. (2018) and also in our work. Our approach is ideologically different from ones in Vunckx et al. (2011) because we use $\mathcal{M}$ to construct models of tracer distributions and then sample "pseudo-sinogram" to mix it with real observed data Y^t. That is MRI data are used only in observation space for ET. This has a practical feature 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" generated from the MRI-based models.

3. A motivating example for NPL in ET

Recently a Gibbs-type sampler was proposed in Filipović et al. (2018) for Bayesian inference for PET-MRI. Despite a number of positive practical features (spatial regularization, use of multimodal data) the problem of slow mixing for the corresponding Markov chain was observed. Below we consider its simplified version which shares the same mixing problem and explain the phenomenon numerically and theoretically.

In algorithms for ETs it is common to introduce data augmentation (latent variables) $n^t = \{n_{ij}^t\}$, where n_{ij}^t is the number of photons emitted from pixel j and detected in LOR i, $n_{ij}^t \sim \mathrm{Po}(ta_{ij}\lambda_j)$, n_{ij}^t are mutually independent for all (i,j); see e.g., Shepp and Vardi (1982).

In view of this physical interpretation of n^t, for variable (n^t, Y^t) the following coherence condition must be satisfied:

$$\sum_{j=1}^{p} n_{ij}^t = Y_i^t \text{ for all } i \in \{1,\dots,p\}. \tag{3.1}$$

From (3.1) it follows that Y^t is a function of n^t, so (Y^t, n^t) is indeed a data augmentation of Y^t. Note that n^t are not observed in a real experiment but n^t greatly simplifies the design of samplers (see e.g., James (2003), Filipović et al. (2018)), because conditional distributions $p(n^t \mid Y^t, A, \lambda, t)$, $p(\lambda \mid n^t, A, t)$ 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:

$$\pi(\lambda) = \prod_{j=1}^{p} \pi_j(\lambda_j),\ \pi_j = \Gamma(\alpha, \beta^{-1}),\ \alpha > 0,\ \beta > 0, \tag{3.2}$$

where α, β are some fixed constants. For the prior in (3.2) and model (2.1) conditional distributions $p(n^t \mid Y^t, A, \lambda, t)$, $p(\lambda \mid n^t, A, t)$ are as follows:

$$\begin{aligned} p(n_{ij}^t \mid Y^t, A, \lambda, t) &= \mathrm{Multinomial}(Y_i^t, p_{i1}(\lambda), \dots, p_{ip}(\lambda)), \\ p_{ij}(\lambda) &= \frac{a_{ij}\lambda_j}{\sum_k a_{ik}\lambda_k},\ i \in \{1,\dots,d\}, \end{aligned} \tag{3.3}$$

$$p(\lambda \mid n^t, Y^t, A, t) = \Gamma\left(\sum_{i=1}^{d} n_{ij}^t + \alpha, (tA_j + \beta)^{-1}\right), \tag{3.4}$$

where A_j is defined in (2.4).

Using (3.3), (3.4) the construction a Gibbs sampler for Bayesian posterior sampling from $p(\lambda \mid Y^t, A, t)$ is straightforward.

Algorithm 1: Gibbs sampler for $p(\lambda \mid Y^t, A, t)$

Data: sinogram Y^t

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

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

  Sample $n_k^t \sim p(n^t \mid Y^t, A, \lambda_{k-1}, t)$ Sample $\lambda_k^t \sim p(\lambda \mid n_k^t, Y^t, A, t)$

end

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

Result: empirical distribution of $\{\lambda_k^t\}_{k=1}^B$ approximates posterior $p(\lambda \mid Y^t, A, t)$

Remark 2. *One may argue that prior in (3.2) 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. However, we consider the mixing rate for the Markov chain in Algorithm 1 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 Bochkina and Green (2014)) that asymptotically for $t \to +\infty$ any prior effect will disappear no matter the choice of $\pi(\lambda)$.*

We choose $h(\lambda)$ to be linear, i.e., $h(\lambda) = h^T\lambda$, for some $h \in \mathbb{R}^p$, and consider the correlations between values of $h(\lambda)$ for subsequent samples from the Markov chain in Algorithm 1

$$\gamma^t(h) = \mathrm{corr}(h(\lambda^t_{k+1}), h(\lambda^t_k) \mid Y^t, t). \tag{3.5}$$

In formula (3.5) we assumed that the chain is in stationary state, i.e. k can be any.

Markov chain for the sampler in Algorithm 1 coincides with data augmentation schemes from Liu (1994), Liu et al. (1994), where the latter are exactly Gibbs samplers with only one layer of latent variables. In Bayesian context $\gamma^t(h)$ is known as fraction of missing information; see Liu (1994). In particular, in Liu (1994) authors gave an exact formula for $\gamma^t(h)$ which can be written for our example as follows:

$$\gamma^t(h) = 1 - \frac{\mathbb{E}[\mathrm{var}(h(\lambda) \mid n^t, Y^t, t) \mid Y^t, t]}{\mathrm{var}(h(\lambda) \mid Y^t, t)}. \tag{3.6}$$

For simplicity assume that

$$\lambda_{*j} > 0 \text{ for all } j \in \{1, \dots, p\}. \tag{3.7}$$

Exact formulas for the nominator and the denominator in (3.6) for arbitrary t seem difficult (if possible) to obtain, however, in the asymptotic regime $t \to +\infty$ one can apply the Bernstein von-Mises type theorem from Bochkina and Green (2014) and arrive to the following simple expression:

$$\gamma(h) = \lim_{t\to+\infty} \gamma^t(h) = 1 - \frac{h^T F^{-1}_{aug}(\lambda_*)h}{h^T F^{-1}_{obs}(\lambda_*)h},\ h \in \mathbb{R}^p,\ \text{ a.s. } Y^t, t \in (0, +\infty). \tag{3.8}$$

where

$$\lambda_* \in \mathbb{R}^p_+ \text{ is the true parameter,} \tag{3.9}$$

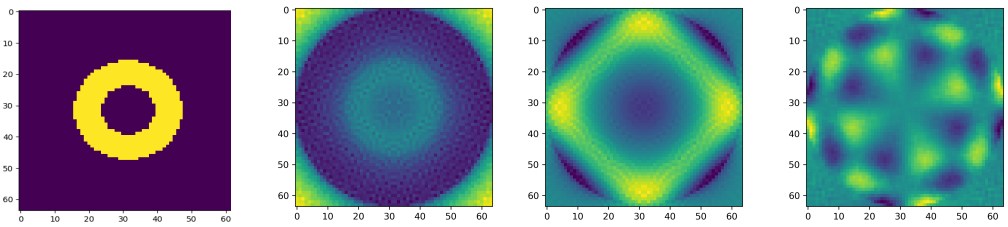
$$F_{obs}(\lambda_*) = \sum_{i=1}^{d} \frac{a_i a_i^T}{\Lambda_i^*} = A^T D^{-1}_{\Lambda^*} A,\ D_{\Lambda^*} = \mathrm{diag}(\dots, \Lambda_i^*, \dots),\ \Lambda_i^* = a_i^T\lambda_*, \tag{3.10}$$

$$F_{aug}(\lambda_*) = \mathrm{diag}(\dots, c_j, \dots),\ c_j = A_j/\lambda_{*j}. \tag{3.11}$$

From (2.5), (3.7) it follows that $\Lambda_i^* > 0$ for all i, therefore division by Λ_i^* in (3.10) is well-defined. Matrices $F_{obs}(\lambda_*)$, $F_{aug}(\lambda_*)$ are the Fisher information matrices at λ_* for Poisson models with observables Y^t, n^t, respectively. Note also that F_{obs} is not invertible in the usual sense, so in (3.8) its pseudo-inversion in the sense of Moore-Penrose is considered.

Remark 3. *Assumption in (3.7) is not practical and a precise analytic formula which extends (3.8) for $\lambda_* \in \partial\mathbb{R}^p_+$ can be established using the results from Bochkina and Green (2014). The point is that model (2.1) is 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 the case in (3.7) since we are mostly interested in mixing times of the Markov chain in areas with positive tracer concentration.*

Let $h_1, \dots, h_p$ be the orthonormal basis of eigenvectors of $F_{obs}(\lambda_*)$ being ordered according to their corresponding eigenvalues $s_1 \geq s_2 \cdots \geq s_p \geq 0$. Intuitively, vectors $\{h_m\}_{m=1}^p$ constitute a basis in space of reconstructed images where higher indices m correspond to higher frequencies on images (see Figure 2).



(a) λ_* (b) h_1 (c) h_2 (d) h_{50}

Fig 2: eigenvectors h_m for $F_{obs}(\lambda_*)$

From (3.8) it follows that

$$\gamma(h_m) = 1 - s_m h_m^T F_{aug}^{-1} h_m. \tag{3.12}$$

Matrix $F_{aug}(\lambda_*)$ is well-conditioned, continuously invertible and the quadratic term in (3.12) admits the following bound:

$$F_{aug}^{-1}(\lambda_*) = \mathrm{diag}(\dots, \frac{\lambda_{*j}}{A_j}, \dots) \Rightarrow h_m^T F_{aug}^{-1}(\lambda_*) h_m \le \frac{\max_j(\lambda_{*j})}{\min_j(A_j)}. \tag{3.13}$$

Regular behavior of F_{aug}^{-1} 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. From (3.10) and the ill-conditioning nature of A it follows that $F_{obs}(\lambda_*)$ is ill-conditioned[1], moreover, $s_m \approx 0$ for large m. From this and (3.12), (3.13) we conclude that

$$\gamma(h_m) \approx 1 \text{ for large } m. \tag{3.14}$$

Formulas (3.5), (3.14) constitute a clear evidence of poor mixing in the Markov chain in Algorithm 1. Though (3.8)-(3.14) were derived for $t \to +\infty$, they reflect well the behavior of the chain for moderate t which is seen from the numerical experiment below (see Supplementary Materials, Section F for details).

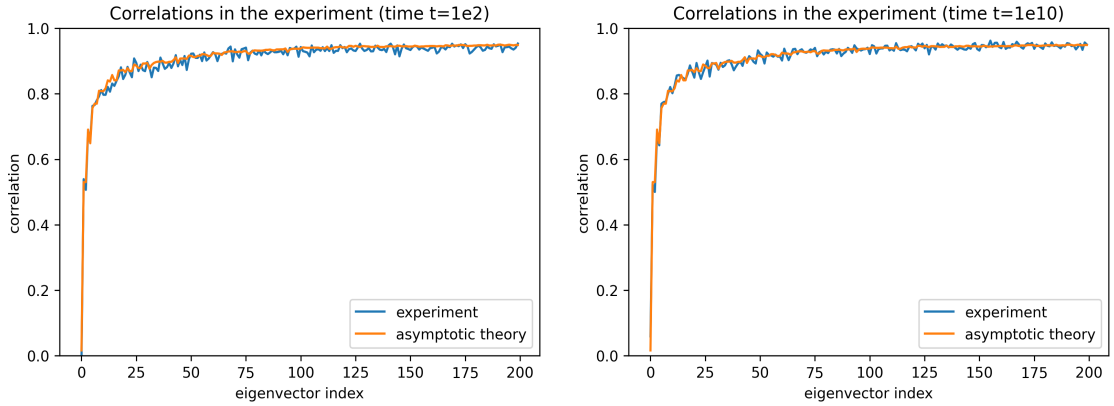


Fig 3: $\mathrm{corr}(h^T\lambda_k^t, h^T\lambda_{k+1}^t \mid Y^t)$ for $t = 10^2, 10^{10}$ for $h = h_m$; blue curve – empirical correlations computed from 2000 samples, orange curve – values for $\gamma(h_m)$ for $m = 1, \dots, 200$ by formula (3.8).

In Figure 3 values given by formula (3.12) are in full correspondence with our numerical results demonstrating that $\mathrm{corr}(h_m^T\lambda_k^t, h_m^T\lambda_{k+1}^t \mid Y^t, t)$ increase fast with m. Here one concludes that mixing is much slower for high-frequency parts of images. Therefore, to estimate reliably, say mean $h^T\lambda$ ($h \in \mathbb{R}^p$ may be a domain mask), one needs almost infinite number of samples if h contains a high-frequency component in terms of $\{h_m\}_{m=1}^p$ (see Supplementary Material, Section E for details). This also can be seen as a recommendation for choosing 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.

Note that such behavior of the sampler is not due to the choice of pixel-wise prior but due to sampling of n_{ij}^t which correspond to observations for the well-posed inverse problem. In this situation a practical advice would be to avoid sampling of missing data in the Markov chain or to use a strong smoothing prior/regularizer (for example by greatly increasing regularization coefficients so that asymptotic arguments in (3.8) 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 Van Dyk and Meng (2001) saying that design of a data augmentation scheme while preserving good mixing in the Markov chain is an "Art", especially in the case of ill-posed inverse problems. In view of poor mixing, complexity of the design and implementation, lack of scalability and high numerical load while using MCMC (Weir (1997), Higdon et al. (1997), Ferreira and Lee (2007), Duan et al. (2018), Filipović et al. (2018)) we turn to NPL as a practical relaxation of Bayesian sampling for the problem of ETs.

[1]In practice the ill-conditioning of $F_{obs}(\lambda_*)$ is commonly observed in ET practice in form of very slow convergence of non-penalized EM-algorithms; see Green (1990).

4. Nonparametric posterior learning for emission tomography

4.1. Nonparametric model for emission tomography

Nonparametric framework for ET 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 in the acquisition geometry of a scanner (e.g., $Z = (\mathbb{R} \times \mathbb{S}^1)/\mathbb{Z}_2$, i.e., all non-oriented straight lines in $\mathbb{R}^2$) for full angle acquisition in a single plane slice. For completeness we assume that Z is equipped with a boundedly-finite measure dz (which reflects the sensitivity of detectors for various lines) and with a metric ρ_Z describing distances between the lines (e.g., ρ_Z 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 random measure Z^t generated by a counting point process:

$$Z^t = \sum_{j=1}^{N^t} \delta_{(t_j, z_j)},\ (t_j, z_j) \in (0, \infty) \times Z,\ t_j \le t_{j+1},\ t_j \le t, \tag{4.1}$$

where

$$N^t \text{ is total number of registered photons,} \tag{4.2}$$

$$\{z_j\}_{j=1}^{N^t},\ \{t_j\}_{j=1}^{N^t} \text{ are the LORs and arrival times of registrations, respectively.} \tag{4.3}$$

In practical literature on PET/SPECT sample Z^t is known as list-mode data, whereas sinogram Y^t is the version of Z^t binned to finite spatial resolution and integrated withing interval $[0, t)$. Under the assumption of temporal stationarity Y^t contains the same amount of information as Z^t since the first one is then a sufficient statistic.

For statistical model of Z^t one takes 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 parameter in LOR $z \in Z$ during time interval $[0, t)$ is $t\Lambda(z)dz = t[A\lambda](z)dz$.

The negative log-likelihood for $\mathcal{PP}^t_{A\lambda}$ and observation Z^t is defined via the following formula (see, e.g., Hohage and Werner (2016), Section 2):

$$\begin{aligned} L(\lambda \mid Z^t, A, \lambda, t) &= -\sum_{j=1}^{N^t} \log(\Lambda(z_j)) + \int_{Z\times[0,t)} \Lambda(z)\, dzdt \\ &= -\int_{Z\times[0,t)} \log(\Lambda) dZ^t + t\int_Z \Lambda(z)\, dz,\ \Lambda(z) = A\lambda(z). \end{aligned} \tag{4.4}$$

4.2. Misspecification and the KL-projection

In reality our model assumption on Z^t is always incorrect (i.e., $\mathcal{PP}^t_{A\lambda}$ being misspecified) and $Z^t \sim \mathcal{PP}^t$ for some point temporal stationary process $\mathcal{PP}^t$, $\mathcal{PP}^t \ne \mathcal{PP}^t_{A\lambda}$ for any $\lambda \succeq 0$. Since the (penalized) maximum log-likelihood estimates are the most popular in ET, we say that the best one can hope to reconstruct using family $\mathcal{PP}^t_{A\lambda}$ is the projection of $\mathcal{PP}^t$ onto $\mathcal{PP}^t_{A\lambda}$ in the sense of Kullback-Leibler divergence:

$$\lambda_*(\mathcal{PP}^t) = \arg\min_{\lambda\succeq 0} \mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}). \tag{4.5}$$

Note that due to temporal stationarity of $\mathcal{PP}^t$, $\mathcal{PP}^t_{A\lambda}$, parameter λ_* in (4.5) is independent of t (t being the proportionality factor in (4.5) so it has no effect on λ_*). Since A is ill-conditioned (see formula (2.6)), in general, λ_* in (4.5) may not be defined uniquely. For this we consider the penalized KL-projection defined by the formula:

$$\lambda_*(\mathcal{PP}^t, \beta^t) = \arg\min_{\lambda\succeq 0} [\mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}) + \beta^t \varphi(\lambda)], \tag{4.6}$$

where β^t is the regularization coefficient and $\varphi(\lambda)$ is a nonparametric extension of penalty from Section 2. From (4.4) and the definition of Kullback-Leibler divergence it follows that

$$\mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}) = -\int_{Z\times[0,t)} \log(\Lambda)\mathbb{E}_{\mathcal{PP}^t}[dZ^t] + t\int_Z \Lambda(z)dz, \tag{4.7}$$

where $\mathbb{E}_{\mathcal{PP}^t}$ is the expectation with respect to $\mathcal{PP}^t$. Putting together (4.6), (4.7), for the penalized KL-projection we get the following formulas:

$$\lambda_* = \arg\min_{\lambda\succeq 0} \mathbb{L}_p(\lambda \mid \mathcal{PP}^t,\, A,\, t,\, \beta^t), \tag{4.8}$$

$$\begin{aligned} \mathbb{L}_p(\lambda \mid \mathcal{PP}^t,\, A,\, t,\, \beta^t) &= -\int_{Z\times[0,t)} \log(\Lambda)\mathbb{E}_{\mathcal{PP}^t}[dZ^t] + t\int_Z \Lambda(z)dz + \beta^t\varphi(\lambda), \\ \Lambda(z) &= A\lambda(z). \end{aligned} \tag{4.9}$$

4.3. Propagation of uncertainty and the generic algorithm

Following the idea from Lyddon et al. (2018), we say that uncertainty on λ propagates from the one on $\mathcal{PP}^t$ via (4.8), (4.9). Let $\pi_{\mathcal{M}}$ be a prior in which we encode our beliefs over a set of possible $\mathcal{PP}^t$'s, that is $\pi_{\mathcal{M}}$ is a nonparametric prior on spatio-temporal point processes on $(0,\infty)\times Z$. In particular, $\pi_{\mathcal{M}}$ is constructed using multimodal data $\mathcal{M}$. Let data be the list-mode Z^t (or the sinogram Y^t), then our prior beliefs on $\mathcal{PP}^t$ can be updated in form of posterior distribution $\pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t)$. In this case the definition of NPL for ET with multimodal data is straightforward as shown below.

Algorithm 2: NPL for ET with multimodal data

Data: list-mode Z^t or sinogram Y^t, $\mathcal{M}$

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

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

  Draw point process $\widetilde{\mathcal{PP}}^t \sim \pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t)$;

  Compute $\widetilde{\lambda}^t_b = \arg\min_{\lambda\succeq 0} \mathbb{L}_p(\lambda \mid \widetilde{\mathcal{PP}}^t, A, t, \beta^t)$ for $\mathbb{L}_p(\cdot)$ defined in (4.9);

end

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

As it has already been outlined in Lyddon et al. (2018), Fong et al. (2019), the above scheme generates i.i.d samples and is trivially parallelizable which is a strong advantage in front of MCMC sampling from pure Bayesian posteriors (see Section 3).

4.4. Construction of $\pi_{\mathcal{M}}(\cdot)$ and of posterior $\pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t)$

Sample $Z^t \sim \mathcal{PP}^t$ is a purely atomic random measure on $(0,\infty)\times Z$ which stands for photon registration events along various lines $z \in Z$ during period $[0,t)$. It is intuitive to assume mutual independence of emission events inside the patient, which is then translated as follows:

$$\begin{aligned} &\text{for any finite family of mutually disjoint bounded Borel sets } \{B_i\}_{i=1}^N,\ B_i \in B(Z), \\ &Z^t(B_i\times[0,t)) = \int_{B_i\times[0,t)} dZ^t,\ i = 1,\dots,N, \text{ are mutually independent.} \end{aligned} \tag{4.10}$$

Measure Z^t which satisfies (4.10) is known as completely random measure; see Daley and Vere-Jones (2007), 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 and registration times differ from sample to sample) the representation theorem of Kingman says that $\mathcal{PP}^t$ is characterized uniquely by a Poisson point process with some intensity measure μ on $Z\times[0,+\infty)$; see Daley and Vere-Jones (2007), Section 10.1, Theorem 10.1.III. Therefore, any prior on $\mathcal{PP}^t$ must be a prior on μ.

In view of the above discussion and temporal stationarity of $\mathcal{PP}^t$ we assume that

$$\mathcal{PP}^t = \mathcal{PP}^t_\Lambda, \text{ for some intensity } \Lambda \text{ on } Z, \text{ that is if } Z^t \sim \mathcal{PP}^t, \text{ then} \tag{4.11}$$
$$Z^t(B \times [0,t)) \sim \text{Po}(t\Lambda(B)),\ \Lambda(B) = \int_B \Lambda(z)\,dz, \text{ for any } B \in B(Z).$$

Note that $Y^t(B) = Z^t(B \times [0,t))$, where Y^t are the sinogram data. The above assumption can also be interpreted that we do not rely completely on design A when inferring on $\mathcal{PP}^t$ (moreover, A is known only approximately in practice).

Hence, to build $\pi_{\mathcal{M}}$ we construct a prior on Λ using $\mathcal{M}$. For the prior on Λ we use the mixture of gamma processes (further denoted by MGP) which can be written as follows:

$$\Lambda_{\mathcal{M}} \sim P_{\mathcal{M}}(\cdot),\ \Lambda \mid \Lambda_{\mathcal{M}} \sim GP(\theta^t \Lambda_{\mathcal{M}}, (\theta^t)^{-1}\mathbb{1}_Z), \tag{4.12}$$

where $\Lambda_{\mathcal{M}}$ is the mixing parameter, $P_{\mathcal{M}}(\cdot)$ is the mixing distribution (hyperprior), θ^t is a positive scalar, $\mathbb{1}_Z$ is the identity function on Z, $GP(\alpha,\beta) = G_{\alpha,\beta}$ is the weighted gamma process on Z (shape α and scale β).

In short, we will use the following notation

$$\pi_{\mathcal{M}}(\cdot) = \text{MGP}(P_{\mathcal{M}}, t, \theta^t \Lambda_{\mathcal{M}}, (\theta^t)^{-1}). \tag{4.13}$$

Note that the scale parameter in the gamma process in (4.12) is constant for all Z and is equal to $(\theta^t)^{-1}$. Such choice allows to center gamma process Λ on $\Lambda_{\mathcal{M}}$, so θ^t controls only the spread (e.g., $\theta^t = 0$ corresponds to improper uniform distribution on Z, $\theta^t = +\infty$ corresponds to prior $\mathcal{PP}^t_{\Lambda_{\mathcal{M}}}$, where $\Lambda_{\mathcal{M}} \sim P_{\mathcal{M}}(\cdot)$).

The key to compute the posterior for MGP in (4.13) is the following theorem which is an adaptation of Theorem 3.1 from Lo (1982).

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

From the result of Theorem 1 it follows that posterior for MGP in (4.13) is also an MGP:

$$\pi_{\mathcal{M}}(\cdot \mid Z^t, t) = \text{MGP}(P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t), t, Y^t + \theta^t \widetilde{\Lambda}^t_{\mathcal{M}}, (\theta^t + t)^{-1}), \tag{4.14}$$

where $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$ is posterior for the mixing parameter. From (4.12)-(4.14) one can see that samples from the MGP posterior are normalized random pseudo-sinograms $\widetilde{\Lambda}^t_{\mathcal{M}}$ in the MRI-based model linearly combined with observed data Z^t. Therefore, regularizing effect of MRI originally takes place in the observation space through pseudo-observations.

Remark 4. *MGP prior in (4.13) and the posterior in (4.14) are direct analogs of MDP (Mixture of Dirichlet processes) prior and posterior from Lyddon et al. (2018), respectively. Weighted gamma processes as priors were also considered in James (2003) for various semiparametric intensity models including very elaborate Poisson model for PET (temporal non-stationarity, detector transition kernels). In particular, in James (2003) a weighted gamma prior was used in the image space (i.e., as a prior on λ) but not in observation space and the sampling from posteriors was based on data augmentation schemes similar to the one in Section 3 for which MCMC is difficult. In our approach most of complexity is moved to construction of a "good" prior in observation space which should be initially centered at the true (KL-optimal) intensity map built from MRI data which also puts zero (or small) mass on $\Lambda \notin R_+(A)$ (see also formula (4.12)).*

4.5. 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 LORs. Mathematically it means that $Z = \left(\bigsqcup_{i=1}^d Z_i\right) \bigsqcup \overline{Z}$, where each set $Z_i \in B(Z)$ corresponds to set of lines which are visible in LOR i, $\overline{Z}$ are the lines which are not visible at all. For each i we define binning of the data and the corresponding intensities by the formulas:

$$\left(\int_{Z_i \times [0,t)} dZ^t, \int_{Z_i} \Lambda(z)\,dz\right) = (Y^t_i, \Lambda_i), \tag{4.15}$$

$$Y_i^t \text{ are mutually independent and } Y_i^t \sim \mathrm{Po}(t\Lambda_i), i \in \{1, \dots, d\}. \tag{4.16}$$

Nonparametric weighted gamma prior and its posterior in (4.13), (4.14), penalized negative log-likelihood in (4.9) are also binned in a similar way with (4.15), so the finite-dimensional version of Algorithm 2 can be written as follows

Algorithm 3: Binned NPL for ET with multimodal data

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

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

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

2 Draw $\widetilde{\Lambda}^t_{\mathcal{M}} = (\widetilde{\Lambda}^t_{\mathcal{M},1}, \dots, \widetilde{\Lambda}^t_{\mathcal{M},d})$ from $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Y^t, t)$;

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

4 Compute $\widetilde{\lambda}^t_b = \arg\min_{\lambda \succeq 0} L_p(\lambda \mid \widetilde{\Lambda}^t_b, A, t, \beta^t/t)$ for $L_p(\cdot)$ defined in (2.10);

5 **end**

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

Remark 5. *In steps 1, 2 intensities $\widetilde{\Lambda}^t_{b,i}$ are sampled from the binned MGP posterior in (4.14). In step 3 we have used the fact that binned version of $\mathbb{L}_p(\cdot)$ from (4.9) coincides with $L_p(\cdot)$ from (2.10). In addition, from formula (2.10) it follows that*

$$L_p(\lambda \mid t\widetilde{\Lambda}^t_b, A, t, \beta^t) = tL_p(\lambda \mid \widetilde{\Lambda}^t_b, A, 1, \beta^t/t) + R, \tag{4.17}$$

where R is a function which is independent of λ. Therefore, minimization in step 3 is directly applied to $L_p(\lambda \mid \widetilde{\Lambda}^t_b, A, 1, \beta^t/t)$ instead of $L_p(\lambda \mid t\widetilde{\Lambda}^t_b, A, t, \beta^t)$. If the complexity of sampling in step 1 is controlled by our choice of $P_{\mathcal{M}}(\cdot)$, step 3 is inevitable, hence, it must be numerically feasible via some scalable optimization algorithm. This is the case for us in view of the well-known in ET the Generalized Expectation-Maximization(GEM)-type algorithm from Fessler and Hero (1995) which 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 our numerical experiment (see Supplementary Material H.1).

4.6. Final algorithm

First, we explain the intuition behind sampling from $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \,|\, Y^t, t)$ in step 1 in Algorithm 3, then, we present the formal and complete procedure.

Using $\mathcal{M}$ (see Figure 1(c)) we construct a model of type (2.1) for which we assume that the isotope's concentration is constant in each segment and has uniform (improper) prior distribution on $\mathbb{R}_+$. If $\lambda_{\mathcal{M}} \in \mathbb{R}_+^{p_{\mathcal{M}}}$ be the corresponding random vector ($p_{\mathcal{M}}$ being the number of segments), then a sample from the prior $P_{\mathcal{M}}(\cdot)$ is defined as $\Lambda_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M}}$, where $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$ is the design for segment-like model of ET computed directly from A (see formulas (C.2), (C.3)). The point is that $p_{\mathcal{M}} \ll p$, so $A_{\mathcal{M}}$ is of moderate size (hence, can be stored in memory), is also injective and well-conditioned. Posterior $P(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Y^t, t)$ is defined via classical Bayes' formula for model $P(Y^t \mid A_{\mathcal{M}}, \lambda_{\mathcal{M}}, t) = \mathrm{Po}(t\Lambda_{\mathcal{M}})$ and the aforementioned prior on $\Lambda_{\mathcal{M}}$. Formal constructions of $P_{\mathcal{M}}(\cdot)$, $P_{\mathcal{M}}(\cdot \mid Y^t, t)$ are given in Supplementary Material, Section C. In practice, for the sake of simplicity we sample from $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \,|\, Y^t, t)$ using the weighted log-likelihood bootstrap (WLB) adapted for ET.

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

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

Input: $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$ from (C.2), (C.3) (well-conditioned)

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

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

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

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

Remark 6. *Since we assume that $A_{\mathcal{M}}$ is well-conditioned, minimizer $\widetilde{\lambda}^t_{\mathcal{M}}$ in step 2 of Algorithm 4 can be efficiently computed via the classical EM-algorithm from Shepp and Vardi (1982).*

Algorithm 5: Binned NPL for ET with MRI data

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

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

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

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

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

4 Compute $\widetilde{\lambda}^t_b = \arg\min_{\lambda \succeq 0} L_p(\lambda \mid \widetilde{\Lambda}^t_b, A, 1, \beta^t/t)$ for $L_p(\cdot)$ defined in (2.10) using the GEM-type algorithm from Fessler and Hero (1995);

5 **end**

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

Remark 7. *Parameter θ^t in Algorithm 5 has the following physical meaning: it is exactly the rate of creation of "pseudo-photons" in the poisson model constructed from MRI data. More precisely, by choosing $\theta^t = \rho t$, $\rho \geq 0$ in step 2 we sum up sinograms Y^t and $t\widetilde{\Lambda}^t_{\mathcal{M}}$ in proportions $1/(1+\rho)$ and $\rho/(1+\rho)$, respectively. For $\theta^t = 0$ we see Algorithm 5 as a version of WLB from Newton and Raftery (1994) being adapted for the ET context; see also Lyddon et al. (2018), Fong et al. (2019), Pompe (2021) for connections between classical WLB and NPL.*

Numerical tests of Algorithm 5 are given in the Supplementary Material, Section G.1.

5. 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, in general, $\lambda_* \in \partial\mathbb{R}^p_+$. This model was investigated in the small noise limit (i.e., when $t \to +\infty$) in pure Bayesian framework in Bochkina and Green (2014) for large class of priors for the well-specified case (i.e., $Y^t \sim P^t_{A,\lambda_*}$ 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 $L(\lambda \mid Y^t, A, t)$ and the non-regularity results in splitting of the posterior in three modes: multivariate exponential (for coordinates which are related to pixels intersected by LORs with zero photon intensities) contracting to zeros with the fastest rate (scaled with t), Gaussian (for pixels where $\lambda_{*,j} > 0$) and half-Gaussian (for pixels with $\lambda_{*,j} = 0$ and pixels being intersected only by LORs with positive intensities) contracting with standard rate (scaled with $\sqrt{t}$).

Our results for consistency and conditional distribution are similar to ones from Bochkina and Green (2014), however, there are several major and minor differences. Asymptotic consistency at λ_* and a very similar splitting are also present in NPL, with the asymptotic distribution being tight

around a strongly consistent estimator $\widehat{\lambda}^t_{sc}$ satisfying additional properties in observation space. The assumptions we put on $\widehat{\lambda}^t_{sc}$ for conditional tightness seem very natural and we discuss them thoroughly in the text. Interestingly, the splitting of the posterior into different modes depends not on λ_* (as it was in Bochkina and Green (2014)) but again on $\widehat{\lambda}^t_{sc}$ because of which yet we fail to demonstrate the asymptotic normality since it requires additional results on behavior of strongly consistent estimators with constraints on the domain. Intuitively, the asymptotic distribution should be similar to the frequentist distribution of MAP estimates from Bochkina and Green (2014): 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 Geyer (1994)). We address this investigation for future and conjecture that classical MLE or penalized MLE (i.e., MAP) from Bochkina and Green (2014) are the right candidates for $\widehat{\lambda}^t_{sc}$.

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 also contain restrictions on the above parameters. At the end, we address the problem of model misspecification for the generalized Poisson model with wrong design which arises twice our setting: first, in Algorithm 4 when sampling $\widetilde{\Lambda}^t_{\mathcal{M}}$ (because we use Y^t with incorrect design $A_{\mathcal{M}}$) and, second, when assume that model (2.1) is wrong, in general.

5.1. Convergence for conditional probabilities.

Let $(\Omega, \mathcal{F}, P)$ be the common probability space on which process Y^t, $t \in (0, +\infty)$ and MGP prior in (4.13) are defined (see Supplementary Material, Section A for details). Let

$$\mathcal{F}^t = \sigma(Y^\tau,\, \tau \in (0, t)) \subset \mathcal{F}, \tag{5.1}$$

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

Definition 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:*

$$P(\|U^t - U\| > \varepsilon \mid \mathcal{F}^t) \to 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{5.2}$$

This type of convergence will be denoted as follows:

$$U^t \xrightarrow{c.p.} U \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{5.3}$$

□

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

$$U^t = o_{cp}(1). \tag{5.4}$$

Definition 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*

$$\sup_{t\in(0,+\infty)} P(\|U^t\| > M \mid \mathcal{F}^t) < \varepsilon. \tag{5.5}$$

In short, in the definition above almost surely Y^t, $t \in (0, +\infty)$ means that statements in (5.2), (5.5) hold for almost every trajectory Y^t, $t \in (0, +\infty)$.

5.2. Consistency

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

$$Y^t \sim P^t_{A,\lambda_*}, \text{ for some } \lambda_* \in \mathbb{R}^p_+ \text{ and all } t \in (0, +\infty), \tag{5.6}$$

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

Theorem 2. *Let Assumption 1 and conditions (2.11), (2.12) for $\varphi(\lambda)$ be satisfied. Let also β^t, θ^t be such that*

$$\beta^t/t \to 0,\ \theta^t/t \to 0 \ \text{ when } t \to +\infty. \tag{5.7}$$

Then,

$$\widetilde{\lambda}_b^t - \lambda_* \xrightarrow{c.p.} w_{A,\lambda_*}(0) \ \text{ when } t \to +\infty, \ \text{ a.s. } Y^t,\ t \in (0,+\infty), \tag{5.8}$$

where $\widetilde{\lambda}_b^t$ is sampled in Algorithm 5, $w_{A,\lambda}(\cdot)$ is defined in (2.14).

Conditional distribution of $\widetilde{\lambda}_b^t$ asymptotically concentrates at 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 $\varphi(\lambda)$ and positivity constraints at λ_*.

There is also an extension of the above result for any generic bootstrap type procedure provided that perturbation of Y^t asymptotically is not too excessive.

Theorem 3. *Let conditions of Theorem 2 be satisfied but Assumption 1. Assume also that*

$$\begin{aligned} &\widetilde{\Lambda}_{b,i}^t \xrightarrow{c.p.} \Lambda_i^* = a_i^T \lambda_*,\ i = 1,\dots,d, \ \text{ for some } \lambda_* \in \mathbb{R}_+^p \\ &\text{when } t \to +\infty, \ \text{ a.s. } Y^t,\ t \in (0,+\infty). \end{aligned} \tag{5.9}$$

Then, formula (5.8) remains valid.

5.3. Tightness and the asymptotic distribution

Assumption 2. *$A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$ is injective.*

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 satisfies the property in (2.4). Consider set $\lambda_{\mathcal{M},*}$ which is defined by the formula:*

$$\lambda_{\mathcal{M},*} = \arg\min_{\lambda_{\mathcal{M}} \succeq 0} L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1), \tag{5.10}$$

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

$$I_0(\Lambda_{\mathcal{M}}^*) = I_0(\Lambda^*),\ \Lambda_{\mathcal{M}}^* = A_{\mathcal{M}} \lambda_{\mathcal{M},*}, \tag{5.11}$$

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

The proposition below states that Assumption 3 is always meaningful and not restrictive at all.

Proposition 1. *Let $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$, $A_{\mathcal{M}}$ has only positive entries and the property in (2.4) holds. Then, for any $\Lambda^* \in \mathbb{R}_+^d$ set of minimizers $\lambda_{\mathcal{M},*}$ defined in (5.10) is always non-empty and constitutes an affine subset of $(p_{\mathcal{M}} - 1)$-dimensional simplex $\Delta_{A_{\mathcal{M}}}^p(\Lambda^*)$ defined by the formula*

$$\Delta_{A_{\mathcal{M}}}^{p_{\mathcal{M}}}(\Lambda^*) = \{\lambda_{\mathcal{M}} \in \mathbb{R}_+^p \mid \sum_{j=1}^{p_{\mathcal{M}}} A_{\mathcal{M},j} \lambda_{\mathcal{M},j} = \sum_{i=1}^{d} \Lambda_i^* \geq 0\},\ A_{\mathcal{M},j} = \sum_{i=1}^{d} a_{\mathcal{M},ij} > 0. \tag{5.12}$$

Moreover, it always holds that

$$I_1(\Lambda^*) \subset I_1(\Lambda_{\mathcal{M}}^*) \ \text{ or equivalently } I_0(\Lambda_{\mathcal{M}}^*) \subset I_0(\Lambda^*), \tag{5.13}$$

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

The non-expansiveness condition is essential for us when we sample $\widetilde{\Lambda}_{\mathcal{M}}^t$ in Algorithm 4 because we know that model $P_{A_{\mathcal{M}},\lambda_{\mathcal{M}}}^t$ is strongly misspecified when we fit data Y^t in it. The aim here is still to have a unique and stable KL-minimizer $\lambda_{\mathcal{M},*}$ so that identifiability holds for $\lambda_{\mathcal{M},*}$ and the prior effect of $\mathcal{M}$ on $\widetilde{\lambda}_b^t$ is not spread ambiguously among different (but equivalent in terms of observations) combinations of tracer in segments of $\mathcal{M}$ (see Figure 1 (c)). This is provided by the theorem below.

Theorem 4 (identifiability in the prior model). *Let Assumptions 2-3 be satisfied. Then, $\lambda_{\mathcal{M},*}$ defined in (5.10) is unique and the following formula holds:*

$$L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) = \mu_{\mathcal{M},*}^T \lambda_{\mathcal{M}} + \frac{1}{2} \sum_{i \in I_1(\Lambda^*)} \Lambda_i^* \frac{(\Lambda_{\mathcal{M},i} - \Lambda^*_{\mathcal{M},i})^2}{(\Lambda^*_{\mathcal{M},i})^2} + o(\|\Pi_{A^T_{\mathcal{M},I_1(\Lambda^*)}}(\lambda_{\mathcal{M}} - \lambda_{\mathcal{M},*})\|^2), \tag{5.14}$$

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

$$\begin{aligned} &\mu_{\mathcal{M},*} = \sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \frac{a_{\mathcal{M},i}}{\Lambda^*_{\mathcal{M},i}} + \sum_{i=1}^{d} a_{\mathcal{M},i}, \\ &\mu_{\mathcal{M},*} \succeq 0,\ \mu_{\mathcal{M},*,j} \lambda_{\mathcal{M},*,j} = 0 \text{ for all } j \in \{1, \dots, p_{\mathcal{M}}\}. \end{aligned} \tag{5.15}$$

In particular, the function $L(\lambda_{\mathcal{M}} \mid \Lambda^*, A, 1)$ is locally strongly convex at $\lambda_{\mathcal{M},*}$, that is, 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

$$L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) \geq C_* \|\lambda_{\mathcal{M}} - \lambda_{\mathcal{M},*}\|^2 \text{ for any } \lambda \in B_* \cap \mathbb{R}_+^{p_{\mathcal{M}}}. \tag{5.16}$$

Result of Theorem 4 is a positive answer to the identification problem when model (2.1) is misspecified in the sense of wrong design. Here, 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 in , Subsection 5.4.

Now we can turn to our main result on the tightness of the posterior.

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

$$\mathcal{V} = \mathrm{Span}\{e_j \mid \exists\, i \in I_0(\Lambda^*) \text{ s.t. } a_{ij} > 0\}, \tag{5.17}$$

$$\mathcal{U} = \mathcal{V}^\perp \cap \mathrm{Span}\{A^T_{I_1(\Lambda^*)}\}, \tag{5.18}$$

$$\mathcal{W} = (\mathcal{V} \oplus \mathcal{U})^\perp \cap \ker A. \tag{5.19}$$

Let

$$\Pi_{\mathcal{V}}, \Pi_{\mathcal{V}}, \Pi_{\mathcal{W}} \text{ be the orthogonal projectors on } \mathcal{V}, \mathcal{V}, \mathcal{W}, \text{ respectively.} \tag{5.20}$$

Theorem 5 (tightness of the asymptotic distribution). *Let assumptions 1-3 be satisfied and assume also that*

$$\varphi \text{ satisfies (2.11), (2.12) and } \varphi \text{ is locally Lipschitz continous.} \tag{5.21}$$

Let $\widetilde{\lambda}_b^t$ 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 $\widehat{\lambda}_{sc}^t$ of λ_* on $\mathcal{V} \oplus \mathcal{U}$ (i.e., $\Pi_{\mathcal{U}\oplus\mathcal{V}} \widehat{\lambda}_{sc}^t \xrightarrow{a.s.} \Pi_{\mathcal{U}\oplus\mathcal{V}} \lambda_*$) such that

$$\widehat{\lambda}_{sc}^t \succeq 0, \tag{5.22}$$

$$\limsup_{t \to +\infty} \left| \sum_{i \in I_1(\Lambda^*)} \sqrt{t} \frac{Y_i^t/t - \widehat{\Lambda}_{sc,i}^t}{\widehat{\Lambda}_{sc,i}^t} a_i \right| < +\infty \text{ a.s. } Y^t,\ t \in (0, +\infty), \tag{5.23}$$

$$t\widehat{\Lambda}_{sc,i}^t \to 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty) \text{ for } i \in I_0(\Lambda^*), \tag{5.24}$$

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

i)

$$t\Pi_{\mathcal{V}}(\widetilde{\lambda}_b^t - \widehat{\lambda}_{sc}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{5.25}$$

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

Statement in (i) claims that in pixels which are interested by LORs with zero intensities (i.e. $\Lambda_i^* = 0$) the posterior distribution contracts to zero with faster rate than for the ones intersected by LORs with positive intensities. Indeed, pixels in subspace $\mathcal{V}$ are strongly forced to be zeros by the positivity constraints (i.e., if $\Lambda_i^* = 0$ and $\lambda_*, a_i \in \mathbb{R}^p_+$, then necessarily $\lambda_{*,j} = 0$ where $a_{ij} > 0$). Statement in (ii) claims that, in general, the posterior concentrates around $\widehat{\lambda}^t_{sc}$ in subspace $\mathcal{U}$ with standard scaling rate $\sqrt{t}$. This is not surprising since $\mathcal{U}$ does not contain projections on $\mathcal{V}$, so the positivity constraints do not give here extra information to achieve the faster contraction rate. Finally, requiring the non-expansiveness condition for the prior (i.e., Assumption 3) may seem surprising at first sight. The intuition behind is that it protects our sampler from creation of "too many" pseudo-photons in LORs where intensity is zero (i.e., $\Lambda_i^* = 0$ implies $Y_i^t \equiv 0$ for the well-specified model) and significantly simplifies the theoretical analysis.

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

$$\widehat{\lambda}^t_{pMLE} = \arg\min_{\lambda \succeq 0} L_p(\lambda \mid Y^t, A, t, \beta^t), \tag{5.26}$$

where $L_p(\cdot)$ is defined in (2.10).

Conjecture 1. *Let assumptions of Theorem 5 be satisfied and $\widehat{\lambda}^t_{sc} = \widehat{\lambda}^t_{pMLE}$, where the latter is defined by formula (5.26). Then, $\widehat{\lambda}^t_{sc}$ is a strongly consistent estimator of λ_* on $\mathcal{V} \oplus \mathcal{V}$ and formulas (5.22)-(5.24) hold.*

The requirement for existence of a strongly consistent estimator for weighted bootstrap is not new and already appears in Ng and Newton (2020). However, in that case the sampling is performed via unconstrained optimization of quadratic functionals though 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 Ng and Newton (2020). According to Kolmogorov's 0-1 Law the statements in (5.23), (5.24) either hold 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 (5.22)-(5.24) are trivially satisfied, for example, if A is diagonal. Finally, the asymptotic structure of Bayesian posterior from Bochkina and Green (2014) gives a strong intuition that the conjecture above should hold: the asymptotic posterior projected on $\mathcal{V}$ has exponential distribution $\mathrm{Exp}(-c_A t \Pi_{\mathcal{V}} \lambda)$ and $\sqrt{t}\Pi_{\mathcal{U}}\lambda$ is normal with mean equals $\sqrt{t}(A^T_{I_1(\Lambda^*)} D^{-1}_{\Lambda^*} A_{I_1(\Lambda^*)})^{-1} A^T_{I_1(\Lambda^*)} D^{-1}_{\Lambda^*}[Y^t/t]$ being also restricted to positivity cone (hence, half-Gaussian), therefore the corresponding MAP estimate asymptotically fits conditions (5.22)-(5.24) being atom at zero for the exponential part and mean for the Gaussian one (up to higher order terms). Formal investigation of Conjecture 1 and of possible $\widehat{\lambda}^t_{sc}$'s 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.

5.4. Misspecification in design and identifiability

Assumption 1 in Subsection 5.2 reflects our belief that model (2.1) is correct. At the same time, for any practitioner in ET it is known that such model is by far approximate: the tracer inside the human body surely does not respect locally constant behavior in each pixel on which our discretized model is based, also, in practice, matrix A is known only approximately, with non-negligible errors, since it contains patient's attenuation map which is reconstructed via a separate MRI or CT scan; see e.g., Stute and Comtat (2013). There also are many other practical issues which are not included in (2.1) such as 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., Levin et al. (1995), Rahmim et al. (2009).

Assuming temporal stationarity of the processwe consider the following scenario for ET:

$$Y^t \sim P^t,\ Y^t \in (\mathbb{N}_0)^d, \tag{5.27}$$

$$\mathbb{E}_{P^t}[Y^t] = \mathrm{var}_{P^t}[Y] = t\Lambda^* \text{ for some } \Lambda^* = (\Lambda^*_1, \dots, \Lambda^*_d) \in \mathbb{R}^d_+. \tag{5.28}$$

Formulas (5.27), (5.28) 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 Sitek and Celler (2015). Most importantly, we do not assume that $\Lambda^* \in R_+(A)$.

The main question now is the identifiability of λ which translated via (2.9), (5.27), (5.28) to the problem of uniqueness in the following minimization problem:

$$\lambda_{A,*}(P^t) = \arg\min_{\lambda \succeq 0} \mathcal{KL}(P^t, P^t_{A,\lambda}) = \arg\min_{\lambda \succeq 0} L(\lambda \mid \Lambda^*, A, 1), \tag{5.29}$$

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

Theorem 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 and injective such that solutions of the optimization problem* (5.29) *constitute a non-empty affine subset of positive dimension of the* $(p-1)$*-simplex* $\Delta_p(\Lambda^*) = \left\{\lambda \in \mathbb{R}^p_+ : \sum_{j=1}^p \lambda_j = \sum_{i=1}^d \Lambda^*_i\right\}$.

Proof. We construct Λ^* and A for $p = 4$, $d = 6$.

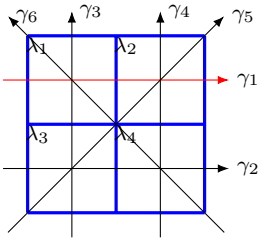


Fig 4: $\mathcal{I}$

Let $\mathcal{I}$ be the image consisting of four square pixels each with side length equal to 1 as shown in Figure 4, 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 4 and matrix A' corresponds to the the classical Radon transform on $\mathcal{I}$, i.e., a'_{ij} is the length of intersection of ray $\gamma_i \in \Gamma$ with pixel j

$$A' = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \end{pmatrix}, \ \det(A'^T A') = 128 \neq 0.$$

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 formula in (5.29) has the following form

$$\lambda_{A,*} = \arg\min_{\lambda \succeq 0} -\log\left(\frac{\lambda_1 + \lambda_2}{2+\sqrt{2}}\right) + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4. \tag{5.30}$$

Note that in (5.30) 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 (5.30) is an affine set of the following form:

$$\lambda_{A,*3} = \lambda_{A,*4} = 0, \ \lambda_{A,*1} + \lambda_{A,*2} = 1 \tag{5.31}$$

which gives the desired non-uniqueness. Theorem is proved. □

At the same time, Theorem 4 provides identifiability and stability (via strong local convexity) for $\lambda_{A,*}$ under the non-expansiveness condition and injectivity of A. The latter assumption can be also relaxed by simply restating the claim of Theorem 4 with analogs (5.14)-(5.16) to hold but only in the subspace $\mathrm{Span}(A^T)$.

6. Discussion

To build the nonparametric posterior learning for the model of ET we have used conjugacy between Poisson and Gamma processes which is analogous to the one in Lyddon et al. (2018), Fong et al. (2019) between Dirichlet and Multinomial processes. This explains why our main calibration parameter ρ ($\theta^t = t\rho$; see Remark 7) has physical interpretation as amount of pseudo-data (pseudo-photons for ET) generated from the posterior process. Possible future improvement of the method is to relax the independence of increments of the Gamma process in the prior and consider processes with correlations, for example, Gamma-weighted Polyà tree priors. Such correlations can be used to smooth sinogram Y^t (i.e., to project it approximately on the stable part of $\mathrm{Span}(A^T)$) using the MRI-based model and, in addition, remove completely regularizer φ from the model. Note that in Algorithm 5 regularization of high frequencies is achieved via control of φ and only low frequencies are regularized by $\mathcal{M}$. Our preliminary results show that new approach improves the resolution while retaining the interpretability of the calibration parameters. This is definitely a next goal for future work.

From the theoretical side a very needed step is to demonstrate Conjecture 1, which is also necessary for theoretical analysis of more complicated prior models discussed above. Work in this direction may also target studies of the first order asymptotics of the posterior (i.e., Edgeworth's expansions) which will be given elsewhere.

Our numerical tests on synthetic data in the Supplementary Material show good coverage of the true signal even for large values of ρ (empirical rule of thumb says that $\rho = 1$ is satisfactory), so new tests on real patient data are needed in future.

Supplementary material

Supplementary material includes the proof of Lemma 2.1, numerical experiments for the Gibbs sampler in Section 3 and for Algorithm 5 (provided with links to the source code), proofs of all theoretical results in Section 5, a remark on the intuition behind the non-expansiveness condition (Assumption 3) and an additional remark on the choice of centering term in Theorem 5.

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. This work is partly supported by the 'MMIPROB' project funded by ITMO Cancer (France).

Appendix 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 4.5), 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 Daley and Vere-Jones (2007), Section 9.2, Example 9.2(b).

Algorithms 4, 5 rely on perturbed intensities $\widetilde{\Lambda}^t_{\mathcal{M}}$ and $\widetilde{\Lambda}^t_b$ for which we show that they can be expressed as functions of random weighting of the list-mode data

$$G^t = \{\delta_{(k,i)} : (k,i) - k^{\text{th}} \text{ photon was detected at detector } i\},$$

where $\delta_{(k,i)} \in \{0,1\}$. Indeed, from step 1 in Algorithm 4 we can see that $\widetilde{\Lambda}^t_{\mathcal{M}}$ is a function of $\widetilde{\Lambda}^t$ for which the following representation holds

$$\widetilde{\Lambda}^t_i = t^{-1} \sum_{k=1}^{N^t} \delta_{(k,i)} \widetilde{w}_k, \; i \in \{1, \dots, d\}, \tag{A.1}$$

$$\{\widetilde{w}_k\}_{k=1}^{N^t} \overset{iid}{\sim} \Gamma(1,1), \tag{A.2}$$

where N^t is the total number of photons.

For $\widetilde{\Lambda}_b^t$ in step 2 of Algorithm 5 we have the following representation:

$$\widetilde{\Lambda}_{b,i}^t = (\theta^t + t)^{-1} \left(\sum_{k=1}^{N^t} \delta_{(k,i)} w_k + w_p \theta^t \Lambda_{\mathcal{M},i}^t \right), \; i \in \{1, \dots, d\}, \tag{A.3}$$

$$\{w_k\}_{k=1}^{N^t}, w_p \overset{iid}{\sim} \Gamma(1,1). \tag{A.4}$$

From formulas (A.1)-(A.4) one can see that perturbations $\widetilde{\Lambda}_{\mathcal{M}}^t$ and $\widetilde{\Lambda}_b^t$ depend on data Y^t and on infinite family of random mutually independent weights $(\{(w_k, \widetilde{w}_k)\}_{k=1}^{\infty}, w_p)$ which are 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 originates to Newton and Raftery (1994); similar ones have been recently used in Ng and Newton (2020).

Appendix B: Limit theorems for stationary Poisson processes.

Let

$$Y^t \sim \mathrm{Po}(\Lambda t), \; \Lambda > 0, \; t \in [0, +\infty). \tag{B.1}$$

The following result is a composition of theorems 9.3, 4.1 and 7.5 (pp. 306, 350, 417, respectively) from Gut (2013).

Theorem B.1. *Let $\{Y^t\}$, $t \in (0, +\infty)$ be the Poisson process defined in (B.1). Then,*

i)

$$\frac{Y^t}{t} \xrightarrow{a.s.} \Lambda \; \textit{as} \; t \to +\infty. \tag{B.2}$$

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 statement holds for almost any trajectory Y^t, $t \in (0, +\infty)$.

Appendix C: Binned NPL for emission tomographies with MRI data

First, we construct $P_{\mathcal{M}}$, then we proceed with construction of $P_{\mathcal{M}}(\widetilde{\Lambda}_{\mathcal{M}}^t \mid Y^t, t)$.

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 $\lambda_s^k \sim \Gamma(1, \infty)$ (uniform (improper) distribution on $\mathbb{R}_+$).
3. Compute random projections

$$\Lambda_{\mathcal{M},i} = \sum_{k=1}^{r} \sum_{s=1}^{p_k} a_{is}^k \lambda_s^k, \; \text{for each } i \in \{1, \dots, d\}. \tag{C.1}$$

where

$$a_{is}^k = \sum_{j=1}^{p} a_{ij} \mathbb{1}\{\text{pixel } j \text{ belongs to segment } s \in S(M_k)\}, \; k \in \{1, \dots, r\}. \tag{C.2}$$

Note that $\Lambda_{\mathcal{M},i}$ in (C.1) is defined through the sum of projections over all images in $\mathcal{M}$. This can be seen as concatenating r models with segmentations :

$$A_{\mathcal{M}} = (A_1, \dots, A_r) \in \mathrm{Mat}\,(d, p_{\mathcal{M}})\,,\ A_k = (a_{ij}^k) \in \mathrm{Mat}(d, p_k),\ p_{\mathcal{M}} = \sum_{k=1}^{r} p_k, \tag{C.3}$$

$$\lambda_{\mathcal{M}} = (\lambda_1^1, \dots, \lambda_{p_1}^1, \dots, \lambda_1^r, \dots, \lambda_{p_r}^r), \tag{C.4}$$

Using notations from (C.3), (C.4), formula (C.1) can be rewritten as follows:

$$\Lambda_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M}},\ \Lambda_{\mathcal{M}} = (\Lambda_{\mathcal{M},1}, \dots, \Lambda_{\mathcal{M},d}). \tag{C.5}$$

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

$$\ker A_{\mathcal{M}} = \{0\},\ \mathrm{cond}(A_{\mathcal{M}}) < c_{\mathcal{M}}, \tag{C.6}$$

where $c_{\mathcal{M}}$ is some moderate constant. The latter assumption reflects the idea that images in $\mathcal{M}$ consist of low number of large segments. In practice, condition (C.6) can be checked via the singular values of $A_{\mathcal{M}}^T A_{\mathcal{M}}$ which, in turn, can be computed due to apriori moderate size of $A_{\mathcal{M}}$. In principle, due to moderate size of $A_{\mathcal{M}}$ and good conditioning it is possible to use MCMC-approach to sample from $P_{\mathcal{M}}(\widetilde{\Lambda}_{\mathcal{M}}^t \mid Y^t, t)$, however, in order to keep the overall implementation as simple as possible we turn to WLB from Newton and Raftery (1994) for approximate posterior sampling.

Appendix D: Remark on recent bootstrap algorithms for ET

A very recent and similar to ours sampling algorithm was proposed in Filipović et al. (2021) provided with a very extensive experiment both on synthetic and real PET-MRI data. The algorithm there is also of boostrap-type, based on optimization of a randomized functional (the KL-distance) and in fact, it coincides up to minor details with Algorithm 5 for $\theta^t \equiv 0$ (i.e., without MRI). Instead, data $\mathcal{M}$ are used there to construct very special penalty $\varphi(\lambda) = \varphi_{\mathcal{M}}(\lambda)$ of Bowsher type (see Subsection 2.4). This penalty satisfes the assumptions in (2.11), (2.12), so our theorems 3, 5 serve as a theoretical foundation also for the algorithms presented there. A nice practical feature of Algorithm 5 is that θ^t has clear physical interpretation of the effect of MRI data on samples (see Remark 7), whereas large number of parameters in Bowsher-type penalties have no such easy interpretations making the problem of their calibration cumbersome for practice.

The aforementioned minor differences between algorithms consist in the way data Y^t (in Filipović et al. (2021)) 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 Algorithm 5 one can see that initially uncertainty propagates via the KL-projection in (4.5) and not concerning at all the problem of limited data. Moreover, we retrieve version of WLB of Newton and Raftery (1994) adapted for ET as a particular case of Algorithm 5 when choosing the scale parameter $\theta^t = 0$ in the nonparametric prior in (4.12) (each photon corresponds to multiplicative perturbation of the data term by $w \sim \Gamma(1,1)$). This is fully coherent with the derivation of NPL in Lyddon et al. (2018) and nonparametric posterior bootstrap with MDP-prior in Fong et al. (2019), where the classical WLB algorithm from Newton and Raftery (1994) is retrieved back as a particular case when choosing the concentration parameter $\alpha = 0$ ($c = 0$ in Fong et al. (2019)) in the nonparametric Dirichlet process prior. On the other hand, the derivation in Filipović et al. (2021) strongly relies on model with finite data and it is claimed that the resulting algorithm is also a version of WLB from Newton and Raftery (1994), however, in this case for us is not clear which randomized functional stands behind this procedure.

Appendix E: Practical interpreptation of slow mixing in MCMC

In practice produced samples by the Markov chain 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 $\Omega \subset \{1, \dots, p\}$. 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:

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

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

$$\begin{aligned}\operatorname{var}(\widehat{f}_{h,N}^t \mid Y^t, t) &= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{s=1}^{N}\operatorname{cov}(h(\lambda_k^t), h(\lambda_s^t) \mid Y^t, t)\\ &\asymp \frac{\sigma^2}{N}(1+2\sum_{k=1}^{\infty}\rho_k^t(h)),\end{aligned} \tag{E.2}$$

where

$$\rho_k^t(h) = \operatorname{corr}(h^T\lambda_1^t, h^T\lambda_{k+1}^t \mid Y^t, t),\ \sigma^2 = \operatorname{var}(h^T\lambda). \tag{E.3}$$

In Liu et al. (1994) it was shown, in particular, that $\rho_k^t(h) \asymp (\gamma^t(h))^k$, so from this and the above formula we get the following expression for the variance of $\widehat{f}_{h,N}^t$ (modulo a universal multiplicative factor):

$$\operatorname{var}(\widehat{f}_{h,N} \mid Y^t, t) \asymp \frac{\sigma^2}{N}\left(\frac{1+\gamma^t(h)}{1-\gamma^t(h)}\right) \approx \frac{\sigma^2}{N}\left(\frac{1+\gamma(h)}{1-\gamma(h)}\right), \tag{E.4}$$

where $\gamma^t(h)$, $\gamma(h)$ are defined in (3.5), (3.8), respectively. The rule of thumb in Aykroyd and Green (1991) 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:

$$\frac{\operatorname{var}(\widehat{f}_{h,N} \mid Y^t, t)}{\sigma^2} < 0.01 \Rightarrow N \gtrsim 100 \times \left(\frac{1+\gamma(h)}{1-\gamma(h)}\right) \to +\infty \text{ for } h = h_m,\ m \gg 1. \tag{E.5}$$

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 $\{h_k\}_{k=1}^p$.

Appendix F: Numerical experiment for the Gibbs-type sampler in ET

λ_* – image of size 64×64 (see Figure F.1),
A – Radon transform matrix of size 4096×4096,
prior $\pi_j = \Gamma(1,1)$,
time $t = 10^2, 10^{10}$ ($\sim$ photons per LOR),
initial point: λ_*,
burn-in samples: 1000,
number of samples for the output: 2000

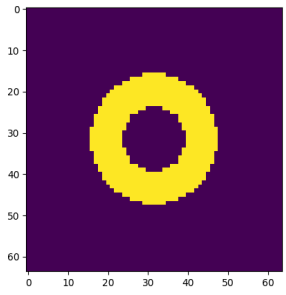


Fig F.1: true distribution λ_*

According to (3.7) we choose $\lambda_* \succ 0$, where $\lambda_*|_{\text{circle}} = 2$, $\lambda_*|_{\text{background}} = 1$ (see Figure F.1), where radius of the inner circle $r_{\text{in}} = 0.25$ and of the outer $r_{\text{out}} = 0.5$, the image corresponds to domain $[-1,1]^2$. Design A is constructed using our implementation of Siddon's algorithm (Siddon (1985)) for parallel beam geometry with 64 projections and 64 parallel lines per projection. Source code in Python of the experiment can be found at https://gitlab.com/eric.barat/npl-pet.

Appendix G: GEM-type algorithm derivation

We mainly follow Wang and Qi (2015) for the derivation of the minimization algorithm based on optimization transfer. Our aim is to build a majoring surrogate of $L_p(\lambda \mid \widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$. Using the fact that $L_p(\lambda \mid \widetilde{\Lambda}_b^t, A, 1, \beta^t/t) = L(\lambda \mid \widetilde{\Lambda}_b^t, A, 1) + \dfrac{\beta^t}{t}\varphi(\lambda)$, we proceed by finding a surrogate for each of both terms in the right hand-side.

G.1. GEM-type algorithm

The attractiveness of Algorithm 5 relies on having an efficient procedure for minimizing $L_p(\lambda \mid \widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$ and $L(\lambda \mid \widetilde{\Lambda}^t, A_{\mathcal{M}}, 1)$. For integer-valued data $Y^t \in \mathbb{N}_0^d$ the $L_p(\lambda \mid Y^t, t)$ coincides with the penalized negative log-likelihood for Poisson-type sample and in this situation, provided penalty $\varphi(\lambda)$ satisfies elementary conditions (convex, C^2 – smooth), fast monotonic GEM algorithms Fessler and Hero (1995), Wang and Qi (2015) can be used.

In our setting intensities $\widetilde{\Lambda}^t$, $\widetilde{\Lambda}_b^t$ 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 Lange et al. (2000) 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 De Pierro (1993) we construct the same majoring surrogate for $L(\lambda \mid \widetilde{\Lambda}_b^t, A, 1)$ as in Fessler and Hero (1995) in a completely algebraic way but now for arbitrary nonnegative term $\widetilde{\Lambda}_b^t$. Further extension to $L_p(\lambda \mid \widetilde{\Lambda}_b^t, A, 1, \beta^t/t)$ is straightforward by considering a separate surrogate for $\varphi(\lambda)$.

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 $\widetilde{\Lambda}_b^t$.

G.2. Majoring surrogate of $L(\lambda \mid \widetilde{\Lambda}_b^t, A, 1)$

In De Pierro (1993) 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 $L(\lambda \mid \widetilde{\Lambda}_b^t, A, 1)$, and denote also $\Lambda_i^{(r)} = a_i^T \lambda^{(r)}$.

Consider the formula

$$\begin{aligned} L(\lambda \mid \widetilde{\Lambda}_b^t, A, 1) &= \sum_{i=1}^{d} f_i(\Lambda_i) \\ &= \sum_{i=1}^{d} f_i \left(\sum_{j=1}^{p} a_{ij} \lambda_j \right) \\ &= \sum_{i=1}^{d} f_i \left(\sum_{j=1}^{p} \left[\frac{a_{ij} \lambda_j^{(r)}}{\Lambda_i^{(r)}} \right] \left[\frac{\lambda_j}{\lambda_j^{(r)}} \Lambda_i^{(r)} \right] \right) \end{aligned}$$

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

$$L(\lambda \mid \widetilde{\Lambda}_b^t, A, 1) \leq Q(\lambda, \lambda^{(r)})$$

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 $Q(\lambda^{(r)},\lambda^{(r)}) = L(\lambda^{(r)} \mid \widetilde{\Lambda}_b^t, A, 1)$. Using the definition of f_i we find that

$$\begin{aligned} Q(\lambda,\lambda^{(r)}) &= \sum_{i=1}^{d}\sum_{j=1}^{p}\left[a_{ij}\lambda_j - \frac{a_{ij}\lambda_j^{(r)}}{\Lambda_i^{(r)}}\widetilde{\Lambda}_{b,i}^t \log\left(\frac{\lambda_j}{\lambda_j^{(r)}}\Lambda_i^{(r)}\right)\right] \\ &= \sum_{j=1}^{p} A_j\left[\lambda_j - \left(\frac{\lambda_j^{(r)}}{A_j}\sum_{i=1}^{d}\frac{a_{ij}\widetilde{\Lambda}_{b,i}^t}{\Lambda_i^{(r)}}\right)\log\lambda_j\right] + const. \end{aligned}$$

where R denotes terms independent of λ.

Function $Q(\lambda,\lambda^{(r)})$ can be rewritten as follows:

$$Q(\lambda,\lambda^{(r)}) \triangleq \sum_{j=1}^{p} A_j\left(\lambda_j - \lambda_{j,L}^{(r+1)}\log\lambda_j\right) \tag{G.1}$$

with

$$\lambda_j^{(r+1),L} \triangleq \frac{\lambda_j^{(r)}}{A_j}\sum_{i=1}^{d}\frac{a_{ij}\widetilde{\Lambda}_{b,i}^t}{\Lambda_i^{(r)}} \tag{G.2}$$

G.3. 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 weights and $\mathcal{N}_j$ is the neighborhood of pixel j.

From Erdogan and Fessler (1999), 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 Wang and Qi (2015) and references therein):

$$\varphi(\lambda) \leq Q_\varphi(\lambda;\lambda^{(r)})$$

where

$$Q_\varphi(\lambda;\lambda^{(r)}) = \frac{1}{2}\sum_{j=1}^{p} p_{j,\varphi}^{(r+1)}(\lambda_j - \lambda_{j,\varphi}^{(r+1)})^2, \tag{G.3}$$

$$p_{j,\varphi}^{(r+1)} = 4\sum_{k\in\mathcal{N}_j} w_{jk}\,\omega_\psi(\lambda_j^{(r)} - \lambda_k^{(r)}), \tag{G.4}$$

$$\lambda_{j,\varphi}^{(r+1)} = \frac{2}{p_{j,\varphi}^{(r+1)}}\sum_{k\in\mathcal{N}_j} w_{jk}\,\omega_\psi(\lambda_j^{(r)} - \lambda_k^{(r)})(\lambda_j^{(r)} + \lambda_k^{(r)}). \tag{G.5}$$

G.4. 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

$$\lambda_j^{(r+1)} = \frac{2\lambda_{j,L}^{(r+1)}}{\sqrt{(b_j^{(r+1)})^2 + 4\beta_j^{(r+1)}\lambda_{j,L}^{(r+1)}} + b_j^{(r+1)}} \tag{G.6}$$

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 6.

Algorithm 6: $\arg\min_{\lambda \succeq 0} L_p(\lambda \mid \widetilde{\Lambda}_b^t, A, 1, \frac{\beta^t}{t})$ by optimization transfer

Data: 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 (G.2);
        compute λ_{j,φ}^{(r+1)} using formula (G.5);
        compute λ_j^{(r+1)} using formula (G.6);
    end
end
```

Output: $\lambda^{(R)}$

Remark G.1. *By setting* $\frac{\beta^t}{t} \to 0$ *in* (G.6)*, we immediately check that* $\lambda^{(r+1)} \to \lambda_L^{(r+1)}$.

Remark G.2. *Parameter* $\widetilde{\lambda}_{\mathcal{M}}^t$ *in Algorithm 4 is easily obtained by iterating formula* (G.2) *with projector* $A_{\mathcal{M}}$ *and random intensities* $\widetilde{\Lambda}^t$

$$\lambda_{\mathcal{M},s}^{(r+1)} = \frac{\lambda_{\mathcal{M},s}^{(r)}}{A_s^{\mathcal{M}}} \sum_{i=1}^{d} \frac{a_{is}^{\mathcal{M}} \Lambda_i^t}{\Lambda_{\mathcal{M},i}^{(r)}} \tag{G.7}$$

Appendix H: Numerical experiment for the NPL in ET

Source code in Python of the experiment can be found at https://gitlab.com/eric.barat/npl-pet

H.1. Penalty φ

For our numerical tests in Section H we choose the well-known in PET imaging log cosh penalty Green (1990) coupled with ℓ_2 convex pairwise difference penalty:

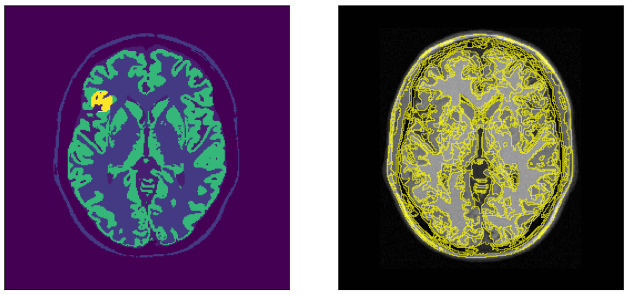
$$\varphi(\lambda) = \sum_{j=1}^{p} \sum_{j' \in \mathcal{N}_j} w_{jj'} \left((1-\nu)\zeta \log\cosh\left(\frac{\lambda_j - \lambda_{j'}}{\zeta}\right) + \frac{\nu}{2}(\lambda_j - \lambda_{j'})^2 \right), \tag{H.1}$$

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 (H.1) is attractive since it bridges together Gaussian prior for pairwise interactions ($\zeta \to +\infty$), and for $\nu = 0$, $\zeta = 0$, it corresponds to pairwise ℓ^1-penalty (Laplace prior). It is easy to check that $\varphi(\lambda)$ in (H.1) is strictly convex except the only direction given by vector $e = \{c(1,\ldots,1), c \in \mathbb{R}\}$. From formula (2.5) it follows that $e \notin \ker A$, therefore conditions (2.11), (2.12) are automatically satisfied.

H.2. Design

We illustrate Algorithm 5 on synthetic PET data based on a realistic phantom from the BrainWeb database Vunckx et al. (2011). 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; see Figure H.1(a). The anatomical MRI (T1) phantom (Figure H.1(b)) does not contain any information relative to the lesion. For segmentation of MRI-images we used ddCRP Blei and Frazier (2011) with a concentration parameter fixed to 10^{-5} leading to a few hundreds of random segments for a 2D brain slice.



(a) λ_* (b) $M_1 \in \mathcal{M}$

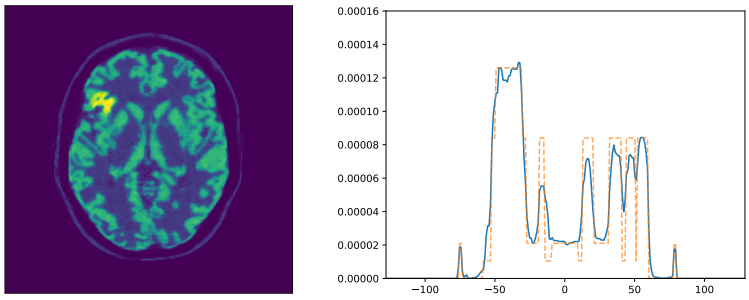
Fig H.1: emission map with lesion hot spot at (a), segmented MRI at (b)

The reconstruction grid was taken 256×256 pixels, i.e., $p = 2^{16}$, being identical to the phantom's one. The observation space consists of LORs derived from a ring of 512 detectors spaced uniformly on a circle. Design A was generated using the Siddon's algorithm Siddon (1985) and $A_{\mathcal{M}}$ was computed from A and segmented image $M_1 \in \mathcal{M}$ using formulas (C.2), (C.3). The intensity λ_* was set so that $\sum_{j=1}^{p} \lambda_{*j} = 5 \cdot 10^5$ 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).

Non-injectivity of A results in the fact that λ_* cannot be reconstructed in principle even from the noiseless sinogram $A\lambda_*$. Result of Theorem 2 in Subsection 5.2 says that the optimal achievable reconstruction (i.e., in presence of infinite amount of data) using the KL-criterion with penalty φ is the following one

$$\lambda_{*opt} = \lambda_* + w_{A,\lambda_*}(0), \tag{H.2}$$

where $w_{A,\lambda_*}(\cdot)$ is defined in (2.14); see Figures H.2 (a), (b) below. Intuitively, $\ker A$ contains only high frequencies, therefore λ_* coincides with λ_{*opt} up to the smallest features on the image (e.g., up to boundaries).



(a) λ_{*opt} (b) profile λ_* vs. λ_{*opt}

Fig H.2: λ_{*opt} at (a), profile through the lesion in λ_{*opt} (blue) vs. λ_* (dotted orange) at (b).

In what follows empirical credible intervals are tested to cover λ_{*opt} but not λ_*. In practice we computed λ_{*opt} as a solution of the following minimization problem

$$\lambda_{*opt} = \arg\min_{\lambda \succeq 0} L_p(\lambda \mid A\lambda_*, A, 1, \beta_{min}), \tag{H.3}$$

where β_{min} was chosen subjectively such that λ_{*opt} does not contain visible numerical artifacts related to the implementation of Siddon's projector. As a result we choose $\beta_{min} = 10^{-3}$. The used minimization algorithm in (H.3) was described in Appendix G.1. For $\varphi(\lambda)$ we use the function from (H.1), where parameters are chosen as follows: $\zeta = 0.05$, $\nu = 0.15$, $\beta^t = 2 \times 10^{-3}$. For $t_1 = 1$, we present results for $\rho = \theta^t/t \in \{0, 0.25, 0.5, 1, 2\}$ (see Remark 7). For $t_2 = 100$ we choose only one value $\rho = 0.05$. For each combination of (t, ρ), Algorithm 5 was generating $B = 1000$ bootstrap draws from which further statistics were computed (empirical mean, standard deviation, etc.).

Finally, the misspecification in the nonparametric prior is mainly due to the fact that the lesion is not reflected in $\mathcal{M}$ and, more generally, to the mismatch between the actual emission map λ_* and the segmentation in $\mathcal{M}$. In this sense our numerical test is the worst-case scenario of using the MRI data in ET.

H.3. Results

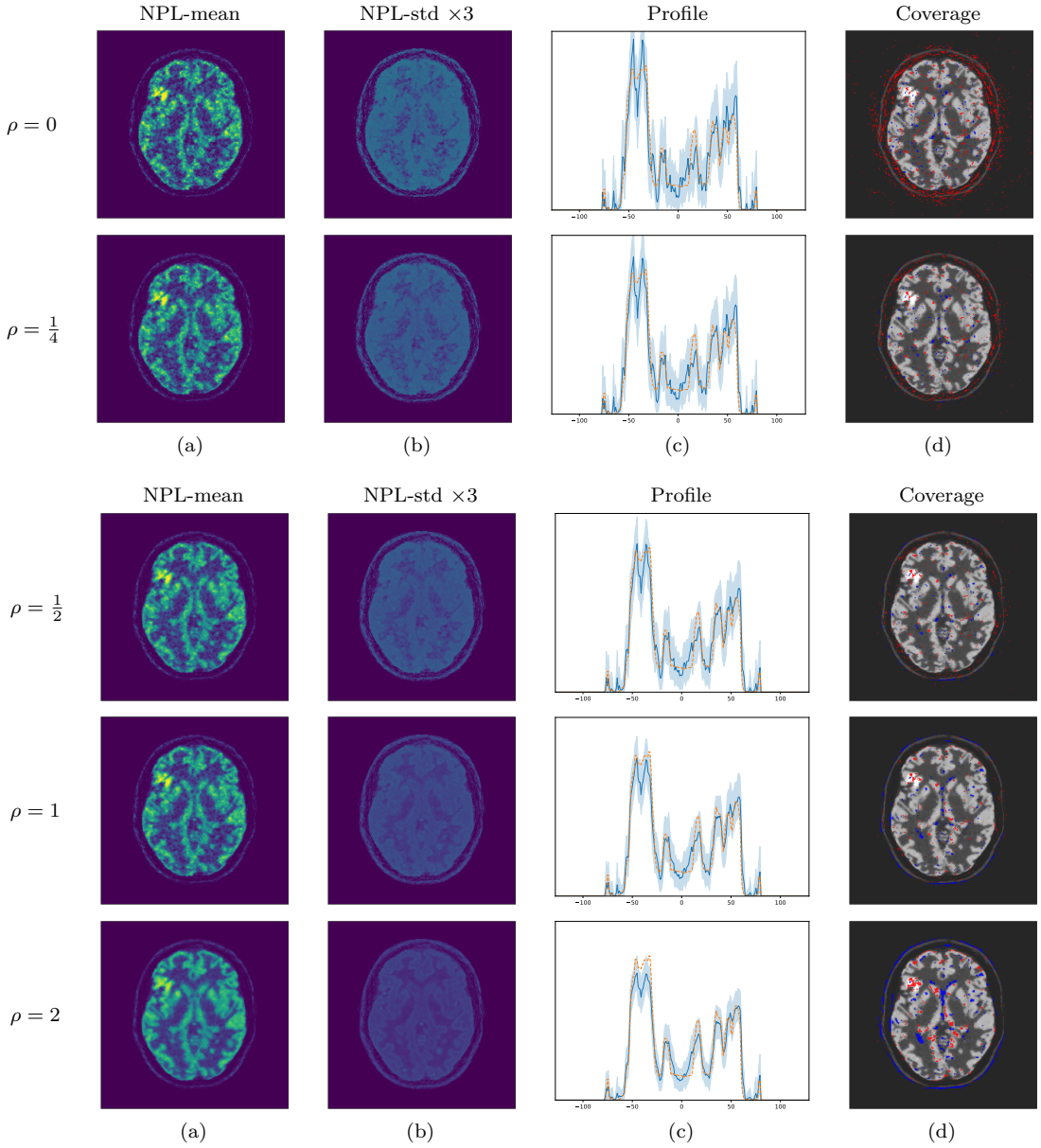


Fig H.3: NPL posterior mean (a), 3× the posterior standard deviation with same color scale as mean (b), posterior 95% band on an horizontal profile through the lesion; blue line – posterior mean, orange dotted – λ_{*opt} (c), pixel-wise coverage success: grayscale – λ_{*opt} is inside the credible band, red, blue colors – λ_{*opt} is above or below, respectively (d).

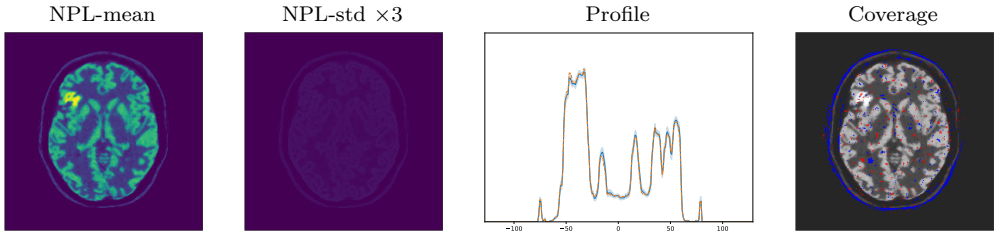


Fig H.4: NPL with $\rho = 0.05$ for $t = 100$

As expected, higher ρ reinforce the effect of MRI on reconstructions and posterior variance decreases with ρ growing ($\text{var}(\widetilde{\lambda}_b^t \mid Y^t, t) \sim (\rho)^{-1}$). As a rule of thumb, it seems reasonable not to exceed $\rho = 1$ since the weight of pseudo-data from the misspecified model in the prior should not exceed the weight of observed data; see Remark 7. We also checked visually that NPL posterior mean with $\rho = 0$ (no MRI) is indistinguishable from the MAP reconstruction with the same penalty tuning (see Section I). This supports the claim in Theorem 5 that the asymptotic distribution is concentrated not around λ_* but a strongly consistent estimator for which we conjecture to coincide asymptotically with the MAP estimate.

In Figure H.3(c) the coverage of λ_{*opt} by pixel-wise 95% credible bands is large almost for all pixels and all values of ρ though the bias in the lesion dominates when $\rho > 1$. This can be explained by the choice of MRI images which do not contain at all of λ_* in the lesion area; see Figures H.1 (a), (b). Visually it seems that $\rho = 1$ is optimal for bias variance trade-off, however, this rule of thumb is reasonable only for moderate value of t (hence, low number of counts in Y^t) and not in the asymptotic regime when $t \to +\infty$. To highlight the latter we also consider the asymptotic behavior of NPL reconstruction by taking $t = 100$ for the regularization parameter $\beta^t/t = 10^{-3}$ (same as for β_{min} in (H.3)) and $\rho = 0.05$ (see Figure H.4). The point is that the case of $t = 100$ corresponds to almost noiseless data, so β^t/t can be chosen in the "optimal way". Pixel-wise posterior bands capture most of the true signal (see Figure H.4(c)), except the blue region at the boundary of the cranium (see Figure H.4(d)). This can be explained by the property of the GEM-algorithm (see Section G.1) where the constructed parabolic majorizing surrogates which approximate poorly zero values at pixels.

Appendix I: Visual comparison between the NPL mean without MRI and the MAP reconstructions

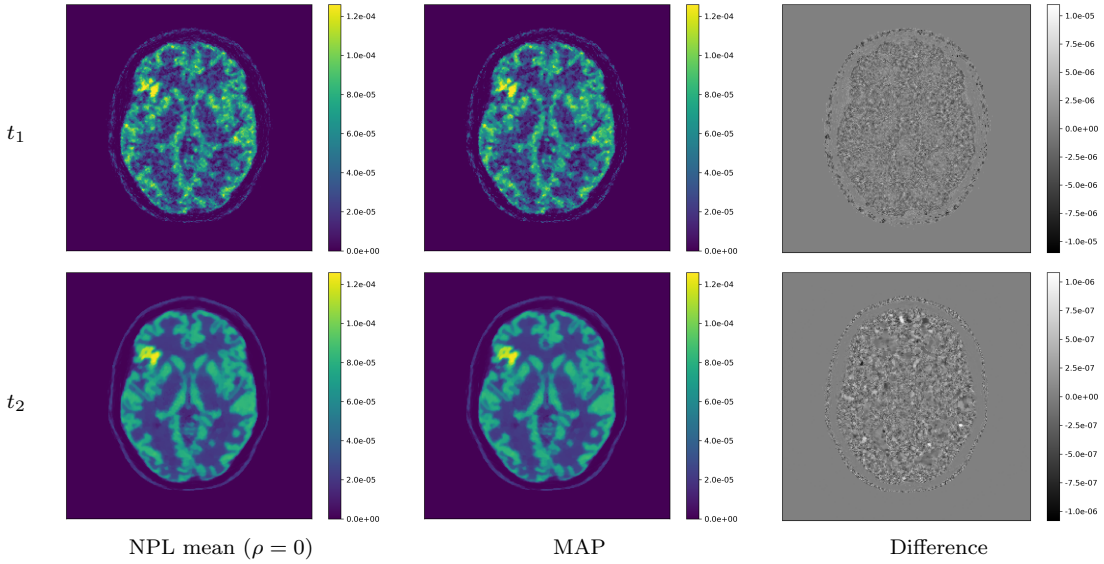


Fig I.1: NPL mean without MRI ($\rho = 0$) compared to MAP reconstruction for $t_1 = 1$ and $t_2 = 100$; $\zeta = 0.05$, $\nu = 0.15$, $\beta^t = 2 \times 10^{-3}$; $B = 8192$ draws for NPL

In Figure I.1, we contrast the empirical mean of NPL without MRI and the MAP estimate with same penalty tuning. For increasing t absolute differences between both images tend to zero (see scales in Figure I.1) which is coherent with the result of Theorem 5 and also supports Conjecture 1 that MAP is the strongly consistent estimator for which (5.22)-(5.24) hold. From the two simulations for $t_1 = 1$ and $t_2 = 100$ one may observe that the empirical contraction rate of absolute differences is of order $t^{-1/2}$. This can be explained by the fact that for regular models with n i.i.d observations (recall that model in (2.1) is regular for pixels intersected by LORs from $I_1(\Lambda^*)$), the next term beyond the normal approximation in the first order Edgeworth's expansion of the posterior decays with rate $n^{-1/2}$ (see Pompe (2021)) which is equivalent to $t^{-1/2}$ in our case.

Appendix J: Remark on centering term of the posterior

Definition J.1. *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:*

$$P(U^t \in A \mid \mathcal{F}^t) \to P(V \in A) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{J.1}$$

This type of convergence will be denoted as follows:

$$U^t \xrightarrow{c.d.} U \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{J.2}$$

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 5 (see lemmas L.8, L.9) we show that

$$\Pi_{\mathcal{U}}(\widetilde{\lambda}_b^t - \widehat{\lambda}_{sc}^t) - u^t(\widetilde{\xi}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty), \tag{J.3}$$

where

$$\widetilde{\xi}^t = (\dots, \sqrt{t}\frac{\widetilde{\Lambda}_{b,i}^t - \widehat{\Lambda}_{sc,i}^t}{\sqrt{\widehat{\Lambda}_{sc,i}^t}}, \dots), \ i \in I_1(\Lambda^*), \tag{J.4}$$

$$u^t(\xi) = \underset{\substack{u:(1-\Pi_{\mathcal{V}})\widehat{\lambda}_{sc}^t + \frac{u}{\sqrt{t}} + w \succeq 0, \\ u \in \mathcal{U},\, w \in \mathcal{W}}}{\arg\min} -u^T (A_{I_1(\Lambda^*)})^T (\widehat{D}_{I_1(\Lambda^*)}^t)^{-1/2} \xi + \frac{1}{2} u^T \widehat{F}_{I_1(\Lambda^*)}^t u, \tag{J.5}$$

$$\widehat{D}_{I_1(\Lambda^*)}^t = \operatorname{diag}(\dots, \widehat{\Lambda}_{sc,i}^t, \dots), \ i \in I_1(\Lambda^*), \tag{J.6}$$

$$\widehat{F}_{I_1(\Lambda^*)}^t = \sum_{i \in I_1(\Lambda^*)} \frac{a_i a_i^T}{\widehat{\Lambda}_{sc,i}^t} = (A_{I_1(\Lambda^*)})^T (\widehat{D}_{I_1(\Lambda^*)}^t)^{-1} A_{I_1(\Lambda^*)}. \tag{J.7}$$

That is the conditional tightness (and also the asymptotic distribution) of $\Pi_{\mathcal{U}}(\widetilde{\lambda}_b^t - \widehat{\lambda}_{sc}^t)$ asymptotically coincides with the one of $u^t(\widetilde{\xi}^t)$ being the minimizer of a quadratic function on a polyhedral set depending on $\widehat{\lambda}_{sc}^t$. In the proof we show that conditional tightness of $u^t(\widetilde{\xi}^t)$ is implied by tightness of $\widetilde{\xi}^t$ (this is especially obvious if the constraints in (J.5) are not active for large t, e.g., when $\lambda_* \succ 0$) and that under the assumptions of the theorem it holds that

$$\begin{gathered}(\dots, \sqrt{t}\frac{\widetilde{\Lambda}_{b,i}^t - \frac{Y_i^t}{t}}{\sqrt{\widehat{\Lambda}_{sc,i}^t}}, \dots) \xrightarrow{c.d.} \mathcal{N}(0, I) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty), \\ I - \text{identity matrix of size } \#I_1(\Lambda^*) \times \#I_1(\Lambda^*).\end{gathered} \tag{J.8}$$

From (J.5)-(J.8) 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 $(\dots, \sqrt{t}\frac{\widehat{\Lambda}_{sc,i}^t - \frac{Y_i^t}{t}}{\sqrt{\widehat{\Lambda}_{sc,i}^t}}, \dots)$, $i \in I_1(\Lambda^*)$. For tightness this term needs to be asymptotically bounded for almost any trajectory Y^t, $t \in (0, +\infty)$, which is exactly asked in (5.23) (in a slightly weakened form).

Now, if we center $\widetilde{\lambda}_b^t$ on λ_* one finds that $\widehat{\lambda}_{sc}^t$ must be replaced everywhere with λ_* in formulas (J.5)-(J.8) and, most importantly, the latter term is now equals $(\dots, \dfrac{Y_i^t - t\Lambda_i^*}{\sqrt{t\Lambda_i^*}}, \dots)$ being asymptotically standard normal (see Section B in Appendix). Therefore, the mean of the asymptotic distribution of $\sqrt{t}\Pi_{\mathcal{U}}(\widetilde{\lambda}_b^t - \lambda_*)$ depends on the trajectory of $(Y_i^t - t\Lambda_i^*)/\sqrt{t\Lambda_i^*}$, $i \in I_1(\Lambda^*)$, 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 $\sqrt{t}\Pi_{\mathcal{U}}(\widetilde{\lambda}_b^t - \lambda_*)$ 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 observed in Theorem 3.3 from Ng and Newton (2020).

Appendix K: 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 $M \in \mathcal{M}$ which is badly aligned with respect to the convex hull of the tracer support. To avoid such situations in practice, we propose to preprocess MRI images before using them in the context of ET which is explained in the end of this section.

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

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

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

$$a_{ij} = \int_{\gamma_i} W(x, \gamma_i)\, \mathbb{1}_j(x)\, dx,\ \gamma_i \in \Gamma,\ j \in \{1, \dots, p\}, \tag{K.2}$$

$$W = W(x, \gamma),\ (x, \gamma) \in \mathbb{R}^2 \times T\mathbb{S}^1,\ 0 < c \le W \le C, \tag{K.3}$$

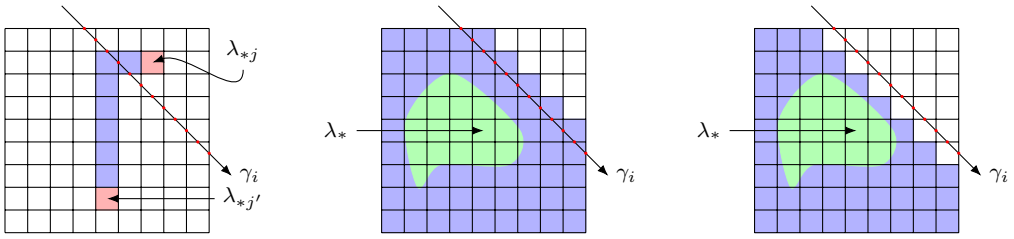
where dx denotes the standard Lebesgue measure on ray γ_i, $\mathbb{1}_j(x)$ is the indicator function of pixel j on the image. Weight $W(x, \gamma)$ is some known sufficiently regular function of spatial coordinates and oriented rays in $\mathbb{R}^2$ which are parameterized by $T\mathbb{S}^1$ (tangent bundle of the unit sphere, see e.g., Natterer (2001)). Projectors defined by the formulas of type (K.2), (K.3) are common in CT and ET practice; see e.g., Siddon (1985), Han et al. (1999). For example, in PET and SPECT weight W is used to model attenuation and nonuniform sensitivity of detectors; see e.g., Quinto (1983), Novikov (2019), Goncharov (2019).

From (C.2), (K.2) it follows that

$$A_{\mathcal{M}} = (a_{M,is}),\ a_{M,is} = \int_{\gamma_i} W(x, \gamma_i) \mathbb{1}_{M,s}(x)\, dx,\ s \in S(M), \tag{K.4}$$

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

Recall that $\lambda_* \in \mathbb{R}_+^p$ is 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 of the signal; two pixels are neighbors if they share an edge (see Figure K.1(a))). 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) $\mathring{\mathrm{DH}}(\lambda_*;\gamma_i)$

Fig K.1

Definition K.1. *Let Γ be the finite family of oriented rays in $\mathbb{R}^2$, A be the projector defined by formulas (K.2), (K.3), $\lambda_* \in \mathbb{R}^p_+$, $\lambda_* \neq 0$ 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*

$$\mathrm{DH}(\lambda_*;\gamma_i) = \{j \in \{1,\dots,p\} \mid \textit{intersection between pixel } j \textit{ and } H(\lambda_*,\gamma_i) \textit{ is of non-zero Lebesgue measure on } \mathbb{R}^2\}. \tag{K.5}$$

Consider

$$\mathring{\mathrm{DH}}(\lambda_*;\gamma_i) = \{j \in \mathrm{DH}(\lambda_*,\gamma_i) \mid \textit{intersection between pixel } j \textit{ and ray } \gamma_i \textit{ is of length zero}\}. \tag{K.6}$$

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

$$\mathrm{DConv}(\lambda_*;\Gamma,A\lambda_*) = \bigcap_{\substack{\gamma_i\in\Gamma,\\ i\in I_0(\Lambda_*)}} D\mathring{H}(\lambda_*;\gamma_i). \tag{K.7}$$

□

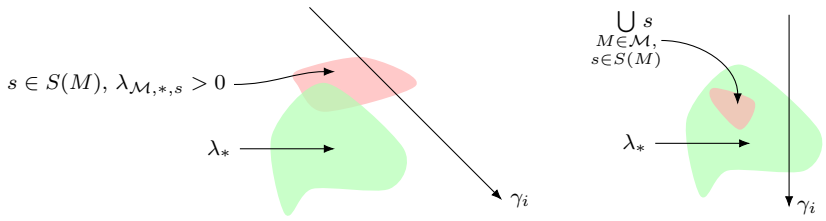
For the geometrical intuition behind definitions $\mathrm{DH}(\cdot)$, $\mathring{\mathrm{DH}}(\cdot)$, $\mathrm{DConv}(\cdot)$, see examples (b), (c) in Figure K.1.

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

$$\text{there exists } i \in I_0(\Lambda^*) \text{ such that } \Lambda^*_{\mathcal{M},i} > 0, \tag{K.8}$$

where $\Lambda^*_{\mathcal{M}}$ is defined in (5.11). From (K.1)-(K.4) and Definition K.1 it follows that in the image for $\lambda_{\mathcal{M},*}$ there is a segment $s \in S(M)$ which intersected by $\gamma_i \in \Gamma$ and such that $\lambda_{\mathcal{M},*,s} > 0$ (see Figure K.2(a)), that is

$$\bigcup_{\substack{M\in\mathcal{M},\\ s\in S(M),\\ \lambda_{\mathcal{M},*,s}>0}} s \not\subset \mathrm{DConv}(\lambda_*;\Gamma,\Lambda^*). \tag{K.9}$$



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

Fig K.2

If we assume that $\lambda_{\mathcal{M},*}$ is also pixel-wise connected, then from (K.9) it follows that

$$\mathrm{DConv}(\lambda_{\mathcal{M},*};\Gamma,A_{\mathcal{M}}\lambda_{\mathcal{M},*}) \not\subset \mathrm{DConv}(\lambda_*;\Gamma,A\lambda_*). \tag{K.10}$$

To conclude, we have just demonstrated the following statement.

Proposition 1. *Let $\lambda_* \in \mathbb{R}^p_+$, $\lambda_* \neq 0$, λ_* is pixel-wise connected and designs A, $A_{\mathcal{M}}$ be of type (K.1)-(K.4). Let $\lambda_{\mathcal{M},*}$ be a solution of the minimization problem in (5.10) and $\lambda_{\mathcal{M},*}$ be also pixel-wise connected. Assume that the non-expansiveness condition (Assumption 3) fails in the sense of (K.8). Then, formula (K.10) holds.*

To avoid the situation in Proposition 1 one may propose to use a significantly smaller segmentation area, for example, such that

$$\bigcup_{\substack{M\in\mathcal{M},\\ s\in S(M)}} s \subsetneq \mathrm{DConv}(\lambda_*;\Gamma,\Lambda^*), \tag{K.11}$$

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 K.2(b).

In view of the latter an ideal choice for $S(M)$ would be such that

$$\mathrm{DConv}(\lambda_{\mathcal{M},*};\Gamma,A_{\mathcal{M}}\lambda_{\mathcal{M},*}) = \mathrm{DConv}(\lambda_*;\Gamma,A\lambda_*). \tag{K.12}$$

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. The term 'mask' is used in practical considerations of ET to denoted restrictions of support of the tracer (e.g., due to medical expertise), so the above procedure theoretically reflects well existing empirical practices.

Appendix L: Proofs

L.1. Proof of Lemma 2.1

Proof. Proof is based on the two following lemmas.

Lemma L.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*

$$S_{A,\lambda}(U) = (\lambda + U + \ker A) \cap \mathbb{R}^p_+ \text{ is convex and compact,} \tag{L.1}$$

where the summation sign denotes the Minkowski sum

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

Lemma L.2. *Let assumptions of Lemma L.1 be satisfied and $d_H(A,B)$ denote the Hausdorff distance between compact sets $A, B \subset \mathbb{R}^p$ being defined by the formula*

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

Let $U \subset \mathrm{Span}(A^T)$ be a compact such that $S_{A,\lambda}(U) \neq \emptyset$. Then,

$$d_H(S_{A,\lambda}(\{u_0\}), S_{A,\lambda}(\{u\})) \to 0 \text{ for } u \to u_0,\ u, u_0 \in U, \tag{L.2}$$

where $S_{A,\lambda}(\cdot)$ is defined in (L.1).

From the result of Lemma L.1 and the assumption in (2.12) it follows that for each $u \in U$ the following problem

$$\begin{aligned} &\text{minimize } \varphi(\lambda + u + w) \text{ w.r.t } w, \\ &\text{subject to: } \lambda + u + w \succeq 0,\ w \in \ker A. \end{aligned} \tag{L.3}$$

admits a unique solution $w(u) \in \ker A$. Indeed, the minimized function in (L.3) is strictly convex function in w and the domain is compact and convex. This proves the first assertion of the lemma.

Now, we prove the continuity of $w(u)$ on its domain. Let u_k be a sequence in U such that $u_k \to u_0$ for some $u_0 \in U$. Let $w_k = w(u_k)$, where the latter are minimizers in (L.3) 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 L.1). Since continuous mapping of a compact is again a compact, all w_k belong to some compact $W_{A,\lambda}(U)$ being the orthogonal projection of $(S_{A,\lambda}(U) - \lambda)$ onto $\ker A$. From compactness of $W_{A,\lambda}(U)$ it follows that w_k contains a converging subsequence $w_m \to w_0$, $w_0 \in W_{A,\lambda}(U)$, where $w_m = w(u_m)$, $m \in \mathbb{N}$.

Since w_m are the minimizers in (L.3), we know that

$$\begin{aligned} &\varphi(\lambda + u_m + w_m) \le \varphi(\lambda + u_m + w), \\ &\text{for all } w \in \ker A, \text{ such that } \lambda + u_m + w \succeq 0. \end{aligned} \tag{L.4}$$

Taking the limit $m \to +\infty$, $u_m \to u_0$, $w_m \to w_0$ we aim to show that

$$\begin{aligned} &\varphi(\lambda + u_0 + w_0) \le \varphi(\lambda + u_0 + w), \\ &\text{for all } w \in \ker A, \text{ such that } \lambda + u_0 + w \succeq 0. \end{aligned} \tag{L.5}$$

Therefore, $w_0 = w(u_0)$ which is unique (by the strict convexity of φ along $\ker A$) and proves the continuity of $w(u)$. The fact that any sequence has a convergent subsequence having the same limit $w(u_0)$ implies that $w_k = w(u_k)$ also converges to $w(u_0)$. However, taking the limit $m \to +\infty$ for each w in (L.4) may not preserve the positivity constraint. To show (L.5), for each w satisfying the positivity constraint in (L.5) we find another sequence $\{w'_m\}$ such that

$$\lambda + u_m + w'_m \succeq 0,\ w'_m \to w \text{ for } m \to +\infty. \tag{L.6}$$

In this case we can replace w with w'_m in (L.4) and take the limit $m \to \infty$ in order to obtain (L.5).

Now, it is left how to choose w'_m so that (L.6) holds. We choose w'_m to be the solution in the following minimization problem

$$\begin{aligned} &\text{minimize } \|(\lambda + u_0 + w) - (\lambda + u_m + w'_m)\| \text{ with respect to } w'_m, \\ &\text{subject to: } w'_m \in \ker A,\ \lambda + u_m + w'_m \succeq 0. \end{aligned} \tag{L.7}$$

Solution w'_m in (L.7) always exists and unique since it corresponds to the euclidean projection of $\lambda + u_0 + w$ onto convex set $S_{A,\lambda}(\{u_m\})$, that is

$$w'_m = \Pi_{\ker A}[\mathrm{Proj}(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\})) - \lambda], \tag{L.8}$$

where $\Pi_{\ker A}$ is the orthogonal projector onto $\ker A$, $\mathrm{Proj}(x, X)$ denotes the euclidean projection of point x onto X. From (L.8) and the fact that $\lambda + u_0 + w \in S_{A,\lambda}(\{u_0\})$ it follows that

$$w'_m - w = \Pi_{\ker A}[\mathrm{Proj}(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\})) - \mathrm{Proj}(\lambda + u_0 + w, S_{A,\lambda}(\{u_0\}))]. \tag{L.9}$$

Using (L.9) and Proposition 5.3 from Attouch and Wets (1993) one can write the following estimate:

$$\|w'_m - w\| \le \rho_m^{1/2} d_{H,\rho_m}(S_{A,\lambda}(\{u_0\}), S_{A,\lambda}(\{u_m\}))^{1/2}, \tag{L.10}$$

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, $d(x, X) = \inf_{x' \in X} d(x, x')$), $d_{H,\rho}(\cdot, \cdot)$ is the bounded Hausdorff distance (see the definition in Section 3 of Attouch and Wets (1993)). In particular, for $d_{H,\rho}$ the following bound holds:

$$d_{H,\rho}(A, B) \le d_H(A, B), \tag{L.11}$$

for any sets A, B.

First, note that $\sup_m \rho_m$ is finite. Indeed, this follows from the fact that $u_m \to u_0$ (hence $\{u_m\}$ is bounded) and following estimates:

$$\begin{aligned} d(\lambda + u_0 + w, S_{A,\lambda}(\{u_m\})) &\le d(\lambda + u_0 + w, 0) + d(0, S_{A,\lambda}(\{u_m\})) \\ &\le \|\lambda + u_0 + w\| + d(0, S_{A,\lambda}(\{u_m\})), \end{aligned} \tag{L.12}$$

$$d(0, S_{A,\lambda_*}(\{u_m\})) \le \max_{j\in\{1,\dots,p\}} \left(\sum_{i=1}^{d} a_i^T(\lambda + u_m) \right) / A_j,\ A_j = \sum_{i=1}^{d} a_{ij}. \tag{L.13}$$

Formula (L.12) is a simple triangle inequality and the estimate in (L.13) follows from the fact that $S_{A,\lambda}(\{u\})$ is the affine subset of $(p-1)$ – simplex defined by the formula

$$\Delta^p_{A,\lambda}(u) = \{\lambda' \in \mathbb{R}^p_+ : \sum_{j=1}^{p} \lambda'_j A_j = \sum_{i=1}^{d} a_i^T(\lambda + u) \ge 0\},\ A_j = \sum_{i=1}^{d} a_{ij} > 0. \tag{L.14}$$

So the inequality in (L.13) express the fact that the furtherst point from the origin to $\Delta^p_{A,\lambda}$ is one of its vertices. From (L.10), (L.11) the fact that $\sup_m \rho_m < +\infty$ and the result of Lemma L.2 it follows that $w'_m \to w$, where $\lambda + u_m + w_m \succeq 0$. Therefore, conditions in (L.6) are satisfied which, in turn, proves (L.5) and the second claim of the lemma.

Lemma is proved. □

L.2. Proof of Lemma L.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 whereas their 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

$$\lambda + u_k + w_k \in \mathbb{R}^p_+,\ \|\lambda + u_k + w_k\| \to \infty. \tag{L.15}$$

From (L.15) and compactness of U it follows, in particular, that

$$w_k \text{ in } \ker A, \|w_k\| \to +\infty. \tag{L.16}$$

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

$$u_{k_n} \to u_0 \in U \text{ for some } u_0, \text{ as } n \to +\infty. \tag{L.17}$$

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

$$w_{k_n} \in \ker A,\ \|w_{k_n}\| \to +\infty \text{ for } n \to +\infty. \tag{L.18}$$

Let

$$\theta_n = \frac{w_{k_n}}{\|w_{k_n}\|},\ \theta_n \in \mathbb{S}^{p-1} \cap \ker A. \tag{L.19}$$

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

$$\theta_m \to \theta_0,\ \theta_0 \in \mathbb{S}^{p-1} \cap \ker A. \tag{L.20}$$

Let $\{u_m\}_{m=1}^{\infty}$ be the corresponding subsequence of $\{u_{k_n}\}_{n=1}^{\infty}$ for index m in formula (L.20). From (L.15)-(L.20) it follows that we have constructed a sequence $\{(u_m, w_m)\}_{m=1}^{\infty}$ such that

$$\lambda + u_m + w_m \in \mathbb{R}^p_+, u_m \in U,\ w_m \in \ker A, \tag{L.21}$$

$$u_m \to u_0,\ \|w_m\| \to +\infty, \tag{L.22}$$

$$\theta_m = \frac{w_m}{\|w_m\|} \to \theta_0 \in \mathbb{S}^{p-1} \cap \ker A. \tag{L.23}$$

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

$$\lambda + s\theta_0 \in \mathbb{R}^p_+ \text{ for any } s > 0, \tag{L.24}$$

where θ_0 is defined in (L.23).

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

$$\lambda + t(u_m + w_m) = \lambda + t(u_m + \|w_m\|\theta_m) \in \mathbb{R}^p_+ \text{ for any } t \in [0,1]. \tag{L.25}$$

Let $s > 0$. By choosing $t = t_m(s) = s/\|w_m\|$ in (L.25) ($t_m(s) \in [0,1]$ for large m; see (L.22)) and using formulas (L.21)-(L.23) we obtain

$$\begin{aligned}(\lambda + s\theta_0) - (\lambda + t_m(s)u_m + t_m(s)\|w_m\|\theta_m) \\ = s(\theta_0 - \theta_m) - s\frac{u_m}{\|w_m\|} \to 0 \text{ for } m \to +\infty.\end{aligned} \tag{L.26}$$

From (L.26) 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 (L.24) cannot hold, because from (2.5) it follows that

$$\text{for any } \theta \in \ker A,\ \theta \neq 0\, \exists j \in \{1,\dots,p\} \text{ s.t. } \theta_j < 0. \tag{L.27}$$

Since $\theta_0 \in \ker A$, by taking $s > 0$ large enough in formula (L.24), we will arrive to the case when $\lambda + s\theta_0 \notin \mathbb{R}^p_+$, which gives the desired contradiction.

Lemma is proved. □

L.3. Proof of Lemma L.2

Proof. The claim of the lemma makes part of Theorem 1 from Walkup and Wets (1969) which, informally says that a closed convex set $K \subset \mathbb{R}^p$ is a polyhedra iff the Hausdorff distance on the space sections by any family of parallel linear subspaces is Lipschitz continuous with respect to the shift vector.

Using notations from Walkup and Wets (1969) we define the following affine mapping

$$\tau_{A,\lambda}(u) = A\lambda + Au,\ u \in \mathbb{R}^p, \tag{L.28}$$

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

Let $K = \mathbb{R}^p_+$ which is obviously a polyhedra in $\mathbb{R}^p$. Next, we define family of sections of K by the formula

$$k(\Lambda) = \tau^{-1}_{A,\lambda}(\Lambda) \cap K,\ \Lambda \in \mathbb{R}^d. \tag{L.29}$$

Essentially, $k(\Lambda)$ is an section of K by $\ker A$ which is shifted by vector u (in some cases $k(\Lambda)$ can be an empty set). In particular, if $\Lambda = \Lambda(u) = A\lambda + Au$ for some $u \in \mathrm{Span}(A^T)$, then it is easy to see that

$$k(\Lambda(u)) = (\lambda + u + \ker A) \cap K = (\lambda + u + \ker A) \cap \mathbb{R}^p_+ = S_{A,\lambda}(\{u\}), \tag{L.30}$$

where $S_{A,\lambda}$ is defined in (L.1).

The result of Theorem 1 from Walkup and Wets (1969) says, in particular, that

$$d_H(k(\Lambda), k(\Lambda')) \leq C\|\Lambda - \Lambda'\|, \tag{L.31}$$

where C is some constant depending on K and A, $d_H(\cdot,\cdot)$ is the standard Hausdorff distance being also extended for empty sets. However, this extension is not needed for us since we always consider parameters $\Lambda(u)$ for u from some $U \subset \mathrm{Span}(A^T)$ with apriori non-empty sets $S_{A,\lambda}(\{u\})$.

From formulas (L.30), (L.31) it follows that

$$d_H(S_{A,\lambda}(\{u\}), S_{A,\lambda}(\{u'\})) \leq C\|A(u - u')\|, \tag{L.32}$$

which directly implies (L.2).

Lemma is proved. □

L.4. Proof of Theorem 1

Proof. Claim follows directly from the result of Theorem 3.1 from Lo (1982). Indeed, having sample $N_1, \dots, N_n$ of size n from a Poisson point process with intensity ν is equivalent having sample $N_1 + \dots N_n$ of size 1 for intensity $n\nu$. 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 Lo (1982) hold for n being replaced with t.

Theorem is proved. □

L.5. Proofs of theorems 2 and 3

First we prove Theorem 3, then we show that if (5.7) holds conditions in (5.9) for Theorem 3 are satisfied which, in turn, automatically proves Theorem 2.

of Theorem 3. Using (2.9), (2.10), the minimization problem in step 3 in Algorithm 5 can be rewritten as as follows:

$$\begin{aligned} \widetilde{\lambda}_b^t &= \arg\min_{\lambda \succeq 0} L_p(\lambda \mid \widetilde{\Lambda}_b^t, A, 1, \beta^t/t) \\ &= \arg\min_{\lambda \succeq 0} \mathcal{L}^t(\lambda), \end{aligned} \tag{L.33}$$

where

$$\begin{aligned} \mathcal{L}^t(\lambda) = &\sum_{i \in I_1(\Lambda^*)} (-\widetilde{\Lambda}_{b,i}^t + \Lambda_i^*) \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) \\ &+ \sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) + (\Lambda_i - \Lambda_i^*) \\ &+ \sum_{i \in I_0(\Lambda^*)} -\widetilde{\Lambda}_{b,i}^t \log(\Lambda_i) + \Lambda_i + \frac{\beta_t}{t}(\varphi(\lambda) - \varphi(\lambda_*)), \end{aligned} \tag{L.34}$$

where $I_0(\cdot)$, $I_1(\cdot)$ are defined in (2.2) and $\Lambda^* = A\lambda_*$.

Next, for the proof we use the following lemma.

Lemma L.3. *Let $\mathcal{L}^t(\lambda)$ be defined in (L.34) and conditions of Theorem 3 be satisfied. Let $C_{A,\delta}(\lambda')$, $\delta > 0$, $\lambda' \succeq 0$, be the cylinder set defined by the formula*

$$C_{A,\delta}(\lambda') = \{\lambda \in \mathbb{R}_+^p,\ \lambda = \lambda' + \delta u + w \mid (u, w) \in \operatorname{Span}(A^T) \times \ker A, \|u\| = 1\}. \tag{L.35}$$

Then,

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

$$\inf_{\lambda \in C_{A,\delta}(\lambda_*)} \mathcal{L}^t(\lambda) \geq C\delta^2 + o_{cp}(1) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.36}$$

where C is a positive constant independent of δ.

ii) there exists a family of random variables $\widetilde{\lambda}^t \in \mathbb{R}_+^p$, $t \in (0, +\infty)$, such that

$$\widetilde{\lambda}^t \xrightarrow{c.p.} \lambda_* \text{ and } \mathcal{L}^t(\widetilde{\lambda}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.37}$$

From the result of Lemma L.3(i) it follows that for all $\lambda \succeq 0$ at distance δ from λ_* in the $\operatorname{Span}(A^T)$ values of $\mathcal{L}^t(\lambda)$ are greater or equal than $C\delta^2$ with conditional probability tending to one a.s. Y^t, $t \in (0, +\infty)$. At the same time, result of Lemma L.3(ii) says that there is $\widetilde{\lambda}^t \in \mathbb{R}_+^p$ which is arbitrarily close to λ_* and $\mathcal{L}^t(\widetilde{\lambda}^t)$ is converges to zero for $t \to +\infty$ with conditional probability also tending to one. The fact that $\mathcal{L}^t(\lambda)$ is convex together with the above arguments and $\widetilde{\lambda}_b^t$ being the unique minimizer of $\mathcal{L}^t(\lambda)$ imply that

$$P(\|\Pi_{A^T}(\widetilde{\lambda}_b^t - \lambda_*)\| < \delta \mid Y^t, t) \to 1 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.38}$$

where Π_{A^T} is the orthogonal projector onto $\mathrm{Span}(A^T)$. Since δ can be chosen arbitrarily small in Lemma L.3 formula (L.38) implies that

$$\Pi_{A^T}(\widetilde{\lambda}_b^t - \lambda_*) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.39}$$

Vector $\widetilde{\lambda}_b^t$ admits in a unique way the following representation

$$\widetilde{\lambda}_b^t = \lambda_* + \widetilde{u}_b^t + \widetilde{w}_b^t, \text{ where } (\widetilde{u}_b^t, \widetilde{w}_b^t) \in \mathrm{Span}(A^T) \times \ker A. \tag{L.40}$$

Using (L.33), (L.34), (L.40) one can see that

$$\widetilde{w}_b^t = \underset{\substack{w:\lambda_* + \widetilde{u}_b^t + w \succeq 0, \\ w \in \ker A}}{\arg\min} \varphi(\lambda_* + \widetilde{u}_b^t + w) = w_{A,\lambda_*}(\widetilde{u}_b^t), \tag{L.41}$$

where $w_{A,\lambda}(\cdot)$ is defined in (2.14). From (L.41), the fact that $\widetilde{u}_b^t \xrightarrow{c.p.} 0$ (see formulas (L.39), (L.40)), continuity of the map $w_{A,\lambda_*}(\cdot)$ (by the result of Lemma 2.1) and the Continuous Mapping Theorem (see, e.g. Van der Vaart (2000), Theorem 2.3, p. 7) it follows that

$$\widetilde{w}_b^t \xrightarrow{c.p.} w_{A,\lambda_*}(0) \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.42}$$

Formula (5.8) follows directly from (L.39)- (L.42).

Theorem is proved. □

of Theorem 2. To prove the theorem we use the following lemma.

Lemma L.4. *Let $\widetilde{\lambda}_b^t$ be defined as in Algorithm 5 and let $\theta^t/t \to 0$ when $t \to +\infty$. Then,*

$$\widetilde{\Lambda}_{b,i}^t \xrightarrow{c.p.} \Lambda_i^* = a_i^T \lambda_* \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.43}$$

In view of (L.43) in Lemma L.4 all assumptions for Theorem 3 are satisfied, which implies formula (5.8).

Theorem is proved. □

L.6. Proof of Lemma L.3

Proof. First we prove (i), then for (ii) we give an explicit formula for $\widetilde{\lambda}^t$ for which (L.37) holds.

First, in formula (L.34) one can see that

$$\inf_{\lambda \in C_{A,\delta}(\lambda_*)} \sum_{i \in I_1(\Lambda^*)} \left(-\widetilde{\Lambda}_{b,i}^t + \Lambda_i^*\right) \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) \xrightarrow{c.p.} 0, \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.44}$$

The above formula follows from the assumption that $\widetilde{\Lambda}_{b,i}^t \xrightarrow{c.p.} \Lambda_i^*$ and that $\log(\Lambda_i/\Lambda_i^*) = \log(1 + \delta a_i^T u/\Lambda_i^*)$ is uniformly bounded for $\lambda \in C_{A,\delta}(\lambda_*)$ from above and below for δ small enough ($u \in \mathrm{Span}(A^T)$, $\|u\| = 1$). For example, to bound all of the logarithmic terms in (L.44) we may choose any δ such that

$$0 < \delta < \min_{i \in I_1(\Lambda^*)} \left(\Lambda_i^* \|a_i\|^{-1}\right). \tag{L.45}$$

Since $\varphi(\lambda)$ satisfies (2.11), (2.12), there exists a constant $M = M(\lambda_*, \delta, A)$ such that

$$\inf_{\lambda \in C_{A,\delta}(\lambda_*)} \varphi(\lambda) \geq M. \tag{L.46}$$

From (5.7), (L.46) it follows that

$$(\beta^t/t) \inf_{C_{A,\delta}(\lambda_*)} (\varphi(\lambda) - \varphi(\lambda_*)) \geq o(1), \text{ when } t \to +\infty. \tag{L.47}$$

Using (L.34), (L.44), (L.47) we obtain the following estimate

$$\begin{aligned} \inf_{\lambda \in C_{A,\delta}(\Lambda_*)} \mathcal{L}^t(\lambda) \geq o_{cp}(1) + \inf_{\lambda \in C_{A,\delta}(\lambda_*)} \sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \log\left(\frac{\Lambda_i}{\Lambda_i^*}\right) + (\Lambda_i - \Lambda_i^*) \\ + \sum_{i \in I_0(\Lambda^*)} -\widetilde{\Lambda}_{b,i}^t \log(\Lambda_i) + \Lambda_i. \end{aligned} \tag{L.48}$$

Note that

$$-\widetilde{\Lambda}^t_{b,i}\log(\Lambda_i)\geq 0 \text{ for } \Lambda_i\leq 1,\ i\in I_0(\Lambda^*). \tag{L.49}$$

From (2.2), (L.35) it follows that we can choose δ sufficiently small so that

$$\Lambda_i\leq 1 \text{ for all } \lambda\in C_{A,\delta}(\lambda_*), i\in I_0(\Lambda^*). \tag{L.50}$$

For example, it suffices to choose δ as follows

$$0<\delta\leq \min_{i\in\{1,\dots,d\}}(\|a_i\|^{-1}). \tag{L.51}$$

Using (L.48), (L.49), for δ satisfying (L.45), (L.51) we obtain

$$\begin{aligned}\inf_{\lambda\in C_{A,\delta}(\Lambda_*)}\mathcal{L}^t(\lambda)\geq \inf_{\lambda\in C_{A,\delta}(\lambda_*)}&\sum_{i\in I_1(\Lambda^*)}-\Lambda_i^*\log\left(\frac{\Lambda_i}{\Lambda_i^*}\right)+(\Lambda_i-\Lambda_i^*)\\ &+\sum_{i\in I_0(\Lambda^*)}\Lambda_i+o_{cp}(1).\end{aligned} \tag{L.52}$$

Now, consider

$$\Phi_{s^*}(s)=-s^*\log(s)+s,\ s>0,\ s^*>0. \tag{L.53}$$

Function $\Phi_{s^*}(s)$ is convex, smooth, has positive non-vanishing second derivative $\Phi''_{s^*}(s)$ and at $s=s^*$ it has its global minimum. Therefore, for any $\varepsilon>0$ small enough (for example, for $\varepsilon<s^*$) there exists positive constant $C(\varepsilon,s^*)$ such that

$$\Phi_{s^*}(s)-\Phi_{s^*}(s^*)\geq C(\varepsilon,s^*)|s-s^*|^2 \text{ for } |s-s^*|<\varepsilon. \tag{L.54}$$

From (L.54) it follows that one can choose $\delta_0>0$ such that

$$\begin{aligned}&\sum_{i\in I_1(\Lambda^*)}-\Lambda_i^*\log\left(\frac{\Lambda_i}{\Lambda_i^*}\right)+(\Lambda_i-\Lambda_i^*)\\ &+\sum_{i\in I_0(\Lambda^*)}\Lambda_i\geq C(\delta_0,\Lambda^*)\sum_{i\in I_1(\Lambda^*)}(\Lambda_i-\Lambda_i^*)^2+\sum_{i\in I_0(\Lambda^*)}\Lambda_i\\ &\text{for } \mid\Lambda_i-\Lambda_i^*\mid<\delta_0,\ i\in I_1(\Lambda^*).\end{aligned} \tag{L.55}$$

Value for δ_0 is precised below. Let $\lambda\in C_{A,\delta}(\lambda_*)$ and $\delta<\delta_0$, that is $\lambda=\lambda_*+\delta u+w$, where $u\in\mathrm{Span}(A^T)$, $\|u\|=1$, $w\in\ker A$. For δ satisfying (L.51) formula (L.50) holds and we get the following estimate:

$$\Lambda_i=a_i^T\lambda=\delta a_i^Tu\geq\delta^2(a_i^Tu)^2\geq 0 \text{ for } i\in I_0(\Lambda^*). \tag{L.56}$$

In (L.56) we used the fact that $\Lambda_i^*=a_i^T\lambda_*=0$, $i\in I_0(\Lambda^*)$.

From (L.55), (L.56) it follows that

$$\begin{aligned}\inf_{\lambda\in C_{A,\delta}(\lambda_*)}&\sum_{i\in I_1(\Lambda^*)}-\Lambda_i^*\log\left(\frac{\Lambda_i}{\Lambda_i^*}\right)+(\Lambda_i-\Lambda_i^*)+\sum_{i\in I_0(\Lambda^*)}\Lambda_i\\ &\geq\min(C(\delta_0,\Lambda^*),1)\delta^2\sum_{i=1}^d(a_i^Tu)^2\\ &\geq\min(C(\delta_0,\Lambda^*),1)\delta^2\sigma^+_{min}(A^TA),\end{aligned} \tag{L.57}$$

where $\sigma^+_{min}(A^TA)$ is the smallest non-zero eigenvalue of A^TA. In particular, in (L.56), (L.57) we have used the property that $u\in\mathrm{Span}(A^T)$ which guarantees that

$$\sum_{i=1}^d(a_i^Tu)^2=u^TA^TAu\geq\sigma^+_{min}(A^TA)>0 \text{ for } \|u\|=1. \tag{L.58}$$

Formula (L.36) follows directly from (L.52), (L.57).

Finally, we choose δ_0 as follows

$$\delta_0 = \frac{1}{2}\min\left[\min_{i\in\{1,\dots,d\}}(\|a_i\|^{-1}), \min_{i\in I_1(\Lambda^*)}(\Lambda_i^*\|a_i\|^{-1}), \min_{i\in I_1(\Lambda^*)}\Lambda_i^*\right], \tag{L.59}$$

so that conditions (L.45), (L.51) are simultaneously satisfied together with (L.55).

Part (i) of Lemma L.3 is proved. Now we prove part (ii) of the lemma.

Let

$$\widetilde{\lambda}^t = \lambda_* + \sum_{i\in I_0(\Lambda^*)} \widetilde{\Lambda}^t_{b,i}\frac{a_i}{\|a_i\|^2}. \tag{L.60}$$

Note that $\widetilde{\lambda}^t \in \mathbb{R}^p_+$ because $a_i \in \mathbb{R}^p_+$ and $\widetilde{\Lambda}^t_{b,i} \geq 0$. Since $\widetilde{\Lambda}^t_{b,i} \xrightarrow{c.p.} 0$ for $i \in I_0(\Lambda^*)$ (by the assumption) we immediately have that

$$\widetilde{\lambda}^t \xrightarrow{c.p.} \lambda_* \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.61}$$

Note that in (L.34) for $\mathcal{L}^t(\lambda)$ all summands are continuous and equal to zero at $\lambda = \lambda_*$ except the logarithmic part

$$g(\lambda) = \sum_{i\in I_0(\Lambda^*)} -\widetilde{\Lambda}^t_{b,i}\log(\Lambda_i),\ \Lambda_i = a_i^T\lambda. \tag{L.62}$$

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 ($\log(x+y) \geq \log(x)$ for $y \geq 0$) it follows that

$$\begin{aligned} g(\widetilde{\lambda}^t) &= \sum_{i\in I_0(\Lambda^*)} -\widetilde{\Lambda}^t_{b,i}\log\left(a_i^T\widetilde{\lambda}^t\right) \\ &\leq \sum_{i\in I_0(\Lambda^*)} -\widetilde{\Lambda}^t_{b,i}\log\left(\widetilde{\Lambda}^t_{b,i}\right) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty \text{ a.s. } Y^t,\ t \in (0,+\infty). \end{aligned} \tag{L.63}$$

Formula (L.63) gives an asymptotic upper bound on $g(\widetilde{\lambda}^t)$ which is equal to zero. For the lower bound we use formulas (L.49), (L.61) 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

$$\begin{aligned} &g(\widetilde{\lambda}^t) \geq 0 \text{ with conditional probability tending to one for } t \to +\infty, \\ &\text{a.s. } Y^t, t \in (0,+\infty). \end{aligned} \tag{L.64}$$

From (L.63), (L.64) it follows that

$$g(\widetilde{\lambda}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0,+\infty). \tag{L.65}$$

From (L.34), (L.60), (L.62), (L.65) it follows that

$$\mathcal{L}^t(\widetilde{\lambda}^t) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in (0,+\infty). \tag{L.66}$$

This proves part (ii) of the lemma.

Lemma is proved. □

L.7. Proof of Lemma L.4

Proof. Recall that

$$\widetilde{\Lambda}^t_{b,i} \mid Y^t, \widetilde{\Lambda}^t_{\mathcal{M}}, t \sim \Gamma(Y^t_i + \theta^t\widetilde{\Lambda}^t_{\mathcal{M},i}, (\theta^t + t)^{-1}),\ i \in \{1,\dots,d\}, \tag{L.67}$$

where $\widetilde{\Lambda}^t_{\mathcal{M}} \mid Y^t, t$ is sampled in Algorithm 4. From the definition of $\widetilde{\Lambda}^t$ in step 1 of Algorithm 4 and necessary optimality conditions in step 2 (see also analogous formula (L.76)) it follows that

$$\sum_{i=1}^d \widetilde{\Lambda}^t_{\mathcal{M},i} = \sum_{i=1}^d \widetilde{\Lambda}^t_i, \tag{L.68}$$

$$\widetilde{\Lambda}^t_{\mathcal{M}} \succeq 0,\ \widetilde{\Lambda}^t \succeq 0,\ E[\widetilde{\Lambda}^t_i \mid Y^t, t] = Y^t_i/t,\ i \in 1,\dots,d. \tag{L.69}$$

Using (L.68), (L.69) we get the following estimate:

$$E[\widetilde{\Lambda}^t_{\mathcal{M},i} \mid Y^t, t] \le \sum_{i=1}^{d} \frac{Y^t_i}{t},\ i \in \{1,\dots,d\}. \tag{L.70}$$

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

$$\begin{aligned} p(\mid \widetilde{\Lambda}^t_{b,i} - \Lambda^*_i \mid > \varepsilon \mid\ Y^t, t) &\le \frac{E[\mid \widetilde{\Lambda}^t_{b,i} - \Lambda^*_i \mid\mid Y^t, t]}{\varepsilon} \\ &\le \frac{E[\mid \widetilde{\Lambda}^t_{b,i} - \frac{Y^t_i + \theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t + t} \mid\mid Y^t, t]}{\varepsilon} + \frac{E[\mid \frac{Y^t_i + \theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t + t} - \Lambda^*_i \mid\mid Y^t, t]}{\varepsilon}. \end{aligned} \tag{L.71}$$

Using the Jensen's inequality $E|X|^2 \ge (E|X|)^2$, formulas (L.67), (L.70), the Strong Law of Large Numbers for Y^t (see Theorem B.1(i) in Section B) and the fact that $\theta^t/t \to 0$, we get the following:

$$\begin{aligned} E[\mid \widetilde{\Lambda}^t_{b,i} - \frac{Y^t_i + \theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t + t} \mid\mid Y^t, t] &\le \left(E[\mid \widetilde{\Lambda}^t_{b,i} - \frac{Y^t_i + \theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t + t} \mid^2 \mid Y^t, t] \right)^{1/2} \\ &= \left(E[\mathrm{var}[(\widetilde{\Lambda}^t_{b,i}) \mid Y^t, \widetilde{\Lambda}^t_{\mathcal{M}}, t] \mid Y^t, t] \right)^{1/2} \\ &= \left(\frac{Y^t_i + \theta^t E[\widetilde{\Lambda}^t_{\mathcal{M},i} \mid Y^t, t]}{(t + \theta^t)^2} \right)^{1/2} \\ &\le \left(\frac{Y^t_i + (\theta^t/t) \sum_{i=1}^d Y^t_i}{(t + \theta^t)^2} \right)^{1/2} \to 0 \text{ a.s. } Y^t,\ t \in (0, +\infty). \end{aligned} \tag{L.72}$$

For estimation of the second term in (L.71) we use formula (L.70), the triangle inequality and again the property that $\theta^t/t \to 0$ to get the following:

$$\begin{aligned} E\left[\left| \frac{Y^t_i + \theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t + t} - \Lambda^*_i \right| Y^t, t \right] &\le \left| \frac{Y^t_i}{\theta^t + t} - \Lambda^*_i \right| + E\left[\frac{\theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t + t} \mid Y^t, t \right] \\ &\le \left| \frac{Y^t_i}{\theta^t + t} - \Lambda^*_i \right| + \frac{\theta^t}{\theta^t + t} \sum_{i=1}^{d} \frac{Y^t_i}{t} \to 0 \text{ a.s. } Y^t, t \in (0, +\infty). \end{aligned} \tag{L.73}$$

Formula (L.43) follows from formulas (L.71)-(L.73).

Lemma is proved. □

L.8. Proof of Proposition 1

Proof. First prove that the set of minimizers in (5.10) is always nonempty and is a subset of the simplex in (5.12). From the Karush-Kuhn-Tucker optimality conditions (see e.g., Bertsekas (1997), Section 3.3) it follows that

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

$$\sum_{i \in I_1(\Lambda^*)} -\Lambda^*_i \frac{a_{\mathcal{M},ij}}{\Lambda^*_{\mathcal{M},i}} + \sum_{i=1}^{d} a_{\mathcal{M},ij} - \mu_{\mathcal{M},*,j} = 0, \tag{L.74}$$

$$\mu_{\mathcal{M},*,j} \lambda_{\mathcal{M},*,j} \equiv 0, \text{ for all } j \in \{1,\dots,p_{\mathcal{M}}\}. \tag{L.75}$$

By multiplying both sides of (L.74) on $\lambda_{\mathcal{M},*,j}$, summing up all equations with respect to j and using (L.75) we obtain the following necessary optimality condition:

$$\begin{aligned} \langle \sum_{i=1}^{d} a_{\mathcal{M},i}, \lambda_{\mathcal{M},*} \rangle &= \sum_{j=1}^{p_{\mathcal{M}}} A_{\mathcal{M},j} \lambda_{\mathcal{M},*,j} = \sum_{i=1}^{d} \Lambda^*_i, \\ A_{\mathcal{M},j} &= \sum_{i=1}^{d} a_{\mathcal{M},ij}. \end{aligned} \tag{L.76}$$

Formula (L.76) proves (5.12). The constraint in (L.76) can be added to the set of constraints in (5.10) without any effect since it is necessary. Because the minimized functional in (5.10) is convex and the domain of constraints is now a convex compact there always exists at least one minimizer.

Demonstration of (5.13) is straightforward. 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 $+\infty$ due to explosion of the logarithmic term. At the same time any interior point $\lambda_{\mathcal{M}} \in \Delta_{A_{\mathcal{M}}}^{p_{\mathcal{M}}}(\Lambda^*)$ (i.e., $\lambda_{\mathcal{M}} \succ 0$) would result in the finite value of the target functional. Hence, inclusions (5.13) always hold.

Proposition is proved. □

L.9. Proof of Theorem 4

Proof. First we prove (5.14), then (5.16) which also implies uniqueness of the minimizer.

Let $\lambda_{\mathcal{M},*} \in \mathbb{R}_+^{p_{\mathcal{M}}}$ be a minimizer in (5.10) (possibly not unique; see also Proposition 1).

Let

$$\lambda_{\mathcal{M}} = \lambda_{\mathcal{M},*} + u_{\mathcal{M}},\ \lambda_{\mathcal{M}} \in \mathbb{R}_+^{p_{\mathcal{M}}}. \tag{L.77}$$

Consider the second order Taylor expansion of $L(\lambda \mid \Lambda^*, A_{\mathcal{M}}, 1)$ in (5.10) in a vicinity of $\lambda_{\mathcal{M},*}$:

$$\begin{aligned} L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) \\ = u_{\mathcal{M}}^T \nabla L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) + \frac{1}{2} \sum_{i \in I_1(\Lambda^*)} \Lambda_i^* \frac{(u_{\mathcal{M}}^T a_{\mathcal{M},i})^2}{(\Lambda_{\mathcal{M},i}^*)^2} \\ + o(\|\Pi_{A_{\mathcal{M},I_1(\Lambda^*)}^T} u_{\mathcal{M}}\|^2), \end{aligned} \tag{L.78}$$

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

$$\nabla L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) = \sum_{i \in I_1(\Lambda^*)} -\Lambda_i^* \frac{a_{\mathcal{M},i}}{\Lambda_{\mathcal{M},i}^*} + \sum_{i=1}^{d} a_{\mathcal{M},i}. \tag{L.79}$$

Karush-Kuhn-Tucker necessary optimality conditions for the problem in (5.10) imply that there exists $\mu_{\mathcal{M},*}$ such that

$$\mu_{\mathcal{M},*} \succeq 0,\ \nabla L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) = \mu_{\mathcal{M},*},\ \mu_{\mathcal{M},*,j} \lambda_{\mathcal{M},*,j} = 0,\ j = 1, \ldots, p. \tag{L.80}$$

From formulas (L.77), (L.80) it follows that

$$\begin{aligned} u_{\mathcal{M}}^T \nabla L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}) = u_{\mathcal{M}}^T \mu_{\mathcal{M},*} = (\lambda_{\mathcal{M}} - \lambda_{\mathcal{M},*})^T \mu_{\mathcal{M},*} \\ = \lambda_{\mathcal{M}}^T \mu_{\mathcal{M},*} \geq 0. \end{aligned} \tag{L.81}$$

Note also that $\mu_{\mathcal{M},*}$ is the optimal Lagrangian multiplier for the problem in (5.10) for which the strong duality holds (e.g., by Slater's condition).

Formulas (5.14), (5.15) follow from (L.78)-(L.81). Next, we prove that (5.16) holds.

Using (L.81) we obtain the following estimate:

$$u_{\mathcal{M}}^T \nabla L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) = u_{\mathcal{M}}^T \mu_{\mathcal{M},*} \geq (u_{\mathcal{M}}^T \mu_{\mathcal{M},*})^2 \text{ if } \|u_{\mathcal{M}}\| \leq \|\mu_{\mathcal{M},*}\|^{-1}. \tag{L.82}$$

From (L.78), (L.82) it follows that

$$\begin{aligned} L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) \\ \geq u_{\mathcal{M}}^T C_{\mathcal{M},*} u_{\mathcal{M}} + o(\|u_{\mathcal{M}}\|^2), \\ \text{for } \|u_{\mathcal{M}}\| \leq \|\mu_{\mathcal{M},*}\|^{-1}, \end{aligned} \tag{L.83}$$

where

$$C_{\mathcal{M},*} = \mu_{\mathcal{M},*} \mu_{\mathcal{M},*}^T + \frac{1}{2} \sum_{i \in I_1(\Lambda^*)} \Lambda_i^* \frac{a_{\mathcal{M},i} a_{\mathcal{M},i}^T}{(\Lambda_{\mathcal{M},i}^*)^2}. \tag{L.84}$$

To finish the proof we use two following lemmas.

Lemma L.5. *Let assumptions of Theorem 4 be satisfied. Let*

$$C_\delta = \inf_{\substack{u_\mathcal{M}:\lambda_{\mathcal{M},*}+u_\mathcal{M}\succeq 0,\\ \|u_\mathcal{M}\|=\delta}} u_\mathcal{M}^T C_{\mathcal{M},*} u_\mathcal{M}. \tag{L.85}$$

Then,

$$C_\delta > 0 \text{ for any } \delta > 0. \tag{L.86}$$

Lemma L.6. *Let $\lambda_{\mathcal{M},*} \in \mathbb{R}_+^{p_\mathcal{M}}$. There exists $\delta_* > 0$ such that for any $u_\mathcal{M} \in \mathbb{R}^{p_\mathcal{M}}$, $0 < |u_\mathcal{M}| \le \delta_*$, $\lambda_{\mathcal{M},*} + u_\mathcal{M} \succeq 0$ it also holds that*

$$\lambda_{\mathcal{M},*} + \delta_* \frac{u_\mathcal{M}}{\|u_\mathcal{M}\|} \succeq 0. \tag{L.87}$$

Let δ_* be the one of Lemma L.6 for chosen $\lambda_{\mathcal{M},*}$. From (L.83), (L.84) and the results of Lemmas L.5, L.6, it follows that

$$\begin{aligned} &L(\lambda_\mathcal{M} \mid \Lambda^*, A_\mathcal{M}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_\mathcal{M}, 1) \\ &\qquad \ge \frac{\delta_* u_\mathcal{M}^T}{\|u_\mathcal{M}\|} C_{\mathcal{M},*} \frac{\delta_* u_\mathcal{M}}{\|u_\mathcal{M}\|} \frac{\|u_\mathcal{M}\|^2}{\delta_*^2} + o(\|u_\mathcal{M}\|^2) \\ &\qquad \ge C_{\delta_*} \frac{\|u_\mathcal{M}\|^2}{\delta_*^2} + o(\|u_\mathcal{M}\|^2),\ C_{\delta_*} > 0, \\ &\qquad \text{for } \lambda_\mathcal{M} = \lambda_{\mathcal{M},*} + u_\mathcal{M} \succeq 0,\ |u_\mathcal{M}| \le \min(\delta_*, |\mu_{\mathcal{M},*}|^{-1}). \end{aligned} \tag{L.88}$$

Formula (L.88) proves the claim in (5.16).

Theorem is proved. □

of Lemma L.5. We use the contradiction argument. Assume that it exists $\delta > 0$ such that $C_\delta = 0$, where C_δ is defined in (L.85). Since the infimum in (L.85) is taken over a compact set, there should exist $u_\mathcal{M}$ such that

$$\|u_\mathcal{M}\| = \delta,\ \lambda_{\mathcal{M},*} + u_\mathcal{M} \succeq 0,\ u_\mathcal{M}^T C_{\mathcal{M},*} u_\mathcal{M} = 0. \tag{L.89}$$

Formulas (L.84), (L.89) imply that

$$u_\mathcal{M}^T a_{\mathcal{M},i} = 0,\ i \in I_1(\Lambda^*),\ u_\mathcal{M}^T \mu_{\mathcal{M},*} = 0. \tag{L.90}$$

Using formulas (5.11) in the non-expansiveness condition, (L.79), (L.81), (L.90) we obtain the following:

$$\begin{aligned} u_\mathcal{M}^T \mu_{\mathcal{M},*} &= \sum_{i\in I_0(\Lambda^*)} u_\mathcal{M}^T a_{\mathcal{M},*,i} = \sum_{i\in I_0(\Lambda^*)} (\lambda_{\mathcal{M},i} - \lambda_{\mathcal{M},*,i})^T a_{\mathcal{M},*,i} \\ &= \sum_{i\in I_0(\Lambda^*)} (\Lambda_{\mathcal{M},i} - \Lambda_{\mathcal{M},*,i}) = \sum_{i\in I_0(\Lambda^*)} \Lambda_{\mathcal{M},i} = 0,\ \Lambda_{\mathcal{M},i} = \lambda_\mathcal{M}^T a_{\mathcal{M},i}. \end{aligned} \tag{L.91}$$

From (L.91) and the fact that $\Lambda_\mathcal{M} \succeq 0$ it follows that

$$\Lambda_{\mathcal{M},i} = u_\mathcal{M}^T a_{\mathcal{M},i} = 0,\ i \in I_0(\Lambda^*). \tag{L.92}$$

Putting formulas (L.90), (L.92) together, we arrive to the following:

$$u_\mathcal{M}^T a_{\mathcal{M},i} = 0 \text{ for } i \in \{1, \dots, d\}. \tag{L.93}$$

The injectivity of $A_\mathcal{M}$ and (L.93) imply that $u_\mathcal{M} = 0$ which contradicts the initial assumption that $\|u_\mathcal{M}\| = \delta > 0$.

Lemma is proved. □

of Lemma L.6. We prove the claim by contradiction.

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

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

$$\delta_* = \frac{1}{2} \min\{\lambda_{\mathcal{M},*,j} \mid \lambda_{\mathcal{M},*,j} > 0\},\ \delta_* > 0. \tag{L.94}$$

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

$$0 < \|u_{\mathcal{M}}\| \leq \delta_*,\ \lambda_{\mathcal{M},*} + u_{\mathcal{M}} \succeq 0 \tag{L.95}$$

and assume that

$$\lambda_{\mathcal{M},*} + \delta_* \frac{u_{\mathcal{M}}}{\|u_{\mathcal{M}}\|} \not\succeq 0 \Leftrightarrow \exists j \in \{1, \ldots, p_{\mathcal{M}}\} \text{ such that } \lambda_{\mathcal{M},*,j} + \delta_* \frac{u_{\mathcal{M},j}}{\|u_{\mathcal{M}}\|} < 0. \tag{L.96}$$

From the fact that $\lambda_{\mathcal{M},*} \succeq 0$ and (L.95), (L.96) it follows that

$$\text{for } j \text{ from (L.96) it holds that } \lambda_{\mathcal{M},*,j} > 0,\ u_{\mathcal{M},j} < 0. \tag{L.97}$$

Using (L.94), (L.96), (L.97) we get the following implication:

$$\frac{\delta_*}{\|u_{\mathcal{M}}\|}(-u_{\mathcal{M},j}) > \lambda_{\mathcal{M},*,j} \geq 2\delta_* \Rightarrow (-u_{\mathcal{M},j}) > 2\|u_{\mathcal{M}}\|. \tag{L.98}$$

The inequality in the right hand-side of (L.98) gives the desired contradiction.

Lemma is proved. □

L.10. Proof of Theorem 5

Proof. In what follows we use the following auxiliary result.

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

$$r(t)(\widetilde{\lambda}^t_{\mathcal{M}} - \lambda_{\mathcal{M},*}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \quad a.s.\ Y^t, t \in (0, +\infty), \tag{L.99}$$

where $\lambda_{\mathcal{M},*}$ is from Theorem 4. Note that formula (L.99) also implies

$$r(t)(\widetilde{\Lambda}^t_{\mathcal{M}} - \Lambda^*_{\mathcal{M}}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \quad a.s.\ Y^t, t \in (0, +\infty), \tag{L.100}$$

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

Remark L.1. *The log-factor for $r(t)$ in Theorem L.1 is necessary for the "almost sure" character of formula (L.100) and, in particular, it is due to the Law of the Iterated Logarithm for trajectory Y^t (see Section B). For our purposes it is sufficient to have the result for rate $r(t) = o(\sqrt{t/\log\log t})$ because $\widetilde{\Lambda}^t_{\mathcal{M}}$ is used in the prior whose effect asymptotically disappears in view of the well-known Bernstein von-Mises phenomenon for Bayesian posteriors; see, e.g. Section 10.2 in Van der Vaart (2000).*

The formula for $\widetilde{\lambda}^t_b$ in step 3 of Algorithm 5 can be rewritten as follows:

$$\widetilde{\lambda}^t_b = \arg\min_{\lambda \succeq 0} A^t(\lambda), \tag{L.101}$$

$$\begin{aligned} A^t(\lambda) &= L_p(\lambda \mid t\widetilde{\Lambda}^t_b, A, t, \beta^t) - L_p(\widehat{\lambda}^t_{sc} \mid t\widehat{\Lambda}^t_{sc}, A, t, \beta^t) \\ &= \sum_{i \in I_1(\Lambda^*)} -t(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}) \log\left(\frac{\Lambda_i}{\widehat{\Lambda}^t_{sc,i}}\right) \\ &+ \sum_{i \in I_1(\Lambda^*)} -t\widehat{\Lambda}^t_{sc,i} \log\left(\frac{\Lambda_i}{\widehat{\Lambda}^t_{sc,i}}\right) + t(\Lambda_i - \widehat{\Lambda}^t_{sc,i}) \\ &+ \sum_{i \in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i} \log(t\Lambda_i) + t\Lambda_i \\ &- \left(\sum_{i \in I_0(\Lambda^*)} -t\widehat{\Lambda}^t_{sc,i} \log(t\widehat{\Lambda}^t_{sc,i}) + t\widehat{\Lambda}^t_{sc,i} \right) \\ &+ \beta^t(\varphi(\lambda) - \varphi(\widehat{\lambda}^t_{sc})),\ \widehat{\Lambda}^t_{sc} = A\widehat{\lambda}^t_{sc} \end{aligned} \tag{L.102}$$

where $\widehat{\lambda}^t_{sc}$ is the strongly consistent estimator from (5.22)-(5.24).

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 approximation argument completes the proof.

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

$$\begin{aligned} B^t(\lambda) = &\sum_{i \in I_1(\Lambda^*)} -t(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i}) \frac{\Lambda_i - \widehat{\Lambda}_{sc,i}}{\widehat{\Lambda}_{sc,i}} + t \frac{(\Lambda_i - \widehat{\Lambda}_{sc,i})^2}{2\widehat{\Lambda}_{sc,i}} \\ &+ \sum_{i \in I_0(\Lambda^*)} t\Lambda_i,\ \Lambda_i = a_i^T \lambda. \end{aligned} \tag{L.103}$$

$$\widetilde{\lambda}^t_{b,app} = \arg\min_{\lambda \succeq 0} B^t(\lambda). \tag{L.104}$$

Process $B^t(\lambda)$ is flat in directions from $\ker A$, therefore, though $\widetilde{\lambda}^t_{b,app}$ in (L.104) always exists, it may not be unique, and, in general, $\widetilde{\lambda}^t_{b,app}$ is set-valued. In what follows, if not said otherwise, for $\widetilde{\lambda}^t_{b,app}$ one chooses any point from the set of minimizers (claims will automatically hold for all points in $\widetilde{\lambda}^t_{b,app}$).

It may happen that $a_i^T \widetilde{\lambda}^t_{b,app} = 0$ for some $i \in I_0(\Lambda^*)$, so $A^t(\widetilde{\lambda}^t_{b,app})$, in general, may not be defined due to the presence of logarithmic terms in (L.102). For this reason we approximate $\widetilde{\lambda}^t_{b,app}$ with another auxiliary point $\widetilde{\lambda}^t_{app}$ defined by the formula:

$$\widetilde{\lambda}^t_{app} = \widetilde{\lambda}^t_{b,app} + \sum_{i \in I_0(\Lambda^*)} \widetilde{\Lambda}^t_{b,i} \frac{a_i}{\|a_i\|^2}, \tag{L.105}$$

where $\widetilde{\Lambda}^t_b$ is from step 2 of Algorithm 5. It is easy to check that value $A^t(\widetilde{\lambda}^t_{app})$ is always well-defined (for $x = 0$ we take convention that $x \log x = 0$).

Let $\mathcal{V}$, $\mathcal{U}$ be the subspaces defined in (5.17), (5.18), respectively. From (L.105) and the definition of $\mathcal{V}, \mathcal{U}$ it follows that

$$\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{app} - \widetilde{\lambda}^t_{b,app}) \equiv 0, \tag{L.106}$$

where $\Pi_{\mathcal{U}}$ is defined in (5.18). For the approximation on $\mathcal{V}$ the following result holds.

Lemma L.7. *Let $\mathcal{V}$ be the subspace defined in (5.17), $\Pi_{\mathcal{V}}$ be defined in (5.20). Then,*

$$t\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \widetilde{\lambda}^t_{app}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{L.107}$$

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

$$D^t_{A,\delta}(\lambda) = \{\lambda' \in \mathbb{R}^p_+ : \lambda' = \lambda + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W},\ \|u\|_2 + \|v\|_1 \le \delta\}, \tag{L.108}$$

$$C^t_{A,\delta}(\lambda) = \{\lambda' \in \mathbb{R}^p_+ : \lambda' = \lambda + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W},\ \|u\|_2 + \|v\|_1 = \delta\}, \tag{L.109}$$

where subspaces $\mathcal{V}, \mathcal{U}, \mathcal{W}$ are defined in (5.17)-(5.19), respectively and $\|\cdot\|_2, \|\cdot\|_1$ denote the standard ℓ_2 and ℓ_1-norms in $\mathbb{R}^p$.

The approximation argument for convex process $A^t(\lambda)$ is due to Hjort and Pollard (2011) and is based on the following implication:

$$\widetilde{\lambda}^t_{app} \in \mathrm{int} D^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}),\ \inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})) > 0 \Rightarrow \widetilde{\lambda}^t_b \in D^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{L.110}$$

From (L.106), (L.107) (in Lemma L.7) and (L.108) one can see that for any $\delta > 0$ it holds that

$$P(\widetilde{\lambda}^t_{app} \in \mathrm{int} D^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}) \mid Y^t, t) \to 1 \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.111}$$

In view of this and (L.110), for the approximation it suffices to establish the following result.

Lemma L.8. *Let $A^t(\lambda)$, $B^t(\lambda)$, $\widetilde{\lambda}^t_b$, $\widetilde{\lambda}^t_{b,app}$, $\widetilde{\lambda}^t_{app}$ be defined in (L.102), (L.103), (L.101), (L.104), (L.105), respectively. Then, for any $\delta > 0$ it holds that*

$$P\left(\inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] > 0 \mid Y^t, t\right) \to 1 \text{ for } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{L.112}$$

From (L.110), (L.112) it follows that

$$\sqrt{t}\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_b - \widetilde{\lambda}^t_{b,app}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty), \tag{L.113}$$

$$t\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_b - \widetilde{\lambda}^t_{b,app}) \xrightarrow{c.p.} 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty). \tag{L.114}$$

Let

$$\lambda = \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ u \in \mathcal{U},\ v \in \mathcal{V},\ w \in \mathcal{W}. \tag{L.115}$$

Process $B^t(\cdot)$ defined in (L.103) has the following form in terms of variables u, v (note that $B^t(\cdot)$ is independent of $w \in \mathcal{W}$):

$$B^t(u,v) = \widetilde{B}^t(u,v) + \widetilde{R}^t(u,v), \tag{L.116}$$

$$\widetilde{B}^t(u,v) = \sum_{i \in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{a_i^T u}{\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}} + \sum_{i \in I_0(\Lambda^*)} a_i^T v, \tag{L.117}$$

$$\begin{aligned}\widetilde{R}^t(u,v) = &\sum_{i \in I_1(\Lambda^*)} -(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})a_i^T v + \frac{(a_i^T v)^2}{2\widehat{\Lambda}^t_{sc,i} t} + \frac{(a_i^T u)(a_i^T v)}{\sqrt{t}\widehat{\Lambda}^t_{sc,i}} \\ &+ \sum_{i \in I_0(\Lambda^*)} t\widehat{\Lambda}^t_{sc,i}.\end{aligned} \tag{L.118}$$

Let

$$(\widetilde{u}^t, \widetilde{v}^t) = \underset{\substack{(u,v):\widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w \succeq 0 \\ u \in \mathcal{U},\, v \in \mathcal{V},\, w \in \mathcal{W}}}{\arg\min} \widetilde{B}^t(u,v) \tag{L.119}$$

In particular, from the definition of $\mathcal{V}$ in (5.17) and from (L.115), (L.117), (L.119) it follows that

$$\frac{\widetilde{v}^t_j}{t} = -\widehat{\lambda}^t_{sc,j} \text{ for } j \text{ s.t. } \exists a_{ij} > 0,\ i \in I_0(\Lambda^*) \Leftrightarrow \Pi_{\mathcal{V}}(\widehat{\lambda}^t_{sc} + \frac{\widetilde{v}^t}{t}) = 0. \tag{L.120}$$

Indeed, formulas (5.17), (5.20), (L.117) imply that the choice in (L.120) satisfies the positivity constraint in (L.119) and at the same time minimizes the linear term $\sum_{i \in I_0(\Lambda^*)} a_i^T v$ since all a_{ij} are non-negative.

Lemma L.9. *Let $\widetilde{u}^t_{b,app}$, $\widetilde{v}^t_{b,app}$ be defined by (L.104) for parametrization in (L.115) and $\widetilde{u}^t$, $\widetilde{v}^t$ be defined by (L.119), respectively. Then,*

$$\widetilde{u}^t - \widetilde{u}^t_{b,app} \xrightarrow{c.p.} 0 \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty), \tag{L.121}$$

$$\widetilde{v}^t - \widetilde{v}^t_{b,app} \xrightarrow{c.p.} 0 \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.122}$$

Hence, in view of (L.113), (L.114) and Lemma L.9 it suffices to demonstrate conditional tightness of $(\widetilde{u}^t, \widetilde{v}^t)$.

Statement in (i), that is formula (5.25)), follows from (L.114), (L.120), (L.122) and the assumption in (5.24).

Now we demonstrate (ii). From (L.119), (L.120) it follows that

$$\widetilde{u}^t = \underset{\substack{u:(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + w \succeq 0 \\ u \in \mathcal{U},\, w \in \mathcal{W}}}{\arg\min} \sum_{i \in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{a_i^T u}{\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}}. \tag{L.123}$$

Since the minmized functional in (L.123) is strongly convex in $u \in \mathcal{U}$ and the set of constraints is also convex, the following mapping is well-defined:

$$\widetilde{u}^t(\xi) = \widetilde{u}(\xi, t) \in \mathcal{U},\ \xi \in \mathbb{R}^{\#I_1(\Lambda^*)},\ t \in (0, +\infty), \tag{L.124}$$

$$\widetilde{u}(\xi, t) = \underset{\substack{u:(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0 \\ u\in\mathcal{U},\, w\in\mathcal{W}}}{\arg\min} -\xi^T(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}A_{I_1(\Lambda^*)}u + \frac{1}{2}u^T\widehat{F}^t_{I_1(\Lambda^*)}u, \tag{L.125}$$

where

$$\widehat{D}^t_{I_1(\Lambda^*)} = \operatorname{diag}(\dots, \widehat{\Lambda}^t_{sc,i}, \dots),\ i \in I_1(\Lambda^*), \tag{L.126}$$

$$\widehat{F}^t_{I_1(\Lambda^*)} = \sum_{i\in I_1(\Lambda^*)} \frac{a_i a_i^T}{\widehat{\Lambda}^t_{sc,i}} = A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1}A_{I_1(\Lambda^*)}. \tag{L.127}$$

Note that for $\xi = (\dots, \sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})/\sqrt{\widehat{\Lambda}^t_{sc,i}}, \dots)$, $i \in I_1(\Lambda^*)$, $\widetilde{u}^t(\xi)$ coincides with $\widetilde{u}^t$ from (L.123). In addition, the minimized functional in (L.125) does not depend on $w \in \mathcal{W}$ which in turn affects only the set of constraints.

Lemma L.10. *Let $\widetilde{u}^t(\xi)$ be the mapping defined in (L.124)-(L.127). Then,*

$$\|\widetilde{u}^t(\xi)\| \le \widehat{c}^t \|A^T_{I_1(\Lambda^*)}\widehat{D}^t_{I_1(\Lambda^*)}\xi\|,\ \xi \in \mathbb{R}^{\#I_1(\Lambda^*)}, \tag{L.128}$$

$$\widehat{c}^t = \|(\widehat{F}^t_{I_1(\Lambda^*)})^{-1}\|_{\mathcal{U}} \cdot \|(\widehat{F}^t_{I_1(\Lambda^*)})^{-1/2}\| \left(\|(\widehat{F}^t_{I_1(\Lambda^*)})^{-1}\|_{\mathcal{U}} + 2 \max_{\sigma\in\sigma_{\mathcal{U}}(\widehat{F}^t_{I_1(\Lambda^*)})} \sigma^{-1/2} \right), \tag{L.129}$$

where $\|\cdot\|_{\mathcal{U}}$ denotes the norm of the operator being reduced to subspace $\mathcal{U}$, $\sigma_{\mathcal{U}}(\cdot)$ denotes the spectrum of the self-adjoint operator acting on $\mathcal{U}$. Moreover,

$$\widehat{c}^t \to c^* \text{ for } t \to +\infty,\ \text{a.s. } Y^t, t \in (0, +\infty),\ c^* < +\infty, \tag{L.130}$$

$$c_* = \|(F^*_{I_1(\Lambda^*)})^{-1}\|_{\mathcal{U}} \cdot \|(F^*_{I_1(\Lambda^*)})^{-1/2}\| \left(\|(F^*_{I_1(\Lambda^*)})^{-1}\|_{\mathcal{U}} + 2 \max_{\sigma\in\sigma_{\mathcal{U}}(F^*_{I_1(\Lambda^*)})} \sigma^{-1/2} \right), \tag{L.131}$$

where

$$D_{I_1(\Lambda^*)} = \operatorname{diag}(\dots, \Lambda^*_i, \dots),\ i \in I_1(\Lambda^*), \tag{L.132}$$

$$F^*_{I_1(\Lambda^*)} = \sum_{i\in I_1(\Lambda^*)} \frac{a_i a_i^T}{\Lambda^*_i} = A_{I_1(\Lambda^*)})^T D^{-1}_{I_1(\Lambda^*)}A_{I_1(\Lambda^*)}. \tag{L.133}$$

Lemma L.11. *Let*

$$\widetilde{\xi}^t = (\dots, \sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})/\sqrt{\widehat{\Lambda}^t_{sc,i}}, \dots),\ i \in I_1(\Lambda^*),\ \widetilde{\xi}^t \in \mathbb{R}^{\#I_1(\Lambda^*)}. \tag{L.134}$$

Then, under the assumptions of Theorem 5, family $A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\widetilde{\xi}^t$ is conditionally tight.

The result of Lemma L.11 together with formulas (L.128)-(L.133) imply that $\widetilde{u}^t = \widetilde{u}^t(\widetilde{\xi}^t)$ is conditionally tight almost surely Y^t, $t \in (0, +\infty)$. Statement (ii) of the lemma follows directly from this and formulas (L.113), (L.121) from lemmas L.8, L.9, respectively.

Theorem is proved.

L.11. Proof of Theorem L.1

Proof. Claim in (L.100) directly follows from (L.99) and the Continuous Mapping Theorem, so we prove only (L.99).

Step 2 in Algorithm 4 can be rewritten as follows:

$$\widetilde{\lambda}^t_{\mathcal{M}} = \underset{\lambda_{\mathcal{M}}\succeq 0}{\arg\min}\, L_{\mathcal{M}}(\lambda_{\mathcal{M}} \mid \widetilde{\Lambda}^t), \tag{L.135}$$

$$L_{\mathcal{M}}(\lambda_{\mathcal{M}} \mid \widetilde{\Lambda}^t) = \sum_{i \in I_1(\Lambda^*)} -\log\left(\frac{\Lambda_{\mathcal{M},i}}{\Lambda^*_{\mathcal{M},i}}\right)(\widetilde{\Lambda}^t_i - \Lambda^*_i) + L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1), \tag{L.136}$$

where $\lambda_{\mathcal{M},*}$ is the point from Theorem 4, $\Lambda^*_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M},*}$, and

$$\begin{aligned} &\widetilde{\Lambda}^t_i \sim \Gamma(Y^t_i, t^{-1}),\ i = 1, \dots, d, \text{ are mutually independent,} \\ &E[\widetilde{\Lambda}^t_i \mid Y^t, t] = Y^t_i/t,\ \mathrm{var}[\widetilde{\Lambda}^t_i \mid Y^t, t] = Y^t_i/t^2,\ i \in \{1, \dots, d\}. \end{aligned} \tag{L.137}$$

Note that

$$L_{\mathcal{M}}(\lambda_{\mathcal{M}} \mid \widetilde{\Lambda}^t) \text{ is convex on } \mathbb{R}^{p_{\mathcal{M}}}_+,\ L_{\mathcal{M}}(\lambda_{\mathcal{M},*} \mid \widetilde{\Lambda}^t) = 0. \tag{L.138}$$

For fixed $t > 0$ consider the following parametrization

$$\lambda_{\mathcal{M}} = \lambda_{\mathcal{M},*} + \frac{u_{\mathcal{M}}}{r(t)},\ \lambda_{\mathcal{M}} \in \mathbb{R}^{p_{\mathcal{M}}}_+,\ r(t) = o(\sqrt{t/\log\log t}). \tag{L.139}$$

Let $\delta > 0$. In view of (L.135), (L.138), (L.139) the following implication holds

$$\inf_{\substack{\lambda_{\mathcal{M}}: \|u_{\mathcal{M}}\| = \delta, \\ \lambda_{\mathcal{M}} \succeq 0}} L_{\mathcal{M}}(\lambda_{\mathcal{M}} \mid \widetilde{\Lambda}^t) > 0 \Rightarrow r(t)\|\widetilde{\lambda}^t_{\mathcal{M}} - \lambda_{\mathcal{M},*}\| < \delta. \tag{L.140}$$

Therefore, to prove (L.99) it is sufficient to show that for any small $\delta > 0$ the conditional probability of the event in the left hand-side of (L.140) tends to one for $t \to +\infty$, a.s. Y^t, $t \in (0, +\infty)$.

Let C_*, δ_* be the values of (5.16) from Theorem 4 and let $\|u_{\mathcal{M}}\| = \delta$, $\delta < \delta_*$.

Using (5.16) and (L.136), (L.139) we get the following estimate:

$$\begin{aligned} L(\lambda_{\mathcal{M}} \mid \widetilde{\Lambda}^t) &\geq \sum_{i \in I_1(\Lambda^*)} -\log\left(1 + \frac{u^T_{\mathcal{M}} a_{\mathcal{M},i}}{r(t)\Lambda^*_{\mathcal{M},i}}\right)(\widetilde{\Lambda}^t_i - \Lambda^*_i) + C_*\delta^2/r^2(t) \\ &\geq C_*\delta^2/r^2(t) - \sum_{i \in I_1(\Lambda^*)} \frac{\mid u^T_{\mathcal{M}} a_{\mathcal{M},i} \mid}{r(t)\Lambda^*_{\mathcal{M},i}} \mid \widetilde{\Lambda}^t_i - \Lambda^*_i \mid \\ &= r^{-2}(t)\left(C_*\delta^2 - \sum_{i \in I_1(\Lambda^*)} \frac{\mid u^T_{\mathcal{M}} a_{\mathcal{M},i} \mid}{\Lambda^*_{\mathcal{M},i}} r(t) \mid \widetilde{\Lambda}^t_i - \Lambda^*_i \mid\right) \\ &\geq r^{-2}(t)\left(C_*\delta^2 - \sum_{i \in I_1(\Lambda^*)} \frac{\delta\|a_{\mathcal{M},i}\|}{\Lambda^*_{\mathcal{M},i}} r(t) \mid \widetilde{\Lambda}^t_i - \Lambda^*_i \mid\right). \end{aligned} \tag{L.141}$$

Note that in (L.141) we have used the property that $\log(1 + x) \leq x$, $x \in (-1, +\infty)$.

Estimate in (L.141) implies the left hand-side of (L.140), for example, if

$$r(t) \mid \widetilde{\Lambda}^t_i - \Lambda^*_i \mid \xrightarrow{c.p.} 0 \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty),\ i \in I_1(\Lambda^*). \tag{L.142}$$

To demonstrate (L.142) we use Markov inequality together with (L.137) and arrive to the following estimate

$$\begin{aligned} P(r(t) \mid \widetilde{\Lambda}^t_i - \Lambda^*_i \mid > \varepsilon \mid Y^t, t) &\leq \frac{r^2(t)E(\mid \widetilde{\Lambda}^t_i - \Lambda^*_i \mid^2 \mid Y^t, t)}{\varepsilon^2} \\ &\leq \frac{2r^2(t)E(\mid \widetilde{\Lambda}^t_i - Y^t_i/t \mid^2 \mid Y^t, t) + 2r^2(t) \mid Y^t_i - \Lambda^*_i \mid^2}{\varepsilon^2} \\ &= \frac{2r^2(t)/t^2 + 2 \mid r(t)(Y^t_i/t - \Lambda^*_i) \mid^2}{\varepsilon^2}, \end{aligned} \tag{L.143}$$

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

$$r^2(t)/t^2 \to 0 \text{ and } r(t)(Y^t_i/t - \Lambda^*_i) \to 0 \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.144}$$

Therefore, from (L.143), (L.144) it follows that formula (L.142) holds which together with (L.141) imply (L.140).

Theorem is proved.

□

□

L.12. Proof of Lemma L.7

Proof. To prove the claim is suffices to show that

$$t\widetilde{\Lambda}^t_{b,i} \xrightarrow{c.p.} 0 \text{ for } i \in I_0(\Lambda^*) \text{ for } t \to +\infty, \text{ a.s. } Y^t,\, t \in (0,+\infty). \tag{L.145}$$

Let $\delta > 0$. Using step 2 in Algorithm 5 and Assumption 1 we obtain

$$\begin{aligned} P(t\widetilde{\Lambda}^t_{b,i} > \delta \mid Y^t, t) &= \int_0^{+\infty} P(t\widetilde{\Lambda}^t_{b,i} > \delta \mid \widetilde{\Lambda}^t_{\mathcal{M},i} = \Lambda, Y^t, t)\, P(\widetilde{\Lambda}^t_{\mathcal{M},i} = \Lambda \mid Y^t, t) d\Lambda \\ &\le \int_0^{+\infty} \min\left(\frac{t\theta^t\Lambda}{(\theta^t+t)\delta}, 1\right) P(\widetilde{\Lambda}^t_{\mathcal{M},i} = \Lambda \mid Y^t, t) d\Lambda \\ &\le \int_0^{\frac{(\theta^t+t)\delta}{t\theta^t}} \frac{t\theta^t\Lambda}{(\theta^t+t)\delta} P(\widetilde{\Lambda}^t_{\mathcal{M},i} = \Lambda \mid Y^t, t) d\Lambda + P\left(\frac{t\theta^t\widetilde{\Lambda}^t_{\mathcal{M},i}}{\theta^t+t} > \delta \mid Y^t, t\right). \end{aligned} \tag{L.146}$$

In (L.146) we have used the Markov inequality for $\Lambda^t_b \mid Y^t, t, \widetilde{\Lambda}^t_{\mathcal{M}}$, $i \in I_0(\Lambda^*)$ for which it is known that $\Lambda^t_{b,i} \mid Y^t, t, \widetilde{\Lambda}^t_{\mathcal{M},i} \sim \Gamma(\theta^t\widetilde{\Lambda}^t_{\mathcal{M},i}, (t+\theta^t)^{-1})$.

The last term in (L.146) tends to zero a.s. Y^t, $t \in (0,+\infty)$ due to (L.100) from Theorem L.1.

Next, we show that the first integral in (L.146) 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 (L.146) is rewritten as follows:

$$\begin{aligned} &\int_0^{\frac{(\theta^t+t)\delta}{t\theta^t}} \frac{t\theta^t\Lambda}{(\theta^t+t)\delta} P(\widetilde{\Lambda}^t_{\mathcal{M},i} = \Lambda \mid Y^t, t) d\Lambda = \\ &= \frac{\delta(\theta^t+t)}{t\theta^t} \int_0^1 sP(\theta^t\widetilde{\Lambda}^t_{\mathcal{M},i} = s\delta(t+\theta^t)/t \,|\, Y^t, t)\, ds. \end{aligned} \tag{L.147}$$

Let $0 < \varepsilon < 1$. Then, by splitting the integral in (L.147) we obtain the following estimate:

$$\begin{aligned} \frac{\delta(\theta^t+t)}{t\theta^t} \int_0^1 sP(\theta^t\Lambda^t_{\mathcal{M},i} = s\delta(t+\theta^t)/t \;\mid\; Y^t, t)\, ds &= \int_0^{\varepsilon} \ldots\, ds + \int_{\varepsilon}^1 \ldots\, ds \\ &\le \varepsilon + P(\theta^t\Lambda^t_{\mathcal{M},i} > \varepsilon\delta(t+\theta^t)/t \;\mid\; Y^t, t). \end{aligned} \tag{L.148}$$

For fixed $\varepsilon > 0$, $\delta > 0$, the second term in (L.148) tends to zero for $t \to +\infty$, a.s. Y^t, $t \in (0,+\infty)$, again due to (L.100) from Theorem L.1. Since ε can be arbitrarily small, it follows that the integral in (L.148) is also arbitrarily small for $t \to +\infty$, a.s. Y^t, $t \in (0,+\infty)$. Hence, the integral in (L.147), and most importantly the right hand-side in (L.146) converge to zero when $t \to +\infty$, a.s. Y^t, $t \in (0,+\infty)$. Since initial δ was chosen arbitrarily, this proves the convergence in (L.145).

Lemma is proved. □

L.13. Proof of Lemma L.8

Proof. Let $\delta > 0$. The left hand-side of (L.110) can be estimated as follows:

$$\begin{aligned} \inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] \ge &\inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - B^t(\lambda)] \\ &+ \inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app})] \\ &+ [B^t(\widetilde{\lambda}^t_{b,app}) - B^t(\widetilde{\lambda}^t_{app})] \\ &+ [B^t(\widetilde{\lambda}^t_{app}) - A^t(\widetilde{\lambda}^t_{app})]. \end{aligned} \tag{L.149}$$

We will show that under the assumptions of Theorem 5 the following holds:

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}[A^t(\lambda)-A^t(\widetilde{\lambda}^t_{app})]\geq \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}[B^t(\lambda)-B^t(\widetilde{\lambda}^t_{b,app})]+o_{cp}(1). \tag{L.150}$$

The first term in right hand-side of (L.150) is expected to be positively separated from zero in view of (L.104), (L.109), and in fact, it gives the main contribution for (L.110) to hold. This is described precisely by the following lemma.

Lemma L.12. *Let $B^t(\lambda)$, $\widetilde{\lambda}^t_{b,app}$ be defined in (L.103), (L.104), respectively. Then, the following formulas hold:*

$$\begin{aligned}&B^t(\lambda)-B^t(\widetilde{\lambda}^t_{b,app})=\sum_{i\in I_1(\Lambda^*)}\frac{t(\Lambda_i-\widetilde{\Lambda}^t_{b,app,i})^2}{2\widehat{\Lambda}^t_{sc,i}}+t\langle\widetilde{\mu}^t_{b,app},\lambda\rangle,\\&\lambda\in\mathbb{R}^p_+,\ \widetilde{\Lambda}^t_{b,app}=A\widetilde{\lambda}^t_{b,app},\end{aligned} \tag{L.151}$$

where

$$\widetilde{\mu}^t_{b,app}=\sum_{i\in I_1(\Lambda^*)}\frac{\widetilde{\Lambda}^t_{b,app,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}a_i+\sum_{i\in I_0(\Lambda^*)}a_i, \tag{L.152}$$

$$\widetilde{\mu}^t_{b,app}\in\mathbb{R}^p_+,\ \widetilde{\mu}^t_{b,app,j}\widetilde{\lambda}^t_{b,app,j}=0\ \textit{for all}\ j\in\{1,\dots,p\}. \tag{L.153}$$

We show that (L.150) and the result of Lemma L.12 imply the statement in (L.112).
Let

$$\lambda(u,v,w)=\widetilde{\lambda}^t_{b,app}+\frac{u}{\sqrt{t}}+\frac{v}{t}+w,\ u\in\mathcal{U},\ v\in\mathcal{V},\ w\in\mathcal{W},\ \lambda(u,v,w)\in\mathbb{R}^p_+. \tag{L.154}$$

Using the parametrization from (L.154), the definition of $C^t_{A,\delta}(\cdot)$ in (L.109) and (L.151)-(L.153) from Lemma L.150 we obtain

$$B^t(\lambda)-B^t(\widetilde{\lambda}^t_{b,app})=K^t(u,v,w)+R^t(u,v,w),\ \lambda=\lambda(u,v,w), \tag{L.155}$$

$$\begin{aligned}K^t(u,v,w)&=\sum_{i\in I_1(\Lambda^*)}\frac{(a_i^Tu)^2}{2\widehat{\Lambda}^t_{sc,i}}+t\langle\widetilde{\mu}^t_{b,app},\,\lambda(u,v,w)\rangle\\&=\sum_{i\in I_1(\Lambda^*)}\frac{(a_i^Tu)^2}{2\widehat{\Lambda}^t_{sc,i}}+t\langle\widetilde{\mu}^t_{b,app},\,\lambda(u,v,w)-\widetilde{\lambda}^t_{b,app}\rangle\\&=\sum_{i\in I_1(\Lambda^*)}\frac{(a_i^Tu)^2}{2\widehat{\Lambda}^t_{sc,i}}+t\langle\widetilde{\mu}^t_{b,app},\,\frac{u}{\sqrt{t}}+\frac{v}{t}\rangle\\&=\sum_{i\in I_1(\Lambda^*)}\frac{(a_i^Tu)^2}{2\widehat{\Lambda}^t_{sc,i}}+\sum_{i\in I_1(\Lambda^*)}\sqrt{t}\frac{\widetilde{\Lambda}^t_{b,app,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}a_i^Tu\\&+\sum_{i\in I_1(\Lambda^*)}\frac{\widetilde{\Lambda}^t_{b,app,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}a_i^Tv+\sum_{i\in I_0(\Lambda^*)}a_i^Tv,\end{aligned} \tag{L.156}$$

$$R^t(u,v,w)=\sum_{i\in I_1(\Lambda^*)}\frac{(a_i^Tu)(a_i^Tv)}{\sqrt{t}\widehat{\Lambda}^t_{sc,i}}+\frac{(a_i^Tv)^2}{2t\widehat{\Lambda}^t_{sc,i}}. \tag{L.157}$$

From the fact that $\widehat{\Lambda}^t_{sc,i}\to\Lambda^*_i$ a.s. $Y^t,t\in(0,+\infty)$ ($\widehat{\lambda}^t_{sc}$ is strongly consistent at λ_* on $\mathcal{U}\oplus\mathcal{V}$ by the assumption), the definition of $C^t_{A,\delta}(\cdot)$ in (L.109) and (L.157) it follows that

$$\sup_{\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}|\,R^t(u,v,w)\,|=o_{cp}(1). \tag{L.158}$$

In view of formulas (L.115), (L.120), the results of Lemmas L.9-L.11 and again the fact that $\widehat{\Lambda}^t_{sc,i}\to\Lambda^*_i$, we find that

$$\frac{\widetilde{\Lambda}^t_{b,app,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}=\frac{\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}+\frac{\widehat{\Lambda}^t_{sc,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}=o_{cp}(1),\ i\in I_1(\Lambda^*). \tag{L.159}$$

Formulas (L.155)-(L.159) imply that

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}[B^t(\lambda)-B^t(\widetilde{\lambda}^t_{b,app})]\geq \inf_{\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} K^t(u,v,w)+o_{cp}(1). \tag{L.160}$$

Now, note that if $\lambda(u,v,w)\succeq 0$ (see formula (L.154)), then

$$\lambda(u,0,w)\succeq 0. \tag{L.161}$$

Indeed, from the definition of $\mathcal{V}$, $\mathcal{U}$, $\mathcal{W}$ in (5.17)-(5.19) it follows that u and v have disjoint set of non-zero components, therefore, setting v to zero for $\lambda(u,v,w)$ cannot break the positivity constraint.

From (L.152), (L.153), (L.161) it follows that

$$\langle\widetilde{\mu}^t_{b,app},\lambda(u,0,w)\rangle=\sum_{i\in I_1(\Lambda^*)}\frac{\widetilde{\Lambda}^t_{b,app,i}-\widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}a_i^Tu\geq 0. \tag{L.162}$$

Note also that $K^t(u,v,w)$ in (L.156) does not change when varying $w\in\mathcal{W}$, so, in what follows we write $K^t(u,v)$ instead. Using formulas (L.156), (L.159), (L.162) and the definition of $C^t_{A,\delta}(\cdot)$ in (L.109) we find that

$$K^t(u,v)\geq\sum_{i\in I_1(\Lambda^*)}\frac{(a_i^Tu)^2}{2\widehat{\Lambda}^t_{sc,i}}+\sum_{i\in I_0(\Lambda^*)}a_i^Tv+o_{cp}(1),\ \lambda=\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{L.163}$$

where the term $o_{cp}(1)$ tends to zero uniformly on $C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$ for $t\to+\infty$, a.s. Y^t, $t\in(0,+\infty)$. From (L.163) and strong consistency of $\widehat{\lambda}^t_{sc}$ on $\mathcal{U}\oplus\mathcal{V}$ it follows that

$$K^t(u,v)\geq c_1\|u\|_2^2+c_2\|v\|_1+o_{cp}(1),\ \text{if }\Pi_{\mathcal{V}}\widetilde{\lambda}^t_{b,app}=0,\ \lambda=\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}), \tag{L.164}$$

where c_1, c_2 are some fixed positive constants which depend only on Λ^* and A. The bound above holds for t large enough a.s. Y^t, $t\in(0,+\infty)$.

Recall that

$$\|u\|_2+\|v\|_1=\delta\text{ for }\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{L.165}$$

Using (L.164), (L.165) it is easy to see that

$$K^t(u,v)\geq c\delta^2+o_{cp}(1),\ \text{if }\Pi_{\mathcal{V}}\widetilde{\lambda}^t_{b,app}=0,\ \lambda=\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}), \tag{L.166}$$

for δ small enough (smaller than some universal constant depending on c_1, c_2), where c is some fixed constant also depending on c_1, c_2 from (L.164). Note that the Karush-Kuhn-Tucker optimality conditions in (L.152), (L.153), formula (L.159) and the definition of space $\mathcal{V}$ in (5.17) imply that

$$P(\Pi_{\mathcal{V}}\widetilde{\lambda}^t_{b,app}=0\mid Y^t,t)\to 1\text{ when }t\to+\infty,\ \text{a.s. }Y^t,t\in(0,+\infty). \tag{L.167}$$

Hence, the event in (L.166) is conditioned on $\{\Pi_{\mathcal{V}}\widetilde{\lambda}^t_{b,app}=0\}$ which has asymptotic conditional probability tending to one a.s. Y^t, $t\in(0,+\infty)$, and it also holds

$$K^t(u,v)\geq c\delta^2+o_{cp}(1),\ \lambda=\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}). \tag{L.168}$$

From (L.150), (L.160), (L.168) it follows that

$$P\left(\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})}[A^t(\lambda)-A^t(\widetilde{\lambda}^t_{app})]>0\mid Y^t,t\right)\to 1\text{ for }t\to+\infty,\ \text{a.s. }Y^t,t\in(0,+\infty). \tag{L.169}$$

It is left to demonstrate the initial statement in (L.150). Consider the first term in the left hand-side of (L.149). Using (L.102), (L.103), the definitions in (L.104), (L.109) and the facts that $\Pi_{\mathcal{V}\oplus\mathcal{U}}(\widehat{\lambda}^t_{sc}-$

$\lambda_*) \xrightarrow{a.s.} 0$, $\Pi_{\mathcal{V}\oplus\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \lambda_*) \xrightarrow{c.p.} 0$, and the Taylor expansion of $A(\lambda)$ at $\widehat{\lambda}^t_{sc}$ up to the second order one gets the following estimate

$$
\begin{aligned}
A^t(\lambda) - B^t(\lambda) \geq & \sum_{i\in I_1(\Lambda^*)} -tC_1 \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \frac{\mid \Lambda_i - \widehat{\Lambda}^t_{sc,i} \mid^2}{\mid \widehat{\Lambda}^t_{sc,i} \mid^2} + \sum_{i\in I_1(\Lambda^*)} -tC_2 \mid \Lambda_i - \widehat{\Lambda}^t_{sc,i} \mid^3 \\
& + \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i} \log(t\Lambda_i) + \sum_{i\in I_0(\Lambda^*)} t\widehat{\Lambda}^t_{sc,i} \log(t\widehat{\Lambda}^t_{sc,i}) - t\widehat{\Lambda}^t_{sc,i} \\
& + \beta^t(\varphi(\lambda) - \varphi(\widehat{\lambda}^t_{sc})),\ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})).
\end{aligned} \tag{L.170}
$$

where C_1, C_2 are some positive constants which depend only design A and Λ^*. The above estimate holds with conditional probability tending to one for $t \to +\infty$ a.s. Y^t, $t \in (0, +\infty)$. In particular, in (L.170) to bound uniformly the error-terms in the Taylor's expansion we have used the following estimates:

$$
\sup_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} \mid \Lambda_i - \widehat{\Lambda}^t_{sc,i} \mid / \mid \widehat{\Lambda}^t_{sc,i} \mid = o_{cp}(1),\ i \in I_1(\Lambda^*), \tag{L.171}
$$

$$
\mid \log(1+x) - x \mid \leq C_1 \mid x \mid^2, \text{ for some } C_1 > 0 \text{ for } \mid x \mid \leq 1/2, \tag{L.172}
$$

$$
\mid -\widehat{s}\log(s/\widehat{s}) + (s - \widehat{s}) - \frac{s^2}{2\widehat{s}} \mid \leq C_2 \mid s - \widehat{s} \mid^3, \tag{L.173}
$$

$$
\text{for some } C_2 = C_2(s_*, \varepsilon) > 0 \text{ and } \mid s - \widehat{s} \mid < \widehat{s}/2, \mid \widehat{s} - s_* \mid < \varepsilon \text{ for some fixed } \varepsilon,\ s_* > 0.
$$

Formulas (L.172), (L.173) describe the standard second order Taylor expansions of the logarithm in vicinity of $x = 0$ and $\widehat{s} = s_*$, respectively. Formula (L.171) can be proved via the following triangle-type inequality:

$$
\begin{aligned}
& \mid \Lambda_i - \widehat{\Lambda}^t_{sc,i} \mid \leq \mid \Lambda_i - \widetilde{\Lambda}^t_{b,app,i} \mid + \mid \widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i} \mid + \mid \widetilde{\Lambda}^t_{b,i} - \Lambda^*_i \mid + \mid \Lambda^*_i + \widehat{\Lambda}^t_{sc,i} \mid, \\
& \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}).
\end{aligned} \tag{L.174}
$$

The first term in the right hand-side of (L.174) is of order $o_{cp}(1)$ in view of the definition in (L.109) and the fact that $\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})$ for some fixed $\delta > 0$. The last two terms are also $o_{cp}(1)$ in view of Lemma L.4 and the fact that $\widehat{\Lambda}^t_{sc,i} \to \Lambda^*_i$ a.s. $Y^t, t \in (0, +\infty)$. Finally, from (L.159) 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 (L.174) is also of order $o_{cp}(1)$. This completes the proof of (L.171).

Using the restriction that $\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app,t})$ two first sums in (L.170) can be estimated as follows:

$$
\begin{aligned}
\sum_{i\in I_1(\Lambda^*)} -tC_1 \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \frac{\mid \Lambda_i - \widehat{\Lambda}^t_{sc,i} \mid^2}{|\widehat{\Lambda}^t_{sc,i}|^2} \geq & \sum_{i\in I_1(\Lambda^*)} -tC_1 \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \\
& \times \left(\frac{2 \mid \Lambda_i - \widetilde{\Lambda}^t_{b,app,i} \mid^2}{\mid \widehat{\Lambda}^t_{sc,i} \mid^2} + \frac{2 \mid \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i} \mid^2}{\mid \widehat{\Lambda}^t_{sc,i} \mid^2} \right) \\
\geq & \sum_{i\in I_1(\Lambda^*)} -tC_1 \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \left(\frac{c\delta^2}{t \mid \widehat{\Lambda}^t_{sc,i} \mid^2} + \frac{2 \mid \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i} \mid^2}{\mid \widehat{\Lambda}^t_{sc,i} \mid^2} \right),
\end{aligned} \tag{L.175}
$$

where c depends only A. Using same argument for the second sum in (L.170) we obtain the following:

$$
\begin{aligned}
\sum_{i\in I_1(\Lambda^*)} -tC_2 \mid \Lambda_i - \widehat{\Lambda}^t_{sc,i} \mid^3 \geq & \sum_{i\in I_1(\Lambda^*)} -8tC_2 \left(\mid \Lambda_i - \widetilde{\Lambda}^t_{b,app,i} \mid^3 + \mid \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i} \mid^3 \right) \\
\geq & \sum_{i\in I_1(\Lambda^*)} -8tC_2 \left(\frac{c\delta^3}{t^{3/2}} + \mid \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i} \mid^3 \right),
\end{aligned} \tag{L.176}
$$

for $\lambda \in C^t_{A,\delta}(\widetilde{\Lambda}^t_{b,app,t})$, where c depends only on A.

From (5.24), (L.115), (L.120), the results of lemmas L.9, L.11 it follows that

$$
t \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \cdot \mid \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i} \mid^2 = o_{cp}(1), \tag{L.177}
$$

$$t \mid \widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i} \mid^3 = o_{cp}(1). \tag{L.178}$$

The above formulas imply that sums in (L.175), (L.176) are bounded from below and of order $o_{cp}(1)$.

The logarithmic term in (L.170) can be estimated as follows:

$$\begin{aligned}\sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) &= \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i}) + t\widetilde{\Lambda}^t_{b,app,i}) \\ &\geq \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t \mid \Lambda_i - \widetilde{\Lambda}^t_{b,app,i} \mid + t\widetilde{\Lambda}^t_{b,app,i}) \\ &\geq \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(c\delta + t\widetilde{\Lambda}^t_{b,app,i}),\ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}).\end{aligned} \tag{L.179}$$

where c is some positive constant depending on A. Using (L.115), (L.120) and (L.122) from Lemma L.9 we obtain

$$\begin{aligned}t\widetilde{\Lambda}^t_{b,app,i} &= t(\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i}) + t\widehat{\Lambda}^t_{sc,i} \\ &= a_i^T\widetilde{v}^t_{b,app} + ta_i^T\widehat{\lambda}^t_{sc} \\ &= a_i^T(\widetilde{v}^t_{b,app} - \widetilde{v}^t) = o_{cp}(1),\ I_0(\Lambda^*)\end{aligned} \tag{L.180}$$

Formulas (L.179), (L.180) imply that

$$\sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) \geq \sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(c\delta + o_{cp}(1)). \tag{L.181}$$

By choosing δ smaller than some fixed constant (e.g., $\delta < c/2$) in (L.181) we find that the right hand-side in (L.181) becomes positive with conditional probability tending to one a.s. Y^t, $t \in (0, +\infty)$. Therefore,

$$\sum_{i\in I_0(\Lambda^*)} -t\widetilde{\Lambda}^t_{b,i}\log(t\Lambda_i) \geq o_{cp}(1),\ \lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}) \text{ for } \delta < c/2. \tag{L.182}$$

In addition, from the initial assumption in (5.24) it directly follows that

$$\sum_{i\in I_0(\Lambda^*)} t\widehat{\Lambda}^t_{sc,i}\log(t\widehat{\Lambda}^t_{sc,i}) - t\widehat{\Lambda}^t_{sc,i} = o_{cp}(1). \tag{L.183}$$

Using (L.170), (L.175)-(L.178), (L.182), (L.183) we finally obtain:

$$\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - B^t(\lambda)] \geq o_{cp}(1) + \beta^t \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (\varphi(\lambda) - \varphi(\widehat{\lambda}^t_{sc})). \tag{L.184}$$

Now, let us consider the third term in the left-hand side of (L.149). Using (L.116)-(L.118) we rewrite it as follows:

$$\begin{aligned}B^t(\widetilde{\lambda}^t_{b,app}) - B^t(\widetilde{\lambda}^t_{app}) &= \widetilde{B}^t(\widetilde{\lambda}^t_{b,app}) - \widetilde{B}^t(\widetilde{\lambda}^t_{app}) \\ &+ \widetilde{R}^t(\widetilde{\lambda}^t_{b,app}) - \widetilde{R}^t(\widetilde{\lambda}^t_{app}).\end{aligned} \tag{L.185}$$

From (L.106), the result of Lemma L.7, (L.116)-(L.118), (L.120), the result of lemmas L.9, L.11 and formula (L.185) it follows directly that

$$B^t(\widetilde{\lambda}^t_{b,app}) - B^t(\widetilde{\lambda}^t_{app}) = o_{cp}(1). \tag{L.186}$$

Now we estimate the last term in the right-hand side of (L.149). Using the same argument as in (L.170)-(L.183) one gets the following estimate:

$$
\begin{aligned}
B^t(\widetilde{\lambda}^t_{app}) - A^t(\widetilde{\lambda}^t_{app}) \geq & \sum_{i\in I_1(\Lambda^*)} -tC_1 \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \frac{\mid \widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i} \mid^2}{\mid \widehat{\Lambda}^t_{sc,i} \mid^2} \\
& + \sum_{i\in I_1(\Lambda^*)} -tC_2 \mid \widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i} \mid^3 \\
& + \sum_{i\in I_0(\Lambda^*)} t\widetilde{\Lambda}^t_{b,i} \log(\widetilde{\Lambda}^t_{app,i}) \\
& + \sum_{i\in I_0(\Lambda^*)} -t\widehat{\Lambda}^t_{sc,i} \log(t\widehat{\Lambda}^t_{sc,i}) + t\widehat{\Lambda}^t_{sc,i} \\
& - \beta^t(\varphi(\widetilde{\lambda}^t_{app}) - \varphi(\widehat{\lambda}^t_{sc})).
\end{aligned} \tag{L.187}
$$

$$
\begin{aligned}
& \geq \sum_{i\in I_1(\Lambda^*)} -tC_1 \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \frac{\mid \widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i} \mid^2}{\mid \widehat{\Lambda}^t_{sc,i} \mid^2} \\
& + \sum_{i\in I_1(\Lambda^*)} -tC_2 \mid \widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i} \mid^3 \\
& + \sum_{i\in I_0(\Lambda^*)} t\widetilde{\Lambda}^t_{b,i} \log(t\widetilde{\Lambda}^t_{b,i}) \\
& + \sum_{i\in I_0(\Lambda^*)} -t\widehat{\Lambda}^t_{sc,i} \log(t\widehat{\Lambda}^t_{sc,i}) + t\widehat{\Lambda}^t_{sc,i} \\
& - \beta^t(\varphi(\widetilde{\lambda}^t_{app}) - \varphi(\widehat{\lambda}^t_{sc,i})),
\end{aligned} \tag{L.188}
$$

where constants C_1, C_2 depend only on A. To pass from (L.187) to (L.188) 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 (L.105), (L.107), (L.115), (L.120), the results of lemmas L.9, L.11 it follows that

$$
t \mid \widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i} \mid \cdot \mid \widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i} \mid^2 = o_{cp}(1), \tag{L.189}
$$

$$
t \mid \widetilde{\Lambda}^t_{app,i} - \widehat{\Lambda}^t_{sc,i} \mid^3 = o_{cp}(1). \tag{L.190}
$$

In addition, using (L.145) in the proof of Lemma L.7 we find that

$$
\sum_{i\in I_0(\Lambda^*)} t\widetilde{\Lambda}^t_{b,i} \log(t\widetilde{\Lambda}^t_{b,i}) = o_{cp}(1). \tag{L.191}
$$

Putting together (L.188)-(L.191) and using again (L.183) we obtain

$$
B^t(\widetilde{\lambda}^t_{app}) - A^t(\widetilde{\lambda}^t_{app}) \geq o_{cp}(1) - \beta^t(\varphi(\widetilde{\lambda}^t_{app}) - \varphi(\widehat{\lambda}^t_{sc})). \tag{L.192}
$$

Formulas (L.149), (L.184), (L.186) (L.192) imply that

$$
\begin{aligned}
\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [A^t(\lambda) - A^t(\widetilde{\lambda}^t_{app})] = & \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [B^t(\lambda) - B^t(\widetilde{\lambda}^t_{b,app})] \\
& + \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} \beta^t(\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{app})) \\
& + o_{cp}(1).
\end{aligned} \tag{L.193}
$$

Lemma L.13. *Let β^t, $\varphi(\cdot)$ satisfy the assumptions of Theorem 5 and $\widetilde{\lambda}^t_{b,app}$, $\widetilde{\lambda}^t_{app}$ be defined in (L.104), (L.105), respectively. Then,*

$$
\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} \beta^t(\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{app})) = o_{cp}(1), \ \ a.s.\ Y^t, t \in (0, +\infty). \tag{L.194}
$$

Formula (L.150) directly follows from (L.193) and the result of Lemma L.13.
Lemma is proved. □

L.14. Proof of Lemma L.9

Proof. To prove the claim we use essentially the same convexity argument as before, for example in Lemma L.8.

Let $\delta > 0$ and

$$\widetilde{\lambda}^t = \widehat{\lambda}^t_{sc} + \frac{\widetilde{u}^t}{\sqrt{t}} + \frac{\widetilde{v}^t}{t} + \widetilde{w}^t,\ \widetilde{\lambda}^t \succeq 0, \tag{L.195}$$

$$\lambda(u,v,w) = \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w,\ (u,v,w) \in \mathcal{U}\times\mathcal{V}\times\mathcal{W},\ \lambda(u,v,w) \succeq 0. \tag{L.196}$$

Recall that

$$\|u - \widetilde{u}^t\|_2 + \|v - \widetilde{v}^t\|_1 = \delta \text{ for } \lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t). \tag{L.197}$$

where $C^t_{A,\delta}(\cdot)$ is defined in (L.109).

Next we show that

$$P(\inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} [B^t(u,v) - B^t(\widetilde{u}^t,\widetilde{v}^t)] > 0 \mid Y^t, t) \to 1 \text{ when } t\to+\infty, \text{ a.s. } Y^t, t \in (0,+\infty) \tag{L.198}$$

which together with the fact that δ can be arbitrarily small and convexity of $B^t(u,v)$, implies the claim of the lemma. Using formulas (L.116)-(L.118) we obtain

$$\begin{aligned} B^t(u,v) - B^t(\widetilde{u}^t,\widetilde{v}^t) &= [\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t)] \\ &\quad + [\widetilde{R}^t(u,v) - \widetilde{R}^t(\widetilde{u}^t,\widetilde{v}^t)], \\ (u,v) \text{ s.t. } \exists w\in\mathcal{W},\ &\lambda(u,v,w) \in C^t_{A,\delta}(\widetilde{\lambda}^t). \end{aligned} \tag{L.199}$$

From the facts that $\widetilde{\Lambda}^t_{b,i} \xrightarrow{c.p.} \Lambda^*_i$ (by Lemma L.4), $\widehat{\Lambda}^t_{sc,i} \xrightarrow{a.s.} \Lambda^*_i$ for $i\in\{1,\dots,d\}$ (see (5.23), (5.24) and (B.2) in Appendix B), the conditional tightness of $\widetilde{u}^t$ (by Lemma L.11) and formulas (5.24), (L.120), (L.197) it follows that

$$\sup_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} \mid \widetilde{R}^t(u,v) - \widetilde{R}^t(\widetilde{u}^t,\widetilde{v}^t) \mid = o_{cp}(1), \tag{L.200}$$

where $\widetilde{R}^t(\cdot)$ is defined in (L.118).

Formulas (L.199), (L.200) imply that

$$\inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} [B^t(u,v) - B^t(\widetilde{u}^t,\widetilde{v}^t)] \geq \inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widetilde{\lambda}^t)} [\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t)] + o_{cp}(1). \tag{L.201}$$

Since the positivity constraints in (L.123) 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

$$(\widetilde{u}^t,\widetilde{w}^t) = \operatorname*{arg\,min}_{\substack{(u,w):(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0\\ u\in\mathcal{U},\, w\in\mathcal{W}}} \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{a_i^T u}{\widehat{\Lambda}^t_{sc,i}} + \frac{(a_i^T u)^2}{2\widehat{\Lambda}^t_{sc,i}}. \tag{L.202}$$

Note that minimizer $\widetilde{u}^t$ in (L.202) coincides with the original solution from (L.123). The problem in (L.202) 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., Bertsekas (1997), Section 3.3) for the optimization problem in (L.202) and the strong duality it follows that

$$\exists\, \widetilde{\mu}^t \succeq 0,\ \widetilde{\mu}^t \in \mathcal{W}^{\perp}, \tag{L.203}$$

$$\sum_{i\in I_1(\Lambda^*)} -\sqrt{t}(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{\Pi_{\mathcal{U}} a_i}{\widehat{\Lambda}^t_{sc,i}} + \frac{\Pi_{\mathcal{U}} a_i a_i^T \widetilde{u}^t}{\widehat{\Lambda}^t_{sc,i}} = \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}},\ \widetilde{\mu}^t_{\mathcal{U}} = \Pi_{\mathcal{U}}\widetilde{\mu}^t, \tag{L.204}$$

$$\widetilde{\mu}^t_j \left([(I-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}]_j + \frac{\widetilde{u}^t_j}{\sqrt{t}} + \widetilde{w}^t_j \right) = 0,\ j\in\{1,\dots,p\}, \tag{L.205}$$

where $(\widetilde{u}^t, \widetilde{w}^t)$ are defined in (L.202). 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 (L.202) is strongly convex in u, so $\widetilde{u}^t$ is always unique, whereas at least one $\widetilde{w}^t$ always exists, however, 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 (L.202) so that positivity constraints are satisfied.

From (L.116), (L.203)-(L.205) it follows that

$$\begin{aligned}
\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t) &= \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}\frac{\widetilde{\Lambda}_{b,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i^T(u-\widetilde{u}^t) + \frac{1}{2}\frac{(a_i^T u)^2 - (a_i^T\widetilde{u}^t)^2}{\widehat{\Lambda}^t_{sc,i}} \\
&+ \sum_{i\in I_0(\Lambda^*)} a_i^T(v-\widetilde{v}^t) \\
&= \sum_{i\in I_1(\Lambda^*)} -\sqrt{t}\frac{\widetilde{\Lambda}_{b,i} - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i^T(u-\widetilde{u}^t) + \frac{1}{2}\frac{(a_i^T(u-\widetilde{u}^t))^2}{\widehat{\Lambda}^t_{sc,i}} \\
&+ \frac{(\widetilde{u}^t)^T a_i a_i^T(u-\widetilde{u}^t)}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} a_i^T(v-\widetilde{v}^t) \\
&= \langle \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}}, u-\widetilde{u}^t\rangle + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)} \frac{|\, a_i^T(u-\widetilde{u}^t)\,|^2}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} a_i^T(v-\widetilde{v}^t).
\end{aligned} \tag{L.206}$$

Note that

$$\langle \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}}, u-\widetilde{u}^t\rangle \geq 0, \tag{L.207}$$

$$v - \widetilde{v}^t \succeq 0, \tag{L.208}$$

for $(u,v) \in \mathcal{U}\times\mathcal{V}$ s.t. $\lambda(u,v,w) = \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w \succeq 0$ for some $w \in \mathcal{W}$.

Indeed, in view of (L.203), (L.205) the left hand-side in (L.207) can be rewritten as follows:

$$\begin{aligned}
\langle \frac{\widetilde{\mu}^t_{\mathcal{U}}}{\sqrt{t}}, u-\widetilde{u}^t\rangle &= \langle \widetilde{\mu}^t_{\mathcal{U}}, \frac{u}{\sqrt{t}} - \frac{\widetilde{u}^t}{\sqrt{t}}\rangle \\
&= \langle (I-\Pi_{\mathcal{V}})\widetilde{\mu}^t, (I-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}}\rangle \\
&= \langle (I-\Pi_{\mathcal{V}})\widetilde{\mu}^t, \widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + \frac{v}{t} + w\rangle \\
&= \langle (I-\Pi_{\mathcal{V}})\widetilde{\mu}^t, \lambda(u,v,w)\rangle
\end{aligned} \tag{L.209}$$

Note also that from (L.203) and the definition of $\mathcal{V}$ in (5.17) it follows that

$$\mu^t_{\mathcal{U}} = (I-\Pi_{\mathcal{V}})\mu^t \succeq 0. \tag{L.210}$$

Formula (L.207) follows directly from (L.209), (L.210) and the fact that $\lambda(u,v,w) \succeq 0$.

In turn, formula (L.208) follows from (L.120).

Formulas (L.197), (L.206)-(L.208) 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:

$$\inf_{(u,v,w):\lambda(u,v,w)\in C^t_{A,\delta}(\widehat{\lambda}^t)} [\widetilde{B}^t(u,v) - \widetilde{B}^t(\widetilde{u}^t,\widetilde{v}^t)] \geq c\delta^2, \tag{L.211}$$

where c is some fixed positive constant depending only on Λ^* and A.

Formula (L.198) follows directly from (L.201), (L.211).

Lemma is proved. □

L.15. Proof of Lemma L.10

Let $\xi \in \mathbb{R}^{\#I_1(\Lambda^*)}$ be a parameter and consider $\widetilde{u}^t(\xi)$ defined in (L.125).

Since the positivity constraints in (L.125) 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

$$(\widetilde{u}^t, \widetilde{w}^t) = \underset{\substack{(u,w):(1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}+\frac{u}{\sqrt{t}}+w\succeq 0 \\ u\in\mathcal{U}, w\in\mathcal{W}}}{\arg\min} -\xi^T C^t u + \frac{1}{2}u^T F^t u, \tag{L.212}$$

where

$$\begin{aligned} &C^t = (\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2} A_{I_1(\Lambda^*)},\ F^t = \widehat{F}^t_{I_1(\Lambda^*)}, \\ &\widehat{D}^t_{I_1(\Lambda^*)},\ \widehat{F}^t_{I_1(\Lambda^*)} \text{ are defined in (L.126), (L.127).} \end{aligned} \tag{L.213}$$

The Lagrangian function for the primal problem in (L.212) is defined by the formula:

$$\mathcal{L}^t(u, w; \mu) = -\xi^T C^t u + \frac{1}{2}u^T F^t u - \mu^T((1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u}{\sqrt{t}} + w), \tag{L.214}$$

$$u \in \mathcal{U},\ w \in \mathcal{W},\ \mu \succeq 0. \tag{L.215}$$

The dual function for $G^t(\mu)$ and solution μ^t for the dual problem are defined by the formulas:

$$G^t(\mu) = \inf_{u\in\mathcal{U},\, w\in\mathcal{W}} \mathcal{L}^t(u, w; \mu),\ \mu^t = \arg\max_{\mu\succeq 0} G^t(\mu). \tag{L.216}$$

From the Karush-Kuhn-Tucker necessary optimality conditions, the fact that the primal problem is strongly convex in $u \in \mathcal{U}$ and the strong duality it follows that

$$\exists (u^t, w^t) \in \mathcal{U}\times\mathcal{W},\ \mu^t \succeq 0,\ \mu^t \in \mathcal{W}^{\perp} \text{ s.t.} \tag{L.217}$$

$$(u^t, w^t) \text{ is a solution for the primal problem in (L.212),} \tag{L.218}$$

$$\mu^t = \mu^t(\xi) \text{ is a solution for the dual problem in (L.216),} \tag{L.219}$$

$$\nabla_{u,w}\mathcal{L}^t(u^t, w^t; \mu^t) = 0, \tag{L.220}$$

$$((1-\Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc,j} + \frac{u^t_j}{\sqrt{t}} + w^t_j)\mu^t_j = 0,\ j \in \{1, \dots, p\}. \tag{L.221}$$

Using (L.214), (L.220) we obtain the following:

$$-\Pi_{\mathcal{U}}(C^t)^T\xi + (\Pi_{\mathcal{U}}F^t\Pi_{\mathcal{U}})u^t - \frac{\Pi_{\mathcal{U}}\mu^t(\xi)}{\sqrt{t}} = 0, \tag{L.222}$$

$$\Pi_{\mathcal{W}}\mu^t = 0, \tag{L.223}$$

where $\Pi_{\mathcal{U}}$, $\Pi_{\mathcal{W}}$ are defined in (5.20). In what follows we use the following notations

$$C^t_{\mathcal{U}} = C^t\Pi_{\mathcal{U}},\ F^t_{\mathcal{U}} = (\Pi_{\mathcal{U}}F^t\Pi_{\mathcal{U}}),\ \mu^t_{\mathcal{U}} = \Pi_{\mathcal{U}}\mu^t. \tag{L.224}$$

Strong consistency of $\widehat{\lambda}^t_{sc}$ on $\mathcal{U}\oplus\mathcal{V}$ and the Continuous Mapping Theorem imply that

$$C^t_{\mathcal{U}} \to C^*_{\mathcal{U}},\ F^t_{\mathcal{U}} \to F^*_{\mathcal{U}} \text{ when } t \to +\infty,\ \text{ a.s. } Y^t,\ t \in (0, +\infty), \tag{L.225}$$

where

$$C^*_{\mathcal{U}} = \Pi_{\mathcal{U}}C^*,\ F^*_{\mathcal{U}} = \Pi_{\mathcal{U}}F^*\Pi_{\mathcal{U}}, \tag{L.226}$$

$$C^* = (D^*_{I_1(\Lambda^*)})^{-1/2}A_{I_1(\Lambda^*)},\ D^*_{I_1(\Lambda^*)} = \operatorname{diag}(\dots, \Lambda^*_i, \dots),\ i \in I_1(\Lambda^*), \tag{L.227}$$

$$F^* = \sum_{i\in I_1(\Lambda^*)} \frac{a_i a_i^T}{\Lambda^*_i} = A_{I_1(\Lambda^*)})^T(D^*_{I_1(\Lambda^*)})^{-1}A_{I_1(\Lambda^*)}. \tag{L.228}$$

Using the notations from (L.224) formula (L.222) can be rewritten as follows:

$$u^t(\xi) = (F^t_{\mathcal{U}})^{-1}(C^t_{\mathcal{U}})^T\xi + (F^t_{\mathcal{U}})^{-1}\frac{\mu^t_{\mathcal{U}}(\xi)}{\sqrt{t}}. \tag{L.229}$$

Note that $F^t_{\mathcal{U}}$ is continuously invertible on $\mathcal{U}$, therefore $(F^t_{\mathcal{U}})^{-1}$ is well-defined. Moreover, $(F^t_{\mathcal{U}})^{-1} \to (F^*_{\mathcal{U}})^{-1}$ for $t \to +\infty$ a.s. Y^t, $t \in (0, +\infty)$. Next, we show that the following estimate always holds:

$$\left|\frac{\mu^t_{\mathcal{U}}(\xi)}{\sqrt{t}}\right| \leq 2 \max_{\sigma \in \sigma_{\mathcal{U}}(F^t_{\mathcal{U}})} \sigma^{-1/2}\|(F^t_{\mathcal{U}})^{-1/2}\|\|(C^t_{\mathcal{U}})^T\xi\|, \tag{L.230}$$

where $\sigma_{\mathcal{U}}(F^t_{\mathcal{U}})$ denotes the spectrum of $F^t_{\mathcal{U}}$ on $\mathcal{U}$ (which in view of (L.225), (L.228) contains only non-zero positive elements starting from some $t \geq t_0$).

We begin with characterization of mapping $\mu^t_{\mathcal{U}}(\xi)$ via the dual problem in (L.216).

First, from (L.214), (L.216) it follows that

$$G^t(\mu) = -\infty \text{ if } \mu \notin \mathcal{W}^{\perp}. \tag{L.231}$$

That is for $\mu \notin \mathcal{W}^{\perp}$ the dual problem is unfeasible. In view of this and the strong duality, formulas in (L.216) can be rewritten as follows:

$$G^t(\mu) = \inf_{u \in \mathcal{U}} \mathcal{L}^t(u, 0; \mu),\ \mu \succeq 0,\ \mu \in \mathcal{W}^{\perp}, \tag{L.232}$$

$$\mu^t = \operatorname*{arg\,max}_{\mu \succeq 0,\, \mu \in \mathcal{W}^{\perp}} G^t(\mu). \tag{L.233}$$

Using (L.214), (L.224), the first order optimality condition in (L.232) has the following form:

$$\begin{aligned} &u^t_{min}(\mu) = (F^t_{\mathcal{U}})^{-1}(C^t_{\mathcal{U}})^T\xi + (F^t_{\mathcal{U}})^{-1}\frac{\mu_{\mathcal{U}}}{\sqrt{t}}, \\ &\mu_{\mathcal{U}} = \Pi_{\mathcal{U}}\mu,\ \mu \succeq 0,\ \mu \in \mathcal{W}^{\perp}. \end{aligned} \tag{L.234}$$

From (L.214), (L.216), (L.232), (L.234) it follows that

$$\begin{aligned} G^t(\mu) = \mathcal{L}^t(u^t_{min}(\mu), 0; \mu) = -\xi^T C^t_{\mathcal{U}} u^t_{min}(\mu) + \frac{1}{2}[u^t_{min}(\mu)]^T F^t_{\mathcal{U}} u^t_{min}(\mu) \\ - \mu^T((1 - \Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} + \frac{u^t_{min}(\mu)}{\sqrt{t}}), \\ \mu_{\mathcal{U}} = \Pi_{\mathcal{U}}\mu,\ \mu \succeq 0,\ \mu \in \mathcal{W}^{\perp}. \end{aligned} \tag{L.235}$$

Formulas (L.234), (L.235) imply that

$$\begin{aligned} &G^t(\mu) = -\frac{1}{2}\frac{\mu^T_{\mathcal{U}}}{\sqrt{t}}(F^t_{\mathcal{U}})^{-1}\frac{\mu_{\mathcal{U}}}{\sqrt{t}} - \xi^T C^t_{\mathcal{U}}(F^t_{\mathcal{U}})^{-1}\frac{\mu_{\mathcal{U}}}{\sqrt{t}} - \mu^T(I - \Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc}, \\ &\mu \succeq 0,\ \mu \in \mathcal{W}^{\perp}. \end{aligned} \tag{L.236}$$

From the facts that $\mu \in \mathcal{W}^{\perp}$, $\mu \succeq 0$ and the definition of $\mathcal{V}$ in (5.17) it follows that

$$\mu_{\mathcal{U}} = (I - \Pi_{\mathcal{V}})\mu = \begin{cases} \mu_j, \text{ if } \sum\limits_{i \in I_0(\Lambda^*)} a_{ij} = 0, \\ 0, \text{ otherwise,} \end{cases} \Rightarrow \mu_{\mathcal{U}} = (I - \Pi_{\mathcal{V}})\mu \succeq 0. \tag{L.237}$$

From (L.237) and the fact that $\widehat{\lambda}^t_{sc} \succeq 0$ it follows that

$$\mu^T(I - \Pi_{\mathcal{V}})\widehat{\lambda}^t_{sc} = [(I - \Pi_{\mathcal{V}})\mu]^T\widehat{\lambda}^t_{sc} = \mu^T_{\mathcal{U}}\widehat{\lambda}^t_{sc} \geq 0. \tag{L.238}$$

From (L.236) one can see that minimizer μ^t in (L.233) may not be unique, however, its projection $\mu^t_{\mathcal{U}}$ is unique since functional $G^t(\mu)$ is strongly convex in $\mu_{\mathcal{U}}$. At the same time, from (L.229) it follows

that only $\mu^t_{\mathcal{U}}$ is essential for $\widetilde{u}^t(\xi)$. In view of (L.229), (L.236), the optimization problem in (L.233) can be rewritten as follows:

$$\frac{\mu^t_{\mathcal{U}}}{\sqrt{t}} = \widetilde{\mu}^t_{\mathcal{U}} = \underset{\mu_{\mathcal{U}} \in \Pi_{\mathcal{U}}(\mathbb{R}^p_+ \cap \mathcal{W}^\perp)}{\arg\min} \frac{1}{2}\|(F^t_{\mathcal{U}})^{-1/2}\mu_{\mathcal{U}} + (F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\xi\|^2 + \sqrt{t}\mu^T_{\mathcal{U}}\widehat{\lambda}^t_{sc}. \tag{L.239}$$

From (L.239) and the fact that $0 \in \Pi_{\mathcal{U}}(\mathbb{R}^p_+ \cap \mathcal{W}^\perp)$ it follows that

$$\frac{1}{2}\|(F^t_{\mathcal{U}})^{-1/2}\widetilde{\mu}^t_{\mathcal{U}} + (F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\xi\|^2 + \sqrt{t}\mu^t_{\mathcal{U}}\widehat{\lambda}^t_{sc} \le \|(F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\xi\|^2, \tag{L.240}$$

where $\widetilde{\mu}^t_{\mathcal{U}}$ is the solution in (L.239). Formulas (L.238), (L.240) imply that

$$|(F^t_{\mathcal{U}})^{-1/2}\widetilde{\mu}^t_{\mathcal{U}} + (F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\xi| \le \|(F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\xi\|. \tag{L.241}$$

which together with inequality $|a+b| \ge |a| - |b|$ imply the following estimate

$$\|(F^t_{\mathcal{U}})^{-1/2}\widetilde{\mu}^t_{\mathcal{U}}\| \le 2\|(F^t_{\mathcal{U}})^{-1/2}(C^t_{\mathcal{U}})^T\xi\|. \tag{L.242}$$

From (5.18), (L.224), (L.225), (L.228) it follows that $F^t_{\mathcal{U}}$ is of full rank on $\mathcal{U}$ (starting from some $t \ge t_0$ a.s. Y^t, $t \in (0, +\infty)$), therefore, for large t matrix $(F^t_{\mathcal{U}})^{-1/2}$ is positive definite, injective on $\mathcal{U}$ and, hence, $|(F^t_{\mathcal{U}})^{-1/2}\widetilde{\mu}^t_{\mathcal{U}}| \ge \min_{\sigma \in \sigma_{\mathcal{U}}(F^t_{\mathcal{U}})} \sigma^{1/2}\|\widetilde{\mu}^t_{\mathcal{U}}\|$, where $\sigma_{\mathcal{U}}(\cdot)$ denotes the spectrum of an operator acting on $\mathcal{U}$.

The above argument with formula (L.242) directly imply (L.230).

Formulas (L.128)-(L.133) follow from (L.224)-(L.228), (L.229), (L.230).

Lemma is proved.

L.16. Proof of Lemma L.11

Proof. In view of step 2 in Algorithm 5 intensities $\widetilde{\Lambda}^t_{b,i}$ can be represented as follows:

$$\widetilde{\Lambda}^t_{b,i} = \frac{1}{\theta^t + t}\sum_{k=1}^{Y^t_i} w_{ik} + \widetilde{r}^t_{b,\mathcal{M},i},\ i \in I_1(\Lambda^*), \tag{L.243}$$

$$\{w_{ik}\}^{\infty,\,d}_{k=1,\,i=1} \text{ are mutually independent, } w_{ik} \sim \Gamma(1,1), \tag{L.244}$$

where

$$\begin{aligned} &\widetilde{r}^t_{b,\mathcal{M},i} \mid \widetilde{\Lambda}^t_{\mathcal{M},i}, Y^t, t \sim \Gamma(\theta^t\Lambda^t_{\mathcal{M},i}, (\theta^t + t)^{-1}),\\ &\widetilde{\Lambda}^t_{\mathcal{M},i} \text{ are sampled in Algorithm 4.} \end{aligned} \tag{L.245}$$

In particular,

$$\sqrt{t}r^t_{b,\mathcal{M},i} = o_{cp}(1). \tag{L.246}$$

Indeed, from (L.70), (L.245) and the Markov inequality it holds that

$$\begin{aligned} P(\sqrt{t}r^t_{b,\mathcal{M},i} > \delta \mid Y^t, t) &\le \frac{\sqrt{t}\theta^t}{\delta(\theta^t + t)}E[\widetilde{\Lambda}^t_{\mathcal{M},i} \mid Y^t, t] \\ &\le \frac{\sqrt{t}\theta^t}{\delta(\theta^t + t)}\sum_{i \in I_1(\Lambda^*)}\frac{Y^t_i}{t} \to 0 \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty), \end{aligned} \tag{L.247}$$

where δ is arbitrary positive value.

Using the Central Limit Theorem for sums of w_{ik} in (L.243), (L.244) and the Strong Law of Large Numbers for Y^t (see Theorem B.1, formula (B.2)) and the fact that $\theta^t = o(\sqrt{t})$, we obtain:

$$\frac{\sqrt{t}}{(\theta^t + t)\sqrt{Y^t_i/t}}\sum_{k=1}^{Y^t_i}(w_{ik} - 1) \xrightarrow{c.d.} \mathcal{N}(0,1) \text{ for } t \to +\infty, \text{ a.s. } Y^t, t \in (0, +\infty). \tag{L.248}$$

Due to mutual independence between w_{ik}, the above convergence holds for all components $i \in I_1(\Lambda^*)$, hence, as for the vector in $\mathbb{R}^{\#I_1(\Lambda^*)}$.

Using formula (L.134) we obtain:

$$\begin{aligned} A^T_{I_1(\Lambda^*)}(\widehat{D}^t_{I_1(\Lambda^*)})^{-1/2}\widetilde{\xi}^t &= \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\frac{\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i \\ &= \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\frac{\widetilde{\Lambda}^t_{b,i}-Y^t_i/t}{\widehat{\Lambda}^t_{sc,i}}a_i + \sum_{i\in I_1(\Lambda^*)} \sqrt{t}\frac{Y^t_i/t-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i. \end{aligned} \tag{L.249}$$

The first sum is conditionally tight in view of the Prokhorov theorem on tightness of weakly convergence sequences and the result in (L.248). Due to (5.23) 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, $t \in (0,+\infty)$.

Lemma is proved. □

L.17. Proof of Lemma L.12

Proof. Since $B^t(\lambda)$ is proportional to t in (L.103), it suffices to prove formula (L.151) for normalized process $B^t(\lambda)/t$ which we denote here by $G^t(\lambda)$, that is

$$\begin{aligned} G^t(\lambda) = &\sum_{i\in I_1(\Lambda^*)} -(\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i})\frac{\Lambda_i-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)}\frac{(\Lambda_i-\widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} \\ &+ \sum_{i\in I_0(\Lambda^*)} \Lambda_i,\ \Lambda_i = a_i^T\lambda,\ i\in\{1,\dots,d\}. \end{aligned} \tag{L.250}$$

Note also that minimizers of B^t and of G^t coincide.

From the necessary Karush-Kuhn-Tucker optimality conditions in (L.104) (see e.g., Bertsekas (1997), Section 3.3) it follows that

$$\begin{aligned} &\exists \widetilde{\lambda}^t_{b,app}, \widetilde{\mu}^t_{b,app} \in \mathbb{R}^p_+ \text{ such that} \\ &-\sum_{i\in I_1(\Lambda^*)}\frac{\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i + \sum_{i\in I_1(\Lambda^*)}\frac{\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}}a_i + \sum_{i\in I_0(\Lambda^*)} a_i - \widetilde{\mu}^t_{b,app} = 0, \end{aligned} \tag{L.251}$$

$$\begin{aligned} &\widetilde{\Lambda}^t_{b,app} = A\widetilde{\lambda}^t_{b,app}, \\ &\widetilde{\mu}^t_{b,app,j}\widetilde{\lambda}^t_{b,app,j} = 0 \text{ for all } j\in\{1,\dots,p\}. \end{aligned} \tag{L.252}$$

Multiplying both sides of (L.251) on $(\widetilde{\lambda}^t_{b,app}-\widehat{\lambda}^t_{sc})$ and using formula (L.252) we obtain following formulas:

$$\begin{aligned} -\langle\widetilde{\mu}^t_{b,app},\widehat{\lambda}^t_{sc}\rangle = &-\sum_{i\in I_1(\Lambda^*)}\frac{(\widetilde{\Lambda}^t_{b,i}-\widehat{\Lambda}^t_{sc,i})(\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i})}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_1(\Lambda^*)}\frac{(\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} \\ &+ \sum_{i\in I_0(\Lambda^*)}\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i}, \end{aligned} \tag{L.253}$$

$$-\langle\widetilde{\mu}^t_{b,app},\widehat{\lambda}^t_{sc}\rangle = \sum_{i\in I_1(\Lambda^*)}\widetilde{\Lambda}^t_{b,i}-\widetilde{\Lambda}^t_{b,app,i} - \sum_{i\in I_0(\Lambda^*)}\widehat{\Lambda}^t_{sc,i}. \tag{L.254}$$

From formulas (L.250), (L.251), (L.253) it follows that

$$G^t(\widetilde{\lambda}^t_{b,app}) = -\langle\widetilde{\mu}^t_{b,app},\widehat{\lambda}^t_{sc}\rangle - \frac{1}{2}\sum_{i\in I_1(\Lambda^*)}\frac{(\widetilde{\Lambda}^t_{b,app,i}-\widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)}\widehat{\Lambda}^t_{sc,i}. \tag{L.255}$$

Using (L.250)-(L.255) we get the following identity:

$$
\begin{aligned}
G^t(\lambda) - G^t(\widetilde{\lambda}^t_{b,app}) = & \sum_{i\in I_1(\Lambda^*)} -(\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{\Lambda_i - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_0(\Lambda^*)} \Lambda_i - \widehat{\Lambda}^t_{sc,i} \\
& + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widehat{\Lambda}^t_{sc,i})^2 + (\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})^2}{\widehat{\Lambda}^t_{sc,i}} + \langle \widetilde{\mu}^t_{b,app}, \widehat{\lambda}^t_{sc}\rangle \\
= & \sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i})^2}{2\widehat{\Lambda}^t_{sc,i}} + \sum_{i\in I_1(\Lambda^*)} \frac{(\widetilde{\Lambda}^t_{b,app,i} - \widehat{\Lambda}^t_{sc,i})(\Lambda_i - \widehat{\Lambda}^t_{sc,i})}{\widehat{\Lambda}^t_{sc,i}} \\
& + \sum_{i\in I_0(\Lambda^*)} \Lambda_i - \widehat{\Lambda}^t_{sc,i} + \sum_{i\in I_1(\Lambda^*)} \widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i} + \sum_{i\in I_0(\Lambda^*)} \widehat{\Lambda}^t_{sc,i} \\
& - \sum_{i\in I_1(\Lambda^*)} (\widetilde{\Lambda}^t_{b,i} - \widehat{\Lambda}^t_{sc,i})\frac{\Lambda_i - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} \\
= & \sum_{i\in I_1(\Lambda^*)} \frac{(\Lambda_i - \widetilde{\Lambda}^t_{b,app,i})^2}{2\Lambda^*_i} + \sum_{i\in I_0(\Lambda^*)} \Lambda_i + \sum_{i\in I_1(\Lambda^*)} \frac{\widetilde{\Lambda}^t_{b,app,i} - \widetilde{\Lambda}^t_{b,i}}{\widehat{\Lambda}^t_{sc,i}}\Lambda_i.
\end{aligned}
\tag{L.256}
$$

Formulas (L.151)-(L.153) follow from (L.250) (L.251), (L.252), (L.256).

Lemma is proved. □

L.18. Proof of Lemma L.13

Proof. Consider the following formula

$$
\inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{app})] = \inf_{\lambda\in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} [\varphi(\lambda - \varphi(\widetilde{\lambda}^t_{b,app})] + [\varphi(\widetilde{\lambda}^t_{b,app}) - \varphi(\widetilde{\lambda}^t_{app})]. \tag{L.257}
$$

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 (L.103)) 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 (L.103), (L.104), (L.115) it follows that $\widetilde{u}^t_{b,app} = \sqrt{t}\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \widehat{\lambda}^t_{sc})$ is unique. At the same time, from (5.24), (L.120) and the result of Lemma L.9 it follows that

$$
\widetilde{v}^t_{b,app} = t\Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \widehat{\lambda}^t_{sc}) = o_{cp}(1), \tag{L.258}
$$

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

$$
\begin{aligned}
& w(u,v) = \operatorname*{arg\,min}_{\substack{w:\lambda_*+u+v+w\geq 0 \\ w\in\mathcal{W}}} \varphi(\lambda_* + u + v + w), \\
& u \in \mathcal{U},\ v \in \mathcal{V} : (\lambda_* + u + v + \mathcal{W}) \cap \mathbb{R}^p_+ \neq \emptyset,
\end{aligned}
\tag{L.259}
$$

where λ_* is the true parameter. From the strict convexity of $\varphi(\cdot)$ along $\ker A$ (by the assumption in (2.12)), the definition of $\mathcal{W}$ in (5.19) and the result of Lemma 2.1 it follows that $w(u,v)$ is one-to-one and continuous in (u,v) on its domain of definition.

Note that

$$
w_* = w(0,0) = w_{A,\lambda_*}(0,0), \tag{L.260}
$$

where $w_{A,\lambda}(\cdot,\cdot)$ is defined in (L.259) ($w_{A,\lambda_*}(0,0)$ appears in Theorems 2, 3). The property that $w_* \in \mathcal{W}$ can be proved by the contradiction argument. Assume that $w_* \in \ker A$ but $w_* \notin \mathcal{W}$ and $w_* \neq 0$. Then, from the definition of $\mathcal{V}, \mathcal{U}, \mathcal{W}$ it follows that

$$
\exists i \in I_0(\Lambda^*),\ j \in \{1,\dots,p\} : a_{ij} > 0,\ w_{*j} > 0. \tag{L.261}
$$

At the same time from the fact that $w_* \in \ker A$ it follows that

$$0 = \sum_{i \in I_0(\Lambda^*)} a_i^T w_* = \sum_{j=1}^{p} \left(\sum_{i \in I_0(\Lambda^*)} a_{ij} \right) w_{*j} \tag{L.262}$$

Formulas (L.261), (L.262) imply that

$$\exists i' \in I_0(\Lambda^*),\ j' \in \{1, \dots, p\} : a_{i'j'} > 0,\ w_{*j'} < 0. \tag{L.263}$$

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 (L.263) contradicts the positivity constraint in (L.260). Thus, $w_* \in \mathcal{W}$.

Let

$$\widetilde{w}^t_{b,app} = w\left(\Pi_{\mathcal{U}}(\widehat{\lambda}^t_{sc} - \lambda_*) + \frac{\widetilde{u}^t_{b,app}}{\sqrt{t}}, \Pi_{\mathcal{V}}(\widehat{\lambda}^t_{sc} - \lambda_*) + \frac{\widetilde{v}^t_{b,app}}{t} \right), \tag{L.264}$$

where $\widetilde{u}^t_{b,app}$, $\widetilde{v}^t_{b,app}$ are defined in (L.104), (L.115), w is the mapping from (L.259). Recall that $\widetilde{\lambda}^t_{b,app}$ from (L.104) can be rewritten via the parametrization in (L.115) as follows

$$\widetilde{\lambda}^t_{b,app} = \widehat{\lambda}^t_{sc} + \frac{\widetilde{u}^t_{b,app}}{\sqrt{t}} + \frac{\widetilde{v}^t_{b,app}}{t} + \widetilde{w}^t_{b,app}, \tag{L.265}$$

where $\widetilde{w}^t_{b,app}$ is chosen in (L.264). For $\widetilde{\lambda}^t_{b,app}$ from (L.265) it holds that

$$\widetilde{\lambda}^t_{b,app} \xrightarrow{c.p.} \lambda_* + w_* \text{ for } t \to +\infty, \text{ a.s. } Y^t,\ t \in (0, +\infty), \tag{L.266}$$

where w_* is defined in (L.260).

Indeed, formula (L.266) 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 (L.120) and results of Lemma L.11) and the continuity of mapping w.

From the local Lipschitz continuity of φ and (L.106), (L.107), (L.266) it follows that there exits some universal constant $L > 0$ such that with conditional probability tending to one a.s. Y^t, $t \in (0, +\infty)$ it holds that:

$$\varphi(\widetilde{\lambda}^t_{b,app}) - \varphi(\widetilde{\lambda}^t_{app}) \le L\|\widetilde{\lambda}^t_{b,app} - \widetilde{\lambda}^t_{app}\|. \tag{L.267}$$

In particular, from (L.106), (L.107), (L.267) it follows that

$$\beta^t(\varphi(\widetilde{\lambda}^t_{b,app}) - \varphi(\widetilde{\lambda}^t_{app})) = o_{cp}(1). \tag{L.268}$$

It is left to show that the first term in (L.257) is also of order $o_{cp}(1)$. For this we use extensively the results from Wets (2003) on the lipshitz-continuity of inf-projections.

The first term in (L.257) can be rewritten as taking the infimum two times:

$$\begin{aligned} \inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{b,app})) = \inf_{\substack{(u,v) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}) \\ (u,v) \in \mathcal{U}\times\mathcal{V}}} [\varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \lambda_*) + \frac{u}{\sqrt{t}}, \Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \lambda_*) + \frac{v}{t}) \\ - \varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \lambda_*), \Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \lambda_*))], \end{aligned} \tag{L.269}$$

where

$$\begin{aligned} \varphi_*(u,v) = \inf_{\substack{w : \lambda_* + u + v + w \ge 0, \\ w \in \mathcal{W}}} \varphi(\lambda_* + u + v + w), \\ u \in \mathcal{U},\ v \in \mathcal{V} : (\lambda_* + u + v + \mathcal{W}) \cap \mathbb{R}^p_+ \ne \emptyset. \end{aligned} \tag{L.270}$$

The expression in the square brackets in (L.269) is essentially the variation of the inf-projection for $\varphi_*(u,v)$ for parameter $(u,v) \in \mathcal{U} \times \mathcal{V}$ in the vicinity of zero along $\mathcal{U} \oplus \mathcal{V}$. 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 Wets (2003) we find that $\varphi_*(u,v)$ is locally Lipschitz continuous.

Indeed, consider the optimization problem in (L.270), where $(u,v) \in \mathcal{U} \times \mathcal{V}$ is a parameter. Then, the problem can be rewritten as follows:

$$\inf_w \varphi_0((u,v);w),\ \varphi_0 : (\mathcal{U} \times \mathcal{V}) \times \mathcal{W} \to \overline{\mathbb{R}}, \tag{L.271}$$

$$\varphi_0((u,v);w) = \begin{cases} \varphi(\lambda_* + u + v + w), & \text{if } \lambda_* + u + v + w \succeq 0, \\ +\infty, & \text{otherwise,} \end{cases} \tag{L.272}$$

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

$$S : \mathcal{U} \times \mathcal{V} \rightrightarrows \mathcal{W} \text{ with } S(u,v) = \{w \in \mathcal{W} : \lambda_* + u + v + w \succeq 0\}, \tag{L.273}$$

where $\rightrightarrows$ denotes the property to be a set-valued mapping. From (L.273) one can see that $\operatorname{gph} S = D$ (gph denotes the graph of a mapping). Therefore, $\operatorname{gph} S$ is polyhedral and, hence, the Proposition 4.1 from Wets (2003) applies to our case (see also Example 9.35 in Rockafellar and Wets (2009)), so mapping S in (L.273) is Lipschitz continuous on $\operatorname{dom} S$ (as set-valued mapping). At the same time, the result of Lemma L.1 implies that feasibility mapping S is locally bounded which yields level boundedness in w locally uniformly in (u,v) of $\varphi_0(\cdot,\cdot)$. The above properties are exactly the same is in Section 4 of Wets (2003), so Theorem 3.4 therein applies to the case of φ_0 from (L.271) and $\varphi_*(u,v) = \inf_w \varphi((u,v);w)$ is locally Lipschitz continuous.

Hence, 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

$$\begin{aligned} &| \varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \lambda_*) + \frac{u}{\sqrt{t}}, \Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \lambda_*) + \frac{v}{t}) - \varphi_*(\Pi_{\mathcal{U}}(\widetilde{\lambda}^t_{b,app} - \lambda_*), \Pi_{\mathcal{V}}(\widetilde{\lambda}^t_{b,app} - \lambda_*)) | \\ &\quad \le L \left(\frac{\|u\|}{\sqrt{t}} + \frac{\|v\|}{t} \right) \le L \left(\frac{\delta}{\sqrt{t}} + c\frac{\delta}{t} \right) \text{ for any } (u,v) \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app}), \end{aligned} \tag{L.274}$$

where c is a positive constant depending only on dimension p.

Using formulas (L.269), (L.274) and the assumption that $\beta^t = o(\sqrt{t})$ we obtain

$$\beta^t \inf_{\lambda \in C^t_{A,\delta}(\widetilde{\lambda}^t_{b,app})} (\varphi(\lambda) - \varphi(\widetilde{\lambda}^t_{b,app})) = o_{cp}(1). \tag{L.275}$$

Formula (L.194) directly follows from (L.268), (L.275).

Lemma is proved. □

References

Attouch, H. and Wets, R. J.-B. (1993). "Quantitative stability of variational systems. II. A framework for nonlinear conditioning." *SIAM Journal on Optimization*, 3(2): 359–381. 33

Aykroyd, R. G. and Green, P. J. (1991). "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. 22

Barrett, H. H., Wilson, D. W., and Tsui, B. M. (1994). "Noise properties of the EM-algorithm. I. Theory." *Phys. Med. Biol.*, 39(5): 833. 1

Bertsekas, D. P. (1997). "Nonlinear programming." *Journal of Operational Research Society*, 48(3): 334. 40, 54, 59

Blei, D. and Frazier, P. (2011). "Distance dependent Chinese restaurant processes." *Journal of Machine Learning Research*, 12(8). 26

Bochkina, N. A. and Green, P. J. (2014). "The Bernstein–von Mises theorem and nonregular models." *The Annals of Statistics*, 42(5): 1850–1878. 1, 2, 6, 7, 13, 14, 17

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

Bowsher, J., Yuan, H., Hedlund, L., Turkington, T., Akabani, G., Badea, A., Kurylo, W., Wheeler, C., Cofer, G., Dewhirst, M., and Johnson, G. (2004). "Utilizing MRI information to estimate F18-FDG distributions in rat flank tumors." In *IEEE Symposium Conference Record Nuclear Science*, volume 4. IEEE. 1, 5

Chun, S. Y., Fessler, J. A., and Dewaraja, Y. K. (2013). "Post-reconstruction non-local means filtering methods using CT side information for quantitative SPECT." *Physics in Medicine & Biology*, 58(17): 6225. 1

Comtat, C., Kinahan, P. E., Fessler, J. A., Beyer, T., Townsend, D. W., Defrise, M., and Michel, C. (2001). "Clinically feasible reconstruction of 3D whole-body PET/CT data using blurred anatomical labels." *Physics in Medicine & Biology*, 47(1): 1. 1

Dahlbom, M. (2001). "Estimation of image noise in PET using the bootstrap method." In *IEEE Nuclear Science Symposium Conference Record*, volume 4. IEEE. 1, 2

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media. 10, 19

De Pierro, A. (1993). "On the relation between the ISRA and the EM algorithm for positron emission tomography." *IEEE Transactions on Medical Imaging*, 12(2): 328–333. 23

Duan, L. L., Johndrow, J. E., and Dunson, D. B. (2018). "Scaling up Data Augmentation MCMC via Calibration." *Journal of Machine Learning Research*, 19(1): 2575–2608. 2, 8

Erdogan, H. and Fessler, J. (1999). "Monotonic algorithms for transmission tomography." *IEEE Transactions on Medical Imaging*, 18(9): 801–814. 24

Ferreira, A. R. and Lee, K. H. (2007). "Single Photon Emission Computed Tomography Example." In *Multiscale Modeling*. Springer Series in Statistics. 1, 2, 8

Fessler, J. and Hero, A. (1995). "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms." *IEEE Transactions on Image Processing*, 4(10): 1417–1429. 3, 12, 13, 23

Fessler, J. A. (1996). "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. 1, 2

Fessler, J. A., Clinthorne, N. H., and Rogers, W. L. (1992). "Regularized emission image reconstruction using imperfect side information." *IEEE Transactions on Nuclear Science*, 39(5): 1464–1471. 1

Filipović, M., Barat, E., Dautremer, T., Comtat, C., and Stute, S. (2018). "PET reconstruction of the posterior image probability, including multimodal images." *IEEE transactions on medical imaging*, 38(7): 1643–1654. 1, 2, 5, 6, 8

Filipović, M., Dautremer, T., Comtat, C., Stute, S., and Barat, E. (2021). "Reconstruction, analysis and interpretation of posterior probability distributions of PET images, using the posterior bootstrap." *Physics in Medicine & Biology*. 1, 5, 21

Fong, E., Lyddon, S., and Holmes, C. (2019). "Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 1952–1962. PMLR. 2, 10, 13, 19, 21

Geyer, C. J. (1994). "On the asymptotics of constrained M-estimation." *The Annals of Statistics*, 22(4): 1993–2010. 14

Ghosh, S., Ungureanu, A. B., Sudderth, E. B., and Blei, D. M. (2011). "Spatial distance dependent Chinese restaurant processes for image segmentation." In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 24*, 1476–1484. Curran Associates, Inc.
URL http://papers.nips.cc/paper/4361-spatial-distance-dependent-chinese-restaurant-processes-for pdf 5

Goncharov, F. (2019). "Weighted Radon transforms and their applications." Ph.D. thesis, Université Paris Saclay (COmUE). 30

Green, P. J. (1990). "Bayesian reconstructions from emission tomography data using a modified EM algorithm." *IEEE Trans. Med. Imag.*, 9: 84–93. 8, 25

Gut, A. (2013). *Probability: a graduate course*. New York, NY: Springer. 20

Han, G., Liang, Z., and You, J. (1999). "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, 1515–1518 vol.3. 30

Haynor, D. R. and Woods, S. D. (1989). "Resampling estimates of precision in emission tomography." *IEEE Transactions on Medical Imaging*, 8(4). 1, 2

Hero, A. O., Piramuthu, R., Fessler, J. A., and Titus, S. R. (1999). "Minimax emission computed tomography using high-resolution anatomical side information and B-spline models." *IEEE Transactions on Information Theory*, 45(3): 920–938. 1

Higdon, D., Bowsher, J., Johnson, V., Turkington, T., Gilland, D., and Jaszczak, R. (1997). "Fully Bayesian estimation of Gibbs hyperparameters for emission computed tomography data." *IEEE Transactions on Medical Imaging*, 16: 516. 1, 2, 8

Hjort, N. L. and Pollard, D. (2011). "Asymptotics for minimisers of convex processes." *arXiv preprint arXiv:1107.3806*. 44

Hohage, T. and Werner, F. (2016). "Inverse problems with Poisson data: statistical regularization theory, applications and algorithms." *Inverse Problems*, 32(9): 093001. 1, 4, 9

James, L. F. (2003). "Bayesian calculus for gamma processes with applications to semiparametric intensity models." *Sankhyā: The Indian Journal of Statistics*, 179–206. 6, 11

Judenhofer, M., Wehrl, H., Newport, D., Catana, C., Siegel, S., Becker, M., Thielscher, A., Kneilling, M., Lichy, M., Eichner, M., Klingel, K., Reischl, G., Widmaier, S., Röcken, M., Nutt, R., Machulla, H., Uluda, K., Cherry, S., Claussen, C., and Pichler, B. (2008). "Simultaneous PET-MRI: a new approach for functional and morphological imaging." *Nature medicine*, 14(4): 459–465. 1

Lange, K., Hunter, D. R., and Yang, I. (2000). "Optimization Transfer Using Surrogate Objective Functions." *Journal of Computational and Graphical Statistics*, 9(1): 1–20. 23

Lartizien, C., Aubin, J.-B., and Buvat, I. (2010). "Comparison of bootstrap resampling methods for 3-D PET imaging." *IEEE Transactions on Medical Imaging*, 29(7): 1442–1454. 1

Levin, C. S., Dahlbom, M., and Hoffman, E. J. (1995). "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. 17

Li, Y. (2011). "Noise propagation for iterative penalized-likelihood image reconstruction based of Fisher information." *Phys. Med. Biol.*, 56(4): 1083. 1

Liu, J. S. (1994). "The fraction of missing information and convergence rate for data augmentation." *Computing Science and Statistics*, 490–497. 7

Liu, J. S., Wong, W. H., and Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes." *Biometrika*, 81(1): 27–40. 7, 22

Lo, A. Y. (1982). "Bayesian nonparametric statistical inference for Poisson point processes." *Zeitschrift fur Wahrscheinlichkeitsteorie und verwandte Gebiete*, 59(1): 55–66. 3, 11, 36

Luna, A., Vilanova, J. C., Hygino da Cruz Jr, L. C., and Rossi, S. E. (2013). *Functional imaging in oncology: biophysical basis and technical approaches - Vol. 1*. Springer Science & Business Media. 1

Lyddon, S., Walker, S., and Holmes, C. (2018). "Nonparametric learning from Bayesian models with randomized objective functions." *Advances in Neural Information Processing Systems*. 2, 10, 11, 13, 19, 21

Marcu, L. G., Moghaddasi, L., and Bezak, E. (2018). "Imaging of tumor characteristics and molecular pathways with PET: developments over the last decade toward personalized cancer therapy." *International Journal of Radiation Oncology Biology Physics*, 102(4): 1165–1182. 1

Natterer, F. (2001). *The mathematics of computerized tomography*. Society for Industrial and Applied Mathematics. 4, 30

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1): 3–26. 2, 13, 20, 21

Ng, T. L. and Newton, M. A. (2020). "Random weighting in LASSO regression." *arXiv preprint arXiv:2002.02629*. 2, 17, 20, 30

Novikov, R. (2019). "5. Non-Abelian Radon transform and its applications." In *The Radon Transform*, 115–128. De Gruyter. 30

Pompe, E. (2021). "Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification." *arXiv preprint arXiv:2103.14445*. 2, 13, 29

Quinto, E. T. (1983). "The invertibility of rotation invariant Radon transforms." *Journal of Mathe-*

matical Analysis and Applications, 91(2): 510–522. 30

Rahmim, A., Tang, J., and Zaidi, H. (2009). "Four-dimensional (4D) image reconstruction strategies in dynamic PET: Beyond conventional independent frame reconstruction." *Medical Physics*, 36(8): 3654–3670. 17

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media. 62

Shepp, L. A. and Vardi, Y. (1982). "Maximum likelihood reconstruction for emission tomography." *IEEE transactions on medical imaging*, 1(2): 113–122. 3, 6, 13

Siddon, R. (1985). "Fast calculation of the exact radiological path for a three-dimensional CT array." *Medical physics*, 12 2: 252–5. 22, 26, 30

Sitek, A. (2012). "Data analysis in emission tomography using emission count posteriors." *Physics in Medicine & Biology*, 52(21): 6779. 1

Sitek, A. and Celler, M. A. (2015). "Limitations of Poisson statistics in describing radioactive decay." *Physica Medica*, 31(8): 1105–1107. 18

Stute, S. and Comtat, C. (2013). "Practical considerations for image-based PSF and blobs reconstruction in PET." *Physics in Medicine & Biology*, 58(11): 3849. 17

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. 2, 37, 43

Van Dyk, D. A. and Meng, X.-L. (2001). "The art of data augmentation." *Journal of Computational and Graphical Statistics*, 10(1): 1–50. 2, 8

Vunckx, K., Atre, A., Baete, K., Reilhac, A., Deroose, C. M., Van Laere, K., and Nuyts, J. (2011). "Evaluation of three MRI-based anatomical priors for quantitative PET brain imaging." *IEEE transactions on medical imaging*, 31(3): 599–612. 1, 5, 26

Walkup, D. W. and Wets, R. J.-B. (1969). "A Lipschitzian characterization of convex polyhedra." *Proceedings of the American Mathematical Society*, 167–173. 35

Wang, G. and Qi, J. (2015). "Edge-Preserving PET Image Reconstruction Using Trust Optimization Transfer." *IEEE Transactions on Medical Imaging*, 34(4): 930–939. 23, 24

Weber, W. A. (2005). "Use of PET for monitoring cancer therapy and for predicting outcome." *Journal of Nuclear Medicine*, 46(6): 983–995. 1

Weir, I. S. (1997). "Fully Bayesian reconstructions from single-photon emission computed tomography data." *Journal of the American Statistical Association*, 92(437): 49–60. 1, 2, 8

Wets, R. J.-B. (2003). "Lipschitz continuity of inf-projections." *Computational Optimization and Applications*, 25(1-3): 269–282. 61, 62