Unconstrained Proximal Operator: the Optimal Parameter for the Douglas-Rachford Type Primal-Dual Methods

Yifan Ran, Wei Dai

Abstract

In this work, we propose an alternative parametrized form of the proximal operator, of which the parameter no longer needs to be positive. That is, the parameter can be a non-zero scalar, a full-rank square matrix, or, more generally, a bijective bounded linear operator. We demonstrate that the positivity requirement is essentially due to a quadratic form. We prove several key characterizations for the new form in a generic way (with an operator parameter). We establish the optimal choice of parameter for the Douglas-Rachford type methods by solving a simple unconstrained optimization problem. The optimality is in the sense that a non-ergodic worst-case convergence rate bound is minimized. We provide closed-form optimal choices for scalar and orthogonal matrix parameters under zero initialization. Additionally, a simple self-contained proof of a sharp linear convergence rate for a \((1/L)\)-cocoercive fixed-point sequence with \(L \in (0, 1)\) is provided (as a preliminary result).

To our knowledge, an operator parameter is new. To show its practical use, we design a dedicated parameter for the 2-by-2 block-structured semidefinite program (SDP). Such a structured SDP is strongly related to the quadratically constrained quadratic program (QCQP), and we therefore expect the proposed parameter to be of great potential use. At last, two well-known applications are investigated. Numerical results show that the theoretical optimal parameters are close to the practical optimums, except they are not a priori knowledge. We then demonstrate that, by exploiting problem model structures, the theoretical optimums can be well approximated. Such approximations turn out to work very well, and in some cases almost reach the underlying limits.

Keywords. proximal operator, fixed-point theory, duality, convergence rate, parameter selection, semidefinite program, Douglas-Rachford splitting, Boolean quadratic program, super-resolution problem

1 Introduction

The proximal operator was introduced by Moreau in 1962 [38] as an extension of the projection operator. Such a notion is implicit, which involves solving a smooth optimization problem. It is a powerful tool that introduces a unified analysis framework. By employing it, many seemingly disparate algorithms are shown to be strongly related, see e.g. [13, 47, 1]. The algorithms that adopt the proximal operator as its basic operation are referred to as the proximal algorithms [42, 46]. They are widely applicable and in general highly efficient. In this work, we consider an important class — the Douglas-Rachford Splitting (DRS) [33, 17, 43, 11, 19]. It is equivalent to two other popular methods: (i) the Alternating Direction Method of Multipliers (ADMM) [4, 23, 26], which can be viewed as a dual variant of the DRS, see e.g. [19, 21]; (ii) the Primal-Dual Hybrid Gradient (PDHG) method [6, 20, 44], which is shown to be completely equivalent to the DRS recently in [41].

For the well-known proximal operator, we show that there exists another parametrized form. Its parameter no longer needs to satisfy the positivity requirement (a positive scalar or a positive definite matrix). One important use of the new form is regarding the optimal parameter selection issue. We show that by employing the new form, finding the optimal parameter can be formulated into an unconstrained optimization problem, and one can obtain some closed-form expressions for some basic cases. In view of this, we would refer to the new form as the ‘unconstrained proximal operator’. We demonstrate that the positivity requirement essentially comes from a...
quadratic form. To see this, let us note that the positive scalar requirement $\lambda > 0$ can be rewritten into $\lambda = \alpha^2$ for some $\alpha \in \mathbb{R}/\{0\}$. Similarly, for the matrix case, the positive definite requirement $M \succ 0$ can be rewritten into $M = R^T R$ for some full-rank square matrix $R$. We will show that the generic form of the unconstrained proximal operator is given by $\text{ProxF}_{\alpha S}$, where $(f \circ S)x \overset{\text{def}}{=} \{ y \mid y \in f(Sx) \}$ and where $S$ is a bijective bounded linear operator (can be reduced to a non-zero scalar $\alpha$ or a full-rank square matrix $R$). Moreover, we prove the relation $\text{ProxF}_{\alpha S^{-1}} = S \circ \text{ProxF}_\alpha$ and refer to the right-hand side as the ‘splitting form’. It is useful for evaluation purpose and one can verify that the unconstrained proximal operator has the same evaluation difficulty as the conventional one.

Regarding the optimal choice of parameter, we rely on the worst-case convergence rate bound. Without any strong assumption, the DRS type of methods admit a $O(1/k)$ worst-case linear rate, see some earlier work in [37, 28, 29, 40], and more recently [16, 22, 14, 15, 30, 25, 39, 49]. Here, we also provide a rate characterization, via a fixed-point view. We introduce a $(1/L)$-cocoercive assumption with $L \in (0,1)$. Our proof appears in a very simple form and can recover the basic $O(1/k)$ case by setting $L = 1$ (the proof will be simplified in this case).

We determine the optimal parameter by minimizing the worst-case convergence rate bound. The rate bound is given by $\|\psi^* - \psi^0\|^2$ where $\psi^*$ is the fixed-point and $\psi^0$ is the initialization. We will demonstrate that such a bound is optimizable with respect to the generic parameter $S$, and the corresponding optimization problem is unconstrained. Furthermore, we present two closed-form optimal choices (under zero initialization) corresponding to the scalar parameter case and the orthogonal matrix parameter case.

To our knowledge, an operator parameter is new. To demonstrate its practical use, we consider a 2-by-2 block-structured semidefinite program (SDP). We first prove that our designed operator parameter (via Hadamard product) is feasible. This involves proving a ‘definiteness invariant’ condition, see Proposition 4.7. It is a special case of the general ‘set invariant’ condition associated with the convex set indicator function, see Proposition 3.7. We find that the 2-by-2 block structure appears to be very common. Particularly, it is a natural consequence of applying the semidefinite relaxation (SDR) to the quadratically constrained quadratic program (QCQP), see [35] for an overview of the technique.

On the application side, we consider two well-known problems — the Boolean quadratic program (BQP) [45, 31, 36] in integer programming and the super-resolution [5, 51] for point sources localisation. Both problems are formulated as 2-by-2 block-structured SDP. We numerically test the theoretical optimal choices, which are shown to be close to the underlying practical optimums. However, they are not a priori knowledge. We then show that, by exploiting problem structures, the theoretical choices can be well approximated. Numerical simulations show that these estimates work surprisingly well. In some data setups, they perform better than the theoretical ones and sometimes almost reach the underlying performance limits (the practical optimums).

### 1.1 Related work on the proximal parameter selection

The proximal parameter, which is often referred to as the step-size parameter, is well-known to have a significant impact on the proximal algorithms convergence speed. It is therefore of great interest to answer the question – what is a good choice, or, if possible, the best choice. Extensive research has been conducted to address the issue and can be roughly divided into two types. The first type has more practical concerns. The goal is to design a practically feasible approach for selecting good parameters. A typical one is [40], where the authors consider ADMM as a dynamic system. There, one first obtains a convergence rate bound by solving an SDP. Then, a grid search over the parameter space is conducted to find the parameter choice (that minimizes the rate bound). In [53, 54], the authors adapt the successful Barzilai-Borwein spectral method for gradient descent and derive adaptive parameter choices for ADMM. The other type of work concerns the theoretically optimal choice. The optimality is in the sense that the convergence factor is minimized (see its definition in e.g. [24, eq(1)]). Earlier work on this aspect is limited to specific problems. In [24, 52], the authors consider the quadratic program (QP), and the optimal ADMM parameters are derived under several specific settings. [50] considers the decentralized consensus problems, where strong assumptions, such as the Lipschitz continuous and the strong convexity, are required. Recently, [48] proposed a general framework that computationally selects the optimal parameter by solving a series of SDPs.

Our work is very much different compared to the prior art. Most importantly, our parameter optimality is in a different sense. We consider minimizing the worst-case convergence rate bound, rather than the convergence factor. This leads to investigating the optimal solution structures rather than the objective function structures (e.g. strong convexity, Lipschitz continuity). Such solution related structures are linked to the problem model (the data
that an alternative way can achieve the same purpose, by considering modified 

provides a feasible operator parameter for the 2-by-2 block-structured SDP. At last, 

(i) \( L \)

(ii) a contraction if it is \( L \)-Lipschitz continuous i.e.,

(iii) nonexpansive if it is 1-Lipschitz continuous, i.e.,

(iv) \( 1/L \)-cocoercive if

(v) firmly nonexpansive if it is 1-cocoercive, i.e.,

Lemma 1.1 (1/2-averaging). Let \( T \) be nonexpansive. Then, the composition \( \frac{1}{2} \mathcal{I} + \frac{1}{2} T \) is firmly nonexpansive. Reversely, suppose \( 2 \mathcal{A} - \mathcal{I} \) is nonexpansive. Then, \( \mathcal{A} \) is firmly nonexpansive.
2 Preliminary Result

In the literature, a worst-case $O(\frac{1}{k+1})$ linear rate is established in various ways. Here, we consider the fixed-point view. A related result without strong assumption can be found in the most recent work [49, Sect. 2.4]. In the following contents, we provide a sharp rate by introducing a $1/L$-cocoercive assumption (see Definition 1.1 (iv)) with constant $L \in (0, 1)$. Our analysis also recovers the basic $O(\frac{1}{k+1})$ rate by setting $L = 1$ (with some simplifications).

Our rate is given by $O(a^k \cdot \frac{1-a}{1-a^{k+1}})$ where $a \overset{\text{def}}{=} \frac{L}{2-L}$, and is guaranteed to be better than $O(\frac{1}{k+1})$. Let us note that by assumption $L \in (0, 1)$, hence $a < L$. One can easily verify that unless $a$ is very close to 1, $a^k$ would dominate the other term, which gives a $O(a^k)$ rate. Let us note that $L = 1$ corresponds to the basic (no strong assumption) case. Then, when $L$ is very close to 1, say 0.99, we may refer to it as almost no strong assumption. A natural guess is that the rate bound value in this case would be roughly the same as when $L = 1$, i.e., the classical analysis would be sufficient. However, below by two numerical examples, we show that there exists notable difference. This justifies the use of our analysis.

- Let $L = 0.99$ and the iteration number $k = 20$. Then, one obtains $a^k \cdot \frac{1-a}{1-a^{k+1}} \approx 0.0387$ and $\frac{1-a}{k+1} \approx 0.0476$.
- Let $L = 0.99$ and the iteration number $k = 100$. Then, one obtains $a^k \cdot \frac{1-a}{1-a^{k+1}} \approx 0.0031$ and $\frac{1}{k+1} \approx 0.0099$.

2.1 A sharp convergence rate

Proposition 2.1 (Strictly decreasing error). Suppose an operator $\mathcal{F}$ is $1/L$-cocoercive with $L \in (0, 1)$. For a set of fixed-point iterates $\{y^k\}$ that converges to $y^* \in \text{Fix}(\mathcal{F})$, the following holds:

$$\frac{2 - L}{L} \|y^{k+1} - y^*\|^2 - \|y^k - y^*\|^2 \leq -\|y^{k+1} - y^k\|^2.$$  \hspace{1cm} (2.1)

Proof. By Definition 1.1 (iv), one has

$$\|\mathcal{F}y^k - \mathcal{F}y^*\|^2 \leq L \langle \mathcal{F}y^k - \mathcal{F}y^*, y^k - y^* \rangle$$  \hspace{1cm} (2.2)

\[ \iff \|y^{k+1} - y^*\|^2 \leq L \langle y^{k+1} - y^*, y^k - y^* \rangle \]  \hspace{1cm} (2.3)

\[ \iff 0 \leq L \langle y^{k+1} - y^*, y^k - y^{k+1} \rangle + (L - 1) \|y^{k+1} - y^*\|^2 \]  \hspace{1cm} (2.4)

\[ \iff 0 \leq L \|y^k - y^*\|^2 - L \|y^{k+1} - y^*\|^2 - L \|y^{k+1} - y^k\|^2 + 2(L - 1) \|y^{k+1} - y^*\|^2 \]  \hspace{1cm} (2.5)

\[ \iff \frac{2 - L}{L} \|y^{k+1} - y^*\|^2 - \|y^k - y^*\|^2 \leq -\|y^{k+1} - y^k\|^2, \]  \hspace{1cm} (2.6)

where the Pythagoras relation

$$2 \langle a - b, c - a \rangle = \|c - b\|^2 - \|a - b\|^2 - \|c - a\|^2$$  \hspace{1cm} (2.7)

is applied to obtain (2.5). The proof is therefore concluded.

Proposition 2.2 (Sharp rate of convergence). Suppose an operator $\mathcal{F}$ is $1/L$-cocoercive with $L \in (0, 1)$. For a set of fixed-point iterates $\{y^k\}$ that converges to $y^* \in \text{Fix}(\mathcal{F})$, it admits a non-ergodic worst-case convergence rate as

$$\|y^{k+1} - y^k\| \leq \left(\frac{L}{2 - L}\right)^k \cdot \frac{1 - \frac{L}{2 - L}}{1 - \left(\frac{L}{2 - L}\right)^{k+1}} \|y^* - y^0\|^2.$$  \hspace{1cm} (2.8)

Proof. For light of notation, let $a \overset{\text{def}}{=} \frac{L}{2-L}$. By Proposition 2.1, we obtain

$$\|y^{t+1} - y^t\|^2 \leq \|y^t - y^*\|^2 - \frac{1}{a} \|y^{t+1} - y^*\|^2.$$  \hspace{1cm} (2.9)

This gives

$$\sum_{t=0}^{k} a^{-t} \|y^{t+1} - y^t\|^2 \leq \|y^* - y^0\|^2.$$  \hspace{1cm} (2.10)
By Definition 1.1 (i), one has
\[ \| y^{k+1} - y^k \|^2 \leq \| y^k - y^{k-1} \|^2. \]  
(2.11)

It follows that
\[ \sum_{t=0}^{k} a^{-t} \| y^{k+1} - y^k \|^2 \leq \sum_{t=0}^{k} a^{-t} \| y^{k+1} - y^t \|^2, \]  
(2.12)
\[ \iff \frac{1 - a^{-(k+1)}}{1 - a^{-1}} \| y^{k+1} - y^k \|^2 \leq \sum_{t=0}^{k} a^{-t} \| y^{k+1} - y^t \|^2, \]  
(2.13)

where we have used the geometric sum formula in the last line and let us note that the quantity \( \| y^{k+1} - y^k \|^2 \) is a constant under \( \sum_t \).

Adding (2.10) and (2.13) yields the convergence rate bound
\[ \| y^{k+1} - y^k \|^2 \leq a^{k} \cdot \frac{1 - a}{1 - a^{k+1}} \| y^* - y^0 \|^2. \]  
(2.14)

Additionally, note that
\[ \sum_{t=0}^{k} a^{-t} > k + 1. \]  
(2.15)

We have the above rate bound satisfies
\[ a^{k} \cdot \frac{1 - a}{1 - a^{k+1}} \| y^* - y^0 \|^2 < \frac{1}{k + 1} \| y^* - y^0 \|^2, \]  
(2.16)

and is therefore a sharp rate. The proof is now concluded.

# 3 Unconstrained Proximal Operator

In this part, we introduce an alternative parametrized form of the well-known proximal operator, where the requirement of its parameter being positive is removed. First, we recall the following basic definitions:

**Definition 3.1.** Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the conventional proximal operator with a scalar parameter \( \lambda \) is given by
\[
\text{Prox}_\lambda f(\cdot) \equiv \arg\min_z f(z) + \frac{1}{2\lambda} \| z - \cdot \|^2,
\]  
(3.1)
where the scalar parameter needs to be positive, i.e., \( \lambda > 0 \).

**Definition 3.2.** Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the conventional proximal operator with a matrix parameter \( H \) is given by
\[
\text{Prox}_H f(\cdot) \equiv \arg\min_z f(z) + \frac{1}{2} \| z - \cdot \|^2_H,
\]  
(3.2)
where the matrix parameter needs to be positive definite, i.e., \( H \succ 0 \), and where \( \| \cdot \|_H \) is a norm induced by the inner product \( \langle \cdot, \cdot \rangle_H \equiv \langle \cdot, H \cdot \rangle \).

Below, we present the unconstrained proximal operator with a generic operator parameter. It can be straightforwardly reduced to the matrix parameter case and the scalar parameter case.

**Proposition 3.1** (Unconstrained proximal operator). Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the unconstrained proximal operator with an operator parameter \( S \) is given by
\[
\text{Prox}_f \circ S(\cdot) \equiv \arg\min_z f \circ S(z) + \frac{1}{2} \| z - \cdot \|^2,
\]  
(3.3)
where the parameter \( S \in \mathcal{B}(\mathcal{H}) \) is bijective, linear and bounded.

**Remarks 3.1** (symmetry). It can be seen from above that the conventional proximal operator consists of a left multiplication \( \lambda f(x) \equiv \{ \lambda y \mid y = f(x) \} \), while the proposed new form involves a right multiplication \( f \circ S(x) \equiv \{ y \mid y = f(Sx) \} \).
3.1 Key Properties

In this part, we prove several important characterizations for the unconstrained proximal operator.

3.1.1 Firm nonexpansiveness

The firm nonexpansiveness is a key property to ensure the proximal algorithms convergence. Here, we prove it for the unconstrained proximal operator. To start, we first introduce the following two lemmas.

Lemma 3.1. [1, Coro. 16.53] Let \( f \in \Gamma_0(\mathcal{H}) \) and \( L \in \mathcal{B}(\mathcal{H}) \). Suppose \( 0 \in \text{sri}(\text{dom}(f) - \text{ran}(L)) \). Then,
\[
\partial(f \circ L) = L^* \circ \partial f \circ L.
\]

Lemma 3.2. [1, Prop. 23.25] Let \( A: \mathcal{H} \rightarrow 2^{\mathcal{H}} \) be maximal monotone. Suppose \( L \in \mathcal{B}(\mathcal{H}) \) is such that \( LL^* \) is invertible. Let \( B = L^*AL \).

(i) \( B: \mathcal{H} \rightarrow 2^{\mathcal{H}} \) is maximal monotone;
(ii) The resolvent \( J_B = I - L^* \circ (LL^* + A^{-1})^{-1} \circ L \).

Now, we are ready to establish the firm nonexpansiveness property.

Proposition 3.2. Let parameter \( S \in \mathcal{B}(\mathcal{H}) \) be bijective. Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), one has
(i) \( \text{Prox}_{f \circ S} = (I + S^* \circ \partial f \circ S)^{-1} \);
(ii) \( \text{Prox}_{f \circ S} \) is firmly nonexpansive and single-valued.

Proof. First, by definition, the following resolvent characterization holds:
\[
\text{Prox}_{f \circ S} = (I + \partial(f \circ S))^{-1}. \tag{3.4}
\]

Since \( S \) is bijective, we have \( \text{ran}(S) = \mathcal{H} = \text{dom}(f) \). Invoking Lemma 3.1, one obtains relation (i).

Given a CCP function \( f \), it is well-known that the subdifferential operator \( \partial f \) is maximal monotone. Since \( S \) is bijective, we have that \( SS^* \) is invertible. In view of Lemma 3.2, the composite operator \( S^* \circ \partial f \circ S \) is therefore maximal monotone. Hence, the operator \( (I + S^* \circ \partial f \circ S)^{-1} \) is firmly nonexpansive and single-valued. The proof is now concluded. \( \square \)

3.1.2 The splitting form

First, we introduce an extended proximal operator.

Definition 3.3 (Extended proximal operator). Let the operator \( A \in \mathcal{B}(\mathcal{H}) \) be injective. Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the extended proximal operator is defined as
\[
\text{Prox}_f^A(z) \overset{\Delta}{=} \arg\min_{z} f(z) + \frac{1}{2} \|Az - \cdot\|^2. \tag{3.5}
\]

Now, we are ready to show what we call ‘the splitting form’.

Proposition 3.3 (The splitting form). Let the parameter \( S \in \mathcal{B}(\mathcal{H}) \) be bijective. Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), one has
\[
\text{Prox}_{f \circ S} = S \circ \text{Prox}_f. \tag{3.6}
\]

Furthermore, we have the following characterizations:
(i) \( \text{Prox}_{f \circ S} = (I + (S^*)^{-1} \circ \partial f \circ S^{-1})^{-1} = I - (I + S \circ \partial f \circ S^*)^{-1} \);
(ii) \( S \circ \text{Prox}_f^S = S \circ (S^* + \partial f)^{-1} \circ S^* = I - (S^*)^{-1} \circ ((S^*S)^{-1} + \partial f^*)^{-1} \circ S^{-1} \).
Proof. It follows from Proposition 3.2 that
\[ \text{Prox}_{f \circ S^{-1}} = (I + (S^*)^{-1} \circ \partial f \circ S^{-1})^{-1}. \] (3.7)
Since \( S \in \mathfrak{B}(\mathcal{H}) \) is bijective, one obtains
\[ \text{Prox}_{f \circ S^{-1}} = (S^* + \partial f \circ S^{-1})^{-1} \circ S^* = S \circ (S^* S + \partial f)^{-1} \circ S^*. \] (3.8)
Let \( x \equiv \text{Prox}_f^S(v) \). Then, by definition
\[ S^* v - S^* S x \in \partial f(x) \iff S^* v \in (S^* S + \partial f)x \iff x = (S^* S + \partial f)^{-1} S^* v. \] (3.9)
This gives
\[ \text{Prox}_f^S = (S^* S + \partial f)^{-1} \circ S^*. \] (3.10)
Comparing it to (3.8) establishes relation (3.6).

The conjugate function \( f^* \) related characterizations follow directly from Lemma 3.2 (ii) by noting that \((\partial f)^{-1} = \partial f^* \). The proof is therefore concluded. \( \square \)

3.1.3 Removing positivity: the quadratic form

Here, we show that the conventional positivity requirement is due to a quadratic form and can be removed. For generality, we first incorporate an operator parameter into the conventional proximal operator.

Definition 3.4 (Generalised conventional proximal operator). Let \( M, S \in \mathfrak{B}(\mathcal{H}) \) be bijective and \( M \equiv S^* S \). Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the conventional proximal operator with a parameter \( M \) can be defined as
\[ \text{Prox}_{M \circ f}(\cdot) \equiv \arg\min f(z) + \frac{1}{2} \| z - \cdot \|^2_M, \] (3.11)
where \( \| \cdot \|_M \) is an inner product induced norm \( \langle \cdot, \cdot \rangle_M \equiv \langle \cdot, M \cdot \rangle \).

Now we are ready to state the relation between the conventional proximal operator and the proposed new form.

Proposition 3.4 (The quadratic relation). Let \( M, S \in \mathfrak{B}(\mathcal{H}) \) be bijective and \( M \equiv S^* S \). Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the following relation holds:
\[ \text{Prox}_{M \circ f}(v) = \text{Prox}_f^S(Sv). \] (3.12)
Proof. By definition,
\[ \text{Prox}_{M \circ f} = (M + \partial f)^{-1} \circ M. \] (3.13)
Recall from (3.10) that
\[ \text{Prox}_f^S = (S^* S + \partial f)^{-1} \circ S^*. \] (3.14)
Hence, we obtain
\[ \text{Prox}_f^S(Sv) = (S^* S + \partial f)^{-1} S^* S v = (M + \partial f)^{-1} \circ M v = \text{Prox}_{M \circ f}(v), \] (3.15)
which concludes the proof. \( \square \)
3.1.4 Moreau identity

In this part, we present the well-known Moreau identity in terms of the unconstrained proximal operator. First, the non-parametrized version of the Moreau identity is given by

\[ v = \text{Prox}_f(v) + \text{Prox}_{f^*}(v). \]  \hfill (3.16)

Its extension to include a positive scalar parameter \( \lambda > 0 \) is given by

\[ v = \text{Prox}_{\lambda f}(v) + \lambda \text{Prox}_{\lambda^{-1}f^*}(\lambda^{-1}v). \]  \hfill (3.17)

And the extension to include a positive definite matrix parameter \( H \succ 0 \) is given by (see e.g. [34, 12])

\[ v = \text{Prox}_{Hf}(v) + H^{-1}\text{Prox}_{H^{-1}f^*}(Hv). \]  \hfill (3.18)

Below, we provide the Moreau identity in terms of the unconstrained proximal operator. We will show it in two ways. First, consider the following lemma:

**Lemma 3.3.** [1, Prop. 13.23 (iv)] Let \( S \in \mathcal{B}(\mathcal{H}) \) be bijective. Then,

\[ (f \circ S^{-1})^* = f^* \circ S^*. \] \hfill (3.19)

For the basic Moreau identity as in (3.16), substitute the function \( f \) by \( f \circ S^{-1} \), and invoke Lemma 3.3. We obtain

\[ v = \text{Prox}_{f \circ S^{-1}}(v) + \text{Prox}_{f^* \circ S}(v) = S \circ \text{Prox}_f^S(v) + (S^*)^{-1} \circ \text{Prox}_{f^*}^{(S^*)^{-1}}(v), \] \hfill (3.20)

where the last equality follows from Proposition 3.3.

Alternatively, we can arrive at the above relation by definition of the proximal operator.

**Proposition 3.5.** Let \( S \in \mathcal{B}(\mathcal{H}) \) be bijective. Given a CCP function \( f \in \Gamma_0(\mathcal{H}) \), the following relation holds:

\[ v = S \circ \text{Prox}_f^S(v) + (S^*)^{-1} \circ \text{Prox}_{f^*}^{(S^*)^{-1}}(v). \] \hfill (3.21)

**Proof.** Let \( x \overset{\text{def}}{=} \text{Prox}_f^S(v) \), \( y \overset{\text{def}}{=} \text{Prox}_{f^*}^{(S^*)^{-1}}(v) \). Showing (3.21) is equivalent to proving

\[ v = Sx + (S^*)^{-1}y \iff y = S^*(v - Sx). \] \hfill (3.22)

By definition, \( x \overset{\text{def}}{=} \arg\min_z f(z) + \frac{1}{2}\|S(z) - v\|^2 \). This implies

\[ S^*(v - Sx) \in \partial f(x) \iff x \in \partial f^*(S^*v - S^*Sx) \]
\[ \iff S^{-1}(v - v + Sx) \in \partial f^*(S^*v - S^*Sx) \]
\[ \iff S^{-1}(v - (S^*)^{-1}S^*(v - S^*x)) \in \partial f^*(S^*v - S^*Sx). \] \hfill (3.23)

By definition, \( y \overset{\text{def}}{=} \arg\min_z f^*(z) + \frac{1}{2}\|(S^*)^{-1}z - v\|^2 \), which gives

\[ S^{-1}(v - (S^*)^{-1}y) \in \partial f^*(y). \] \hfill (3.24)

Comparing (3.23) and (3.24), we obtain the relation

\[ y = S^*(v - Sx). \] \hfill (3.25)

That is, we have shown (3.22). The proof is therefore concluded. \( \square \)

3.2 Evaluation with generic parameter: two closed-form expressions

For the non-parametrized or the scalar-parametrized proximal operator, the closed-form expressions are well-established, see e.g. [13, 2]. However, for a more general parameter (typically the matrix parameter), the closed-form expressions are rarely available. In the literature, researchers therefore often consider to approximately evaluate the proximal operator. This requires an additional error tolerance assumption, see [8, 9, 10, 19, 46] for some theoretical results, and [7, 34] for some applications.

Here, we show two special cases where, by introducing additional parameter restrictions, the closed-form expressions are available.
3.2.1 $l_1$ norm

Consider the sub-differential of the $l_1$ norm. It admits the following characterization:

$$\partial \|x\|_1 = \{ \text{sgn}(x) + w | x \odot w = 0, \|w\|_\infty \leq 1 \}. \quad (3.26)$$

With the above characterization, we obtain the result below.

**Proposition 3.6.** Let $D$ be a full-rank square matrix with orthogonal columns. Then,

$$(D^TD)^{-1}T(D^Tv) = \text{Prox}_D^{\|\cdot\|_1}(v) \triangleq \text{argmin}_x \|x\|_1 + \frac{1}{2}\|Dx - v\|^2,$$  \quad (3.27)

where $T$ is a soft thresholding operator, defined as

$$T(u) \triangleq \text{sgn}(u) \odot |u - 1|_+,$$

and where $| \cdot |$ denotes the absolute value; $1$ denotes a ones vector; $\{ \cdot \}_+$ is the operation of setting all negative entries to zeros.

**Proof.** By definition, (3.27) holds if and only if

$$D^Tv - T(D^Tv) \in \partial \|(D^TD)^{-1}T(D^Tv)\|_1.$$  \quad (3.28)

Our goal is to prove the above relation.

To start, define the following decomposition:

$$D^Tv \triangleq y_0 + y_1, \quad (3.29)$$

where $y_0$ (resp., $y_1$) has the absolute value of all its non-zero entries larger than 1 (resp., less or equal to 1). Such a decomposition implies the relation $y_0 \odot y_1 = 0$. Applying the soft thresholding operator $T$ to above yields

$$T(D^Tv) = y_0 - \text{sgn}(y_0). \quad (3.30)$$

Then, the left hand side of (3.28) can be written as

$$D^Tv - T(D^Tv) = \text{sgn}(y_0) + y_1. \quad (3.31)$$

Now consider the right-hand side of (3.28). First, recall the subdifferential characterization of the $l_1$ norm

$$\partial \|x\|_1 = \{ \text{sgn}(x) + w | x \odot w = 0, \|w\|_\infty \leq 1 \}. \quad (3.32)$$

Since by definition $\|y_1\|_\infty \leq 1$ and $y_0 \odot y_1 = 0$, we have

$$\text{sgn}(y_0) + y_1 \in \partial \|y_0\|_1. \quad (3.33)$$

Substituting (3.31) to the above yields

$$D^Tv - T(D^Tv) \in \partial \|y_0\|_1.$$  \quad (3.34)

Comparing the above to the final goal (3.28), all what left is to show

$$\partial \|y_0\|_1 = \partial \|(D^TD)^{-1}T(D^Tv)\|_1.$$  \quad (3.35)

This is equivalent to showing

$$\text{sgn}(y_0) = \text{sgn}((D^TD)^{-1}T(D^Tv)). \quad (3.36)$$

To do so, since by assumption $D$ has orthogonal columns, let

$$(D^TD)^{-1} \triangleq \text{Diag}(z), \quad (3.37)$$

where $z$ is a positive vector. The right-hand side of (3.36) can be written into

$$\text{sgn}(\text{Diag}(z)(y_0 - \text{sgn}(y_0))) = \text{sgn}(z \odot \text{sgn}(y_0) \odot (|y_0| - 1)) = \text{sgn}(y_0) \odot \text{sgn}(z \odot (|y_0| - 1)). \quad (3.38)$$

Since by definition $(|y_0| - 1)$ and $z$ are positive vectors, their Hadamard product therefore yields a positive vector. That said,

$$\text{sgn}(y_0) \odot \text{sgn}(z \odot (|y_0| - 1)) = \text{sgn}(y_0) \odot 1 = \text{sgn}(y_0), \quad (3.39)$$

which proves (3.36). The proof is therefore concluded. \qed
3.2.2 Set indicator function

Let $C$ be a non-empty convex subset of $\mathcal{H}$. Its associated indicator function is given by

$$\delta_C(x) \overset{\text{def}}{=} \begin{cases} 0 & x \in C, \\ +\infty & x \notin C. \end{cases}$$ (3.40)

Below, we show that if a generic parameter satisfies a ‘set invariant’ condition, then one has a closed-form expression.

**Proposition 3.7.** Let $S \in \mathcal{B}(\mathcal{H})$ be bijective. Suppose the following ‘set invariant’ condition holds:

$$S(x) \in C, \forall x \in C \quad \text{and} \quad S(x) \notin C, \forall x \notin C$$ (3.41)

Then,

$$\Pi_C = \text{Prox}_{\delta_C \circ S}.$$ (3.42)

**Proof.** By definition

$$\text{Prox}_{\delta_C \circ S}(\cdot) \overset{\text{def}}{=} \arg\min \delta_C \circ S(x) + \frac{1}{2}\|x - \cdot\|^2.$$ (3.43)

Suppose the set invariant condition (3.41) holds. Then,

$$\delta_C \circ S(x) = \begin{cases} 0 & x \in C, \\ +\infty & x \notin C. \end{cases}$$ (3.44)

That said, the composition $\delta_C \circ S$ reduces to $\delta_C$. Hence,

$$\text{Prox}_{\delta_C \circ S} = \text{Prox}_{\delta_C} = \Pi_C.$$ (3.45)

The proof is therefore concluded.

**Remarks 3.2 (applicability).** From above, we see that the proximal operator reduces to the projection operator and the parameter vanishes. Clearly, the above result is useless for the case where the objective function only consists of the indicator functions (of different convex sets). Apart from this special case, the above result can be useful. In Section 4.6, we show that the ‘set invariant’ condition reduces to the ‘definiteness invariant’ condition for SDP, and an operator parameter is designed based on it.

4 Primal-Dual Proximal Algorithms

In this section, we consider the following composite problem:

$$\text{minimize} \quad f(x) + g(x),$$ (P)

where the following basic assumptions are imposed:

- (i) $f, g \in \Gamma_0(\mathcal{H})$;
- (ii) ri(dom($f$) $\cap$ dom($g$)) and the solution set are nonempty.

The Fenchel-Rockafellar dual of (P) is given by

$$\text{maximize} \quad -f^*(-\lambda) - g^*(\lambda).$$ (D)
4.1 Equivalent Algorithms: an Augmented Lagrangian View

In this section, we show how the unconstrained proximal operator can be accommodated to the DRS type of algorithms. We consider these algorithms unifiedly via an extended augmented Lagrangian.

To this end, we first rewrite problem (Ψ) into a constrained form

\[
\begin{align*}
\text{minimize} & \quad f(x) + g(z), \\
\text{subject to} & \quad x - z = 0.
\end{align*}
\]

Let \( S \in \mathcal{B}(\mathcal{H}) \) be bijective. Consider the following augmented Lagrangian:

\[
L_S(x, z; \lambda) = f(x) + g(z) + \langle \lambda, x - z \rangle + \frac{1}{2}||S(x - z)||^2.
\]

Its saddle point can be found via the following iterates, which correspond to the Alternating Direction Method of Multipliers (ADMM) algorithm.

\[
\begin{align*}
x^{k+1} &= \text{argmin}_x L_S(x, z^k; \lambda^k) = \text{Prox}^S_\lambda(Sz^k - (S^*)^{-1}\lambda^k), \\
z^{k+1} &= \text{argmin}_z L_S(x^{k+1}, z; \lambda^k) = \text{Prox}^S_\lambda(Sx^{k+1} + (S^*)^{-1}\lambda^k), \\
\lambda^{k+1} &= \lambda^k + S^*S(x^{k+1} - z^{k+1}).
\end{align*}
\]

(ADMM)

Introducing a new point \( \psi^{k+1} \equiv Sz^{k+1} + (S^*)^{-1}\lambda^k \), we can rewrite the above into the following 4-point form:

\[
\begin{align*}
x^{k+1} &= \text{Prox}^S_\lambda(2Sz^k - \psi^k), \\
\psi^{k+1} &= Sz^{k+1} + (S^*)^{-1}\lambda^k, \\
z^{k+1} &= \text{Prox}^S_{\lambda^k}(\psi^{k+1}), \\
\lambda^{k+1} &= S^*(\psi^{k+1} - Sz^{k+1}).
\end{align*}
\]

(4-point)

Then, one can show that the sequence \( \{\psi^k\} \) converges to a fixed-point and therefore establishes the algorithm convergence.

**Proposition 4.1** (fixed-point characterization). The algorithm in (4-point) admits the following fixed-point characterization and the fixed-point iterates converge to a fixed-point (if it exists):

\[
\psi^{k+1} = F^{S}_\lambda \psi^k,
\]

where

\[
F^{S}_\lambda \equiv \frac{1}{2}I + \frac{1}{2}(2S \circ \text{Prox}^S_\lambda - I) \circ (2S \circ \text{Prox}^S_\lambda - I).
\]

**Proof.** By Proposition 3.2, \( \text{Prox}_{f \circ S^{-1}} \) is firmly nonexpansive. Then, it follows from Proposition 3.3 that \( S \circ \text{Prox}^S_\lambda \) is firmly nonexpansive. In view of Lemma 1.1, the operator \( F^{S}_\lambda \) is 1/2-averaged and thus firmly nonexpansive. The firm nonexpansiveness of the fixed-point operator \( F^{S}_\lambda \) then implies convergence. The proof is now concluded. \( \square \)

4.1.1 2-point DRS

Here we show that the well-known DRS algorithm can be easily derived from (4-point). Eliminating variables \( x, \lambda \) from (4-point), one obtains

\[
\begin{align*}
\psi^{k+1} &= S\text{Prox}^S_\lambda(2Sz^k - \psi^k) + \psi^{k+1} - Sz^{k+1}, \\
z^{k+1} &= \text{Prox}^S_{\psi^{k+1}}(\psi^{k+1}).
\end{align*}
\]

(4.5)

Let \( y^{k+1} \equiv \text{Prox}^S_{\psi^{k+1}}(\psi^{k+1}) \). Then,

\[
\begin{align*}
y^{k+1} &= \text{Prox}^S_{\psi^{k+1}}(\psi^{k+1}), \\
\psi^{k+1} &= S\text{Prox}^S_\lambda(2Sy^{k+1} - \psi^k) + \psi^{k} - Sy^{k+1}.
\end{align*}
\]

(2-point DRS)
4.1.2 Primal-dual iterations
Applying different variable eliminations, one can obtain some other formulations. Below we show one that resembles
the PDHG iterates. Eliminating variables $z, \psi$ from \textbf{(4-point)}, one has
\begin{align*}
x^{k+1} &= \text{Prox}_f^S(Sx^k + (S^*)^{-1}(\lambda^{k-1} - 2\lambda^k)), \\
\lambda^{k+1} &= \text{Prox}_{S^*}^g(Sx^{k+1} + (S^*)^{-1}\lambda^k),
\end{align*}
(PD)

4.1.3 Primal-dual fixed-point iterations
For our later convenience regarding the optimal parameter issue, here we provide another formulation which consists
of the primal-dual solution pair and the fixed-point. Eliminating variable $z$ from \textbf{(4-point)} gives
\begin{align*}
x^{k+1} &= \text{Prox}_f^S(\psi_k - 2(S^*)^{-1}\lambda^k), \\
\psi^{k+1} &= Sx^{k+1} + (S^*)^{-1}\lambda^k, \\
\lambda^{k+1} &= S^*(\psi^{k+1} - \text{SProx}_g^S(\psi^{k+1}))
\end{align*}
(PDF)

4.2 Optimality condition
In this part, we present a single optimality criterion for the DRS type of algorithms. It can be of particular interest
to ADMM, as the original criterion in [4, Sect. 3.3] is evaluated in a sophisticated manner (multiple conditions and
error thresholds). Without loss of generality, we prove our result using the iterates from \textbf{(4-point)}.

**Proposition 4.2** (DRS optimality condition). The following relation holds for \textbf{(4-point)}:
\begin{align*}
x^{k+1} = z^k &\iff 0 \in \partial f(x^{k+1}) + \partial g(x^{k+1}). 
\end{align*}
\textbf{(4.6)}

**Proof.** Suppose $x^{k+1} = z^k$ holds. By the update rule of $x$, we have
\begin{align*}
x^{k+1} &= \arg\min_x f(x) + \frac{1}{2}\|Sx - (2Sz^k - \psi^k)\|^2. 
\end{align*}
\textbf{(4.7)}

This implies
\begin{align*}
S^*(2Sz^k - \psi^k - Sx^{k+1}) &\in \partial f(x^{k+1}) 
\end{align*}
\textbf{(4.8)}

By the update rule of $z$
\begin{align*}
z^k &= \arg\min_z g(z) + \frac{1}{2}\|Sz - \psi^k\|^2, 
\end{align*}
\textbf{(4.9)}

we obtain
\begin{align*}
S^*(\psi^k - Sz^k) &\in \partial g(z^k). 
\end{align*}
\textbf{(4.10)}

Adding \textbf{(4.8)}, \textbf{(4.10)} gives
\begin{align*}
S^*(Sz^k - Sx^{k+1}) &\in \partial f(x^{k+1}) + \partial g(z^k). 
\end{align*}
\textbf{(4.11)}

Invoking the assumption $x^{k+1} = z^k$, we arrive at
\begin{align*}
0 &\in \partial f(x^{k+1}) + \partial g(x^{k+1}) 
\end{align*}
\textbf{(4.12)}

Conversely, suppose $0 \in \partial f(x^{k+1}) + \partial g(x^{k+1})$ holds. That said, the algorithm has converged. Hence, we obtain
\begin{align*}
\psi^{k+1} &= \psi^k. 
\end{align*}

Then, by the update rule of $\psi$
\begin{align*}
\psi^{k+1} &= Sx^{k+1} + (S^*)^{-1}\lambda^k = Sx^{k+1} + \psi^k - Sz^k, 
\end{align*}
\textbf{(4.13)}

and applying $\psi^{k+1} = \psi^k$, one obtains
\begin{align*}
x^{k+1} &= z^k. 
\end{align*}
\textbf{(4.14)}

The proof is now concluded. \hfill \square
4.3 Implementation

Below, we summarize the implementation details of the formulations in previous sections.

Algorithm 1  ADMM via the unconstrained proximal operator

1: Choose a bijective parameter $S$, an error threshold $\epsilon$, and arbitrary $z^0, \lambda^0$.
2: while $\|x^{k+1} - z^k\| > \epsilon_1$ do
3: $x^{k+1} = \text{Prox}_S^f(Sz^k - (S^*)^{-1}\lambda^k))$,
4: $z^{k+1} = \text{Prox}_g^S(Sx^{k+1} + (S^*)^{-1}\lambda^k)$,
5: $\lambda^{k+1} = \lambda^k + S^*S(x^{k+1} - z^{k+1})$.
6: end while

Algorithm 2  DRS via the unconstrained proximal operator

1: Choose a bijective parameter $S$, an error threshold $\epsilon$, and arbitrary $\psi^0$.
2: while $\|\psi^{k+1} - \psi^k\| > \epsilon$ do
3: $y^{k+1} = \text{Prox}_g^S(\psi^k)$,
4: $\psi^{k+1} = S\text{Prox}_f^S(2Sy^{k+1} - \psi^k) + \psi^k - Sy^{k+1}$.
5: end while

Algorithm 3  PD via the unconstrained proximal operator

1: Choose a bijective parameter $S$, an error threshold $\epsilon$, and arbitrary $x^0, \lambda^{-1}, \lambda^0$.
2: while $\|S(x^{k+1} - x^k) + (S^*)^{-1}(\lambda - \lambda^{-1})\| > \epsilon$ do
3: $x^{k+1} = \text{Prox}_f^S(Sx^k + (S^*)^{-1}(\lambda^{-1} - 2\lambda))$,
4: $\lambda^{-1} = \text{Prox}_g^S((S^*)^{-1}(Sx^{k+1} + (S^*)^{-1}\lambda))$.
5: end while

Algorithm 4  PDF via the unconstrained proximal operator

1: Choose a bijective parameter $S$, an error threshold $\epsilon$, and arbitrary $\psi^0, \lambda^0$.
2: while $\|\psi^{k+1} - \psi^k\| > \epsilon$ do
3: $x^{k+1} = \text{Prox}_f^S(\psi^k - 2(S^*)^{-1}\lambda^k)$,
4: $\psi^{k+1} = Sx^{k+1} + (S^*)^{-1}\lambda$,
5: $\lambda^{k+1} = S^*(\psi^{k+1} - S\text{Prox}_g^S(\psi^{k+1}))$.
6: end while

Remarks 4.1 (initialization). It is worth mentioning that while theoretically the algorithms can converge with arbitrary initialization, the convergence speed can also be arbitrarily slow. A common choice is a zero initialization.

4.4 An alternative view

In previous sections, we start with an extended augmented Lagrangian and illustrate how the unconstrained proximal operator can be applied to the DRS type of algorithms. Here, we show an alternative way via modified primal and dual problems to achieve the same purpose.

4.4.1 Modified primal and dual problems

The primal problem $(P)$ can be rewritten into

$$\min_x f \circ S^{-1}(Sx) + g \circ S^{-1}(Sx).$$

where we impose the following basic assumptions:

- (i) $S \in \mathcal{B}(\mathcal{H})$ is bijective;
- (ii) $f, g \in \Gamma_0(\mathcal{H})$;
- (iii) $\text{ri}(\text{dom}(f) \cap \text{dom}(g))$ and the solution set are nonempty.
To derive the dual problem, rewrite \((\mathcal{P}_S)\) into
\[
\begin{align*}
\text{minimize} & \quad f \circ S^{-1}(Sx) + g \circ S^{-1}(Sz), \\
\text{subject to} & \quad x - z = 0. \tag{4.15}
\end{align*}
\]
Its associated Lagrangian is given by
\[
\begin{align*}
\hat{L}_S(x, z; \lambda) & \equiv f \circ S^{-1}(Sx) + g \circ S^{-1}(Sz) + \langle \lambda, x - z \rangle \\
& = f \circ S^{-1}(Sx) + g \circ S^{-1}(Sz) + \langle (S^*)^{-1}\lambda, S(x - z) \rangle. \tag{4.16}
\end{align*}
\]
It follows that
\[
\begin{align*}
\min_{x,z} \sup_{\lambda} \hat{L}(x, z; \lambda) &= \max_{\lambda} \inf_{x,z} \langle (S^*)^{-1}\lambda, Sx \rangle + f \circ S^{-1}(Sx) - \langle (S^*)^{-1}\lambda, Sz \rangle + g \circ S^{-1}(Sz) \\
& = \max_{\lambda} - f^* \circ (S^*)^{-1}\lambda - g^* \circ (S^*)^{-1}\lambda, \tag{4.17}
\end{align*}
\]
where the last line invoked Lemma 3.3. The new dual problem is therefore given by
\[
\begin{align*}
\max_{\lambda} - f^* \circ (S^*)^{-1}\lambda - g^* \circ (S^*)^{-1}\lambda. \tag{\mathcal{D}_S}
\end{align*}
\]

### 4.4.2 Substitutions

For the modified primal and dual problems \((\mathcal{P}_S), (\mathcal{D}_S)\), one can observe new functions \(f \circ S^{-1}, f^* \circ S^*\) (also \(g \circ S^{-1}, g^* \circ S^*\)) and new variables \(Sx, Sz, (S^*)^{-1}\lambda\). That is, to derive the unconstrained proximal operator version of the DRS type of algorithms, one can simply perform variables substitutions.

For example, consider the following standard DRS formulation:
\[
\begin{align*}
y^{k+1} &= \text{Prox}_f(y^k), \\
\psi^{k+1} &= \text{Prox}_f(2y^{k+1} - \psi^k) + \psi^k - y^{k+1}. \tag{4.18}
\end{align*}
\]
Inserting the new functions and variables, one has
\[
\begin{align*}
S\psi^{k+1} &= \text{Prox}_{g \circ S^{-1}}(\psi^k), \\
\psi^{k+1} &= \text{Prox}_{f \circ S^{-1}}(2S\psi^{k+1} - \psi^k) + \psi^k - S\psi^{k+1}. \tag{4.19}
\end{align*}
\]
Applying Proposition 3.3, one can see that the above has the same update rule as Algorithm 2. Other formulations in previous sections can be obtained in an identical manner.

### 4.5 Optimal parameter

In this part, we find the optimal choice of parameter by minimizing the worst-case convergence rate bound (see the bound in Proposition 2.2). We emphasize that no strong assumption is required. Prior to minimizing the bound, we need to first show that the bound is optimizable with respect to the parameter, i.e., the bound only consists of the parameter and some other independent components.

#### 4.5.1 Optimizable rate bound

For the DRS type of algorithms without any strong assumption, one has the following linear rate:
\[
\|\psi^{k+1} - \psi^k\|^2 \leq \frac{1}{k+1} \|\psi^* - \psi^0\|^2, \tag{4.20}
\]
where \(\psi^0\) is a pre-fixed initialization. Since \(\psi^0\) is pre-fixed, the bound \(\|\psi^* - \psi^0\|^2\) is optimizable w.r.t. \(S\) if \(\psi^*\) is.

To this end, first recall from Algorithm 4 that the relation between the fixed-point \(\psi^*\) and the primal-dual solution pair \((x^*, \lambda^*)\) is given by
\[
\psi^* = Sx^* + (S^*)^{-1}\lambda^*. \tag{4.21}
\]
Then, by the following two facts, we can claim that \(\psi^*\) is optimizable w.r.t. \(S\), and hence the bound is optimizable.
• The solution pair \((x^*, \lambda^*)\) is invariant to any feasible choice of parameter \(S\). A quick way to verify this is via the extended augmented Lagrangian as in (4.2). One would notice that \(S\) only appears in the augmented term, and such a term vanishes when the saddle point is attained.

• Different choices of \(S\) yield different fixed-points, i.e., \(\psi^*\) is not fixed w.r.t. \(S\). This follows from combining the relation (4.21) and the above fact that the optimal solution pair \((x^*, \lambda^*)\) is fixed.

4.5.2 The generic optimal choice

**Proposition 4.3** (The optimal parameter). For the DRS type of algorithms summarized in Section 4.3, the optimal choice of the proximal parameter \(S_{opt}\) is given by

\[
S_{opt} = \arg\min_S \|Sx^* - \psi^0\|^2 + \|(S^*)^{-1}\lambda^* - \psi^0\|^2, \tag{4.22}
\]

**Proof.** Since the rate bound is optimizable, the optimal parameter can be found by solving the following problem:

\[
S_{opt} = \arg\min_S \|Sx^* + (S^*)^{-1}\lambda^* - \psi^0\|^2 \\
= \arg\min_S \|Sx^*\|^2 + \|(S^*)^{-1}\lambda^*\|^2 - 2(Sx^* + (S^*)^{-1}\lambda^*, \psi^0) + \|\psi^0\|^2 \\
= \arg\min_S \|Sx^*\|^2 - 2(Sx^*, \psi^0) + \|\psi^0\|^2 + \|(S^*)^{-1}\lambda^*\|^2 - 2((S^*)^{-1}\lambda^*, \psi^0) + \|\psi^0\|^2 \\
= \arg\min_S \|Sx^* - \psi^0\|^2 + \|(S^*)^{-1}\lambda^* - \psi^0\|^2.
\]

The proof is now concluded.

**Proposition 4.4** (Acceleration gain). Suppose \(S_{opt}\) is a solution to (4.22). Then, compared to the non-parametrized case, employing \(S_{opt}\) improves the worst-case convergence rate by a factor of

\[
\xi_S = \frac{\|S_{opt}x^* + (S_{opt}^*)^{-1}\lambda^* - \psi^0\|^2}{\|x^* + \lambda^* - \psi^0\|^2}. \tag{4.23}
\]

**Proof.** The non-parametrized case corresponds to setting parameter \(S\) to the identity operator. The above result then follows instantly from the rate bound expression \(\|Sx^* + (S^*)^{-1}\lambda^* - \psi^0\|^2\). The proof is now concluded.

4.5.3 Closed-form expressions

Without specifying the operator parameter, the generic problem (4.22) does not admit a closed-form solution. On the other hand, when we specify it to the scalar and matrix cases, some nice closed-form expressions can be obtained.

First, we consider the scalar parameter case with a zero initialization.

**Proposition 4.5** (Optimal scalar choice). Consider problem (4.22) with a zero initialization. Then, the optimal scalar parameter is given by

\[
\alpha^* = \pm \sqrt{\frac{\|\lambda^*\|}{\|x^*\|}}. \tag{4.24}
\]

**Proof.** Substituting a scalar parameter \(\alpha\) to problem (4.22), one obtains

\[
\alpha^* = \arg\min_{\alpha} \|\alpha x^*\|^2 + \|\lambda^*/\alpha\|^2 = \arg\min_{\alpha} \alpha^2\|x^*\|^2 + \frac{1}{\alpha^2}\|\lambda^*\|^2.
\]

For the above objective function, its minimum is attained when \(\alpha^2\|x^*\|^2 = \frac{1}{\|\lambda^*\|^2}\), which gives (4.24). The proof is now concluded.

**Remarks 4.2.** (4.24) shows that the optimal choices can be both positive and negative. This is a direct consequence of the zero initialization.
Following from Proposition 4.4, the acceleration gain in the scalar parameter case is given by

$$\xi_\alpha = \frac{2\|x^*\|\|\lambda^*\|}{\|x^* + \lambda^*\|^2} + 2\langle x^*, \lambda^* \rangle$$  \hspace{1cm} (4.25)$$

Next, we consider the matrix parameter case with a zero initialization. We require an additional orthogonal-column condition to obtain a closed-form expression.

**Proposition 4.6** (Optimal orthogonal matrix parameter). Let $D$ be a full-rank square matrix with orthogonal columns. Let $D^T D \overset{\text{def}}{=} \text{Diag}(d)$ where $d$ is a positive vector (of which the entries record the energy of each column of $D$). Let $d_i$ denote the $i$-th entry of $d$. Its optimal choice is given by

$$d_i^* = \begin{cases} \frac{\lambda_i^*/x_i^*}{|x_i^*|} & x_i^* \neq 0, \\ +\infty & x_i^* = 0. \end{cases} \hspace{1cm} (4.26)$$

where $x_i^*$ and $\lambda_i^*$ are the $i$-th entries of the primal and dual optimal solutions, respectively.

**Proof.** For ease of demonstration, let $D \overset{\text{def}}{=} Q\text{Diag}(z)$, where $z^T z = d$ and $Q$ is an orthonormal matrix, i.e., $Q^T Q = I$. By Proposition 4.3, the optimization problem is given by

$$\argmin_D \|Dx^*\|^2 + \|(D^{-1})^T \lambda^*\|^2$$

(4.27)

It follows that

$$\|Dx^*\|^2 + \|(D^{-1})^T \lambda^*\|^2 = \|Q \text{Diag}(z) x^*\|^2 + \|(\text{Diag}(z))^{-1} Q \lambda^*\|^2$$

(4.28)

$$= \|\text{Diag}(z) x^*\|^2 + \|(\text{Diag}(z))^{-1} \lambda^*\|^2$$

(4.29)

$$= \sum_i |z_i x_i^*|^2 + \sum_i |\lambda_i^*/z_i|^2$$

(4.30)

$$= \sum_i d_i |x_i^*|^2 + \sum_i |\lambda_i^*|^2/d_i.$$  \hspace{1cm} (4.31)

Let us note that (4.31) implies that the original problem (4.27) is decomposable. The $i$-th optimal entry $d_i^*$ is therefore given by solving the problem

$$\argmin_{d_i} d_i |x_i^*|^2 + |\lambda_i^*|^2/d_i.$$  \hspace{1cm} (4.32)

Additionally, note that $x$ could be sparse, i.e., some of its entries can be zeros. In which case, the corresponding entries of $d^*$ should be chosen as $+\infty$ in theory. The proof is therefore concluded. \hfill \Box

**Remarks 4.3** (non-unique selection). It is worth noting that, without additional information, the optimal matrix parameter $D$ is not unique. The only thing fixed is the column-wise energy. Clearly, among all the potential choices, a diagonal matrix parameter is the simplest.

Following from Proposition 4.4, the acceleration gain is given by

$$\xi_D = \frac{2\sum_i |x_i^*\lambda_i^*| + 2\langle x^*, \lambda^* \rangle}{\|x^* + \lambda^*\|^2} = \frac{2\sum_i (|x_i^*\lambda_i^*| + x_i^*\lambda_i^*)}{\|x^* + \lambda^*\|^2}.$$  \hspace{1cm} (4.33)

**Remarks 4.4** (acceleration comparison). Comparing the matrix parameter case to the scalar case, one can easily verify that $\xi_D \leq \xi_\alpha$, i.e., the matrix parameter can provide extra acceleration.

4.5.4 Translating the dual variable — a solely primal view

In previous sections, we show that the optimal parameter choice can be nicely characterized by the primal-dual solution pair. Here, we show the feasibility to transfer the dual solution to a primal objective gradient. This leaves a solely primal view. This is useful when we approximate the theoretical optimal choice into a priori knowledge (one only needs to exploit the primal problem structure).
First, the primal and the dual solutions can be characterized as
\[ x^* = \text{Prox}^S_g(\psi^*), \quad \lambda^* = S^*(\psi^* - S\text{Prox}^S_g(\psi^*)), \]  
where the first relation follows from combining the update rule \( z^* = \text{Prox}^S_g(\psi^*) \) (see (4-point)) and the relation \( x^* = z^* \) (see Proposition 4.2); the second characterization directly follows from the update rule of \( \lambda \), see Algorithm 4.

Then, recall the definition
\[ \text{Prox}^S_g(\psi^*) = \arg\min_z g(z) + \frac{1}{2}\|S(z) - \psi^*\|^2. \]  
This implies
\[ S^*(\psi^* - S\text{Prox}^S_g(\psi^*)) \in \partial g(\text{Prox}^S_g(\psi^*)) \]  
\[ \iff \lambda^* \in \partial g(x^*) \]  
\[ \iff \lambda^* \equiv \nabla g(x^*), \]  
where \( \nabla g \) denotes the actual choice of the subgradient \( \partial g \).

Denote the actual choice of the subgradient \( \partial f \) to be \( \nabla f \). It follows that \( 0 = \nabla f(x^*) + \nabla g(x^*) \). Then, we have the following translation rule:
\[ \lambda^* = \nabla g(x^*) = -\nabla f(x^*). \]  

4.6 A dedicated operator parameter

To our knowledge, associating an operator parameter to the proximal operator is new. A natural question is whether this generalized notion is useful, i.e., is there any need to employ an operator parameter? In this part, we answer the question affirmatively. We will propose a useful operator parameter tailored for the 2-by-2 block-structured SDP.

4.6.1 Problem model

Consider the following semidefinite program (SDP):
\[
\begin{align*}
\text{minimize} & \quad f(X) \\
\text{subject to} & \quad X \in S^{N+K},
\end{align*}
\]  
where \( X \equiv \begin{bmatrix} X_1 & X_0 \\ X_0^T & X_2 \end{bmatrix} \), \( X_1 \in \mathbb{R}^{N \times N}, X_0 \in \mathbb{R}^{N \times K}, X_2 \in \mathbb{R}^{K \times K} \) and the objective function \( f \) is block-wise separable, i.e., \( f(X) \equiv f_1(X_1) + f_0(X_0) + f_2(X_2) \). Let \( S_+ \) denote the positive semidefinite cone and \( \delta_{S_+} \) as its indicator function.

The above can be rewritten into the standard \( f + g \) form
\[
\begin{align*}
\text{minimize} & \quad f(X) + \delta_{S_+}(X).
\end{align*}
\]  

The DRS type of algorithms can therefore apply.

4.6.2 The 2-by-2 block structure

We find that the required 2-by-2 block structure arises in various fields. Particularly, it is a natural consequence of applying the semidefinite relaxation (SDR), where a rank-one constraint \( X = xx^H \) is relaxed into
\[
X \succeq xx^H \iff \begin{bmatrix} X & x \\ x^H & 1 \end{bmatrix} \succeq 0.
\]  

The highlight is that the SDR in principle can be applied to all quadratically constrained quadratic programs (QCQPs), see [35] for an overview of the technique. This indicates that the result in this work could have broader impact.
4.6.3 Parameter design

To employ a generic proximal parameter for solving problem (4.41), a set invariant condition needs to be satisfied to obtain a closed-form expression (recall Proposition 3.7). For SDP, such a set invariant condition is specified into a definiteness invariant condition.

**Proposition 4.7 (Generic form).** Let parameter $S$ be defined as

$$S(X) \triangleq \begin{bmatrix} \alpha_1^2 1_{N \times N} & \alpha_1 \alpha_2 1_{N \times K} \\ \alpha_1 \alpha_2 1_{K \times N} & \alpha_2^2 1_{K \times K} \end{bmatrix} \odot \begin{bmatrix} X_1 & X_0 \\ X_0^T & X_2 \end{bmatrix}, \quad \alpha_1, \alpha_2 \in \mathbb{R}/\{0\}. \tag{4.43}$$

Then, the above parameter is self-dual and the following holds:

$$S(X) \in S_{++}^{N+K}, \forall X \in S_{++}^{N+K} \quad \& \quad S(X) \notin S_{++}^{N+K}, \forall X \notin S_{++}^{N+K}, \tag{4.44}$$

**Proof.** First, by the symmetric form in the definition, clearly $S$ is self-dual, i.e., $S = S^*$. Then, following the generalised Schur complement argument (see [3, A.5.5]), we have

$$\begin{bmatrix} \alpha_1^2 X_1 & \alpha_1 \alpha_2 \alpha_1 \alpha_2 X_0 \\ \alpha_1 \alpha_2 X_0^T & \alpha_2^2 X_2 \end{bmatrix} \succeq 0 \iff \alpha_1^2 X_1 \succeq 0, \quad \alpha_1 \alpha_2 (I - X_1 X_1^T) X_0 = 0, \quad \alpha_2^2 (X_2 - X_0^T X_1^T X_0) \succeq 0, \tag{4.45}$$

where $\cdot^\dagger$ denotes the pseudo-inverse. Clearly, for any $\alpha_1, \alpha_2 \in \mathbb{R}/\{0\}$, the definiteness of $S(X)$ remains the same as $X$. The proof is hence concluded. □

Introduce the following partition for the dual variable:

$$\Lambda \overset{\text{def}}{=} \begin{bmatrix} \Lambda_1 & \Lambda_0 \\ \Lambda_0^T & \Lambda_2 \end{bmatrix}. \tag{4.46}$$

Following from Proposition 4.3, the optimal choice of parameter under zero initialization is given by

$$S_{\text{opt}} = \arg \min_S \|S X^*\|^2 + \|(S^*)^{-1} \Lambda^*\|^2$$

$$= \arg \min_{\alpha_1, \alpha_2} \alpha_1^2 \|X_1^*\|^2 + \frac{1}{\alpha_1^2} \|\Lambda_1^*\|^2 + 2 \alpha_1 \alpha_2 \|X_0^*\|^2 + \frac{2}{\alpha_1 \alpha_2} \|\Lambda_0^*\|^2 + \alpha_2^2 \|X_2^*\|^2 + \frac{1}{\alpha_2^2} \|\Lambda_2^*\|^2, \tag{4.47}$$

The above involves minimizing a bivariate real positive polynomial. To our knowledge, there is no closed-form solution. In fact, it is already tricky to just solve the problem. One possible way to solve (4.47) is to first reformulate it into an SDP via the Gram matrix representation, see [18]. Then, one employs some off-the-shelf solvers such as CVX to find a solution in an iterative manner. On the other hand, if it is just for validation purpose, an easier way to find the minimum is by plotting the graph of the polynomial (since there is only two variables). We will later employ this approach to obtain the theoretical ground truth in our numerical simulations.

Although we do not have the closed-form solution pair ($\alpha_{\text{1opt}}, \alpha_{\text{2opt}}$) for problem (4.47), we do know the closed-forms for the separate cases ($\tilde{\alpha}_{\text{1opt}}, 1$) and $(1, \tilde{\alpha}_{\text{2opt}})$, where $\tilde{\alpha}$ is added for distinguishing purpose. The idea is to combine $\tilde{\alpha}_{\text{1opt}}$ and $\tilde{\alpha}_{\text{2opt}}$ in a certain way to approximate ($\alpha_{\text{1opt}}, \alpha_{\text{2opt}}$).

4.6.4 A new form

Before continue to the separate-case optimal choices, we would like to point out that the generic definition of $S$ as in (4.43) is not convenient to use. Particularly, it is not straightforward to see how it is connected to the scalar parameter case, and hence hard to justify the advantage of the proposed parameter. To this end, we will use the following alternative form throughout the rest of the paper.

Let $\beta \overset{\text{def}}{=} \alpha_2/\alpha_1$. Then, (4.43) can be rewritten into

$$S(X) \overset{\text{def}}{=} \begin{bmatrix} \beta \, 1_{N \times N} & 1_{N \times K} \\ 1_{K \times N} & \beta \, 1_{K \times K} \end{bmatrix} \odot \begin{bmatrix} X_1 & X_0 \\ X_0^T & X_2 \end{bmatrix}. \tag{4.48}$$

It is worth noticing that, by definitions, the signs of $\alpha$ and $\beta$ have to be the same, i.e., $\text{sgn}(\alpha) = \text{sgn}(\beta)$. Without loss of generality, we will choose them being positive $\alpha, \beta \in \mathbb{R}_{++}$ throughout the rest of the paper.
4.6.5 Parameter selection

Following from Proposition 4.3, the separate-case optimal choices \((\tilde{\alpha}_{opt}, 1)\) and \((1, \tilde{\beta}_{opt})\) under zero initialization admit the following closed-form expressions:

\[
\tilde{\alpha}_{opt} = \sqrt{\frac{\|\Sigma\|}{\|X^*\|}}, \quad \tilde{\beta}_{opt} = \sqrt{\frac{\|X^*_1\|^2 + \|\Lambda^*_1\|^2}{\|X^*_2\|^2 + \|\Lambda^*_2\|^2}}. \tag{4.49}
\]

Ideally, if parameter \(\alpha\) and \(\beta\) have independent effect on the convergence rate bound. Then, the joint choice would be a simple combination of the above separate-case choices as \((\tilde{\alpha}_{opt}, \tilde{\beta}_{opt})\). However, this is not the case. To see this, substituting the new definition \((4.48)\) into \((4.47)\), one has

\[
(\alpha_{opt}, \beta_{opt}) = \underset{\alpha, \beta}{\text{argmin}} \alpha^2\|X^*\|^2 + \frac{1}{\alpha^2}\|\Sigma\|^2 + \beta^2(\|X^*_1\|^2 + \|\Lambda^*_1\|^2) + \frac{1}{\beta^2}(\|X^*_2\|^2 + \|\Lambda^*_2\|^2). \tag{4.50}
\]

Combine the two separate-case choices in \((4.49)\) into \((\tilde{\alpha}_{opt}, \tilde{\beta}_{opt})\). One can see that this is not the solution to the above problem, since the corresponding objective value is clearly not at the minimum. Nevertheless, such a combined choice \((\tilde{\alpha}_{opt}, \tilde{\beta}_{opt})\) turns out to be useful. In our applications, we find that the underlying optimal choice \((\alpha_{opt}, \beta_{opt})\) appears close to this combination \((\tilde{\alpha}_{opt}, \tilde{\beta}_{opt})\). Moreover, we observe that \(\text{dist}(\alpha_{opt}, 1) \leq \text{dist}(\tilde{\alpha}_{opt}, 1)\) and \(\text{dist}(\beta_{opt}, 1) \leq \text{dist}(\tilde{\beta}_{opt}, 1)\). These two observations provide us an approximation strategy to slightly adjust \((\tilde{\alpha}_{opt}, \tilde{\beta}_{opt})\).

4.6.6 Acceleration gain

Here, we first introduce a useful result for the SDP. For concreteness, we prove it using the unconstrained proximal operator, with an operator parameter satisfying a definiteness invariant condition, see a feasible example to be introduced in Proposition 4.7.

Lemma 4.1 (Orthogonal primal-dual solution pair). Let \((X^*, \Lambda^*)\) be the primal-dual solution pair for solving problem \(f(X) + \delta_{B_N^+}(X)\) via the DRS type of algorithms. Then, it satisfies

\[
X^* \succeq 0, \quad \Lambda^* \preceq 0, \quad \langle X^*, \Lambda^* \rangle = 0. \tag{4.51}
\]

Proof. Without loss of generality, we employ the algorithm iterates from \((4\text{-point})\) and a feasible parameter \(S\) (definiteness invariant).

First, by definition the primal solution \(X^*\) has to be positive semidefinite, i.e., \(X^* \succeq 0\). This can be verified through

\[
SX^* = S\Sigma^* = S\text{Prox}^S_g(\Psi^*) = \Pi_{B_N^+}(\Psi^*), \tag{4.52}
\]

where \(g(X) = \delta_{B_N^+}(X)\) and the relation \(S\text{Prox}^S_g = \Pi_{B_N^+}\) follows from Proposition 3.7.

By the update rule of \(\Lambda\), we obtain

\[
\Lambda^* = S^*(\Psi^* - SX^*) = S^*(\Psi^* - \Pi_{B_N^+}(\Psi^*)) = S^* \circ \Pi_{B_N^+}(\Psi^*). \tag{4.53}
\]

By assumption \(S\) is definiteness invariant, we therefore obtain \(\Lambda^* \preceq 0\).

At last, we have

\[
\langle X^*, \Lambda^* \rangle = \langle S^{-1} \circ \Pi_{B_N^+}(\Psi^*), S^* \circ \Pi_{B_N^+}(\Psi^*) \rangle = \langle \Pi_{B_N^+}(\Psi^*), \Pi_{B_N^+}(\Psi^*) \rangle = 0. \tag{4.54}
\]

The proof is now concluded.

Owing to the orthogonality \(\langle X^*, \Lambda^* \rangle = 0\), the acceleration gain (see Proposition 4.4) for the two separate-case choices as in \((4.49)\) admit the following nice forms:

\[
\xi_\alpha = \frac{\|\alpha_{opt}X^* + \Sigma^* / \alpha_{opt}\|^2}{\|X^* + \Sigma^*\|^2} = \frac{2}{\alpha^2_{opt} + \alpha_{opt}^2}, \tag{4.55}
\]
and
\[
\xi_{\beta} = \frac{\beta_{\text{opt}}^2(\|X_2^\star\|^2 + \|\Lambda_1^\star\|^2) + (\|X_1^\star\|^2 + \|\Lambda_2^\star\|^2)/\beta_{\text{opt}}^2 + 2(\|X_0^\star\|^2 + \|\Lambda_0^\star\|^2)}{\|X^\star\|^2 + \|\Lambda^\star\|^2} = \frac{2 + c}{\beta_{\text{opt}}^2 + \beta_{\text{opt}}^{-2} + c},
\]
where \( c \equiv 2(\|X_0^\star\|^2 + \|\Lambda_0^\star\|^2)/\sqrt{(\|X_2^\star\|^2 + \|\Lambda_1^\star\|^2)(\|X_1^\star\|^2 + \|\Lambda_2^\star\|^2)} \).

5 Applications and Simulations

In this section, we consider two well-known applications — the Boolean quadratic program (BQP) in integer programming and the super-resolution (SR) for point sources localisation. We start with the BQP which has a simpler problem structure. Then, we investigate the SR problem, where the estimation theory is the most powerful. We will focus on two aspects: (i) Evaluating the practical performance of the optimal parameters; (ii) Deriving some a priori knowledge parameter choices based on the proposed theory and the problem structure.

5.1 Boolean quadratic program

Boolean quadratic program (BQP) is a basic problem in digital communications that is NP-hard. The basic formulation is in a QCQP form
\[
\begin{align*}
\text{minimize} \quad & \|Ax - b\|^2 \\
\text{subject to} \quad & x_i^2 = 1, \quad \forall i,
\end{align*}
\]
where \( A \in \mathbb{R}^{N \times K}, b \in \mathbb{R}^K, x \in \mathbb{R}^N. \)

Define a new variable \( X \equiv xx^\text{H}. \) The set of constraints \( x_i^2 = 1, \forall i \) can be combined into diag\((X) = 1_{N \times 1}. \)

Applying the semidefinite relaxation (SDR) as in (4.42), we obtain the following relaxed formulation:
\[
\begin{align*}
\text{minimize} \quad & \langle X, G_f \rangle \\
\text{subject to} \quad & X \in S_+^{N+1}, \\
& \text{diag}(X) = 1_{(N+1) \times 1},
\end{align*}
\]
where \( X \equiv \begin{bmatrix} X_1 & x \\ x^\text{T} & 1 \end{bmatrix}, \quad G_f \equiv \begin{bmatrix} A^\text{T}A & -A^\text{T}b \\ -b^\text{T}A & 0 \end{bmatrix} \) (we omit the constant term \( \|b\|^2 \) in (5.1) for convenience).

To better apply the theory, we rewrite (5.2) into the standard \( f + g \) form
\[
\text{minimize} \quad f(X) + \delta_{S_+^{N+1}}(X),
\]
where \( f(X) \equiv \langle X, G_f \rangle \) with \( \text{dom}(f) = \{ X \in \mathbb{R}^{(N+1) \times (N+1)} | \text{diag}(X) = 1_{(N+1) \times 1} \}. \) Below, we present the DRS algorithm for solving the problem.

**Algorithm 5** Boolean quadratic program via DRS with a dedicated operator parameter

1. Let parameter \( \mathcal{S} \) be defined as in (4.48); set \( \Psi^0 = 0. \)

2. while \( \|Y^{k+1} - X^{k+1}\| > \epsilon \) do

3. \( \mathcal{S} Y^{k+1} = \Pi_{S_+^{N+1}}(\Psi^k); \)

4. \( X^{k+1} = S^{-1}(2S Y^{k+1} - \Psi^k) - (S^* S)^{-1} G_f; \) \( \text{diag}(X^{k+1}) = 1_{(N+1) \times 1}; \)

5. \( \Psi^{k+1} = S X^{k+1} + \Psi^k - S Y^{k+1}; \)

6. end while
5.1.1 Theoretical optimal choices

First, we consider the theoretical optimal choices of parameters. Introduce the following partition:

\[ \Lambda = \begin{bmatrix} \Lambda_1 & \lambda_0 \\ \lambda_0^T & \lambda_2 \end{bmatrix}. \]  

(5.4)

By (4.49), we obtain the separate-case optimal choices for BQP as

\[ \tilde{\alpha}_{opt} = \sqrt{\frac{\| \Lambda^* \|}{\| X^* \|}}, \quad \tilde{\beta}_{opt} = \sqrt[4]{\frac{\| X^1 \|^2 + \lambda_2^2}{\| \Lambda^* \|^2 + 1}}. \]  

(5.5)

One can easily see that the above choices are not a priori knowledge. Below, we approximate them by exploiting the BQP problem structure.

5.1.2 A priori knowledge estimates

To approximate (5.5), let us first note that the constraints \( X \geq 0 \) and \( \text{diag}(X) = 1_{(N+1) \times 1} \) together imply the following energy bounds:

\[ N \leq \| X^1 \|^2 \leq N^2, \quad N + 1 \leq \| X^* \|^2 \leq (N + 1)^2, \]  

(5.6)

where the upper bound is attained when \( X^* \) is rank-1, and the lower bound is attained when \( X^* \) has \((N + 1)\) equal positive eigenvalues. Furthermore,

- When \( X^* \) is rank-1, the SDR is tight. In this case, (5.3) is equivalent to (5.1) and one has

\[ X^* = \begin{bmatrix} x^* x^*^T & x^* \\ x^*^T & 1 \end{bmatrix}, \quad x^* x^*^T = 1_{N \times N}. \]  

(5.7)

It follows that \( \| X^1 \|^2 = \| 1_{N \times N} \|^2 = N^2 \) and \( \| X^* \|^2 = \| 1_{(N+1) \times (N+1)} \|^2 = (N + 1)^2 \).

- When the SDR is not tight, \( X^* \) has more than one eigenvalue. In this case, the upper bounds in (5.6) are not attained, i.e., \( \| X^1 \|^2 < N^2 \) and \( \| X^* \|^2 < (N + 1)^2 \). Meanwhile, we do expect that the SDR is not too loose, since otherwise the problem (the relaxed formulation) is ill-posed. That said, \( \| X^1 \|^2 \) and \( \| X^* \|^2 \) should be relatively close to their upper bounds. Moreover, due to the root operation in (5.5), the approximation will be error-tolerant. Therefore, we may directly treat \( \| X^1 \|^2 = N^2 \) and \( \| X^* \|^2 = (N + 1)^2 \) for convenience.

Next, we consider the dual variable. Standard Lagrange analysis gives

\[ \Lambda^* = -G_f + \text{Diag}(\mu^*), \]  

(5.8)

where \( \mu \) is the Lagrange multiplier associated with the diagonal constraint and \( \text{Diag}(\cdot) \) is the operation that creates a diagonal matrix out of a vector input. Recall from problem (5.2) that \( G_f \) is known exactly. The only issue left is the structure of \( \mu^* \). In fact, as we will show below, \( \mu^* \) is related to \( G_f \). Depending on the size of the entries of \( G_f \), we have two different strategies for parameter selection.

- Small \( \| G_f \| \) case

Suppose \( G_f \) has small entries such that \( \| G_f \| \ll N \). Theoretically, we will show that the element-wise absolute value of \( \mu \) cannot be too large. Then, by (5.8), we will have that \( \| \Lambda^* \| \) should be comparable to \( \| G_f \| \). In which case, we may simply approximate \( \| \Lambda^* \| \) using \( \| G_f \| \). Following from (5.5), we obtain

\[ \tilde{\alpha}_{est} = \sqrt{\frac{\| G_f \|}{N + 1}}, \quad \tilde{\beta}_{est} = \sqrt[4]{\frac{N^2}{1 + \| \Lambda^* \|^2}}. \]  

(5.9)

We recommend the following joint choice:

\[ (\alpha_{est}, \beta_{est}) = (\sqrt{2} \tilde{\alpha}_{est}, \tilde{\beta}_{est} / \sqrt{2}). \]  

(5.10)
To see why $\mu^*$ cannot have too large entries in this case, first by Lemma 4.1, one has

$$-\Lambda^* = G_f - \text{Diag}(\mu^*) \succeq 0.$$  \hfill (5.11)

By Schur complement, one obtains

$$\mu_2 \leq 0, \quad A^T A - \text{Diag}(\mu_1^*) \succeq 0.$$  \hfill (5.12)

where $\mu^* \equiv [\mu_1^*, \mu_2]^T$. This gives a necessary condition

$$\mu_1^* \leq \text{diag}(A^T A).$$  \hfill (5.13)

Above says that all entries of $\mu_1^*$ are upper bounded by some small positive values.

Also, by Lemma 4.1, we have

$$\langle X^*, -G_f + \text{Diag}(\mu^*) \rangle = 0,$$  \hfill (5.14)

which gives

$$\sum_i \mu_i^* = \langle X^*, G_f \rangle,$$  \hfill (5.15)

where the constraint $\text{diag}(X^*) = 1_{(N+1) \times 1}$ is invoked. Above implies that the sum of the entries of $\mu^*$ is small in absolute value (recall the structure of $X^*$ from (5.7)). Combining (5.13) and (5.15), and that $\mu_2 \leq 0$, one can verify that the entries of $\mu$ cannot be too large in absolute value.

- **Large $\|G_f\|$ case**

Suppose $G_f$ has large entries such that $\|G_f\| \gg N$. We believe this is an ill-posed case, since numerically we observe that $\|X^*\|^2$ is small, roughly at 1/4 of its upper bound value (recall that the relaxation is tight only if the upper bound is attained).

We find that the parameter $\beta$ does not work well in this case and should be set to 1 for convenience. On the other hand, parameter $\alpha$ significantly affects the convergence rate, and therefore should be carefully chosen. Numerically, we find $\sqrt{\|A^*\|} \approx \sqrt{\|G_f\|}$. Following (5.5), we provide the following estimate:

$$\tilde{\alpha}_{\text{est}} = \sqrt{\frac{\|G_f\|}{N+1}}.$$  \hfill (5.16)

5.1.3 Numerical results

We set the data dimension to $N=40$, $K=50$. Let the entries of $A$ and $b$ be generated from the normal distribution $\mathcal{N}(0, \sigma^2 A)$ and $\mathcal{N}(0, \sigma^2 b)$, respectively. For the small $\|G_f\|$ case, we set $\sigma_A = 0.05$, $\sigma_b = 1$, and the estimated parameter choices are given by (5.9) and (5.10). For the large $\|G_f\|$ case, we set $\sigma_A = 30$, $\sigma_b = 1$. Only parameter $\alpha$ is employed in this case, with the estimated choice given by (5.16).

We solve the problem by employing Algorithm 5. For concreteness, we do not invoke the stopping criteria when measuring the iteration complexity. Instead, the algorithm stops when the MSE of the output compared to the ground truth (generated by the off-the-shelf solver CVX [27] under the best precision mode) is smaller than the threshold $\epsilon = 10^{-6}$. The figures can be reproduced using MATLAB by setting the random number generator to default, i.e., `rng('default').`

In Figure 1a, 1b, 1c, we consider the small $\|G_f\|$ case. Figure 1a shows the practical performance of Algorithm 5 in the iteration complexity sense. The pair $(1, 1)$ corresponds to the non-parametric case where more than $5 \times 10^4$ iterations are required to achieve the MSE threshold. The estimate $(\tilde{\alpha}_{\text{est}}, 1)$ almost reaches the underlying limit $(\tilde{\alpha}^*, 1)$, where roughly 450 iterations are needed. The estimate $(1, \tilde{\beta}_{\text{est}})$ needs around 320 iterations, and the underlying limit $(\tilde{\beta}^*, 1)$ requires roughly 160 iterations. Under this setting, both estimates outperform the theoretical optimums. While this is not always the case, it occurs quite often during our numerical experiments. Figure 1b plots the worst-case convergence rate bound values, where the curves are convex and smooth, and resemble the practical ones.
Figure 1: (a) Practical performance: number of iterations by employing $\alpha$ and $\beta$ separately; (b) Theoretical curves: convergence rate bound values by employing $\alpha$ and $\beta$ separately; (c) Convergence rates plot via MSE (small $\|G_f\|$ case): parameter $\alpha$ and $\beta$. (d) Convergence rates plot via MSE (large $\|G_f\|$ case): parameter $\alpha$.

Figure 1c, 1d evaluate the algorithm convergence rates under different MSE thresholds. In both figures, the solid lines correspond to the a priori knowledge estimates and the dashed lines correspond to the underlying performance limits. In Figure 1c, we observe that the joint-case choices ($\alpha_{est}, \beta_{est}$) and ($\alpha^*, \beta^*$) outperform the others. Figure 1d considers the large $\|G_f\|$ case. As aforementioned, parameter $\beta$ appears not useful in this setting and is set to 1. We observe that the estimate ($\tilde{\alpha}_{est}, 1$) and the theoretical optimum ($\tilde{\alpha}_{opt}, 1$) have similar performance, with nearly two orders of magnitudes acceleration compared to the non-parametric case (1, 1). Meanwhile, the underlying limit ($\tilde{\alpha}^*, 1$) indicates that another order of magnitude acceleration potential could be further explored.

5.2 Super-resolved point sources localisation

In this part, we consider a source localisation problem that receives significant amount of attention in recent years and continues to be an active area of research. It is referred to as the super-resolution problem \cite{5,51}. In a nutshell, one aims to recover the point sources locations (encoded by a spike train) from some partial Fourier measurements. The highlight is that the encoded source locations are considered in a continuum and the recovery is exact under some mild conditions.

First, consider the following spike train of $K$ components:

$$\xi(t) = \sum_{k=1}^{K} c_k \delta_{\tau_k}(t).$$  \hspace{1cm} (5.17)

where $\delta_\tau$ is a Dirac measure at time instant $\tau$ (which encodes the locations) and $c_k \in \mathbb{R}$ is the amplitude of the $k$-th spike. Following the literature, we consider the location parameter $\tau$ in a normalized range, i.e., $\tau \in [0, 1)$.  


Applying the Fourier transform to (5.17), we obtain

\[ x_n^* = \sum_k c_k e^{-i2\pi nk}, \quad n = 0, 1, \ldots, N - 1. \]  

(5.18)

It is assumed that only part of the measurements are available

\[ \{x_j^*\}_{j \in \Omega}, \quad \Omega \subset \{0, \ldots, N - 1\}, \]  

(5.19)

where \( \Omega \) denotes the index set. Following the literature (see e.g. [51]), the problem admits the following equivalent SDP characterization:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2N} \langle T(u), I_N \rangle + \frac{1}{2} t \\
\text{subject to} & \quad \begin{bmatrix} T(u) & x \\ x^H & t \end{bmatrix} \in S^{N+1}_+ \\
& \quad x_j = x_j^*, \quad \forall j \in \Omega,
\end{align*}
\]

(5.20)

where \( T : \mathbb{C}^N \to \mathbb{C}^{N \times N} \) denotes the Toeplitz mapping.

We rewrite the above into a compact form

\[
\begin{align*}
\text{minimize} & \quad \langle X, G_f \rangle \\
\text{subject to} & \quad X \in S^{N+1}_+ \\
& \quad \{X\}_\Omega = \{X^*\}_\Omega,
\end{align*}
\]

(5.21)

where \( X \defeq \begin{bmatrix} T(u) & x \\ x^H & t \end{bmatrix} \), \( G_f \defeq \begin{bmatrix} \frac{1}{2N} I_N & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \). Below, we list the detailed implementation steps in a DRS style.

The stopping criteria follows from Proposition 4.2 (where \( Y_{k+1} \defeq Z_k \)).

**Algorithm 6** Super-resolution via DRS with a dedicated operator parameter

1: Let parameter \( S \) be defined as in (4.48); set \( \Psi^0 = 0 \).

2: while \( \|Y^{k+1} - X^{k+1}\| > \epsilon \) do

3: \( SY^{k+1} = \Pi_{S^{N+1}}(\Psi^k) \)

4: \( \bar{\Psi}^{k+1} \defeq 2SY^{k+1} - \Psi^k; \quad \bar{\Psi}^{k+1}_1 = T(T^*T)^{-1}T^* \bar{\Psi}^{k+1}_1 \) // Update the top-left block (due to Toeplitz mapping)

5: \( X^{k+1} = S^{-1}\bar{\Psi}^{k+1} - (S^*S)^{-1}G_f; \quad \{X^{k+1}\}_\Omega = \{X^*\}_\Omega \)

6: \( \Psi^{k+1} = SX^{k+1} + \Psi^k - SY^{k+1} \)

7: end while

It is worth mentioning the behaviour of the adjoint Toeplitz operator \( T^* : \mathbb{C}^{N \times N} \to \mathbb{C}^N \). Let \( p \defeq T^*(Q) \). Its k-th element is given by \( p_k = \langle \Theta_k, Q \rangle \) where \( \Theta_k \) is an elementary symmetric Toeplitz matrix with ones on the k-th and (-k)-th diagonals and zeros elsewhere. Moreover, the operator \( T^*T \) admits a simple form \( T^*T(v) = n \odot v \) where \( n \defeq [N, 2(N - 1), 2(N - 2), \ldots, 2]^T \).

5.2.1 Theoretical optimal choices

Here, we first provide the theoretical optimal choices of parameters. Introduce the following partition:

\[
\Lambda^* \defeq \begin{bmatrix} \Lambda^*_1 & \Lambda^*_0 \\ \Lambda^*_0 & \Lambda^*_2 \end{bmatrix}.
\]

(5.22)
By (4.49), we obtain the separate-case optimal choices for super-resolution as

\[
\tilde{\alpha}_{\text{opt}} = \sqrt{\frac{\|\Lambda^*\|}{\|X^*\|}}, \quad \tilde{\beta}_{\text{opt}} = \sqrt{\frac{\|T(u^*)\|^2 + \|\Lambda^*_1\|^2}{t^2 + \|\Lambda^*_1\|^2}}.
\]  

(5.23)

One can verify that the above choices are not a priori knowledge. Below, we approximate them by exploiting the super-resolution problem structure.

5.2.2 A priori knowledge estimates

In this part, we show that by exploiting problem structures very good approximations can be made. Quite interestingly, these estimates tend to have better practical performance than the theoretical optimal choices.

• Primal variable energy (large)

First, due to the Toeplitz structure, the primal variable admits the following Vandermonde decomposition (see e.g., [51]):

\[
X^* \equiv \begin{bmatrix} T(u^*) \\ x^* \\ t^* \\ 1 \end{bmatrix} = \sum_{k=1}^K |c_k| \begin{bmatrix} a(\tau_k) \\ a(\tau_k) \end{bmatrix}^H,
\]

(5.24)

where

\[
a(\tau_k) \overset{\text{def}}{=} \begin{bmatrix} e^{i2\pi(0)\tau_k} & \ldots & e^{i2\pi(N-1)\tau_k} \end{bmatrix}^T.
\]

(5.25)

Note that a Toeplitz matrix by definition has all its diagonal elements being the same. Denote the first entry of \(u^*\) as \(u^*_1\). The diagonal elements of \(T(u^*)\) then all equal to \(u^*_1\). It follows that

\[
u^*_1 = t^* = \sum_{k=1}^K |c_k|.
\]

(5.26)

This implies

\[
\text{trace}(T(u^*)) = N \sum_{k=1}^K |c_k|, \quad \text{trace}(X^*) = (N+1) \sum_{k=1}^K |c_k|.
\]

(5.27)

Above says that the trace, or equivalently the sum of the eigenvalues, is fixed. Since \(T(u^*)\) and \(X^*\) are positive semidefinite, we therefore have that their energies are maximized when they are rank-one; their energies are minimized when all the eigenvalues are the same. Since for this problem \(\text{rank}(T(u^*)) = \text{rank}(X^*) = K\), we arrive at

\[
\frac{N}{K} \sum_{k=1}^K |c_k| \leq \|T(u^*)\| < N \sum_{k=1}^K |c_k|, \quad \frac{N+1}{K} \sum_{k=1}^K |c_k| \leq \|X^*\| < (N+1) \sum_{k=1}^K |c_k|.
\]

(5.28)

Define the following factor as the average magnitude:

\[
m_{\text{avg}} \overset{\text{def}}{=} \frac{1}{K} \sum_{k=1}^K |c_k|.
\]

(5.29)

Then, we can rewrite (5.28) into

\[
Nm_{\text{avg}} \leq \|T(u^*)\| < N \sum_{k=1}^K |c_k|, \quad (N+1)m_{\text{avg}} \leq \|X^*\| < (N+1) \sum_{k=1}^K |c_k|.
\]

(5.30)

It is worth noticing that the average magnitude \(m_{\text{avg}}\) cannot be too small in practice. Otherwise, due to the inevitable noise, it is impossible to recover the underlying spikes/sources and the problem is hence ill-posed. On the other hand, the measurement number \(N\) is often very large in practice. This says that the lower bounds from above is large. We therefore conclude that the energy of the primal variable should be considered large.
• **Dual variable energy (small)**

The following holds for the dual solution:

\[ T^*(\Lambda_1^\star) = T^*(-\frac{1}{2N}I_N) = -\frac{1}{2}e_1, \]  

(5.31)

where \( e_1 \) denotes the first standard basis with one on its first entry and zeros elsewhere, and where \( T^* \) denotes the adjoint Toeplitz operator. Recall the dual variable translation rule as in (4.39), one obtains

\[ \lambda_2^\star = -\frac{1}{2}, \]  

(5.32)

which gives \( \|\lambda_2^\star\|^2 = 1/4. \)

Recall from Lemma 4.1 that the dual variable \( \Lambda^\star \) is negative semidefinite. By Schur complement, one obtains

\[ \Lambda_1^\star \preceq 0, \quad \Lambda_1^\star + \frac{1}{2}\lambda_0^\star\lambda_0^H \preceq 0. \]  

(5.33)

The above has two implications. First, since \( \Lambda_1^\star \) is negative semidefinite, its energy \( \|\Lambda_1^\star\|^2 \) is maximized when the eigenvalues are most concentrated; it is minimized when the eigenvalues are evenly spread out. That said,

\[ \frac{1}{4N} \leq \|\Lambda_1^\star\|^2 \leq \frac{1}{4}. \]  

(5.34)

Second, since \( \lambda_0^\star\lambda_0^H \) is positive semidefinite with a single eigenvalue \( \|\lambda_0\|^2 \), the second relation in (5.33) implies that

\[ \frac{1}{4}\|\lambda_0\|^4 \leq \|\Lambda_1^\star\|^2 \leq \frac{1}{4}. \]  

(5.35)

That said, \( \|\lambda_0\|^2 \) is at most 1.

The dual variable energy can therefore be bounded as

\[ \frac{1}{4N} + \frac{1}{4} \leq \|\Lambda^\star\|^2 \equiv \|\Lambda_1^\star\|^2 + \|\lambda_2^\star\|^2 + 2\|\lambda_0\|^2 \leq \frac{5}{2}, \]  

(5.36)

which, compared to the primal variable, should be considered small.

• **The separate-case choices**

After bounding the primal and dual variables energies, we are now ready to provide the estimates. First, we consider parameter \( \alpha \). Recall that

\[ \tilde{\alpha}_{opt} = \sqrt{\frac{\|\Lambda^\star\|}{\|\mathbf{X}^\star\|}}. \]  

(5.37)

By (5.36), we have \( \sqrt{\|\Lambda^\star\|} \in (\sqrt{0.5}, \sqrt{2.5}) \), which is roughly between 0.71 and 1.26. Meanwhile, by (5.30), \( \sqrt{\|\mathbf{X}^\star\|} \) is at least \( \sqrt{(N+1)m_{avg}} \), which is quite large. Therefore, it is safe to treat \( \tilde{\alpha}_{opt} \approx \sqrt{1/\|\mathbf{X}^\star\|} \). Invoking the bounds for \( \|\mathbf{X}^\star\| \) in (5.30), we then have \( \tilde{\alpha}_{opt} \) roughly in the following range:

\[ \left[ \frac{1}{\sqrt{(N+1)\sum K c_k}}, \quad \frac{1}{\sqrt{(N+1)m_{avg}}} \right]. \]  

(5.38)

Recall the definition of \( m_{avg} \) from (5.29). We have that the above upper bound is exactly \( K \) times larger than the lower bound. Then, clearly when \( K = 1 \) the upper bound and the lower bound are the same. When \( K \) increases, numerically we observe that \( \tilde{\alpha}_{opt} \) slowly moves (from the upper bound) towards the lower bound. That said, the upper bound is key to our approximation, i.e., \( m_{avg} \) is the key. In fact, we can have some rough idea of \( m_{avg} \) via the partial observation as in (5.18). That is, if the observation vector has large energy, \( m_{avg} \) should be large and vice versa.
For a more accurate approximation of $m_{\text{avg}}$, we need additional information. For example, if we know that the spike magnitude $c$ subjects to the zero-mean normal distribution $\mathcal{N}(0, \sigma^2)$. Then, its absolute value $|c|$ follows the folded normal distribution [32] with the mean value being $\sigma\sqrt{2/\pi} \approx 0.8\sigma$. That said, a natural approximation for $m_{\text{avg}}$ is $0.8\sigma$. Suppose $K$ is moderate, say 10. In this case, $\tilde{\alpha}_{\text{opt}}$ is typically slightly smaller than the upper bound. We suggest the following choice:

$$\tilde{\alpha}_{\text{est}} = \frac{1}{\sqrt{(N+1)\sigma}}$$  (5.39)

For other $K$, one may slightly adjust the above choice.

Now, we consider parameter $\beta$. Recall that

$$\tilde{\beta}_{\text{opt}} = 4\sqrt{\frac{\|\mathcal{T}(u^*)\|^2 + \|\lambda^*_2\|^2}{t^* + \|\Lambda^*_1\|^2}}.$$  (5.40)

Employing the energy bounds (5.28), (5.34), and that $\lambda^*_2 = -\frac{1}{2}$ as in (5.32), we obtain

$$\sqrt{\frac{N^2t^* + 1 + \frac{1}{4}}{t^* + \frac{1}{4}}} \leq \tilde{\beta}_{\text{opt}} \leq 4\sqrt{\frac{N^2t^* + 1 + \frac{1}{4}}{t^* + \frac{1}{4}}},$$  (5.41)

which can be rewritten into

$$\sqrt{\frac{N^2 + \frac{1}{4}}{1 + \frac{1}{4}}} \leq \tilde{\beta}_{\text{opt}} \leq \sqrt{\frac{N^2 + \frac{1}{4}}{1 + \frac{1}{4}}},$$  (5.42)

By definition (5.26) and (5.29), we have $t^* = Km_{\text{avg}}$. As aforementioned, the average magnitude $m_{\text{avg}}$ should be not too small, we hence have $t^*$ not too small. Then, at least we can exclude the following two extreme cases: $1/(4t^2) \gg 1$, $t^* \gg N^2/K^2$. That said, the $t^*$ related terms are at most comparable (not dominant) to other terms. With the presence of the 4-th order root operation (implies error tolerance), it is safe to say that $\tilde{\beta}_{\text{opt}}$ is roughly in the following range:

$$[\sqrt{N/K}, \sqrt{N}].$$  (5.43)

Clearly, when $K = 1$, we can directly treat $\tilde{\beta}_{\text{opt}}$ to be $\sqrt{N}$. When $K$ starts to increase from 1, numerically we observe that $\tilde{\beta}_{\text{opt}}$ slowly moves from the above upper bound towards the lower bound. For example, given a moderate $K$, say 15, we find an appropriate estimation of $\tilde{\beta}_{\text{opt}}$ being $\sqrt{N/2}$.

- **The joint choice**

Numerically, comparing the joint choice $(\alpha_{\text{opt}}, \beta_{\text{opt}})$ to the separate-case ones $(\bar{\alpha}_{\text{opt}}, 1)$, $(1, \tilde{\beta}_{\text{opt}})$, we observe that $\text{dist}(\alpha_{\text{opt}}, 1) < \text{dist}(\bar{\alpha}_{\text{opt}}, 1)$, $\text{dist}(\beta_{\text{opt}}, 1) < \text{dist}(\tilde{\beta}_{\text{opt}}, 1)$. Also, we find that $\alpha_{\text{opt}}, \beta_{\text{opt}}$ are close to their corresponding separate-case choices. Therefore, we will perform a shrinkage operation to obtain the approximations.

Since $\alpha_{\text{opt}} < 1$ and $\beta_{\text{opt}} > 1$, we simply choose $\alpha_{\text{est}}$ being the upper bound of $\bar{\alpha}_{\text{opt}}$ (to reduce $\text{dist}(\bar{\alpha}_{\text{opt}}, 1)$), and similarly $\beta_{\text{est}}$ being the lower bound of $\tilde{\beta}_{\text{opt}}$. This yields,

$$(\alpha_{\text{est}}, \beta_{\text{est}}) = \left( \frac{1}{\sqrt{(N+1)m_{\text{avg}}}}, \sqrt{\frac{N}{K}} \right).$$  (5.44)

As aforementioned, if $\{c_k\}_{k=1}^K$ subjects to the normal distribution $\mathcal{N}(0, \sigma)$, then $m_{\text{avg}}$ can be approximated as $0.8\sigma$, which gives

$$(\alpha_{\text{est}}, \beta_{\text{est}}) = \left( \frac{1}{\sqrt{0.8(N+1)\sigma}}, \sqrt{\frac{N}{K}} \right).$$  (5.45)

In the simulations, we find that the above choice works very well. However, in the super-resolution problem, $K$ is typically not assumed as a priori knowledge. Therefore, one may either pre-estimate $K$ via some well-established techniques in the literature, or simply make a rough guess as we observe that the effect of $K$ is limited.
5.2.3 Numerical results

We choose the measurement number to be \( N = 50 \). We uniformly generate \( K = 10 \) source locations \( \{\tau_k\}_{k=1}^K \) from the normalized interval \([0, 1)\), with a minimum separation of \( 1/N \) guaranteed. The spike amplitudes \( \{c_k\}_{k=1}^K \) are generated from zero-mean normal distribution \( \mathcal{N}(0, \sigma^2) \), with \( \sigma = 2 \). We assume 80% randomly chosen (without replacement) entries are observed, i.e., \( |\Omega| = 0.8N \). The estimates are chosen to be \( \hat{\alpha}_{\text{est}} = 1/\sqrt{(N + 1)\sigma} \), see (5.39); \( \hat{\beta}_{\text{est}} = \sqrt{2N/3} \), see (5.43); \( (\hat{\alpha}_{\text{est}}, \hat{\beta}_{\text{est}}) = (1/\sqrt{0.8(N + 1)\sigma}, \sqrt{N/K}) \), see (5.45).

We solve the problem by employing Algorithm 6. For concreteness, we do not invoke the stopping criteria when measuring the iteration complexity. Instead, the algorithm stops when the MSE of the output compared to the ground truth (generated by the off-the-shelf solver CVX [27] under the best precision mode) is smaller than the threshold \( \epsilon = 10^{-6} \). The figures can be reproduced using MATLAB by setting the random number generator to default, i.e., rng('default').

Figure 2: (a) Practical performance: number of iterations by employing \( \alpha \) and \( \beta \) separately; (b) Theoretical curves: convergence rate bound values by employing \( \alpha \) and \( \beta \) separately; (c) The joint choice: iteration number graph plotted from employing both \( \alpha \) and \( \beta \), with the practical optimal choice and the estimated choice identified; (d) Convergence rate plot: MSE decreasing rates for several choices of parameters.

Figure 2a records the practical performance in the iteration complexity sense. The pair (1, 1) corresponds to the non-parametric case where slightly more than \( 5 \times 10^4 \) iterations are required to achieve the MSE threshold of \( 10^{-6} \). Under the current setting, the estimates and the theoretical optimal choices have similar near-optimum performance. For parameter \( \alpha \), the choices \( (\hat{\alpha}_{\text{est}}, 1) \) and \( (\hat{\alpha}_{\text{opt}}, 1) \) require roughly 550 iterations. For parameter \( \beta \), the choices \( (1, \hat{\beta}_{\text{est}}) \) and \( (1, \hat{\beta}_{\text{opt}}) \) require roughly 5000 iterations. Figure 2b plots the worst-case convergence rate bound values, where the curves are convex and smooth. While the curves are quite similar to the practical ones (as in Figure 2a), parameter \( \beta \) performs better in practice. Figure 2c evaluates the iteration number against parameter \( \alpha \) and \( \beta \). The underlying optimum \( (\alpha^*, \beta^*) \) is roughly at \((0.14, 2.79)\) with around 180 iterations; the
estimate \((\alpha_{est}, \beta_{est})\) is roughly at \((0.11, 2.24)\) with around 210 iterations.

Figure 2d plots the algorithm convergence rates against the MSE, where the solid lines correspond to the estimates as a priori knowledge and the dashed lines correspond to the underlying limits that are not known beforehand. Under the current setting, we see that all three estimates, namely \(\alpha_{est}, \beta_{est}, (\alpha_{est}, \beta_{est})\) are very close to their limits. Comparing to the non-parametric choice \((1, 1)\), the estimate \((1, \beta_{est})\) brings roughly one order of magnitude acceleration; meanwhile, \((\alpha_{est}, 1)\) brings around two orders of magnitudes acceleration. The joint estimate \((\alpha_{est}, \beta_{est})\) brings roughly two and half orders of magnitudes acceleration gain compared to the non-parametric case.

6 Conclusion

In this work, we proposed a new parametrized form of the well-known proximal operator, and we refer to it as the unconstrained proximal operator. Its basic form is given by \(\text{Prox}_{f \circ S}\), where \(S\) is a generic operator parameter and where \((f \circ S)x \equiv \{y \mid y = f(Sx)\}\). Comparing it to the conventional one \(\text{Prox}_{\lambda f}\), one can observe a nice symmetry. The highlight for the new form is that the positivity requirement for its parameter is removed. We show that this is essentially due to a quadratic relation. For example, given a positive scalar \(\lambda \in \mathbb{R}_+\), one can always find some \(\alpha \in \mathbb{R}/\{0\}\) such that \(\lambda = \alpha^2\). Moreover, several key characterizations for the new form are established. Particularly, regarding the evaluation issue, we prove the relation \(\text{Prox}_{f \circ S} = S \circ \text{Prox}_f\). We refer to the right-hand side as the splitting form, and one can verify that its evaluation difficulty is the same as the conventional proximal operator. To our knowledge, employing an operator proximal parameter is new. To justify its use, we design an operator parameter dedicated to the SDP with a 2-by-2 block structure. We believe the tailored parameter has great potential use since such a block structure appears to be very common, particularly as a consequence of applying the semidefinite relaxation (SDR) to the quadratic constrained quadratic program (QCQP).

By employing the unconstrained proximal operator, the optimal parameter issue (for the DRS type methods) is formulated into a simple unconstrained optimization problem. Unlike in the literature where one optimizes the so-called convergence factor (associated with sophisticated schemes and to our knowledge no closed-form expressions available), here we minimize the worst-case convergence rate bound. One significant difference is that we now exploit the optimal solution structure (problem model structure) rather than the objective function structure (e.g. strong convexity, smoothness). By this approach, we avoid introducing any strong assumption such as strong convexity and smoothness, also the optimal parameter choices admit closed-form expressions in some basic cases. Furthermore, we illustrate through two applications that the optimal parameter choices can be well approximated into a priori knowledge by carefully investigating the problem model structures.

To support the theory, we consider two well-known applications — the Boolean quadratic program and the super-resolution problem. Both problems are formulated as 2-by-2 block-structured SDP, and a dedicated operator parameter is employed. In both applications, we find the theoretical optimal choices close to the practical optimums, except they are not known beforehand. We further show that, by exploiting problem structures, the optimal choices (under zero initialization) can be well approximated. The estimates turn out to work very well and in some cases even reach the underlying limits (the practical optimums).

References

[1] H. H. Bauschke and P. L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, 2017.
[2] A. Beck. First-order methods in optimization. SIAM, 2017.
[3] S. Boyd, S. P. Boyd, and L. Vandenberghe. Convex optimization. Cambridge university press, 2004.
[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning, 3(1):1–122, 2011.
[5] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. Communications on pure and applied Mathematics, 67(6):906–956, 2014.
[6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[7] C. Chaux, J.-C. Pesquet, and N. Pustelnik. Nested iterative algorithms for convex constrained image recovery problems. *SIAM Journal on Imaging Sciences*, 2(2):730–762, 2009.

[8] P. L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504, 2004.

[9] P. L. Combettes. Iterative construction of the resolvent of a sum of maximal monotone operators. *J. Convex Anal*, 16(4):727–748, 2009.

[10] P. L. Combettes, D. Dung, and B. C. Vu. Dualization of signal recovery problems. *Set-Valued and Variational Analysis*, 18(3-4):373–404, 2010.

[11] P. L. Combettes and J.-C. Pesquet. A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, 2007.

[12] P. L. Combettes and N. N. Reyes. Moreau’s decomposition in banach spaces. *Mathematical Programming*, 139(1):103–114, 2013.

[13] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[14] D. Davis and W. Yin. Convergence rate analysis of several splitting schemes. In *Splitting methods in communication, imaging, science, and engineering*, pages 115–163. Springer, 2016.

[15] D. Davis and W. Yin. Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, 42(3):783–805, 2017.

[16] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[17] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

[18] B. Dumitrescu. *Positive trigonometric polynomials and signal processing applications*, volume 103. Springer, 2007.

[19] J. Eckstein and D. P. Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

[20] E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[21] M. Fortin and R. Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*. Elsevier, 2000.

[22] G. França and J. Bento. An explicit rate bound for over-relaxed ADMM. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2104–2108. IEEE, 2016.

[23] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[24] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson. Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems. *IEEE Transactions on Automatic Control*, 60(3):644–658, 2014.

[25] P. Giselsson. Tight global linear convergence rate bounds for Douglas–Rachford splitting. *Journal of Fixed Point Theory and Applications*, 19(4):2241–2270, 2017.
[26] R. Glowinski and A. Marroco. Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité d’une classe de problèmes de Dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

[27] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.

[28] B. He and X. Yuan. On the O(1/n) convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[29] B. He and X. Yuan. On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. *Numerische Mathematik*, 130(3):567–577, 2015.

[30] M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

[31] C. Lemaréchal and F. Oustry. *Semidefinite relaxations and Lagrangian duality with application to combinatorial optimization*. PhD thesis, INRIA, 1999.

[32] F. C. Leone, L. S. Nelson, and R. Nottingham. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

[33] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[34] Y. Liu, Y. Xu, and W. Yin. Acceleration of Primal–Dual Methods by Preconditioning and Simple Subproblem Procedures. *Journal of Scientific Computing*, 86(2):1–34, 2021.

[35] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3):20–34, 2010.

[36] J. Malick. The spherical constraint in boolean quadratic programs. *Journal of Global Optimization*, 39(4):609–622, 2007.

[37] R. D. Monteiro and B. F. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

[38] J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l’Académie des sciences*, 255:2897–2899, 1962.

[39] W. M. Moursi and L. Vandenberghe. Douglas–Rachford splitting for the sum of a Lipschitz continuous and a strongly monotone operator. *Journal of Optimization Theory and Applications*, 183(1):179–198, 2019.

[40] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan. A general analysis of the convergence of admm. In *International Conference on Machine Learning*, pages 343–352. PMLR, 2015.

[41] D. O’Connor and L. Vandenberghe. On the equivalence of the primal-dual hybrid gradient method and Douglas–Rachford splitting. *Mathematical Programming*, 179(1):85–108, 2020.

[42] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[43] D. W. Peaceman and H. H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for industrial and Applied Mathematics*, 3(1):28–41, 1955.

[44] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1133–1140. IEEE, 2009.

[45] S. Poljak, F. Rendl, and H. Wolkowicz. A recipe for semidefinite relaxation for (0, 1)-quadratic programming. *Journal of Global Optimization*, 7(1):51–73, 1995.

[46] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.
[47] E. K. Ryu and S. Boyd. A primer on monotone operator methods. *Appl. Comput. Math.*, 15(1):3–43, 2016.

[48] E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

[49] E. K. Ryu and W. Yin. Large-scale convex optimization via monotone operators (2020). URL https://large-scale-book.mathopt.com/LSCOMO.pdf (visited on 03/2021), 2021.

[50] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[51] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

[52] A. Teixeira, E. Ghadimi, I. Shames, H. Sandberg, and M. Johansson. Optimal scaling of the admm algorithm for distributed quadratic programming. In *52nd IEEE Conference on Decision and Control*, pages 6868–6873. IEEE, 2013.

[53] Z. Xu, M. Figueiredo, and T. Goldstein. Adaptive admm with spectral penalty parameter selection. In *Artificial Intelligence and Statistics*, pages 718–727. PMLR, 2017.

[54] Z. Xu, M. A. Figueiredo, X. Yuan, C. Studer, and T. Goldstein. Adaptive relaxed admm: Convergence theory and practical implementation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7389–7398, 2017.