Unbiased Bregman-Risk Estimators: Application to Regularization Parameter Selection in Tomographic Image Reconstruction

Elias S. Helou, Sandra A. Santos, and Lucas E. A. Simões
SME/ICMC/USP, Postal Box 668, 13560-970, São Carlos, SP, Brazil
E-mail: elias@icmc.usp.br

Abstract. Unbiased estimators are introduced for averaged Bregman divergences which generalize Stein’s Unbiased (Predictive) Risk Estimator, and the minimization of these estimators is proposed as a regularization parameter selection method for regularization of inverse problems. Numerical experiments are presented in order to show the performance of the proposed technique. Experimental results indicate a useful occurrence of a concentration of measure phenomena and some implications of this hypothesis are analyzed.

Keywords: Bregman divergences, Regularization, Parameter selection, Tomographic image reconstruction, Concentration of measure

Submitted to: Inverse Problems

1. Introduction

Many problems in science and engineering can be formulated as a system of nonlinear equations of the form

$$A(x) \approx b,$$  \hspace{1cm} (1)

where $x \in \mathbb{R}^n$ is the vector of unknowns, $A: \mathbb{R}^n \mapsto \mathbb{R}^m$ is the system function arising from a mathematical model for the problem, and $b \in \mathbb{R}^m$ is the vector of observed data, which contains noise, that is, it is given by

$$b = A(x^*) + \epsilon,$$  \hspace{1cm} (2)

where $x^* \in \mathbb{R}^n$ is the exact solution and $\epsilon$ is some unknown vector of random variables. The methodology we will propose can be applied to several noise models, including Poisson distributed, the sum of Gauss distributed and Poisson distributed, exponential family distributed and elliptically contoured distributed data. Example
applications include tomographic image reconstruction [31, 36, 44] and image denoising and deblurring [12, 45].

Because ill-conditioned inverse problems arise so often in applications, methods for obtaining meaningful results from noisy data have been devised. These techniques are the so-called regularization methods [21] and always require a regularization parameter to be selected by the user. In the present paper, we develop techniques for estimating certain expected errors and we apply these techniques to the problem of selecting parameters for nonlinear approaches to regularization.

Among the first regularization methods, we have Tikhonov regularization [21], which consists of using, as an approximation to the unknown solution, the minimizer $x^\gamma_{\text{Tikhonov}}$ of

$$
\min \|Ax - b\|_2^2 + \gamma \|x\|_2^2.
$$

(3)

Here, $\gamma \geq 0$ is the regularization parameter, which determines how much smoothness is imposed on the regularized solution, and we are assuming that the problem is linear, i.e., $A$ is a matrix. It is possible to verify that the Tikhonov regularized solution is given by

$$
x^\gamma_{\text{Tikhonov}} = (A^T A + \gamma \mathbb{I})^{-1} A^T b,
$$

(4)

where we write $\mathbb{I}$ for the identity matrix, which will always have its dimension clear from the context. Unlike Tikhonov’s, many regularization techniques are nonlinear. For example, maximum likelihood and penalized maximum likelihood methods are common for tomographic image reconstruction from emission data [2, 14, 17, 29, 30, 33]. In this case, the regularized approximation is the solution of a more general convex optimization problem, possibly of the form

$$
\min \|Mx\|_1 \\
\text{s.t.} \quad -\log \mathbb{P}(b|x) \leq \ell, \quad x \in \mathbb{R}^n
$$

(5)

where $M$ is some sparsifying decomposition (such as a wavelet transform) and $\mathbb{P}(\cdot|x)$ is the probability density function of the data for a given image $x$. The parameter now is $\ell > 0$, which determines our “tolerance to unlikelihood”, within which the (hopefully) sparsest solution in the transformed space will be selected. We will study the general situation where the regularized solution is given by

$$
x^\gamma := B_\gamma(b),
$$

(6)

where $B_\gamma : \mathbb{R}^m \mapsto \mathbb{R}^n$ is a function implicitly defined by the regularization method. Precise conditions on $B_\gamma$ for our techniques to be applicable depend on the noise model, and will be discussed later in the text, when appropriate.

Because confidence in the regularized solution $x^\gamma$ requires careful choice of $\gamma$, several techniques to accomplish this task have been developed and studied for linear regularization, under the assumptions of uncorrelated Gaussian or deterministic
Bregman-Risk Estimators: Application to Parameter Selection

Data from imaging technologies, however, often do not fit well under the pure Gaussian hypothesis, because signal detection for imaging techniques is usually performed as a photon-counting process, thereby leading to data that can be modeled as Poisson variables [53], possibly with added Gaussian noise from underlying signal amplifying electronics [49]. Methods for regularization parameter selection under Poissonian, mixed Poissonian-Gaussian and more general noise hypotheses have also been devised, including those designed for nonlinear regularization [6, 11, 20, 40].

Many parameter selection techniques rely on Unbiased Predictive Risk Estimators (UPREs) or Unbiased Risk Estimators (UREs), that is, computable unbiased estimators for

\[
\mathbb{E}\|A(x^\gamma) - A(x^*)\|^2_2 \text{ or } \mathbb{E}\|x^\gamma - x^*\|^2_2, \tag{7}
\]

where \(\mathbb{E} := \mathbb{E}_b\) is the probabilistic expectation over the random variable \(b\), on which \(x^\gamma\) depends. Such U(P)REs are minimized with respect to the regularization parameter to yield a selection rule [9, 18, 23]. When \(A\) is linear, the regularization is obtained via an ordered filter (encompassing a large class of linear regularization operators), and the noise is Gaussian, it can be shown that selecting the minimizer of certain unbiased estimators, such as the UPRE regularization parameter, relates to the URE [39] as the magnitude of the error diminishes. It is unclear what is the most general setting where this kind of strong relationship between UPREs and UREs will hold.

Notice that the averaged quantities in (7) are not directly computable and Stein’s Lemma [34, 50] is the tool which allows obtaining U(P)REs, originally for the i.i.d. Gaussian case, but later generalized for a variety of noise models [3, 20, 27, 38, 40, 43]. We will focus on predictive-type estimators, but it is possible to obtain projected UREs from these operators, at least in the linear model case [20].

Contributions of the paper The main contribution of the present paper is the introduction of attainable unbiased estimators for general risk functions, which include:

\[
\mathbb{E}D_f(A(x^*), A(x^\gamma)), \tag{8}
\]

for \(f : \mathbb{R}^n \mapsto \mathbb{R}\), where \(D_f\) is a Bregman divergence [13]:

\[
D_f(x, y) = f(x) - f(y) - \nabla f(y)^T(x - y). \tag{9}
\]

Bregman divergences are usually defined with strictly convex \(f\), in which case \(D_f(x, y) = 0 \iff x = y\) and \(D_f(x, y) \geq 0\) for every pair \(x\) and \(y\). If \(f\) is strictly convex but not differentiable, a subgradient [32] can be used instead of the gradient \(\nabla f\) in order to obtain the same properties.

Our techniques are not limited to provide estimates for \(\mathbb{E}D_f(A(x^*), A(x^\gamma))\) when \(f\) is convex. If \(f\) is not strictly convex, however, the result may not be as meaningful for the purpose of regularization parameter selection because it might happen that a
minimizer of $D_f(x, y)$ occurs when $x \neq y$, in which case the Bregman divergence fails to behave as a measure of separation between vectors. We discuss computationally effective implementations of the estimators and its application to parameter selection in tomographic image reconstruction. We also present numerical experiments in order to show the effectiveness of the methodology when applied to Total-Variation regularized reconstruction in computed tomography.

It is worth noticing that although there are generalizations of Stein’s Lemma to several noise models, all such results have been, to the best of our knowledge, applied to the approximate minimization of the averaged squared error of (7), with the exception of [41], where the averaged Kullback-Leibler (KL) divergence is considered, for the case where the parameter to be selected is the iteration number of a likelihood maximizing algorithm. Although the work of Massa and Bevenuto [41] was unknown to us during the preparation of the present manuscript, our results relate to those in [41] in an interesting manner, as we will discuss later.

A generalization of the unbiased quadratic or KL predictive estimators for quantities like (8) with (9) enables the use of different divergences [16], which may be more appropriate to the problem at hand. Our research is motivated by the successful use of several Bregman divergences in many contexts [4, 5, 15, 25, 48]. We, therefore, introduce a technique that enables the use of many different Bregman-based risk measures as of the form (8). In the numerical experiments, besides the classic mean-squared error, we have used Itakura-Saito and Kullback-Leibler related divergences as well.

Finally, we discuss the reasons behind the behaviour of the method under the viewpoint of the concentration of measure phenomenon. Although we do not prove any concentration inequality in this work, we do analyze what would be the consequences of such concentration phenomena to our method if they actually hold true, and present some preliminary numerical experimentation in order to illustrate the ideas.

2. Stein-Like Estimates

We will now rely on unbiased estimators for quantities of the form

$$\mathbb{E} \left[ h(b)^T \beta \right],$$

where $h : \mathbb{R}^m \mapsto \mathbb{R}^m$ and $\mathbb{E} b = \beta$. For that, the knowledge of the probabilistic laws for $b$ is required. For example, manyfold application of Stein’s Lemma [50, Lemma 2] leads to the following, where $b \sim \mathcal{N}(\beta, \sigma^2 I)$ denotes a vector $b$ of independent random variables such that each of its components $b_i$ is normally distributed with mean $\beta_i$ and variance $\sigma^2$:

**Lemma 1.** Let $\beta \in \mathbb{R}^m$ and $b \sim \mathcal{N}(\beta, \sigma^2 I)$ and consider $h : \mathbb{R}^m \mapsto \mathbb{R}^m$ such that $h$ is weakly differentiable and, for $i \in \{1, 2, \ldots, m\}$, $\mathbb{E} \left| \frac{\partial h_i}{\partial b_i}(b) \right| < \infty$. Then

$$\mathbb{E} \left[ h(b)^T (b - \beta) \right] = \sigma^2 \mathbb{E} \left[ \sum_{i=1}^{m} \frac{\partial h_i}{\partial b_i}(b) \right].$$

(11)
We will now apply this result to the nonlinear cases (2) and (6) with expected Bregman divergence $D_f$ as a risk measure. First we rewrite:

$$
D_f(A(x^*), A(x^)) = f(A(x^*)) - f(A(x^)) - \nabla f(A(x^*))^T (A(x^*) - A(x^))
$$

$$
= f(A(x^*)) - f(A(x^)) - \nabla f(A(x^*))^T (b - A(x^*)) + \nabla f(A(x^*))^T (b - A(x^*))
$$

$$
= f(A(x^*)) - f(b) + D_f(b, A(x^*)) + \nabla f(A(x^*))^T (b - A(x^*)). \tag{12}
$$

Then we can prove the following result.

**Proposition 1.** Suppose $b \sim N(A(x^*), \sigma^2 I)$ and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $B_\gamma : \mathbb{R}^m \mapsto \mathbb{R}^n$, $x^* \in \mathbb{R}^n$ be given. Define $x^\gamma := B_\gamma(b)$ and denote

$$
g_\gamma := \nabla f \circ A \circ B_\gamma. \tag{13}
$$

Assume that $f$, $A$ and $B_\gamma$ are such that $g_\gamma$ as defined in (13) is weakly differentiable, $\mathbb{E} f(b) < \infty$, $\mathbb{E} D_f(b, A(x^\gamma)) < \infty$ and that for $i \in \{1, 2, \ldots, m\}$, $\mathbb{E}_{b_i} \left| \frac{\partial g}{\partial b_i}(b) \right| < \infty$. Then we have:

$$
\mathbb{E} D_f(A(x^*), A(x^\gamma)) = K + \mathbb{E} D_f(b, A(x^\gamma)) + \sigma^2 \mathbb{E} \left[ \sum_{i=1}^m \frac{\partial g_i}{\partial b_i}(b) \right], \tag{14}
$$

where $K$ is a constant independent of $\gamma$.

**Proof.** Let $K := f(A(x^*)) - \mathbb{E} f(b)$, then computing expectations on both sides of (12) we have

$$
\mathbb{E} D_f(A(x^*), A(x^\gamma)) = K + \mathbb{E} D_f(b, A(x^\gamma)) + \mathbb{E} \left[ \nabla f(A(x^\gamma))^T (b - A(x^*)) \right]
$$

$$
= K + \mathbb{E} D_f(b, A(x^\gamma)) + \mathbb{E} \left[ g_\gamma(b)^T (b - A(x^*)) \right].
$$

Using Lemma 1 to replace the last term on the right gives the desired result. \qed

The above proposition shows that, from the viewpoint of obtaining an estimator to the average risk for nonlinear models and/or reconstruction techniques, applying Stein’s Lemma to a more general risk function has the same difficulty of applying it to the Mean Squared Error (MSE). Furthermore, there seems to exist compelling reasons to use other risk measures [16]. Indeed, a varied set of Bregman divergences have been successfully used in several applications, such as principal component analysis [15]; on-line density estimation [4]; machine learning [5, 48], and speech processing [25].

Let us now examine the Poisson case. Suppose that $b$ is Poisson distributed with mean $\beta$. We denote this as $b \sim \mathcal{P}(\beta)$. Also, if $b$ is a vector of random variables such that $b_i \sim \mathcal{P}(\beta_i)$, we simplify the notation by $b \sim \mathcal{P}(\beta)$. Now, let $b \sim \mathcal{P}(\beta)$ and $h : \mathbb{R} \mapsto \mathbb{R}$ be such that $\mathbb{E}_{b}[h(b)] < \infty$, then we have [46]:

$$
\mathbb{E}_b[\beta h(b)] = \mathbb{E}_b[\beta h(b - 1)]. \tag{15}
$$

This equation can be used to prove the following result [40, Property 2]:
Lemma 2. Let $\beta \in \mathbb{R}_+^m$, $b \sim \mathcal{P}(\beta)$ and consider $h : \mathbb{R}^m \rightarrow \mathbb{R}^m$ such that for $i \in \{1, 2, \ldots, m\}$, $E[h_i(b)] < \infty$ and $E[h(b)^Tb] < \infty$. Then

$$E[h(b)^T(b - \beta)] = E\left[b^T(h(b) - h^{[-1]}(b))\right],$$

with $h^{[\xi]}$, for $\xi \in \mathbb{R}$, given componentwise as

$$h_i^{[\xi]}(b) := h_i(b + \xi e^i),$$

where $e^i$ denotes the $i$-th column of the $m \times m$ identity matrix.

Now using Lemma 2 instead of Lemma 1, we have the following result, the proof of which we omit for similarity with the proof of Proposition 1.

Proposition 2. Suppose $b \sim \mathcal{P}(A(x^*))$ and let $f$, $A$, $x^*$, $x^\gamma$, and $g_\gamma$ be as in Proposition 1. Assume $g_\gamma^{[-1]}$ follows the notation of (17). Further assume that $f$, $A$ and $B_\gamma$ are such that $E f(b) < \infty$, $ED_f(b, A(x^\gamma)) < \infty$, $E[g_\gamma(b)^Tb] < \infty$ and that for $i \in \{1, 2, \ldots, m\}$, $E[h_i(b)] < \infty$. Then we have

$$ED_f(A(x^*), A(x^\gamma)) = K + ED_f(b, A(x^\gamma)) + E[b^T(g_\gamma(b) - g_\gamma^{[-1]}(b))],$$

where $K$ is a constant independent of $\gamma$.

By now, the pattern has hopefully become evident to the reader. The idea is that given an unbiased, computable from the data, estimator for

$$E \left[ \nabla f(A(x^\gamma))^T (b - A(x^*)) \right],$$

one can straightforwardly obtain, up to a constant, an estimator for $ED_f(A(x^*), A(x^\gamma))$ by taking (12) in consideration. Practical estimators for quantities such as (19) exist for a variety of noise models. For example, we can mention papers [40, 43] for the mixed Poisson-Gaussian case; [20] for the exponential family case (which includes Gaussian, Poisson, binomial, gamma and inverse Gaussian distributions), and [27, 38] for elliptically distributed errors.

2.1. Computation of Stein-Like Estimators

While equations (14) and (18) do not rely on the unattainable quantities $x^*$ or $A(x^*)$, both pose computational difficulties. In formula (14) there is the need to compute the partial derivatives $\partial g_i / \partial x_i$, which depend on the derivatives of the reconstruction method. In most cases, no analytical expression for these derivatives will be available. Even for linear reconstruction methods, such as the Filtered BackProjection (FBP) algorithm, analytical expressions are not available and Monte-Carlo techniques will likely be used, as in [43, 47]. These approaches take advantage of the large number of terms in the summation and use random vectors to estimate the trace of the Jacobian with good accuracy and relatively low computational cost. Options such as numerically
approximating each of the summands by finite differences are unfeasible as they would require the solution of a large number of related problems.

The Monte-Carlo principle we use is given by the equality [43, 47]:

\[
\mathbb{E}_{\omega} \left[ \lim_{\epsilon \to 0} \frac{1}{\epsilon} \omega^T \text{diag}(z) \left( g_\gamma(b + \epsilon \omega) - g_\gamma(b) \right) \right] = z^T \partial g_\gamma(b),
\]  

(20)

where \( z \in \mathbb{R}^m \), \( \omega \in \mathbb{R}^m \) is such that \( \mathbb{E}_{\omega} \omega = 0 \) and \( \mathbb{E}_{\omega} \omega \omega^T = I \), and \( \partial g_\gamma \) is defined componentwise as

\[
\partial_i g_\gamma = \frac{\partial g_i}{\partial b_i}.
\]

(21)

Therefore, if the noise model is Gaussian, one could define the following estimator

\[
\text{G-UPBRE}_{\epsilon}^f(\gamma) := D_f(b, A(x^\gamma)) + \frac{\sigma^2}{\epsilon} \omega^T \left( g_\gamma(b + \epsilon \omega) - g_\gamma(b) \right),
\]

(22)

where UPBRE stands for Unbiased Predictive Bregman Risk Estimators. Thus, because of (14) and (20), we have

\[
\mathbb{E}_{b, \omega} \left[ \lim_{\epsilon \to 0} \text{G-UPBRE}_{\epsilon}^f(\gamma) \right] = \mathbb{E} D_f(A(x^\gamma), A(x^\gamma)) - K,
\]

(23)

where \( K \) does not depend on \( \gamma \).

A discussion on the selection of the discretization parameter \( \epsilon \) to be used in a practical approximation of the limit inside the expectation on the left-hand side of (20) can be found, for example, in [43, 47]. In both references, it has been found experimentally that under weak differentiability hypothesis, the approximation

\[
\frac{1}{\epsilon} \omega^T \text{diag}(z) \left( g_\gamma(b + \epsilon \omega) - g_\gamma(b) \right) \approx z^T \partial g_\gamma(b)
\]

appears to hold consistently within a wide range of values of \( \epsilon \). Interestingly, for non-differentiable regularization methods, there still seems to be a (narrower) range for \( \epsilon \) where the technique provides useful results, even without theoretical backup. In [43], it was also shown that the most favorable probability distribution for \( \omega \), in the sense that it minimizes the variance of the resulting estimator on the left-hand side of (24) for vanishing \( \epsilon \), is the one for which the components \( \omega_i \) of \( \omega \) are independently distributed with \( \mathbb{P}(\{\omega_i = -1\}) = \mathbb{P}(\{\omega_i = 1\}) = 1/2 \).

Now assume a Poissonian noise model. With \( \omega \) as before, let us then define

\[
\text{P-UPBRE}_{\epsilon}^f(\gamma) := D_f(b, A(x^\gamma)) + \frac{1}{\epsilon} \omega^T \text{diag}(b) \left( g_\gamma(b + \epsilon \omega) - g_\gamma(b) \right).
\]

(25)

Therefore, application of (20) leads to

\[
\mathbb{E}_{\omega} \left[ \lim_{\epsilon \to 0} \text{P-UPBRE}_{\epsilon}^f(\gamma) \right] = D_f(b, A(x^\gamma)) + b^T \partial g_\gamma(b).
\]

(26)
Furthermore, notice that a first-order Taylor expansion for the last term in (18) yields
\[ \mathbb{E} [b^T (g_\gamma(b) - g_\gamma^{-1}(b))] \approx \mathbb{E} [b^T \partial g_\gamma(b)]. \] (27)

Finally, computing the expectation with respect to \( b \) in both sides of (26), taking (27) into consideration, and then using (18), we get:
\[ \mathbb{E}_{b,\omega} \left[ \lim_{\epsilon \downarrow 0} \text{P-UPBRE}_\epsilon^f (\gamma) \right] \approx \mathbb{E} D_f(A(x^\gamma), A(x^\gamma)) - K, \] (28)
which is an approximate result, unlike (23). The approximation error should be relatively small since for Poisson random variables unity perturbations are likely to be small relatively to the size of the perturbed variables. Accordingly, numerical experimentation has found the approximation (27) to be accurate enough for practical applications [43].

It is not the purpose of the present paper to go through all the noise models possibly covered by the technique. Instead, we will focus our experimental work in the Poissonian case, which is the dominant noise type, e.g., in emission tomography. It is necessary, however, to notice that more sophisticated circumstances, such as a combination of Poissonian and Gaussian noise models, may lead to complications in the numerical computation of unbiased Stein-like estimators other than the simple first-derivative trace estimation. On the other hand, this issue has already been previously addressed within reasonable detail in the literature [43] and should not be too much of a concern to the practitioner.

3. Numerical Experimentation

3.1. The Radon Transform

Tomography is the production of cross-sectional images of objects in a minimally invasive manner. Several techniques have been devised in order to achieve this goal, many of which are modeled via the so-called Radon Transform (RT). The RT of a function \( \eta : \mathbb{R}^2 \to \mathbb{R} \), denoted as \( \mathcal{R}[\eta] \) is defined as:
\[ \mathcal{R}[\eta](\theta, t) := \int_{\mathbb{R}} \eta \left( t \left( \frac{\cos \theta}{\sin \theta} \right) + s \left( \frac{\sin \theta}{\cos \theta} \right) \right) ds. \] (29)

A well known example of tomographic reconstruction technique that can be modeled with the help of the Radon transform is X-ray Computed Tomography (XCT). Experiments involving XCT are presented below. Both synthetic and real world datasets are used.

3.2. XCT

Our transmission datasets were collected at the Brazilian Synchrotron Light Laboratory (LNLS). In this kind of setup, three measurements are made for each path \( L \):
Bregman-Risk Estimators: Application to Parameter Selection

- $I_{\text{dark}}(L)$: expected number of photons detected with the source turned off;
- $I_{\text{flat}}(L)$: expected number of photons detected with the source turned on but without object between source and detector;
- $N_{\text{count}}(L)$: number of photons detected with the source turned on and with the object between source and detector.

Photons detected during the dark scan are assumed to be part of the background radiation and are, therefore, detected in addition to the photons generated by the source. The model becomes

$$I_{\text{count}}(L) = I_{\text{flat}}(L) e^{-\int L \mu ds} + I_{\text{dark}}(L).$$

(30)

Notice that $I_{\text{flat}}(L)$ and $I_{\text{dark}}(L)$ do not involve the imaged object and can thus be estimated rather accurately. This is not true, however, for $I_{\text{count}}(L)$ and the Poisson random variable $N_{\text{count}}(L) \sim \mathcal{P}(I_{\text{count}}(L))$ is measured instead.

Using this principle, we can estimate the RT by

$$R[\mu](\theta_i, t_i) \approx -\log \left( \frac{N_{\text{count}}(L_i) - I_{\text{dark}}(L_i)}{I_{\text{flat}}(L_i)} \right),$$

(31)

where $(\theta_i, t_i)$ parametrize the $i$-th line $L_i$ from source to detector according to the definition of the RT. To be precise,

$$L_i := \{ t_i \left( \frac{\cos \theta_i}{\sin \theta_i} \right) + s \left( \frac{-\sin \theta_i}{\cos \theta_i} \right) : s \in \mathbb{R} \}. $$

(32)

Assuming the original image $\mu : \mathbb{R}^2 \to \mathbb{R}_+$ lies in a finite dimensional vector space generated by some basis $\{\mu^1, \mu^2, \ldots, \mu^n\}$, then it can be written as $\mu = \sum_{j=1}^n x_j \mu^j$. Noticing, moreover, that the number of measurements is always finite in practice, one can reduce the problem of tomographic reconstruction to a linear system of equations:

$$Rx = y,$$

(33)

where $x = (x_1, x_2, \ldots, x_n)^T$, the matrix $R$ is given componentwise by

$$r_{ij} = R[\mu^j](\theta_i, t_i),$$

(34)

and the elements $y_i$ of $y$ are the corresponding Radon data, that is, $y_i = R[\mu](\theta_i, t_i)$. In practice, the above linear system of equations will be replaced by

$$Rx = \tilde{y},$$

(35)

where $\tilde{y}$ is the experimentally obtained data.

The set of sampled Radon coordinates $(\theta_i, t_i)$ was as follows. Let

$$T := \left\{ -1, -1 + \frac{2}{2047}, -1 + 2 \frac{2}{2047}, -1 + 3 \frac{2}{2047}, \ldots, -1 + 2046 \frac{2}{2047}, 1 \right\}$$

(36)
Bregman-Risk Estimators: Application to Parameter Selection

and

\[ \Theta := \left\{ 0, \frac{\pi}{512}, 2 \frac{\pi}{512}, 3 \frac{\pi}{512}, \ldots, 511 \frac{\pi}{512} \right\}. \tag{37} \]

Then

\[ \{ (\theta_1, t_1), (\theta_2, t_2), \ldots, (\theta_m, t_m) \} = \Theta \times T. \tag{38} \]

The data acquisition was not performed directly through a photon counting sensor. Instead, a scintillator crystal [1] was used, which had a photon yield at 300 K (27\(^\circ\)C) of 25 Ph/KeV. The mean energy of the X-ray photons of the UVX line was around 13 KeV. The optical coupling between the CCD detector and the scintillator achieved a 2% photon detection rate. Ignoring the crystal’s finite temporal resolution, using an idealized monochromatic model for the light source, and ignoring the CCD’s electric noise, data can be corrected dividing the CCD’s visible light photon count by 6.5 in order to estimate the X-ray photon count, which is a Poisson variable to which our methodology can be applied. We could also have used a Poissonian plus Gaussian model [40] in the same way we have used the pure Poissonian model, but we wanted to keep our focus on the comparison between different Bregman divergences instead of on the noise modeling.

3.3. Reconstruction

Our discretization of the image space \([-1, 1]^2\] was in a grid of 512 \times 512 square pixels. This intuitive non-overlapping basis of square pixels allows for reasonably efficient implementations of the matrix-vector products of the form \(Rx\) and \(R^T y\).

With the discretization of the problem fully defined by the basis functions and sampling scheme, we estimate the pixel values \(x_i\) by solving

\[
\min_{x \in \mathbb{R}^2_n} \frac{1}{2} \| R x - \tilde{y} \|_2^2 + \gamma TV(x), \tag{39}
\]

with \(\gamma \geq 0\) and \(TV\) being the total variation

\[
TV(x) := \sum_{i=1}^n \sum_{j=1}^n \sqrt{(x_{i,j} - x_{i,j-1})^2 + (x_{i,j} - x_{i-1,j})^2}, \tag{40}
\]

where we again have used the lexicographic identification \(x_{i,j} = x_{512(i-1)+j}\) and, by convention, the boundary condition \(x_{0,j} = x_{i,0} = 0\).

We end summarizing the reconstruction method. For that, let us first denote:

- \(b := \left( N_{\text{count}}(L_1), N_{\text{count}}(L_2), \ldots, N_{\text{count}}(L_{512 \times 2048}) \right)^T; \)
- \(f := \left( I_{\text{flat}}(L_1), I_{\text{flat}}(L_2), \ldots, I_{\text{flat}}(L_{512 \times 2048}) \right)^T; \)
- \(d := \left( I_{\text{dark}}(L_1), I_{\text{dark}}(L_2), \ldots, I_{\text{dark}}(L_{512 \times 2048}) \right)^T. \)

Then, the steps for computing \(B_\gamma(b)\) are
(i) Compute \( \tilde{y}_i = -\log \left( \frac{b_i - d_i}{f_i} \right) \) for all \( i \in \{1, 2, \ldots, 512 \times 2048\} \);

(ii) Return the minimizer of (39). The Fast Iterative Soft-Thresholding Algorithm (FISTA) [8] was used to obtain the numerical minimizer.

Because the input of this method is a vector of independent Poisson variables, we should be able to apply the estimator \( \text{P-UPBRE}^f \) developed above in order to find an estimate of the optimal value for the regularization parameter \( \gamma \). The final ingredient is the forward operator \( A \) which is given componentwise by

\[
A_i(x) = d_i + f_i e^{-(R x)_i},
\]

since this is the expected photon count over line \( L_i \) for the image \( \sum_{j=1}^n x_j \mu^j \).

3.4. Bregman Functions

Throughout the numerical experimentation, we have tried three different Bregman divergences, two of which are modified versions of the Itakura-Saito [22] and the Kullback-Leibler [37] divergences. These divergences are obtained using the following Bregman functions, respectively:

\[
\begin{align*}
    f_{ms}(x) &:= \sum_{i=1}^n x_i^2, \\
    f_{kl}(x) &:= \sum_{i=1}^n x_i \ln(x_i), \quad \text{and} \quad f_{is}(x) := -\sum_{i=1}^n \ln(x_i).
\end{align*}
\]

Furthermore, we define

\[
\ln(x) := \begin{cases} 
    \ln(x) & x \geq \varepsilon \\
    \ln(\varepsilon) + \frac{1}{\varepsilon}(x - \varepsilon) - \frac{1}{2\varepsilon^2}(x - \varepsilon)^2 & x < \varepsilon
\end{cases}.
\]

We have used \( \varepsilon = 10^{-1} \) in all our experiments.

The second order approximation \( \ln(x) \) for the logarithm near the negative orthant was used because it is not possible to use \( \ln(x) \) directly, as in the original definition of both the Itakura-Saito and Kullback-Laibler divergences, due to difficulties in the boundary of the domain of definition of the resulting functionals. The original version of the Itakura-Saito divergence, for example, reads

\[
\sum_{i=1}^n \left\{ \frac{x_i}{y_i} - \ln \left( \frac{x_i}{y_i} \right) - 1 \right\},
\]

which is not well defined if either \( x_i = 0 \) or \( y_i = 0 \) for some \( i \in \{1, 2, \ldots, n\} \), a common situation in, e.g., emission tomography imaging tasks, where the data will likely contain several components \( b_i = 0 \).
Figure 1. Plots of the estimator $P\text{-UPBRE}_f^\epsilon(\gamma)$ over a range of values of $\gamma$ for three different values of $\epsilon$ and three different functions $f$. Top: $P\text{-UPBRE}_f^{ms}(\gamma)$. Center row: $P\text{-UPBRE}_f^{kl}(\gamma)$. Bottom: $P\text{-UPBRE}_f^{is}(\gamma)$. Left: slice shown in the left column of Figure 2. Right: slice shown in the center column of Figure 2.

3.5. Real world data results

In this subsection we will reconstruct images of slices of an apple seed scanned at the UVX tomography line of the LNLS. This first experiment is intended to assess the behaviour of the method with respect to changes in the numerical differentiation parameter $\epsilon$ and to changes in the Bregman function. Also, it serves as a proof of concept of the methodology applied to a realistic situation.

We first compute $P\text{-UPBRE}_f^\epsilon(\gamma)$ for $\epsilon \in \{10^{-2}, 10^{-1}, 10^0\}$ and for 33 values of
Figure 2. From top to bottom: images reconstructed with no regularization; images reconstructed with the regularization parameter set as the minimizer of $P \cdot \text{UPBRE}_{f_{10^{-1}}(\gamma)}$ over the tested values of $\gamma$; images reconstructed with the regularization parameter set as the minimizer of $P \cdot \text{UPBRE}_{f_{10^{-1}}}(\gamma)$ over the tested values of $\gamma$; images reconstructed with the regularization parameter set as the minimizer of $P \cdot \text{UPBRE}_{f_{10^{-1}}(\gamma)}$ over the tested values of $\gamma$. Left and center: reconstructions of different slices of an apple seed. Right: detail of the center reconstruction.
γ logarithmically spaced in the range \([10^{-5}, 3 \cdot 10^{-4}]\). This range was selected for best visualization of the most relevant region of the domain. The value of the estimator rapidly increases outside the displayed range, which is good, for example, for numerical minimization tasks. Each of the values for the discretization parameter ε gives rise to a curve \((γ, P-UPBRE_ε^f(γ))\). In Figure 1 we see plots of these curves for \(ε \in \{10^{-2}, 10^{-1}, 10^0\}\) (grouped in the same graphic) for all the functions \(f_{ms}, f_{kl}, \) and \(f_{is}\) and for two different slices of the apple seed. There we can see that the method is reasonably robust to the choice of the numerical differentiation parameter. When this parameter becomes too small, an oscillation behavior due to numerical and floating point errors is noticeable in the curve. Tuning of the parameter can be done by gradually increasing the parameter until the oscillatory behaviour is eliminated.

The minimizer of P-UPBRE_{10^{-1}}^f should be a sound choice for the regularization parameter for the tomographic reconstruction problem. Figure 2 shows that images reconstructed using such a minimizer as the regularization parameter indeed present a good balance between noise-removal and feature retention. Some of the artifacts seen in the images are from imperfections in the acquisition setup, such as the ring-shaped artifacts and the streaks. These are not supposed to be eliminated by the regularization. The noise, on the other hand, should be reduced. This can indeed be seen to be the case.

A major contribution of the present paper is to generalize the idea of P-UPBRE_{10^{-1}}^f to more general Bregman divergences. This is why we have reconstructed images from the same datasets using the minimizers of P-UPBRE_{10^{-1}}^{f_{ms}}, P-UPBRE_{10^{-1}}^{f_{kl}}, \) and P-UPBRE_{10^{-1}}^{f_{is}} as regularization parameters. It is possible to notice that even under the coarse sampling of the parameter space that we have used, there seems to be some noticeable, although not very large, differences among the selected regularization parameters. The next set of experiments is designed to make sense of these differences in a more controlled setting, where comparisons against a ground truth, made possible by the simulated nature of the experiment, are made.

### 3.6. Simulated data results

For the following set of experiments, data was simulated. The dark and flat scans were taken from the first seed slice dataset described above, but the count was simulated as a random Poisson variable:

\[
N_{\text{count}}(L_i) \sim \mathcal{P} \left( I_{\text{flat}}(L_i) e^{-R[\mu^\dagger](\theta_i, t_i)} + I_{\text{dark}}(L_i) \right),
\]

where \(\mu^\dagger\) is a 2048 x 2048 discretization of the Shepp-Logan [36] head phantom. Because the reconstruction used a 512 x 512 discretization, the inverse crime [35] is avoided.

Each experiment consisted of generating the dataset, then minimizing P-UPBRE_{f_ε}(γ) for some fixed \(f\) and \(ε\), then minimizing \(∥x^γ - x^*∥_2\) over \(γ\), and then minimizing \(D_f(A(x^*), A(x^γ))\). This was repeated 20 times for each pair \((f, ε)\). The
result is summarized in Figure 3, where it can be seen that the regularization parameter does seem to somehow depend on the function used for the Bregman divergence.

For ease of exposition, let us denote

$$
\gamma^* := \arg \min_{\gamma} \| x^* - x^{\gamma} \|, \quad \gamma_f^* := \arg \min_{\gamma} D_f(x^*, x^{\gamma}),
$$

and

$$
\gamma_{f,\epsilon}^* := \arg \min_{\gamma} \text{P-UPBRE}_\epsilon^f(\gamma).
$$

Notice that these are minimization problems in one variable. Moreover, because we are performing simulated experiments where $x^*$ is known, the objective function is easily computable. Therefore, the approximation of the above minimizers is not a complicated task and we have used the one-dimensional improved golden-section method as implemented in the minimize_scalar routine of the SciPy package.

These minimizers happen to be random variables that concentrate rather tightly around its median, as shown in the experiments. We have repeated the simulated data
Bregman-Risk Estimators: Application to Parameter Selection

Figure 4. Left: relative differences $|\gamma^* - \gamma^*_f,\epsilon|/\gamma^*$ between each minimizer $\gamma^*_f,\epsilon$ of P-UPBRE$L_\epsilon(\gamma)$ and the respective minimizer $\gamma^*$ of the squared error $\|x^\gamma - x^*\|^2_2$. Right: relative differences $|\gamma^* - \gamma^*_{f,\epsilon}|/\gamma^*$ between each minimizer $\gamma^*_{f,\epsilon}$ of P-UPBRE$L_\epsilon(\gamma)$ and the respective minimizer $\gamma^*$ of the predictive Bregman error $D_f(\mathbf{A}(x^*), \mathbf{A}(x^\gamma))$.

generation 20 times for each pair

$$(f, \epsilon) \in \{(f_{\text{ms}}, 10^{-1}), (f_{\text{ms}}, 10^{-2}), (f_{\text{kl}}, 10^{-1}), (f_{\text{ls}}, 10^{-1})\}$$

and we numerically computed $\gamma^*$, $\gamma^*_f$, and $\gamma^*_f,\epsilon$ for each of these simulated datasets. Figure 3 brings boxplots of the minimizers. We can observe that in fact there is not much variation, which is a useful property as we shall see in the next section.

For now we would like to focus on the fact that the minimizers $\gamma^*_f,\epsilon$ are consistently below the minimizers $\gamma^*_f$, which is not surprising as the numerical differentiation scheme will affect the smoothness of the finite difference part of the estimator. Noticeably, in this example $\gamma^*_f,\epsilon$ tends to be larger than $\gamma^*$, which compensates for this fact and makes $\gamma^*_f,\epsilon$ the most accurate of the estimators for the optimal regularization parameter $\gamma^*$ as can be seen on the left of Figure 4. This is so even though the difference $|\gamma^*_{f,\epsilon} - \gamma^*_{f,10^{-1}}|$ is larger than $|\gamma^*_{f} - \gamma^*_{f,10^{-1}}|$ for the other $f$, which can be seen on the right of Figure 4.

In order to verify these results with other images, discretization and noise setups, we have reconstructed a mouse head slice from the high-resolution tomographic atlas published in [42]. This was done with data simulated in three different ways, two of which purposely not trying to avoid the inverse crime. The other reconstruction of the mouse head slice used the same setup as before with the Shepp-Logan phantom. Finally, we have also reconstructed the Shep-Logan phantom, this time simulating noisier data.

The results can be seen in Figure 7, where the minimization of the estimator based on the Itakura-Saito divergence seems to consistently have a slightly superior accuracy. The only exception might be a statistical artifact of the small sample size of 10 simulations for each combination of noise level, image, and discretization. Figure 6 shows some of the reconstructions that were obtained by the minimization of some of the proposed estimators. Figure 5 shows the original images we chose to reconstruct and the “best” reconstruction.
Notice that in the case of the mouse head, the reconstruction obtained using \( \gamma \) that minimizes \( \| x^* - x^\gamma \|_2^2 \) is noticeably smoother than those obtained minimizing P-UPBRE_\epsilon. The reason for this seems to be that the image \( x^* \) used in the comparison is in fact a smoothed version of \( x^\dagger \), the one used to generate the data, because the resolution of \( x^* \) is lower than the resolution of \( x^\dagger \). This means that the model discrepancy confuses the estimator, which takes only the noise model into consideration. It appears that our technique allows for some of the systematic error introduced by the model inaccuracy to be reduced by minimizing a different Bregman divergence instead of the squared norm.

It is interesting to observe that our methodology generalizes the approach of [41] in the sense that if we plug the KL divergence in our general approach we get the same estimator obtained in [41], both before and after applying the Monte-Carlo procedure (which we took from [40]). It should be noted, however, that stronger claims are proven in [41] about the particular regularization technique being used, which provide further insight on the nature of the approximation. Indeed, they have shown that, as the Poisson parameters increase (i.e., the relative noise level decreases), approximation (27) becomes more accurate for that particular regularization technique (early stopping of the EM algorithm).

Finally, we remark that the experiments we have presented regarding transmission tomography are unique because they deal with a model where the acquired data is a nonlinear function of the Poisson variables. Our experiments with simulated and real data show that predictive error-based are flexible and robust enough to cope with such circumstances.

4. Concentration of Measure

Thanks to the efforts of many, from the pioneering insights of Vitali Milman to the refined work of Michel Talagrand, the initial non-asymptotic concentration results of Chernoff and of Hoeffding evolved to the following general idea [51]:

“A random variable that depends (in a ‘smooth’ way) on the influence of many independent variables (but not too much on any of them) is essentially constant.”

In order to understand the relevance of this phenomenon to the techniques described in the present paper, in what follows we will consider the consequences of concentrations of two different random variables that fit the description “depends (in a ‘smooth’ way) on the influence of many independent variables (but not too much on any of them)” to our methodology.

First we notice that there is in principle no reason to believe that \( \mathbb{E}_b D_f\left( A(x^*), A(B_\gamma(b)) \right) \) would be close to \( D_f\left( A(x^*), A(B_\gamma(b)) \right) \) for a given \( b \). As it is known, of course, the ultimate goal should be to minimize \( D_f\left( A(x^*), A(B_\gamma(b)) \right) \), not \( \mathbb{E}_b D_f\left( A(x^*), A(B_\gamma(b)) \right) \). However, in many practical applications (such as
imaging problems), \( D_f(\mathbf{A}(\mathbf{x}^*), \mathbf{A}(\mathbf{B}_\gamma(\mathbf{b}))) \) is naturally smoothly dependent on several independent random variables (such as millions of data pixels \( \mathbf{b} \)) but not too much on any of them (each data pixel \( b_i \) has limited influence in the final result). That is, the concentration of measure principle states that the probability that we have large

\[
\left| \mathbb{E}_b D_f(\mathbf{A}(\mathbf{x}^*), \mathbf{A}(\mathbf{B}_\gamma(\mathbf{b}))) - D_f(\mathbf{A}(\mathbf{x}^*), \mathbf{A}(\mathbf{B}_\gamma(\mathbf{b}))) \right|
\]

is exponentially small.

In practice, we do not compute \( \mathbb{E}_b D_f(\mathbf{A}(\mathbf{x}^*), \mathbf{A}(\mathbf{B}_\gamma(\mathbf{b}))) \) directly. Instead, we make use of (14) or (18) because the quantities on the right-hand side of these equations are in principle computable without knowledge of the exact solution \( \mathbf{x}^* \). Indeed, we do not even use \( \mathbb{E} \left[ \mathbf{D}_f(\mathbf{b}, \mathbf{A}(\mathbf{x}^*)) + \sigma^2 \mathbb{E} \left[ \sum_{i=1}^{m} \frac{\partial g_i}{\partial b_i}(\mathbf{b}) \right] \right] \) (we will use the Gaussian case for concreteness, but the discussion applies to the other cases too). Instead, we again reasonably assume that the probability is very small that

\[
\left| \mathbb{E} \left[ \mathbf{D}_f(\mathbf{b}, \mathbf{A}(\mathbf{x}^*)) + \sigma^2 \mathbb{E} \left[ \sum_{i=1}^{m} \frac{\partial g_i}{\partial b_i}(\mathbf{b}) \right] - D_f(\mathbf{b}, \mathbf{A}(\mathbf{x}^*)) - \sigma^2 \sum_{i=1}^{m} \frac{\partial g_i}{\partial b_i}(\mathbf{b}) \right] \right|
\]

is large. Thus, taking (14) into consideration, we should expect that there is only a small probability that the following difference is large

\[
D_f(\mathbf{A}(\mathbf{x}^*), \mathbf{A}(\mathbf{B}_\gamma(\mathbf{b}))) - K - D_f(\mathbf{b}, \mathbf{A}(\mathbf{x}^*)) - \sigma^2 \sum_{i=1}^{m} \frac{\partial g_i}{\partial b_i}(\mathbf{b}).
\]
Figure 6. From top to bottom: images reconstructed with no regularization; images reconstructed with the regularization parameter set as the numerical minimizer of $P\text{-UPBRE}_{10^{-1}}^{f_{ms}}(\gamma)$; images reconstructed with the regularization parameter set as the numerical minimizer of $P\text{-UPBRE}_{10^{-1}}^{f_{kl}}(\gamma)$; images reconstructed with the regularization parameter set as the numerical minimizer of $P\text{-UPBRE}_{10^{-1}}^{f_{is}}(\gamma)$. Left: Shepp-Logan phantom. Center: mouse head. Right: detail of the center image.
This is an admittedly vague affirmation. Indeed, we do not have the goal to provide concrete concentration inequalities here. Instead, our analysis will start from the following question: assuming that some function $\varphi : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ somehow concentrates around $\mathbb{E}_b \varphi(b, \gamma)$, does the minimizer $\gamma^*$ of $\varphi(b, \gamma)$ concentrate around the minimizer $\gamma^*$ of $\mathbb{E}_b \varphi(b, \gamma)$? In order to further simplify the analysis we will assume a discrete parameter space, leaving the continuous case for future research.

Since we will be mainly concerned with $\varphi(b, \gamma)$ as a function of $\gamma$, we denote $\varphi_b(\gamma) := \varphi(b, \gamma)$ from now on. We will assume that, for each $b$, the function $\varphi_b$ is locally Lipschitz continuous on the variable $\gamma$. Moreover, assume that there exists $\lambda > 0$ such that the function $\mathbb{E}_b \varphi_b : \mathbb{R}_+ \to \mathbb{R}$ satisfies:

$$\mathbb{E}_b \varphi_b(\gamma) \geq \mathbb{E}_b \varphi_b(\gamma^*) + \frac{\lambda}{2} \|\gamma - \gamma^*\|_2^2, \quad \text{for all } \gamma \geq 0.$$  (50)
A sufficient condition for the above inequality to hold is strong convexity of $E\varphi_b(\gamma)$ with respect to $\gamma$, but this is not necessary, what is really required is that the function $\Phi(\gamma) := E\varphi_b(\gamma)$ is not too “flat” close to its (unique) minimizer $\gamma^*$. The following form of this inequality will be more frequently used below:

$$E\varphi_b(\gamma^*) - E\varphi_b(\gamma) \leq -\frac{\lambda}{2}||\gamma - \gamma^*||^2, \quad \text{for all } \gamma \geq 0. \quad (51)$$

We will assume that $\varphi_b(\gamma)$ satisfies, uniformly over all possible values of $\gamma$, the following concentration inequalities for small enough $t > 0$:

$$\mathbb{P}[\varphi_b(\gamma) > E\varphi_b(\gamma) + t] \leq \exp\left(-\frac{t^2}{4V}\right) \quad (52)$$

and

$$\mathbb{P}[\varphi_b(\gamma) < E\varphi_b(\gamma) - t] \leq \exp\left(-\frac{t^2}{4V}\right) \quad (53)$$

for some $V > 0$. For conditions that might ensure the validity of such inequalities, see, e.g., [19, Theorem 8.2].

Let us then compute a bound on the probability that $\varphi_b(\gamma)$ is larger than $E\varphi_b(\gamma^*)$ by $c > 0$ units:

$$\mathbb{P}[\varphi_b(\gamma) \geq E\varphi_b(\gamma^*) + c] = \mathbb{P}[\varphi_b(\gamma) - \varphi_b(\gamma^*) + \varphi_b(\gamma^*) \geq E\varphi_b(\gamma^*) + c]$$

$$\leq \mathbb{P}[\varphi_b(\gamma^*) \geq E\varphi_b(\gamma^*) + \varphi_b(\gamma^*) - \varphi_b(\gamma) + c]$$

$$\leq \mathbb{P}[\varphi_b(\gamma^*) \geq E\varphi_b(\gamma^*) - L||\gamma - \gamma^*|| + c], \quad (54)$$

where the last inequality comes from the Lipschitz property of $\varphi_b$. If $c > L||\gamma - \gamma^*||$, then we can use (52) and obtain

$$\mathbb{P}[\varphi_b(\gamma) \geq E\varphi_b(\gamma^*) + c] \leq \exp\left(-\frac{(c - L||\gamma - \gamma^*||)^2}{4V}\right). \quad (55)$$

Now, let us compute a bound for the probability that $\varphi_b(\gamma)$ is smaller than $E\varphi_b(\gamma^*)$ by a difference of $0 < c < \lambda||\gamma - \gamma^*||^2/2$:

$$\mathbb{P}[\varphi_b(\gamma) \leq E\varphi_b(\gamma^*) + c] = \mathbb{P}[\varphi_b(\gamma) \leq E\varphi_b(\gamma) + E\varphi_b(\gamma^*) - E\varphi_b(\gamma) + c]$$

$$\leq \mathbb{P}\left[\varphi_b(\gamma) \leq E\varphi_b(\gamma) - \frac{\lambda}{2}||\gamma - \gamma^*||^2 + c\right]$$

$$\leq \exp\left(-\frac{\left(\frac{\lambda}{2}||\gamma - \gamma^*||^2 - c\right)^2}{4V}\right). \quad (56)$$

So, suppose that we have a closed interval $[\gamma_0, \gamma_\ell] \subset \mathbb{R}_+$ such that $\gamma^* \in [\gamma_0, \gamma_\ell]$, and that $\gamma_0 < \gamma_1 < \ldots < \gamma_\ell$ produces a partition to $[\gamma_0, \gamma_\ell]$. Also, for some $d > 0$, consider the following two sets:

$$\Gamma := \{\gamma \geq 0 : ||\gamma - \gamma^*|| \leq d\} \quad \text{and} \quad \bar{\Gamma} := \{\gamma \geq 0 : ||\gamma - \gamma^*|| > d\}. \quad (57)$$

Then, let us define the following events for any given $Ld < c < \lambda d^2/2$:
A: there is at least one element in \{\gamma_0, \gamma_1, \ldots, \gamma_\ell\} such that \gamma_j \in \Gamma;

B: there is at least one element \gamma_j \in \{\gamma_0, \gamma_1, \ldots, \gamma_\ell\} such that it belongs to \Gamma, and additionally, \varphi_b(\gamma_j) < \mathbb{E}\varphi_b(\gamma^*) + c;

C: all elements \gamma_i \in \{\gamma_0, \gamma_1, \ldots, \gamma_\ell\} that belong to \hat{\Gamma} satisfy \varphi_b(\gamma_i) > \mathbb{E}\varphi_b(\gamma^*) + c.

Hence, letting

\[ \hat{\gamma} \in \arg\min_{\gamma \in \{\gamma_0, \ldots, \gamma_\ell\}} \varphi_b(\gamma), \tag{58} \]

we have

\[
\begin{align*}
\mathbb{P} [\hat{\gamma} \in \Gamma | A] & \geq \mathbb{P} [B \cap C | A] \\
& = 1 - \mathbb{P} [-B \cup -C | A] \\
& \geq 1 - \mathbb{P} [-B | A] - \mathbb{P} [-C | A] \\
& \geq 1 - \exp \left( -\frac{(c - Ld)^2}{4V} \right) - \exp \left( -\frac{(\frac{1}{2}d^2 - c)^2}{4V} \right). \tag{59}
\end{align*}
\]

In case that \( d > 0 \) is large enough for \( Ld < (\lambda/4)d^2 < \lambda d^2/2 \) to hold, then we can take \( c = (\lambda/4)d^2 \), and find a lower bound for \( \mathbb{P} [\hat{\gamma} \in \Gamma | A] \) in terms of \( d > 0 \):

\[
\mathbb{P} [\hat{\gamma} \in \Gamma | A] \geq 1 - \exp \left( -\frac{(\frac{1}{4}d^2 - Ld)^2}{4V} \right) - \exp \left( -\frac{(\frac{1}{4}d^2)^2}{4V} \right).
\]

With the aim of providing an illustration for these ideas, consider the function \( \varphi_b(\gamma) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R} \) stated as \( \varphi_b(\gamma) = (b_1 + 1)^2\gamma^2 - (b_2 + 1)^2\gamma + \frac{b_2}{2} + 2b_2^2 \), in which \( b_j \in \mathcal{N}(0, \sigma^2), \ j = 1, 2. \)

By taking the standard deviations \( \sigma \in \{0.1, 0.2, 0.3\} \) and randomly generating samples of 200 points in each case, the concentration of measure phenomenon may be observed in the plots of Figure 8. The sets of minimizers of \( \varphi_b(\gamma) := \varphi_b(\gamma) \), given by \( \hat{\gamma}^* = (\frac{b_2 + 1)^2}{2(b_1 + 1)^2} \), are displayed in Figure 9 for each sampling. Notice that, since the average vector of coefficients is \( \bar{b} = (0, 0) \), we obtain \( \varphi_b(\gamma) = \gamma^2 - \gamma \), whose minimizer is 0.5. The minimizers of \( \mathbb{E}\varphi_b \) for each choice of the deviation \( \sigma \), namely 0.499592, 0.503212, and 0.610599, are closer to the ideal value of 0.5 than the average of the minimizers of \( \varphi_b(\gamma) \), which are 0.519001, 0.65556, and 2.34117. We should stress that although the ideal value is 0.5, our interest here is to compare how close a minimizer of a single realization of \( \varphi_b \) is from the minimizer of \( \mathbb{E}\varphi_b \), since this is the subject of concern in this section.

For such an example, the constant of strong convexity may be set as \( \lambda = 2 \). Denoting by \( r \) the radius of the neighborhood to establish the local Lipschitz constant for \( \varphi_b(\gamma) \) around \( \gamma^* \), we reach \( L = (1 + 2r)(1 + \sigma)^2 \). Setting \( r = 0.1 \), the three choices for \( \sigma \) yield \( L \in \{1.452, 1.728, 2.028\} \). Defining \( V = \sigma^2 \) and \( d = 9L/(2\lambda) \), the desired relationships \( Ld < (\lambda/4)d^2 < \lambda d^2/2 \) hold. Thus, the lower bounds we have computed for \( \mathbb{P}[\hat{\gamma} \in \Gamma | A] \) are respectively given by 0.999848, 0.987820, and 0.975685.
Figure 8. Graphs of the sampled functions $\varphi_b(\gamma)$, together with the curve $y = E\varphi_b(\gamma)$.

Figure 9. Sorted distribution of minimizers of $\varphi_b(\gamma)$ with its average depicted in the dashed line. The minimizer of $E\varphi_b$ corresponds to the continuous horizontal line. Notice that the vertical ranges are distinct.

5. Conclusions

We have introduced a generalization of the unbiased predictive risk estimator which allows the use of more general Bregman divergences than the squared norm of the difference. The minimization of these estimators leads to a regularization parameter selection method for inverse problems, which we have applied to the image reconstruction problem in computed tomography. Simulated and real-world experiments corroborate the intuition that the flexibility to select the most appropriate Bregman divergence for the problem in hand might be useful. Finally, we have analysed what the consequences of a concentration inequality in the estimator would be for its minimizer and we concluded that if the estimator does indeed concentrate around its expected value, than so does the minimizer of this estimator.

Acknowledgments

This work was partially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grants 310893/2019-4 and 305010/2020-4 and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grants 2018/24293-0, 2016/22989-2 and 2013/07375-0.

We are also indebted to Prof. Juliana Cobre for reading and commenting on an early version of the manuscript, and to the anonymous referees for pointing rooms for improving the presentation of our work.
References

[1] Crytur LuAG:Ce. https://www.crytur.cz/materials/luagce/. Accessed: 2021-04-04.

[2] Sangtae Ahn and Jeffrey A. Fessler. Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms. *IEEE Transactions on Medical Imaging*, 22(5):613–626, 2003. doi:10.1109/TMI.2003.812251.

[3] R. Averkamp and C. Houdré. Stein estimate for infinitely divisible laws. *ESAIM: Probability and Statistics*, 10:269–276, 2006. doi:10.1051/ps:2006011.

[4] Katy S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001. doi:10.1023/A:1010896012157.

[5] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005. URL http://dl.acm.org/citation.cfm?id=1194902.

[6] Johnathan M. Bardsley and John Goldes. Regularization parameter selection methods for ill-posed Poisson maximum likelihood estimation. *Inverse Problems*, 25(9):095005, 2009. doi:10.1088/0266-5611/25/9/095005.

[7] Frank Bauer and Mark A. Lukas. Comparing parameter choice methods for regularization of ill-posed problems. *Mathematics and Computers in Simulation*, 81(9):1795–1841, 2011. doi:10.1016/j.matcom.2011.01.016.

[8] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009. doi:10.1109/TIP.2009.2028250.

[9] Amel Benazza-Benyahia and Jean-Christophe Pesquet. Building robust wavelet estimators for multicomponent images using Stein’s principle. *IEEE Transactions on Image Processing*, 14(11):1814–1830, 2005. doi:10.1109/TIP.2005.857247.

[10] Federico Benvenuto and Bangti Jin. A parameter choice rule for Tikhonov regularization based on predictive risk. *Inverse Problems*, 36(6):065004, 2020. doi:10.1088/1361-6420/ab6d58.

[11] Mario Bertero, Patrizia Boccacci, Giorgio Talenti, Riccardo Zanella and Luca Zanni. A discrepancy principle for Poisson data. *Inverse Problems*, 26(10):105004–105023, 2010. doi:10.1088/0266-5611/26/10/105004.

[12] Alan C. Bovik, editor. *Handbook of Image & Video Processing*. Elsevier Academic Press, 2005.

[13] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.
REFERENCES

USSR Computational Mathematics and Mathematical Physics, 7(3):200–217, 1967. doi:10.1016/0041-5553(67)90040-7.

[14] Jolyon Browne and Álvaro Rodolfo De Pierro. A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography. IEEE Transactions on Medical Imaging, 15(5):687–699, 1996. doi:10.1109/42.538946.

[15] Michael Collins, Sanjoy Dasgupta and Robert E. Schapire. A generalization of principal components analysis to the exponential family. Advances in Neural Information Processing Systems, 14, 2001. URL http://papers.nips.cc/paper/2078-a-generalization-of-principal-components-analysis-to-the-exponential-family.

[16] Imre Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. The Annals of Statistics, 19(4):2032–2066, 1991. URL http://www.jstor.org/stable/2241918.

[17] Álvaro Rodolfo De Pierro and Michel Eduardo Beleza Yamagishi. Fast EM-like methods for maximum “a posteriori” estimates in emission tomography. IEEE Transactions on Medical Imaging, 20(4):280–288, 2001. doi:10.1109/42.921477.

[18] L. Desbat and D. Girard. The “minimum reconstruction error” choice of regularization parameters: Some more efficient methods and their application to deconvolution problems. SIAM Journal on Scientific Computing, 16(6):1387–1403, 1995. doi:10.1137/0916080.

[19] Devdatt P. Dubashi and Alessandro Panconesi. Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge, 2009.

[20] Yonina C. Eldar. Generalized sure for exponential families: Applications to regularization. IEEE Transactions on Signal Processing, 57(2):471–481, 2009. doi:10.1109/TSP.2008.2008212.

[21] Heinz W. Engl, Martin Hanke and Andreas Neubauer. Regularization of Inverse Problems. Kluwer Academic Publishers, 2000.

[22] Cédric Févotte, Nancy Bertin and Jean-Louis Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Computation, 3(21):793–830, 2009. doi:10.1162/neco.2008.04-08-771.

[23] Nikolas P. Galatsanos and Aggelos K. Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. IEEE Transactions on Image Processing, 1(3):322–336, 1992. doi:10.1109/83.148606.

[24] Gene Golub, Michael Heath and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21(2):215–223, 1979. doi:10.1080/00401706.1979.10489751.
REFERENCES

[25] Robert M. Gray, Andrés Buzo, Jr. Augustine H. Gray and Yasuo Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):367–376, 1980. doi:10.1109/TASSP.1980.1163421.

[26] Petter Hall and D. Michael Titterington. Common structure of techniques for choosing smoothing parameters in regression problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):184–198, 1987. URL http://www.jstor.org/stable/2345419.

[27] Mahmoud Hamada and Emiliano A. Valdez. CAPM and option pricing with elliptically contoured distributions. *The Journal of Risk and Insurance*, 75(2):387–409, 2008. doi:10.1111/j.1539-6975.2008.00265.x.

[28] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992. doi:10.1137/1034115.

[29] Elias S. Helou, Yair Censor, Tai-Been Chen, I-Liang Chern, Álvaro R. De Pierro, Ming Jiang and Henry H.-S. Lu. String-averaging expectation-maximization for maximum likelihood estimation in emission tomography. *Inverse Problems*, 30(5):055003, 2014. doi:10.1088/0266-5611/30/5/055003.

[30] Elias Salomão Helou Neto and Álvaro Rodolfo De Pierro. Convergence results for scaled gradient algorithms in positron emission tomography. *Inverse Problems*, 21(6):1905–1914, 2005. doi:10.1088/0266-5611/21/6/007.

[31] Gabor T. Herman. *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*. Academic Press, 1980.

[32] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms*. A Series of Comprehensive Studies in Mathematics. Springer-Verlag, Berlin, 1993.

[33] H. Malcolm Hudson and Richard S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Transactions on Medical Imaging*, 13(4):601–609, 1994. doi:10.1109/42.363108.

[34] W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379. University of California Press, 1961. URL http://projecteuclid.org/euclid.bsmsp/1200512173.

[35] Jari Kaipio and Erkki Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007. doi:10.1016/j.cam.2005.09.027.

[36] Avinash C. Kak and Malcolm Slaney. *Principles of Computerized Tomographic Imaging*. IEEE press, 1988.

[37] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi:10.1214/aoms/1177729694.
[38] ZINOVYI LANDSMAN AND JOHANNA NEŠLEHOVÁ. Stein’s lemma for elliptical random vectors. Journal of Multivariate Analysis, 99:912–927, 2008. doi:10.1016/j.jmva.2007.05.006.

[39] HOUSEN LI AND FRANK WERNER. Empirical risk minimization as parameter choice rule for general linear regularization methods. Annales de l’Institut Henri Poincaré - Probabilités et Statistiques, 56(1):405–427, 2020. doi:10.1214/19-AIHP966.

[40] FLORIAN LUISIER, THIERRY BLU AND MICHAEL UNSER. Image denoising in mixed Poisson–Gaussian noise. IEEE Transactions on Image Processing, 20(3):696–708, 2011. doi:10.1109/TIP.2010.2073477.

[41] PAOLO MASSA AND FEDERICO BENVENUTO. Predictive risk estimation for the exact minimization algorithm with poisson data. Inverse Problems, 37(4):045013, 2021. doi:10.1088/1361-6420/abe950.

[42] JAN MATULA, MARKETA TESAROVA, TOMAS ZIKMUND, MARKETA KAUCKA, IGOR ADAMEYKO AND JOZEF KAISER. X-ray microtomography-based atlas of mouse cranial development. GigaScience, 3(10):1–6, 2021. doi:10.1093/gigascience/giab012.

[43] YOANN LE MONTAGNER, ELSA D. ANGELINI AND JEAN-CHRISTOPHE OLIVOMARIN. An unbiased risk estimator for image denoising in the presence of mixed Poisson–Gaussian noise. IEEE Transactions on Image Processing, 23(3):1255–1268, 2014. doi:10.1109/TIP.2014.2300821.

[44] FRANK NATTERER. The Mathematics of Computerized Tomography. Wiley, 1986.

[45] FRANK NATTERER AND FRANK WÜBBELING. Mathematical Methods in Image Reconstruction. SIAM, 2001.

[46] JAMES C.-M. PENG. Simultaneous estimation of the parameters of independent Poisson distribution. Technical Report EFS NSF 78, Stanford, Dec. 1975. URL https://purl.stanford.edu/gj419gc5599.

[47] SATISH RAMANI, THIERRY BLU AND MICHAEL UNSER. Monte-Carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms. IEEE Transactions on Image Processing, 17(9):1540–1554, 2008. doi:10.1109/TIP.2008.2001404.

[48] SI SI, DACHENG TAO AND BO GENG. Bregman divergence-based regularization for transfer subspace learning. IEEE Transactions on Knowledge and Data Engineering, 22(7):929–942, 2010. doi:10.1109/TKDE.2009.126.

[49] DONALD L. SNYDER AND ABED M. HAMMOUD. Image recovery from data acquired with a charge-coupled-device camera. Journal of the Optical Society of America A, 10(5):1014–1023, 1993. doi:10.1364/JOSAA.10.001014.

[50] CHARLES M. STEIN. Estimation of the mean of a multivariate normal distribution. The Annals of Statistics, 9(6):1135–1151, 1981. URL http://www.jstor.org/stable/2240405.
[51] Michel Talagrand. A new look at independence. *The Annals of Probability*, 1(24):1–34, 1996. URL https://www.jstor.org/stable/2244830.

[52] Alan M. Thompson, John C. Brown, Jim W. Kay and D. Michael Titterington. A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):326–339, 1991. doi:10.1109/34.88568.

[53] Y. Vardi, Larry A. Shepp and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985. URL http://www.jstor.org/pss/2288030.