SIAM J. MATH. DATA SCI.
Vol. 4, No. 4, pp. 1320–1346

Operator Shifting for General Noisy Matrix Systems*

Philip A. Etter[†] and Lexing Ying[‡]

Abstract. In the computational sciences, one must often estimate model parameters from data subject to noise and uncertainty, leading to inaccurate results. In order to improve the accuracy of models with noisy parameters, we consider the problem of reducing error in a linear system with the operator corrupted by noise. Our contribution in this paper is to extend the elliptic *operator shifting* framework from Etter and Ying, 2020 to the general nonsymmetric matrix case. Roughly, the operator shifting technique is a matrix analogue of the James–Stein estimator. The key insight is that a shift of the matrix inverse estimate in an appropriately chosen direction will reduce average error. In our extension, we interrogate a number of questions—namely, whether or not shifting towards the origin for general matrix inverses always reduces error as it does in the elliptic case. We show that this is usually the case, but that there are three key features of the general nonsingular matrices that allow for counterexamples not possible in the symmetric case. We prove that when these possibilities are eliminated by the assumption of noise symmetry and the use of the residual norm as the error metric, the optimal shift is always towards the origin, mirroring results from Etter and Ying, 2020. We also investigate behavior in the small noise regime and other scenarios. We conclude by presenting numerical experiments (with accompanying source code) inspired by reinforcement learning to demonstrate that operator shifting can yield substantial reductions in error.

Key words. operator shifting, random matrices, Monte Carlo, polynomial expansion, asymmetric matrices, noise reduction

MSC codes. 65F99, 62A99, 60B20

DOI. 10.1137/21M1416849

1. Introduction.

Numerical linear algebra is a crucial foundation for research across a massive breadth of technical domains. It forms the computational bedrock of everything from data science to computational physics. Even nonlinear problems are usually solved via linear approximation. One typically writes such systems via matrix notation,

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,, \tag{1.1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$ for $n \in \mathbb{N}$.

However, linear systems are often imperfect. Scientific problems can be subject to noise in the underlying data or model parameters, sampling error, or even epistemic uncertainty—each

[*] Received by the editors May 7, 2021; accepted for publication (in revised form) August 22, 2022; published electronically December 20, 2022.
https://doi.org/10.1137/21M1416849

Funding: The work of the second author is partially supported by the National Science Foundation under award DMS1818449 and DMS-2011699.

[†] Institute for Computational and Mathematical Engineering, Stanford University, Palo Alto, CA, USA (paetter@proton.me).

[‡] Department of Mathematics, Stanford University, Palo Alto, CA, USA (lexing@stanford.edu).

potentially giving rise to errors in predictions or inferences (see, for example, [15, 13]). So in reality, one is more often confronted by a system

$$\hat{\mathbf{A}}\hat{\mathbf{x}} = \mathbf{b}\,, \tag{1.2}$$

where one constructs $\hat{\mathbf{A}}$ from data to approximate the true $\mathbf{A}$. Hence, $\hat{\mathbf{x}} = \hat{\mathbf{A}}^{-1}\mathbf{b} \in \mathbb{R}^n$ is the solution one actually obtains when solving the observed system naively. If the uncertainty or noise is severe enough, the discrepancy between $\hat{\mathbf{x}}$ and $\mathbf{x}$ may be a real practical concern.

These situations are fairly common in the computational sciences. As an example, $\hat{\mathbf{A}}$ might be a Laplacian for a Markov chain that is not known outright, but must be sampled via trajectories through the state space. Or perhaps, $\hat{\mathbf{A}}$ might be the scattering operator through a background that is estimated from data.

Regardless of the specific application, there are a wide variety of techniques available for obtaining a better estimate of the true $\mathbf{x}$. For example, one may suppose a certain distribution for $\mathbf{x}$, as is common in such techniques as Tikhonov regularization [18]. In this paper, however, we take a fundamentally different tact. Instead of trying to apply postprocessing or Bayesian regularization to $\hat{\mathbf{x}}$, we will instead examine this problem from the standpoint of building an improved estimator for the matrix $\hat{\mathbf{A}}^{-1}$. The fundamental question we seek to investigate in this paper and the prequel [7] is whether there exist operations that can make $\hat{\mathbf{A}}^{-1}$ potentially more accurate.

1.1. Operator shifting.

As this paper is an extension of Etter and Ying 2020 [7], some discussion of the previous results of operator shifting is inevitable. We will attempt to give the high level details in this section.

The fundamental idea of operator shifting is to *shift* the estimator $\hat{\mathbf{A}}^{-1}$ by an appropriately chosen function $\hat{\mathbf{K}}(\hat{\mathbf{A}})$ of $\hat{\mathbf{A}}$,

$$\tilde{\mathbf{A}}_\beta^{-1} = \hat{\mathbf{A}}^{-1} - \beta\,\hat{\mathbf{K}}\,. \tag{1.3}$$

In continuity with previous work, we refer to $\hat{\mathbf{K}}$ as the *shift matrix* and the scalar quantity $\beta \in \mathbb{R}$ as the *shift factor*. After choosing the shift matrix, one optimizes β such that the error

$$\mathbb{E}\|\tilde{\mathbf{A}}_\beta^{-1} - \mathbf{A}^{-1}\|^2 \tag{1.4}$$

is minimized with respect to some matrix norm $\|\cdot\|$. The reader will note that performing this optimization is impossible outright, as it requires knowledge of the quantity $\mathbf{A}^{-1}$ we are trying to estimate. This issue is not fatal, however, and can be effectively addressed via a bootstrap procedure that we will discuss later.

With regards to the choice of shift matrix $\hat{\mathbf{K}}$, the simplest choice is simply $\hat{\mathbf{K}}(\hat{\mathbf{A}}) = \hat{\mathbf{A}}^{-1}$. For $\beta \in (0, 1)$, this choice corresponds to shrinking the operator towards zero. Indeed, the original intent behind operator shifting was to produce an analogue of the high-dimensional James–Stein estimator [9] for matrices. The reasoning involved is that since the underlying space is high-dimensional, the error $\hat{\mathbf{A}} - \mathbf{A}$ will likely be close to orthogonal to $\mathbf{A}$ in the Frobenius inner product. Thus, shrinking $\hat{\mathbf{A}}$ towards the origin logically brings one closer to $\mathbf{A}$ in expectation.

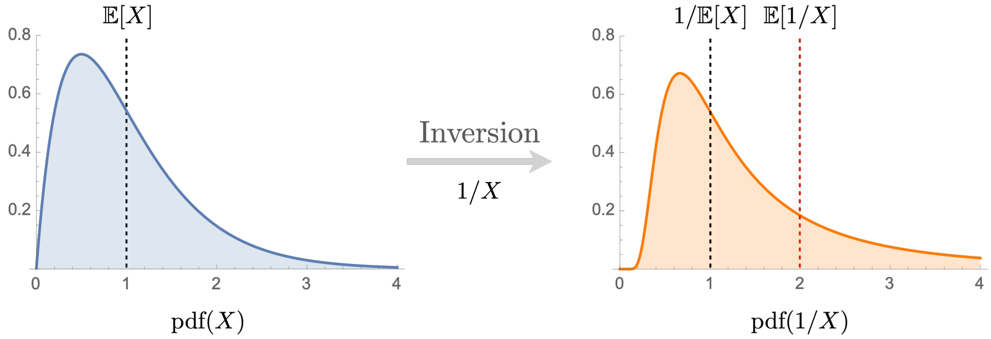

Figure 1. *An example of the upward bias induced by inversion. If we take a single sample of the scalar random variable $X \sim \Gamma(2, 1/2)$, and invert it, the probability distribution of $1/X$ has an expectation double that of $1/\mathbb{E}[X]$. Hence, estimating $1/\mathbb{E}[X]$ naively will likely give a significant overestimate. The same principle can apply when X is a random matrix.*

In contrast to the James–Stein setting, however, we must also contend with the presence of the matrix inversion operator $(\cdot)^{-1}$—as our goal is to estimate $\mathbf{A}^{-1}$ and not $\mathbf{A}$. In this respect, there is an extra wrinkle of complexity that one must deal with.

Fortunately, in the case of positive-definite symmetric matrices, matrix inversion has very nice structure. Most importantly, it is convex with respect to the Löwner order. This means that the naive $\hat{\mathbf{A}}^{-1}$ will always dominate $\mathbf{A}^{-1}$ if $\hat{\mathbf{A}}$ is unbaised (this is analogous to Jensen's inequality; see Figure 1). The confluence of these two factors—high dimensionality and the convexity of $(\cdot)^{-1}$—suggest that shrinking towards the origin is the most natural operation to perform on $\hat{\mathbf{A}}^{-1}$.

Indeed, the results of Etter and Ying, 2020 [7] bear this out. In particular, the results of the this paper demonstrate that for positive-definite symmetric matrices, shrinking towards the origin always reduces error.

Theorem 1.1 (informal, Etter and Ying [7]). *For all distributions on $\hat{\mathbf{A}}$ for which $\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}$ and $\mathbb{E}[\hat{\mathbf{A}}^{-2}]$ exists, when $\hat{\mathbf{K}}(\hat{\mathbf{A}}) = \hat{\mathbf{A}}^{-1}$ we have that $\beta^* \in (0, 1]$ for both the Frobenius norm and residual*[1] *norms.*

The other primary contributions of Etter and Ying 2020 include the following:

- Efficient Monte Carlo estimation of β with monotonic polynomial approximations in the residual norm (to be defined in section 3).
- Lower bounds relating the optimal reduction in error to the variance of the noise in $\hat{\mathbf{A}}$.
- Lower bounds for how far β^* is away from 0.

Of course, the shift $\hat{\mathbf{K}}(\hat{\mathbf{A}}) = \hat{\mathbf{A}}^{-1}$ is only one of a huge number of potential shifts. Still, a simple dimensional analysis suggests that $\hat{\mathbf{K}}(\hat{\mathbf{A}})$ should always be a homogeneous function of $\hat{\mathbf{A}}^{-1}$. So, it is natural to consider shifts of the form

$$\hat{\mathbf{K}}(\hat{\mathbf{A}}) = \mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{R} \tag{1.5}$$

[1] Defined as $\|\mathbf{X}\|^2 = \|\mathbf{AX}\|_F^2$.

for constant matrices $\mathbf{B}, \mathbf{R} \in \mathbb{R}^{n\times n}$. As it turns out, analogues of Theorem 1.1 and the bullet points above are provable for this larger class of shifts (as proved in [7]).

1.2. Novel contributions and paper overview.

We stress that, prior to the work done herein, all of the above applies strictly to *symmetric positive-definite matrices* only. The central question of this paper is *to what extent the theory of operator shifting for positive-definite matrices can be extended to general nonsingular matrices***.**

In particular, we will investigate the following questions:

1. Is it the case that shrinking the operator $\hat{\mathbf{A}}^{-1}$ towards zero always produces better results as in the positive-definite case?
    - *If not*, what are the salient structural differences between the positive-definitive cone and the general matrix group that allows for counterexamples?
    - What do these counterexamples look like?
    - What are the practical consequences of a potentially negative optimal shift factor?
2. How large of a class of operator shifts can we extend the theory to (i.e., does the theory generalize to shifts of the form (1.5))?
3. Does bootstrapped operator shifting on general matrices empirically reduce error?

We partially answer question (1) in section 3. Our work shows that under certain conditions, it is true that shrinking the operator always produces a reduction in error. We list all of the conditions in subsection 3.1 and give a proof in subsection 3.2. Furthermore, as we have discovered, each one of these conditions has a corresponding illustrative counterexample that both exemplifies the critical structural features of the general matrix group and demonstrates the necessity of the aforementioned conditions. We present these examples in section 4. Some of these examples depend on the presence of "large noise," hence, we dedicate section 5 to examining what happens when higher order noise terms are negligible. Then, we answer question (2) in section 6 and provide the main theorem for this paper, Theorem 6.1.

In section 7, we introduce the machinery that we will use to approximate the optimal shift factor β^* using a bootstrap optimization procedure. We then use this machinery to give practical algorithms for the approximation of β^* using Monte Carlo in section 8. We also provide a counterexample of how $\beta^* < 0$ can cause the bootstrapping procedure to fail arbitrarily badly in estimating β^* in subsection 8.2.

For the numerical experiments section of this paper we draw upon problems from reinforcement learning (RL). RL problems frequently require one to approximately solve linear systems (value function estimation for Markov decision processes[2]) and linear programs (policy optimization for Markov decision processes). However, the underlying problems can usually only be estimated from data due to both memory and sampling restrictions, making RL the perfect domain in which to apply our technique. Our numerical experiments in section 9 demonstrate that operator shifting can provide substantial error reduction on simple value function estimation problems.

2. Related work.

As discussed in the previous section, operator shifting is heavily inspired by the work of James and Stein. Stein's original paper [16] proved the relatively shocking

[2] Markov decision processes are Markov processes whose transition probabilities are determined by a controller.

conclusion that in dimensions ≥ 3, the standard estimator is actually *inadmissible* for the quadratic loss, as shrinking the estimate towards any fixed point by an appropriately chosen amount will always reduce loss in expectation. This idea was later refined by James and Stein in their paper on estimation under quadratic loss [9]. Fundamentally, we view our work as taking this idea and applying it to the novel setting of matrices corrupted by noise.

We remark that there are a number of connections with the field of statistical inverse problems. For example, one is often interested in estimating an object from incomplete or noisy measurements. One relevant example is the area of *semiblind deconvolution*, where one has measurements of a unknown function convolved with a kernel that is known with some uncertainty (as opposed to *blind deconvolution*, where one knows nothing about the kernel). This uncertainty in the underlying operator is a shared feature between our work and this body of literature; however, we should note that both our formalism and the semiblind convolution approach are quite different.

The operative approach of related papers in statistical inverse problems tends to be to introduce regularization on both the operator and the recovery target. An example would be the pioneering work of Golub and Van Loan on total least squares (TLS) [8]. Golub and Van Loan optimize over both perturbations to linear features and feature weights themselves, minimizing a residual term together with a regularization on the feature perturbations. Another example includes techniques from semiblind deconvolution, where one introduces a free estimate of the kernel with an appropriate regularization term into the inverse problem optimization [5]. Other approaches (i.e., double regularization) involve introducing a free estimate of the operator but constraining the free estimate so that it doesn't differ from the observed operator by too much [4]. In a gross oversimpliciation that we will perform for readability, we will characterize the above approaches as roughly solving a variant of an optimization problem that looks like

$$\min_{\mathbf{x},\mathbf{E}} \|(\hat{\mathbf{A}} + \mathbf{E})\mathbf{x} - \mathbf{b}\|^2 + R_1(\mathbf{x}) + R_2(\mathbf{E}) \,, \tag{2.1}$$

where here $\mathbf{E}$ denotes a correction to the linear features of $\hat{\mathbf{A}}$, and R_1 and R_2 denote appropriate regularizers on the recovery target $\mathbf{x}$ and the matrix correction $\mathbf{E}$.

But while these works are related, there are stark differences between our respective formalisms and approach. The primary difference between our setting and TLS is that we assume we are operating in the regime where $\hat{\mathbf{A}}$ is nonsingular and square, whereas TLS is typically applied in underdetermined scenarios. In semiblind deconvolution settings, the choice of regularizer R_2 and optimization process both depend heavily on the assumption that the operator $\hat{\mathbf{A}}$ comes from a kernel convolution. We make no such assumptions about the specific character of $\hat{\mathbf{A}}$ in our work, though it may certainly be the case that our technique functions better for some problems than others. In addition, other works also do not frame error reduction in terms of producing an estimator for a random matrix in the way that we do here, and hence their analyses are focused more on the optimization methods themselves and less on the underlying probabilistic effects that arise from noisy operators. As a final note, we observe that optimizing over $\mathbf{E}$ is in many practical scenarios infeasible. On most RL problems, for example, even storing the operator in memory would be prohibitively expensive—however, this fact is something that operator shifting can deal with fairly well, since it only needs to optimize over a single parameter β rather than an entire matrix $\mathbf{E}$. Moreover, the sheer number of degrees of freedom $\mathbf{E}$ added to the optimization in our setting has a danger of contributing to overfitting unless one is careful with regularization.

Other statistical inverse problems literature pertaining to noisy or uncertain operators include situations where the forward operator is too computationally expensive to use in an optimization procedure and is replaced by a learned proxy [11]. This is not directly relevant to our problem at hand, but notable nonetheless. Another situation studied in the literature is when one has a set of noisy input-output pairs of the underlying operator. One can use these input-output pairs to construct a regularizer for solving the inverse problem [2]. These works are both very different to the approach we take in this paper.

In addition to statistical inverse problems, the field of *model uncertainty* is tangentially relevant. Model uncertainty—both its quantification and representation—is an important topic in many branches of computational science, from structural dynamics [15] to weather and climate prediction [13]. However, work in model uncertainty is typically tied very closely to a specific domain. In contrast, our work here does not make domain specific assumptions.

Uncertainty quantification (UQ) is another relevant, but ultimately tangential subject. UQ is concerned with quantifying the probability distributions associated with calculations or physical processes. For example, one may be interested in the variance of a set of outputs given a distribution of noise on a set of inputs. Practitioners can quantify this through a variety of means—including, but not limited to, Monte Carlo techniques [12], stochastic Galerkin projection [20], or collocation [19]. In our situation, however, we are more interested in the reduction of error rather than quantifying its distribution.

The central problem in this paper is also not too dissimilar to the setting of matrix completion seen in [6, 10]. In matrix completion, one usually seeks to recover a low-rank ground truth matrix from observations that have been corrupted by additive noise. Regardless, the respective settings of operator shifting and matrix completion are still different. The operator shifting setting operates purely on full-rank matrices, and not those of low-rank.

We should briefly mention that the mathematical branch of random matrix theory (RMT) studies the spectral properties of random matrix ensembles [1, 17]. However, RMT results usually apply only when the entries of the random matrices are independent and in the large matrix limit. We find these assumptions to be too stringent for the problem at hand.

In conclusion, we do not believe that the setting we introduce in this paper has been studied in the proposed fashion before. There is little precedent in the literature for the operator shifting method beyond the original paper [7].

3. Theoretical guarantees.

In order to provide a theory for nonsymmetric operator shifting that mirrors the theory for symmetric operator shifting, we focus on the inverse operator error

$$\mathcal{E}_{\mathbf{B},\mathbf{R}}(\tilde{\mathbf{A}}_\beta^{-1}) \equiv \mathbb{E}\left[\|\tilde{\mathbf{A}}_\beta^{-1} - \mathbf{A}^{-1}\|^2_{\mathbf{B},\mathbf{R}}\right]. \tag{3.1}$$

Here, the $\|\cdot\|_{\mathbf{B},\mathbf{R}}$ is a generalized version of a matrix inner product norm for symmetric positive-definite $\mathbf{B}$ and $\mathbf{R}$. The corresponding matrix inner product, $\langle\cdot,\cdot\rangle_{\mathbf{B},\mathbf{R}}$, we define as

$$\begin{aligned} \langle \mathbf{X}, \mathbf{Y}\rangle_{\mathbf{B},\mathbf{R}} &\equiv \operatorname{tr}(\mathbf{R}\mathbf{X}^T\mathbf{B}\mathbf{Y}) = \operatorname{tr}(\mathbf{R}^{1/2}\mathbf{X}^T\mathbf{B}\mathbf{Y}\mathbf{R}^{1/2})\,,\\ \|\mathbf{X}\|^2_{\mathbf{B},\mathbf{R}} &\equiv \langle \mathbf{X},\mathbf{X}\rangle_{\mathbf{B},\mathbf{R}} = \operatorname{tr}((\mathbf{B}^{1/2}\mathbf{X}\mathbf{R}^{1/2})^T(\mathbf{B}^{1/2}\mathbf{X}\mathbf{R}^{1/2})) = \|\mathbf{B}^{1/2}\mathbf{X}\mathbf{R}^{1/2}\|_F^2\,, \end{aligned} \tag{3.2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Note that when $\mathbf{B}$ and $\mathbf{R}$ are the identity, this simply becomes the standard Frobenius inner product. In this way, the above norm is a natural generalization of the Frobenius norm for matrix operators. Just like with the Frobenius norm, one can interpret the $\mathbf{B}, \mathbf{R}$ norm via the use of expectations. Namely, if $\mathbf{b} \sim P$ is a random vector with second moment matrix $\mathbb{E}[\mathbf{b}\mathbf{b}^T] = \mathbf{R}$, then it is the case that

$$\|\mathbf{X}\|^2_{\mathbf{B},\mathbf{R}} = \mathbb{E}_{\mathbf{b}\sim P}\|\mathbf{X}\mathbf{b}\|^2_{\mathbf{B}}\,, \tag{3.3}$$

where $\|\cdot\|^2_{\mathbf{B}}$ is the vector norm induced by the symmetric positive-definite matrix $\mathbf{B}$, i.e., $\|\mathbf{x}\|^2_{\mathbf{B}} = \mathbf{x}^T\mathbf{B}\mathbf{x}$. This means that one may interpret (3.1) as being the average squared error of the solution of the linear noisy linear system (1.2) if the right-hand side is sampled from the distribution P. In mathematical notation, we may write

$$\mathcal{E}_{\mathbf{B},\mathbf{R}}(\tilde{\mathbf{A}}_\beta^{-1}) = \mathbb{E}_{\hat{\mathbf{A}}}\mathbb{E}_{\mathbf{b}\sim P}\left[\|\tilde{\mathbf{A}}_\beta^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b}\|^2_{\mathbf{B}}\right]. \tag{3.4}$$

The goal of operator shifting is to approximate the value of β^* that minimizes the above error, namely,

$$\beta^* \equiv \operatorname*{argmin}_\beta \mathcal{E}_{\mathbf{B},\mathbf{R}}(\tilde{\mathbf{A}}_\beta^{-1})\,. \tag{3.5}$$

While exact optimization of this quantity is out of reach for the aforementioned reason that one does not explicitly know $\mathbf{A}$, one can develop intuition for how β^* should behave through the use of mathematical theory, and then use bootstrap Monte Carlo to approximate it.

In accordance with our discussion of the previous work on symmetric operator shifting from our introduction, the primary theoretical question we seek to answer is whether shifting towards the origin (i.e., $\beta^* > 0$) can be expected to always decrease error as it does in the symmetric positive-definite case.

To begin to answer this question, we perform a simple calculation,

$$\beta^* = \frac{\mathbb{E}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}}}{\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}}\,, \tag{3.6}$$

hence the sign of β^* is equivalent to the sign of $\mathbb{E}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}}$, and we would therefore like a shift method to exhibit $\mathbb{E}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}} \geq 0$. We will begin by studying the simplest choice of operator shift,

$$\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}\,. \tag{3.7}$$

The overshooting effect demonstrated in Figure 1 gives one reason to believe that this choice of shift is a reasonable one, as it shrinks the inverse operator towards zero.

3.1. Conditions.

Unlike the symmetric case, to prove a rigorous statement about the sign of β^* for (3.7) in the nonsymmetric case, one must place a number of additional conditions on the constituent components of the model. We will discuss the necessity of these conditions in more detail in section 4—in short, each of these conditions has a counterexample associated with it that causes the theory to fail when the conditions are not assumed. Throughout the following proofs and discussion we will denote the noise in the matrix $\hat{\mathbf{A}}$ with the symbol $\hat{\mathbf{Z}}$,

$$\hat{\mathbf{Z}} \equiv \hat{\mathbf{A}} - \mathbf{A}\,. \tag{3.8}$$

In order to prove nonnegativity of β^*, we introduce the following constraints:

1. *Mean-zero noise*: We assume that the noise matrix $\hat{\mathbf{Z}}$ is mean zero, i.e., $\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}$.
2. *Isotropy*: We assume that $\mathbf{R} = \mathbb{E}_P[\mathbf{b}\mathbf{b}^T] = \mathbf{I}$. This means that there is no preferred direction in which we care about the accuracy of the estimator $\tilde{\mathbf{A}}_\beta^{-1}$.
3. *Noise symmetry*: We assume that the distribution of the matrix $\hat{\mathbf{A}}$ is symmetric about its mean, namely, that $\hat{\mathbf{Z}}$ has the same distribution as $-\hat{\mathbf{Z}}$.
4. *Residual norm*: We specifically choose our norm of interest $\mathbf{B}$ to be the residual norm $\mathbf{B} = \mathbf{A}^T\mathbf{A}$. The residual norm is often used as an objective in nonsymmetric iterative methods.

We note that for the theory of elliptic operator shifting, items (2) through (4) are not necessary—and so their apparent necessity in the nonsymmetric case is a curious mathematical phenomenon of the general nonsingular matrices $GL(\mathbb{R}^n)$.

3.2. Proof.

With the conditions outlined above, we proceed to prove the positivity of the shift factor β^*.

Theorem 3.1. *Let $\hat{\mathbf{A}}$ be a random matrix, invertible almost everywhere, such that $\mathbb{E}[\hat{\mathbf{A}}^{-2}]$ exists. Under the conditions outlined in subsection 3.1, the optimal shift factor is always nonnegative.*

Proof. We must verify

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{I}} \geq 0\,. \tag{3.9}$$

Expanding,

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{I}} = \mathrm{tr}\mathbb{E}\left[\hat{\mathbf{A}}^{-T}\mathbf{A}^T\mathbf{A}(\hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1})\right]\,. \tag{3.10}$$

Since $\mathbf{A}$ is invertible and $\hat{\mathbf{A}}$ is invertible almost everywhere, let us define the matrix $\hat{\mathbf{Y}}$,

$$\hat{\mathbf{Y}} = \hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I} = \hat{\mathbf{Z}}\mathbf{A}^{-1}\,. \tag{3.11}$$

Note that this definition implies that $\mathbb{E}[\hat{\mathbf{Y}}] = \mathbf{0}$ as well as that the distribution of $\hat{\mathbf{Y}}$ is symmetric, since the distribution of $\hat{\mathbf{Z}}$ is symmetric by condition (3).

One can rearrange to obtain an expression for $\hat{\mathbf{A}}^{-1}$,

$$\hat{\mathbf{A}}^{-1} = \mathbf{A}^{-1}(\mathbf{I} + \hat{\mathbf{Y}})^{-1}\,. \tag{3.12}$$

And we substitute the above expression into (3.10),

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{I}} = \mathrm{tr}\mathbb{E}\left[(\mathbf{I} + \hat{\mathbf{Y}})^{-T}(\mathbf{I} + \hat{\mathbf{Y}})^{-1} - (\mathbf{I} + \hat{\mathbf{Y}})^{-T}\right]\,. \tag{3.13}$$

Since the distribution of $\hat{\mathbf{Y}}$ is symmetric, it suffices to verify that

$$\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}+\mathbf{Y})^{-1}-(\mathbf{I}+\mathbf{Y})^{-T}\right]+\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y})^{-T}(\mathbf{I}-\mathbf{Y})^{-1}-(\mathbf{I}-\mathbf{Y})^{-T}\right]\geq 0 \tag{3.14}$$

or, alternatively,

$$\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}+\mathbf{Y})^{-1}+(\mathbf{I}-\mathbf{Y})^{-T}(\mathbf{I}-\mathbf{Y})^{-1}\right]\geq\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}+(\mathbf{I}-\mathbf{Y})^{-T}\right] \tag{3.15}$$

for all matrices $\mathbf{Y}$ for which $\mathbf{I}+\mathbf{Y}$ and $\mathbf{I}-\mathbf{Y}$ are nonsingular.

In order to verify (3.15), we begin by considering the matrix

$$((\mathbf{I}-\mathbf{Y})^{-T}-(\mathbf{I}+\mathbf{Y})^{-1})((\mathbf{I}-\mathbf{Y})^{-1}-(\mathbf{I}+\mathbf{Y})^{-T})\succeq\mathbf{0}\,. \tag{3.16}$$

Since this matrix is positive-definite (by virtue of the fact that it has the form $\mathbf{M}^T\mathbf{M}$), it follows that the trace of the above matrix is positive,

$$\operatorname{tr}\left[((\mathbf{I}+\mathbf{Y})^{-T}-(\mathbf{I}-\mathbf{Y})^{-1})((\mathbf{I}+\mathbf{Y})^{-1}-(\mathbf{I}-\mathbf{Y})^{-T})\right]\geq 0\,. \tag{3.17}$$

The above inequality can be rearranged,

$$\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}+\mathbf{Y})^{-1}+(\mathbf{I}-\mathbf{Y})^{-1}(\mathbf{I}-\mathbf{Y})^{-T}\right]\geq\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y}^2)^{-T}+(\mathbf{I}-\mathbf{Y}^2)^{-1}\right]\,. \tag{3.18}$$

Note that, by the cyclic property of the trace, that the left-hand sides of both (3.15) and (3.18) are identical. Therefore, it suffices to prove that

$$\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y}^2)^{-T}+(\mathbf{I}-\mathbf{Y}^2)^{-1}\right]=\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}+(\mathbf{I}-\mathbf{Y})^{-T}\right], \tag{3.19}$$

from which (3.15) will follow.

To prove (3.19), we note that

$$\begin{aligned}
&\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}+(\mathbf{I}-\mathbf{Y})^{-T}\right]\\
&\quad=\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}-\mathbf{Y})^{T}(\mathbf{I}-\mathbf{Y})^{-T}+(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}+\mathbf{Y})^{T}(\mathbf{I}-\mathbf{Y})^{-T}\right]\\
&\quad=\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}-\mathbf{Y}+\mathbf{I}+\mathbf{Y})^{T}(\mathbf{I}-\mathbf{Y})^{-T}\right]\\
&\quad=2\operatorname{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}-\mathbf{Y})^{-T}\right]\\
&\quad=2\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y}^2)^{-T}\right]\\
&\quad=\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y}^2)^{-T}+(\mathbf{I}-\mathbf{Y}^2)^{-1}\right].
\end{aligned} \tag{3.20}$$

This proves the nonnegativity of the optimal shift factor. ■

We remark that the critical step that requires isotropy is the assertion that the left-hand sides of both (3.15) and (3.18) are equal, namely, that

$$\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y})^{-1}(\mathbf{I}-\mathbf{Y})^{-T}\right]=\operatorname{tr}\left[(\mathbf{I}-\mathbf{Y})^{-T}(\mathbf{I}-\mathbf{Y})^{-1}\right]. \tag{3.21}$$

This is no longer necessarily true if we replace the isotropic trace operator $\operatorname{tr}(\cdot)$ with an anisotropic operator $\operatorname{tr}(\mathbf{R}^{1/2}(\cdot)\mathbf{R}^{1/2})$ for general $\mathbf{R}$.

We also note that the theorem above only proves $\beta^* \geq 0$ and not $\beta^* > 0$. Getting to $\beta^* > 0$ requires an additional condition.

Corollary 3.2. *Under the conditions of Theorem 3.1, the optimal shift factor is positive if and only if* $\mathbb{P}((\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})^T \neq -(\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})) > 0$.

Proof. Note that the sole inequality in Theorem 3.1 is

$$\mathbb{E}\mathrm{tr}\left[((\mathbf{I} + \hat{\mathbf{Y}})^{-T} - (\mathbf{I} - \hat{\mathbf{Y}})^{-1})((\mathbf{I} + \hat{\mathbf{Y}})^{-1} - (\mathbf{I} - \hat{\mathbf{Y}})^{-T})\right] \geq 0\,. \tag{3.22}$$

It thus suffices to show the equivalence between the two events

$$\left\{\mathrm{tr}\left[((\mathbf{I} + \hat{\mathbf{Y}})^{-T} - (\mathbf{I} - \hat{\mathbf{Y}})^{-1})((\mathbf{I} + \hat{\mathbf{Y}})^{-1} - (\mathbf{I} - \hat{\mathbf{Y}})^{-T})\right] = 0\right\} \tag{3.23}$$

and

$$\left\{(\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})^T = -(\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})\right\} = \left\{\hat{\mathbf{Y}}^T = -\hat{\mathbf{Y}}\right\}. \tag{3.24}$$

Clearly, if $\hat{\mathbf{Y}}$ is antisymmetric, then this implies (3.23). Conversely, if (3.23) holds, then because the matrix inside the trace is positive semidefinite, the only way that the trace can be zero is if

$$(\mathbf{I} + \hat{\mathbf{Y}})^{-T} - (\mathbf{I} - \hat{\mathbf{Y}})^{-1} = \mathbf{0}\,. \tag{3.25}$$

Rearranging the above gives $\hat{\mathbf{Y}}^T = -\hat{\mathbf{Y}}$. Therefore (3.22) is strictly positive in expectation if and only if $\mathbb{P}((\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})^T \neq -(\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I}))) > 0$, proving the corollary. ■

It is interesting to note that juxtaposition of the above result with the results for SPD matrices from Etter and Ying, 2020 [7]. For SPD matrices, the optimal shift factor is always positive unless $\hat{\mathbf{A}} = \mathbf{A}$ almost surely. On the other hand, Corollary 3.2 tells us that for general matrices, the further $\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I}$ is on average from its negative transpose $-(\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})^T$, the larger the gap in (3.22) and, hence, the larger the value of β^* and the better operator shifting will perform. In the worst case scenario, it is possible to force $\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I}$ to be antisymmetric almost surely, in which case, operator shifting will be no better than the naive estimate.

4. The necessity of conditions.

In this section, we provide counterexamples of how violating the conditions in subsection 3.1 can lead to situations where the optimal shift factor β^* for $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}$ is negative. This provides a concrete lens of how the group $GL(\mathbb{R}^n)$ differs from the SPD cone $S(\mathbb{R}^n)_+$, as such situations are not possible in the SPD case. These examples demonstrate how one might use the structure of $GL(\mathbb{R}^n)$ to construct situations where increasing the "noise" in our estimator can actually lead to more accurate results.

4.1. The necessity of isotropy.

First, we investigate the necessity of the isotropy condition. Suppose that $\mathbf{A} = \mathbf{I}$ and $\hat{\mathbf{Z}}$ has a two-atom distribution with atoms

$$\mathbf{Z}_1 = \begin{bmatrix} 0 & k \\ 0 & -k \end{bmatrix}, \qquad \mathbf{Z}_2 = \begin{bmatrix} 0 & -k \\ 0 & k \end{bmatrix}, \tag{4.1}$$

where each atom occurs with equal probability, and $k \gg 1$. Clearly in this situation, the error distribution of $\hat{\mathbf{Z}}$ is symmetric and has mean zero. Therefore, all of the other conditions in subsection 3.1 are met. The shifted operator $\tilde{\mathbf{A}}_\beta^{-1}$ is given by $\tilde{\mathbf{A}}_\beta^{-1} = (1-\beta)\hat{\mathbf{A}}^{-1}$ and the random matrix $\hat{\mathbf{A}}$ has two outcomes with equal probability,

$$\mathbf{A}_1 = \begin{bmatrix} 1 & k \\ 0 & -k+1 \end{bmatrix}, \qquad \mathbf{A}_2 = \begin{bmatrix} 1 & -k \\ 0 & k+1 \end{bmatrix}. \tag{4.2}$$

These outcomes have inverses

$$\mathbf{A}_1^{-1} = \begin{bmatrix} 1 & k/(k+1) \\ 0 & -1/(k-1) \end{bmatrix}, \qquad \mathbf{A}_2^{-1} = \begin{bmatrix} 1 & k/(k+1) \\ 0 & 1/(k-1) \end{bmatrix}. \tag{4.3}$$

This means that in either case, we have

$$\hat{\mathbf{A}}^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} + O(1/k). \tag{4.4}$$

With this matrix ensemble, we can take the distribution P to be deterministic, such that

$$\mathbf{b} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}. \tag{4.5}$$

This immediately makes the problem with this setup evident, as

$$\hat{\mathbf{A}}^{-1}\mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + O(1/k), \qquad \mathbf{A}^{-1}\mathbf{b} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}. \tag{4.6}$$

It is therefore clear that the objective

$$\mathcal{E}_{\mathbf{A}^T\mathbf{A}}(\tilde{\mathbf{A}}_\beta) = \mathcal{E}_{\mathbf{I}}(\tilde{\mathbf{A}}_\beta) = \mathbb{E}\|(1-\beta)\hat{\mathbf{A}}^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b}\|_2^2 \tag{4.7}$$

must achieve its minimum at

$$\beta^* = -1 + O(1/k). \tag{4.8}$$

One might initially conclude that the ability of this example to undermine Theorem 3.1 has something to do with the assumption that k is very large. This is not the case. The largeness of k is assumed only for illustrative purposes. One can verify numerically (see Figure 2) for the above choice of $\hat{\mathbf{A}}$, that $\mathbb{E}[\hat{\mathbf{A}}^{-T}\hat{\mathbf{A}}^{-1} - \hat{\mathbf{A}}^{-T}]$ has a negative eigenvalue for all $k \neq \pm 1, 0$. This means that if we let $\mathbf{v}$ be the corresponding eigenvector and take $\mathbf{R} = \mathbf{v}\mathbf{v}^T$, the quantity determining the sign of β^* (see (3.9)) becomes

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{R}} = \mathbf{v}^T\mathbb{E}[\hat{\mathbf{A}}^{-T}\hat{\mathbf{A}}^{-1} - \hat{\mathbf{A}}^{-T}]\mathbf{v} < 0\,. \tag{4.9}$$

Therefore, for any $k \neq \pm 1, 0$, one can find a $\mathbf{b}$ such that β^* is negative.

In particular, one should note that without a requirement of positive-definiteness, it is possible to create a situation where $\hat{\mathbf{A}}$ is always "larger" than $\mathbf{A}$. This runs counter to the intuition behind operator shifting in the SPD case [7], where such a situation is not possible.

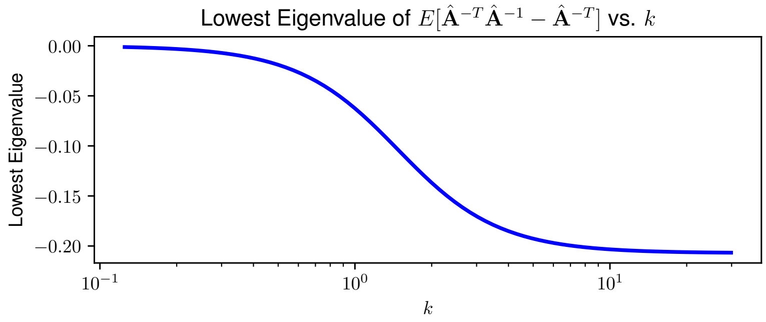

Figure 2. *The lowest eigenvalue of* $\mathbb{E}[\hat{\mathbf{A}}^{-T}\hat{\mathbf{A}}^{-1} - \hat{\mathbf{A}}^{-T}]$ *as defined in subsection* 4.1.

4.2. Outlier masking and the importance of noise symmetry.

The second condition that one needed to prove the results in section 3 is the presence of symmetry in the noise distribution. In this section, we will see how if this is not required, it is possible to construct a counterexample where $\beta^* < 0$ for the shift (3.7). For this example, we take $\mathbf{A} = \mathbf{I}$ and let $\hat{\mathbf{Z}}$ have a distribution of three equally probable atoms, given by

$$\mathbf{Z}_1 = \mathbf{I}, \qquad \mathbf{Z}_2 = k\mathbf{I}, \qquad \mathbf{Z}_3 = -(k+1)\mathbf{I}\,, \tag{4.10}$$

where $k \gg 1$. Note that this distribution is mean zero. Computing the atoms of the distribution $\hat{\mathbf{A}}^{-1}$,

$$\mathbf{A}_1^{-1} = \frac{1}{2}\mathbf{I}, \qquad \mathbf{A}_2^{-1} = \frac{1}{k+1}\mathbf{I}, \qquad \mathbf{A}_3^{-1} = -\frac{1}{k}\mathbf{I}\,. \tag{4.11}$$

Now, note that in order to minimize the quantity

$$\mathcal{E}_{\mathbf{A}^T\mathbf{A}}(\tilde{\mathbf{A}}_\beta) = \mathbb{E}\|\tilde{\mathbf{A}}_\beta^{-1} - \mathbf{I}\|_F^2 = \frac{2}{3}\left(\frac{1-\beta}{2} - 1\right)^2 + \frac{2}{3}\left(\frac{1-\beta}{k+1} - 1\right)^2 + \frac{2}{3}\left(-\frac{1-\beta}{k} - 1\right)^2, \tag{4.12}$$

one can verify by taking the derivative and setting it to zero that

$$\beta^* = -1 + O(1/k)\,. \tag{4.13}$$

Therefore, the optimal shift will grow the inverse operator instead of shrinking it.

The message of this example is that *outliers in the matrix noise can mask distribution imbalances in the region near* $\mathbf{A}$ that can cause $\mathbb{E}[\hat{\mathbf{A}}^{-1}]$ to both lie in the direction of $\mathbf{A}^{-1}$ while at the same time being dominated by $\mathbf{A}^{-1}$. Indeed, we have that $\mathbb{E}[\hat{\mathbf{A}}^{-1}] \approx \frac{1}{2}\mathbf{I} \preceq \mathbf{I} = \mathbf{A}^{-1}$ (it bears repeating that such a feat is impossible in the SPD setting where $\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1}$ [7]). The importance of noise symmetry is that it forces the distribution of $\hat{\mathbf{Z}}$ to be balanced in the region around $\mathbf{A}$, even if the distribution contains large outliers.

4.3. The importance of conditioning and a counterexample for the Frobenius norm.

Our final counterexample concerns the use of the Frobenius norm in the objective rather than the residual norm. In the SPD case, one can prove the positivity of the optimal shift factor for a large range of different objective norms [7]. However, as we will see in this section, there are ways to break norms other than the residual norm in the nonsymmetric case. We will focus here on giving an example that shows how using the Frobenius norm instead of the residual norm makes it possible to have a negative optimal shift factor for the shift (3.7).

To begin, consider the ground truth matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -\epsilon \\ \epsilon & 0 \end{bmatrix}, \tag{4.14}$$

where $0 < \epsilon \ll 1$. For the noise $\hat{\mathbf{Z}}$, we reuse the noise distribution from subsection 4.1—consider a two-atom distribution with atoms

$$\mathbf{Z}_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad \mathbf{Z}_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \tag{4.15}$$

where each atom has equal probability. Note that this distribution is symmetric and mean zero. Since ϵ is extremely small, this means that the distribution of $\hat{\mathbf{A}}$ will have two atoms whose inverses are approximately

$$\mathbf{A}_1^{-1} \approx \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \mathbf{A}_2^{-1} \approx \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}. \tag{4.16}$$

Note that the inverses of these atoms have the same Frobenius norm as the atoms themselves. In sharp contrast, the ill-conditioning of $\mathbf{A}$ means that $\mathbf{A}^{-1}$ is an order of magnitude larger than $\mathbf{A}$,

$$\mathbf{A}^{-1} = \begin{bmatrix} 0 & -\epsilon^{-1} \\ \epsilon^{-1} & \epsilon^{-2} \end{bmatrix}. \tag{4.17}$$

Immediately, we see that in order for β^* to minimize

$$\mathcal{E}_{\mathbf{I}}(\tilde{\mathbf{A}}_\beta) = \frac{1}{2}\|(1-\beta)\mathbf{A}_1^{-1} - \mathbf{A}^{-1}\|_F^2 + \frac{1}{2}\|(1-\beta)\mathbf{A}_2^{-1} - \mathbf{A}^{-1}\|_F^2\,, \tag{4.18}$$

one must therefore have $\beta^* \sim -\epsilon^{-2}$, for which $\mathcal{E}_{\mathbf{I}}(\tilde{\mathbf{A}}_{\beta^*}) \sim \epsilon^{-1}$. For other growth orders of β, one has $\mathcal{E}_{\mathbf{I}}(\tilde{\mathbf{A}}_\beta) \gg \epsilon^{-1}$ as ϵ grows small. It is therefore clear that for ϵ small enough, β^* will be negative.

In contrast to the L_2 norm, note that the residual norm matrix for this problem is given by

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1+\epsilon^2 & \epsilon \\ \epsilon & 0 \end{bmatrix}. \tag{4.19}$$

Therefore, we see that the reason why Theorem 3.1 holds in the residual norm but not in the L_2 norm, is because the residual norm places significantly less weight on the part of the matrix $\mathbf{A}^{-1}$ that contributes to the $\mathbf{A}^{-1}$'s large increase of magnitude over $\mathbf{A}$. This means that the residual norm $\mathbf{A}^T\mathbf{A}$ accounts for ill-conditioning on $\mathbf{A}$ in a way that the L_2 norm does not.

This example therefore also demonstrates the importance of conditioning in the ground truth matrix $\mathbf{A}$. It is perfectly possible that $\mathbf{A}$ lies close to a singular matrix, while the outcomes of $\hat{\mathbf{A}}$ are moved away from singularity by the noise imparted by $\hat{\mathbf{Z}}$. If this is the case, $\mathbb{E}[\hat{\mathbf{A}}^{-1}]$ will be small in magnitude compared to $\mathbf{A}^{-1}$ and shrinking the operator $\hat{\mathbf{A}}^{-1}$ further will not reduce the average error in the Frobenius sense.

Note that SPD setting [7] avoids this issue, since if $\hat{\mathbf{A}}$ is SPD everywhere, it is impossible for $\mathbb{E}[\hat{\mathbf{A}}]$ to be close to the origin without a significant chunk of the probability distribution also lying close to the origin. This ensures that $\mathbb{E}[\hat{\mathbf{A}}^{-1}]$ will always spectrally dominate $\hat{\mathbf{A}}^{-1}$, and shifting the operator $\hat{\mathbf{A}}^{-1}$ towards $\mathbf{0}$ will reduce error.

5. The small noise regime allows for nonsymmetric noise.

As we saw in subsection 4.2, the presence of large outliers can completely mask local imbalances in the noise distribution near $\mathbf{A}$. These local imbalances can be severe enough to invalidate Theorem 3.1. Naturally, the example presented in subsection 4.2 is quite extreme, so one might ask if the issue inherent is not necessarily the *symmetry* of the noise, but rather the *magnitude* of the noise. To answer this question, we consider the *small noise regime*, where deviations in $\hat{\mathbf{Y}} = \hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I}$ are very small relative to $\mathbf{I}$. It turns out, that if one assumes that terms of order $O(\mathbb{E}[\|\hat{\mathbf{Y}}\|_F^3])$ are negligible—what we term the small noise regime—then the symmetry assumption is unnecessary, as we will see momentarily.

For this discussion, we duplicate the setting of Theorem 3.1, except we replace the condition that the noise distribution is symmetric with the condition that noise terms of the order $O(\mathbb{E}[\|\hat{\mathbf{Y}}\|_F^3])$ are negligible. Recall that the statement necessary for Theorem 3.1 to be true was

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{I}} = \mathrm{tr}\mathbb{E}\left[(\mathbf{I}+\hat{\mathbf{Y}})^{-T}(\mathbf{I}+\hat{\mathbf{Y}})^{-1} - (\mathbf{I}+\hat{\mathbf{Y}})^{-T}\right] \geq 0\,. \tag{5.1}$$

We define the function f,

$$f(\mathbf{Y}) \equiv \mathrm{tr}\left[(\mathbf{I}+\mathbf{Y})^{-T}(\mathbf{I}+\mathbf{Y})^{-1} - (\mathbf{I}+\mathbf{Y})^{-T}\right]. \tag{5.2}$$

Taking a second order Taylor expansion of f about $\mathbf{Y} = \mathbf{0}$, we obtain

$$f(\hat{\mathbf{Y}}) = f(\mathbf{0}) + \delta f(\mathbf{0})\,\hat{\mathbf{Y}} + \frac{1}{2}\delta^2 f(\mathbf{0})\,(\hat{\mathbf{Y}},\hat{\mathbf{Y}}) + O(\|\hat{\mathbf{Y}}\|_F^3). \tag{5.3}$$

Note that $f(\mathbf{0}) = 0$ and that the first term is linear in $\hat{\mathbf{Y}}$. Hence, in expectation, both of these terms vanish and we are left with

$$\mathbb{E}[f(\hat{\mathbf{Y}})] = \frac{1}{2}\mathbb{E}[\delta^2 f(\mathbf{0})\,(\hat{\mathbf{Y}},\hat{\mathbf{Y}})] + O(\mathbb{E}[\|\hat{\mathbf{Y}}\|_F^3]). \tag{5.4}$$

A calculation of $\delta^2 f(\mathbf{0})\,(\hat{\mathbf{Y}},\hat{\mathbf{Y}})$ gives

$$\begin{aligned}\mathbb{E}[\delta^2 f(\mathbf{0})\,(\hat{\mathbf{Y}},\hat{\mathbf{Y}})] &= \mathbb{E}\mathrm{tr}\left[2\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}^T + 2\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + 2\hat{\mathbf{Y}}\hat{\mathbf{Y}} - 2\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}^T\right] \\ &= 2\,\mathbb{E}\mathrm{tr}\left[\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + \hat{\mathbf{Y}}\hat{\mathbf{Y}}\right] \\ &= \mathbb{E}\mathrm{tr}\left[\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}^T + \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + \hat{\mathbf{Y}}\hat{\mathbf{Y}}^T + \hat{\mathbf{Y}}\hat{\mathbf{Y}}\right] \\ &= \mathbb{E}\mathrm{tr}\left[(\hat{\mathbf{Y}}+\hat{\mathbf{Y}}^T)^T(\hat{\mathbf{Y}}+\hat{\mathbf{Y}}^T)\right] \geq 0\,.\end{aligned} \tag{5.5}$$

Indeed, if $\hat{\mathbf{Y}}$ is not antisymmetric almost surely, then the above is in fact a strict inequality. This mirrors the result of Corollary 3.2. Regardless, we know that (5.1) must be true to second order.

6. Main theorem for more general shifts.

As mentioned in the introduction, in the elliptic case [7], one can prove a reduction in error for a variety of different shifts. In particular, the energy norm in the elliptic shifting setting has an extensive theory regarding the approximation of β^*. This therefore begs the question, do more general shifts of the form (1.5) retain their nice properties in the nonsymmetric shifting setting—where the residual norm plays the role of the energy norm in the elliptic shifting setting? Subsection 4.1 provides a definitive answer to this question: it is not possible unless the modified second moment is isotropic and the chosen norm is the residual norm.

Nonetheless, while the wide number of choices regarding norms and shifts do not translate to the nonsymmetric shifting setting, the operator shifting framework does provide a way for handling anisotropic $\mathbf{R} \neq \mathbf{I}$, namely, one chooses the operator shift

$$\hat{\mathbf{K}} \equiv \hat{\mathbf{A}}^{-1}\mathbf{R}^{-1}\,. \tag{6.1}$$

It is immediate that the results of Theorem 3.1 hold for this choice of shift—as the $\mathbf{R}^{-1}$ will cancel the $\mathbf{R}$ in the error objective. We restate this conclusion into the main theorem of this paper.

Theorem 6.1 (main theorem). *Let $\hat{\mathbf{A}}$ be a random matrix, invertible almost everywhere, such that $\mathbb{E}[\hat{\mathbf{A}}^{-2}]$ exists. Suppose that the distribution of $\hat{\mathbf{A}}$ is symmetric about its mean, and that $\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}$. Let $\mathbf{b} \sim P$ with nonsingular second moment matrix $\mathbb{E}[\mathbf{b}\mathbf{b}^T] = \mathbf{R}$. Consider the residual error*

$$\mathcal{E}_{\mathbf{A}^T\mathbf{A}}(\tilde{\mathbf{A}}_\beta) = \mathbb{E}_{\hat{\mathbf{A}}}\mathbb{E}_{\mathbf{b}\sim P}\left[\|\tilde{\mathbf{A}}_\beta^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b}\|^2_{\mathbf{A}^T\mathbf{A}}\right] = \mathbb{E}\left[\|\tilde{\mathbf{A}}_\beta^{-1} - \mathbf{A}^{-1}\|^2_{\mathbf{A}^T\mathbf{A},\mathbf{R}}\right]. \tag{6.2}$$

Consider the operator shift

$$\tilde{\mathbf{A}}_\beta^{-1} = \hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{A}}^{-1}\mathbf{R}^{-1}\,. \tag{6.3}$$

Then the β minimizing (6.2) is always nonnegative. Furthermore, the optimal β is strictly positive if and only if $\mathbb{P}((\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})^T \neq -(\hat{\mathbf{A}}\mathbf{A}^{-1} - \mathbf{I})) > 0$.

Proof. As in Theorem 3.1, we must verify

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}\mathbf{R}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{R}} \geq 0\,. \tag{6.4}$$

But note, by the definition in (3.2) and symmetry of $\mathbf{R}$, that

$$\langle \hat{\mathbf{A}}^{-1}\mathbf{R}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{R}} = \langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{I}}. \tag{6.5}$$

The condition (6.4) is therefore equivalent to

$$\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A}^T\mathbf{A},\mathbf{I}} \geq 0\,, \tag{6.6}$$

which we proved in Theorem 3.1. Likewise, the claim about positivity was proved in Corollary 3.2. ■

7. Bootstrap formalism.

In this section, we prepare to give an algorithm for estimating β^* by providing a mathematical formalism that will enable us to write down an algorithm using bootstrap Monte Carlo. In our formalism, we assume that there exists some underlying parameter space Ω (with a sigma algebra Σ) that produces matrices $\mathbf{M} = \mathcal{M}(\omega)$ through a measurable mapping $\mathcal{M} : \Omega \longrightarrow GL(\mathbb{R}^n)$. Here $GL(\mathbb{R}^n)$ denotes the group of nonsingular matrices in $\mathbb{R}^{n\times n}$. Elements $\omega \in \Omega$ may represent any number of things, e.g., measurements of a scattering background, edge weights, vertex positions, etc., that carry sufficient information to generate their respective matrices $\mathbf{M} = \mathcal{M}(\omega)$. For example, $\omega \in \Omega$ may be a weighted graph, and $\mathcal{M}(\omega)$ may denote its Laplacian.

In this parameter space Ω, we assume that there exists some *unobserved* true system parameters $\omega^* \in \Omega$ that produce the true matrix $\mathbf{A} = \mathcal{M}(\omega^*)$. The central assumption of our bootstrap procedure is the ability to sample $\hat{\mathbf{A}}$ if the unobserved true parameters ω^* were known—much in the same way one can use sufficient statistics to bootstrap samples from a distribution.

To codify this, there must exist a parameterized family of distributions D_ω over $GL(\mathbb{R}^n)$ that describes the observed randomness in the system if ω were to be the true system parameters. In practice, the following algorithms tend to be easier to implement if one takes the view that noise acts on parameter space Ω rather than directly on the matrices themselves. For most problems, this makes intrinsic sense, i.e., there may be noise in the edge weights in a graph, noise in measured scattering background, etc. Therefore, we assume the existence of a parameterized family of distributions $\mathbb{P}_\omega$ on Ω whose pushforward under $\mathcal{M}$ gives D_ω. That is, we use $\mathbb{P}_\omega$ to talk about the noise as distributed over parameters, whereas we use D_ω to talk about noise as distributed over the corresponding matrices.

Now given the (unobserved) true parameters ω^*, we suppose that we are charged with the following restatement of the central problem of this paper: we observe a single noisy parameter sample $\hat{\omega} \sim \mathbb{P}_{\omega^*}$ and the corresponding noisy matrix $\hat{\mathbf{A}} = \mathcal{M}(\hat{\omega})$ (which has distribution D_{ω^*}), and would like to generate a better estimate of the true matrix inverse $\mathbf{A}^{-1}$, where $\mathbf{A} = \mathcal{M}(\omega^*)$. Because we do not have access to the unobserved ω^*, we cannot draw additional samples from D_{ω^*}. However, we do have access to the observed $\hat{\omega}$, so we are free to draw bootstrap samples from $D_{\hat{\omega}}$—as we will in the subsequent section.

A concrete example of this formalism is given in subsection 8.2.

8. Monte Carlo estimation.

To build an algorithm from this theory, we attempt to estimate the optimal shift factor with bootstrap Monte Carlo. The quantity of interest is

$$\beta^* = \frac{\mathbb{E}_{D_{\omega^*}} \langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle_{\mathbf{A}^T\mathbf{A},\mathbf{R}}}{\mathbb{E}_{D_{\omega^*}} \|\hat{\mathbf{K}}\|^2_{\mathbf{A}^T\mathbf{A},\mathbf{R}}} . \tag{8.1}$$

As mentioned previously, it is not possible to compute this quantity directly, since we only have a single sample $\hat{\mathbf{A}}$ and we do not have access to the ground truth $\mathbf{A}$ or the distribution D_{ω^*} of $\hat{\mathbf{A}}$. We must therefore try to approximate these quantities as best as possible with the available information. Suppose that $\hat{\omega} \in \Omega$ is the parameter instance that generates the matrix $\hat{\mathbf{A}} = \mathcal{M}(\hat{\omega})$. To bootstrap (8.1), we replace all instances of $\mathbf{A}$ with $\hat{\mathbf{A}}$ and draw random instances from $D_{\hat{\omega}}$ instead of D_{ω^*}. We get

$$\tilde{\beta} = \frac{\mathbb{E}_{D_{\hat{\omega}}} \langle \hat{\mathbf{K}}_b, \hat{\mathbf{A}}_b^{-1} - \hat{\mathbf{A}}^{-1} \rangle_{\hat{\mathbf{A}}^T\hat{\mathbf{A}},\mathbf{R}}}{\mathbb{E}_{D_{\hat{\omega}}} \|\hat{\mathbf{K}}_b\|^2_{\hat{\mathbf{A}}^T\hat{\mathbf{A}},\mathbf{R}}}, \tag{8.2}$$

where $\hat{\mathbf{K}}_b = \hat{\mathbf{K}}(\hat{\mathbf{A}}_b)$ and $\hat{\mathbf{A}}_b$ is drawn from $D_{\hat{\omega}}$. This can then be discretized using Monte Carlo (i.i.d. is independent and identically distributed):

$$\hat{\beta} = \frac{\sum_{i=1}^m \mathbf{b}_i^T \hat{\mathbf{K}}_{b,i}^T \hat{\mathbf{A}}^T \hat{\mathbf{A}} (\hat{\mathbf{A}}_{b,i}^{-1} - \hat{\mathbf{A}}^{-1}) \mathbf{b}_i}{\sum_{i=1}^m \mathbf{b}_i^T \hat{\mathbf{K}}_{b,i}^T \hat{\mathbf{A}}^T \hat{\mathbf{A}} \hat{\mathbf{K}}_{b,i} \mathbf{b}_i}, \qquad \mathbf{b}_i \sim P \text{ i.i.d.}, \quad \hat{\mathbf{A}}_{b,i} \sim D_{\hat{\omega}} \text{ i.i.d.}. \tag{8.3}$$

If one opts to use the shift provided by (6.1), the expression simplifies to

$$\hat{\beta} = 1 - \frac{\sum_{i=1}^m \mathbf{b}_i^T \hat{\mathbf{A}}_{b,i}^{-T} \hat{\mathbf{A}}^T \mathbf{b}_i}{\sum_{i=1}^m \mathbf{q}_i^T \hat{\mathbf{A}}_{b,i}^{-T} \hat{\mathbf{A}}^T \hat{\mathbf{A}} \hat{\mathbf{A}}_{b,i}^{-1} \mathbf{q}_i}, \qquad \begin{aligned} \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ i.i.d.}, \\ \mathbf{q}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{-1}) \text{ i.i.d.}, \\ \hat{\mathbf{A}}_{b,i} &\sim D_{\hat{\omega}} \text{ i.i.d.} \end{aligned} \tag{8.4}$$

Note that $\mathbf{b}_i$ and $\mathbf{q}_i$ do not necessarily have to be normal—they must just have the same second moment matrix as the normal distributions specified. For good measure, one might also threshold the above expression to guarantee $\hat{\beta} \geq 0$. Finally, one can now estimate the true solution to the system (1.1) via

$$\tilde{\mathbf{x}}_{\hat{\beta}} = \tilde{\mathbf{A}}_{\hat{\beta}}^{-1} \mathbf{b} = (\hat{\mathbf{A}}^{-1} - \hat{\beta} \hat{\mathbf{A}}^{-1} \mathbf{R}^{-1}) \mathbf{b} = \hat{\mathbf{A}}^{-1} (\mathbf{b} - \hat{\beta} \mathbf{R}^{-1} \mathbf{b}). \tag{8.5}$$

Figure 3 presents a flowchart of the process.

8.1. Approximation via Taylor expansion.

Note that every Monte Carlo sample in (8.4) requires inverting a matrix system. There are times where this may be too computationally expensive to be feasible. However, if we are in the small noise regime, one may take a Taylor expansion of $\hat{\mathbf{A}}^{-1}$ about $\mathbf{A}^{-1}$—as this means one only has to factorize an operator once for the whole estimation process. The expansion to second order is given by

$$\hat{\mathbf{A}}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \hat{\mathbf{Z}} \mathbf{A}^{-1} + 2 \mathbf{A}^{-1} \hat{\mathbf{Z}} \mathbf{A}^{-1} \hat{\mathbf{Z}} \mathbf{A}^{-1} + O(\|\hat{\mathbf{Z}}^3\|_F). \tag{8.6}$$

Inserting this expression into (8.4) yields

$$\hat{\beta} \approx 1 - \frac{\sum_{i=1}^m \mathbf{b}_i^T [\mathbf{I} + 2(\hat{\mathbf{Z}}_{b,i} \hat{\mathbf{A}}^{-1})^2] \mathbf{b}_i}{\sum_{i=1}^m \mathbf{q}_i^T [\mathbf{I} + (\hat{\mathbf{Z}}_{b,i} \hat{\mathbf{A}}^{-1})^T (\hat{\mathbf{Z}}_{b,i} \hat{\mathbf{A}}^{-1}) + 4(\hat{\mathbf{Z}}_{b,i} \hat{\mathbf{A}}^{-1})^2] \mathbf{q}_i}; \tag{8.7}$$

note that we have omitted linear terms, because terms linear in $\hat{\mathbf{Z}}$ will be zero in expectation, by virtue of the fact that $\mathbb{E}[\hat{\mathbf{Z}}] = 0$. It is possible that higher orders of truncation may produce better results; however, unlike the elliptic shifting setting [7], it is difficult to prove guarantees about the quality of these truncated expansions—as one cannot use the machinery of positive-definite polynomials available in the elliptic case.

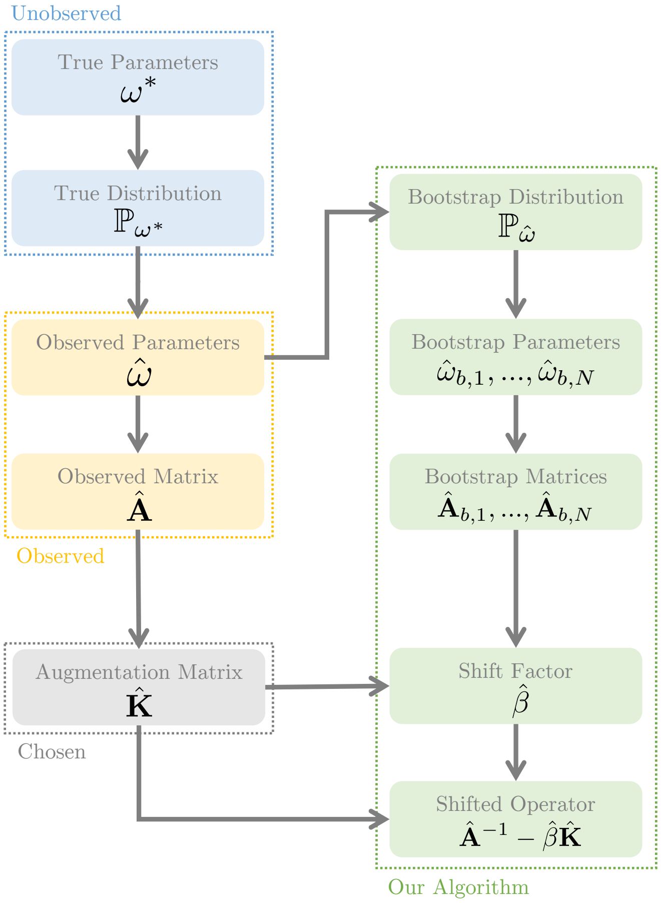
Unobserved
True Parameters ω^*
True Distribution $\mathbb{P}_{\omega^*}$
Observed Parameters $\hat{\omega}$
Observed Matrix $\hat{\mathbf{A}}$
Observed
Augmentation Matrix $\hat{\mathbf{K}}$
Chosen
Bootstrap Distribution $\mathbb{P}_{\hat{\omega}}$
Bootstrap Parameters $\hat{\omega}_{b,1},\ldots,\hat{\omega}_{b,N}$
Bootstrap Matrices $\hat{\mathbf{A}}_{b,1},\ldots,\hat{\mathbf{A}}_{b,N}$
Shift Factor $\hat{\beta}$
Shifted Operator $\hat{\mathbf{A}}^{-1}-\hat{\beta}\hat{\mathbf{K}}$
Our Algorithm

Figure 3. *A flowchart of the probabilistic setting of operator shifting as well as the algorithm itself. Operator shifting aims to find a β that gives an optimal reduction in error given a shift matrix $\hat{\mathbf{K}}$.*

8.2. How bootstrap can fail when $\beta^* < 0$. To offer a concrete example of why $\beta^* > 0$ is a desirable trait to have when performing operator shifting,we offer a counterexample where bootstrapping can be made arbitrarily bad at estimating β^* when β^* is allowed to be negative. Naturally, since there are three primary modes of failure for $\beta^* \geq 0$ (namely, the cases discussed in section 4), the example will be a bootstrapped version of subsection 4.3.

To convert the example from subsection 4.3 into a bootstrap problem we consider the sample space $\Omega = \mathbb{R}$ and let the mapping $\mathcal{M}$ be given by

$$\mathcal{M} : \omega \mapsto \begin{bmatrix} 1 & \omega \\ -\omega & 0 \end{bmatrix}. \tag{8.8}$$

For parameter ω, the noise distribution is given by

$$\mathbb{P}_\omega = \frac{1}{2}\left[\delta_{\omega-1} + \delta_{\omega+1}\right], \tag{8.9}$$

where δ_ω is the Dirac delta distribution at ω. In Figure 4 below, we see that bootstrapping the optimal shift factor in this example always returns an average shift factor that is positive,

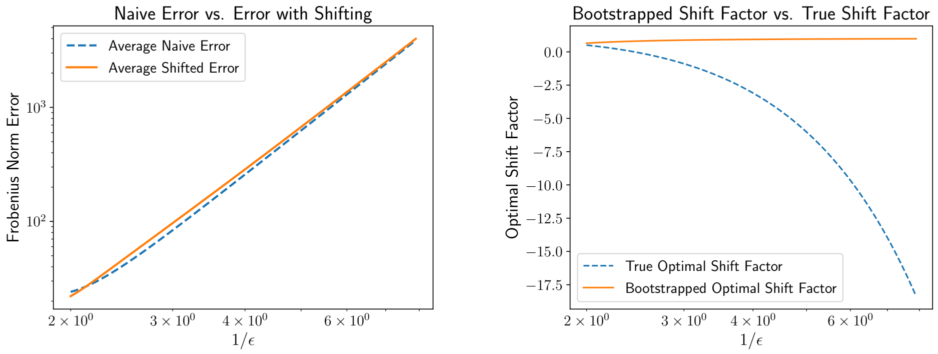

Figure 4. *This is an example of where bootstrapping can misestimate the optimal shift factor β^* when $\beta^* < 0$. In particular, we note that the bootstrapped algorithm always returns a positive shift factor on average. On the left: a plot of the error of the naive estimator $\hat{\mathbf{A}}^{-1}$ versus the error of the estimator $\tilde{\mathbf{A}}^{-1}_{\tilde{\beta}^*}$. On the right: the true optimal shift factor β^* versus the average shift factor returned by bootstrapping. We observe that the error becomes slightly worse when bootstrapped operator shifting is applied because the algorithm shrinks the matrix rather than enlarging it.*

telling the algorithm to shrink the inverse operator, whereas the true optimal shift factor is unboundedly negative as one takes $\omega^* = \epsilon \to 0$, meaning that it would have been optimal to enlarge the inverse operator instead.

Notably, the examples in section 6 and subsection 4.2 do not exhibit similar failure when bootstrapped. We have tested them both and bootstrapping does quite well in recovering the optimal shift factor of $\beta^* \approx -1$. We believe that this discrepancy in the effectiveness of bootstrapping has to do with the fact that the above example is substantially worse conditioned.

9. Numerical experiments.

To confirm our theoretical results, we will examine the asymmetric operator shift algorithm applied to the problem of more accurately computing a value function for a Markov chain whose probability transition function must be estimated from data.

9.1. Background.

For context, a discrete Markov chain X_i for $i \in \mathbb{Z}_{\geq 0}$ on a finite state set V is a stochastic process that satisfies the Markov property

$$\mathbb{P}(X_i = v_i \mid X_{i-1} = v_{i-1}, \ldots, X_1 = v_1) = \mathbb{P}(X_i = v_i \mid X_{i-1} = v_{i-1})\,. \tag{9.1}$$

A discrete Markov chain is time homogeneous if the right-hand side of (9.1) does not depend on the time i. In this situation, the Markov chain is completely characterized by its probability transition matrix

$$\mathbf{P}_{v,u} = \mathbb{P}(X_i = u \mid X_{i-1} = v)\,, \tag{9.2}$$

as well as a distribution $\mathbb{P}(X_0 = v)$ of the initial state X_0. There are countless examples of such Markov chains from the fields of probability, statistics, RL, and physics. Good references for the RL flavored problems we will be introducing shortly include [14, 3].

One is often interested in computing or approximating functionals of the process X_i. For example, one may think of X_i as an agent navigating through the set V via the means of some fixed policy, where the agent obtains a fixed reward $r(v)$ whenever it transitions from the state v. In many situations in RL, one is often interested in the average discounted reward $Q(v)$ the agent obtains over its life cycle when beginning at $X_0 = v$, namely,

$$Q(v) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r(X_i) \mid X_0 = v\right]. \tag{9.3}$$

$Q(v)$ is typically referred to as the *value function* of X_i with respect to reward function $r(v)$. It is used to gauge the quality of the agent's policy at maximizing the discounted reward it receives over its lifetime. The quantity $\gamma \in (0,1)$ is known as a *discount factor* and determines how much immediate reward is valued against future reward. One notes that the function Q is linear in r. One may use first transition analysis to express the relationship between Q and r in matrix form. Isolating the first term in the infinite sum (9.3) gives

$$Q(v) = r(v) + \gamma \sum_u \mathbb{P}(X_1 = u \mid X_0 = v) Q(u)\,. \tag{9.4}$$

Alternatively,

$$\mathbf{q} = \mathbf{r} + \gamma \mathbf{P}\mathbf{q}\,, \tag{9.5}$$

where $\mathbf{q}$ is the vector with entries $\mathbf{q}_v = Q(v)$ and $\mathbf{r}$ is the vector with entries $\mathbf{r}_v = r(v)$. This finally produces a linear system for the value function

$$\mathbf{A}\mathbf{q} \equiv (\mathbf{I} - \gamma\mathbf{P})\mathbf{q} = \mathbf{r}\,. \tag{9.6}$$

However, there are many situations where the transition matrix $\mathbf{P}$ may not be known exactly. In RL, for example, one does not have access to $\mathbf{P}$ itself, but rather a number of finite realizations of the process X_i as it traverses the state space. Therefore, instead of having access to the ground truth $\mathbf{P}$, one usually has access to a noisy version $\hat{\mathbf{P}}$ thereof. A naive solve will give

$$(\mathbf{I} - \gamma\hat{\mathbf{P}})\hat{\mathbf{q}} = \mathbf{r}\,. \tag{9.7}$$

We will see how operator shifting can be used to reduce the average error between our estimate $\hat{\mathbf{q}}$ and the ground truth $\mathbf{q}$ in the residual norm.

9.2. Noise model. One popular way to approximate a Markov chain's probability transition matrix from a sample $\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_t$ of the Markov chain is to approximate the probability $\mathbb{P}(X_i = v \mid X_{i-1} = u)$ by examining the fraction of times X_i transitions to v from u out of the total number of times X_i transitions from u, i.e.,

$$\mathbb{P}(X_i = v \mid X_{i-1} = u) = \frac{\mathbb{P}(X_i = v \text{ and } X_{i-1} = u)}{\mathbb{P}(X_{i-1} = u)} \approx \frac{\#\{i \mid \hat{X}_i = v \text{ and } \hat{X}_{i-1} = u\}}{\#\{i \mid \hat{X}_i = u \text{ and } i \geq 1\}}\,. \tag{9.8}$$

But while this estimator is commonly used—it is difficult to ensure coverage of the state space in a way that is natural, as one must take care to ensure that $\#\{i \mid \hat{X}_i = u \text{ and } i \geq 1\}$ is positive for every u.

Therefore, to model the uncertainty in a Markov chain constructed from data, we assume that, instead of observing the first t states of the Markov chain X_i, we instead observe the first N transitions of the Markov chain X_i out of every state v. We denote the first N transitions out of a state v as $\hat{Y}_{v,1}, \hat{Y}_{v,2}, \ldots, \hat{Y}_{v,N}$. If the Markov chain is irreducible, $\hat{Y}_{v,1}, \hat{Y}_{v,2}, \ldots, \hat{Y}_{v,N}$ are well-defined almost surely, and have distribution

$$\hat{Y}_{v,1}, \hat{Y}_{v,2}, \ldots, \hat{Y}_{v,N} \sim \mathbb{P}(X_i \mid X_{i-1} = v)\,, \qquad \text{i.i.d.} \tag{9.9}$$

One may then easily estimate the probability transition matrix by using the empirical distribution of $\mathbb{P}(X_i \mid X_{i-1} = v)$ to construct each row of $\hat{\mathbf{P}}$,

$$\hat{\mathbf{P}}_{v,u} \equiv \frac{\#\{\hat{Y}_{v,i} = u\}}{N} \approx \mathbb{P}(X_i = u \mid X_{i-1} = v) = \mathbf{P}_{v,u}. \tag{9.10}$$

Note that $\hat{\mathbf{P}}$ is unbiased and hence that

$$\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}\,. \tag{9.11}$$

We pose the question: is it possible to use operator shifting to increase the accuracy of the naive estimate of the value function Q? Our numerical results suggest that the answer to this question is yes.

Before delving into our numerical results, however, we should remark that this noise model is almost always not symmetric about its expectation. Nonetheless, our numerical results suggest that even though this assumption is violated, operator shifting can still substantially reduce error in practice.

9.3. Numerical results.

To test the theory, we consider three different Markov chains on one dimensional (1D), two dimensional (2D), and three dimensional (3D) grids, respectively. For our experiments, we use the operator shift $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}$,

$$x\tilde{\mathbf{A}}^{-1} = \hat{\mathbf{A}}^{-1} - \hat{\beta}\hat{\mathbf{K}} = (1-\hat{\beta})\hat{\mathbf{A}}^{-1} = (1-\hat{\beta})(\mathbf{I} - \gamma\hat{\mathbf{P}})^{-1}\,, \tag{9.12}$$

where we use the bootstrapped Taylor expanded approximation $\hat{\beta}$ from (8.7), i.e.,

$$\hat{\beta} \approx 1 - \frac{\sum_{i=1}^{m} \mathbf{b}_i^T[\mathbf{I} + 2(\hat{\mathbf{Z}}_{b,i}\hat{\mathbf{A}}^{-1})^2]\mathbf{b}_i}{\sum_{i=1}^{m} \mathbf{q}_i^T[\mathbf{I} + (\hat{\mathbf{Z}}_{b,i}\hat{\mathbf{A}}^{-1})^T(\hat{\mathbf{Z}}_{b,i}\hat{\mathbf{A}}^{-1}) + 4(\hat{\mathbf{Z}}_{b,i}\hat{\mathbf{A}}^{-1})^2]\mathbf{q}_i}\,, \qquad \begin{array}{r} \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ i.i.d.,} \\ \mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{-1}) \text{ i.i.d.,} \\ \hat{\mathbf{A}}_{b,i} \sim D_{\hat{\omega}} \text{ i.i.d.} \end{array} \tag{9.13}$$

We compute our shifted solution via

$$\tilde{\mathbf{q}} = \tilde{\mathbf{A}}^{-1}\mathbf{r}\,. \tag{9.14}$$

Note that in principle, one could use the full Monte Carlo expression (i.e., (8.4)) instead of the Taylor expanded version above. However, for practical problems, the full Monte Carlo expression in (8.4) requires a matrix inversion for every sample. This is usually prohibitively expensive for large-scale computation. We will therefore focus our numerical results solely on the Taylor expanded version, as we expect it to have more utility in practice. Furthermore, in agreement with results from [7], the descrepancy between 2nd order Taylor approximation and full approximation for the subsequent problems is quite marginal. The reader may execute the provided source code to observe this for themself.

9.3.1. Random walk with drift on a 1D grid.

For this example, we consider the Markov chain of a lazy random walk with drift on a 1D periodic grid graph. We let the state space V be the set of integers $\{1, 2, \ldots, K\}$, where $K = 16$ and take the probability transition matrix to be

$$\mathbb{P}^{(1\mathrm{D})}_{\ell,r}(X_i = u \mid X_{i-1} = v) = \begin{cases} \ell, & u = v - 1 \mod K, \\ 1 - \ell - r, & u = v \mod K, \\ r, & u = v + 1 \mod K, \\ 0, & \text{otherwise.} \end{cases} \tag{9.15}$$

We test a handful of different values for (ℓ, r) as well as a handful of different values for the sample count N that determines the noise in the matrix $\hat{\mathbf{P}}$ (higher N means less noise). For the reward function,we consider two cases: for the first, we consider a deterministic reward function given by

$$r^{(1\mathrm{D})}(v) = \sin(4\pi v/K)\,. \tag{9.16}$$

For the second, we consider a random isotropic reward vector distribution, where $\mathbf{r}^{(1\mathrm{D})} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In addition to the above collection of transition matrices, we also consider random walks with the ability to skip a vertex in the grid,

$$\mathbb{P}^{(1\mathrm{D})}_{\ell_1,\ell_2,r_1,r_2}(X_i = u \mid X_{i-1} = v) = \begin{cases} \ell_1, & u = v - 2 \mod K, \\ \ell_2, & u = v - 1 \mod K, \\ 1 - \ell_1 - \ell_2 - r_1 - r_2, & u = v \mod K, \\ r_1, & u = v + 1 \mod K, \\ r_2, & u = v + 2 \mod K, \\ 0, & \text{otherwise,} \end{cases} \tag{9.17}$$

as well as transition matrices corresponding to a random walk on a complete graph,

$$\mathbb{P}^{(1\mathrm{D})}_{\text{complete}}(X_i = u \mid X_{i-1} = v) = \frac{1}{K}\,. \tag{9.18}$$

For these 1D experiments, we use the discount factor

$$\gamma^{(1D)} = 0.99\,. \tag{9.19}$$

We present the results of our numerical experiments for the specified reward vectors (anisotropic) in Table 1 as well as for isotropic reward vectors in Table 2.

9.3.2. Random walk with drift on 2D and 3D grids.

Now let us consider the Markov chain of a lazy random walk with nonuniform drift on a 2D periodic grid graph. We let the

Table 1

Deterministic reward function comparison: A performance comparison between the accuracy of the shifted estimator $\tilde{\mathbf{x}}$ for the solution $\mathbf{x}$ versus the naive estimator $\hat{\mathbf{x}}$. The error is measured as a percentage with respect to the residual norm of the true solution. Since the error is calculated via Monte Carlo, we provide a 95% confidence interval in the $\pm 2\sigma$ column.

Chain	Samples (N)	Naive Error	$\pm 2\sigma$	Shifted Error	$\pm 2\sigma$
$\mathbb{P}^{(1\mathrm{D})}_{1/4,1/4}$	16	166%	±13.5%	53.5%	±1.35%
	32	48.8%	±1.50%	30.1%	±0.53%
	64	20.1%	±0.45%	16.0%	±0.27%
$\mathbb{P}^{(1\mathrm{D})}_{1/6,2/6}$	16	115%	±10.2%	42.2%	±1.57%
	32	31.7%	±1.11%	20.8%	±0.47%
	64	12.7%	±0.31%	10.4%	±0.20%
$\mathbb{P}^{(1\mathrm{D})}_{0,1/2}$	16	11.7%	±1.06%	8.70%	±0.67%
	32	4.02%	±0.12%	3.49%	±0.04%
	64	1.75%	±0.04%	1.63%	±0.03%
$\mathbb{P}^{(1\mathrm{D})}_{1/8,1/8,1/8,1/8}$	16	55.2%	±2.61%	29.9%	±0.59%
	32	19.5%	±0.44%	15.0%	±0.24%
	64	8.55%	±0.15%	7.55%	±0.12%
$\mathbb{P}^{(1\mathrm{D})}_{\mathrm{complete}}$	16	6.63%	±0.18%	6.09%	±0.09%
	32	3.25%	±0.05%	3.10%	±0.04%
	64	1.43%	±0.02%	1.41%	±0.02%
$\mathbb{P}^{(2\mathrm{D})}_{\mathrm{unif}}$	12	206%	±42.4%	66.9%	±3.26%
	24	91.3%	±14.3%	47.5%	±3.00%
	48	43.3%	±5.63%	30.1%	±2.29%
$\mathbb{P}^{(2\mathrm{D})}_{\mathrm{nonunif}}$	12	113%	±20.8%	52.7%	±7.91%
	24	51.7%	±7.91%	33.8%	±3.08%
	48	24.7%	±3.40%	19.7%	±2.05%
$\mathbb{P}^{(3\mathrm{D})}_{\mathrm{unif}}$	4	92.7%	±30.9%	48.3%	±7.25%
	8	38.8%	±9.09%	27.8%	±4.12%
	16	18.0%	±3.66%	15.2%	±2.36%
$\mathbb{P}^{(3\mathrm{D})}_{\mathrm{nonunif}}$	4	77.8%	±22.6%	44.3%	±7.00%
	8	33.4%	±7.86%	25.1%	±4.04%
	16	15.7%	±3.20%	13.5%	±2.29%

state space V be the set of tuples $\{1,2,\ldots,K\} \times \{1,2,\ldots,K\}$, where $K = 16$ and we consider two probability transition matrices: the first a standard uniform random walk,

$$\mathbb{P}^{(2\mathrm{D})}_{\mathrm{unif}}(X_i = u \mid X_{i-1} = v) = \begin{cases} 1/4, & |u - v| = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{9.20}$$

and the second a random walk with a nonuniform drift,

$$\mathbb{P}^{(2\mathrm{D})}_{\mathrm{nonunif}}(X_i = u \mid X_{i-1} = v) = \begin{cases} \frac{1}{4} \pm \frac{1}{8}\sin(2\pi v_x/K), & u = v \mp (1,0), \\ \frac{1}{4} \pm \frac{1}{8}\sin(2\pi v_y/K,) & u = v \mp (0,1), \\ 0, & \text{otherwise}. \end{cases} \tag{9.21}$$

Table 2

Frobenius error comparison: A performance comparison between the accuracy of the shifted operator estimator $\tilde{\mathbf{x}}$ for the solution $\mathbf{x}$ versus the naive estimator $\hat{\mathbf{x}}$. The reward function is sampled from the distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$ (hence, the error is the Frobenius operator error in expectation). The error is measured as a percentage with respect to the average residual norm of the true solutions from the prior. Since the error is calculated via Monte Carlo, we provide a 95% confidence interval in the $\pm 2\sigma$ column.

Chain	Samples (N)	Naive Error	$\pm 2\sigma$	Shifted Error	$\pm 2\sigma$
$\mathbb{P}^{(\mathrm{1D})}_{1/4,1/4}$	16	105%	±25.4%	48.4%	±6.16%
	32	30.8%	±2.95%	22.8%	±1.70%
	64	12.7%	±1.00%	11.1%	±0.77%
$\mathbb{P}^{(\mathrm{1D})}_{1/6,2/6}$	16	64.1%	±16.0%	33.9%	±5.52%
	32	17.2%	±1.77%	13.5%	±1.14%
	64	6.84%	±0.52%	6.13%	±0.41%
$\mathbb{P}^{(\mathrm{1D})}_{0,1/2}$	16	11.0%	±3.30%	8.49%	±2.40%
	32	3.77%	±0.33%	3.34%	±0.26%
	64	1.64%	±0.11%	1.54%	±0.10%
$\mathbb{P}^{(\mathrm{1D})}_{1/8,1/8,1/8,1/8}$	16	55.1%	±2.62%	29.9%	±0.59%
	32	19.5%	±0.44%	15.0%	±0.24%
	64	8.55%	±0.15%	7.55%	±0.12%
$\mathbb{P}^{(\mathrm{1D})}_{\mathrm{complete}}$	16	6.21%	±0.29%	5.81%	±0.25%
	32	2.98%	±0.13%	2.89%	±0.12%
	64	1.46%	±0.06%	1.43%	±0.06%
$\mathbb{P}^{(\mathrm{2D})}_{\mathrm{unif}}$	12	25.7%	±7.51%	20.4%	±4.89%
	24	11.3%	±2.93%	10.2%	±2.38%
	48	5.35%	±1.31%	5.07%	±1.19%
$\mathbb{P}^{(\mathrm{2D})}_{\mathrm{nonunif}}$	12	20.2%	±6.17%	16.7%	±4.89%
	24	8.96%	±2.43%	8.20%	±2.04%
	48	4.26%	±1.09%	4.07%	±1.00%
$\mathbb{P}^{(\mathrm{3D})}_{\mathrm{unif}}$	4	35.3%	±35.0%	26.2%	±20.8%
	8	14.6%	±11.8%	12.7%	±9.11%
	16	6.76%	±5.03%	6.33%	±4.43%
$\mathbb{P}^{(\mathrm{3D})}_{\mathrm{nonunif}}$	4	33.4%	±35.4%	25.1%	±21.6%
	8	13.9%	±12.0%	1.22%	±9.36%
	16	6.43%	±5.12%	6.03%	±4.53%

For the reward function, we consider both a deterministic reward

$$r^{(\mathrm{2D})}(v) = -\sin(2\pi v_x/K)\sin(2\pi v_y/K)\,, \tag{9.22}$$

as well as an isotropic reward vector distribution, where $\mathbf{r}^{(\mathrm{2D})} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. For these 2D experiments, we use the discount factor

$$\gamma^{(2D)} = 0.99\,. \tag{9.23}$$

We define similar transition matrices on 3D periodic grid graphs, where $V = \{1, \ldots, K\}^3$ and $K = 8$:

$$\mathbb{P}^{(3\mathrm{D})}_{\mathrm{unif}}(X_i = u \mid X_{i-1} = v) = \begin{cases} 1/6, & |u - v| = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{9.24}$$

as well as an analogous 3D random walk with a nonuniform drift:

$$\mathbb{P}^{(3\mathrm{D})}_{\mathrm{nonunif}}(X_i = u \mid X_{i-1} = v) = \begin{cases} \frac{1}{6} \pm \frac{1}{8}\sin(2\pi v_x/K), & u = v \mp (1,0,0), \\ \frac{1}{6} \pm \frac{1}{8}\sin(2\pi v_y/K), & u = v \mp (0,1,0), \\ \frac{1}{6} \pm \frac{1}{8}\sin(2\pi v_z/K), & u = v \mp (0,0,1), \\ 0 & \text{otherwise}. \end{cases} \tag{9.25}$$

For the reward function, we consider both a deterministic reward

$$r^{(3\mathrm{D})}(v) = -\sin(2\pi v_x/K)\sin(2\pi v_y/K)\sin(2\pi v_z/K)\,, \tag{9.26}$$

as well as an isotropic reward vector distribution, where $\mathbf{r}^{(3\mathrm{D})} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For these 3D experiments, we use the discount factor

$$\gamma^{(3D)} = 0.9\,. \tag{9.27}$$

We present the results of our numerical experiments for the specified reward vectors (anisotropic) in Table 1 as well as for isotropic reward vectors in Table 2.

9.4. Discussion. As we see in Tables 1 and 2, operator shifting can provide significant reductions in error for a variety of different Markov chain problems, measured by the reduction of residual norm error for both deterministic and random value functions (Table 1). As predicted by the theory, the method also reduces the error in isotropic residual matrix norm, as seen in Table 2, but these improvements seem more marginal. This behavior is present for all different levels of sample count N we tested. Therefore, while there are theoretical limitations in the nonsymmetric case of operator shifting that make the theory less powerful than the SPD case, operator shifting still functions quite well on the Markov chain problems we've tested it on. However, we note that the set of Markov chain matrices is only a small subset of possible matrices; hence, how well operator shifting performs in practice on other nonsymmetric problems with noise remains to be seen and is a potential avenue for future work.

Note that the confidence intervals for the 2D and 3D problems in the isotropic setting seen in Table 2 are quite large. This is despite the fact that we use a very large number of samples (256,000 for 2D, and 25,600 for 3D) to estimate the isotropic error. However, the theory tells us that the naive isotropic error will always be greater than the shifted isotropic error, so this chart is provided more-so as a way to gauge the magnitude of the error reduction, rather than the existence of an error reduction.

10. Conclusion.

We conclude this paper by noting that we have accomplished two main goals. First, we have investigated the extent to which the SPD operator shifting theory of [7] can be applied to the general nonsymmetric matrix case. We have found that under the assumptions of noise symmetry and right-hand side isotropy, the optimal shift factor is always positive. This answers the question of whether or not operator shifting towards the origin always reduces error as it does in the positive-definite symmetric case. Moreover, we have fully characterized the pathological situations in which this does not happen. We have also investigated the small noise regime, where we showed that it is possible to discard the noise symmetry assumption.

Second, we have shown empirically that operator shifting can *still* reduce error for noisy Markov chain problems, even when the aforementioned theoretical assumptions are not satisfied. In particular, our numerical experiments do not satisfy the symmetry assumption, and the results for the deterministic reward function (Table 1) do not satisfy the isotropic assumption. Moreover, this reduction holds true across a number of different Markov chains.

One may continue this work by attempting to apply some form of operator shifting to real problems—for example, in control theory or RL, where the underlying Markov decision process may not be fully known and must be estimated from data. Another more theoretical possibility would be to investigate if the operator shifting framework can be applied to optimization to create optimization algorithms that are less vulnerable to noise in the objective function, as is common in many real world applications. Other potentially interesting avenues of work may include extending operator shifting to an infinite dimensional setting, or trying to learn an appropriate shift $\hat{\mathbf{K}}$ from data.

11. Source code.

For reproducibility and reference purposes, we provide an accompanying implementation of our algorithms and numerical experiments at https://github.com/UniqueUpToPermutation/OperatorShifting.

REFERENCES

[1] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*, Cambridge Stud. Adv. Math. 118, Cambridge University Press, Cambridge, 2010.

[2] A. Aspri, Y. Korolev, and O. Scherzer, *Data driven regularization by projection*, Inverse Problems, 36 (2020), 125009.

[3] D. Bertsekas, *Reinforcement Learning and Optimal Control*, Athena Scientific, Belmont, MA, 2019.

[4] I. R. Bleyer and R. Ramlau, *A double regularization approach for inverse problems with noisy data and inexact operator*, Inverse Problems, 29 (2013), 025004.

[5] A. Buccini, M. Donatelli, and R. Ramlau, *A semiblind regularization algorithm for inverse problems with application to image deblurring*, SIAM J. Sci. Comput., 40 (2018), pp. A452–A483.

[6] E. J. Candes and Y. Plan, *Matrix completion with noise*, Proc. IEEE, 98 (2010), pp. 925–936.

[7] P. A. Etter and L. Ying, *Operator Shifting for Noisy Elliptic Systems*, preprint, arXiv:2010.09656, 2020.

[8] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[9] W. James and C. Stein, *Estimation with quadratic loss*, in Breakthroughs in Statistics, Springer, New York, 1992, pp. 443–460.

[10] R. Keshavan, A. Montanari, and S. Oh, *Matrix completion from noisy entries*, in Advances in Neural Information Processing Systems, 22 (2009), pp. 952–960.

[11] S. Lunz, A. Hauptmann, T. Tarvainen, C.-B. Schönlieb, and S. Arridge, *On learned operator correction in inverse problems*, SIAM J. Imaging Sci., 14 (2021), pp. 92–127.

[12] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, *An Introduction to Sampling via Measure Transport*, preprint, arXiv:1602.05023, 2016.

[13] T. Palmer, G. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and M. Leutbecher, *Representing model uncertainty in weather and climate prediction*, Annu. Rev. Earth Planet. Sci., 33 (2005), pp. 163–193.

[14] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 2014.

[15] C. Soize, *A comprehensive overview of a non-parametric probabilistic approach of model uncertainties for predictive models in structural dynamics*, J. Sound Vibration, 288 (2005), pp. 623–652.

[16] C. Stein, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, CA, 1956, pp. 197–206.

[17] T. Tao, *Topics in Random Matrix Theory*, Grad. Stud. Math. 132, American Mathematical Society, Providence, RI, 2012.

[18] A. N. Tikhonov, *On the solution of ill-posed problems and the method of regularization*, Dokl. Akad. Nauk, 151 (1963), pp. 501–504.

[19] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.

[20] D. Xiu and G. E. Karniadakis, *The Wiener-Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.