Research in
the Mathematical Sciences

RESEARCH

Operator shifting for noisy elliptic systems

Philip A. Etter[1*] and Lexing Ying[2]

*Correspondence:
paetter@meta.com
[1] Meta Reality Labs, Redmond, WA, USA
Full list of author information is available at the end of the article

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 an elliptic linear system with the operator corrupted by noise. We assume the noise preserves positive definiteness, but otherwise, we make no additional assumptions about the structure of the noise. Under these assumptions, we propose the *operator shifting* framework, a collection of easy-to-implement algorithms that augment a noisy inverse operator by subtracting an additional shift term. In a similar fashion to the James–Stein estimator, this has the effect of drawing the noisy inverse operator closer to the ground truth by reducing both bias and variance. We develop bootstrap Monte Carlo algorithms to estimate the required shift magnitude for optimal error reduction in the noisy system. To improve the tractability of these algorithms, we propose several approximate polynomial expansions for the operator inverse and prove desirable convergence and monotonicity properties for these expansions. We also prove theorems that quantify the error reduction obtained by operator shifting. In addition to theoretical results, we provide a set of numerical experiments on four different graph and grid Laplacian systems that all demonstrate the effectiveness of our method.

Keywords: Operator shifting, Random matrices, Monte Carlo, Polynomial expansion, Elliptic systems

1 Introduction

There are a plethora of different situations in the natural, mathematical, and computer sciences that necessitate computing the solution to a linear system of equations given by

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$ for $n \in \mathbb{N}$. When both the matrix $\mathbf{A}$ and $\mathbf{b}$ are known, there are many decades of research on how to solve the system Eq. (1) efficiently. Unfortunately, for a variety of reasons, it is often the case that the true matrix $\mathbf{A}$ is not known exactly, and must be estimated from data (see [15,18]). In this situation, there is an error between the unobserved true matrix $\mathbf{A}$ and the matrix $\hat{\mathbf{A}}$ one constructs from data. The discrepancy between $\mathbf{A}$ and $\hat{\mathbf{A}}$ is often referred to as *model uncertainty*, as it stems from incomplete or inaccurate information about the underlying system. This model uncertainty means that with a naive application of the inverse of the observed matrix $\hat{\mathbf{A}}$, one is not solving the desired system Eq. (1), but rather, the system

$$\hat{\mathbf{A}}\hat{\mathbf{x}} = \mathbf{b}, \tag{2}$$

where $\hat{\mathbf{x}} = \hat{\mathbf{A}}^{-1}\mathbf{b} \in \mathbb{R}^n$ is the solution we observe when we solve the observed system naively. Often, we will write

$$\hat{\mathbf{A}} = \mathbf{A} + \hat{\mathbf{Z}}, \tag{3}$$

where one can think of the matrix $\hat{\mathbf{Z}}$ as constituting the noise or sampling error in our measurements of the system Eq. (1). Hence, the sampling error $\hat{\mathbf{Z}}$ between $\mathbf{A}$ and $\hat{\mathbf{A}}$ translates into an error between the true solution $\mathbf{x}$ and the naively estimated solution $\hat{\mathbf{x}}$.

The question of interest in this paper is whether, using the information available to us, we can find a better approximation $\tilde{\mathbf{x}}$ for the true solution $\mathbf{x}$ by modifying how we solve the sampled system Eq. (2). "Better" here means in the sense of average error measured in the norm of some symmetric positive definite matrix $\mathbf{B}$, i.e., that we have

$$\mathcal{E}_{\mathbf{B}}(\tilde{\mathbf{x}}) < \mathcal{E}_{\mathbf{B}}(\hat{\mathbf{x}}), \tag{4}$$

where the error functional $\mathcal{E}_{\mathbf{B}}(\cdot)$ is defined as

$$\mathcal{E}_{\mathbf{B}}(\hat{\mathbf{x}}) \equiv \mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{B}}^2] = \mathbb{E}[\|\hat{\mathbf{A}}^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b}\|_{\mathbf{B}}^2], \tag{5}$$

where the norm $\|\cdot\|_{\mathbf{B}}$ is defined $\|\mathbf{x}\|_{\mathbf{B}}^2 = \mathbf{x}^T\mathbf{B}\mathbf{x}$. The two norms of particular interest to us are the L^2 norm (for obvious reasons), i.e., $\mathbf{B} = \mathbf{I}$, as well as the energy norm, i.e., $\mathbf{B} = \mathbf{A}$, as the latter is an important metric of error in many physical problems.

Many traditional techniques approach this problem by imposing Bayesian regularization conditions on the sampled solution $\hat{\mathbf{x}}$ (e.g., Tikhonov regularization [22]) or applying post-processing on $\hat{\mathbf{x}}$. In this paper, we take a fundamentally different tact. Instead of thinking about the problem of improving the individual estimates $\hat{\mathbf{x}}$ of solutions $\mathbf{x}$, we propose herein a framework for thinking about the problem in terms of linear operators. We contend that this paradigm shift is quite useful—as it is often the case that one may be interested in solving more than just one system of the form Eq. (2) given a single estimate $\hat{\mathbf{A}}$ of the matrix $\mathbf{A}$. In this situation, it often makes more sense to think of improving the estimator $\hat{\mathbf{A}}^{-1}$ rather than improving individual estimators $\hat{\mathbf{x}}$, although the two are obviously related. In light of this, we will amend our earlier objective Eq. (5) slightly. Namely, instead of achieving low error on just a single right-hand side $\mathbf{b}$, we want to simultaneously perform well on a whole collection of possible right-hand sides of interest. For this reason, we suppose that $\mathbf{b}$ is sampled from a distribution $\mathcal{B}$ and that our goal is to reduce the average error over this distribution,

$$\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}}[\|\hat{\mathbf{A}}^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b}\|_{\mathbf{B}}^2]. \tag{6}$$

In the interest of building out this new perspective, we propose a novel method we call *operator shifting*. The idea of operator shifting is to add an augmenting term to the sampled inverse operator $\hat{\mathbf{A}}^{-1}$, yielding a family of operators

$$\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}}(\hat{\mathbf{A}}^{-1}) \tag{7}$$

parameterized by an shift factor $\beta \in \mathbb{R}$, for a choice of shift operator $\hat{\mathbf{K}}(\hat{\mathbf{A}}^{-1}) \in \mathbb{R}^{n\times n}$ depending on the problem setting. Note that the shift operator is a function of the sampled matrix $\hat{\mathbf{A}}$. Our new approximation for $\mathbf{x}$ is then given by

$$\tilde{\mathbf{x}}_\beta = (\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})\mathbf{b} \tag{8}$$

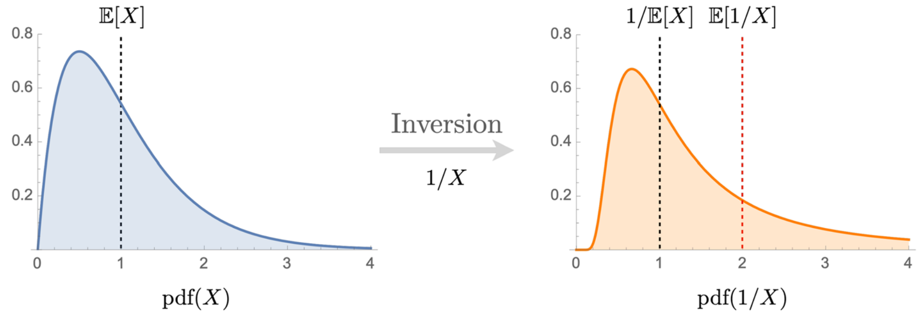

Fig. 1 An example of the overshooting effect. If we take a single sample of the scalar random variable $X \sim \Gamma(2, 1/2)$, and invert it, the pdf of the inverted $1/X$ has an expectation that is significantly larger (x2) than the inversion $1/\mathbb{E}[X]$. This means that naively trying to estimate $1/\mathbb{E}[X]$ with only a single sample will likely give a significant overestimate. The same principle also applies when X is a random matrix

Through judicious selection of the shift operator $\hat{\mathbf{K}}$, we show that one can estimate a β that will reduce error by a factor that depends on the variance of the naive solution $\hat{\mathbf{x}}$. As we will see, the power of operator shifting lies in the fact that the technique works under very minimal assumptions on the randomness structure of $\hat{\mathbf{A}}$; in general, the only assumption we need to guarantee error reduction is that $\hat{\mathbf{A}}$ is an unbiased estimator of $\mathbf{A}$, and even this assumption can be relaxed.

The most obvious choice of shift operator is perhaps to shift the naive estimate $\hat{\mathbf{A}}^{-1}$ toward the origin by taking $\hat{\mathbf{K}}(\hat{\mathbf{A}}) = \hat{\mathbf{A}}^{-1}$. There are two fundamental reasons why one might expect this to be a good choice of shift—the first concerns the bias of the estimate $\hat{\mathbf{A}}^{-1}$ and the second concerns the variance.

1. **Bias**: For symmetric positive definite matrices, the matrix inversion operation is convex with respect to the Löwner order[1]. A matrix analogue of Jensen's inequality, therefore, suggests that, depending on the variance in $\hat{\mathbf{A}}$, $\hat{\mathbf{A}}^{-1}$ will substantially overestimate $\mathbf{A}^{-1}$ on average (i.e., $\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1}$). Hence, it makes to shift $\hat{\mathbf{A}}^{-1}$ toward the origin in order to reduce the bias in $\hat{\mathbf{A}}^{-1}$. We provide an illustration of this bias in Fig. 1.
2. **Variance**: Shrinking the estimate toward a fixed point (i.e., the origin) simultaneously has the effect of reducing variance in the estimator. This is analogous to the seminal work of James and Stein [11] that demonstrated the standard mean estimator is inadmissible, as shrinking the estimator slightly toward the origin always reduces average error.

Therefore, the confluence of these two factors suggests that we should expect a reduction in both bias and variance and hence a more accurate estimator as a result. Indeed, in this paper, we prove that, with only minimal assumptions on the randomness of $\hat{\mathbf{A}}$, the optimal reduction in error always comes from a shift toward the origin.

We structure the remainder of the paper as follows. First, we give an overview of related work in Sect. 2, then we set out basic assumptions and notations in Sect. 3. In Sect. 4, we prove bounds that quantify how much error the operator shifting technique can reduce in various norms (e.g., the Frobenius norm). Afterward, Sect. 5 focuses specifically on

[1] Recall the definition of the Löwner order: $A \preceq B$ when $x^T A x \leq x^T B x$ for all vectors $x \in \mathbb{R}^n$.

error in the energy norm (i.e., when $\mathbf{B} = \mathbf{A}$, the norm defined by the elliptic operator itself) and provides the analogous bounds. Afterward, Sect. 6 introduces the bootstrap formalism we will use in Sect. 7 to produce a computable Monte Carlo estimator for the optimal β, and hence for $\mathbf{A}^{-1}$. But since this Monte Carlo estimator requires a full matrix solve for every sample, we turn for the remainder of the paper to the problem of efficient computation of the optimal β. We show in Sect. 8 that the energy norm objective has a polynomial expansion with monotonicity properties that are immensely useful for efficiently computing a good choice of β. Unfortunately, this expansion does not converge for all matrices—but we show in Sect. 9 that one can shift the base point of the polynomial expansion while maintaining monotonicity. Then, Sect. 10 considers this base-point shifting approach on a variable point-wise basis. And finally, in Sect. 11, we present numerical experiments to verify the theoretical results in this paper.

Note that we consider only elliptic systems in this paper, i.e., requiring that $\mathbf{A}$ is symmetric positive definite and $\hat{\mathbf{A}}$ is symmetric positive definite almost surely—however, one could theoretically apply the techniques we present herein to asymmetric systems as well, but we do not provide any theoretical guarantees in the asymmetric case.

Finally, in order to help readers quickly implement our method without getting caught up in all of the surrounding mathematical details, we provide the quick start Sect. 8.3 to give readers an alternate entry point to the algorithm we present in this paper. For the accompanying source code for this paper, please see Sect. 13.

2 Related work

The spirit of our approach is heavily influenced by James–Stein Estimation [11]. In Stein's original paper, [19], he demonstrated the (at the time) shocking phenomenon that the standard mean estimator is actually *inadmissible*[2] for the quadratic loss in dimensions ≥ 3. The reason behind this has to do with the fact that one can always advantageously trade-off bias for a reduction in variance by shrinking the estimator toward any fixed point. At a fundamental level, one can frame our work as taking this idea and applying it to the novel setting of matrices corrupted by noise.

Some particularly relevant work pertains to debiasing distributed second-order optimization. Second-order optimization methods often rely on solving a symmetric linear system involving the Hessian of an object (positive definite if the objective is strongly convex). However, in many machine learning applications, the objective is composed of a summation of terms over a massive corpus of data, such that computing the true Hessian is extremely expensive. Instead, practitioners often turn to stochastic optimization methods that subsample the objective and its derivatives by using only a small section of the corpus at a time. However, for an optimization problem given by

$$\min_x \sum_{i=0}^{m} f_i(x) , \tag{9}$$

the true Hessian and approximated Hessian are given as follows:

$$\mathbf{H} = \sum_{i=0}^{m} \mathbf{H}_i \qquad \hat{\mathbf{H}} = \sum_{i=0}^{m} \frac{\hat{p}_i}{\mathbb{E}[\hat{p}_i]} \mathbf{H}_i , \tag{10}$$

[2] An inadmissible estimator $\hat{\mu}$ for a quantity μ is one for which there exists an alternate estimator $\hat{\mu}'$ that always achieves better loss regardless of the value of μ.

where $\mathbf{H}_i$ is the Hessian of f_i and $p_i \in \{0, 1\}$ is a random variable that determines if the ith item in the corpus is in the current mini-batch. The naive estimator $\hat{\mathbf{H}}$ has an upward bias and there has been work in the literature on how to de-bias the estimator using determinantal averaging [8]. However, this approach is fundamentally limited to matrix ensembles of the form Eq. (10). In this paper, the types of noise we consider are far more general.

Other relevant work has been done in the field of matrix sketching. Matrix sketching is a technique to reduce the complexity of a least-squares/linear problem by using random sketches of the rows/columns of the matrix. This process can likewise produce estimates that are substantially biased. One can attempt to address this bias by modifying regularization or other problem parameters [9]. This can be applied to the aforementioned second-order optimization problem by using a Hessian sketch. However, again, the technique is tied to a very specific type of matrix noise.

There has been related work on the mathematical analysis of linear algebra algorithms in the noisy regime. For example, [3] studies the randomized Kaczmarz algorithm on so-called "Doubly-Noisy Linear Systems" (i.e., systems with noise both in $\mathbf{A}$ and $\mathbf{b}$). The setting is very similar to the one we have presented; however, the approach of the work is a through the lens of a specific solver (e.g., Kaczmarz), rather than statistical estimator approach we take herein. Moreover, due to the difficulties of such analyses, results are typically limited to specific types of multiplicative noise.

Beyond the world of James–Stein estimation and operator de-biasing, there are a number of immediate connections between the work done herein and previous work in the field of statistical inverse problems. In various inverse problems, one is interested in estimation from noisy or incomplete measurements. For example, *semi-blind deconvolution* involves trying to reconstruct a function convolved with a kernel where the kernel is known, but with some uncertainty. Note that this is distinct from fully *blind deconvolution* where one has no information about the kernel. In the sense that the measurement operator is corrupted by noise or uncertainty, and the goal is to recover the underlying object by inverting a linear system, this setting is quite similar to our own and hence worth mentioning.

A common approach to these problems is to induce regularization on both the operator and the recovery target. For example, Total Least Squares algorithms as pioneered by Golub and Van Loan [10] optimize over small perturbations to the noisy operator as well as the linear regression weights. Similar approaches specific to semi-blind deconvolution include introducing a free estimate of the underlying kernel with regularization to match the observed data [5]. Another technique in semi-blind deconvolution is to treat the full operator as a free variable and introduce optimization constraints to make sure that the operator and the observations do not deviate by too much [4].

Unfortunately, these types of techniques that operate over the operator suffer from a number of flaws. The most obvious is that introducing $\sim n^2$ additional free variables into an optimization problem also introduces a substantial additional computational cost. Along with this computational cost also comes a much more severe chance of overfitting unless regularization is handled appropriately. Furthermore, these regularization techniques implicitly depend on good Bayesian priors for what the underlying target and the operator should look like. In the absence of good priors, this optimization avenue may not be as viable. In contrast, all optimizations performed in the operator shifting framework we present here are only over a single variable β, and hence are not subject to these concerns.

Other situations in the statistical inverse problem literature that involve noisy or uncertain operators include circumstances where the forward operator may be far too expensive to apply directly, and hence must be replaced by a learned proxy for efficient computation [13]. Another setting in the literature is when one has a set of noisy input-output pairs of the underlying operator. Work has been done on using these input-output pairs to construct regularizers for solving the inverse problem [2]. Nonetheless, these approaches and settings are quite different from the approach and setting we present in this paper.

Beyond the field of statistical inverse problems, a pertinent area of the literature related to our work is *model uncertainty*. Quantifying and representing model uncertainty is important in many different fields of computational science, ranging from structural dynamics [18] to weather and climate prediction [15]. However, work relating to model or parameter uncertainty is usually domain-specific and focuses more on establishing a model for uncertainty than it does on trying to reduce error in the resulting predictions. In contrast, our work focuses entirely on reducing error, rather than quantifying it. Our work is also not restricted to a particular domain, class of problems, or randomness structure, as long as those problems are linear.

We note that our setting shares some similarities with the problem of uncertainty quantification (UQ). However, the problem we face here is different from the standard uncertainty quantification setting in a subtle but very important way. In UQ, one is usually given a distribution $\mathcal{P}$ and a map T and asked to estimate statistics about the pushforward distribution $T_*\mathcal{P}$ (i.e., expectation, standard error, etc.). Practitioners typically accomplish this task via Monte Carlo techniques [14] or some form of stochastic Galerkin projection [24] or collocation method [23]. However, for our purposes, we are more interested in the image of the statistic $\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}$ under matrix inversion, rather than quantifying the pushforward of the distribution of $\hat{\mathbf{A}}$ under matrix inversion.

The central problem in this paper is also not dissimilar to the setting of matrix completion seen in [6,12]. In matrix completion, one usually seeks to recover a low-rank ground truth matrix $\mathbf{M}_{ij}$ from observations that have been corrupted by additive noise, e.g., $\mathbf{N} = \mathbf{M} + \mathbf{Z}$. If $\mathcal{P}_\Omega$ denotes the subset sampling operator on matrix space, then one is trying to recover $\mathbf{M}$ from

$$\mathcal{P}_\Omega(\mathbf{N}) = \mathcal{P}_\Omega(\mathbf{M}) + \mathcal{P}_\Omega(\mathbf{Z}). \tag{11}$$

However, the operator shifting and matrix completion settings are subtly different. The matrix completion analogue of $\mathbf{A}$ is the actual linear operator $\mathcal{P}_\Omega$ and not the matrix $\mathbf{M}$. Morally, one may think of the matrix competition problem as solving the underdetermined linear system

$$\mathcal{P}_\Omega(\mathbf{M}) = \mathcal{P}_\Omega(\mathbf{N}) \tag{12}$$

by assuming a low-rank regularity on $\mathbf{M}$. The randomness in this problem lies completely in the right-hand side $\mathbf{N}$, and not in the actual linear operator $\mathcal{P}_\Omega$.

We also draw attention to the related field of perturbation matrix analysis. In this setting, one is usually interested in proving results about how various properties of matrices change under a perturbation to the elements of the matrix. A seminal example of work in this field is the Davis-Kahan theorem [7], which quantifies the extent to which the invariant sub-spaces of a matrix change under perturbations. In a similar vein, work in backward stability analysis revolves around understanding the behavior of the solution of a linear system under perturbations to the matrix. However, backward stability analysis typically adopts a worst-case mentality in analysis. In contrast, we care about average case error—and more importantly, how one can reduce it.

We should briefly mention that the mathematical branch of random matrix theory (RMT) studies the spectral properties of random matrix ensembles [1,21]. 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 addition to these tangentially related settings, we also call attention to the similarity of some of our techniques to those in harmonic analysis. It is well known that the method of summation of an infinite series can affect the conditions under which it converges, as well as the quality of the convergence. For example, the Fourier series of a continuous function f on the unit interval [0, 1] may not converge pointwise to f if summed naively. But Fejér's theorem (see [20]) states that Césaro and Abel sums of the Fourier series of an integrable function f converge uniformly to f at any point of continuity. Our work takes on a similar flavor in that it revolves heavily around the convergence properties of partial sums of the infinite series expansion of the matrix function $f(\mathbf{A}) = \mathbf{A}^{-1}$. These partial sums are critical to accelerating an otherwise expensive Monte Carlo computation. Hence we develop methods of partial summation that have desirable properties—such as convergence and monotonicity.

In conclusion, we do not believe that the setting we introduce in this paper, where the operator is noisy, has been studied in the proposed fashion before. There is little precedent in the literature for the operator shifting method we present herein.

3 Basic assumptions and notation

For the sake of transparency, before we go any further, we will make a number of assumptions on the nature of randomness on $\hat{\mathbf{A}}$—as this will help clarify the setting. We will use D_{ω^*} to denote the distribution of $\hat{\mathbf{A}}$. Throughout this paper, we will use $S_+(\mathbb{R}^n)$ to denote the set of symmetric positive definite matrices in $\mathbb{R}^{n\times n}$. We make the following extremely lax assumptions about the randomness of $\hat{\mathbf{A}}$:

1. *Almost-Surely Positive Definite*: We assume that $\hat{\mathbf{A}} \in S_+(\mathbb{R}^n)$ almost surely. We believe this is a very reasonable assumption if $\hat{\mathbf{A}}$ is generated from an elliptic problem whose parameters are subject to noise, it is extremely unlikely that any value of the underlying problem parameters will destroy ellipticity.
2. *Unbiased, or Downward-Biased*: We assume that $\hat{\mathbf{A}}$ is an unbiased estimate of $\mathbf{A}$, i.e., that $\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}$. More generally, all of the machinery applies equally well when $\mathbb{E}[\hat{\mathbf{A}}] \preceq \mathbf{A}$.
3. *Finiteness of the Inverse Second Moment*: We assume that $\mathbb{E}[\hat{\mathbf{A}}^{-2}] \prec \infty$. Note that this is necessary to ensure that our measure of error Eq. (5) actually exists for arbitrary choice of $\mathbf{b}$.

We note that these assumptions are surprisingly lax. Most importantly, we *do not* assume that entries of $\hat{\mathbf{A}}$ are independent. In the context of the theory to be presented herein, this assumption is irrelevant and not needed. Moreover, for all of the numerical examples we present, the entries of $\hat{\mathbf{A}}$ will in fact be correlated random variables. We believe this helps reinforce the generality of the operator shifting framework.

4 Operator shifting in operator inner product norms

To begin, we want to shift the perspective of estimating the result of a matrix solve $\mathbf{A}^{-1}\mathbf{b}$ into the problem of estimating the matrix inverse $\mathbf{A}^{-1}$ itself. As we will see, these are actually the same problem, and the matrix inverse estimation viewpoint is more in line with the traditional perspective of the Stein estimation. Indeed, in practice, one may not be simply interested in a single right-hand side $\mathbf{b}$, but rather, producing a good inverse operator for a wide variety of potential right-hand sides $\mathbf{b}$. As discussed in our introduction, we encode this desire by changing our error metric to have $\mathbf{b}$ be sampled from a known distribution $\mathcal{B}$ and then measuring the average error under this distribution,

$$\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}[\|\hat{\mathbf{A}}^{-1}\mathbf{b}-\mathbf{A}^{-1}\mathbf{b}\|_{\mathbf{B}}^2]. \tag{13}$$

When $\mathbf{b}$ is made into a random variable, the result actually induces a metric on the space of operators $\mathbb{R}^{n\times n}$. To see this, let $\mathbf{R}\equiv\mathbb{E}[\mathbf{b}\mathbf{b}^T]$ denote the second moment matrix of the distribution $\mathcal{B}$ and consider the following manipulations,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}[\|\hat{\mathbf{A}}^{-1}\mathbf{b}-\mathbf{A}^{-1}\mathbf{b}\|_{\mathbf{B}}^2]\\
&\quad=\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}[\mathbf{b}^T(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})^T\mathbf{B}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})\mathbf{b}]\\
&\quad=\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\operatorname{tr}[\mathbf{b}^T(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})^T\mathbf{B}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})\mathbf{b}]\\
&\quad=\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\operatorname{tr}[(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})^T\mathbf{B}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})\mathbf{b}\mathbf{b}^T]\\
&\quad=\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\operatorname{tr}[(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})^T\mathbf{B}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})\,\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}(\mathbf{b}\mathbf{b}^T)]\\
&\quad=\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\operatorname{tr}[(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})^T\mathbf{B}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})\mathbf{R}]\\
&\quad=\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\operatorname{tr}[\mathbf{R}^{1/2}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})^T\mathbf{B}(\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1})\mathbf{R}^{1/2}]
\end{aligned} \tag{14}$$

The natural metric and norm on operator space that corresponds to this notion of error is therefore defined by:

$$\begin{aligned}
&\langle\mathbf{X},\mathbf{Y}\rangle_{\mathbf{B},\mathbf{R}}\equiv\operatorname{tr}[\mathbf{R}^{1/2}\mathbf{X}^T\mathbf{B}\mathbf{Y}\mathbf{R}^{1/2}]\\
&\|\mathbf{X}\|_{\mathbf{B},\mathbf{R}}^2\equiv\langle\mathbf{X},\mathbf{X}\rangle_{\mathbf{B},\mathbf{R}},
\end{aligned} \tag{15}$$

where $\mathbf{X},\mathbf{Y}\in\mathbb{R}^{n\times n}$. Note that the often-used Frobenius norm $\|\cdot\|_F$ and corresponding inner product $\langle\cdot,\cdot\rangle_F$ defined by

$$\begin{aligned}
&\langle\mathbf{X},\mathbf{Y}\rangle_F\equiv\operatorname{tr}(\mathbf{X}^T\mathbf{Y}),\\
&\|\mathbf{X}\|_F^2\equiv\langle\mathbf{X},\mathbf{X}\rangle_F
\end{aligned} \tag{16}$$

is a special case of this class of norms that we obtain when $\mathbf{B}=\mathbf{R}=\mathbf{I}$.

Therefore, the pivot from thinking about obtaining a lower error in a specific $\mathbf{b}$ to obtaining a lower error on a collection of $\mathbf{b}$ essentially changes our problem to an estimation problem for $\hat{\mathbf{A}}^{-1}$ in the $\|\cdot\|_{\mathbf{B},\mathbf{R}}$ norm. Corresponding to this change in outlook, we will use the notation

$$\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})\equiv\mathbb{E}\|\hat{\mathbf{A}}^{-1}-\mathbf{A}^{-1}\|_{\mathbf{B},\mathbf{R}}^2=\mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}[\|\hat{\mathbf{A}}^{-1}\mathbf{b}-\mathbf{A}^{-1}\mathbf{b}\|_{\mathbf{B}}^2], \tag{17}$$

to denote the $(\mathbf{B},\mathbf{R})$-error of the estimator $\hat{\mathbf{A}}^{-1}$. When considering the Frobenius norm, we will simply use the notation $\mathcal{E}_F(\cdot)$.

Now, let us return to the central technique of this paper and introduce an operator shift to the sampled operator inverse $\hat{\mathbf{A}}^{-1}$,

$$\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}}(\hat{\mathbf{A}}^{-1}). \tag{18}$$

A quick dimensional analysis of the above quantity suggests that $\hat{\mathbf{K}}(\hat{\mathbf{A}}^{-1})$ should be a linear function of $\hat{\mathbf{A}}^{-1}$. Therefore, it makes sense to study operator shifts of the form

$$\hat{\mathbf{K}}(\hat{\mathbf{A}}^{-1}) = \mathbf{C}\hat{\mathbf{A}}^{-1}\mathbf{D}, \tag{19}$$

where $\mathbf{C}, \mathbf{D}$ are matrices.

To start, let us consider the choice of $\mathbf{C} = \mathbf{D} = \mathbf{B}$ where $\mathbf{B}$ is symmetric positive definite. When $\mathbf{B} = \mathbf{I}$, this is analogous to Stein shrinkage but for matrices instead of vector quantities. We will see that there are more choices of $\mathbf{C}$ and $\mathbf{D}$ available that yield interesting theoretical results for error reduction; however, the case $\hat{\mathbf{K}}(\hat{\mathbf{A}}^{-1}) = \mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{B}$ is the simplest to analyze and has the most interesting error bound. We note that this is analogous to simply applying a Stein shrinkage factor to the naive estimator $\hat{\mathbf{A}}^{-1}$. As we laid out in our introduction, we should expect that for some $\beta > 0$, we should achieve a smaller estimation error of $\mathbf{A}^{-1}$. The central lemma which bears this intuitive expectation out is the following Löwner Order Inversion Lemma,

Lemma 1 (Löwner Order Inversion) *Suppose that* $\mathbf{A} \in S_+(\mathbb{R}^n)$ *and* $\hat{\mathbf{A}} \in S_+(\mathbb{R}^n)$ *almost surely. Moreover, suppose that,* $\mathbf{A}$ *spectrally dominates* $\hat{\mathbf{A}}$ *in expectation, i.e.,*

$$\mathbb{E}[\hat{\mathbf{A}}] \preceq \mathbf{A}, \tag{20}$$

then, matrix inversion inverts the expected Löwner order, i.e.,

$$\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1} \tag{21}$$

The proof of this lemma simply depends upon the convexity of the function $(\cdot)^{-1}$ on the cone of positive definition matrices. However, we relegate this proof to the "Appendix" so that we can continue on to the main result of this section,

Theorem 1 (Operator Shifting Bounds) *Under the assumptions in Sect. 3, consider operator shifting in any* $(\mathbf{B}, \mathbf{B})$ *matrix norm* $\|\cdot\|_{\mathbf{B},\mathbf{B}}$ *(for example, the Frobenius norm). The operator shift* $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}$ *has an optimal shift factor that satisfies:*

$$1 \geq \sqrt{\frac{\mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{B},\mathbf{B}}}} \geq \beta^* \geq \frac{\mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{B},\mathbf{B}}} \geq 0. \tag{22}$$

The corresponding optimal reduction in relative error is given by

$$\max_{\beta\in\mathbb{R}} \frac{\mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1}) - \mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})}{\mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1})} \geq \frac{\mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{B},\mathbf{B}}}, \tag{23}$$

where $\mathcal{E}(\hat{\mathbf{X}})$ *is the mean squared error of matrix estimator* $\hat{\mathbf{X}}$ *in the* $\|\cdot\|_{\mathbf{B},\mathbf{B}}$ *matrix norm.*

This theorem tells us that if we approximate a good shift factor that comes close to β^* we should expect a reduction in error that is proportional to the error relative to the average squared norm. Moreover, it tells us roughly how large we should expect the optimal shift factor to be. If one already has a good estimate of the ratio $\mathcal{E}_{\mathbf{B},\mathbf{B}}(\hat{\mathbf{A}}^{-1})/\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{B},\mathbf{B}}$, one could use this as an approximate shift factor. Let us now establish this result in the following proof,

Proof For the scope of this proof, we use $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ to denote the $\mathbf{B}, \mathbf{B}$ matrix inner product $\langle \cdot, \cdot \rangle_{\mathbf{B},\mathbf{B}}$ and matrix norm $\| \cdot \|_{\mathbf{B}}$ with the subscript suppressed, with $\mathcal{E}(\cdot)$ denoting the corresponding estimator error $\mathcal{E}_{\mathbf{B},\mathbf{B}}(\cdot)$ for $\mathbf{A}^{-1}$.

To begin this proof we want to compute an expression for the shift factor β that optimizes the error. Expanding the error gives us

$$\mathcal{E}(\hat{\mathbf{A}}^{-1} - \beta \hat{\mathbf{K}}) = \mathcal{E}(\hat{\mathbf{A}}^{-1}) - 2\beta\, \mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle + \beta^2\, \mathbb{E}\|\hat{\mathbf{K}}\|^2 , \tag{24}$$

and hence, the optimal shift factor β^* is given by

$$\beta^* = \frac{\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle}{\mathbb{E}\|\hat{\mathbf{K}}\|^2} , \tag{25}$$

and the corresponding optimal error is

$$\mathcal{E}(\hat{\mathbf{A}}^{-1} - \beta^* \hat{\mathbf{K}}) = \mathcal{E}(\hat{\mathbf{A}}^{-1}) - \frac{(\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle)^2}{\mathbb{E}\|\hat{\mathbf{K}}\|^2} , \tag{26}$$

We therefore expand the quantity $\mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle$ above,

$$\begin{aligned}
\mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle &= \mathbb{E}\,\mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{B}^{1/2}) - \mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{1/2}) \\
&= \mathbb{E}\,\mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{B}\mathbf{A}^{-1/2}) - \mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{1/2}) \\
&= \mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{B}\,\mathbb{E}[\hat{\mathbf{A}}^{-1}]\,\mathbf{B}\mathbf{A}^{-1/2}) - \mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{1/2}) \\
&\geq \mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1/2}) - \mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{1/2}) \\
&= \mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{1/2}) - \mathrm{tr}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{1/2}) = 0
\end{aligned} \tag{27}$$

where we have used the fact that $\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1}$ by Lemma 1. Now returning to Eq. (26), we obtain

$$\begin{aligned}
\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle &= \mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle \\
&\geq \mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle - \mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle \\
&= \mathbb{E}\langle \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle \\
&= \mathcal{E}(\hat{\mathbf{A}}^{-1}) .
\end{aligned} \tag{28}$$

For a bound in the opposite direction, we simply invoke Cauchy-Schwarz:

$$\begin{aligned}
\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1} \rangle &\leq \sqrt{\mathbb{E}\|\hat{\mathbf{K}}\|^2\, \mathbb{E}\|\hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|^2} \\
&= \sqrt{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2\, \mathcal{E}(\hat{\mathbf{A}}^{-1})}
\end{aligned} \tag{29}$$

Therefore, the desired result follows immediately from Eqs. (25) and (26). □

It is possible to extend this result to a wider range of possible choices of shifts where $\mathbf{C}$ and $\mathbf{D}$ are not identities. The proof sketch remains more or less the same but requires a few extra steps. Unfortunately, not all possible choices of shifts admit an analogy to the above proof. There is a compatibility constraint on $\mathbf{C}$ and $\mathbf{D}$ that forces the result to play especially nice with the $\langle \cdot, \cdot \rangle_{\mathbf{B},\mathbf{R}}$ inner product, namely,

$$\mathbf{R}\mathbf{D}^T = \mathbf{C}^T\mathbf{B}, \qquad (\mathbf{R}\mathbf{D}^T) = (\mathbf{R}\mathbf{D}^T)^T, \qquad \mathbf{R}\mathbf{D}^T \succeq \mathbf{0} . \tag{30}$$

Some examples when this may be the case are as follows:

1. The trivial case where $\mathbf{R} = \mathbf{B} = \mathbf{D} = \mathbf{C} = \mathbf{I}$.
2. The case where $\mathbf{R} = \mathbf{B}$ and $\mathbf{D} = \mathbf{C} = \mathbf{I}$.

3. The case where $\mathbf{C} = \mathbf{R}$, $\mathbf{D} = \mathbf{B}$ and $[\mathbf{B}, \mathbf{R}] = 0$.
4. The case where $\mathbf{C} = \mathbf{B}^{-1}$ and $\mathbf{D} = \mathbf{R}^{-1}$.

However, with these constraints, we can essentially repeat the previous theorem for a wider selection of possible operator shifts and obtain the following more general result,

Theorem 2 *Under the assumptions in Sect. 3, consider operator shifting in the $\|\cdot\|_{\mathbf{B},\mathbf{R}}$-norm. Any operator shift $\hat{\mathbf{K}} = \mathbf{C}\hat{\mathbf{A}}^{-1}\mathbf{D}$ such that $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times n}$ satisfy the compatibility conditions Eq. (30) has an optimal shift factor that satisfies:*

$$\sqrt{\frac{\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{BC},\mathbf{DRD}^T}}} \geq \beta^* \geq \frac{\mathcal{E}_{\mathbf{C}^T\mathbf{B},\mathbf{RD}^T}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{BC},\mathbf{DRD}^T}} \geq 0\,. \tag{31}$$

And the corresponding optimal reduction in error is given by

$$\max_{\beta\in\mathbb{R}} \frac{\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - \mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})}{\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \geq \frac{\mathcal{E}_{\mathbf{C}^T\mathbf{B},\mathbf{RD}^T}(\hat{\mathbf{A}}^{-1})^2}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{BC},\mathbf{DRD}^T}\,\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}\,, \tag{32}$$

where $\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{X}})$ is the mean squared error of matrix estimator $\hat{\mathbf{X}}$ in the $\|\cdot\|_{\mathbf{B},\mathbf{R}}$-norm.

The proof of this statement is relegated to the “Appendix” because it is not that fundamentally different from the proof that we just gave for the more specialized statement.

We observe that the theorems above tell us how much error one could expect if we could produce a perfect estimate of the shift factor β. However, this is unfortunately not possible. If one examines the optimal shift factor for (as an example) the Frobenius error,

$$\beta^* = \frac{\mathbb{E}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_F}{\mathbb{E}\|\hat{\mathbf{K}}\|_F^2}\,, \tag{33}$$

we see very clearly that this expression depends on $\mathbf{A}$, a quantity that we don’t know. This means that it must be approximated. The bounds of Theorem 1 give us some idea of roughly how large β will be, as it is very likely that the user will have a good idea of how large the relative error $\mathcal{E}_F(\hat{\mathbf{A}}^{-1})/\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|_F^2$ is, as it corresponds to the amount of noise in the estimate.

Alternatively, another method to approximate β^* is to try to bootstrap it using synthetic samples of $\hat{\mathbf{A}}^{-1}$. Naturally, one cannot draw additional samples from the distribution D_{ω^*}; however, it is usually the case that by observing $\hat{\mathbf{A}}$, we have some ideas of the parameters that generate the distribution D_{ω^*} and hence can draw synthetic samples from an approximate distribution $D_{\hat{\omega}}$ that can be used to build a Monte Carlo estimate for β^*. However, we will table this discussion until later in the paper when we talk about algorithmic implementations of operator shifting. For now, let us focus primarily on theoretical results.

5 Operator shifting in the energy norm

The previous section represents a class of operator shifts that one might use when the norm $\mathbf{B}$ is actually known; however, for many elliptic problems, the norm defined by the true matrix $\mathbf{A}$ itself is an important error norm. For example, in many physical problems, $\mathbf{x}^T\mathbf{A}\mathbf{x}$ measures the energy of a state $\mathbf{x}$ and hence can be even more important as a metric than L^2. Moreover, the case of $\mathbf{A}$ is special because the optimal shift factor reads

$$\beta^* = \frac{\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}}}{\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{A},\mathbf{R}}}, \tag{34}$$

and hence the $\mathbf{A}^{-1}$ in the numerator will cancel with the $\mathbf{A}$ in the $\langle\cdot,\cdot\rangle_{\mathbf{A},\mathbf{R}}$-inner product, where once again $\mathbf{R}$ is the second moment matrix of $\mathbf{b}$. This choice of norm leads more or less to a very similar set of bounds on the optimal shift factor β^* but for a slightly different set of norms,

Theorem 3 *Under the assumptions in Sect. 3, consider operator shifting in any* $(\mathbf{A}, \mathbf{R})$ *matrix norm* $\|\cdot\|_{\mathbf{A},\mathbf{R}}$ *(for example, the energy norm). The operator shift* $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}$ *has an optimal shift factor that satisfies:*

$$1 \geq \sqrt{\frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{R}}}} \geq \beta^* \geq \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{R}}} \geq 0. \tag{35}$$

The corresponding optimal reduction in relative error is given by

$$\max_{\beta\in\mathbb{R}} \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - \mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})}{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \geq \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{R}}} \tag{36}$$

where $\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{X}})$ *is the mean squared error of matrix estimator* $\hat{\mathbf{X}}$ *in the* $\|\cdot\|_{\mathbf{A},\mathbf{R}}$*-norm.*

Proof Almost exactly the same as Theorem 1. The one place where the proof diverges is the equation Eq. (27). In this setting, we instead have:

$$\begin{aligned}\mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle &= \mathbb{E}\,\mathrm{tr}(\mathbf{R}^{1/2}\mathbf{A}^{-1}\mathbf{A}\hat{\mathbf{A}}^{-1}\mathbf{R}^{1/2}) - \mathrm{tr}(\mathbf{R}^{1/2}\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{R}^{1/2})\\ &= \mathbb{E}\,\mathrm{tr}(\mathbf{R}^{1/2}\hat{\mathbf{A}}^{-1}\mathbf{R}^{1/2}) - \mathrm{tr}(\mathbf{R}^{1/2}\mathbf{A}^{-1}\mathbf{R}^{1/2})\\ &= \mathrm{tr}(\mathbf{R}^{1/2}\,\mathbb{E}[\hat{\mathbf{A}}^{-1}]\,\mathbf{R}^{1/2}) - \mathrm{tr}(\mathbf{R}^{1/2}\mathbf{A}^{-1}\mathbf{R}^{1/2})\\ &\geq \mathrm{tr}(\mathbf{R}^{1/2}\mathbf{A}^{-1}\mathbf{R}^{1/2}) - \mathrm{tr}(\mathbf{R}^{1/2}\mathbf{A}^{-1}\mathbf{R}^{1/2}) = 0\end{aligned} \tag{37}$$

□

We note that the admissible norms for the above theorem are slightly different than those of *Theorem* 1. In particular, while the proof of *Theorem* 1 required that the norm matrix $\mathbf{B}$ and second moment matrix $\mathbf{R}$ be identical, one is allowed to choose any second moment matrix $\mathbf{R}$ in the above theorem as long as $\mathbf{B} = \mathbf{A}$.

Now let us consider other possible operator shifts besides the trivial one $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}$. For a generalized version of the above theorem, using the energy norm also means that the possible operator shifts we can make and the conditions they must satisfy are slightly different. Indeed, for the energy norm, we consider only shifts in the form

$$\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}\mathbf{C}, \tag{38}$$

where $\mathbf{C}$ satisfies the compatibility conditions:

$$(\mathbf{R}\mathbf{C}^T) = (\mathbf{R}\mathbf{C}^T)^T, \qquad \mathbf{R}\mathbf{C}^T \succeq \mathbf{0}. \tag{39}$$

This type of shift gives the following theorem:

Theorem 4 *Under the assumptions in Sect. 3, consider operator shifting in energy norm* $\|\cdot\|_{\mathbf{A},\mathbf{R}}$. *Any operator shift* $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}\mathbf{C}$ *such that* $\mathbf{C}$ *satisfies the compatibility conditions*

Eq. (39) *has an optimal shift factor that satisfies:*

$$1 \geq \sqrt{\frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{C}^T\mathbf{R}\mathbf{C}}}} \geq \beta^* \geq \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}\mathbf{C}^T}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{C}^T\mathbf{R}\mathbf{C}}} \geq 0\,. \tag{40}$$

And the corresponding optimal reduction in relative error is given by

$$\max_{\beta\in\mathbb{R}} \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - \mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})}{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \geq \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}\mathbf{C}^T}(\hat{\mathbf{A}}^{-1})^2}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{C}^T\mathbf{R}\mathbf{C}}\,\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \tag{41}$$

where $\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{X}})$ *is the mean squared error of matrix estimator* $\hat{\mathbf{X}}$ *in the* $\|\cdot\|_{\mathbf{A},\mathbf{R}}$*-norm.*

We religate the proof of this theorem to "Appendix." Now, we finally turn to the problem of actually estimating the quantity β^*.

6 Bootstrap formalism

To be able to approximate the optimal shift factor β^* using Bootstrap Monte Carlo and write down a final algorithm for the operator shifting ideas presented above, we must first establish a formalism that allows one to generate synthetic samples of $\hat{\mathbf{A}}^{-1}$.

To build the formalism, we assume that there exists an underlying parameter space Ω (with sigma algebra Σ), where the parameters $\omega \in \Omega$ contain a description of the system that produces the matrices above (e.g., ω may be measurements of a scattering background, edge weights, vertex positions, etc.). We suppose the relationship between parameters and matrices is given by a measurable map

$$\mathcal{M} : \Omega \longrightarrow S_+(\mathbb{R}^n)\,. \tag{42}$$

For example, $\omega \in \Omega$ may be a weighted graph, and $\mathcal{M}(\omega) \in S_+(\mathbb{R}^n)$ may denote a minor of its Laplacian. We suppose that there exist some *unobserved* true system parameters $\omega^* \in \Omega$ that produce the true matrix $\mathbf{A} = \mathcal{M}(\omega^*)$. We also suppose that there exists a known family of distributions $\mathbb{P}_\omega$ over Ω indexed by $\omega \in \Omega$ that describe the observed randomness in the system if ω were to be the true system parameters. It is this relationship between ω^* and the distribution $\mathbb{P}_{\omega^*}$ that we assume is known as part of the model (but not the true system parameters ω^* themselves). Once this family has been specified, the distribution of $\hat{\mathbf{A}}$ is given by $\mathcal{M}_\#\mathbb{P}_{\omega^*}$, where $\mathcal{M}_\#$ denotes the pushforward. We define $D_{\omega^*} \equiv \mathcal{M}_\#\mathbb{P}_{\omega^*}$.

To frame the full problem, we assume that we are given a single sample $\hat{\omega}$ from $\mathbb{P}_{\omega^*}$ with corresponding matrix $\hat{\mathbf{A}} = \mathcal{M}(\hat{\omega})$ and we would like to use operator shifting to obtain a more accurate estimate of the inverse operator $\mathbf{A} = \mathcal{M}(\omega^*)$. This, of course, necessitates estimating the optimal shift factor,

$$\beta^* = \frac{\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}}}{\mathbb{E}_{\hat{\mathbf{A}}\sim D_{\omega^*}}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}}\,. \tag{43}$$

Naturally, it is not possible for us to estimate this quantity directly with Monte Carlo, as we do not know the true parameters ω^* and hence cannot draw synthetic samples from D_{ω^*}.

However, while D_{ω^*} is unknown, we assume that the family of distributions $\mathbb{P}_\omega$ itself is known—that is, given a ω, we can sample synthetic data from the distribution $\mathbb{P}_\omega$. This means that to approximate the optimal shift factor, we can try to approximate β^* by

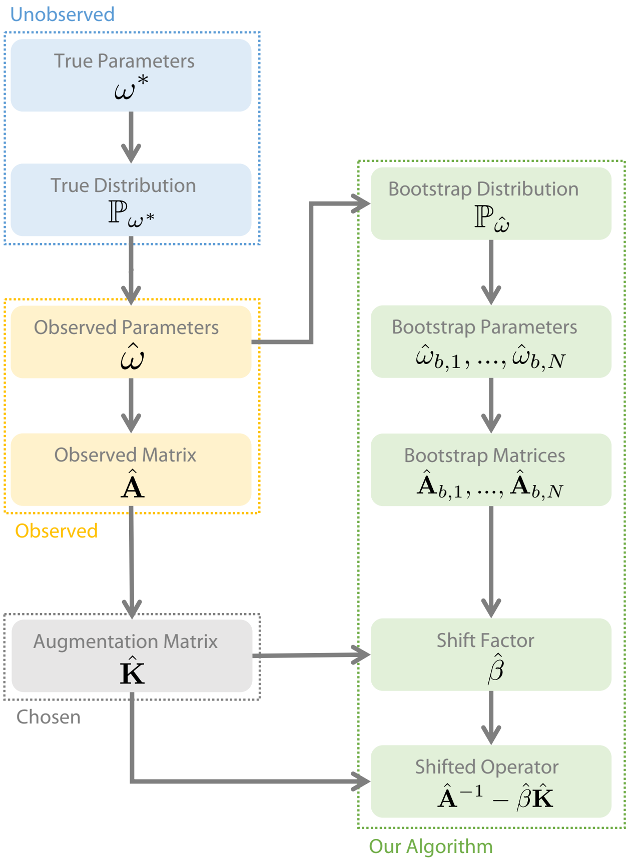
Unobserved
True Parameters ω^*
True Distribution $\mathbb{P}_{\omega^*}$
Observed Parameters $\hat{\omega}$
Observed Matrix $\hat{\mathbf{A}}$
Observed
Augmentation Matrix $\hat{\mathbf{K}}$
Chosen
Bootstrap Distribution $\mathbb{P}_{\hat{\omega}}$
Bootstrap Parameters $\hat{\omega}_{b,1}, \ldots, \hat{\omega}_{b,N}$
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

Fig. 2 A pictographic diagram of the flow of our algorithm. Arrows denote logical dependencies. Quantities in blue are unobserved truth quantities. Quantities in yellow are observed (noisy) data. Quantities in gray are chosen by the user, and quantities in green are computed or bootstrapped via our algorithm

drawing approximate Monte Carlo samples from the approximate distribution $\mathbb{P}_{\hat{\omega}}$. We will give all the details of this algorithm in the next section.

7 Estimating the optimal shift factor

To convert the above into a general algorithm, we need to first do two things. The first is to convert β^* into a form that is more amenable to Monte Carlo evaluation. Obviously, computing the trace of a dim × dim matrix is too expensive in most settings. Therefore, we evaluate traces by using the probabilistic form of the trace, i.e., if $\mathbf{R} \in S_+(\mathbb{R}^n)$, then

$$\mathrm{tr}(\mathbf{R}^{1/2}\mathbf{X}\mathbf{R}^{1/2}) = \mathbb{E}_{\hat{\mathbf{q}}}[\hat{\mathbf{q}}^T \mathbf{X}\hat{\mathbf{q}}], \tag{44}$$

where $\hat{\mathbf{q}}$ is sampled from any distribution with second moment matrix $\mathbf{R}$. We will use the notation that $\langle \cdot, \cdot \rangle_{\mathbf{B}}$ for $\mathbf{B} \in S_+(\mathbb{R}^n)$ denotes the $\mathbf{B}$ vector norm,

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{B}} \equiv \mathbf{x}^T \mathbf{B}\mathbf{y} \, . \tag{45}$$

With Eq. (44), we can evaluate matrix inner products in the $\langle \cdot, \cdot \rangle_{\mathbf{B},\mathbf{R}}$ by using expectations of the corresponding $\langle \cdot, \cdot \rangle_{\mathbf{B}}$ vector norm,

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{B},\mathbf{R}} = \mathbb{E}_{\hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{R})} \langle \mathbf{X}\hat{\mathbf{q}}, \mathbf{Y}\hat{\mathbf{q}} \rangle_{\mathbf{B}} \, . \tag{46}$$

Algorithm 1 Operator Shifting (**GS**)

```
Input: A right-hand side $\mathbf{b}$, an operator sample $\hat{\mathbf{A}} \sim D_{\omega^*}$ with corresponding parameters $\hat{\omega} \in \Omega$, a choice of second moment matrix $\mathbf{R}$, a choice of norm $\mathbf{B}$, sample count $M$.
Output: An estimate $\tilde{\mathbf{x}}$ of $\mathbf{A}^{-1}\mathbf{b}$.

Draw $M$ i.i.d. bootstrap samples $\hat{\mathbf{A}}_{b,1}, \ldots, \hat{\mathbf{A}}_{b,M} \sim D_{\hat{\omega}}$.
Draw $M$ i.i.d. bootstrap samples $\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_M \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.
Assign
    $\hat{\beta}^*(\hat{\mathbf{A}}) = \frac{\sum_{i=1}^{M} \langle \hat{\mathbf{K}}(\hat{\mathbf{A}}_{b,i})\hat{\mathbf{q}}_i, (\hat{\mathbf{A}}_{b,i}^{-1} - \hat{\mathbf{A}}^{-1})\hat{\mathbf{q}}_i \rangle_{\mathbf{B}}}{\sum_{i=1}^{M} \|\hat{\mathbf{K}}(\hat{\mathbf{A}}_{b,i})\hat{\mathbf{q}}_i\|_{\mathbf{B}}^2},$
Assign $\tilde{\mathbf{x}} \leftarrow (\hat{\mathbf{A}}^{-1} - \hat{\beta}^* \mathbf{R}\hat{\mathbf{A}}^{-1}\mathbf{B})\mathbf{b}$
Return $\tilde{\mathbf{x}}$.
```

With this, we can rewrite the expression Eq. (43) as

$$\beta^* = \frac{\mathbb{E}_{\hat{\mathbf{A}} \sim D_{\omega^*}, \hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{R})} \langle \hat{\mathbf{K}}\hat{\mathbf{q}}, (\hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1})\hat{\mathbf{q}} \rangle_{\mathbf{B}}}{\mathbb{E}_{\hat{\mathbf{A}} \sim D_{\omega^*}, \hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{R})} \|\hat{\mathbf{K}}\hat{\mathbf{q}}\|_{\mathbf{B}}^2}, \tag{47}$$

where the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ can always be substituted for any other distribution with the same second moment. We note that the above quantity is impossible to compute outright because we do not know the ground truth $\mathbf{A}$ or the distribution D_{ω^*}. To work around this limitation, we approximate β^* by bootstrapping the above quantity with observed data $\hat{\mathbf{A}}$, and replacing $\mathbf{A}$ with an observed $\hat{\mathbf{A}}$ and the distribution D_{ω^*} with $D_{\hat{\omega}}$. This nets us the approximation

$$\tilde{\beta}^*(\hat{\mathbf{A}}) = \frac{\mathbb{E}_{\hat{\mathbf{A}}_b \sim D_{\hat{\omega}}, \hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{R})} \langle \hat{\mathbf{K}}(\hat{\mathbf{A}}_b)\hat{\mathbf{q}}, (\hat{\mathbf{A}}_b^{-1} - \hat{\mathbf{A}}^{-1})\hat{\mathbf{q}} \rangle_{\mathbf{B}}}{\mathbb{E}_{\hat{\mathbf{A}} \sim D_{\hat{\omega}}, \hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{R})} \|\hat{\mathbf{K}}(\hat{\mathbf{A}}_b)\hat{\mathbf{q}}\|_{\mathbf{B}}^2}, \tag{48}$$

where $\hat{\mathbf{A}}_b$ denotes a bootstrapped sample from the distribution $D_{\hat{\omega}}$. Since bootstrapping tends to work well when estimating scalar quantities, we believe that this approximation step is justified. Now, the above can be estimated with Monte Carlo,

$$\hat{\beta}^*(\hat{\mathbf{A}}) = \frac{\sum_{i=0}^{M} \langle \hat{\mathbf{K}}(\hat{\mathbf{A}}_{b,i})\hat{\mathbf{q}}_i, (\hat{\mathbf{A}}_{b,i}^{-1} - \hat{\mathbf{A}}^{-1})\hat{\mathbf{q}}_i \rangle_{\mathbf{B}}}{\sum_{i=0}^{M} \|\hat{\mathbf{K}}(\hat{\mathbf{A}}_{b,i})\hat{\mathbf{q}}_i\|_{\mathbf{B}}^2}, \tag{49}$$

where

$$\begin{aligned} \hat{\mathbf{A}}_{b,1}, \ldots, \hat{\mathbf{A}}_{b,M} &\sim D_{\hat{\omega}} && \text{i.i.d.,} \\ \hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_M &\sim \mathcal{N}(0, \mathbf{R}) && \text{i.i.d.} \end{aligned} \tag{50}$$

This gives us our general purpose operator shifting algorithm, which we give in full detail in Algorithm 1. Moreover, we give a pictographic representation of the algorithm in Fig. 2.

8 Efficient estimation using truncated expansions

The reader will note that an implementation of operator shifting will involve applying a different M Monte Carlo samples in Eq. (49). Naturally, this can be quite expensive for very large operators. Hence, in this section, we turn to the problem of making Monte Carlo samples more efficient. Fortunately, the energy norm has a number of properties that make it particularly attractive when it comes to efficient computations. In particular, under certain assumptions on the distribution of the randomness in $\hat{\mathbf{A}}$, we will prove that

β can be approximated effectively by using a modified $2k$th order Taylor expansion for $\hat{\mathbf{A}}^{-1}$. This means that one can perform Monte-Carlo computation of β effectively without needing to invert a full linear system for each sample.

We will operate in the framework of Sect. 5, but specialize our discussion to the operator shift given by

$$\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}, \tag{51}$$

Repeating the computation done in the previous two sections, we have that the optimal shift factor is given by

$$\beta^* = \frac{\mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}}}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{R}}}, \tag{52}$$

For brevity of notation, we introduce a shorthand for the expected $\mathbf{R}$-modulated trace,

$$\langle \hat{\mathbf{X}} \rangle_{\mathbf{R}} = \mathbb{E}\,\mathrm{tr}(\mathbf{R}^{1/2}\hat{\mathbf{X}}\mathbf{R}^{1/2}). \tag{53}$$

With this notation, we have:

$$\beta^* = \frac{\langle \hat{\mathbf{A}}^{-1}\mathbf{A}\hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}} - \langle \hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}}}{\langle \hat{\mathbf{A}}^{-1}\mathbf{A}\hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}}}, \tag{54}$$

The properties that make this setting amenable for computation are related to the Taylor series of the numerator and denominator of the above expression. To demonstrate, we can expand the numerator and denominator term using the Taylor expansion of $\hat{\mathbf{A}}^{-1}$ about base-point $\mathbf{A}^{-1}$,

$$\begin{aligned}\hat{\mathbf{A}}^{-1} &\sim \mathbf{A}^{-1} - \mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} - \mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \ldots \\ &= \mathbf{A}^{-1/2}\left[\sum_{k=0}^{\infty}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\right]\mathbf{A}^{-1/2}.\end{aligned} \tag{55}$$

However, note that for this infinite Taylor series to converge, one must restrict the domain of $\hat{\mathbf{A}}$. Just like in the single variable case, the Taylor series only converges absolutely on the event $\{\hat{\mathbf{A}} \prec 2\mathbf{A}\}$. We prove this in a lemma,

Lemma 2 *Let $\hat{\mathbf{X}} \in S_+(\mathbb{R}^n)$ be a random matrix such that $\mathbb{E}[\hat{\mathbf{X}}^{-2}]$ exists and $\hat{\mathbf{X}} \preceq (2-\varepsilon)\mathbf{Y}$ almost surely for $\mathbf{Y} \in S_+(\mathbb{R}^n)$ and $\varepsilon > 0$. Consider the infinite Taylor series for $\hat{\mathbf{X}}^{-1}$ and $\hat{\mathbf{X}}^{-2}$, respectively, about base-point $\mathbf{Y}$, i.e.,*

$$\begin{aligned}\hat{\mathbf{X}}^{-1} &\sim \mathbf{Y}^{-1/2}\left[\sum_{k=0}^{\infty}(-\mathbf{Y}^{-1/2}(\hat{\mathbf{X}}-\mathbf{Y})\mathbf{Y}^{-1/2})^k\right]\mathbf{Y}^{-1/2}, \\ \hat{\mathbf{X}}^{-2} &\sim \mathbf{Y}^{-1/2}\left[\sum_{k=0}^{\infty}(k+1)(-\mathbf{Y}^{-1/2}(\hat{\mathbf{X}}-\mathbf{Y})\mathbf{Y}^{-1/2})^k\right]\mathbf{Y}^{-1/2}.\end{aligned} \tag{56}$$

Both series converge in mean-squared Frobenius norm to their respective limits.

A proof of this fact is relegated to the "Appendix." This places a damper on our ability to use the Taylor expansion of $\hat{\mathbf{A}}^{-1}$ with impunity over all of $S_+(\mathbb{R}^n)$. For simplicity, however, we will assume for now that the true distribution D_{ω^*} is supported on the event $\{\hat{\mathbf{A}} \prec (2-\varepsilon)\mathbf{A}\}$. It turns out, as we will show in Sect. 9, that one can remove this assumption by instead expanding about a variable base-point $\alpha(\hat{\mathbf{A}})\mathbf{A}$ for some large enough factor $\alpha(\hat{\mathbf{A}}) \in \mathbb{R}$.

Therefore, when we have $\operatorname{supp}(D_{\omega^*}) \subset \{\hat{\mathbf{A}} \prec (2-\varepsilon)\mathbf{A}\}$, we can expand $\langle \hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}}$,

$$\begin{aligned}
\langle \hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}} &= \left\langle \mathbf{A}^{-1/2}\left(\sum_{k=0}^{\infty}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\right)\mathbf{A}^{-1/2}\right\rangle_{\mathbf{R}} \\
&= \sum_{k=0}^{\infty}\left\langle \mathbf{A}^{-1/2}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\mathbf{A}^{-1/2}\right\rangle_{\mathbf{R}} \\
&= \sum_{k=0}^{\infty}\left\langle (-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\right\rangle_{\mathbf{A}^{-1/2}\mathbf{R}\mathbf{A}^{-1/2}} \\
&= \sum_{k=0}^{\infty}\langle \hat{\mathbf{X}}^k\rangle_{\mathbf{S}},
\end{aligned} \tag{57}$$

where we have defined

$$\hat{\mathbf{X}} \equiv -\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2}, \qquad \mathbf{S} \equiv \mathbf{A}^{-1/2}\mathbf{R}\mathbf{A}^{-1/2}. \tag{58}$$

Note quickly that the assumption that $\mathbb{E}[\hat{\mathbf{A}}] \preceq \mathbf{A}$ implies $\mathbb{E}[\hat{\mathbf{X}}] \succeq 0$. The assumption that $\mathbf{0} \prec \hat{\mathbf{A}} \prec (2-\varepsilon)\mathbf{A}$ gives us that

$$-(1-\varepsilon)\mathbf{I} \prec \hat{\mathbf{X}} \prec \mathbf{I}. \tag{59}$$

From line one to two in Eq. (57), we may interchange the $\langle\cdot\rangle_{\mathbf{R}}$ operator and the infinite sum by virtue of the fact that $\langle\cdot\rangle_{\mathbf{R}}$ is continuous with respect to the mean squared Frobenius norm,

$$\langle \hat{\mathbf{X}}\rangle_{\mathbf{R}} = \mathbb{E}\operatorname{tr}(\mathbf{R}^{1/2}\hat{\mathbf{X}}\mathbf{R}^{1/2}) = \mathbb{E}\operatorname{tr}(\mathbf{R}\hat{\mathbf{X}}) \le \|\mathbf{R}\|_F\sqrt{\mathbb{E}\|\hat{\mathbf{X}}\|_F^2}, \tag{60}$$

where the inequality above is by Cauchy-Schwarz.

We can similarly expand $\langle \hat{\mathbf{A}}^{-1}\mathbf{A}\hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}}$,

$$\begin{aligned}
&\langle \hat{\mathbf{A}}^{-1}\mathbf{A}\hat{\mathbf{A}}^{-1}\rangle_{\mathbf{R}} \\
&= \left\langle \mathbf{A}^{-1/2}\left(\sum_{k=0}^{\infty}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\right)\left(\sum_{k=0}^{\infty}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\right)\mathbf{A}^{-1/2}\right\rangle_{\mathbf{R}} \\
&= \left\langle \mathbf{A}^{-1/2}\left(\sum_{k=0}^{\infty}(k+1)(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\right)\mathbf{A}^{-1/2}\right\rangle_{\mathbf{R}} \\
&= \sum_{k=0}^{\infty}(k+1)\left\langle \mathbf{A}^{-1/2}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}\mathbf{A}^{-1/2})^k\mathbf{A}^{-1/2}\right\rangle_{\mathbf{R}}, \\
&= \sum_{k=0}^{\infty}(k+1)\langle \hat{\mathbf{X}}^k\rangle_{\mathbf{S}}
\end{aligned} \tag{61}$$

where on lines two to three we have used the property that $\left(\sum_k z^k\right)\left(\sum_k z^k\right) \sim \sum_k (k+1)z^k$. lemma 2 tells us the above series converges in the mean Frobenius norm, and the fact that $\langle\cdot\rangle_{\mathbf{S}}$ is continuous with respect to the expected squared Frobenius norm lets us interchange summation and the $\langle\cdot\rangle_{\mathbf{S}}$ operator.

Thus, plugging everything into Eq. (54), we obtain that

$$\beta^* = \frac{\sum_{k=0}^{\infty} k\,\langle \hat{\mathbf{X}}^k\rangle_{\mathbf{S}}}{\sum_{k=0}^{\infty}(k+1)\,\langle \hat{\mathbf{X}}^k\rangle_{\mathbf{S}}} \tag{62}$$

The form Eq. (62) suggests a possible way of avoiding the need to invert a linear system for every Monte Carlo sample involved in approximating β^*. Instead of attempting to approximate the quantity β^* directly, one can truncate the series in Eq. (62) with an appropriate windowing function to obtain a series of truncated shift factors, defined as

$$\beta_N \equiv \frac{\sum_{k=0}^{\infty} \omega^N(k)\, \langle \hat{\mathbf{X}}^k \rangle_{\mathbf{S}}}{\sum_{k=0}^{\infty} \omega_*^N(k)\, \langle \hat{\mathbf{X}}^k \rangle_{\mathbf{S}}}, \tag{63}$$

where $\omega^N(k) : \mathbb{Z}_{\geq 0} \longrightarrow \mathbb{R}$ and $\omega_*^N(k) : \mathbb{Z}_{\geq 0} \longrightarrow \mathbb{R}$ are two appropriately defined collections of discrete windowing functions, each with *bounded support*, such that the collection has the property that $\omega^N(k) \to k$ and $\omega_*^N(k) \to k+1$ as $N \to \infty$. It turns out, as we will discuss in the next section, that regardless of the randomness structure of the distribution D_{ω^*} (as long as it is bounded), one can choose an appropriate series of windowing functions $\omega^N(k), \omega_*^N(k)$ such that

$$0 \leq \beta_1 \leq \beta_2 \leq ... \leq \beta_N \leq ... \leq \beta^* \leq 1, \tag{64}$$

which means that using any of the truncated shift factors β_N underestimates the value of β^* and hence still decreases the value of the objective $\mathcal{E}((1-\beta)\hat{\mathbf{A}}^{-1})$ from its base value of $\mathcal{E}(\hat{\mathbf{A}}^{-1})$, i.e.,

$$\mathcal{E}(\hat{\mathbf{A}}^{-1}) \geq \mathcal{E}((1-\beta_1)\hat{\mathbf{A}}^{-1}) \geq ... \geq \mathcal{E}((1-\beta_N)\hat{\mathbf{A}}^{-1}) \geq ... \geq \mathcal{E}((1-\beta^*)\hat{\mathbf{A}}^{-1}). \tag{65}$$

Before we continue, note that one can rewrite the truncated shift factors β_N in a form more amenable for computation, namely

$$\beta_N = \frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \omega^N(k)\, \hat{\mathbf{q}}^T \mathbf{A}^{-1} (\hat{\mathbf{Z}}\mathbf{A}^{-1})^k \hat{\mathbf{q}}\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \omega_*^N(k)\, \hat{\mathbf{q}}^T \mathbf{A}^{-1} (\hat{\mathbf{Z}}\mathbf{A}^{-1})^k \hat{\mathbf{q}}\right]}, \tag{66}$$

where $\hat{\mathbf{q}}$ is sampled from a distribution with second moment matrix $\mathbf{R}$ (perhaps $\mathcal{N}(\mathbf{0}, \mathbf{R})$) and is independent from $\hat{\mathbf{A}} \sim D_{\omega^*}$.

8.1 Monotonic estimates of the shift factor

Our analyses of the monotonicity of the β_N rely upon the following lemma,

Lemma 3 *Let $a_1, a_2, ..., a_k, ... \in \mathbb{R}_{\geq 0}$ and $b_1, b_2 ..., b_k, ... \in \mathbb{R}_{\geq 0}$ be two sequences of nonnegative real numbers with $b_1 > 0$, and consider the truncated sum ratios*

$$\beta_N \equiv \frac{\sum_{k=1}^{N} a_k}{\sum_{k=1}^{N} b_k}, \tag{67}$$

then, if it is the case that

$$\frac{a_k}{b_k} \geq \frac{a_{k-1}}{b_{k-1}}, \tag{68}$$

for all k (e.g., the ratios a_k/b_k are monotonically increasing), then the sequence $\beta_1, \beta_2, ..., \beta_k, ...$ is monotonically increasing.

To construct the discrete windowing functions $\omega^N(k), \omega_*^N(k)$, it is instructive to think of the generating polynomials corresponding to $\omega^N(k), \omega_*^N(k)$, i.e.,

$$\Omega^N(x) \equiv \sum_{k=0}^{\infty} \omega^N(k)\, x^k, \qquad \Omega_*^N(x) \equiv \sum_{k=0}^{\infty} \omega_*^N(k)\, x^k. \tag{69}$$

We can rewrite Eq. (63) as

$$\beta_N = \frac{\langle \Omega^N(\hat{\mathbf{X}}) \rangle_{\mathbf{S}}}{\langle \Omega_*^N(\hat{\mathbf{X}}) \rangle_{\mathbf{S}}}. \tag{70}$$

Note that we have used the fact that $\Omega^N(x)$, $\Omega^N_*(x)$ are polynomial generating functions of bounded degrees to interchange summation and expectation.

Our intent now is to find a sequence of polynomials $\Omega^N(x)$, $\Omega^N_*(x)$ with the properties

$$\Omega^N(x) \nearrow \sum_{k=0}^{\infty} k\, x^k\,, \qquad \Omega^N_*(x) \nearrow \sum_{k=0}^{\infty} (k+1)\, x^k\,, \qquad \text{for } |x| < 1, \text{ as } N \to \infty, \tag{71}$$

such that the sequence in Eq. (70) allows us to invoke Lemma 3. We do this by constructing $\Omega^N(x)$, $\Omega^N_*(x)$ from smaller primitive polynomials $\Delta^j(x)$, $\Delta^j_*(x)$ such that

$$\Omega^N(x) = \sum_{j=0}^{N} \Delta^j(x)\,, \qquad \Omega^N_*(x) = \sum_{j=0}^{N} \Delta^j_*(x)\,, \tag{72}$$

$$\mathbb{E}\left[\Delta^j(\hat{\mathbf{X}})\right] \succeq 0\,, \qquad \mathbb{E}\left[\Delta^j_*(\hat{\mathbf{X}})\right] \succ 0\,, \tag{73}$$

$$\Delta^j(x) = \frac{2j-1}{2j}\Delta^j_*(x)\,, \qquad \text{for } j \geq 1\,, \tag{74}$$

$$\Delta^0(x) = 0\,. \tag{75}$$

With this, we can expand Eq. (70) into the required form of Lemma 3,

$$\beta_N = \frac{\sum_{j=0}^N \langle \Delta^j(\hat{\mathbf{X}})\rangle_{\mathbf{S}}}{\sum_{j=0}^N \langle \Delta^j_*(\hat{\mathbf{X}})\rangle_{\mathbf{S}}} \equiv \frac{\sum_{j=0}^N a_j}{\sum_{j=0}^N b_j}\,. \tag{76}$$

Note that property Eq. (73) implies $a_j \geq 0$ and $b_j > 0$, and the property Eq. (74) implies, for $j \geq 1$,

$$\frac{a_j}{b_j} = \frac{\langle \Delta^j(\hat{\mathbf{X}})\rangle_{\mathbf{S}}}{\langle \Delta^j_*(\hat{\mathbf{X}})\rangle_{\mathbf{S}}} = \frac{2j-1}{2j}\frac{\langle \Delta^j_*(\hat{\mathbf{X}})\rangle_{\mathbf{S}}}{\langle \Delta^j_*(\hat{\mathbf{X}})\rangle_{\mathbf{S}}} = \frac{2j-1}{2j}\,, \tag{77}$$

and for $j = 0$, we have $a_0/b_0 = 0$. Hence, the ratio a_j/b_j is monotonically increasing in j and hence satisfies the requirement Eq. (68) of Lemma 3. Therefore, the existence of such primitive polynomials $\Delta^N(x)$, $\Delta^N_*(x)$ immediately implies that

$$\beta_N \to \beta^* \text{ as } N \to \infty\,, \tag{78}$$

$$0 \leq \beta_1 \leq \beta_2 \leq \beta_3 \leq \ldots \leq \beta_N \leq \ldots \leq \beta^* \leq 1\,, \tag{79}$$

where Eq. (78) follows from Eq. (71); the fact that $\beta^* \leq 1$ follows from from $\beta_N \to \beta^*$ and the fact that $a_j \leq b_j$, and hence the numerator of $\sum_{j=0}^N a_j / \sum_{j=0}^N b_j$ is always bounded by the denominator, implying $\beta_N \leq 1$ for all N; and the fact that $\beta_N \geq 0$ for any N comes from nonnegativity of the numerator and denominator of β_N.

To show that such primitive polynomials $\Delta^N(x)$ and $\Delta^N_*(x)$ actually exist, we consider the following definition,

$$\Delta^0(x) \equiv 0\,, \qquad \Delta^0_*(x) \equiv 1\,, \tag{80}$$

$$\Delta^1(x) \equiv x + \frac{1}{2}x^2\,, \qquad \Delta^1_*(x) \equiv 2x + x^2\,, \tag{81}$$

$$\Delta^j(x) \equiv (2j-1)\left(\frac{1}{2}x^{2j-2} + x^{2j-1} + \frac{1}{2}x^{2j}\right)\,, \qquad \text{for } k \geq 2\,, \tag{82}$$

$$\Delta^j_*(x) \equiv 2j\left(\frac{1}{2}x^{2j-2} + x^{2j-1} + \frac{1}{2}x^{2j}\right)\,, \qquad \text{for } k \geq 2\,. \tag{83}$$

To show this family of primitive polynomials satisfies the desired properties, note that, for $j \geq 2$,

$$\Delta^j(x) = \frac{2j-1}{2}x^{2j-2}(x+1)^2 \geq 0\,. \tag{84}$$

This implies $\Delta^j(\hat{\mathbf{X}}) \succeq 0$. Moreover, we can only have $\Delta^j(\hat{\mathbf{X}}) = \mathbf{0}$ if all of the eigenvalues of $\hat{\mathbf{X}}$ are either 0 or -1. Note that a -1 eigenvalue in $\hat{\mathbf{X}}$ is impossible by virtue of the fact that $-\mathbf{I} \prec \hat{\mathbf{X}} \prec \mathbf{I}$. Therefore, $\Delta^j(\hat{\mathbf{X}}) = \mathbf{0}$ is only possible if $\hat{\mathbf{X}} = \mathbf{0}$. However, this cannot be the case almost surely, as $\hat{\mathbf{X}} = \mathbf{0}$ implies $\hat{\mathbf{A}} = \mathbf{A}$. Therefore, with probability greater than 0, we have that $\Delta^j(\hat{\mathbf{X}}) \succ 0$, implying

$$\mathbb{E}[\Delta^j(\hat{\mathbf{X}})] \succ \mathbf{0}, \qquad \mathbb{E}[\Delta^j_*(\hat{\mathbf{X}})] \succ \mathbf{0}. \tag{85}$$

Furthermore, for $j = 1$, we have

$$\mathbb{E}[\Delta^1(\hat{\mathbf{X}})] = \mathbb{E}[\hat{\mathbf{X}}] + \frac{1}{2}\mathbb{E}[\hat{\mathbf{X}}^2] \succ \mathbf{0}, \tag{86}$$

where we have used the fact that $\mathbb{E}[\hat{\mathbf{X}}] \succeq 0$ (from the fact that $\mathbb{E}[\hat{\mathbf{A}}] \preceq \mathbf{A}$) and the fact that $\hat{\mathbf{X}}^2 \succ \mathbf{0}$ with probability greater than 0 (unless $\hat{\mathbf{A}} = \mathbf{A}$ a.s.).

Finally, to show Eq. (71), we simply note that, for k odd, and N large enough, it is the case that

$$\begin{aligned} [x^k]\Omega^N(x) &= [x^k]\Delta^{(k+1)/2}(x) = k, \\ [x^k]\Omega^N_*(x) &= [x^k]\Delta^{(k+1)/2}_*(x) = k+1, \end{aligned} \tag{87}$$

since $\Delta^{(k+1)/2}(x)$ is the only primitive polynomial with a x^k term in $\Omega^N(x)$, and likewise for $\Omega^N_*(x)$. For $k \geq 2$ even, we have that

$$\begin{aligned} [x^k]\Omega^N(x) &= [x^k](\Delta^{k/2}(x) + \Delta^{k/2+1}(x)) = \frac{k-1}{2} + \frac{k+1}{2} = k, \\ [x^k]\Omega^N_*(x) &= [x^k](\Delta^{k/2}_*(x) + \Delta^{k/2+1}_*(x)) = \frac{k}{2} + \frac{k+2}{2} = k+1. \end{aligned} \tag{88}$$

Thus, the polynomials $\Omega^N(x)$ and $\Omega^N_*(x)$ have all the desired properties. We restate the results of the past two sections in a theorem,

Theorem 5 *Under the assumptions in Sect. 3, consider operator shifting with shift $\hat{\mathbf{K}} = \hat{\mathbf{A}}$ in energy norm $\|\cdot\|_{\mathbf{A},\mathbf{R}}$. Suppose that the random matrix $\hat{\mathbf{A}} \in S_+(\mathbb{R}^n)$ satisfies $\mathbf{0} \prec \hat{\mathbf{A}} \prec (2-\varepsilon)\mathbf{A}$ almost surely. Then let β_N be the truncated approximations to the optimal shift factor β^*, i.e.,*

$$\beta_N = \frac{\sum_{k=0}^{2N} \omega^N(k)\langle \hat{\mathbf{X}}^k\rangle_{\mathbf{S}}}{\sum_{k=0}^{2N} \omega^N_*(k)\langle \hat{\mathbf{X}}^k\rangle_{\mathbf{S}}} = \frac{\mathbb{E}\left[\sum_{k=0}^{2N} \omega^N(k)\,\hat{\mathbf{q}}^T\mathbf{A}^{-1}(\hat{\mathbf{Z}}\mathbf{A}^{-1})^k\hat{\mathbf{q}}\right]}{\mathbb{E}\left[\sum_{k=0}^{2N} \omega^N_*(k)\,\hat{\mathbf{q}}^T\mathbf{A}^{-1}(\hat{\mathbf{Z}}\mathbf{A}^{-1})^k\hat{\mathbf{q}}\right]}, \tag{89}$$

where $\omega^N(k)$ and $\omega^N_*(k)$ are given by

$$\omega^N(k) = \begin{cases} k & k < 2N \\ \frac{k-1}{2} & k = 2N \\ 0 & o.w. \end{cases}, \qquad \omega^N_*(k) = \begin{cases} k+1 & k < 2N \\ \frac{k}{2} & k = 2N \\ 0 & o.w. \end{cases}. \tag{90}$$

Under these assumptions, we have that

$$\begin{gathered} \beta_N \nearrow \beta^* \text{ as } N \to \infty, \\ 0 \leq \beta_1 \leq \beta_2 \leq \beta_3 \leq \ldots \leq \beta_N \leq \ldots \leq \beta^* \leq 1, \\ \mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) \geq \mathcal{E}_{\mathbf{A},\mathbf{R}}((1-\beta_1)\hat{\mathbf{A}}^{-1}) \geq \ldots \geq \mathcal{E}_{\mathbf{A},\mathbf{R}}((1-\beta_N)\hat{\mathbf{A}}^{-1}) \geq \ldots \geq \mathcal{E}_{\mathbf{A},\mathbf{R}}((1-\beta^*)\hat{\mathbf{A}}^{-1}), \end{gathered}$$

where $\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{X}})$ denotes the mean squared error of matrix estimator $\hat{\mathbf{X}}$ in the $\|\cdot\|_{\mathbf{A},\mathbf{R}}$-norm.

8.2 Hard windowing

The tradeoff for monotone convergence to the true shift factor β^* is that the windowing functions $\omega^N(k)$ and $\omega_*^N(k)$ presented above—which we will refer to as *soft windowing functions*—may be too conservative at low orders. When this is the case, one may instead choose to use *hard windowing functions* that perform a hard truncation of the infinite Taylor series. That is, one may choose to instead use

$$\omega^N(k) = \begin{cases} k & k \leq 2N \\ 0 & \text{o.w.} \end{cases}, \qquad \omega_*^N(k) = \begin{cases} k+1 & k \leq 2N \\ 0 & \text{o.w.} \end{cases}. \tag{91}$$

Under the conditions of Theorem 5, this choice of windowing function will still guarantee the convergence $\beta_N \to \beta^*$. However, we lose the monotonicity guarantees of the soft windowing functions unless one makes very stringent assumptions on the underlying distribution. That being said, in practice this technique can perform quite well, as indicated in our numerical experiments in Sect. 11. To distinguish between truncated energy norm shifting with soft and hard windows, we will use the abbreviations **ES-T-S** and **ES-T-H** for truncated energy norm augmentation with soft and hard windows, respectively.

8.3 Quick start

To help readers with implementation, we provide explicit formulas for the shift factor β for low truncation orders, as well as a pseudo-code implementation of the different variants of energy norm augmentation.

8.3.1 Explicit formulas for low orders

First, we provide formulas for low orders of the algorithm presented in the previous section. In the subsequent formulas, we let

$$\hat{\mathbf{Z}} \equiv \hat{\mathbf{A}} - \mathbf{A}, \qquad \hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \qquad \hat{\mathbf{A}} \sim D_{\omega^*}, \qquad \hat{\mathbf{q}} \perp\!\!\!\perp \hat{\mathbf{A}}. \tag{92}$$

1. **ES-T-S, Order 2**:

$$\beta_1^{\text{ES-T-S}} = \frac{\mathbb{E}\left[\hat{\mathbf{q}}^T(\frac{1}{2}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1})\hat{\mathbf{q}}\right]}{\mathbb{E}\left[\hat{\mathbf{q}}^T(\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + 2\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1})\hat{\mathbf{q}}\right]}. \tag{93}$$

2. **ES-T-S, Order 2, Mean-Zero Error**:
In many cases, the error matrix $\hat{\mathbf{Z}}$ may be mean zero, i.e., $\mathbb{E}[\hat{\mathbf{Z}}] = \mathbf{0}$. When this happens, the above expression has an even simpler form,

$$\beta_1^{\text{ES-T-S}} = \frac{1}{2}\frac{\mathbb{E}\left[\hat{\mathbf{q}}^T(\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1})\hat{\mathbf{q}}\right]}{\mathbb{E}\left[\hat{\mathbf{q}}^T(\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1})\hat{\mathbf{q}}\right]}. \tag{94}$$

3. **ES-T-H, Order 2**:

$$\beta_1^{\text{ES-T-H}} = \frac{\mathbb{E}\left[\hat{\mathbf{q}}^T(2\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1})\hat{\mathbf{q}}\right]}{\mathbb{E}\left[\hat{\mathbf{q}}^T(3\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + 2\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1})\hat{\mathbf{q}}\right]}. \tag{95}$$

4. **ES-T-H, Order 2, Mean-Zero Error**: In many cases, the error matrix $\hat{\mathbf{Z}}$ may be mean zero, i.e., $\mathbb{E}[\hat{\mathbf{Z}}] = \mathbf{0}$. When this happens, the above expression has an even simpler form,

$$\beta_1^{\text{ES-T-H}} = \frac{\mathbb{E}\left[\hat{\mathbf{q}}^T(2\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1})\hat{\mathbf{q}}\right]}{\mathbb{E}\left[\hat{\mathbf{q}}^T(3\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1}\hat{\mathbf{Z}}\mathbf{A}^{-1} + \mathbf{A}^{-1})\hat{\mathbf{q}}\right]}. \tag{96}$$

8.4 Algorithm

We give the full meta-algorithm for all favors of energy norm augmentation in Algorithm 2. Note that in Algorithm 2, like in Algorithm 1, we replace expectations with bootstrapped Monte Carlo estimators. If one wants to use the simplified expressions provided above in Sect. 8.3.1, one must similarly replace the expectations with sampled and bootstrapped versions. This process is fairly straightforward; for example, for ES-T-H, Order 2, Mean-Zero Error, we get

$$\hat{\beta}_1^{\text{ES-T-H}} = \frac{\sum_{i=0}^{M} \hat{\mathbf{q}}_i^T (2\hat{\mathbf{A}}^{-1}(\hat{\mathbf{A}}_{b,i} - \hat{\mathbf{A}})\hat{\mathbf{A}}^{-1}(\hat{\mathbf{A}}_{b,i} - \hat{\mathbf{A}})\hat{\mathbf{A}}^{-1})\hat{\mathbf{q}}_i}{\sum_{i=0}^{M} \hat{\mathbf{q}}_i^T (3\hat{\mathbf{A}}^{-1}(\hat{\mathbf{A}}_{b,i} - \hat{\mathbf{A}})\hat{\mathbf{A}}^{-1}(\hat{\mathbf{A}}_{b,i} - \hat{\mathbf{A}})\hat{\mathbf{A}}^{-1} + \hat{\mathbf{A}}^{-1})\hat{\mathbf{q}}_i}, \tag{97}$$

where $\hat{\mathbf{A}}_{b,i}$ and $\hat{\mathbf{q}}_i$ are defined as in algorithm 2.

9 Shifted base-point estimation

Obviously, the issue with the above theorem is that the restriction that $\text{supp}(D_{\omega^*}) \subset \{\hat{\mathbf{A}} \prec (2-\varepsilon)\mathbf{A}\}$ is quite restrictive from a problem standpoint; there are many natural problems that do not fall into this setting. Recall that this assumption comes from the fact that the infinite Taylor series for $\hat{\mathbf{A}}^{-1}$ about base-point $\mathbf{A}$ only converges when $\hat{\mathbf{A}} \prec (2-\varepsilon)\mathbf{A}$.

We address this issue with a technique we call *shifted base-point estimation*. The key idea is to grow the region of convergence of the infinite Taylor series by changing the base point of the Taylor series expansion. If we make the assumption that the distribution D_{ω^*} is bounded, then there must exist some $\alpha \geq 1$ such that $\hat{\mathbf{A}} \prec \alpha\mathbf{A}$ for every $\hat{\mathbf{A}}$ in the support of D_{ω^*}. Lemma 2 then tells us that we are justified in taking an infinite Taylor expansion about base-point $\alpha\mathbf{A}$,

$$\hat{\mathbf{A}}^{-1} = \mathbf{A}^{-1/2}\left[\sum_{k=0}^{\infty} \frac{1}{\alpha^{k+1}}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}_\alpha\mathbf{A}^{-1/2})^k\right]\mathbf{A}^{-1/2}. \tag{98}$$

where $\hat{\mathbf{Z}}_\alpha \equiv \hat{\mathbf{A}} - \alpha\mathbf{A}$. In general, the best values of α are those that are as small as possible while maintaining that the support of D_{ω^*} lies within $\{\hat{\mathbf{A}} \prec \alpha\mathbf{A}\}$, as the accuracy of a truncated series becomes less far away from the base point.

With the above, one can repeat the calculations of Sect. 8 practically verbatim to derive the infinite series expression for the optimal shift factor,

$$\beta^* = \frac{\sum_{k=0}^{\infty}(k+1-\alpha)\,\alpha^{-k}\,\langle(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}_\alpha\mathbf{A}^{-1/2})^k\rangle_{\mathbf{S}}}{\sum_{k=0}^{\infty}(k+1)\,\alpha^{-k}\,\langle(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}_\alpha\mathbf{A}^{-1/2})^k\rangle_{\mathbf{S}}}. \tag{99}$$

for notational simplicity, define

$$\hat{\mathbf{X}}_\alpha \equiv \alpha^{-1}(-\mathbf{A}^{-1/2}\hat{\mathbf{Z}}_\alpha\mathbf{A}^{-1/2}). \tag{100}$$

Note that

$$\hat{\mathbf{X}}_\alpha = \mathbf{I} - \alpha^{-1}\mathbf{A}^{-1/2}\hat{\mathbf{A}}\mathbf{A}^{-1/2}. \tag{101}$$

From the fact that $0 \prec \hat{\mathbf{A}} \prec \alpha\mathbf{A}$, it follows that

$$0 \prec \hat{\mathbf{X}}_\alpha \prec \mathbf{I}. \tag{102}$$

Therefore, the expression for the optimal shift factor becomes

$$\beta^* = \frac{\sum_{k=0}^{\infty}(k+1-\alpha)\,\alpha^{-k}\,\langle\hat{\mathbf{X}}_\alpha^k\rangle_{\mathbf{S}}}{\sum_{k=0}^{\infty}(k+1)\,\alpha^{-k}\,\langle\hat{\mathbf{X}}_\alpha^k\rangle_{\mathbf{S}}}. \tag{103}$$

Algorithm 2 Energy-Norm Operator Shfiting Meta-algorithm

```
Input: A right-hand side $\mathbf{b}$, an operator sample $\hat{\mathbf{A}} \sim D_{\omega^*}$ with corresponding parameters $\hat{\omega} \in \Omega$, a choice of second moment matrix $\mathbf{R}$, a choice of matrix $\mathbf{C}$ satisfying the compatibility conditions, sample count $M$.
Output: An estimate $\tilde{\mathbf{x}}$ of $\mathbf{A}^{-1}\mathbf{b}$.
Factorize/preprocess $\hat{\mathbf{A}}$ to precompute $\hat{\mathbf{A}}^{-1}$ if necessary.
Draw $M$ i.i.d. bootstrap samples $\hat{\mathbf{A}}_{b,1}, ..., \hat{\mathbf{A}}_{b,M} \sim D_{\hat{\omega}}$.
Draw $M$ i.i.d. bootstrap samples $\hat{\mathbf{q}}_1, ..., \hat{\mathbf{q}}_M \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.
if using Truncated Energy-Norm Shifting (ES-T) then
    if using Soft Truncation (ES-T-S) then
        Let
            $\omega^N(k) = \begin{cases} k & k < 2N \\ \frac{k-1}{2} & k = 2N \\ 0 & \text{o.w.} \end{cases}, \qquad \omega_*^N(k) = \begin{cases} k+1 & k < 2N \\ \frac{k}{2} & k = 2N \\ 0 & \text{o.w.} \end{cases}.$
    else if using Hard Truncation (ES-T-H) then
        Let
            $\omega^N(k) = \begin{cases} k & k \leq 2N \\ 0 & \text{o.w.} \end{cases}, \qquad \omega_*^N(k) = \begin{cases} k+1 & k \leq 2N \\ 0 & \text{o.w.} \end{cases}.$
    end if
    Assign
        $\hat{\beta}^* \leftarrow \frac{\sum_{i=0}^M \sum_{k=0}^\infty \omega^N(k)\, \hat{\mathbf{q}}_i^T \mathbf{C}^T \hat{\mathbf{A}}^{-1}((\hat{\mathbf{A}}_{b,i} - \hat{\mathbf{A}})\hat{\mathbf{A}}^{-1})^k \hat{\mathbf{q}}_i}{\sum_{i=0}^M \sum_{k=0}^\infty \omega_*^N(k)\, \hat{\mathbf{q}}_i^T \mathbf{C}^T \hat{\mathbf{A}}^{-1}((\hat{\mathbf{A}}_{b,i} - \hat{\mathbf{A}})\hat{\mathbf{A}}^{-1})^k \mathbf{C}\hat{\mathbf{q}}_i},$
    where
else if using Untruncated Energy-Norm Shifting (ES) then
    Assign
        $\hat{\beta}^* \leftarrow \frac{\sum_{i=0}^M \hat{\mathbf{q}}_i^T \mathbf{C}^T (\hat{\mathbf{A}}_{b,i}^{-1} \hat{\mathbf{A}} \hat{\mathbf{A}}_{b,i}^{-1} - \hat{\mathbf{A}}_{b,i}^{-1})\hat{\mathbf{q}}_i}{\sum_{i=0}^M \hat{\mathbf{q}}_i^T \mathbf{C}^T (\hat{\mathbf{A}}_{b,i}^{-1} \hat{\mathbf{A}} \hat{\mathbf{A}}_{b,i}^{-1})\mathbf{C}\hat{\mathbf{q}}_i},$
end if
Clamp $\hat{\beta}^* \leftarrow \max(0, \hat{\beta}^*)$.
Assign $\tilde{\mathbf{x}} \leftarrow (\hat{\mathbf{A}}^{-1} - \hat{\beta}^* \hat{\mathbf{A}}^{-1}\mathbf{C})\mathbf{b}$
Return $\tilde{\mathbf{x}}$.
```

From here, we follow the same schema to define the truncation of the infinite series above,

$$\beta_N = \frac{\sum_{k=0}^\infty \omega_\alpha^N(k)\, \alpha^{-k}\, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}}}{\sum_{k=0}^\infty \omega_{\alpha,*}^N(k)\, \alpha^{-k}\, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}}}. \tag{104}$$

and the form we will use for Monte Carlo,

$$\beta_N = \frac{\mathbb{E}\left[\sum_{k=0}^\infty \omega_\alpha^N(k)\, \alpha^{-k}\, \hat{\mathbf{q}}^T \mathbf{A}^{-1}(\hat{\mathbf{Z}}_\alpha \mathbf{A}^{-1})^k \hat{\mathbf{q}}\right]}{\mathbb{E}\left[\sum_{k=0}^\infty \omega_{\alpha,*}^N(k)\, \alpha^{-k}\, \hat{\mathbf{q}}^T \mathbf{A}^{-1}(\hat{\mathbf{Z}}_\alpha \mathbf{A}^{-1})^k \hat{\mathbf{q}}\right]}, \tag{105}$$

where $\omega_\alpha^N(k)$ and $\omega_{\alpha,*}^N(k)$ are new window functions that converge to $k + 1 - \alpha$ and $k + 1$, respectively. To show that this has the same properties as the truncated shift factors in the previous section, we simply repeat the proof from the previous section, but with a few small changes.

First, we repeat the previous section to obtain the polynomial expression for β_N,

$$\beta_N = \frac{\langle \Omega_\alpha^N(\hat{\mathbf{X}}_\alpha)\rangle_{\mathbf{S}}}{\langle \Omega_{\alpha,*}^N(\hat{\mathbf{X}}_\alpha)\rangle_{\mathbf{S}}}, \tag{106}$$

where, once again

$$\Omega_\alpha^N(x) = \sum_{j=0}^{N} \Delta_\alpha^j(x), \qquad \Omega_{\alpha,*}^N(x) = \sum_{j=0}^{N} \Delta_{\alpha,*}^j(x), \tag{107}$$

For brevity of notation, we define the quantity,

$$\eta \equiv 1 - \alpha^{-1}. \tag{108}$$

Now, our monotonicity analysis in this section is based upon the observation that for $x \geq 0$, it is the case that

$$x^j(x-\eta) \geq \eta^j(x-\eta), \tag{109}$$

and therefore, it is also the case that, for $x \geq 0$,

$$x^j(x^k - \eta^k) = x^j(x^{k-1} + \eta x^{k-2} + \ldots + \eta^{k-2}x + \eta^{k-1})(x-\eta) \geq k\eta^{j+k-1}(x-\eta). \tag{110}$$

Whereas the analysis in the previous section built a monotonic sequence of polynomials that were positive everywhere, the above formula allows us to build a monotonic sequence of polynomials that are positive in expectation, but not necessarily positive everywhere. To do this, we first note that by our $\mathbb{E}[\hat{\mathbf{A}}] \preceq \mathbf{A}$ assumption,

$$\mathbb{E}[\hat{\mathbf{X}}_\alpha] = \mathbb{E}[\mathbf{I} - \alpha^{-1}\mathbf{A}^{-1/2}\hat{\mathbf{A}}\mathbf{A}^{-1/2}] \succeq \mathbf{I} - \alpha^{-1}\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2} = \eta\mathbf{I}. \tag{111}$$

Hence, the above polynomial inequalities imply that

$$\mathbb{E}[\hat{\mathbf{X}}_\alpha^j(\hat{\mathbf{X}}_\alpha^k - \eta^k\mathbf{I})] \succeq \mathbb{E}[k\eta^{j+k-1}(\hat{\mathbf{X}}_\alpha - \eta\mathbf{I})] \succeq \mathbf{0}. \tag{112}$$

This allows us to use the matrix polynomials $\hat{\mathbf{X}}_\alpha^j(\hat{\mathbf{X}}_\alpha^k - \eta^k\mathbf{I})$ as building blocks for a series that converges monotonically to the desired β^*. The final observation that one needs to build the series is the fact that

$$\sum_{k=0}^{\infty} \eta^k = \frac{1}{1-\eta} = \alpha. \tag{113}$$

With this established, we finally define the primitive polynomials

$$\Delta_\alpha^k(x) \equiv kx^k - \eta x^{k-1} - \eta^2 x^{k-2} - \ldots - \eta^{k-1}x - \eta^k = (k+1)x^k - \sum_{j=0}^{k} \eta^j x^{k-j}. \tag{114}$$

By Eq. (112), we have that

$$\mathbb{E}[\Delta_\alpha^k(\hat{\mathbf{X}}_\alpha)] = \mathbb{E}\left[(k+1)\hat{\mathbf{X}}_\alpha^k - \sum_{j=0}^{k} \eta^j \hat{\mathbf{X}}_\alpha^{k-j}\right] = \sum_{j=0}^{k} \mathbb{E}[\hat{\mathbf{X}}_\alpha^{k-j}(\hat{\mathbf{X}}_\alpha^j - \eta^j)] \succeq \mathbf{0}. \tag{115}$$

However, if one examines the individual terms of the composite sum $\sum_{k=0}^{\infty} \Delta_\alpha^k(x)$ by powers of x, one observes that, for $x \in [0, 1)$,

$$[x^j] \sum_{k=0}^{\infty} \Delta_\alpha^k(x) = j - \sum_{k=1}^{\infty} \eta^k = j + 1 - \alpha. \tag{116}$$

Ergo, for $x \in [0, 1)$, we have that

$$\Omega_\alpha^N(x) = \sum_{k=0}^{N} \Delta_\alpha^k(x) \nearrow \sum_{k=0}^{\infty} (k + 1 - \alpha) x^k . \tag{117}$$

And therefore, if we also take

$$\Delta_{\alpha,*}^k(x) = (k+1)x^k , \qquad \Omega_{\alpha,*}^N(x) = \sum_{k=0}^{N} \Delta_{\alpha,*}^k(x) = \sum_{k=0}^{N} (k+1) x^k , \tag{118}$$

and note that

$$\beta^* = \frac{\sum_{k=0}^{\infty} (k + 1 - \alpha) \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}}}{\sum_{k=0}^{\infty} (k+1) \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}}} , \tag{119}$$

we can conclude that

$$\beta_N = \frac{\langle \Omega_\alpha^N(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}}}{\langle \Omega_{\alpha,*}^N(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}}} \to \beta^* . \tag{120}$$

and from the fact that the numerator is a sum of positive terms, and the fact that $\Omega_\alpha^N(\hat{\mathbf{X}}_\alpha) \preceq \Omega_{\alpha,*}^N(\hat{\mathbf{X}}_\alpha)$ by construction, it therefore follows that

$$0 \leq \beta_N \leq 1 , \qquad 0 \leq \beta^* \leq 1 . \tag{121}$$

To achieve a proof of monotonicity of the β_N, we appeal to Lemma 3, which necessitates that we verify the inequality

$$a_k b_{k-1} = \langle \Delta_\alpha^k(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} \, \langle \Delta_{\alpha,*}^{k-1}(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} \geq \langle \Delta_\alpha^{k-1}(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} \, \langle \Delta_{\alpha,*}^k(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} = a_{k-1} b_k . \tag{122}$$

To do this, let us subtract and expand the above terms

$$\begin{aligned}
a_k b_{k-1} - a_{k-1} b_k &= \langle \Delta_\alpha^k(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} \, \langle \Delta_{\alpha,*}^{k-1}(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} - \langle \Delta_\alpha^{k-1}(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} \, \langle \Delta_{\alpha,*}^k(\hat{\mathbf{X}}_\alpha) \rangle_{\mathbf{S}} \\
&= k \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \left((k+1) \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} - \sum_{j=0}^{k} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^{k-j} \rangle_{\mathbf{S}} \right) \\
&\quad - (k+1) \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \left(k \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} - \sum_{j=0}^{k-1} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} \right) \\
&= (k+1) \sum_{j=0}^{k-1} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} - k \sum_{j=0}^{k} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \, \langle \hat{\mathbf{X}}_\alpha^{k-j} \rangle_{\mathbf{S}} \\
&= k \sum_{j=0}^{k-1} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} - k \sum_{j=0}^{k-1} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \, \langle \hat{\mathbf{X}}_\alpha^{k-j} \rangle_{\mathbf{S}} \\
&\quad + \sum_{j=0}^{k-1} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \, \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} - k \, \eta^k \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \, \langle \mathbf{I} \rangle_{\mathbf{S}} .
\end{aligned} \tag{123}$$

We now appeal to the following lemma, which allows us to compare terms across the two sums above,

Lemma 4 *Let $\hat{\mathbf{X}}$ be a random matrix such that $\hat{\mathbf{X}} \succeq \mathbf{0}$ a.s. For $i \geq j$ and $r \geq 0$, and any symmetric positive semi-definite matrix $\mathbf{S} \succeq \mathbf{0}$, we have that*

$$\langle \hat{\mathbf{X}}^{i+r} \rangle_{\mathbf{S}} \, \langle \hat{\mathbf{X}}^{j-r} \rangle_{\mathbf{S}} \geq \langle \hat{\mathbf{X}}^i \rangle_{\mathbf{S}} \, \langle \hat{\mathbf{X}}^j \rangle_{\mathbf{S}} . \tag{124}$$

The proof of this fact is relegated to the "Appendix." However, applying this lemma to the above Eq. (123) gives us

$$\begin{aligned} a_k b_{k-1} - a_{k-1} b_k &\geq \sum_{j=0}^{k-1} \eta^j \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \, \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} - k \, \eta^k \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \langle \mathbf{I} \rangle_{\mathbf{S}} \\ &= \sum_{j=0}^{k-1} \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \left(\eta^j \, \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} \right) - \sum_{j=0}^{k-1} \left(\eta \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \right) \left(\eta^{k-1} \langle \mathbf{I} \rangle_{\mathbf{S}} \right) . \end{aligned} \tag{125}$$

Finally, we note that Eq. (112) gives us $\eta^j \mathbb{E}[\hat{\mathbf{X}}_\alpha^{k-j-1}] \succeq \eta^{k-1} \mathbf{I}$ and therefore

$$\eta^j \, \langle \hat{\mathbf{X}}_\alpha^{k-j-1} \rangle_{\mathbf{S}} \geq \eta^{k-1} \langle \mathbf{I} \rangle_{\mathbf{S}} \, . \tag{126}$$

Moreover, Eq. (112) also gives us that $\mathbb{E}[\hat{\mathbf{X}}_\alpha^k] \succeq \eta \mathbb{E}[\hat{\mathbf{X}}_\alpha^{k-1}]$ and therefore

$$\langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}} \geq \eta \, \langle \hat{\mathbf{X}}_\alpha^{k-1} \rangle_{\mathbf{S}} \tag{127}$$

Noting the above two inequalities are between positive numbers and then substituting the above two inequalities into Eq. (112) gives the desired result

$$a_k b_{k-1} - a_{k-1} b_k \geq 0 \, . \tag{128}$$

Thus, the truncated estimators β_N form a positive monotonic sequence that converges to β^*. To summarize, we restate the results we have just proved into a theorem,

Theorem 6 *Under the assumptions in Sect. 3, consider operator shifting with shift* $\hat{\mathbf{K}} = \hat{\mathbf{A}}$ *in energy norm* $\| \cdot \|_{\mathbf{A},\mathbf{R}}$. *Suppose that the random matrix* $\hat{\mathbf{A}} \in S_+(\mathbb{R}^n)$ *satisfies* $\mathbf{0} \prec \hat{\mathbf{A}} \prec \alpha \mathbf{A}$ *almost surely. Then let* β_N *be the truncated approximations to the optimal shift factor* β^*, *i.e.,*

$$\beta_N = \frac{\sum_{k=0}^N \omega_\alpha^N(k) \, \alpha^{-k} \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}}}{\sum_{k=0}^N \omega_{\alpha,*}^N(k) \, \alpha^{-k} \, \langle \hat{\mathbf{X}}_\alpha^k \rangle_{\mathbf{S}}} = \frac{\mathbb{E}\left[\sum_{k=0}^N \omega_\alpha^N(k) \, \alpha^{-k} \, \hat{\mathbf{q}}^T \mathbf{A}^{-1} (\hat{\mathbf{Z}}_\alpha \mathbf{A}^{-1})^k \hat{\mathbf{q}} \right]}{\mathbb{E}\left[\sum_{k=0}^N \omega_{\alpha,*}^N(k) \, \alpha^{-k} \, \hat{\mathbf{q}}^T \mathbf{A}^{-1} (\hat{\mathbf{Z}}_\alpha \mathbf{A}^{-1})^k \hat{\mathbf{q}} \right]} , \tag{129}$$

where $\omega_\alpha^N(k)$ *and* $\omega_{\alpha,*}^N(k)$ *are given by*

$$\omega_\alpha^N(k) = \begin{cases} (k+1) - \sum_{j=k}^N \eta^{j-k} & k \leq N \\ 0 & o.w. \end{cases} , \qquad \omega_{\alpha,*}^N(k) = \begin{cases} k+1 & k \leq N \\ 0 & o.w. \end{cases} , \tag{130}$$

and $\eta = 1 - \alpha^{-1}$. *Under these assumptions, we have that*

$$\begin{gathered} \beta_N \nearrow \beta^* \text{ as } N \to \infty , \\ 0 \leq \beta_1 \leq \beta_2 \leq \beta_3 \leq \ldots \leq \beta_N \leq \ldots \leq \beta^* \leq 1 , \\ \mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) \geq \mathcal{E}_{\mathbf{A},\mathbf{R}}((1-\beta_1)\hat{\mathbf{A}}^{-1}) \geq \ldots \geq \mathcal{E}_{\mathbf{A},\mathbf{R}}((1-\beta_N)\hat{\mathbf{A}}^{-1}) \geq \ldots \geq \mathcal{E}_{\mathbf{A},\mathbf{R}}((1-\beta^*)\hat{\mathbf{A}}^{-1}) , \end{gathered}$$

where $\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{X}})$ *denotes the mean squared error of matrix estimator* $\hat{\mathbf{X}}$ *in the* $\| \cdot \|_{\mathbf{A},\mathbf{R}}$*-norm.*

10 Accelerating shifted base-point estimation

In practice, while the formula Eq. (129) provides a positive, monotonically increasing series of estimates β_N for the optimal β^* which only use N powers of the matrix $\hat{\mathbf{Z}}_\alpha \mathbf{A}^{-1}$, note that the larger one takes the factor α, the poorer the accuracy of the truncated approximation near the matrix $\mathbf{A}$, where most of the probability distribution is concentrated. Therefore, while we get a guarantee of an estimate that will decrease the value of the objective $\mathcal{E}_{\mathbf{A}}^{\text{Bayes}}(\cdot)$, the convergence to the optimal factor β^* might be very slow as a result, necessitating larger and larger powers of $\hat{\mathbf{Z}}_\alpha \mathbf{A}^{-1}$. Thus, in practice, it is often a good idea to let the quantity α be a function of the sample $\hat{\mathbf{A}}$ such that $\hat{\mathbf{A}} \preceq \alpha(\hat{\mathbf{A}})\mathbf{A}$. This means that, instead of using the estimator β_N in Eq. (129) above, we use

$$\bar{\beta}_N = \frac{\mathbb{E}\left[\alpha(\hat{\mathbf{A}})^{-2} \sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}})}(k)\, \alpha(\hat{\mathbf{A}})^{-k}\, \hat{\mathbf{q}}^T \mathbf{A}^{-1} (\hat{\mathbf{Z}}_{\alpha(\hat{\mathbf{A}})} \mathbf{A}^{-1})^k \hat{\mathbf{q}}\right]}{\mathbb{E}\left[\alpha(\hat{\mathbf{A}})^{-2} \sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}),*}(k)\, \alpha(\hat{\mathbf{A}})^{-k}\, \hat{\mathbf{q}}^T \mathbf{A}^{-1} (\hat{\mathbf{Z}}_{\alpha(\hat{\mathbf{A}})} \mathbf{A}^{-1})^k \hat{\mathbf{q}}\right]}, \tag{131}$$

where one choice of $\alpha(\hat{\mathbf{A}})$ is

$$\alpha(\hat{\mathbf{A}}) = \|\mathbf{A}^{-1/2} \hat{\mathbf{A}} \mathbf{A}^{-1/2}\|_2, \tag{132}$$

i.e., the smallest value for which $\hat{\mathbf{A}} \preceq \alpha(\hat{\mathbf{A}})\mathbf{A}$, and the windowing functions $\omega^N_\alpha(k)$ and $\omega^N_{\alpha,*}(k)$ are defined as in Eq. (130). In practice, one may choose to approximate $\alpha(\hat{\mathbf{A}})$ instead of computing it exactly. Note in Eq. (131) the reintroduction of the $\alpha(\hat{\mathbf{A}})^{-2}$ terms in the numerator and denominator; originally, these terms passed out of the expectation and canceled, but now the explicit dependence on $\hat{\mathbf{A}}$ prevents this cancelation from happening.

Computing $\|\mathbf{A}^{-1/2} \hat{\mathbf{A}} \mathbf{A}^{-1/2}\|_2$ can be done with power method. In particular, with probability 1, if $\hat{\mathbf{v}} \in \mathbb{R}^n$ is sampled from a distribution continuous with respect to the Lebesgue measure on its support, we have that

$$\begin{aligned} \alpha(\hat{\mathbf{A}}) &= \lim_{k \to \infty} \frac{\|(\mathbf{A}^{-1/2} \hat{\mathbf{A}} \mathbf{A}^{-1/2})^k \hat{\mathbf{v}}\|_2}{\|(\mathbf{A}^{-1/2} \hat{\mathbf{A}} \mathbf{A}^{-1/2})^{k-1} \hat{\mathbf{v}}\|_2} \\ &= \lim_{k \to \infty} \sqrt{\frac{\hat{\mathbf{v}}^T \mathbf{A}^{1/2} (\mathbf{A}^{-1}\hat{\mathbf{A}})^k \mathbf{A}^{-1} (\hat{\mathbf{A}}\mathbf{A}^{-1})^k \mathbf{A}^{1/2} \hat{\mathbf{v}}}{\hat{\mathbf{v}}^T \mathbf{A}^{1/2} (\mathbf{A}^{-1}\hat{\mathbf{A}})^{k-1} \mathbf{A}^{-1} (\hat{\mathbf{A}}\mathbf{A}^{-1})^{k-1} \mathbf{A}^{1/2} \hat{\mathbf{v}}}}. \end{aligned} \tag{133}$$

Since $\mathbf{A}$ is non-singular, transforming the random variable $\hat{\mathbf{v}}$ by $\mathbf{A}^{1/2}$ transforms the corresponding distribution into a distribution continuous with respect to the Lebesgue measure on its support. Therefore, it is sufficient to compute/approximate

$$\alpha(\hat{\mathbf{A}}) = \lim_{k \to \infty} \sqrt{\frac{\hat{\mathbf{v}}^T (\mathbf{A}^{-1}\hat{\mathbf{A}})^k \mathbf{A}^{-1} (\hat{\mathbf{A}}\mathbf{A}^{-1})^k \hat{\mathbf{v}}}{\hat{\mathbf{v}}^T (\mathbf{A}^{-1}\hat{\mathbf{A}})^{k-1} \mathbf{A}^{-1} (\hat{\mathbf{A}}\mathbf{A}^{-1})^{k-1} \hat{\mathbf{v}}}} = \lim_{k \to \infty} \frac{\|(\hat{\mathbf{A}}\mathbf{A}^{-1})^k \hat{\mathbf{v}}\|_{\mathbf{A}^{-1}}}{\|(\hat{\mathbf{A}}\mathbf{A}^{-1})^{k-1} \hat{\mathbf{v}}\|_{\mathbf{A}^{-1}}}, \tag{134}$$

and as a result, we do not actually need to know $\mathbf{A}^{1/2}$ or $\mathbf{A}^{-1/2}$ to be able to compute the correct value of α.

Now, to produce an algorithm, we follow the template of Sect. 7—we bootstrap $\mathbf{A}$ by replacing it with our sampled $\hat{\mathbf{A}}$ and bootstrap the expectation by using the distribution $D_{\hat{\omega}}$ instead of the true distribution D_{ω^*}. This nets us the approximate estimator

$$\tilde{\beta}_N(\hat{\mathbf{A}}) = \frac{\mathbb{E}_{\hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{L}), \hat{\mathbf{A}}_b \sim D_{\hat{\omega}}}\left[\sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}_b)}(k)\, \alpha(\hat{\mathbf{A}}_b)^{-k-2}\, \hat{\mathbf{q}}^T \hat{\mathbf{A}}^{-1} (\hat{\mathbf{Z}}_{b,\alpha(\hat{\mathbf{A}}_b)} \hat{\mathbf{A}}^{-1})^k \hat{\mathbf{q}}\right]}{\mathbb{E}_{\hat{\mathbf{q}} \sim \mathcal{N}(\mathbf{0},\mathbf{L}), \hat{\mathbf{A}}_b \sim D_{\hat{\omega}}}\left[\sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}_b),*}(k)\, \alpha(\hat{\mathbf{A}}_b)^{-k-2}\, \hat{\mathbf{q}}^T \hat{\mathbf{A}}^{-1} (\hat{\mathbf{Z}}_{b,\alpha(\hat{\mathbf{A}}_b)} \hat{\mathbf{A}}^{-1})^k \hat{\mathbf{q}}\right]}, \tag{135}$$

where $\hat{\mathbf{Z}}_{b,\alpha(\hat{\mathbf{A}}_b)} = \hat{\mathbf{A}} - \alpha(\hat{\mathbf{A}}_b)\hat{\mathbf{A}}_b$. The above quantity can be estimated by Monte Carlo by computing

$$\hat{\beta}_N(\hat{\mathbf{A}}) = \frac{\sum_{i=0}^{M} \sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}_{b,i})}(k)\, \alpha(\hat{\mathbf{A}}_{b,i})^{-k-2}\, \hat{\mathbf{q}}_i^T \hat{\mathbf{A}}^{-1} (\hat{\mathbf{Z}}_{b,\alpha(\hat{\mathbf{A}}_{b,i})} \hat{\mathbf{A}}^{-1})^k \hat{\mathbf{q}}_i}{\sum_{i=0}^{M} \sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}_{b,i}),*}(k)\, \alpha(\hat{\mathbf{A}}_{b,i})^{-k-2}\, \hat{\mathbf{q}}_i^T \hat{\mathbf{A}}^{-1} (\hat{\mathbf{Z}}_{b,\alpha(\hat{\mathbf{A}}_{b,i})} \hat{\mathbf{A}}^{-1})^k \hat{\mathbf{q}}_i}, \tag{136}$$

where

$$\begin{aligned} \hat{\mathbf{A}}_{b,1}, ..., \hat{\mathbf{A}}_{b,M} &\sim D_{\hat{\omega}} \qquad && \text{i.i.d.,} \\ \hat{\mathbf{q}}_1, ..., \hat{\mathbf{q}}_M &\sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \qquad && \text{i.i.d.} \end{aligned} \tag{137}$$

The full algorithm is presented in algorithm 3.

Unfortunately, we do not believe that the monotonic guarantees of the previous two sections carry over when acceleration is applied. While it is not difficult to prove that the terms underneath the expectations in Eq. (131) become more accurate point-wise in $\hat{\mathbf{A}}$ (as we are shifting the base point of the Taylor expansion closer to the point we are evaluating), it may be possible to construct contrived examples where this produces less accurate estimates of the expectations. However, we strongly believe that in almost all practical use cases, one should expect a significant improvement in accuracy in using this technique, as the reduction in truncation error is extremely substantial.

Algorithm 3 Accel. Shifted Truncated En.-Norm Augmentation (**ES-TRA**)

Input: A right-hand side $\mathbf{b}$, an operator sample $\hat{\mathbf{A}} \sim D_{\omega^*}$ with corresponding parameters $\hat{\omega} \in \Omega$, a choice of second moment matrix $\mathbf{R}$, a choice of matrix $\mathbf{C}$ satisfying the compatibility conditions, sample count M.

Output: An estimate $\tilde{\mathbf{x}}$ of $\mathbf{A}^{-1}\mathbf{b}$.

```
Factorize/preprocess $\hat{\mathbf{A}}$ to precompute $\hat{\mathbf{A}}^{-1}$ if necessary.
Draw $M$ i.i.d. bootstrap samples $\hat{\mathbf{A}}_{b,1}, ..., \hat{\mathbf{A}}_{b,M} \sim D_{\hat{\omega}}$.
For each $\hat{\mathbf{A}}_{b,i}$, perform power method to assign
    $\alpha(\hat{\mathbf{A}}_{b,i}) \leftarrow \lim_{k\to\infty} \frac{\|(\hat{\mathbf{A}}_{b,i}\hat{\mathbf{A}}^{-1})^k \hat{\mathbf{v}}\|_{\hat{\mathbf{A}}^{-1}}}{\|(\hat{\mathbf{A}}_{b,i}\hat{\mathbf{A}}^{-1})^{k-1} \hat{\mathbf{v}}\|_{\hat{\mathbf{A}}^{-1}}}.$
Draw $M$ i.i.d. bootstrap samples $\hat{\mathbf{q}}_1, ..., \hat{\mathbf{q}}_M \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.
Assign
    $\hat{\beta}^* \leftarrow \frac{\sum_{i=0}^{M} \sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}_{b,i})}(k)\, \alpha(\hat{\mathbf{A}}_{b,i})^{-k-2}\, \hat{\mathbf{q}}_i^T \mathbf{C}^T \hat{\mathbf{A}}^{-1}(\mathbf{I} - \alpha(\hat{\mathbf{A}}_{b,i})\hat{\mathbf{A}}_{b,i}\hat{\mathbf{A}}^{-1})^k \hat{\mathbf{q}}_i}{\sum_{i=0}^{M} \sum_{k=0}^{\infty} \omega^N_{\alpha(\hat{\mathbf{A}}_{b,i}),*}(k)\, \alpha(\hat{\mathbf{A}}_{b,i})^{-k-2}\, \hat{\mathbf{q}}_i^T \mathbf{C}^T \hat{\mathbf{A}}^{-1}(\mathbf{I} - \alpha(\hat{\mathbf{A}}_{b,i})\hat{\mathbf{A}}_{b,i}\hat{\mathbf{A}}^{-1})^k \mathbf{C}\hat{\mathbf{q}}_i},$
    where
    $\omega^N_\alpha(k) = \begin{cases} (k+1) - \sum_{j=k}^{N}(1-\alpha^{-1})^{j-k} & k \le N \\ 0 & \text{o.w.} \end{cases}, \quad \omega^N_{\alpha,*}(k) = \begin{cases} k+1 & k \le N \\ 0 & \text{o.w.} \end{cases},$
Clamp $\hat{\beta}^* \leftarrow \max(0, \hat{\beta}^*)$.
Assign $\tilde{\mathbf{x}} \leftarrow (\hat{\mathbf{A}}^{-1} - \hat{\beta}^* \hat{\mathbf{A}}^{-1}\mathbf{C})\mathbf{b}$
Return $\tilde{\mathbf{x}}$.
```

11 Numerical experiments

In this section, we present numerical experiments to benchmark the above methods. We compare a number of different variations of operator shifting:

1. **Naive**: Naive solve of the system $\hat{\mathbf{A}}\hat{\mathbf{x}} = \mathbf{b}$, by inverting the system directly without modifying the operator $\hat{\mathbf{A}}$.
2. **GS** *(General Operator Shifting)*: The method presented in Sect. 4 and algorithm 1, where we take $\mathbf{R} = \mathbf{B} = \mathbf{I}$ and let the prior on $\mathbf{b}$ be the standard normal distribution.

3. **ES** *(Energy-Norm Operator Shifting)*: The method presented in Sect. 8 without any truncation (i.e., computing the shift factor β^* directly using bootstrap and Monte-Carlo), where we take $\mathbf{R} = \mathbf{I}$ and let the distribution of $\mathbf{b}$ be the standard normal distribution.
4. **ES-T** *(Truncated Energy Operator Shifting)*: The method presented in Sect. 8.1. In the numerical results, we test different orders of truncation. The order here denotes the highest power of a bootstrapped matrix sample which appears in the computation for the approximate shift factor. Furthermore, we will also test both **soft** (ES-T-S) and **hard** (ES-T-H) truncation windows, as discussed in Sect. 8.2.
5. **ES-TRA** *(Truncated Rebased Accelerated Energy Operator Shifting)*: The method presented in Sect. 10 and algorithm 3. The order of truncation denotes the highest power of a bootstrapped matrix sample that appears in the computation. Unlike with ES-T, we will only benchmark the windowing function presented in 130. Like above, we take $\mathbf{R} = \mathbf{I}$ and let the distribution of $\mathbf{b}$ be the standard normal distribution.

In our numerical experiments, we measure two metrics of error:

1. **R. MSE** *(Relative Mean Squared Error)*: This is a normalized version of the error function $\mathcal{E}(\cdot)$ with norm matrix $\mathbf{B} = \mathbf{I}$,
$$\text{R. MSE} \equiv \frac{\mathcal{E}_F(\tilde{\mathbf{x}})}{\|\mathbf{A}^{-1}\|_F^2} = \frac{\mathbb{E}\|(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}}) - \mathbf{A}^{-1}\|_F^2}{\|\mathbf{A}^{-1}\|_F^2}. \tag{138}$$
Therefore, this quantity measures both the relative error of $\tilde{\mathbf{x}}$ from the true solution $\mathbf{x}$ in L^2, as well as the relative error from our augmented operator $\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}}$ to the true operator $\mathbf{A}^{-1}$ in the Frobenius norm. We evaluate this quantity with Monte-Carlo and provide a 2σ estimate of the error of the Monte-Carlo procedure.
2. **Rel. EMSE** *(Relative Energy-Norm Mean Squared Error)*: This is defined like the above, except it is defined using the Energy norm $\|\cdot\|_{\mathbf{A}}$,
$$\text{Rel. EMSE} \equiv \frac{\mathcal{E}_{\mathbf{A},\mathbf{I}}(\tilde{\mathbf{x}})}{\|\mathbf{A}^{-1}\|_{\mathbf{A},\mathbf{I}}^2}, \tag{139}$$
this quantity may be of more interest than Rel. MSE in many problems, as for many elliptic systems, it more heavily penalizes high-frequency noise.

11.1 1D and 2D Poisson equation on a noisy background

Our first benchmark will be the Poisson equation, given by

$$\begin{aligned} \nabla \cdot (a(x)\nabla u(x)) = b(x), &\qquad \text{on } \mathcal{D}, \\ u(x) = 0, &\qquad \text{on } \partial\mathcal{D}, \end{aligned} \tag{140}$$

where $a(x) > 0$ is a function determined by the physical background of the system. We discretize this equation using finite differences as follows: let $G_{\mathcal{D}} = (V, E)$ be a regular grid on the domain $\mathcal{D}$, with vertices V and edges E. Let $\mathbf{E} \in \mathbb{R}^{V \times E}$ be the (arbitrarily oriented) incidence operator of the grid, i.e.,

$$\mathbf{E}_{v,e} = \begin{cases} \pm 1 & v \text{ is incident to } e \\ 0 & \text{otherwise} \end{cases}, \tag{141}$$

Table 1 Comparison of augmentation methods for a 1D Poisson problem on 128 grid points, where $a(x) = 1$ and $\hat{z}_e \sim \mathcal{U}\{0.5, 1.5\}$

Method	Order	Window	R. MSE (%)	$\pm 2\sigma$ (%)	R. EMSE (%)	$\pm 2\sigma$ (%)
Naive	–	–	12	±0.0352	55.1	±0.1
GS	–	–	0.59	±0.0203	24.6	±0.448
ES	–	–	4.32	±0.124	20	±0.362
ES-T	2	Soft	4.77	±0.155	39.7	±0.723
ES-T	4	Soft	1.26	±0.044	21.5	±0.39
ES-T	6	Soft	3.11	±0.0946	20.1	±0.364
ES-T	2	Hard	0.79	±0.0319	22.9	±0.42
ES-T	4	Hard	2.71	±0.0855	20.3	±0.367
ES-TRA	2	–	0.798	±0.029	22.7	±0.406
ES-TRA	4	–	2.88	±0.0913	20.2	±0.366
ES-TRA	6	–	4.01	±0.117	20	±0.354

$\mathbf{E}_{v,e}$ is positive for one of the v incident to e and negative for the other. The discrete approximation for the differential operator in Eq. (140) is given by

$$\mathbf{L} = -\mathbf{E}\mathbf{W}\mathbf{E}^T \,, \tag{142}$$

where $\mathbf{W} \in \mathbb{R}^{E\times E}$ is a diagonal matrix whose e, eth entry is the function a evaluated at the midpoint of e.

We suppose that we only have noisy measurements of the physical background, i.e., that the matrix $\mathbf{W}$ is subject to some randomness. Hence, in practice, we only have access to an approximate

$$\hat{\mathbf{L}} = -\mathbf{E}\hat{\mathbf{W}}\mathbf{E}^T \,, \tag{143}$$

where $\hat{\mathbf{L}}$ is drawn from a distribution D_{ω^*}, where $\omega^* = (a(x_e))_{e\in E}$, i.e., the background a evaluated at all the edge midpoints x_e. Note, to use the operator shifting method, one must prescribe a class of distributions D_ω that we may sample from given background samples ω.

In particular, the noisy background model we use for this benchmark perturbs every observation with independent multiplicative noise,

$$\hat{\mathbf{W}}_{e,e} = \hat{\omega}_e = \hat{z}_e \omega_e \,, \tag{144}$$

where $\hat{z}_e \sim \mathcal{Z}$ i.i.d. for some positive distribution $\mathcal{Z}$ to be specified. To enforce Dirichlet boundary conditions, we solve

$$\begin{aligned} \hat{\mathbf{L}}_{\mathrm{int}(G_{\mathcal{D}}),\mathrm{int}(G_{\mathcal{D}})}\mathbf{u}_{\mathrm{int}(G_{\mathcal{D}})} + \hat{\mathbf{L}}_{\mathrm{int}(G_{\mathcal{D}}),\partial G_{\mathcal{D}}}\mathbf{u}_{\partial G_{\mathcal{D}}} &= \mathbf{b}, \\ \mathbf{u}_{\partial G_{\mathcal{D}}} &= 0, \end{aligned} \tag{145}$$

where $\mathrm{int}(G_{\mathcal{D}}) \subset V$ denotes the interior of the grid $G_{\mathcal{D}}$ and $\partial G_{\mathcal{D}} \subset V$ denotes the boundary, and $\hat{\mathbf{L}}_{A,B}$ for $A \subset V$ and $B \subset V$ denotes the A, B-minor of $\hat{\mathbf{L}}$. Hence, this becomes

$$\hat{\mathbf{A}}\hat{\mathbf{x}} = \mathbf{b}, \tag{146}$$

where $\hat{\mathbf{A}} = \hat{\mathbf{L}}_{\mathrm{int}(G_{\mathcal{D}}),\mathrm{int}(G_{\mathcal{D}})}$ and $\mathbf{b}$ is the function $b(x)$ sampled at the interior vertices of $G_{\mathcal{D}}$.

In Tables 1, 2, and 3, we see the results of operator shifting applied to the above Poisson equation problem. As we can see, all our methods produce a substantial improvement in

Table 2 Comparison of augmentation methods for a 1D Poisson problem on 128 grid points, where $a(x) = 1$ and $\hat{z}_e \sim \Gamma(\mu = 1, \sigma = 0.45)$

Method	Order	Window	R. MSE (%)	$\pm 2\sigma$ (%)	R. EMSE (%)	$\pm 2\sigma$ (%)
Naive	–	–	7.52	±0.0247	58	±0.145
GS	–	–	0.802	±0.0317	32.5	±0.803
ES	–	–	6.5	±0.189	24.9	±0.546
ES-T	2	Soft	3.28	±0.161	46.6	±2.75
ES-T	4	Soft	1.15	±0.0384	29.4	±0.721
ES-T	6	Soft	5.22	±0.177	25.1	±0.524
ES-T	2	Hard	1.07	±0.0385	29.7	±0.716
ES-T	4	Hard	6.27	±0.183	25	±0.541
ES-TRA	2	–	1.28	±0.0435	29	±0.801
ES-TRA	4	–	7.04	±0.203	24.7	±0.507
ES-TRA	6	–	22.9	±0.633	31.8	±0.586

Table 3 Comparison of augmentation methods for a 2D Poisson problem on 128 x 128 grid points, where $a(x) = 1$ and $\hat{z}_e \sim \mathcal{U}\{0.4, 1.6\}$

Method	Order	Window	R. MSE (%)	$\pm 2\sigma$ (%)	R. EMSE (%)	$\pm 2\sigma$ (%)
Naive	–	–	6.43	±0.105	45.3	±0.11
GS	–	–	0.234	±0.00974	25	±0.64
ES	–	–	4.31	±0.772	20	±0.456
ES-T	2	Soft	2.57	±0.465	35.6	±0.908
ES-T	4	Soft	0.876	±0.133	21.8	±0.504
ES-T	6	Soft	2.59	±0.515	20.3	±0.561
ES-T	2	Hard	0.422	±0.039	23.1	±0.464
ES-T	4	Hard	2.13	±0.353	20.5	±0.519
ES-TRA	2	–	0.845	±0.117	22	±0.503
ES-TRA	4	–	3.62	±0.586	20.1	±0.499
ES-TRA	6	–	5.61	±0.887	20.2	±0.471

both relative MSE and relative energy-norm MSE—with GS obtaining the largest reduction in L^2 error and ES obtaining the largest reduction in energy-norm error, as is to be expected. Moreover, note that the truncated methods ES-T and ES-TRA quickly approach the efficacy of ES as one increases the truncation order, with an order of 6 usually being enough to obtain an error comparable to baseline ES (which requires significantly more computation for large-scale problems). Note also, that the energy error of ES-T is always monotonically decreasing, which agrees with Theorem 5. Moreover, note that the error of ES-TRA is not always monotonically decreasing. The unfortunate reality is that, while ES-TRA is guaranteed to converge to ES as the order becomes large, this convergence may be uneven and is not guaranteed to be monotonic like ES-T with a soft window. We also note that the performance of our technique is comparable across different problems (i.e., 1D vs. 2D), as well as across different models of randomness (i.e., discrete vs. gamma).

11.2 Graph Laplacian systems with noisy edge weights

One may extend the model in the above section to general graphs $G = (V, E)$. However, convention typically dictates that the Laplacian should be positive definite instead of negative definite, i.e.,

$$\hat{\mathbf{L}} = \mathbf{E}\hat{\mathbf{W}}\mathbf{E}^T. \tag{147}$$

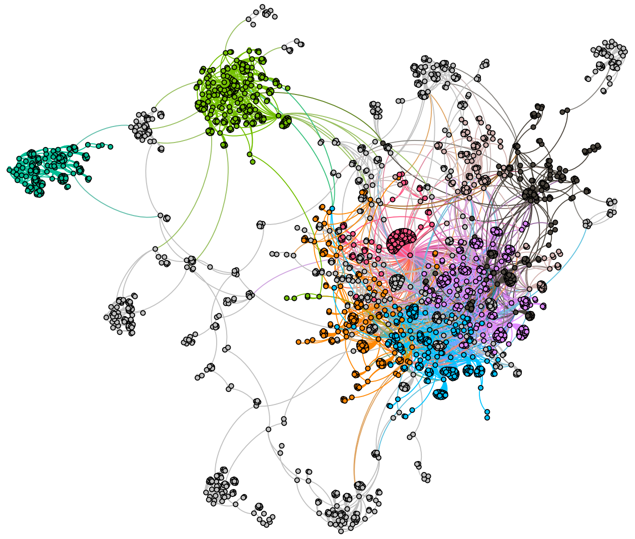

Fig. 3 A visualization of the *fb-pages-food* graph used in our numerical experiments. We give our performance results on this graph in Tables 4 and 6

Table 4 Comparison of augmentation methods for a graph Laplacian system

Method	Order	Window	R. MSE (%)	$\pm 2\sigma$ (%)	R. EMSE (%)	$\pm 2\sigma$ (%)
Naive	–	–	46.9	±0.386	46.2	±0.14
GS	–	–	17.5	±1.27	19.1	±0.464
ES	–	–	18.1	±1.37	19.1	±0.502
ES-T	2	Soft	34.9	±2.79	34.6	±0.999
ES-T	4	Soft	18.9	±1.1	20	±0.433
ES-T	6	Soft	16.9	±1.03	18.6	±0.407
ES-T	2	Hard	20.1	±1.11	21	±0.438
ES-T	4	Hard	17.8	±1.2	19	±0.452
ES-TRA	2	–	20.6	±1.18	21.6	±0.451
ES-TRA	4	–	18	±1.13	19.3	±0.431
ES-TRA	6	–	18.2	±1.17	19.3	±0.449

This particular weighted graph is the fb-pages-food graph [17], visualized in Fig. 3. In this benchmark we have $\hat{z}_e \sim \mathcal{U}\{0.5, 1.5\}$

In this model, we suppose that we are given a weighted graph G; however, the true edge weights of the graph, denoted by w_e, are unknown to us—but we have access to a noisy observation $\hat{w}_e$ of w_e. Like in the previous setting, we suppose that the observations are independent. Therefore, the diagonal weight matrix $\hat{\mathbf{W}}$ again has entries given by

$$\hat{\mathbf{W}}_{e,e} = \hat{\omega}_e = \hat{z}_e \omega_e \,, \tag{148}$$

where $\hat{z}_e \sim \mathcal{Z}$ i.i.d. for some distribution $\mathcal{Z}$ to be specified. Like in the previous example, we solve a Dirichlet problem, arbitrarily selecting approximately six vertices as our boundary ∂G, whose values we set to zero. Thus, $\hat{\mathbf{A}} = \hat{\mathbf{L}}_{\text{int}(G),\text{int}(G)}$ with $\text{int}(G) = V \setminus \partial G$ like before, and we again solve Eq. (146) with operator shifting.

We see the results of this computation in Tables 4 and 5. The graphs shown are from the Network Repository [16]. We note that the method performs quite similarly on this

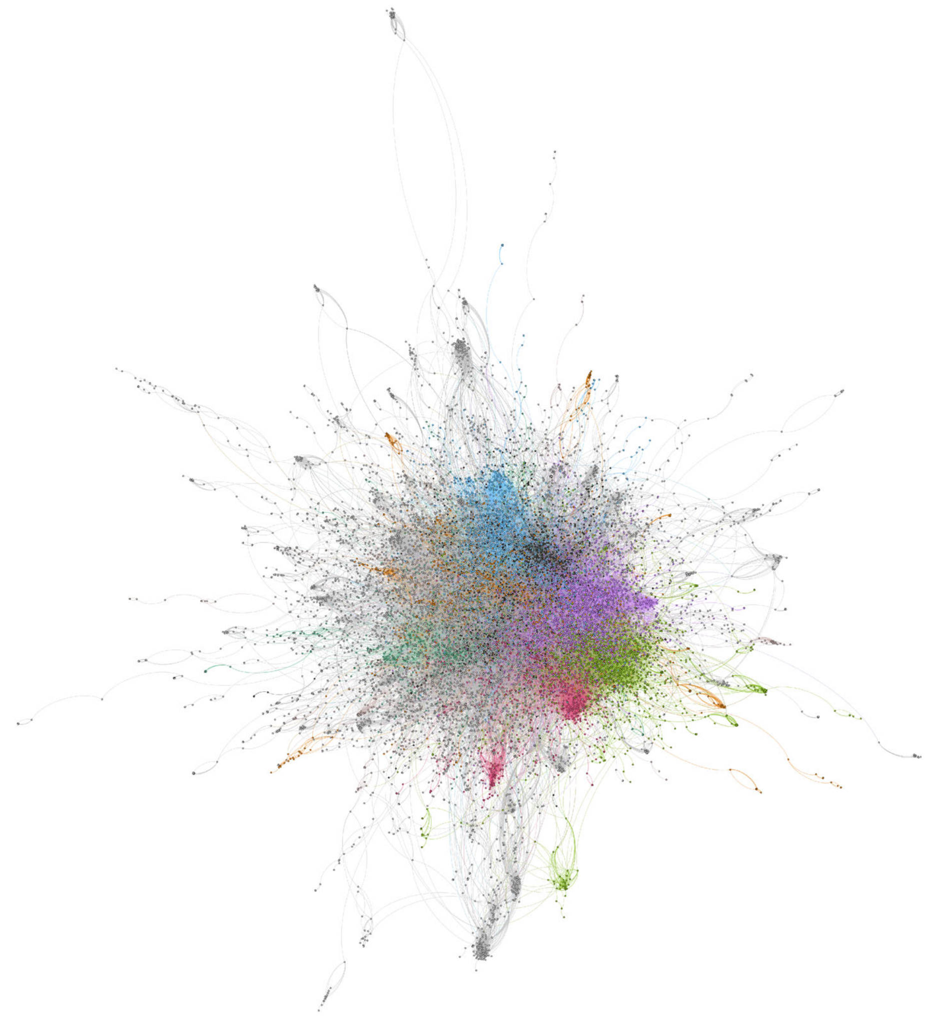

Fig. 4 A visualization of the *fb-pages-company* graph used in our numerical experiments. We give our performance results on this graph in Tables 4 and 6

Table 5 Comparison of augmentation methods for a graph Laplacian system

Method	Order	Window	R. MSE (%)	$\pm 2\sigma$ (%)	R. EMSE (%)	$\pm 2\sigma$ (%)
Naive	–	–	33.8	±3.16	32.4	±1.28
GS	–	–	13.3	±4.92	16.8	±2.22
ES	–	–	13.7	±5.7	16.1	±2.15
ES-T	2	Soft	32.5	±17.9	27.8	±7.67
ES-T	4	Soft	18	±11.3	18.1	±4.78
ES-T	6	Soft	22.3	±12	20.1	±4.76
ES-T	2	Hard	11.8	±4.59	15.3	±2.18
ES-T	4	Hard	13.1	±4.49	14.9	±1.7
ES-TRA	2	–	11.6	±4.08	15.8	±1.81
ES-TRA	4	–	19.8	±5.88	22	±2.13
ES-TRA	6	–	35.4	±11.3	34.7	±4.17

This particular weighted graph is the fb-pages-company graph [17], visualized in Fig. 4. In this benchmark we have $\hat{z}_e \sim \mathcal{U}\{0.5, 1.5\}$

Table 6 Comparison of augmentation methods for a sparsified graph Laplacian system

Method	Order	Window	R. MSE (%)	$\pm 2\sigma$ (%)	R. EMSE (%)	$\pm 2\sigma$ (%)
Naive	–	–	18.5	±0.0843	26.2	±0.125
GS	–	–	12.6	±0.386	16.9	±0.613
ES	–	–	13.8	±0.258	16.9	±0.38
ES-T	2	Soft	16.9	±0.771	23.7	±1.18
ES-T	4	Soft	12.7	±0.448	17.5	±0.737
ES-T	6	Soft	13	±0.379	17.2	±0.611
ES-T	2	Hard	14.1	±0.528	20.1	±0.802
ES-T	4	Hard	12.2	±0.326	16.4	±0.498
ES-TRA	2	–	12.6	±0.404	17	±0.627
ES-TRA	4	–	13.7	±0.315	17	±0.493
ES-TRA	6	–	15.3	±0.318	17.8	±0.367

This particular weighted graph is the fb-pages-food graph [17], visualized in Fig. 3. In this benchmark we have $\hat{z}_e \sim \frac{1}{0.75}\text{Ber}(0.75)$ with $\gamma = 1$

problem as it does on the grid Laplacian case—this shows that the performance of the method is consistent across different types of problems.

11.3 Heat steady-state with sparsified graph Laplacians

In many areas of computer science, one can use graph sparsification techniques to reduce the complexity of a Laplacian system solve if one is able to tolerate some degree of approximation. These graph sparsification techniques work by randomly selecting some subset of the edges of the graph G to remove and then re-weighting the remaining edges to obtain a sparsified graph $\hat{G}$. We consider the problem of approximating the solution to a Laplacian system on G using the Laplacian of $\hat{G}$. In particular, suppose we are interested in the steady-state heat distribution given by

$$(\mathbf{L} + \gamma \mathbf{I})\mathbf{u} = \mathbf{b}, \tag{149}$$

where $\gamma > 0$ is the coefficient of heat decay and $\mathbf{b}$ is the vector describing heat introduced to the system per unit time. However, we only have access to the topology of the sparsified $\hat{G}$ and its Laplacian $\hat{\mathbf{L}}$. Naively, one could solve

$$(\hat{\mathbf{L}} + \gamma \mathbf{I})\hat{\mathbf{u}} = \mathbf{b}. \tag{150}$$

Of course, this naive solution carries a certain amount of error. Note that we can apply operator shifting to $\hat{\mathbf{L}} + \gamma \mathbf{I}$ to obtain a more accurate solution.

In particular, for this numerical experiment, we use the sparsification model

$$\hat{\mathbf{W}}_{e,e} = \hat{\omega}_e = \hat{z}_e \omega_e \tag{151}$$

where $\hat{z}_e \sim p^{-1}\text{Ber}(p)$ i.i.d., for $p \in (0, 1)$.

We see in Table 6 that our methods allow for a substantial reduction in energy-norm mean squared error like in the previous two scenarios. However, this scenario seems to be more difficult for the augmentation process. Particularly, the L^2 reduction is not as high as in previous examples. Regardless, the fact that operator shifting functions under this regime of noise shows us that operator shifting is a technique that can be broadly applied to various problems.

Table 7 Comparison of pros/cons of different augmentation methods presented in this paper

Method	Computation	L^2	Energy	Convergence	Monotone
Naive	Lowest	–	–	–	–
GS	High	Best	Good	–	–
ES	High	Good	Best	–	–
ES-T-S	Low	Good	Better	When $\hat{\mathbf{A}} \prec 2\mathbf{A}$	Always
ES-T-H	Low	Good	Better+	When $\hat{\mathbf{A}} \prec 2\mathbf{A}$	Empirically
ES-TRA	Moderate	Good-	Better-	Pointwise	No

L^2 and energy denote reduction in L^2 and energy-norm error, respectively. (S) and (H) denote hard and soft windows, respectively. Convergence denotes whether or not the method converges to ES when the order is taken to be large, monotone denotes whether or not the truncated shift factors β_N of the method are monotonic

12 Conclusion

In this paper, we have presented a novel method for reducing error in elliptic systems corrupted by noise that requires only a single sample of a corrupted system. We have introduced the GS and ES methods, as well as the ES-T and ES-TRA methods, for efficiently approximating ES. Moreover, we have proved multiple important theorems that underlie our methods—this includes the error reduction bounds in Theorems 2 and 4 for the GS and ES methods, respectively, as well as monotone convergence guarantees Theorems 5 and 6 that provide justification and intuition for the ES-T and ES-TRA methods.

Furthermore, we have demonstrated in our numerical experiments that the operator shifting methods we presented are effective in many different scenarios and different noise models—consistently providing a 2× reduction in energy mean-squared error, and often a significantly higher reduction in L^2 error. We have also shown that ES-T and ES-TRA converge relatively quickly to ES, which makes these truncated methods good alternatives when solving a large number of matrix systems is computationally intractable.

Our numerical results also make clear the relative benefits and trade-offs of the different augmentation methods; these are seen in Table 7. As per these trade-offs, we recommend using ES if computation is not an issue. If computation is an issue, we recommend using hard-window (or soft-window) ES-T, depending on the scenario, and if this approximation seems not to be performing well, or the noise distribution is heavy-tailed, then we recommend using ES-TRA.

While the operator shifting framework offers a new approach to reducing error in noisy elliptic systems, there are still a number of interesting avenues for further exploration. The most obvious is, of course, the extension of the operator shifting framework machinery to the case of asymmetric systems. Unfortunately, while there is nothing preventing one from using the same approach for asymmetric systems, the question of how one would analyze such an algorithm remains open. The machinery developed within does not relatively apply, since the move from symmetric to asymmetric systems breaks a number of core tools used throughout. Since many systems of interest are indeed asymmetric, this is an important direction for future research. In addition, while we leave the optional choice of matrices **B**, **R**, **C**, **D** up to the reader—it is yet unclear how one should approach making a choice for these optional parameters in general. Finally, to judge the performance of the method in real-world problems, one could apply the techniques we've developed within to an application area where elliptic systems are corrupted by randomness—possible aforementioned applications include structural dynamics [18] or whether modeling [15], among a plethora of others.

Funding

Funding was provided by National Science Foundation (Grant Nos. DMS1818449, DMS-2011699).

Data availability

For reproducibility and reference purposes, we provide an implementation of all the algorithms in this paper and the corresponding benchmarks herein, along with instructions on how to reproduce our work, https://github.com/UniqueUpToPermutation/OperatorShifting.

Author details

[1]Meta Reality Labs, Redmond, WA, USA, [2]Mathematics Department, Stanford University, Stanford, CA, USA.

A Proofs of miscellaneous lemmas and theorems

Lemma 1 (Löwner Order Inversion) *Suppose that* $\mathbf{A} \in S_+(\mathbb{R}^n)$ *and* $\hat{\mathbf{A}} \in S_+(\mathbb{R}^n)$ *almost surely. Moreover, suppose that,* $\mathbf{A}$ *spectrally dominates* $\hat{\mathbf{A}}$ *in expectation, i.e.,*

$$\mathbb{E}[\hat{\mathbf{A}}] \preceq \mathbf{A}, \tag{20}$$

then, matrix inversion inverts the expected Löwner order, i.e.,

$$\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1} \tag{21}$$

Proof Consider the exact second-order Taylor expansion of the inverse functional on the space of positive definite matrices,

$$\hat{\mathbf{A}}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\hat{\mathbf{A}} - \mathbf{A})\mathbf{A}^{-1} + \mathbf{A}_*^{-1}(\hat{\mathbf{A}} - \mathbf{A})\mathbf{A}_*^{-1}(\hat{\mathbf{A}} - \mathbf{A})\mathbf{A}_*^{-1}, \tag{152}$$

where $\mathbf{A}_*$ is a matrix between $\mathbf{A}$ and $\hat{\mathbf{A}}$. Note that the last term is positive semi-definite because $\mathbf{A}_*$ is positive definite. Therefore,

$$\hat{\mathbf{A}}^{-1} \succeq \mathbf{A}^{-1} - \mathbf{A}^{-1}(\hat{\mathbf{A}} - \mathbf{A})\mathbf{A}^{-1}. \tag{153}$$

Taking expectations of both sides and using the fact that $\mathbb{E}[\hat{\mathbf{A}} - \mathbf{A}] \preceq \mathbf{0}$ yields

$$\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbb{E}[\hat{\mathbf{A}} - \mathbf{A}]\mathbf{A}^{-1} \succeq \mathbf{A}^{-1}. \tag{154}$$

□

Theorem 2 *Under the assumptions in Sect. 3, consider operator shifting in the* $\|\cdot\|_{\mathbf{B},\mathbf{R}}$*-norm. Any operator shift* $\hat{\mathbf{K}} = \mathbf{C}\hat{\mathbf{A}}^{-1}\mathbf{D}$ *such that* $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times n}$ *satisfy the compatibility conditions Eq.* (30) *has an optimal shift factor that satisfies:*

$$\sqrt{\frac{\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{B}\mathbf{C},\mathbf{D}\mathbf{R}\mathbf{D}^T}}} \geq \beta^* \geq \frac{\mathcal{E}_{\mathbf{C}^T\mathbf{B},\mathbf{R}\mathbf{D}^T}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{B}\mathbf{C},\mathbf{D}\mathbf{R}\mathbf{D}^T}} \geq 0. \tag{31}$$

And the corresponding optimal reduction in error is given by

$$\max_{\beta\in\mathbb{R}} \frac{\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - \mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})}{\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \geq \frac{\mathcal{E}_{\mathbf{C}^T\mathbf{B},\mathbf{R}\mathbf{D}^T}(\hat{\mathbf{A}}^{-1})^2}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{B}\mathbf{C},\mathbf{D}\mathbf{R}\mathbf{D}^T}\,\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}, \tag{32}$$

where $\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{X}})$ *is the mean squared error of matrix estimator* $\hat{\mathbf{X}}$ *in the* $\|\cdot\|_{\mathbf{B},\mathbf{R}}$*-norm.*

Proof We would like to repeat the results of the previous section, except now we want to choose the shift factor β that optimizes the $(\mathbf{B}, \mathbf{R})$-error. Just like in the previous section, we obtain

$$\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}}) = \mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - 2\beta\,\mathbb{E}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}} + \beta^2\,\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}, \tag{155}$$

and hence, the optimal shift factor β^* is given by

$$\beta^* = \frac{\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}}}{\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}}, \tag{156}$$

and the corresponding optimal error is

$$\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta^*\hat{\mathbf{K}}) = \mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - \frac{(\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}})^2}{\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}}, \tag{157}$$

Let us expand the quantity

$$\begin{aligned}\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}} &= \mathbb{E}\langle \mathbf{C}\hat{\mathbf{A}}^{-1}\mathbf{D}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}} \\ &= \mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{D}^T\mathbf{B},\mathbf{R}\mathbf{C}^T}\end{aligned} \tag{158}$$

We want to repeat the argument of the theorem in the previous section. Namely, we would like to have

$$\mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{D}^T\mathbf{B},\mathbf{R}\mathbf{C}^T} \geq 0, \tag{159}$$

so that we can complete the square in Eq. (158). To prove this fact, we will use $\mathbf{M} = \mathbf{M}^T$ to denote $\mathbf{D}^T\mathbf{B} = \mathbf{R}\mathbf{C}^T \succeq \mathbf{0}$. Now, we simply need to do some manipulations inside the trace,

$$\begin{aligned}\mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{M},\mathbf{M}} &= \mathbb{E}\,\mathrm{tr}(\mathbf{M}\mathbf{A}^{-1}\mathbf{M}(\hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1})) \\ &= \mathbb{E}\,\mathrm{tr}(\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\hat{\mathbf{A}}^{-1}) - \mathrm{tr}(\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{A}^{-1}) \\ &= \mathbb{E}\,\mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{M}\hat{\mathbf{A}}^{-1}\mathbf{M}\mathbf{A}^{-1/2}) - \mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{A}^{-1/2}) \\ &= \mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{M}\,\mathbb{E}[\hat{\mathbf{A}}^{-1}]\mathbf{M}\mathbf{A}^{-1/2}) - \mathrm{tr}(\mathbf{A}^{-1/2}\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{A}^{-1/2}).\end{aligned} \tag{160}$$

Since $\mathbb{E}[\hat{\mathbf{A}}^{-1}] \succeq \mathbf{A}^{-1}$ by Lemma 1, it follows that:

$$\mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{M},\mathbf{M}} \geq 0. \tag{161}$$

Using this fact and returning to Eq. (158), we obtain

$$\begin{aligned}\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}} &= \mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{D}^T\mathbf{B},\mathbf{R}\mathbf{C}^T} \\ &\geq \mathbb{E}\langle \hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{D}^T\mathbf{B},\mathbf{R}\mathbf{C}^T} \\ &\quad - \mathbb{E}\langle \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{D}^T\mathbf{B},\mathbf{R}\mathbf{C}^T} \\ &= \mathbb{E}\langle \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{D}^T\mathbf{B},\mathbf{R}\mathbf{C}^T} \\ &= \mathcal{E}_{\mathbf{C}^T\mathbf{B},\mathbf{R}\mathbf{D}^T}(\hat{\mathbf{A}}^{-1}).\end{aligned} \tag{162}$$

Similarly, an expansion of the term $\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}$ gives:

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}} &= \mathbb{E}\,\mathrm{tr}(\mathbf{R}^{1/2}\mathbf{D}^T\hat{\mathbf{A}}^{-1}\mathbf{C}^T\mathbf{B}\mathbf{C}\hat{\mathbf{A}}^{-1}\mathbf{D}\mathbf{R}^{1/2}) \\ &= \mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{B}\mathbf{C},\mathbf{D}\mathbf{R}\mathbf{D}^T}\end{aligned} \tag{163}$$

For a bound in the opposite direction, we simply invoke Cauchy-Schwarz:

$$\begin{aligned}\mathbb{E}\langle \hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{B},\mathbf{R}} &\leq \sqrt{\mathbb{E}\|\hat{\mathbf{K}}\|^2_{\mathbf{B},\mathbf{R}}\,\mathbb{E}\|\hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|^2_{\mathbf{B},\mathbf{R}}} \\ &= \sqrt{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{C}^T\mathbf{B}\mathbf{C},\mathbf{D}\mathbf{R}\mathbf{D}^T}\,\mathcal{E}_{\mathbf{B},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}\end{aligned} \tag{164}$$

Therefore, the desired result follows immediately from Eqs. (156) and (157). □

Theorem 4 *Under the assumptions in Sect. 3, consider operator shifting in energy norm $\|\cdot\|_{\mathbf{A},\mathbf{R}}$. Any operator shift $\hat{\mathbf{K}} = \hat{\mathbf{A}}^{-1}\mathbf{C}$ such that $\mathbf{C}$ satisfies the compatibility conditions Eq. (39) has an optimal shift factor that satisfies:*

$$1 \geq \sqrt{\frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{C}^T\mathbf{R}\mathbf{C}}}} \geq \beta^* \geq \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}\mathbf{C}^T}(\hat{\mathbf{A}}^{-1})}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{C}^T\mathbf{R}\mathbf{C}}} \geq 0\,. \tag{40}$$

And the corresponding optimal reduction in relative error is given by

$$\max_{\beta\in\mathbb{R}} \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1}) - \mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1} - \beta\hat{\mathbf{K}})}{\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \geq \frac{\mathcal{E}_{\mathbf{A},\mathbf{R}\mathbf{C}^T}(\hat{\mathbf{A}}^{-1})^2}{\mathbb{E}\|\hat{\mathbf{A}}^{-1}\|^2_{\mathbf{A},\mathbf{C}^T\mathbf{R}\mathbf{C}}\,\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{A}}^{-1})} \tag{41}$$

where $\mathcal{E}_{\mathbf{A},\mathbf{R}}(\hat{\mathbf{X}})$ is the mean squared error of matrix estimator $\hat{\mathbf{X}}$ in the $\|\cdot\|_{\mathbf{A},\mathbf{R}}$-norm.

Proof This proof is more or less a carbon copy of the proof of Theorem 2. The only difference is when lower bounding

$$\begin{aligned}\mathbb{E}\langle\hat{\mathbf{K}}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}} &= \mathbb{E}\langle\hat{\mathbf{A}}^{-1}\mathbf{C}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}}\\ &= \mathbb{E}\langle\hat{\mathbf{A}}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}\mathbf{C}^T}\end{aligned} \tag{165}$$

The crucial inequality we need to complete the square as in the previous proof is

$$\mathbb{E}\langle\mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}\mathbf{C}^T} \geq 0\,. \tag{166}$$

expanding the quantity on the left-hand side

$$\begin{aligned}&\mathbb{E}\langle\mathbf{A}^{-1}, \hat{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\rangle_{\mathbf{A},\mathbf{R}\mathbf{C}^T}\\ &\quad= \mathbb{E}\,\mathrm{tr}((\mathbf{R}\mathbf{C}^T)^{1/2}\hat{\mathbf{A}}^{-1}(\mathbf{R}\mathbf{C}^T)^{1/2}) - \mathrm{tr}((\mathbf{R}\mathbf{C}^T)^{1/2}\mathbf{A}^{-1}(\mathbf{R}\mathbf{C}^T)^{1/2})\\ &\quad= \mathrm{tr}((\mathbf{R}\mathbf{C}^T)^{1/2}\,\mathbb{E}[\hat{\mathbf{A}}^{-1}](\mathbf{R}\mathbf{C}^T)^{1/2}) - \mathrm{tr}((\mathbf{R}\mathbf{C}^T)^{1/2}\mathbf{A}^{-1}(\mathbf{R}\mathbf{C}^T)^{1/2}) \geq 0\,.\end{aligned} \tag{167}$$

Thus, the result follows as in Theorem 2. □

Lemma 5 *Let $\hat{\mathbf{W}}$ be a symmetric random matrix that satisfies $(1-\varepsilon)\mathbf{I} \preceq \hat{\mathbf{W}} \prec \mathbf{I}$ almost surely and $\mathbb{E}[(\mathbf{I} - \hat{\mathbf{W}})^{-2}]$ exists. Then it is the case that:*

$$\mathbb{E}\|\hat{\mathbf{W}}^k\|_F^2 = o(1/k^2)\,. \tag{168}$$

Proof Note that $\hat{\mathbf{W}}$ is symmetric and hence can always has a spectral decomposition

$$\hat{\mathbf{W}} = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^T\,. \tag{169}$$

Using the above decomposition, for any positive $\gamma > 0$, we can split the matrix $\hat{\mathbf{W}}$ into two matrices $\hat{\mathbf{W}}_{\geq\gamma} + \hat{\mathbf{W}}_{<\gamma}$ with the properties

$$\gamma\mathbf{I} \preceq \hat{\mathbf{W}}_{\geq\gamma} \prec \mathbf{I}, \qquad -(1-\varepsilon) \preceq \hat{\mathbf{W}}_{<\gamma} \prec \gamma\mathbf{I}\,. \tag{170}$$

We do this by defining $\hat{\mathbf{\Lambda}}_{\geq\gamma}$ and $\hat{\mathbf{\Lambda}}_{<\gamma}$ to be $\hat{\mathbf{\Lambda}}$ but with all entries zeroed that don't fall within the ranges $[\gamma, 1)$ and $[-1+\varepsilon, \gamma)$, respectively. Then we have that $\hat{\mathbf{\Lambda}} = \hat{\mathbf{\Lambda}}_{\geq\gamma} + \hat{\mathbf{\Lambda}}_{<\gamma}$ and therefore, we can define:

$$\hat{\mathbf{W}}_{\geq\gamma} \equiv \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}_{\geq\gamma}\hat{\mathbf{Q}}^T\,, \qquad \hat{\mathbf{W}}_{<\gamma} \equiv \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}_{<\gamma}\hat{\mathbf{Q}}^T\,. \tag{171}$$

Moreover, since $\hat{\mathbf{\Lambda}}^k = \hat{\mathbf{\Lambda}}^k_{\geq\gamma} + \hat{\mathbf{\Lambda}}^k_{<\gamma}$, we have that

$$\hat{\mathbf{W}}^k = \hat{\mathbf{W}}^k_{\geq\gamma} + \hat{\mathbf{W}}^k_{<\gamma}\,. \tag{172}$$

Hence, it follows that:

$$k^2\|\hat{\mathbf{W}}^k\|_F^2 \le 2k^2\|\hat{\mathbf{W}}^k_{\ge\gamma}\|_F^2 + 2k^2\,\|\hat{\mathbf{W}}^k_{<\gamma}\|_F^2\,. \tag{173}$$

Furthermore, since $\|\hat{\mathbf{W}}^k_{<\gamma}\|_F^2$ is the sum of the eigenvalues of $\hat{\mathbf{W}}^{2k}_{<\gamma}$, which are all bounded by $\max(1-\varepsilon,\gamma)^{2k}$, it follows again that:

$$k^2\|\hat{\mathbf{W}}^k\|_F^2 \le 2k^2\,\|\hat{\mathbf{W}}^k_{\ge\gamma}\|_F^2 + 2nk^2\cdot\max(1-\varepsilon,\gamma)^{2k}\,. \tag{174}$$

For the remaining term $\|\hat{\mathbf{W}}^k_{\ge\gamma}\|_F^2$, we note that for $x\in[0,1)$,

$$k^2x^k \le 2\sum_{i=1}^k ix^k \le 2\sum_{i=1}^k ix^i\,, \tag{175}$$

whereas the exact Taylor expansion for $1/(1-x)^2$ to $k+1$th order has the form:

$$\frac{1}{(1-x)^2} = \sum_{i=1}^k ix^i + (k+1)y^{k+1} \ge \sum_{i=1}^k ix^i\,, \tag{176}$$

where $0\le y\le x$. Thus, for $x\in[0,1)$,

$$k^2x^k \le \frac{2}{(1-x)^2}\,. \tag{177}$$

Hence, since all eigenvalues of $\hat{\mathbf{W}}^k_{\ge\gamma}$ lie in the range $[\gamma,1)$, it follows that

$$\begin{aligned} k^2\hat{\mathbf{W}}^{2k}_{\ge\gamma} &= k^2\hat{\mathbf{W}}^{2k}_{\ge\gamma}\,(1-\mathbb{1}(\hat{\mathbf{W}}\preceq\gamma\mathbf{I}))\\ &\preceq \frac{1}{2}(\mathbf{I}-\hat{\mathbf{W}}_{\ge\gamma})^{-2}\,(1-\mathbb{1}(\hat{\mathbf{W}}\preceq\gamma\mathbf{I}))\\ &\preceq \frac{1}{2}(\mathbf{I}-\hat{\mathbf{W}})^{-2}\,(1-\mathbb{1}(\hat{\mathbf{W}}\preceq\gamma\mathbf{I})), \end{aligned} \tag{178}$$

where $\mathbb{1}(\hat{\mathbf{W}}\preceq\gamma\mathbf{I})$ is the indicator function for the event $\{\hat{\mathbf{W}}\preceq\gamma\mathbf{I}\}$. The first line is by virtue of the fact that $\hat{\mathbf{W}}^{2k}_{\ge\gamma}$ is zero on the set $\{\hat{\mathbf{W}}\preceq\gamma\mathbf{I}\}$. Thus, we may substitute this into Eq. (174) to obtain:

$$k^2\mathbb{E}\|\hat{\mathbf{W}}^k\|_F^2 \le 2\mathbb{E}[\mathrm{tr}((\mathbf{I}-\hat{\mathbf{W}})^{-2})\,(1-\mathbb{1}(\hat{\mathbf{W}}\preceq\gamma\mathbf{I}))] + 2nk^2\cdot\max(1-\varepsilon,\gamma)^{2k}\,. \tag{179}$$

Now, we choose γ to be $\gamma = 1-1/\sqrt{k}$. This gives

$$k^2\mathbb{E}\|\hat{\mathbf{W}}^k\|_F^2 \lesssim \mathbb{E}[\mathrm{tr}((\mathbf{I}-\hat{\mathbf{W}})^{-2})\,(1-\mathbb{1}(\hat{\mathbf{W}}\preceq(1-1/\sqrt{k})\mathbf{I}))] + k^2(1-1/\sqrt{k})^{2k}\,. \tag{180}$$

The two terms above are quite easy to bound, note that

$$k^2(1-1/\sqrt{k})^{2k} \le k^2\exp(-2k/\sqrt{k}) = k^2\exp(-2\sqrt{k}) = o(1)\,. \tag{181}$$

Conversely, we have

$$\mathrm{tr}((\mathbf{I}-\hat{\mathbf{W}})^{-2})\,\mathbb{1}(\hat{\mathbf{W}}\preceq(1-1/\sqrt{k})\mathbf{I}) \nearrow \mathrm{tr}((\mathbf{I}-\hat{\mathbf{W}})^{-2})\,. \tag{182}$$

Therefore, it follows from the monotone convergence theorem and the convergence of $\mathbb{E}[(\mathbf{I}-\hat{\mathbf{W}})^{-2}]$ that

$$\mathbb{E}[\mathrm{tr}((\mathbf{I}-\hat{\mathbf{W}})^{-2})\,(1-\mathbb{1}(\hat{\mathbf{W}}\preceq(1-1/\sqrt{k})\mathbf{I}))] = o(1)\,. \tag{183}$$

Plugging Eqs. (181) and (183) into Eq. (180) gives the desired result,

$$\mathbb{E}\|\hat{\mathbf{W}}^k\|_F^2 = o(1/k^2)\,. \tag{184}$$

□

Lemma 2 *Let $\hat{\mathbf{X}} \in S_+(\mathbb{R}^n)$ be a random matrix such that $\mathbb{E}[\hat{\mathbf{X}}^{-2}]$ exists and $\hat{\mathbf{X}} \preceq (2-\varepsilon)\mathbf{Y}$ almost surely for $\mathbf{Y} \in S_+(\mathbb{R}^n)$ and $\varepsilon > 0$. Consider the infinite Taylor series for $\hat{\mathbf{X}}^{-1}$ and $\hat{\mathbf{X}}^{-2}$, respectively, about base-point $\mathbf{Y}$, i.e.,*

$$\begin{aligned}\hat{\mathbf{X}}^{-1} &\sim \mathbf{Y}^{-1/2}\left[\sum_{k=0}^{\infty}(-\mathbf{Y}^{-1/2}(\hat{\mathbf{X}}-\mathbf{Y})\mathbf{Y}^{-1/2})^k\right]\mathbf{Y}^{-1/2},\\ \hat{\mathbf{X}}^{-2} &\sim \mathbf{Y}^{-1/2}\left[\sum_{k=0}^{\infty}(k+1)(-\mathbf{Y}^{-1/2}(\hat{\mathbf{X}}-\mathbf{Y})\mathbf{Y}^{-1/2})^k\right]\mathbf{Y}^{-1/2}.\end{aligned} \tag{56}$$

Both series converge in mean-squared Frobenius norm to their respective limits.

Proof Via a transformation of variables, it suffices to prove the statements

$$(\mathbf{I}-\hat{\mathbf{W}})^{-1} = \sum_{k=0}^{\infty}\hat{\mathbf{W}}^k, \qquad (\mathbf{I}-\hat{\mathbf{W}})^{-2} = \sum_{k=0}^{\infty}(k+1)\hat{\mathbf{W}}^k, \tag{185}$$

in the mean squared Frobenius norm when $-(1-\varepsilon)\mathbf{I} \prec \hat{\mathbf{W}} \prec \mathbf{I}$ almost surely and $\mathbb{E}[(\mathbf{I}-\hat{\mathbf{W}})^{-2}] \prec \infty$. Let us write:

$$\begin{aligned}\mathbb{E}\left\|(\mathbf{I}-\hat{\mathbf{W}})^{-1}-\sum_{k=0}^{N}\hat{\mathbf{W}}^k\right\|_F^2 &\leq \left(\mathbb{E}\left\|(\mathbf{I}-\hat{\mathbf{W}})^{-1}\right\|_F^2\right)\left(\mathbb{E}\left\|\mathbf{I}-(\mathbf{I}-\hat{\mathbf{W}})\sum_{k=0}^{N}\hat{\mathbf{W}}^k\right\|_F^2\right)\\ &= \left(\mathbb{E}\left\|(\mathbf{I}-\hat{\mathbf{W}})^{-1}\right\|_F^2\right)\left(\mathbb{E}\left\|\hat{\mathbf{W}}^{N+1}\right\|_F^2\right)\\ &= \operatorname{tr}(\mathbb{E}(\mathbf{I}-\hat{\mathbf{W}})^{-2})\left(\mathbb{E}\left\|\hat{\mathbf{W}}^{N+1}\right\|_F^2\right)\\ &\lesssim \mathbb{E}\left\|\hat{\mathbf{W}}^{N+1}\right\|_F^2 \to 0.\end{aligned} \tag{186}$$

The first inequality above is by Cauchy–Schwarz and convergence in the last line is by Lemma 5.

$$\begin{aligned}&\mathbb{E}\left\|(\mathbf{I}-\hat{\mathbf{W}})^{-2}-\sum_{k=0}^{N}(k+1)\hat{\mathbf{W}}^k\right\|_F^2\\ &\leq \left(\mathbb{E}\left\|(\mathbf{I}-\hat{\mathbf{W}})^{-1}\right\|_F^2\right)^2\left(\mathbb{E}\left\|\mathbf{I}-(\mathbf{I}-\hat{\mathbf{W}})^2\sum_{k=0}^{N}(k+1)\hat{\mathbf{W}}^k\right\|_F^2\right)\\ &= \left(\operatorname{tr}(\mathbb{E}(\mathbf{I}-\hat{\mathbf{W}})^{-2})\right)^2\left(\mathbb{E}\left\|\mathbf{I}-(\mathbf{I}-\hat{\mathbf{W}})^2\sum_{k=0}^{N}(k+1)\hat{\mathbf{W}}^k\right\|_F^2\right)\\ &\lesssim \left(\mathbb{E}\left\|\mathbf{I}-(\mathbf{I}-\hat{\mathbf{W}})^2\sum_{k=0}^{N}(k+1)\hat{\mathbf{W}}^k\right\|_F^2\right)\\ &= \mathbb{E}\left\|\mathbf{I}-(\mathbf{I}-2\hat{\mathbf{W}}+\hat{\mathbf{W}}^2)\sum_{k=0}^{N}(k+1)\hat{\mathbf{W}}^k\right\|_F^2\\ &= \mathbb{E}\left\|(N+1)\hat{\mathbf{W}}^{N+1}-N\hat{\mathbf{W}}^{N+2}\right\|_F^2\\ &\leq 2\mathbb{E}\|(N+1)\hat{\mathbf{W}}^{N+1}\|_F^2 + 2\mathbb{E}\|N\hat{\mathbf{W}}^{N+2}\|_F^2 \to 0,\end{aligned} \tag{187}$$

where we have once again invoked Lemma 5 for the convergence on the last line. Note that we use Cauchy–Schwarz twice on the first line above. □

Lemma 3 *Let $a_1, a_2, ..., a_k, ... \in \mathbb{R}_{\geq 0}$ and $b_1, b_2..., b_k, ... \in \mathbb{R}_{\geq 0}$ be two sequences of nonnegative real numbers with $b_1 > 0$, and consider the truncated sum ratios*

$$\beta_N \equiv \frac{\sum_{k=1}^N a_k}{\sum_{k=1}^N b_k}, \tag{67}$$

then, if it is the case that

$$\frac{a_k}{b_k} \geq \frac{a_{k-1}}{b_{k-1}}, \tag{68}$$

for all k (e.g., the ratios a_k/b_k are monotonically increasing), then the sequence $\beta_1, \beta_2, ..., \beta_k, ...$ is monotonically increasing.

Proof Consider the following series of equivalent inequalities,

$$\begin{aligned} \beta_N &\geq \beta_{N-1}, \\ \frac{\sum_{k=1}^N a_k}{\sum_{k=1}^N b_k} &\geq \frac{\sum_{k=1}^{N-1} a_k}{\sum_{k=1}^{N-1} b_k}, \\ \left(\sum_{k=1}^N a_k\right)\left(\sum_{k=1}^{N-1} b_k\right) &\geq \left(\sum_{k=1}^N b_k\right)\left(\sum_{k=1}^{N-1} a_k\right), \\ \left(a_N + \sum_{k=1}^{N-1} a_k\right)\left(\sum_{k=1}^{N-1} b_k\right) &\geq \left(b_N + \sum_{k=1}^{N-1} b_k\right)\left(\sum_{k=1}^{N-1} a_k\right), \\ \sum_{k=1}^{N-1} a_N b_k &\geq \sum_{k=1}^{N-1} b_N a_k \end{aligned} \tag{188}$$

The last inequality above is clearly true because the terms in the sum on the left dominate their corresponding terms on the right. Therefore, the first inequality is also true. □

Lemma 4 *Let $\hat{\mathbf{X}}$ be a random matrix such that $\hat{\mathbf{X}} \succeq \mathbf{0}$ a.s. For $i \geq j$ and $r \geq 0$, and any symmetric positive semi-definite matrix $\mathbf{S} \succeq \mathbf{0}$, we have that*

$$\langle \hat{\mathbf{X}}^{i+r} \rangle_{\mathbf{S}} \langle \hat{\mathbf{X}}^{j-r} \rangle_{\mathbf{S}} \geq \langle \hat{\mathbf{X}}^{i} \rangle_{\mathbf{S}} \langle \hat{\mathbf{X}}^{j} \rangle_{\mathbf{S}}. \tag{124}$$

Proof We make a series of simplifications. The first assumption is that $\hat{\mathbf{X}}$ is a uniform random variable over a set of (not necessarily distinct) outcomes $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, i.e., has distribution

$$\mathcal{D} = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{X}_i}, \tag{189}$$

where $\delta_{\mathbf{X}_k}$ is the delta distribution supported at $\mathbf{X}_k$. Since any continuous distribution can be approximated by a series of discrete distributions of this above form, it suffices to prove the statement for discrete distributions of the form above. Under this assumption, the inequality Eq. (124) becomes

$$\sum_{k,l} \langle \mathbf{X}_l^i \rangle_{\mathbf{S}} \langle \mathbf{X}_k^j \rangle_{\mathbf{S}} \leq \sum_{i,l} \langle \mathbf{X}_l^{i+r} \rangle_{\mathbf{S}} \langle \mathbf{X}_k^{j-r} \rangle_{\mathbf{S}}. \tag{190}$$

Therefore, it suffices to consider individual pairs $\{i, l\}$ under the sum and show that for any $\mathbf{A}, \mathbf{B} \succeq \mathbf{0}$,

$$\begin{aligned} &\langle \mathbf{B}^i \rangle_{\mathbf{S}} \langle \mathbf{A}^j \rangle_{\mathbf{S}} + \langle \mathbf{A}^i \rangle_{\mathbf{S}} \langle \mathbf{B}^j \rangle_{\mathbf{S}} \\ &\quad \leq \langle \mathbf{A}^{j-r} \rangle_{\mathbf{S}} \langle \mathbf{B}^{i+r} \rangle_{\mathbf{S}} + \langle \mathbf{B}^{j-r} \rangle_{\mathbf{S}} \langle \mathbf{A}^{i+r} \rangle_{\mathbf{S}}. \end{aligned} \tag{191}$$

Let $\lambda_i(\mathbf{A})$ and $\lambda_i(\mathbf{B})$ denote the eigenvalues of $\mathbf{A}, \mathbf{B}$, respectively. Note that since $\langle\cdot\rangle_{\mathbf{S}}$ is a linear functional it satisfies

$$\langle \mathbf{A}^j \rangle_{\mathbf{S}} = \sum_i s_i \lambda_i(\mathbf{A})^j , \tag{192}$$

for some $s_i \geq 0$ that don't depend on j. Therefore, Eq. (191) amounts to

$$\begin{aligned} &\sum_{i,l} s_i s_l' \left[\lambda_i(\mathbf{A})^j \lambda_l(\mathbf{B})^k + \lambda_i(\mathbf{B})^j \lambda_l(\mathbf{A})^k \right] \\ &\quad \leq \sum_{i,l} s_i s_l' \left[\lambda_i(\mathbf{A})^{j-r} \lambda_l(\mathbf{B})^{k+r} + \lambda_i(\mathbf{B})^{j-r} \lambda_l(\mathbf{A})^{k+r} \right] . \end{aligned} \tag{193}$$

Since s_i and s_i' are nonnegative, it suffices to prove that for any nonnegative $a, b \geq 0$,

$$a^j b^k + b^j a^k \leq a^{j-r} b^{k+r} + b^{j-r} a^{k+r} \tag{194}$$

To prove Eq. (194), define the function

$$C_{a,b,s}(\Delta) = a^{s-\Delta} b^{s+\Delta} + b^{s-\Delta} a^{s+\Delta} = 2 a^s b^s \cosh\left(\Delta \log(a/b)\right) . \tag{195}$$

If we take $s = (k+j)/2$, the claim Eq. (194) can be rephrased as

$$C_{a,b,s}\left(\frac{k-j}{2}\right) \leq C_{a,b,s}\left(\frac{k-j}{2} + r\right) , \tag{196}$$

so it suffices to prove $C_{a,b,s}$ is monotonic in Δ for $\Delta \geq 0$—and this follows from the fact that $\cosh(x)$ is monotonically increasing for $x \geq 0$ and monotonically decreasing for $x \leq 0$. □

Received: 30 May 2023 Accepted: 20 October 2023 Published online: 12 November 2023

References

1. Anderson, G.W., Guionnet, A., Zeitouni, O.: An Introduction to Random Matrices, vol. 118. Cambridge University Press, Cambridge (2010)
2. Aspri, A., Korolev, Y., Scherzer, O.: Data driven regularization by projection. Inverse Prob. **36**(12), 125009 (2020)
3. Bergou, E.H., Boucherouite, S., Dutta, A., Li, X., Ma, A.: A note on randomized Kaczmarz algorithm for solving doubly-noisy linear systems. arXiv preprint arXiv:2308.16904 (2023)
4. Bleyer, I.R., Ramlau, R.: A double regularization approach for inverse problems with noisy data and inexact operator. Inverse Prob. **29**(2), 025004 (2013)
5. Buccini, A., Donatelli, M., Ramlau, R.: A semiblind regularization algorithm for inverse problems with application to image deblurring. SIAM J. Sci. Comput. **40**(1), 452–483 (2018)
6. Candes, E.J., Plan, Y.: Matrix completion with noise. Proc. IEEE **98**(6), 925–936 (2010)
7. Davis, C., Kahan, W.M.: The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal. **7**(1), 1–46 (1970)
8. Derezinski, M., Mahoney, M.W.: Distributed estimation of the inverse hessian by determinantal averaging. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
9. Derezinski, M., Bartan, B., Pilanci, M., Mahoney, M.W.: Debiasing distributed second order optimization with surrogate sketching and scaled regularization. Adv. Neural. Inf. Process. Syst. **33**, 6684–6695 (2020)
10. Golub, G.H., Van Loan, C.F.: An analysis of the total least squares problem. SIAM J. Numer. Anal. **17**(6), 883–893 (1980)
11. James, W., Stein, C.: Estimation with quadratic loss. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics, pp. 443–460. Springer, New York (1992)
12. Keshavan, R., Montanari, A., Oh, S.: Matrix completion from noisy entries. In: Advances in Neural Information Processing Systems, pp. 952–960 (2009)
13. Lunz, S., Hauptmann, A., Tarvainen, T., Schonlieb, C.-B., Arridge, S.: On learned operator correction in inverse problems. SIAM J. Imaging Sci. **14**(1), 92–127 (2021)
14. Marzouk, Y., Moselhy, T., Parno, M., Spantini, A.: An introduction to sampling via measure transport. arXiv preprint arXiv:1602.05023 (2016)
15. Palmer, T., Shutts, G., Hagedorn, R., Doblas-Reyes, F., Jung, T., Leutbecher, M.: Representing model uncertainty in weather and climate prediction. Annu. Rev. Earth Planet. Sci. **33**, 163–193 (2005)
16. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI. http://networkrepository.com (2015)
17. Rozemberczki, B., Davies, R., Sarkar, R., Sutton, C.: Gemsec: graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019, pp. 65–72. ACM (2019)

18. Soize, C.: A comprehensive overview of a non-parametric probabilistic approach of model uncertainties for predictive models in structural dynamics. J. Sound Vib. **288**(3), 623–652 (2005)
19. Stein, C., et al.: 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, pp. 197–206 (1956)
20. Stein, E.M., Shakarchi, R.: Fourier Analysis: An Introduction, vol. 1. Princeton University Press, Princeton (2011)
21. Tao, T.: Topics in Random Matrix Theory, vol. 132. American Mathematical Society, Providence (2012)
22. Tikhonov, A.N.: On the solution of ill-posed problems and the method of regularization. In: Doklady Akademii Nauk, vol. 151, pp. 501–504. Russian Academy of Sciences (1963)
23. Xiu, D., Hesthaven, J.S.: High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**(3), 1118–1139 (2005)
24. Xiu, D., Karniadakis, G.E.: The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.