This advance publication article will be replaced by the finalized version after proofreading.
Convex and Differentiable Formulation for Inverse Problems in Hilbert Spaces with Nonlinear Clipping Effects

Natsuki UENO, Nonmember, Shoichi KOYAMA, and Hiroshi SARUWATARI, Members

SUMMARY

We propose a useful formulation for ill-posed inverse problems in Hilbert spaces with nonlinear clipping effects. Ill-posed inverse problems are often formulated as optimization problems, and nonlinear clipping effects may cause nonconvexity or nondifferntiability of the objective functions in the case of commonly used regularized least squares. To overcome these difficulties, we present a tractable formulation in which the objective function is convex and differentiable with respect to optimization variables, on the basis of the Bregman divergence associated with the primitive function of the clipping function. By using this formulation in combination with the representer theorem, we need only to deal with a finite-dimensional, convex, and differentiable optimization problem, which can be solved by well-established algorithms. We also show two practical examples of inverse problems where our theory can be applied, estimation of band-limited signals and time-harmonic acoustic fields, and evaluate the validity of our theory by numerical simulations.

key words: inverse problem, Hilbert space, representer theorem, Bregman divergence, convex optimization

1. Introduction

An inverse problem is a process of determining or estimating a quantity of interest from its indirect observation, called a forward problem, whose theory has been under intensive investigation in engineering, applied mathematics, and many other fields. Here, we consider the following types of forward problem:

\[ s = (f_1(L_1u), \ldots, f_N(L_Nu)), \]  

where \( N \) is a natural number, \( u \in \mathcal{U} \) is a quantity of interest, \( \mathcal{U} \) is a real Hilbert space consisting of all possible candidates for quantities of interest, \( s \in \mathcal{S} \) is observed data, \( \mathcal{S} \subseteq \mathbb{R}^N \) is a set consisting of all possible candidates for observed data, \( L_1, \ldots, L_N : \mathcal{U} \to \mathbb{R} \) are bounded linear functionals, and \( f_1, \ldots, f_N : \mathbb{R} \to \mathbb{R} \) are continuous and monotonically increasing (not necessarily strictly monotonically increasing and therefore not necessarily invertible) functions representing nonlinear clipping effects (hereafter referred to as clipping functions). Three examples of clipping functions are shown in Fig. 1. The goal of solving the inverse problem is to find the unknown quantity \( u \) from the given \( L_1, \ldots, L_N, f_1, \ldots, f_N \), and available (often noisy) data \( s \).

A wide variety of physical measurements in practical situations can be modeled by Eq. (1) in a unified way. A typical example is a restoration problem of a band-limited continuous-time signal from its discrete and clipped samples.

Here, \( L_1, \ldots, L_N \) represent sampling, i.e., pointwise evaluations, of signals, and \( f_1, \ldots, f_N \) represent clipping effects caused by sensors, amplifiers, and analog-to-digital converters. There are several related works on signal reconstruction methods and theories using nonclipped samples [1]–[3], clipped but uniform samples [4]. Several reconstruction methods for clipped signals [5]–[9] can be applied also in cases of nonuniform samples. Moreover, Rencker et al. [9] provided a discussion on influence of observational noises before clipping effects. However, formulation in these methods has discontinuous sensitivity against the observational noises after clipping effects, and a signal reconstruction method considering noises after clipping effects remains to be established and investigated. In addition, Eq. (1) includes many other examples such as measurements of various physical fields, e.g., electric, magnetic, and acoustic fields. For example, in an acoustic field measurement using multiple (omnidirectional or directional) microphones [10], [11], the directivity and frequency response of the microphones can be modeled by \( L_1, \ldots, L_N \). Note that \( L_1, \ldots, L_N \) are not limited to pointwise evaluations of functions, as this example shows.

In general, this inverse problem is ill-posed, i.e., the forward problem is not bijective and/or not continuous. Therefore, the following types of optimization problem are often used for determining the most reasonable solution:

\[ \min_{u \in \mathcal{U}} Q_s(u) := D_f(L_u, s) + \frac{\delta}{2} \| u \|_{\mathcal{U}}^2. \]  

where \( L : \mathcal{U} \to \mathbb{R}^N \) and \( f : \mathbb{R}^N \to \mathcal{S} \) are simplified notations of \( L_1, \ldots, L_N \) and \( f_1, \ldots, f_N \), respectively, i.e., \( f(\mathbf{L}_u) = (f_1(L_1u), \ldots, f_N(L_Nu)) \). \( D_f : \mathbb{R}^N \times \mathcal{S} \to (-\infty, +\infty] \) is a loss function, \( \delta \in (0, +\infty) \) is a regularization parameter, and \( \| \cdot \|_{\mathcal{U}} \) is the norm induced by the inner product \( (\cdot, \cdot)_{\mathcal{U}} \) on \( \mathcal{U} \). Here, \( Q_s : \mathcal{U} \to (-\infty, +\infty] \) is an objective function for evaluating the reasonability of \( u \) under the given \( s \), which is composed of two terms on the right-hand side of Eq. (2). The first term evaluates the consistency between \( L_u \) and \( s \) under \( f \). On the other hand, the second term is added to avoid divergence of the solution and determined independently of the observation. The following regularized least squares is one of the most commonly used formulations in the form of Eq. (2):

\[ \min_{u \in \mathcal{U}} Q_s(u) := \frac{1}{2} \| f(L_u) - s \|_{\mathbb{R}^N}^2 + \frac{\delta}{2} \| u \|_{\mathcal{U}}^2. \]  

Various regularization methods have been proposed and an-
analyzed in innumerable works (see, for example, [12]–[15] for an overview).

The optimization problem of Eq. (2) or (3) includes two difficulties. One is that \( \mathcal{U} \) may have infinite dimensions. This problem can be overcome using the representer theorem [16]–[21], which states that the optimal solution of Eq. (2) can be represented as an \( N \)-linear combination of certain vectors determined by \( L_1, \ldots, L_N \). Therefore, we need only to seek the optimal (finitely many) coefficients of these vectors. Although this theorem is typically applied to kernel methods [22], [23] in machine learning because of its affinity for reproducing kernel Hilbert spaces [24], [25], it is also useful in physical measurements. For example, the authors proposed an acoustic field estimation method based on the representer theorem for cases when \( f \) is linear [26]. When the observation includes nonlinear clipping effects, however, there still remains another difficulty even with the aid of the representer theorem; even a commonly used regularized least squares, i.e., Eq. (3), is not necessarily convex and differentiable with respect to the optimization variables owing to the nonlinearity or nondifferentiability of \( f \). Therefore, an alternative formulation of the loss function that is easy to optimize is desired from a practical viewpoint.

In this paper, we propose a tractable formulation where the loss function is convex and differentiable with respect to the optimization variables and has a divergence-like property, i.e., it is minimized if and only if \( s = f(Lu) \). By using our proposed formulation in combination with the representer theorem, we need only to deal with a finite-dimensional, convex, and differentiable optimization problem without any constraints, which can be solved easily by using well-established algorithms such as gradient descent, nonlinear conjugate gradient descent, and quasi-Newton methods [27]–[31]. This loss function is designed using the primitive functions of the clipping functions and also related closely to the concept of the Bregman divergence. Even in cases where another loss function such as Eq. (3) is considered to be suitable from a stochastic or other viewpoint, a good initial solution can be obtained by using our proposed formulation owing to its convexity. We also provide two practical examples, estimation problems of band-limited signals from their time-sampled values and time-harmonic acoustic fields using pressure-gradient microphones, and evaluate the validity of our theory by numerical experiments.

The rest of this paper is organized as follows. In Sect. 2, we introduce the representer theorem. In Sect. 3, we propose a convex and differentiable formulation of the loss function in Eq. (2). In Sect. 4, we show two practical examples of inverse problems in which our theory can be applied, and the results of their numerical experiments are provided. Finally, we present our conclusion in Sect. 5.

2. Representer Theorem

We introduce the representer theorem, which guarantees that Eq. (2) can be reduced to an optimization problem with respect to \( N \) variables. Since there are several related results [16]–[21], all of which are called representer theorems, we provide here a suitable form for Eq. (2) also with extension to local solutions [18]. This extension is essential because we also consider nonconvex formulation such as Eq. (3) with initialization using our proposed convex formulation. This theorem is effective, especially when \( \mathcal{U} \) has infinite dimensions; however, it is also useful when \( \mathcal{U} \) has finite but much larger dimensions than \( N \) because it may contribute to a significant reduction of computational cost in the optimization.

2.1 Preliminaries

First, we introduce the Riesz representation theorem for Hilbert spaces (see, e.g., [32], [33], for a proof).

**Theorem 1:** Let \( (\mathcal{U}, \langle \cdot, \cdot \rangle_{\mathcal{U}}) \) be a Hilbert space over \( \mathbb{R} \) (or \( \mathbb{C} \)) and \( \Phi : \mathcal{U} \to \mathbb{R} \) be a bounded linear functional. Then, there exists a unique \( \varphi \in \mathcal{U} \) satisfying

\[
\Phi u = \langle \varphi, u \rangle_{\mathcal{U}} \quad \forall u \in \mathcal{U}. \tag{4}
\]

From the Riesz representation theorem, for the bounded linear operator \( L \) defined in Sect. 1, there exists a unique sequence \( (v_1, \ldots, v_N) \in \mathcal{U}^N \) satisfying

\[
L_n u = \langle v_n, u \rangle_{\mathcal{U}} \quad \forall u \in \mathcal{U}, n \in \{1, \ldots, N\}. \tag{5}
\]

Hereafter, we call \( v_n \) the representation vector of \( L_n \) and \((v_1, \ldots, v_N)\) the representation vector sequence of \( L \).

Next, we define a Gram operator as follows.

**Definition 2:** Let \( (\mathcal{U}, \langle \cdot, \cdot \rangle_{\mathcal{U}}) \) be a Hilbert space over \( \mathbb{R} \) (or \( \mathbb{C} \)), \( N \) be a natural number, and \( v_n \in \mathcal{U} \) be the representation vector of a bounded linear functional \( L_n : \mathcal{U} \to \mathbb{R} \) for...
each \( n \in \{1, \ldots, N\} \). Then, a linear mapping \( K : \mathbb{K}^N \rightarrow \mathbb{K}^N \) whose representation matrix \( K \) is given as

\[
K := \begin{bmatrix}
K_{1,1} & \cdots & K_{1,N} \\
\vdots & \ddots & \vdots \\
K_{N,1} & \cdots & K_{N,N}
\end{bmatrix}
\]

(6)

with

\[
K_{n,m} := \langle v_n, v_m \rangle_{\mathcal{U}} = L_{nm} v_n \quad \forall n, m \in \{1, \ldots, N\}
\]

(7)

is called a Gram operator generated by \( (v_1, \ldots, v_N) \). This matrix \( K \) is always self-adjoint and positive-semidefinite, and it is positive-definite if and only if \( (v_1, \ldots, v_N) \) is linearly independent. Furthermore, the semi-inner product and seminorm on \( \mathbb{K}^N \) induced by \( K \) are denoted by \( \langle \cdot, \cdot \rangle_K \) and \( \| \cdot \|_K \), respectively, i.e.,

\[
\langle \alpha, \beta \rangle_K := \langle \alpha, K \beta \rangle_{\mathbb{K}^N} \quad \forall \alpha, \beta \in \mathbb{K}^N,
\]

(8)

\[
\| \alpha \|_K := \sqrt{\langle \alpha, \alpha \rangle_K} \quad \forall \alpha \in \mathbb{K}^N.
\]

(9)

2.2 Statement

In accordance with a prior work [17], we provide the statement for slightly more general conditions than those in Eq. (2), it can be easily confirmed that Eq. (2) satisfies these conditions. For an essential part of the proof, see [18].

**Theorem 3:** Assume the following conditions.

- \( (\mathcal{U}, \langle \cdot, \cdot \rangle_{\mathcal{U}}) \) : Hilbert space over \( \mathbb{K} \) (\( \mathbb{R} \) or \( \mathbb{C} \))
- \( \| \cdot \|_{\mathcal{U}} \) : norm on \( \mathcal{U} \) induced by the inner product \( \langle \cdot, \cdot \rangle_{\mathcal{U}} \)
- \( N \) : natural number
- \( L : \mathcal{U} \rightarrow \mathbb{K}^N \) bounded linear operator
  - \( (v_1, \ldots, v_N) \) : representation vector sequence of \( L \)
  - \( K \) : Gram operator generated by \( (v_1, \ldots, v_N) \)
  - \( \| \cdot \| : \text{seminorm on } \mathbb{K}^N \) induced by \( K \)
- \( E : \mathbb{K}^N \rightarrow (-\infty, +\infty] \) : arbitrary function
- \( g : [0, +\infty) \rightarrow (-\infty, +\infty] \) : strictly monotonically increasing function

Then, any global/local optimal point \( u^{\text{opt}} \in \mathcal{U} \) of the optimization problem

\[
\min_{u \in \mathcal{U}} \quad Q(u) := E(Lu) + g(\|u\|_{\mathcal{U}})
\]

(10)

admits the following representation:

\[
u^{\text{opt}} = \sum_{n=1}^{N} u^{(\text{opt})}_n v_n.
\]

(11)

Here, \( u^{(\text{opt})} := (u_1^{(\text{opt})}, \ldots, u_N^{(\text{opt})}) \in \mathbb{K}^N \) is given as a global/local optimal point of the optimization problem

\[
\min_{a \in \mathbb{K}^N} \quad Q(a) := E(Ka) + g(\|a\|_K).
\]

(12)

More precisely, the following two propositions hold:

1. If and only if \( u^{\text{opt}} \in \mathcal{U} \) is a global optimal point of Eq. (10), \( u^{\text{opt}} \) can be represented in the form of \( u^{\text{opt}} = \sum_{n=1}^{N} u^{(\text{opt})}_n v_n \) for some \( u^{(\text{opt})} := (u_1^{(\text{opt})}, \ldots, u_N^{(\text{opt})}) \in \mathbb{K}^N \) that is a global optimal point of Eq. (12).

2. If and only if \( u^{\text{opt}} \in \mathcal{U} \) is a local optimal point of Eq. (10), \( u^{\text{opt}} \) can be represented in the form of \( u^{\text{opt}} = \sum_{n=1}^{N} u^{(\text{opt})}_n v_n \) for some \( u^{(\text{opt})} := (u_1^{(\text{opt})}, \ldots, u_N^{(\text{opt})}) \in \mathbb{K}^N \) that is a local optimal point of Eq. (12).

**Remark 4:** With respect to the term “local”, we consider the standard topologies on \( \mathcal{U} \) and \( \mathbb{K}^N \) induced by the norms \( \| \cdot \|_{\mathcal{U}} \) and \( \| \cdot \|_{\mathbb{K}^N} \), respectively. In particular, \( u^{\text{opt}} \in \mathcal{U} \) is called a local optimal point of Eq. (10) if and only if \( Q(u^{\text{opt}}) < +\infty \) and there exists some \( \epsilon \in (0, +\infty) \) satisfying \( Q(u^{\text{opt}}) \leq Q(u) \forall u \in B(u^{\text{opt}}; \epsilon) \). Here, for \( u \in \mathcal{U} \) and \( \epsilon \in (0, +\infty) \), \( B(u; \epsilon) \in \mathcal{U} \) denotes the open ball centered at \( u_0 \) with radius \( \epsilon \), i.e., \( B(u; \epsilon) := \{ u \in \mathcal{U} \mid \|u - u_0\|_{\mathcal{U}} < \epsilon \} \).

3. Convex Formulation

In this section, we present a tractable formulation of the loss function \( D \) and provide a concrete discussion for solving the optimization problem by computation.

3.1 Formulation

Let \( F_1, \ldots, F_N : \mathbb{R} \rightarrow \mathbb{R} \) denote primitive functions of \( f_1, \ldots, f_N \), respectively. Note that \( F_1, \ldots, F_N \) are differentiable and convex because \( f_1, \ldots, f_N \) are continuous and monotonically increasing. In addition, the ranges of \( f_1, \ldots, f_N \) are respectively denoted by \( S_1, \ldots, S_N \subseteq \mathbb{R} \). Then, we define the loss function \( D_f \) as

\[
D_f(z, s) := \sum_{n=1}^{N} d_n(z_n, s_n)
\]

(13)

\[
\forall z := (z_1, \ldots, z_N) \in \mathbb{R}^N, s := (s_1, \ldots, s_N) \in S
\]

with

\[
d_n(z_n, s_n) := \frac{1}{\sigma_n} [F_n(z_n) - F_n(f_n^{-1}(s_n)) - s_n(z_n - f_n^{-1}(s_n))]
\]

(14)

where \( \sigma_1, \ldots, \sigma_N \in (0, +\infty) \) are dispersion parameters representing the observational uncertainty. Although \( f_1, \ldots, f_N \) is not uniquely invertible in general, this is a well-defined function because any choice of \( f_n^{-1}(s_n) \) in Eq. (14) yields the same value (see Appendix A). This loss function has the following divergence-like properties.

1. \( D_f(z, s) \geq 0 \forall z \in \mathbb{R}^N, s \in S \).
2. \( D_f(z, s) = 0 \) if and only if \( s = f(z) \).
Proofs of these properties are provided in Appendix B. In cases of a strictly monotonically increasing \( f_n \), \( d_n \) can also be represented as \( d_n(z_n, s_n) = B_{f_n}(z_n, f_n^{-1}(s_n))/\sigma_n^2 \forall z_n, s_n \in \mathbb{R}, \) where \( B_{f_n} : \mathbb{R} \times \mathbb{R} \to [0, +\infty) \) denotes the Bregman divergence associated with \( f_n \). Moreover, if \( f_n \) is the identity function, \( d_n \) is a weighted squared error function defined as \( d_n(z_n, s_n) = (z_n - s_n)^2/(2\sigma_n^2) \forall z_n, s_n \in \mathbb{R} \). Using Eq. (14), we obtain the objective function as

\[
Q_s(u) := \sum_{n=1}^{N} \frac{1}{\sigma_n^2} [F_n(L_nu) - F_n(f_n^{-1}(s_n))] - s_n(L_nu - f_n^{-1}(s_n))] + \frac{\delta}{2} \|u\|_U^2. \tag{15}
\]

This can be simplified by omitting the terms irrelevant to the optimization variable as

\[
Q_s(u) = \sum_{n=1}^{N} \frac{1}{\sigma_n^2} (F_n(K_nu) - s_nK_n) + \frac{\delta}{2} \|u\|_U^2 + C, \tag{16}
\]

where \( C \in \mathbb{R} \) is a constant.

From Theorem 3, the minimization problem of Eq. (16) on \( \mathcal{U} \) can be recast to the minimization problem of the following objective function on \( \mathbb{R}^N \):\(^\dagger\)

\[
Q_s^{(\alpha)}(\alpha) := \sum_{n=1}^{N} \frac{1}{\sigma_n^2} (F_n(K_n\alpha) - s_nK_n\alpha) + \frac{\delta}{2} \|\alpha\|_K^2 + C, \tag{17}
\]

where \( K_n : \mathbb{R}^N \to \mathbb{R} \) is a linear mapping whose representation matrix (a row vector) \( K_n \) is given by

\[
K_n := [K_{n,1}, \ldots, K_{n,N}] \quad \forall n \in \{1, \ldots, N\}. \tag{18}
\]

Then, \( Q_s^{(\alpha)} \) is a convex function because the functions \( F_n(K_n\alpha), -s_nK_n\alpha, \) and \( \|\alpha\|_K^2 \) are convex with respect to \( \alpha \) (all of the following functions are convex: linear functions, seminorm functions\(^\dagger\), composition of a convex function and linear transformation, square of a convex nonnegative function, and nonnegative weighted sums of convex functions \([30]\)).

Note that Eq. (16) can be defined even when \( s \notin S \) (due to observational noises) by ignoring the constant \( C \). Also in this case, the existence of the solution is guaranteed because \( Q_s \) (where \( C \) is omitted) is continuous and coercive (see Appendix C); therefore, one can use the proposed formulation regardless of whether or not \( s \) is in \( S \). It should be also noted that most related works \([6]–[9]\) cannot be used in such cases.

3.2 Derivation of Gradient

As described above, the objective function \( Q_s^{(\alpha)} \) is convex and has no constraint. In addition, \( Q_s^{(\alpha)} \) is partially differentiable, and the gradient of \( Q_s^{(\alpha)} \) is given by

\[
\frac{\partial}{\partial \alpha_i} Q_s^{(\alpha)}(\alpha) = \frac{1}{\sigma_n^2} (f_n(K_n\alpha) - s_nK_{n,i}) + \delta K_i \alpha \quad \forall i \in \{1, \ldots, N\}. \tag{19}
\]

Therefore, many well-established iterative algorithms using the gradient of \( Q_s^{(\alpha)} \), such as gradient descent, nonlinear conjugate gradient descent, and quasi-Newton methods, are available for the minimization of \( Q_s^{(\alpha)} \).

Furthermore, if \( f_1, \ldots, f_N \) are \( \eta \)-Lipschitz continuous with \( \eta \in [0, +\infty) \), the gradient of \( Q_s^{(\alpha)} \) is also Lipschitz continuous with a Lipschitz constant \( \eta M_K^2/\sigma_{\min}^2 + \delta M_K \), where \( M_K \) denotes the maximum eigenvalue of the representation matrix of \( K \) and \( \sigma_{\min} \) denotes the minimum value of \( \{\sigma_1, \ldots, \sigma_N\} \). In this case, the global convergence is guaranteed in many iterative algorithms including a gradient descent with a fixed step size \([27]\) and an accelerated gradient descent \([28]\).

4. Application Examples and Numerical Experiments

We provide two examples of inverse problems to which our proposed formulation can be applied. Hereafter, for any measurable set \( \mathcal{X} \subseteq \mathbb{R} \) and measurable function \( \rho : \mathcal{X} \to \mathbb{C} \) with respect to the Lebesgue measure, \( \int_{\mathcal{X}} \rho(x) \, d\mu \) denotes the Lebesgue integral of \( \rho \) on \( \mathcal{X} \).

4.1 Estimation of band-limited signals

First, we provide an example of an estimation problem of band-limited signals from their irregularly time-sampled values with soft clipping effects.

4.1.1 Formulation

A band-limited signal can be modeled as a function \( u : \mathbb{R} \to \mathbb{R} \) in the form of

\[
u(t) = \frac{1}{\sqrt{2\pi}} \int_{\omega \in \Omega} \hat{u}(\omega) \exp(i\omega t) \, d\mu \quad \forall t \in \mathbb{R} \tag{20}
\]

with a square-integrable function \( \hat{u} : \Omega \to \mathbb{C} \), where \( \Omega := [-\omega_{\max}, \omega_{\max}] \) and \( \omega_{\max} \in (0, +\infty) \) denotes the maximum angular frequency of target signals. Therefore, we define a real Hilbert space \( \mathcal{U} \) and its inner product \( \langle \cdot, \cdot \rangle_\mathcal{U} \) as

\[
\mathcal{U} := \left\{ u : \mathbb{R} \to \mathbb{R} \big| \exists \hat{u} : \Omega \to \mathbb{C} \right. \\
\left. \text{s.t.} \int_{\omega \in \Omega} |\hat{u}(\omega)|^2 \, d\mu < +\infty, \ u = \Re \hat{u} \right\}, \tag{21}
\]

and

\[
\mathcal{L} := \left\{ \Omega \to \mathbb{R} \mid \exists \hat{u} : \Omega \to \mathbb{C} \right. \\
\left. \text{s.t.} \int_{\omega \in \Omega} |\hat{u}(\omega)|^2 \, d\mu < +\infty, \ \Re \hat{u} \right\}. \tag{22}
\]
Table 1: Settings in band-limited signal estimation.

<table>
<thead>
<tr>
<th>( \omega_{\text{max}} )</th>
<th>( \omega_{\text{max}} := 2\pi f_{\text{max}} ) with ( f_{\text{max}} = 100 \text{ Hz} ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u )</td>
<td>( u(t) := \frac{\sin(\omega_{\text{max}}t - t)}{\pi} ).</td>
</tr>
<tr>
<td>( (a_{1\text{B}}, \ldots, a_{200}) )</td>
<td>Sampled independently from the univariate real normal distribution with mean 0 and variance 1.</td>
</tr>
<tr>
<td>( N )</td>
<td>( N := 500 ).</td>
</tr>
<tr>
<td>( (t_1, \ldots, t_N) )</td>
<td>Sampled independently from the uniform distribution on ([0, 1]) s.</td>
</tr>
<tr>
<td>( \tau )</td>
<td>( \tau := 1 ) (clipped), ( \tau := +\infty ) (nonclipped).</td>
</tr>
</tbody>
</table>

Observation noise

- Added to each \( f_n(L_nu) \) for \( n \in \{1, \ldots, N\} \) independently from the univariate real normal distribution with mean 0 and variance \( 10^{-3} \times \|f(L_nu)\|_{L_2}^2 / N \).

\[ \delta, (\sigma_1^2, \ldots, \sigma_N^2) \]

where \( \delta := 10^{-1.5}, \sigma_n := 1 \forall n \in \{1, \ldots, N\} \).

Evaluation criterion

- Normalized mean square error (NMSE): 10 log(1E/S), where

\[ E := \sum_{n=1}^{N} \|u(t) - u(t)\|_2^2, \]

\[ S := \sum_{n=1}^{N} \|u(t)\|_2^2. \]

\[ (t_{\text{eval}}, \ldots, t_{\text{eval}}) \]

Equally spaced points from 0 to 1 s with intervals of 0.001 s (\( N_f = 1001 \)).

\[ u_1, u_2 \in U := \int \sum_{\mu} u_1(t)u_2(t) \mu, \quad \text{(22)} \]

respectively, where \( \mathcal{S} \) denotes the transform of functions defined as Eq. (20).

We consider the forward problem in the form of Eq. (1) with

\[ L_nu := u(t_n) \quad \forall u \in \mathcal{U}, n \in \{1, \ldots, N\}, \quad \text{(23)} \]

\[ f_n(z) := \tau \tanh(z/\tau) \quad \forall z \in \mathbb{R}, n \in \{1, \ldots, N\}, \quad \text{(24)} \]

where \( (t_1, \ldots, t_N) \in \mathbb{R}^N \) are sampling times and \( \tau \in (0, +\infty] \) represents a clipping level (\( \tau = +\infty \) means that \( f_1, \ldots, f_N \) are the identical functions).

In this case, \( t_n \) and \( K_{n_1, n_2} \) are given by

\[ v_n(t) = \frac{\omega_{\text{max}}}{\pi} \sin(\omega_{\text{max}}(t - t_n)) \quad \forall t \in \mathbb{R}, n \in \{1, \ldots, N\}, \quad \text{(25)} \]

and

\[ K_{n_1, n_2} = \frac{\omega_{\text{max}}}{\pi} \sin(\omega_{\text{max}}(t_{n_2} - t_{n_1})) \quad \forall n_1, n_2 \in \{1, \ldots, N\}, \quad \text{(26)} \]

respectively, where \( \sin(\cdot) \) is the unnormalized sinc function defined as \( \sin(0) := 1 \) and \( \sin(z) := \sin(z)/z \) for \( z \in \mathbb{R}\setminus\{0\} \). These are derived by the fact that the sinc function is the reproducing kernel of \( \mathcal{U} \), as noted in [34], [35].

4.1.2 Experimental evaluation

We conducted numerical experiments using Julia v.1.2.0 whose settings are given in Table 1. Here, noisy observation was simulated. Note that observation noise was added after the clipping effects (therefore, mere inversion of \( f \) causes significant amplification of the noise). We compared the following five conditions: clipped (Proposed, SC: soft consistency, RLS: regularized least squares, Proposed+RLS) and nonclipped (Proposed). Under the clipped (Proposed, RLS, and Proposed+RLS) conditions, \( \tau \) was set as 1. Under the nonclipped (Proposed) condition, on the other hand, \( \tau \) was set as +\( \infty \). Under the clipped (Proposed) and nonclipped (Proposed) conditions, the optimal solutions \( u^{\text{(opt)}} \) in \( \mathcal{U} \) were obtained by minimizing Eq. (17) with an initial value of zero. Under the clipped (SC) condition, the objective function was defined using the soft consistency loss function [9] as

\[ Q_{\text{SC}, s}(u) := \sum_{n=1}^{N} \frac{1}{2\sigma_n^2} (L_nu - \Pi_{C_n}(L_nu))^2 + \frac{\delta}{2} \|u\|_{L_2}^2, \quad \text{(27)} \]

where \( C_n \subseteq \mathbb{R} \) is the feasible set defined as \( C_n := \{ z \in \mathbb{R} | f(z) = u_n \} \), and \( \Pi_{C_n}(\cdot) \) denotes the orthogonal projection into \( C_n \) for each \( n \in \{1, \ldots, N\} \) (the Euclidean-distance-based loss function is used here although the method of [9] can be combined with other loss functions such as the Huber loss function). On the basis of Theorem 3, the optimal solutions were obtained by minimizing

\[ Q_{\text{SC}, s}(\alpha) := \sum_{n=1}^{N} \frac{1}{2\sigma_n^2} (K_n\alpha - \Pi_{C_n}(K_n\alpha))^2 + \frac{\delta}{2} \|\alpha\|_{K}^2, \quad \text{(28)} \]

with an initial value of zero. Here, for each \( n \in \{1, \ldots, N\} \), the signal \( s_n \) was discarded when it was outside the range of \( f_n \) owing to the observational noise since \( C_n \) cannot be defined for such \( s_n \). Under the clipped (RLS) condition, the objective function was defined as the following regularized least squares:

\[ Q_{\text{LSQ}, s}(u) := \sum_{n=1}^{N} \frac{1}{2\sigma_n^2} (f_n(L_nu) - s_n)^2 + \frac{\delta}{2} \|u\|_{L_2}^2. \quad \text{(29)} \]

On the basis of Theorem 3, the optimal solutions were obtained by minimizing

\[ Q_{\text{LSQ}, s}(\alpha) := \sum_{n=1}^{N} \frac{1}{2\sigma_n^2} (f_n(K_n\alpha) - s_n)^2 + \frac{\delta}{2} \|\alpha\|_{K}^2, \quad \text{(30)} \]
signals in the first trial in Fig. 2. The estimation accuracy for clipped (SC) was lowest, which means its lack of robustness against observational noises after clipping effects. Also the estimation accuracy for clipped (RLS) was low, which was considered to be due to its nonconvexity. On the other hand, the estimation accuracy for clipped (Proposed) was higher than that for clipped (RLS), and it was further improved by clipped (Proposed+RLS), which was close to that for nonclipped (Proposed). One possible reason why clipped (Proposed+RLS) outperformed clipped (Proposed) is because the formulation of RLS is well suited for Gaussian noises; it can be interpreted formally as a maximum a posteriori (MAP) estimation when the observational noises follow Gaussian distributions. From these results, one can see that an accurate estimation was achieved using the formulation of Eq. (17), which could also be used as an initial value in the optimization of a different formulation.

4.2 Estimation of time-harmonic acoustic fields

Next, we provide an example of an estimation problem of time-harmonic acoustic fields using pressure-gradient microphones (also called velocity microphones) with hard clipping effects.

4.2.1 Formulation

Consider the following two-dimensional acoustic wave equation:

$$\left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) u(x, y, t) = 0 \quad \forall x, y, t \in \mathbb{R}. \quad (31)$$

Here, $u(x, y, t) \in \mathbb{R}$ denotes the sound pressure at a position $(x, y) \in \mathbb{R}^2$ and a time $t \in \mathbb{R}$, and $c \in (0, +\infty)$ denotes the speed of sound. One of the solutions for Eq. (31) is a complex sinusoidal plane wave function defined as

$$\varphi(x, y, t; \theta) := \exp \left( i \omega_0 \left( t - \frac{x \cos \theta + y \sin \theta}{c} \right) \right) \quad \forall x, y, t \in \mathbb{R}, \quad (32)$$

where $\omega_0 \in (0, +\infty)$ denotes an angular frequency and $\theta \in [0, 2\pi)$ denotes a polar angle of the traveling direction of the plane wave. As a superposition of plane waves, the time-harmonic incident acoustic field with angular frequency $\omega_0$ can be well modeled as

$$u(x, y, t) = \int_{\theta \in [0, 2\pi)} \text{Re} \left( \hat{\varphi}(x, y, t; \theta) \right) d\mu$$

$$\forall x, y, t \in \mathbb{R} \quad (33)$$

with a square-integrable function $\hat{\varphi} : [0, 2\pi) \rightarrow \mathbb{C}$. We define a real Hilbert space $\mathcal{U}$ and its inner product $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ as

$$\mathcal{U} := \left\{ u : \mathbb{R} \rightarrow \mathbb{R} \ ; \ \forall u : [0, 2\pi) \rightarrow \mathbb{C} \right\}$$

$$\text{s.t.} \quad \int_{\theta \in [0, 2\pi)} |\hat{u}(\theta)|^2 d\mu < +\infty, \ u = \mathcal{F}u \quad (34)$$

and

$$\langle u_1, u_2 \rangle_{\mathcal{U}} := \int_{\theta \in [0, 2\pi)} \text{Re} \left( \hat{u}_1(\theta)^* \hat{u}_2(\theta) \right) d\mu \quad (35)$$

respectively, where $\mathcal{F}$ denotes the transform of functions defined as Eq. (33) and $\hat{u}_1, \hat{u}_2 : [0, 2\pi) \rightarrow \mathbb{C}$ are functions satisfying $u_1 = \mathcal{F}u_1$ and $u_2 = \mathcal{F}u_2$.

Suppose $M$ pressure-gradient microphones are located at $(x_1, y_1), \ldots, (x_M, y_M) \in \mathbb{R}^2$, and let $t_1, \ldots, t_p \in \mathbb{R}$ with a natural number $P$ be the sampling times of the $m$th microphone for each $m \in \{1, \ldots, M\}$. We consider the forward problem in the form of Eq. (1) with

$$L_n u := A_m \frac{\partial}{\partial d(\theta_m)} u(x_m, y_m, t_p)$$

$$\forall n := (m, p) \in \{ (1, 1), \ldots, (M, P_M) \}, \quad (36)$$

$$f_n(z) := \begin{cases} -\tau & z \leq -\tau \\ \tau & z \geq \tau \\ \tau & z \leq z < \tau \\ -\tau & \tau \leq \tau \leq \tau \end{cases}$$

$$\forall n := (m, p) \in \{ (1, 1), \ldots, (M, P_M) \}. \quad (37)$$

Here, $A_m \in \mathbb{R}$ is a gain of the $m$th microphone, $\partial / \partial d(\theta_m)$ denotes the directional derivative along the direction $d(\theta_m) := (\cos \theta_m, \sin \theta_m)$ defined as

$$\frac{\partial}{\partial d(\theta_m)} u(x_m, y_m, t_p) := \left( \cos \theta_m \frac{\partial}{\partial x} + \sin \theta_m \frac{\partial}{\partial y} \right) u(x_m, y_m, t_p). \quad (38)$$

and $\tau \in (0, +\infty]$ is a clipping level. Note that $f_1, \ldots, f_N$ are noninvertible functions for $\tau \neq +\infty$. In this case, $v_n$ and $K_{n_1, n_2}$ are given respectively by

$$v_n(x, y, t) := 2\pi A_n \frac{\omega_0}{c} \int_1^\infty \frac{\omega_0}{c} \sqrt{(x-x_m)^2 + (y-y_m)^2}$$

$$\cdot \cos(\zeta (x-x_m, y-y_m) - \omega_0 (t-t_p))$$

$$\forall x, y, t \in \mathbb{R}, n := (m, p) \in \{ (1, 1), \ldots, (M, P_M) \} \quad (39)$$

and

<table>
<thead>
<tr>
<th>Condition</th>
<th>NMSE: mean ± standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clipped (Proposed)</td>
<td>$-19.67 \pm 3.64$ dB</td>
</tr>
<tr>
<td>Clipped (SC)</td>
<td>$-15.94 \pm 3.87$ dB</td>
</tr>
<tr>
<td>Clipped (RLS)</td>
<td>$-16.84 \pm 2.19$ dB</td>
</tr>
<tr>
<td>Clipped (Proposed+RLS)</td>
<td>$-21.70 \pm 3.25$ dB</td>
</tr>
<tr>
<td>Nonclipped (Proposed)</td>
<td>$-25.66 \pm 3.70$ dB</td>
</tr>
</tbody>
</table>
We conducted numerical experiments using Julia v.1.2.0 whose settings are given in Table 3. Among the five conditions investigated in Sect. 4.1.2, clipped (Proposed and SC) with \( \tau = 0.5 \) and nonclipped (Proposed) with \( \tau = +\infty \) were compared in this experiment because Eq. (30) is not differentiable with respect to the optimization variables. Here, under the clipped (SC) condition, each of \( s_n \) for \( n \in \{1, \ldots, N\} \) was projected into \( S_n \), i.e., \([-\tau, \tau]\), as a preprocessing (note that this preprocessing was unavailable in the experiments in Sect. 4.1.2 since \( S_n \) is an open set in that case). The results from 50 trials are shown in Table 4. The clipping distortion defined as \( 10 \log_{10} \| (Lu) - Lu \|_{2,N}^2 / \| Lu \|_{2,N}^2 \) was also calculated, and its mean \pm standard deviation for 50 trials was \(-10.18 \pm 3.87 \) dB. In addition, we show the true and estimated acoustic fields in the first trial in Fig. 3. Here, pairs of red and blue circles represent the position and orientation of the pressure-gradient microphones; the direction from the blue circle to the red circle corresponds to the direction of the derivative that the pressure-gradient microphone observed. Even under the hard clipping effects, clipped (Proposed) achieved an estimation performance relatively close to that of nonclipped (Proposed), compared with the value of the clipping distortion. We can also see that clipped (Proposed) outperformed clipped (SC), as was the case in the experiments in Sect. 4.1.2.

5. Conclusion

We presented a new useful formulation for ill-posed inverse problems in Hilbert spaces with nonlinear clipping effects. A loss function was designed so that it was convex and differentiable with respect to the optimization variables. By using this formulation in combination with the representer...
In this section, we derived a tractable optimization problem that is easy to optimize using well-established algorithms. Finally, we provided two practical examples of inverse problems with soft and hard clipping effects, and the experimental evaluations confirmed the validity of our proposed formulation.

**Acknowledgements**

This work was supported by JSPS KAKENHI Grant Number JP18J21926, JST PRESTO Grant Number JPMJPR18J4, and SECOM Science and Technology Foundation.

**References**


![Fig. 3: Results of time-harmonic acoustic field estimation at t = 0 s in the first trial. The NMSEs were (b) –17.83 and (c) –6.82 dB. The clipping distortion was –8.07 dB.](image-url)
Table 3: Settings in time-harmonic acoustic field estimation.

| $\omega_0$ | $\omega_0 := 2\pi f_0$ with $f_0 = 100$ Hz. |
| $c$ | $c = 340$ m/s. |
| $u(\alpha_{10} \ldots, \alpha_{0})$ | Sampled independently from the univariate circular-symmetric complex normal distribution with mean 0 and variance 1. |
| $M, (P_1, \ldots, P_M)$ | $M := 12, P_m := 10 \forall m \in \{1, \ldots, M\}$ (N = 120). |
| $(x_1, \ldots, x_M)$, $(y_1, \ldots, y_M)$ | Sampled independently from the uniform distribution on $[-2, 2]$ m. |
| $(\theta_1, \ldots, \theta_M)$ | Sampled independently from the uniform distribution on $[0, 2\pi]$. |
| $(A_1, \ldots, A_M)$ | $A_m = c/\omega_0 \forall m \in \{1, \ldots, M\}$. |
| $(t_1, \ldots, t_M)$ | $t_m = 0.5$ (clipped), $t_m = +\infty$ (nonclipped). |
| Observation noise | Added to each $f_n(Lu)u$ for $n \in \{1, 1, \ldots, (M, P_M)\}$ independently from the univariate real normal distribution with mean 0 and variance $10^{-3} \|f(Lu)\|_{L_2}^2/N$. |
| $\delta, \sigma^2_n$ | $\delta := 10^{-2.5}, \sigma^2_n := 1$ $\forall n \in \{1, 1, \ldots, (M, P_M)\}$. |
| Evaluation criterion | $\text{NMSE} := 10 \log_{10} \text{E} S$, where $E := \sum_{i=1}^{N_k} \sum_{n=1}^{N_k} \sum_{k=1}^{N_k} |u(x, y, t) - u^{(\text{eval})}(x, y, t)|^2$. |
| $(x_1^{(\text{eval})}, \ldots, x_N^{(\text{eval})})$, $(y_1^{(\text{eval})}, \ldots, y_N^{(\text{eval})})$ | Equally spaced points from $-2$ m to $2$ m with intervals of 0.05 m ($N_k = 81$). |
| $(t_1^{(\text{eval})}, \ldots, t_N^{(\text{eval})})$ | Equally spaced points from 0 s to 0.01 s with intervals of 0.001 s ($N_T = 11$). |

Table 4: Results of time-harmonic acoustic field estimation (50 trials).

| Condition | NMSE: mean ± standard deviation |
| Clipped (Proposed) | $-12.33 \pm 5.01$ dB |
| Clipped (SC) | $-6.97 \pm 3.65$ dB |
| Nonclipped (Proposed) | $-17.25 \pm 4.30$ dB |

for arbitrary fixed \( u_0 \in \mathcal{U} \setminus \{0\} \), and then we will prove that \( J \) is coercive, i.e., \( \lim_{t \to +\infty} J(t) = +\infty \) and \( \lim_{t \to -\infty} J(t) = +\infty \). First, the derivative of \( J \), denoted by \( J' \), is given by

\[
J'(t) = \frac{1}{\sigma_n} \left( f_n(tL_n u_0) - s_n L_n u_0 + \delta \| u_0 \|_q^2 \right). \tag{A.6}
\]

Since \( f_n(tL_n u_0)L_n u_0 \) is monotonically increasing with respect to \( t \) for each \( n \in \{1, \ldots, N\} \), \( J' \) is also monotonically increasing, and moreover we obtain

\[
\lim_{t \to -\infty} J'(t) = -\infty, \tag{A.7}
\]

\[
\lim_{t \to +\infty} J'(t) = +\infty. \tag{A.8}
\]

Therefore, we obtain the coerciveness of \( J \) and also \( Q_x \) (where the constant \( C \) is omitted).

**Appendix D: Derivation of representation vector sequence and Gram operator**

Let \( u \in \mathcal{U} \) be represented as Eq. (33) with a square-integrable function \( \hat{u} : [0, 2\pi) \to \mathbb{C} \) and \( \varphi_{\theta_m} \) be defined as

\[
\varphi_{\theta_m}(x, y, t; \theta) := \frac{\partial}{\partial \theta} \varphi(x, y, t; \theta) \quad \forall x, y, t \in \mathbb{R}, m \in \{1, \ldots, M\}. \tag{A.9}
\]

Then, by interchanging the integral and partial derivative (their interchangeability is proved later), we obtain

\[
\frac{\partial}{\partial \theta} u(x, y, t) = \int_{\theta \in [0, 2\pi)} \text{Re}(\hat{u}(\theta)\varphi_{\theta_m}(x, y, t; \theta)) \, d\mu \quad \forall x, y, t \in \mathbb{R}, m \in \{1, \ldots, M\}. \tag{A.10}
\]

Here, for each \( n := (m, p) \in \{(1, 1), \ldots, (M, P_M)\} \), let \( \tilde{v}_n \) and \( v_n \) be respectively defined as

\[
\tilde{v}_n(\theta) := A_m \varphi_{\theta_m}(x_n, y_m, t_p; \theta) \quad \forall \theta \in [0, 2\pi), \tag{A.11}
\]

\[
v_n = \delta \tilde{v}_n, \tag{A.12}
\]

which immediately yield

\[
\langle v_n, u \rangle_{\mathcal{H}} = A_m \int_{\theta \in [0, 2\pi)} \text{Re}(\hat{u}(\theta)\varphi_{\theta_m}(x_n, y_m, t_p; \theta) \varphi(x_n, y_n, t; \theta)) \, d\mu \quad \forall x_n, y_n, t \in \mathbb{R}, \tag{A.13}
\]

\[
= L_n u. \tag{A.14}
\]

Therefore, \( v_n \) and \( K_{n_1, n_2} \) can be obtained as

\[
v_n(x, y, t) = A_m \int_{\theta \in [0, 2\pi)} \text{Re}(\varphi_{\theta_m}(x_n, y_m, t_p; \theta) \varphi(x_n, y_n, t; \theta)) \, d\mu.
\]

The following relation

\[
\langle \tilde{u}_X(x+h) - \tilde{u}_X(x) \rangle = \int_{\theta \in [0, 2\pi)} \text{Re}(\hat{u}(\theta)\varphi_X(x+h, \theta) - \varphi_X(x, \theta)) \, d\mu \leq \left( \int_{\theta \in [0, 2\pi)} |\hat{u}(\theta)|^2 \, d\mu \right)^{\frac{1}{2}} \int_{\theta \in [0, 2\pi)} |\varphi_X(x+h, \theta) - \varphi_X(x, \theta)|^2 \, d\mu. \tag{A.21}
\]
Since \( \varphi_X(x + h, \theta) \) converges to \( \varphi_X(x, \theta) \) as \( h \to 0 \) uniformly in \( \theta \in [0, 2\pi] \), we obtain
\[
\lim_{h \to 0} |\tilde{u}_X(x + h) - \tilde{u}_X(x)| = 0 \quad \forall x \in \mathbb{R}.
\] (A.22)

On the other hand, from the mean value theorem, for any \( h \in \mathbb{R} \setminus \{0\} \), there exists some \( \theta(h) \in [0, 1] \) satisfying
\[
\frac{\varphi(x + h, y, t; \theta) - \varphi(x, y, t; \theta)}{h} = \varphi_X(x + \theta(h)h, \theta).
\] (A.23)

Therefore, we obtain
\[
u_X(x, y, t) = \lim_{h \to 0} \frac{1}{h} \int_{\theta \in [0, 2\pi]} \text{Re} \{\hat{u}(\theta)(\varphi(x + h, y, t; \theta) - \varphi(x, y, t; \theta))\} d\mu
\]
\[
= \lim_{h \to 0} \int_{\theta \in [0, 2\pi]} \text{Re} \{\hat{u}(\theta) \varphi_X(x + \theta(h)h, \theta)\} d\mu
\]
\[
= \lim_{h \to 0} \tilde{u}_X(x + \theta(h)h)
\]
\[
= \tilde{u}_X(x) \quad \forall x \in \mathbb{R}.
\] (A.24)

The same property also holds for a partial derivative with respect to \( y \). Thus, we obtain the interchangeability of the integral and partial derivative in Eq. (A.10).

---

**Natsuki Ueno** received the B.E. degree in engineering from Kyoto University, Kyoto, Japan, in 2016, and the M.S. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2018. He is currently a Ph.D. student at Graduate School of Information Science and Technology, The University of Tokyo. His research interests include spatial audio and acoustic signal processing.

**Shoichi Koyama** received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 2007, 2009, and 2014, respectively. In 2009, he was a Researcher in acoustic signal processing with the Nippon Telegraph and Telephone Corporation. He moved to the University of Tokyo in 2014, and since 2018, he has been an Assistant Professor (Lecturer). From 2016 to 2018, he was also a Visiting Researcher and JSPS overseas research fellow with Paris Diderot University (Paris7), Institut Langevin, Paris, France. His research interests include acoustic inverse problems, sound field analysis and synthesis, and spatial audio. He is a member of the Acoustical Society of America, the Audio Engineering Society, the Institute of Electronics, Information and Communication Engineers, and the Acoustical Society of Japan (ASJ). He was the recipient of Itakura Prize Innovative Young Researcher Award by ASJ in 2015, and the Research Award by Funai Foundation for Information Technology in 2018.

**Hiroshi Saruwatari** received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and Nara Institute of Science and Technology, Japan, in 2000. From 2014, he is currently a Professor of The University of Tokyo, Japan. His research interests include acoustic inverse problems, sound field analysis and synthesis, and spatial audio. He is a member of the Acoustical Society of America, the Audio Engineering Society, the Institute of Electronics, Information and Communication Engineers, and the Acoustical Society of Japan (ASJ). He was the recipient of Itakura Prize Innovative Young Researcher Award by ASJ in 2015, and the Research Award by Funai Foundation for Information Technology in 2018.