MULTISCALE MODEL. SIMUL.
Vol. 21, No. 3, pp. 1170–1209

AGGREGATION METHODS FOR COMPUTING STEADY STATES IN STATISTICAL PHYSICS*

GABRIEL EARLE† AND BRIAN VAN KOTEN‡

Abstract. We give a new proof of local convergence of a multigrid method called iterative aggregation/disaggregation (IAD) for computing steady states of Markov chains. Our proof leads naturally to precise and interpretable estimates of the asymptotic rate of convergence. We study IAD as a model of more complex methods from statistical physics for computing nonequilibrium steady states, such as the nonequilibrium umbrella sampling method of Warmflash, Bhimalapuram, and Dinner [*J. Chem. Phys.*, 127 (2007), 154112]. We explain why it may be possible to use methods like IAD to efficiently calculate steady states of processes in statistical physics and how to choose parameters to optimize efficiency.

Key words. nonequilibrium, steady states, statistical physics, multigrid, sampling

MSC codes. 60J22, 82C80, 65C05

DOI. 10.1137/22M1530628

1. Introduction.

We prove local convergence of iterative aggregation/disaggregation (IAD) for computing steady states of Markov chains, and we estimate the asymptotic rate of convergence. IAD was devised in the 1960's to solve economic input-output models; see the references given in [21, 34]. Substantially, equivalent methods were independently developed in the 1980's to calculate steady states of Markov chains [3, 4, 14, 15]. In the 2000's, similar ideas arose for the third time as a part of more complex methods for calculating nonequilibrium steady states and reaction rates in statistical physics [1, 2, 9, 11, 35, 36]. We study IAD as a simple model of these complex methods. We explain why it may be possible to use methods like IAD to efficiently calculate steady states in statistical physics and how to choose parameters to optimize efficiency. We hope others will apply our results to understand IAD in other contexts.

We call a Markov process *nonequilibrium* if it is irreversible. A physical system subject to nonconservative forces or external flows of energy and matter would typically be modeled by a nonequilibrium process, e.g., a single-molecule experiment where a protein is subjected to a flow of ions [8, 20]. In principle, to sample the steady state distribution of any ergodic process, reversible or irreversible, one can take the average over a long trajectory. In practice, however, trajectory averages converge to the steady state very slowly when obstacles like bottlenecks inhibit exploration of the state space. For example, a process modeling a protein may spend most of the time vibrating around some stable folded state, undergoing transitions between different folded states only rarely. Such a process is said to be *metastable.*

Computing the steady state of a metastable, nonequilibrium process is especially difficult. Reliable methods have been devised to efficiently compute steady states of

*Received by the editors October 25, 2022; accepted for publication (in revised form) March 31, 2023; published electronically September 8, 2023.
https://doi.org/10.1137/22M1530628
Funding: This research was partially supported by NSF DMS-2012207.
†Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 USA (gearle@umass.edu).
‡Corresponding author. Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 USA (bvankoten@umass.edu).

reversible, metastable processes, e.g., parallel tempering [12, 31], umbrella sampling [17, 30, 33], metadynamics [18], and adaptive biasing [5]. These methods are essential tools for simulating systems in equilibrium. Unfortunately, however, none of them can compute nonequilibrium steady states. Each requires either reversibility or knowledge of the steady state density, and when computing nonequilibrium steady states one typically knows only the generator of the process. By contrast, in equilibrium, the steady state has the Boltzmann density, which can almost always be calculated up to a normalizing constant.

Recently, analogous methods have been devised to compute nonequilibrium steady states and dynamical quantities such as reaction rates. We consider one class derived from umbrella sampling, including nonequilibrium umbrella sampling (NEUS) [36], trajectory parallelization and tilting [35], weighted ensemble with a direct solve [2], exact milestoning [1], trajectory stratification [9], and injection measures [11]. The exact objectives and details of these methods differ significantly, but they are all essentially stochastic evolving particle systems that approximate IAD (or a similar deterministic dynamics) in the limit of a large number of particles. We refer the reader to [11] for details. We ask whether approximating IAD is a suitable goal for an algorithm designed to compute steady states in statistical physics.

IAD is like an algebraic multigrid method, but it is nonlinear and nonsymmetric, which significantly complicates its analysis; cf. Appendix D.3. In the earliest convergence analysis of IAD known to us, Mandel and Sekerka proved local convergence for a class of problems including the solution of input-output models but not steady states of Markov chains [21]. Later work verified local convergence for Markov chains under various conditions and for various versions of the IAD algorithm [11, 16, 22, 23, 24, 25]. We are not aware of a proof of global convergence that holds under general conditions. However, see [24] for a proof of global convergence under somewhat restrictive conditions and examples where IAD fails to converge. The efficiency of IAD has been studied in special cases, including nearly completely decomposable chains [15] and cyclic chains [29].

We contribute a new proof of local convergence of IAD under weak conditions that are easy to verify. The usual conditions that guarantee convergence of a Markov chain to a steady state, irreducibility and aperiodicity, do not suffice to prove even local convergence of IAD; cf. Appendix B.2 and [24]. If P is the transition matrix of the chain, we prove local convergence when P and $P^{\mathrm{t}}P$ are irreducible; cf. Theorem 4.13. It is equivalent to assume that the chain is strictly contracting in a certain norm; cf. Lemma 4.4. A sufficient condition is that P be irreducible and have a positive diagonal.

Our proof of local convergence leads to precise and interpretable estimates of the asymptotic rate of convergence; cf. Theorem 5.4 and Corollary 5.6. Based on these estimates, we have developed some general advice to guide the choice of parameters in IAD; see the discussion following Corollary 5.6. We apply our theory of the rate of convergence in section 6 to explain why it may be possible to use methods like IAD to efficiently compute steady states of reversible or irreversible processes arising in statistical physics and computational chemistry. To be precise, in section 6, we introduce a family of Markov chains analogous to the metastable diffusion processes that are widely used as models of molecular systems. We then explain why IAD can sometimes efficiently calculate the steady states of such chains and how to choose the parameters in practice. We illustrate our conclusions with numerical experiments. We developed our rate estimates with these examples from statistical physics in mind, but our results are general, and we hope others will apply them to understand IAD in other contexts.

2. Notation. Here, we summarize our notation. Notation is also explained below when it first appears.

- $P \in \mathbb{R}^{N\times N}$ will be an irreducible, column stochastic transition matrix with invariant distribution $\mu \in \mathbb{R}^N$.
- For any matrix or vector M, $M \geq 0$ means all entries of M are nonnegative. $M > 0$ means all entries are positive.
- $\mathbb{1}$ will denote a vector of all ones, and I will denote the identity matrix. The dimension of $\mathbb{1}$ or I will be determined by the context.
- For any $A \subset \{1,\dots,N\}$, $\mathbb{1}_A$ will denote the characteristic function of A. That is, $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ if $x \notin A$.
- For any vector $\nu \in \mathbb{R}^k$, we let $\operatorname{diag}(\nu) \in \mathbb{R}^{k\times k}$ denote the diagonal matrix with $\operatorname{diag}(\nu)_{ii} = \nu_i$ for $i = 1,\dots,k$.
- $\|\cdot\|_\nu$ and $\langle,\rangle_\nu$ denote the $\ell^2(\nu)$-norm and inner product, respectively; see Definition 4.1. For $M \in \mathbb{R}^{k\times k}$, $M^{*,\nu}$ denotes the adjoint of M with respect to the $\ell^2(\nu)$-inner product. In some proofs, we simplify notation, letting $\|\cdot\|$, $\langle,\rangle$, and M^* denote the $\ell^2(1/\mu)$-norm, inner product, and adjoint, respectively.
- For any operator $M \in \mathbb{R}^{N\times N}$, $\operatorname{Rg}(M)$ denotes the range of M.

3. The iterative aggregation/disaggregation method. Iterative aggregation/disaggregation (IAD) is a numerical method for computing the steady state distribution of a Markov chain. Let $P \in \mathbb{R}^{N\times N}$ be the *column*[1] stochastic transition probability matrix of a discrete-time Markov chain on the state space

$$\Omega = \{1,\dots,N\}.$$

We call Ω the *fine space.* We assume that P is irreducible,[2] so there is a unique steady state probability vector $\mu \in \mathbb{R}^N$ solving

$$P\mu = \mu.$$

In each step of IAD, one calculates a coarse approximation to P based on a user-specified partition of Ω into disjoint sets $\{S_i : i = 1,\dots,n\}$. We call each set S_i a *coarse state,* and we call the set $\{1,\dots,n\}$ of indices the *coarse space.* To define the coarse approximation, we specify operators mapping between the sets of probability vectors on the fine and coarse spaces. The *aggregation operator* maps vectors on the fine space to vectors on the coarse space.

DEFINITION 3.1. *We define the aggregation operator* $A : \mathbb{R}^N \to \mathbb{R}^n$ *by*

$$(A\nu)_i := \nu^{\mathrm{t}}\mathbb{1}_{S_i} = \sum_{x\in S_i} \nu_x$$

for any $i = 1,\dots,n$ *and* $\nu \in \mathbb{R}^N$.

[1] A matrix is (row) stochastic if its entries are nonnegative and each row sums to one. A matrix is column stochastic if it is nonnegative and each column sums to one. Note that the usual convention in the probability literature is for the transition matrix to be *row* stochastic, so if X_t is a Markov chain with transition matrix P, then $\mathbb{P}[X_{t+1} = j | X_t = i] = P_{ij}$. Following the literature on IAD, we adopt the opposite convention, taking $\mathbb{P}[X_{t+1} = j | X_t = i] = P_{ji}$.

[2] We recall that a column stochastic matrix $P \in \mathbb{R}^{n\times n}$ is irreducible if and only if for any $i, j \in \{1,\dots,n\}$, there exists a $k \in \mathbb{N}$ so that $(P^k)_{ji} > 0$. The Perron–Frobenius theorem guarantees that an irreducible column stochastic matrix has a unique positive steady state probability vector μ so that $P\mu = \mu$.

Note that when ν is a probability vector, $A\nu_i$ is simply the probability of S_i under ν. Moreover, if ν is a probability vector, then so is $A\nu$.

The *disaggregation operator* maps vectors on the coarse space to vectors on the fine space. It depends on the current approximation $\tilde{\mu}$ of the steady state μ.

DEFINITION 3.2. *Given a probability vector $\tilde{\mu} \in \mathbb{R}^N$ with $A\tilde{\mu} > 0$ and a coarse state S_i, define the conditional distribution $\tilde{\mu}(\cdot|S_i)$ by*

$$\tilde{\mu}(j|S_i) = \frac{\tilde{\mu}_j \mathbb{1}_{S_i}(j)}{A\tilde{\mu}_i}$$

for $j \in \Omega$. Here, $\mathbb{1}_{S_i}$ denotes the characteristic function of S_i. Define the disaggregation operator $D(\tilde{\mu}) : \mathbb{R}^n \to \mathbb{R}^N$ by

$$D(\tilde{\mu})z_j := \sum_{i=1}^{n} z_i \tilde{\mu}(j|S_i)$$

for any $z \in \mathbb{R}^n$.

Note that if z is a probability vector, so is $D(\tilde{\mu})z$. Also, observe that $D(\tilde{\mu})$ is defined only when $A\tilde{\mu} > 0$. This will always be the case in practice under our assumptions; cf. Lemma 3.6.

Given an approximation $\tilde{\mu}$ of μ, the coarse approximation $C(\tilde{\mu})$ to P is defined by composing P with A and $D(\tilde{\mu})$.

DEFINITION 3.3. *Let $\tilde{\mu} \in \mathbb{R}^N$ be a probability vector with $A\tilde{\mu} > 0$. We define the coarse approximation $C(\tilde{\mu}) \in \mathbb{R}^{n \times n}$ by*

$$C(\tilde{\mu}) = APD(\tilde{\mu}).$$

The coarse approximation $C(\tilde{\mu})$ is a column stochastic matrix. To see this, note that $C(\tilde{\mu})$ maps probability vectors to probability vectors, since each of A, P, and $D(\tilde{\mu})$ maps probability vectors to probability vectors. In each step of IAD, one solves for the steady state of $C(\tilde{\mu})$. It is convenient to establish some general notation for this operation.

DEFINITION 3.4. *For M an irreducible and column stochastic matrix, we let $z(M)$ denote the unique probability vector solving*

$$z(M) = Mz(M).$$

Now let $\mu^0 \in \mathbb{R}^N$ be a user-specified initial approximation of μ. In IAD, one alternates coarse correction and smoothing steps. Given a probability $\mu^k \in \mathbb{R}^N$, the *coarse correction* step is to compute

$$\mu^{k+\frac{1}{2}} = D(\mu^k)z(C(\mu^k)). \tag{3.1}$$

To compute $z(C(\mu^k))$ in practice, one can use the algorithm outlined in Appendix A. The *smoothing* step is to compute

$$\mu^{k+1} = P\mu^{k+\frac{1}{2}}. \tag{3.2}$$

Note that the smoothing step is the same as one step of the power method for calculating μ and also the same as evolving $\mu^{k+\frac{1}{2}}$ by one step under the forwards equation for the chain.

Many variations of IAD have appeared in the literature. We will not treat all of them. Some versions apply to substochastic problems such as economic input-output models [21, 34]. Others use different smoothers [16, 22, 25], apply multiple smoothing iterations at each step [16, 22, 25], smooth the aggregation or disaggregation operators [6], use a hierarchy of several coarse approximations in a multigrid V- or W-cycle [7], or solve infinite-dimensional problems using more general aggregation and disaggregation operators [11, 25]. We do not consider these possibilities.

We now summarize our assumptions and show that IAD is well-posed.

Assumption 3.5. We assume the following:

1. P is irreducible.
2. The initial approximation μ^0 of the steady state μ is strictly positive.

LEMMA 3.6. *If Assumption* 3.5 *holds, then the iterates* μ^k *produced by IAD are defined for all* $k \in \mathbb{N}$. *In particular,* $A\mu^k > 0$ *and* $C(\mu^k)$ *is irreducible, so* $D(\mu^k)$ *and* $z(C(\mu^k))$ *are defined.*

Proof. See Appendix B.1. ∎

If one does not assume $\mu^0 > 0$, then $C(\mu^0)$ may be reducible, in which case the steady state $z(C(\mu^0))$ need not be unique. This can occur even when P is both irreducible and aperiodic; see Appendix B.2 for an example.

Finally, we present a complete version of IAD with a termination criterion similar to those typically used in practice.

The user must specify the following:

1. a column stochastic and irreducible transition matrix $P \in \mathbb{R}^{N\times N}$,
2. a partition $\{S_i : i = 1, \ldots, n\}$ of $\{1, \ldots, N\}$ into disjoint sets,
3. a probability vector $\mu^0 \in \mathbb{R}^N$ with $\mu^0 > 0$, and
4. an error tolerance $\tau > 0$.

Given these data, IAD proceeds as follows:

1. Set $\mu^{\text{old}} = \mu^0$.
2. Calculate $z(C(\mu^{\text{old}}))$ using the algorithm in Appendix A. Set
$$\mu^{\text{new}} = PD(\mu^{\text{old}})z(C(\mu^{\text{old}})).$$
3. If
$$\max_{i\in 1,\ldots,N} \frac{|\mu_i^{\text{new}} - \mu_i^{\text{old}}|}{\mu_i^{\text{old}}} \le \tau \text{ and } \max_{i\in 1,\ldots,N} \frac{|P\mu_i^{\text{new}} - \mu_i^{\text{new}}|}{P\mu_i^{\text{new}}} \le \tau,$$
then output μ^{new}. Otherwise, set $\mu^{\text{old}} = \mu^{\text{new}}$, and return to step 2 above.

---

4. Local convergence of IAD. In this section, we prove local convergence of IAD. That is, we show that if the initial approximation μ^0 is sufficiently close to the true steady state μ, then μ^k converges to μ. We begin with an analysis of the smoothing step of IAD in section 4.1, which is the power method. In section 4.2, we prove local convergence of IAD.

4.1. The power method. Here, we prove convergence of the power method (or, equivalently, convergence of the forwards equation of the chain) to the steady state μ. Of course, convergence of the power method is already well understood. The

details of our particular proof will be instrumental in our analysis of the efficiency of IAD, however. We begin by defining convenient norms.

DEFINITION 4.1. *Let $\nu \in \mathbb{R}^k$ with $\nu > 0$. We define the $\ell^2(\nu)$-norm and inner product by*

$$\langle x, y\rangle_\nu = \sum_{i=1}^{k} x_i y_i \nu_i \quad \text{and} \quad \|x\|_\nu = \langle x, x\rangle_\nu^{\frac{1}{2}}$$

for $x, y \in \mathbb{R}^k$. Given $M \in \mathbb{R}^{k\times k}$, we let $\|M\|_\nu$ denote the induced operator norm. We let $M^{*,\nu}$ be the adjoint matrix so that

$$\langle Mx, y\rangle_\nu = \langle x, M^{*,\nu} y\rangle_\nu$$

for all $x, y \in \mathbb{R}^k$.

We will use the $\ell^2(1/\mu)$-norm to measure discrepancies between probability measures on Ω, for example the error $\mu^k - \mu$ after k steps of IAD. The $\ell^2(\mu)$-norm will arise when we analyze the efficiency of IAD. As a first step in our analysis of the power method, we relate adjoints in the $\ell^2(1/\mu)$-inner product to time reversals.

LEMMA 4.2. *Let $P \in \mathbb{R}^{N\times N}$ be column stochastic and irreducible. The time reversal of P is $P^{*,1/\mu}$. In particular,*

$$P^{*,1/\mu} = \operatorname{diag}(\mu) P^{\mathrm{t}} \operatorname{diag}(1/\mu) \tag{4.1}$$

is column stochastic and irreducible and has invariant distribution μ.

Proof. See Appendix C.1. ∎

The spectrum of the operator $P^{*,1/\mu}P$ will play a crucial role in our proof of convergence of the power method and in our efficiency analysis of IAD. By Lemma 4.2, $P^{*,1/\mu}P$ is column stochastic with invariant distribution μ. Therefore, $\mathbb{1}$ is a left eigenvector of $P^{*,1/\mu}P$ with eigenvalue 1, and μ is the corresponding right eigenvector. (Here, $\mathbb{1} \in \mathbb{R}^N$ denotes the vector whose entries are all equal to one.) Moreover, $P^{*,1/\mu}P$ is self-adjoint and positive semidefinite with respect to the $\ell^2(1/\mu)$-inner product. Therefore, $\sigma(P^{*,1/\mu}P) \subset [0, 1]$. Let

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$$

be the eigenvalues of $P^{*,1/\mu}P$ listed in decreasing order and with repetition if any have multiplicity greater than one. Let $v_1, \ldots, v_N$ be the corresponding right eigenvectors normalized so that $\|v_i\|_{1/\mu} = 1$ for all $i = 1, \ldots, N$. Since $P^{*,1/\mu}P$ is self-adjoint, the eigenvectors are an orthonormal basis of $\mathbb{R}^N$ with the $\ell^2(1/\mu)$-inner product, so

$$\langle v_i, v_j\rangle_{1/\mu} = \delta_{ij},$$

and we have the diagonalization

$$P^{*,1/\mu}P = \mu\mathbb{1}^{\mathrm{t}} + \sum_{k=2}^{N} \lambda_k v_k v_k^{\mathrm{t}} \operatorname{diag}(1/\mu). \tag{4.2}$$

We will refer to this diagonalization frequently. Note that the left eigenvectors of $P^{*,1/\mu}P$ are

$$v_1' = \mathbb{1}, v_2' = \operatorname{diag}(1/\mu)v_2, \ldots, v_N' = \operatorname{diag}(1/\mu)v_N.$$

The left eigenvectors play an important role in our efficiency analysis of IAD.

We now show that when P is irreducible, the power method for computing μ is strictly contracting in the $\ell^2(1/\mu)$-norm if and only if $P^{\mathrm{t}}P$ is irreducible.

DEFINITION 4.3. *For $P \in \mathbb{R}^{N\times N}$ an irreducible column stochastic matrix and $\nu^0 \in \mathbb{R}^N$ a probability vector, we define the power method iteration*

$$\nu^{k+1} = P\nu^k.$$

LEMMA 4.4. *Let $P \in \mathbb{R}^{N\times N}$ be an irreducible column stochastic matrix. Let $\nu^0 \in \mathbb{R}^N$ be a probability vector, and let ν^k be the corresponding sequence of power method iterates in Definition* 4.3. *Define*

$$\hat{P} = P - \mu\mathbb{1}^{\mathrm{t}}.$$

We have

$$\nu^{k+1} - \mu = \hat{P}(\nu^k - \mu). \tag{4.3}$$

Moreover,

$$\|\hat{P}\|_{1/\mu} = \sqrt{\lambda_2},$$

and $\lambda_2 < 1$ if and only if $P^{\mathrm{t}}P$ is irreducible.

Proof. See Appendix C.2. ∎

Note that the asymptotic rate of convergence of the power method is

$$\lim_{m\to\infty} \|\hat{P}^m\|^{\frac{1}{m}} = \rho(\hat{P}).$$

For irreversible chains, $\rho(\hat{P})$ may be significantly less than the contraction constant $\|\hat{P}\|_{1/\mu}$ of the power method in the $\ell^2(1/\mu)$-norm. See section 6.4 for an example. However, for reversible chains, we have $\rho(\hat{P}) = \|\hat{P}\|_{1/\mu}$, since $\hat{P}^{*,1/\mu} = \hat{P}$.

In our convergence analysis of IAD, we assume that $P^{\mathrm{t}}P$ is irreducible. By formula (4.1) for $P^{*,1/\mu}$, it is equivalent to assume that $P^{*,1/\mu}P$ is irreducible. A sufficient, but not necessary, condition is that P be irreducible with a positive diagonal. We note that if P is irreducible but $P^{\mathrm{t}}P$ is not, then $\bar{P} = \frac{1}{2}(I + P)$ is irreducible, has a positive diagonal, and has the same unique steady state as P. Therefore, to compute the steady state of P one could apply IAD with $\bar{P}$ in place of P, and local convergence would then be guaranteed by the results below.

We now give an example to illustrate what can go wrong when $P^{\mathrm{t}}P$ is reducible. Consider a right shift on three states:

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

This is an irreducible Markov chain, and the steady state is the uniform distribution $\frac{1}{3}\mathbb{1} \in \mathbb{R}^3$. The time reversal is the left shift $P^{*,\frac{1}{3}\mathbb{1}} = P^{\mathrm{t}} = P^{-1}$. Therefore, $P^{*,\frac{1}{3}\mathbb{1}}P = I$ is reducible even though both P and $P^{*,\frac{1}{3}\mathbb{1}}$ are irreducible. In this case, $\|\hat{P}\|_{1/\mu} = 1$, and the power method does not converge, since P is periodic.

The right shift in our last example is irreducible but periodic. There also exist irreducible, aperiodic chains so that $P^{\mathrm{t}}P$ is reducible. For such chains, the power

method is convergent, but it is not a strict contraction in the $\ell^2(1/\mu)$-norm. See Appendix B.2 and [24] for an example of an irreducible, aperiodic chain so that $P^{\mathrm{t}}P$ is reducible and IAD is not locally convergent.

4.2. Local convergence of IAD. Here, we prove local convergence of IAD. We begin with a convenient reformulation of the steady state problem for a Markov chain.

LEMMA 4.5. *Let $M \in \mathbb{R}^{k\times k}$ be an irreducible column stochastic matrix, and let $v, w \in \mathbb{R}^k$ with $\mathbb{1}^{\mathrm{t}}v \neq 0$ and $z(M)^t w \neq 0$. The matrix $I - M + vw^{\mathrm{t}}$ is invertible, and*

$$z(M) = (I - M + vw^{\mathrm{t}})^{-1}vw^{\mathrm{t}}z(M).$$

Proof. See Appendix D.1. ∎

We use two special cases of Lemma 4.5. First, the steady state $\mu \in \mathbb{R}^N$ is the unique solution x of

$$(I - P + \mu\mathbb{1}^{\mathrm{t}})x = \mu. \tag{4.4}$$

Second, the coarse steady state $z(C(\mu^k)) \in \mathbb{R}^n$ is the unique solution of

$$(I - C(\mu_k) + A\mu\mathbb{1}^{\mathrm{t}})x = A\mu. \tag{4.5}$$

Note that one cannot solve the linear equations (4.4) and (4.5) in practice to compute μ and $z(C(\mu^k))$, since both the matrices and the right-hand-sides depend on the unknown μ. We use (4.4) and (4.5) only to derive the following recursive formula for the error after the coarse correction.

LEMMA 4.6. *For any probability vector $\nu \in \mathbb{R}^N$ with $\nu > 0$, the matrix $A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\nu)$ is invertible, and we may define*

$$S(\nu) := D(\nu)[A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\nu)]^{-1}A(I - P + \mu\mathbb{1}^{\mathrm{t}}).$$

The operator $S(\nu)$ is a projection with $\mathrm{Rg}(S(\nu)) = \mathrm{Rg}(D(\nu))$. We call $S(\nu)$ the coarse projection. We have

$$\mu^{k+\frac{1}{2}} - \mu = (I - S(\mu^k))(\mu^k - \mu). \tag{4.6}$$

Proof. See Appendix D.2. ∎

In Appendix D.3, we interpret IAD as an adaptive algebraic multigrid method for solving (4.4). The coarse projection $S(\mu^k)$ is exactly the coarse grid correction in this interpretation. The restriction operator is A, and the prolongation operator is $D(\mu^k)$. Note that $S(\nu)$ is a projection on $\mathrm{Rg}(D(\nu))$. This is the most important algebraic fact in our analysis below. All of our results below would hold if $S(\mu)$ were *any* projection on $\mathrm{Rg}(D(\mu))$, except for Theorem 5.5.

To derive a recursive formula for the error after a complete step of IAD, we simply compose formula (4.6) for the error after the coarse correction with formula (4.3) for the propagation of the error under the power method.

LEMMA 4.7. *Define the* error propagation operator

$$J(\mu^k) := \hat{P}(I - S(\mu^k)).$$

We have

$$\mu^{k+1} - \mu = J(\mu^k)(\mu^k - \mu).$$

Proof. By (4.3) and (4.6), we have

$$\mu^{k+1} - \mu = \hat{P}(\mu^{k+\frac{1}{2}} - \mu) = \hat{P}(I - S(\mu^k))(\mu^k - \mu).$$

Similar formulas for the propagation of the error appear in [21] and subsequent work [22, 23, 25]. ∎

We now show that $J(\mu)$ has norm less than one with respect to a certain operator norm. Local convergence follows. To construct the right norm, we decompose $\mathbb{R}^N$ into $\mathrm{Rg}(D(\mu))$ and its orthogonal complement in $\ell^2(1/\mu)$. The orthogonal projection $\Pi(\mu)$ defined below will be useful.

DEFINITION 4.8. *Given a probability vector $\nu \in \mathbb{R}^N$ with $\nu > 0$, we define the orthogonal coarse projection*

$$\Pi(\nu) := D(\nu)A.$$

Lemma 4.9 summarizes the properties of $\Pi(\nu)$.

LEMMA 4.9. *Let $\nu \in \mathbb{R}^N$ be a positive probability vector. The orthogonal coarse projection $\Pi(\nu)$ is the orthogonal projection on $\mathrm{Rg}(D(\nu))$ with respect to $\langle , \rangle_{1/\nu}$. It is also a reversible, column stochastic matrix with invariant distribution ν.*

Proof. For completeness, we give a proof in Appendix D.4. Equivalent observations appear in [21] and subsequent work [22, 23, 25]. ∎

We will show that for the $\|\cdot\|_\varepsilon$ norm defined below, $\|J(\mu)\|_\varepsilon < 1$ for sufficiently small ε.

DEFINITION 4.10. *For $\varepsilon > 0$, we define the ε-inner product and norm on $\mathbb{R}^N$ by*

$$\langle x, y\rangle_\varepsilon := \langle x, (I - \Pi(\mu))y\rangle_{1/\mu} + \varepsilon\langle x, \Pi(\mu)y\rangle_{1/\mu} \quad \text{and} \quad \|x\|_\varepsilon := \langle x, x\rangle_\varepsilon^{\frac{1}{2}}.$$

For $M \in \mathbb{R}^{N\times N}$, we let $\|M\|_\varepsilon$ denote the induced operator norm.

To verify that $\langle , \rangle_\varepsilon$ is an inner product, observe that it is symmetric, since $\Pi(\mu)$ is an orthogonal projection, and therefore $\Pi(\mu) = \Pi(\mu)^{*,1/\mu}$. It is nondegenerate, since

$$\begin{aligned}\|x\|_\varepsilon^2 &\geq \min\{1, \varepsilon\}\left\{\|(I - \Pi(\mu))x\|_{1/\mu}^2 + \|\Pi(\mu)x\|_{1/\mu}^2\right\} \\ &= \min\{1, \varepsilon\}\|x\|_{1/\mu}^2,\end{aligned}$$

using again that $\Pi(\mu)$ is an orthogonal projection. Bilinearity is inherited from $\langle , \rangle_{1/\mu}$.

We now show that when $\varepsilon > 0$ is small, $\|J(\mu)\|_\varepsilon$ is approximately $\|(I - \Pi(\mu))J(\mu)\|_{1/\mu}$.

LEMMA 4.11. *We have*

$$\|J(\mu)\|_\varepsilon \leq \sqrt{\|(I - \Pi(\mu))J(\mu)\|_{1/\mu}^2 + \varepsilon\|\Pi(\mu) - S(\mu)\|_{1/\mu}^2}.$$

Proof. See Appendix D.5. ∎

We will estimate $\|(I-\Pi(\mu))J(\mu)\|_{1/\mu}$. By Lemma 4.11, if $\|(I-\Pi(\mu))J(\mu)\|_{1/\mu} < 1$, then $\|J(\mu)\|_\varepsilon < 1$ for ε sufficiently small.

THEOREM 4.12. *Assume that P and $P^{\mathrm{t}}P$ are irreducible and that at least one coarse state contains more than one fine state. We have*

$$\|(I-\Pi(\mu))J(\mu)\|_{1/\mu}^2 = 1 - \inf_{z\in \mathrm{Rg}(I-S(\mu))} \frac{\langle z,(I-\hat{P}^{*,1/\mu}\hat{P})z\rangle_{1/\mu}}{\|(I-\Pi(\mu))z\|_{1/\mu}^2} \tag{4.7}$$

$$\leq 1 - \frac{1}{\|(I-\Pi(\mu))(I-\hat{P}^{*,1/\mu}\hat{P})^{-1}(I-\Pi(\mu))\|_{1/\mu}} \tag{4.8}$$

$$< 1.$$

Proof. See Appendix D.6. ∎

We now prove local convergence.

THEOREM 4.13. *Assume that P and $P^{\mathrm{t}}P$ are irreducible. For $\varepsilon > 0$ sufficiently small,*

$$\|J(\mu)\|_\varepsilon < 1. \tag{4.9}$$

For any $\varepsilon > 0$ small enough so that $\|J(\mu)\|_\varepsilon < 1$ and any $\eta \in (0, 1-\|J(\mu)\|_\varepsilon]$, there exists $r > 0$ so that if $\|\mu^0 - \mu\|_\varepsilon \leq r$, then

$$\|\mu^k - \mu\|_\varepsilon \leq (\|J(\mu)\|_\varepsilon + \eta)^k \|\mu^0 - \mu\|_\varepsilon$$

for all $k \in \mathbb{N}$.

Proof. See Appendix D.7. ∎

5. The rate of convergence. Here, we analyze the asymptotic rate of convergence of IAD. First, we show that the rate is bounded by the spectral radius $\rho(J(\mu))$. We then derive an upper bound on the spectral radius based on the results in section 4. Our upper bound is appealing and easy to interpret, but it significantly overestimates $\rho(J(\mu))$ for some irreversible processes. See section 6.4 for an example. Therefore, we also derive an exact formula for $\rho(J(\mu))$. Our exact formula could be the basis for a better understanding of the rate of convergence of IAD for irreversible processes, and it yields an interpretable exact expression for $\rho(J(\mu))$ when P is reversible.

We now show that $\rho(J(\mu))$ bounds the asymptotic rate of convergence. By the asymptotic rate of convergence, we mean the expression on the left-hand side of (5.1) below.

LEMMA 5.1. *Let P and $P^{\mathrm{t}}P$ be irreducible, let $r > 0$ be as in Theorem 4.13, and assume that $\|\mu^0 - \mu\|_\varepsilon < r$. For any norm $\|\cdot\|$ on $\mathbb{R}^N$, we have*

$$\limsup_{n\to\infty} \|\mu^k - \mu\|^{1/k} \leq \rho(J(\mu)). \tag{5.1}$$

Proof. See Appendix E.1. ∎

We now estimate $\rho(J(\mu))$. Our approach is based on comparing the orthogonal coarse projection $\Pi(\mu)$ with the orthogonal projection on the eigenvectors associated with the largest eigenvalues of $P^{*,1/\mu}P$.

DEFINITION 5.2. *Fix some $k < N$, and let Q be the $\ell^2(1/\mu)$-orthogonal projection on the eigenvectors $v_1, \ldots, v_k$ associated with the k largest eigenvalues of $P^{*,1/\mu}P$. That is, define*

$$Q = \mu \mathbb{1}^{\mathrm{t}} + \sum_{i=2}^{k} v_i (v_i')^{\mathrm{t}}. \tag{5.2}$$

Here, $\{v_i : i = 1, \dots, N\}$ *and* $\{v_i' : i = 1, \dots, N\}$ *are the right and left eigenvectors of* $P^{*,1/\mu}P$. *In formula* (5.2)*, we normalize the right eigenvectors so that* $\|v_i\|_{1/\mu} = 1$ *and we take* $v_i' = \mathrm{diag}(1/\mu)v_i$ *as in* (4.2).

Note that $Q = Q^{*,1/\mu}$ and $Q^2 = Q$, so Q is an orthogonal projection in $\ell^2(1/\mu)$. Moreover, by (4.2), we have $QP^{*,1/\mu}P = P^{*,1/\mu}PQ$, and $\sigma(QP^{*,1/\mu}P) = \{\lambda_1, \dots, \lambda_k\}$.

When thinking about Q, we suggest that the reader keep the family of Markov chains defined in section 6.2 in mind. We devised this family of chains as a simple model of the reversible metastable diffusion processes encountered in molecular simulation. For these chains, $P^{*,1/\mu}P$ typically has a small number k of eigenvalues that are very close to one. The rest of the spectrum is much farther from one. That is,

$$\frac{1-\lambda_k}{1-\lambda_{k+1}} \ll 1.$$

In our examples, we choose Q to be the projection associated with these k largest eigenvalues.

Our estimate of $\rho(J(\mu))$ is expressed in terms of the angle from $\mathrm{Rg}(Q^{\mathrm{t}})$ to $\mathrm{Rg}(\Pi(\mu)^{\mathrm{t}})$ in the $\ell^2(\mu)$-inner product.

LEMMA 5.3. *We have* $0 \le \|(I - \Pi(\mu)^{\mathrm{t}})Q^{\mathrm{t}}\|_\mu \le 1$, *and therefore we may define an angle* $\theta \in [0, \pi/2]$ *by*

$$\sin(\theta) := \|(I - \Pi(\mu)^{\mathrm{t}})Q^{\mathrm{t}}\|_\mu. \tag{5.3}$$

Note that here the norm is weighted by μ *not* $1/\mu$.

Proof. Both Π^{t} and Q^{t} are orthogonal projections with respect to the $\ell^2(\mu)$-inner product, since Π and Q are orthogonal projections with respect to $\ell^2(1/\mu)$. Therefore, $\|(I - \Pi(\mu)^{\mathrm{t}})Q^{\mathrm{t}}\|_\mu \le \|(I - \Pi(\mu)^{\mathrm{t}})\|_\mu \|Q^{\mathrm{t}}\|_\mu = 1$, since any orthogonal projection must have norm equal to one. ∎

One can show that the angle defined above coincides with the typical definition

$$\begin{aligned}\sin(\theta) &:= \mathrm{gap}(\mathrm{Rg}(Q^{\mathrm{t}}), \mathrm{Rg}(\Pi^{\mathrm{t}})) \\ &= \max_{\substack{u \in \mathrm{Rg}(Q^{\mathrm{t}}) \\ \|u\|_\mu = 1}} \min_{\substack{w \in \mathrm{Rg}(\Pi^t) \\ \|w\|_\mu = 1}} \|u - w\|_\mu \\ &= \max_{\substack{u \in \mathrm{Rg}(Q^{\mathrm{t}}) \\ \|u\|_\mu = 1}} \|(I - \Pi^t)u\|_\mu.\end{aligned}$$

We do not prove this, since it will not be important below.

We understand θ as a measure of how well one can approximate elements of $\mathrm{Rg}(Q^{\mathrm{t}})$ within $\mathrm{Rg}(\Pi(\mu)^{\mathrm{t}})$. Note that

$$\mathrm{Rg}(\Pi^{\mathrm{t}}(\mu)) = \mathrm{span}\{\mathbb{1}_{S_1}, \dots, \mathbb{1}_{S_n}\}$$

and that

$$\mathrm{Rg}(Q^{\mathrm{t}}) = \mathrm{span}\{v_1', \dots, v_k'\}.$$

That is, $\mathrm{Rg}(\Pi(\mu)^{\mathrm{t}})$ is spanned by the characteristic functions of the coarse states, and $\mathrm{Rg}(Q^{\mathrm{t}})$ is spanned by the first k left eigenvectors of $P^{*,1/\mu}P$. Therefore, θ will be small when each of the first k left eigenvectors of $P^{*,1/\mu}P$ is well-approximated by a

linear combination of characteristic functions of coarse states. Equivalently, θ is small when each of the first k left eigenvectors can be approximated by a function that is constant on the coarse states.

We now estimate the asymptotic rate of convergence.

THEOREM 5.4. *Assume that P and $P^{\mathrm{t}}P$ are irreducible and that at least one coarse state contains more than one fine state. We have*

$$\rho(J(\mu))^2 \le \|(I-\Pi(\mu))J(\mu)\|_{1/\mu}^2$$

$$\le 1 - \frac{1}{\|(I-\Pi(\mu))(I-\hat{P}^{*,1/\mu}\hat{P})^{-1}(I-\Pi(\mu))\|_{1/\mu}} \tag{5.4}$$

$$\le 1 - \frac{1}{\sin^2(\theta)\frac{1}{1-\lambda_2} + \cos^2(\theta)\frac{1}{1-\lambda_{k+1}}}. \tag{5.5}$$

Proof. See Appendix E.2. ∎

In our examples in section 6, we consider metastable chains for which $P^{*,1/\mu}P$ has a small number of eigenvalues very close to one, and we choose k to be the number of such eigenvalues. We note, however, that Theorem 5.4 holds for any k. To obtain a useful estimate, one has to choose k carefully. If k is too large, then $\sin(\theta)$ will be close to one, giving only $\sqrt{\lambda_2}$ as an upper bound on $\rho(J(\mu))$.

Note that the right-hand side of (5.5) increases from λ_{k+1} to λ_2 as $\sin^2(\theta)$ increases from zero to one. Thus, the asymptotic rate of convergence of IAD is never larger than $\sqrt{\lambda_2}$, which is the contraction constant of the power method in the $\ell^2(1/\mu)$-norm. For reversible chains, $\sqrt{\lambda_2}$ is also the asymptotic rate of convergence of the power method, since any reversible P is self-adjoint with respect to the $\ell^2(1/\mu)$-inner product by Lemma 4.2 and therefore $\|\hat{P}\|_{1/\mu} = \rho(\hat{P})$. Thus, for reversible chains, the asymptotic rate of convergence of IAD is never greater than the asymptotic rate for the power method.

For irreversible chains, however, the asymptotic rate of convergence $\rho(\hat{P})$ of the power method may be less than the contraction constant $\|\hat{P}\|_{1/\mu}$. Theorem 5.4 may significantly overestimate $\rho(J(\mu))$ in such cases. For example, suppose there is only a single coarse state $S_1 = \Omega$. In that case, IAD reduces to the power method, and $J(\mu) = \hat{P}$. Moreover, $I - \Pi(\mu) = I - \mu\mathbb{1}^{\mathrm{t}}$, so $(I-\Pi(\mu))J(\mu) = \hat{P}$. Note that our upper bounds in Theorem 5.4 are in fact upper bounds on $\|(I-\Pi(\mu))J(\mu)\|_{1/\mu}$. Therefore, if there is only one coarse state, neither of our upper bounds can be smaller than $\|\hat{P}\|_{1/\mu}$. See Figures 6.5 and 6.6 for additional examples where our upper bounds overestimate $\rho(J(\mu))$.

Since our upper bound in Theorem 5.4 may significantly overestimate the spectral radius for some irreversible chains, we also give an exact formula for the spectrum of $J(\mu)$. This formula could lead to a better understanding of IAD in the irreversible case, and it also leads to an interpretable exact formula for $\rho(J(\mu))$ for reversible processes.

THEOREM 5.5. *Assume that P and $P^{\mathrm{t}}P$ are irreducible. Assume that there is more than one coarse state and at least one coarse state contains more than one fine state. The spectrum of $J(\mu)$ is given by*

$$\sigma(J(\mu)) = \left(1 - \frac{1}{\sigma((I-\Pi(\mu))(I-\hat{P})^{-1}(I-\Pi(\mu)))\setminus\{0\}}\right) \cup \{0\}. \tag{5.6}$$

Proof. See Appendix E.3. ∎

For reversible processes, Theorem 5.5 has the following corollary.

COROLLARY 5.6. *Let P be reversible, and assume that P and $P^{\mathrm{t}}P$ are irreducible. Assume that there is more than one coarse state and at least one coarse state contains more than one fine state. We have*

$$\rho(J(\mu)) = 1 - \frac{1}{\|(I-\Pi(\mu))(I-\hat{P})^{-1}(I-\Pi(\mu))\|_{1/\mu}} \tag{5.7}$$

$$\leq 1 - \frac{1}{\sin^2(\theta)\frac{1}{1-\sqrt{\lambda_2}} + \cos^2(\theta)\frac{1}{1-\sqrt{\lambda_{k+1}}}}. \tag{5.8}$$

Proof. See Appendix E.4. ∎

We include the angle upper bound (5.8) in Corollary 5.6 to demonstrate that the exact formula (5.7) can be interpreted in the same way as the norm upper bound (5.4) in Theorem 5.4. There does not appear to be a meaningful difference between the two angle upper bounds in Theorem 5.4 and Corollary 5.6 when λ_2 and λ_{k+1} are both close to one. In fact, one can show that the two angle upper bounds are asymptotic in various limits as λ_2 and λ_{k+1} tend to one.

We now list three implications of our theory for the choice of coarse states. First, Corollary 5.6 suggests that *for a reversible process one should choose coarse states so that the left eigenvectors of P corresponding to eigenvalues close to one are well-approximated by vectors that are constant on the coarse states.* In section 6, we explain how to interpret this statement for processes like those used in molecular modeling. Similarly, Theorem 5.4 suggests that for irreversible processes one should choose coarse states so that the leading left eigenvectors of $P^{*,1/\mu}P$ are well-approximated. Our examples in section 6.4 indicate that this may be good advice for some irreversible chains but that it could be misleading for some very irreversible chains; cf. Figure 6.6.

Second, since the quality of approximation is measured in the $\ell^2(\mu)$-norm, *one only needs an accurate approximation in regions of high probability under μ.* Regions of low probability will not have a significant influence on $\sin(\theta)$ unless some of the first k eigenvectors are concentrated in those regions. As a consequence, the efficiency of IAD is not always as sensitive to the choice of coarse states as one might expect, and in some cases a very naive choice of coarse states can work quite well. See section 6.5 for an example.

Third, Corollary 5.6 proves that *for reversible chains $\rho(J(\mu))$ decreases whenever the coarse states are refined.* We say that a set of coarse states $\mathcal{R} = \{T_1, \ldots, T_m\}$ is a refinement of $\mathcal{C} = \{S_1, \ldots, S_n\}$ if each S_i can be expressed as a union of T_j's. Let $\Pi_{\mathcal{R}}$ and $\Pi_{\mathcal{C}}$ be the orthogonal coarse projections $\Pi(\mu)$ for the two partitions $\mathcal{R}$ and $\mathcal{C}$, respectively. To see that the spectral radius $\rho(J(\mu))$ for the refined partition $\mathcal{R}$ is less than or equal to the spectral radius for the coarse partition $\mathcal{C}$, observe that $\mathrm{Rg}(\Pi_{\mathcal{R}}) \supset \mathrm{Rg}(\Pi_{\mathcal{C}})$, so

$$\Pi_{\mathcal{C}}\Pi_{\mathcal{R}} = \Pi_{\mathcal{R}}\Pi_{\mathcal{C}} = \Pi_{\mathcal{C}},$$

since both $\Pi_{\mathcal{C}}$ and $\Pi_{\mathcal{R}}$ are $\ell^2(1/\mu)$-orthogonal projections. Therefore,

$$\begin{aligned}
&\|(I-\Pi_{\mathcal{R}})(I-\hat{P})^{-1}(I-\Pi_{\mathcal{R}})\|_{\frac{1}{\mu}} \\
&\quad = \|(I-\Pi_{\mathcal{R}})(I-\Pi_{\mathcal{C}})(I-\hat{P})^{-1}(I-\Pi_{\mathcal{C}})(I-\Pi_{\mathcal{R}})\|_{\frac{1}{\mu}} \\
&\quad \leq \|I-\Pi_{\mathcal{R}}\|_{\frac{1}{\mu}}\|(I-\Pi_{\mathcal{C}})(I-\hat{P})^{-1}(I-\Pi_{\mathcal{C}})\|_{\frac{1}{\mu}}\|I-\Pi_{\mathcal{R}}\|_{\frac{1}{\mu}} \\
&\quad \leq \|(I-\Pi_{\mathcal{C}})(I-\hat{P})^{-1}(I-\Pi_{\mathcal{C}})\|_{\frac{1}{\mu}},
\end{aligned}$$

since $I - \Pi_{\mathcal{R}}$ is an $\ell^2(1/\mu)$-orthogonal projection and so $\|I - \Pi_{\mathcal{R}}\|_{\frac{1}{\mu}} = 1$. It follows by Corollary 5.6 that the spectral radius for the refined partition is less than for the coarse partition.

For irreversible chains, the spectral radius may increase with refinement. See section 6.4 for an example. However, in our examples, we still observe a clear (but not monotone) trend toward lower spectral radii with increasing refinement. We also note that all of the upper bounds on the spectral radius in Theorem 5.4 must decrease with refinement even when P is irreversible.

6. Examples related to modeling molecules. Here, we apply the theory developed in section 5 to develop an understanding of the rate of convergence of IAD for processes similar to those used as molecular models. To begin, we review certain important properties of molecular models, and we define a simple family of Markov chains with similar properties. We then calculate $\rho(J(\mu))$ and the upper bounds in Theorem 5.4 for some members of this family and for various choices of coarse states. Our theory explains the observed dependence of the rate of convergence on the choice of coarse states for all but the most irreversible (and least metastable) chains.

6.1. Molecular models. Molecular modeling begins with the specification of a potential energy $V : \mathbb{R}^M \to \mathbb{R}$ defined on the space of all configurations of the atoms comprising the system. Based on the potential, one defines a stochastic process to model the evolution of the system. For example, the overdamped Langevin dynamics

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2kT}\, dB_t$$

may be used to model a system in contact with a heat bath at temperature T. (Here, k is Boltzmann's constant.) Refer to [19] for details. We recall the following well-known properties of overdamped Langevin:

- Under some conditions on V, the unique steady state of X_t is the *Boltzmann distribution*

$$\pi(dx) = Z^{-1} \exp\left(\frac{V(x)}{kT}\right) dx, \quad \text{where } Z^{-1} = \int_{\mathbb{R}^{3N}} \exp(-\beta V(x))\, dx.$$

- X_t is *reversible.*
- If the potential energy V has several local minima, then when the temperature T is low, X_t is *metastable.* In particular, trajectories tend to vibrate around local minima of V, undergoing transitions between minima only rarely. Under some conditions on V, in the limit as $T \to 0$, each local minimum of V corresponds to an eigenvalue of the generator of X_t that converges exponentially to zero. The remainder of the spectrum remains bounded away from zero uniformly in T. The eigenvectors corresponding to the eigenvalues that converge to zero are approximately constant on the basins of attraction of the minima. See [19, section 2.5] for details.

Overdamped Langevin is reversible, but we take a particular interest in irreversible models, since these are the hardest to sample. For example, consider

$$dX_t = (-\nabla V(X_t) + \alpha F(X_t))\, dt + \sqrt{2kT}\, dB_t, \tag{6.1}$$

where F is a nonconservative force; i.e., F is not the gradient of a potential function. Here, X_t is irreversible [19, section 5.1.2]. There is no general, closed-form expression for the steady state density of (6.1). In particular, the steady state is not the Boltz-

mann distribution. This is one reason why sampling nonequilibrium steady states is difficult.

6.2. Simple Markov chain model of overdamped Langevin. We define a family of Markov chains on a one-dimensional grid with properties similar to overdamped Langevin. Let $V : \mathbb{R} \to \mathbb{R}$, $T > 0$, $[a, b] \subset \mathbb{R}$, and $N \in \mathbb{N}$. Define the discrete Boltzmann distribution $\mu \in \mathbb{R}^N$ by

$$\mu_i = Z^{-1} \exp\left(-\frac{V\left(a + \frac{b-a}{N} i\right)}{T} \right), \tag{6.2}$$

where $Z = \sum_{i=1}^{N} \exp(-\frac{V(a+\frac{b-a}{N} i)}{T})$. Define the transition matrix

$$\begin{aligned}
P_{ii} &:= \frac{1}{2}\left(\frac{\mu(i)}{\mu(i-1)+\mu(i)} + \frac{\mu(i)}{\mu(i+1)+\mu(i)} \right) && \text{for all } i \in \Omega, \\
P_{i+1,i} &:= \frac{1}{2} \frac{\mu(i+1)}{\mu(i+1)+\mu(i)} && \text{for all } i \in \Omega, \\
P_{i-1,i} &:= \frac{1}{2} \frac{\mu(i-1)}{\mu(i-1)+\mu(i)} && \text{for all } i \in \Omega, \text{ and} \\
P_{ji} &:= 0 && \text{otherwise.}
\end{aligned} \tag{6.3}$$

In the definition of P, we impose periodic boundary conditions, associating 0 with N, 1 with $N+1$, etc. This family of Markov chains was proposed in [32] as a model of overdamped Langevin and other metastable processes often encountered in statistical physics. We also define a similar family of chains on a two-dimensional grid; see Appendix F and section 6.5.

The Markov chain P has properties similar to overdamped Langevin: Observe that P is in detailed balance with the discrete Boltzmann distribution μ, so P is reversible and has invariant distribution μ. In our examples below, we choose T small, and in that case P is metastable, as demonstrated in [32]. Moreover, in the examples given in section 6.3, for each local minimum of V, there is one eigenvalue of P that lies very close to one and the remainder of the spectrum lies much farther from one. The left eigenvectors of P associated with the eigenvalues that lie close to one are approximately constant on the basins of attraction of the minima. We will not prove that these properties of the spectrum hold in general (or even formulate them precisely), but we note that they do hold in our examples.

We also define irreversible Markov chains that are analogous to overdamped Langevin with a nonconservative force (6.1). Define the *right shift* $W \in \mathbb{R}^{N \times N}$ by

$$\begin{aligned}
W_{i+1,i} &:= 1 \quad \text{for all } i \in \Omega, \text{ and} \\
W_{j,i} &:= 0 \quad \text{otherwise,}
\end{aligned} \tag{6.4}$$

taking periodic boundary conditions as in the definition of P. We consider chains of the form $(1-\alpha)P + \alpha W$ for $\alpha \in (0, 1)$.

6.3. IAD for a metastable, reversible chain on a one-dimensional grid. We now test our theory on a highly metastable, reversible problem. We will see that for any sufficiently refined choice of coarse states, IAD converges quickly compared with the power method. However, for some very poor choices of coarse states, IAD converges at essentially the same rate as the power method. We explain these results in detail using the rate estimate in terms of θ (5.5) and the properties of molecular models outlined above.

TABLE 6.1
The largest five eigenvalues of $P^{*,1/\mu}P$ for the reversible, one-dimensional chain P of section 6.3. We report $\sqrt{\lambda_k}$ instead of λ_k, since it is $\sqrt{\lambda_k}$ that appears in Theorem 5.4. We have added the third column for comparison with Figure 6.3.

k	$\sqrt{\lambda_k}$	$-\log_{10}(1-\sqrt{\lambda_k})$
1	1	∞
2	0.999992	5.09
3	0.991441	2.07
4	0.986243	1.86
5	0.979807	1.69

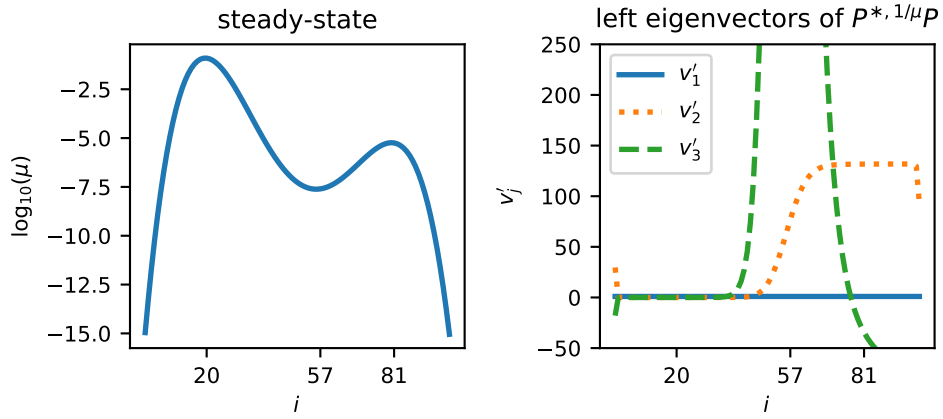

FIG. 6.1. *Left: the steady state μ. Right: the eigenvectors of $P^{*,1/\mu}P$. Note that v'_1 and v'_2 are approximately constant on the basins of attraction of the local minima of V, which correspond to maxima of the steady state distribution. Here, the local minima of μ at $i = 57$ and $i = 0$ separate the basins of attraction. Note that $i = 0$ and $i = 100$ are identified, since we impose periodic boundary conditions.*

Let μ be the discrete Boltzmann distribution defined in (6.2) with

$$V(x) = (1 - x^2)^2 + \frac{1}{2}x,$$

$N = 100$, $[a, b] = [-1.7, 1.55]$, and $T = 1/10$. Let P be the corresponding reversible transition matrix defined by (6.3). Here, the potential V has two minima, so we expect that exactly two eigenvalues of $P^{*,1/\mu}P = P^2$ will lie very close to one with the remainder significantly farther from one. Table 6.1 confirms that this is indeed the case. The left eigenvectors are displayed in Figure 6.1. Based on our discussion of molecular models above, we expect that the eigenvectors corresponding to the two largest eigenvalues should be approximately constant on the basins of attraction of V. Here, on the grid used to define μ, V has a local maximum at $i = 57$. It has a global maximum at $i = 0$, which is identified with $i = 100$ by periodicity. These maxima divide the state space into two basins of attraction. Observe that the first two eigenvectors, $v'_1 = \mathbb{1}$ and v'_2, are roughly constant on the basins of attraction. The third is not.

We compute $\rho(J(\mu))$ and the upper bounds in Theorem 4.13 for several different choices of coarse states. First, we test uniform grids. We show for this simple, one-dimensional system that IAD converges quickly whenever the coarse states are sufficiently refined. Therefore, one does not need detailed prior knowledge of the

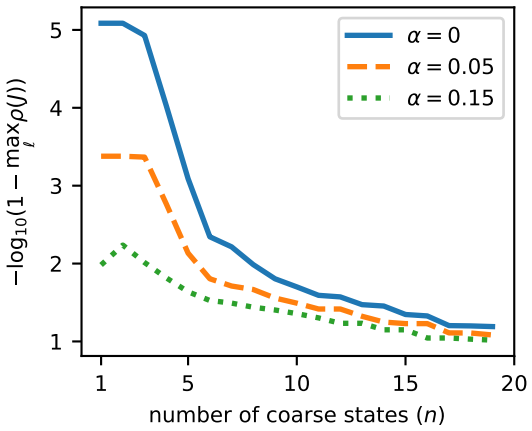

FIG. 6.2. *The maximum of $\rho(J(\mu))$ over all $l = 0, \ldots, \lfloor 100/n \rfloor$ for $n = 1, \ldots, 20$ for the uniform grids of coarse states defined in* (6.5). *The curve labeled $\alpha = 0$ is for the reversible chain of section* 6.3. *The other curves are for the irreversible chains in section* 6.4.

eigenvectors of P to choose good coarse states for this problem. For any $n \in \mathbb{N}$ and $\ell \in \{0, \ldots, \lfloor 100/n \rfloor\}$, we define the uniform grid of coarse states

$$S_J = \left\{ \left\lfloor J\frac{100}{n} \right\rfloor + \ell, \ldots, + \left\lfloor (J+1)\frac{100}{n} \right\rfloor + \ell - 1 \right\} \tag{6.5}$$

for $J = 0, \ldots, n-2$, and

$$S_n = \{0, \ldots, \ell - 1\} \cup \left\{ \left\lfloor (n-1)\frac{100}{n} \right\rfloor, \ldots, 99 \right\}.$$

For example, for $n = 2$ and $\ell = 5$, the coarse states would be $S_0 = \{5, \ldots, 54\}$ and $S_1 = \{0, \ldots, 4\} \cup \{55, \ldots, 99\}$. In Figure 6.2, we report the maximum of $\rho(J(\mu))$ over all $l = 0, \ldots, \lfloor 100/n \rfloor$ for $n = 1, \ldots, 20$. We see a clear decreasing trend with a growing number of coarse states.

We now investigate the dependence of $\rho(J(\mu))$ on the choice of coarse states in more detail. We compute the spectral radius $\rho(J(\mu))$ and our upper bounds for a family of coarse states of the form

$$S_1 = \{0, \ldots, \ell\} \quad \text{and} \quad S_2 = \{\ell + 1, \ldots, 99\}, \tag{6.6}$$

with $\ell = 0, \ldots, 98$. We display the results in Figure 6.3. Different locations ℓ of the boundary between coarse states result in different angles θ, depending on how well v_2' can be approximated by vectors that are constant on the coarse states. We expect θ to be small when the boundary between the coarse states coincides with the boundary between the basins of attraction, and this happens when $\ell = 57$. Note that when ℓ is close to 57, $\rho(J(\mu)) \approx \sqrt{\lambda_3}$, and it is as if one has eliminated the larger eigenvalue $\sqrt{\lambda_2}$. When ℓ is far from 57, $\rho(J(\mu)) \approx \sqrt{\lambda_2}$, and IAD will converge at approximately the same rate as the power method. Note that both of the upper bounds in Theorem 5.4 yield precise estimates of $\rho(J(\mu))$, but the norm bound (5.4) is so precise as to be indistinguishable from the spectral radius $\rho(J(\mu))$.

Note that the optimal coarse states in the family (6.6) considered above coincide with the basins of attraction of V. We wish to emphasize that it is not in general

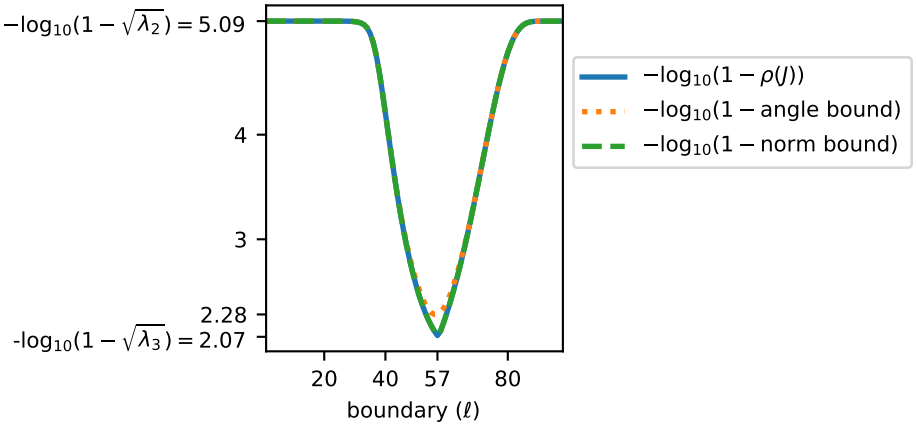

FIG. 6.3. *The spectral radius $\rho(J(\mu))$, the norm upper bound* (5.4)*, and the angle upper bound* (5.5) *for the reversible, one-dimensional chain. Each of these numbers x is very close to one for all values of ℓ, so we plot $-\log_{10}(1-x)$. The variable ℓ on the horizontal axis relates to the definition of the coarse states; cf.* (6.6)*. The spectral radius and the norm bound are indistinguishable in this figure.*

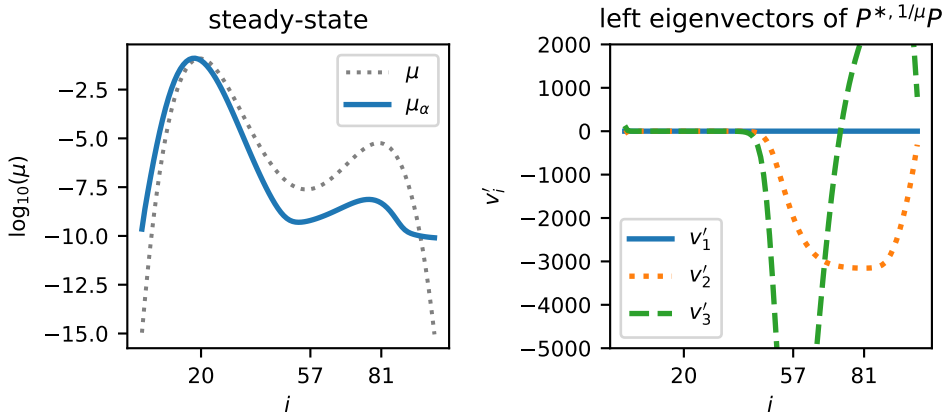

FIG. 6.4. *Left: the steady state distribution μ_α of P_α for $\alpha = 0.05$. For comparison, we have included the discrete Boltzmann distribution μ as well. Right: the first three left eigenvectors of $P_\alpha^{*,1/\mu_\alpha} P_\alpha$ for $\alpha = 0.05$.*

necessary to choose the coarse states to be the basins of attraction. It is only necessary that the leading left eigenvectors of P be well-approximated by functions that are constant on the coarse states. Note that this will be true whenever the coarse states are sufficiently refined.

6.4. IAD for irreversible chains on a one-dimensional grid. We now consider irreversible perturbations of the last process. Define

$$P_\alpha = (1-\alpha)P + \alpha W,$$

where P is as above, W is the right shift matrix (6.4), and $\alpha \in [0,1]$. Let μ_α be the steady state of P_α. We compute $\rho(J(\mu_\alpha))$ and the upper bounds for the families of coarse states defined above in section 6.3. We will see that our bounds are not as precise for irreversible chains as reversible chains. For $\alpha = 0.05$, our bounds

TABLE 6.2
The spectral radius $\rho(\hat{P}_\alpha)$ for $\alpha = 0, 0.05, 0.15$. Note that for $\alpha = 0$, P_α is simply the reversible chain defined in section 6.3. Observe that for $\alpha = 0.15$, the chain converges very quickly compared with $\alpha = 0$.

α	$\rho(\hat{P}_\alpha)$	$-\log_{10}(1-\rho(\hat{P}_\alpha))$
0	0.999992	5.09
0.05	0.999581	3.38
0.15	0.989564	1.98

overestimate the true rate of convergence but correctly predict the dependence of the rate on the choice of coarse states. For $\alpha = 0.15$, our bounds do not seem to yield any useful information about the dependence of the rate of convergence on the coarse states. However, we note that $P_{0.15}$ is not metastable: we have $\rho(\hat{P}_{0.15}) \approx 0.99$ compared with $\rho(\hat{P}) \approx 0.99999$; cf. Table 6.2. Therefore, one does not need a sophisticated method like IAD to estimate the steady state of a kernel like $P_{0.15}$, so $P_{0.15}$ is not of much interest as a test case for IAD. We include results for $\alpha = 0.15$ simply to illustrate the limitations of our theory.

In Figure 6.2, we report the maximum of $\rho(J(\mu_\alpha))$ over all $l = 0, \ldots, \lfloor 100/n \rfloor$ for $n = 1, \ldots, 20$. We see a clear decreasing trend with a growing number of coarse states. However, note that the spectral radius does not decrease monotonically with the number of coarse states for $\alpha = 0.15$. In particular, for some choices of coarse states with $n = 2$, $\rho(J(\mu_{0.15}))$ is larger than $\rho(\hat{P}_{0.15})$. For such poor choices of coarse states, IAD would converge more slowly than the power method.

In Figure 6.5, we report $\rho(J(\mu_\alpha))$ and the upper bounds for the family of shifted coarse states (6.6) for $\alpha = 0.05$. We report the steady state μ_α and the three leading left eigenvectors of P_α in Figure 6.5. Note the similarity with the eigenvectors of P_0 in Figure 6.1. Our upper bounds correctly predict the dependence of $\rho(J(\mu_\alpha))$ on ℓ. However, note that when ℓ is far from the optimal value, $\rho(J(\mu_\alpha))$ is almost the same as $\rho(\hat{P}_\alpha)$, which is the asymptotic rate of convergence of the power method. Our estimates in Theorem 5.4 predict the slower rate of convergence $\sqrt{\lambda_2}$, which is the contraction constant of the power method in $\ell^2(1/\mu)$. Recall that all of our upper bounds on $\rho(J(\mu))$ are in fact upper bounds on $\|(I - \Pi(\mu))J(\mu)\|_{1/\mu}$. Note that although our bounds are generally quite close to $\|(I - \Pi(\mu))J(\mu)\|_{1/\mu}$, they sometimes significantly overestimate $\rho(J(\mu))$; cf. the discussion after the statement of Theorem 5.4.

In Figure 6.6, we report $\rho(J(\mu_\alpha))$ and the upper bounds for the family of shifted coarse states (6.6) for $\alpha = 0.15$. Here, our upper bounds do not seem to yield any useful information about the dependence of the convergence rate on the choice of coarse states. We propose that more precise estimates based on the exact formula for the spectral radius given in Theorem 5.5 could be developed to understand the rate of convergence in this case. We leave this for future work. Note also that for some values of ℓ, we have $\rho(J(\mu_\alpha)) > \rho(\hat{P}_\alpha)$, which indicates that IAD would converge more slowly than the power method.

6.5. IAD for a metastable chain on a two-dimensional grid. We now consider a metastable chain on a two-dimensional grid. Molecular models and other models in computational chemistry usually involve stochastic processes on spaces having thousands or millions of dimensions. Of course, when the dimension is so high, one cannot expect to cover space by a uniform grid of coarse states. Therefore, in many sampling strategies, one chooses a low-dimensional coordinate to discretize.

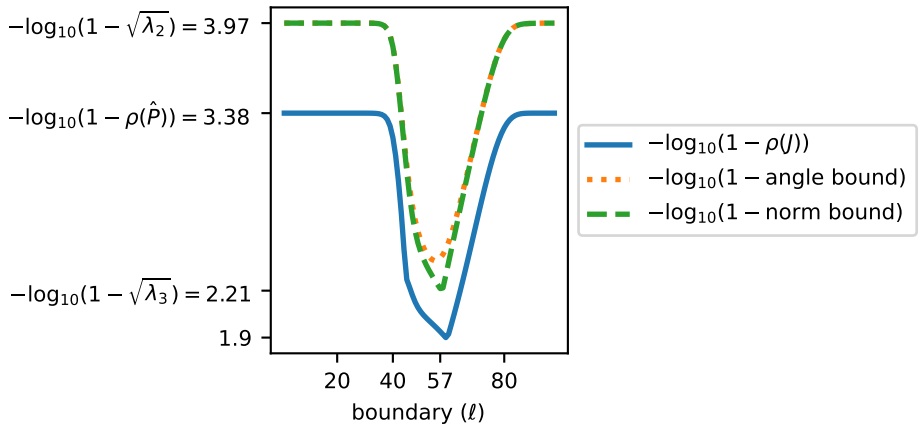

FIG. 6.5. *The spectral radius $\rho(J(\mu_\alpha))$, the norm upper bound* (5.4)*, and the angle upper bound* (5.5) *with $k=2$ for the irreversible, one-dimensional chain P_α with $\alpha=0.05$. Each of these numbers x is very close to one for all values of ℓ, so we report $-\log_{10}(1-x)$. The variable ℓ on the horizontal axis relates to the definition of the coarse states; cf.* (6.6).

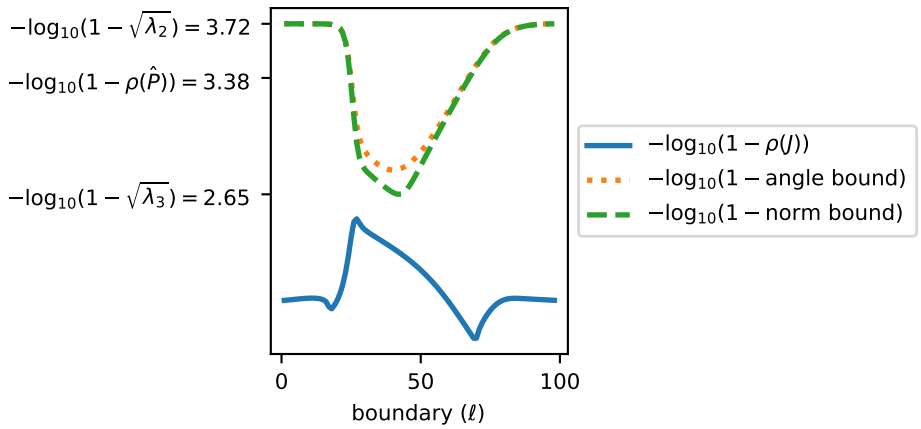

FIG. 6.6. *The spectral radius $\rho(J(\mu_\alpha))$, the norm upper bound* (5.4)*, and the angle upper bound* (5.5) *with $k=2$ for the irreversible, one-dimensional chain P_α with $\alpha=0.15$. Each of these numbers x is very close to one for all values of ℓ, so we report $-\log_{10}(1-x)$. The variable ℓ on the horizontal axis relates to the definition of the coarse states; cf.* (6.6).

We demonstrate for a model two-dimensional system that one can attain a significant reduction in the rate of convergence by discretizing a single variable into a small number of coarse states.

Define $V:\mathbb{R}^2\to\mathbb{R}$ by

$$V(x,y) = 3\exp\left(-x^2-\left(y-\frac{1}{3}\right)^2\right) - 3\exp\left(-x^2-\left(y^2-\frac{5}{3}\right)^2\right) - 5\exp\left(-(x-1)^2-y^2\right) - 5\exp\left(-(x+1)^2-y^2\right).$$

See Figure 6.7. This simple potential function was proposed for a study of reaction rates in [28]. Let μ be the discrete Boltzmann distribution on a two-dimensional grid with V as above and with $[a,b]\times[c,d] = [-1.7,1.7]\times[-1.7,2]$, $N=50$, and

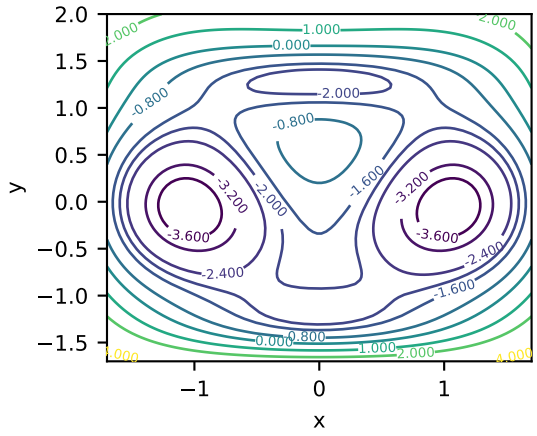

FIG. 6.7. *A Contour map of the potential function V. Note the two global minima on the left and right and the local minimum along a path between the other two at the top.*

TABLE 6.3
The largest five eigenvalues of $P^{*,1/\mu}P$ for the reversible, two-dimensional chain P of section 6.5. We report $\sqrt{\lambda_k}$ instead of λ_k, since it is $\sqrt{\lambda_k}$ that appears in Theorem 5.4.

k	$\sqrt{\lambda_k}$
1	1
2	0.999997
3	0.999488
4	0.997511
5	0.994219

$T = 1/4$. Let P be the corresponding reversible Markov chain. The definitions of the discrete Boltzmann distribution and the reversible dynamics are analogous to the one-dimensional case. See Appendix F for details. We report the largest five eigenvalues of P in Table 6.3, and we display the left eigenvectors v_2' and v_3' in Figures 6.8 and 6.9.

Define the coarse states

$$S_I = \left\{ (i,j) \in \{1,\dots,N\}^2 : \left\lfloor \frac{i}{3} \right\rfloor = j \right\} \text{ for } I = 0,1,2. \tag{6.7}$$

Here, P is a Markov chain on the state space $\{1,\dots,50\} \times \{1,\dots,50\}$, and the coarse states discretize only the first variable, not the second. The outlines of the coarse states are visible in the left panel of Figure 6.8. We report $\rho(J(\mu))$, the norm upper bound (5.4), and the angle upper bound (5.5) for $k = 2$ and $k = 3$ for this choice of coarse states in Table 6.4. Note that $\sin^2(\theta) \approx 0.002$ is quite small. We display the eigenvector v_2' and its best approximation in the $\ell^2(\mu)$-norm in Figure 6.8. Although the two vectors may not appear to be aligned, they are in fact very close in the $\ell^2(\mu)$-norm, since the regions where they differ have very small probability under μ and the maximum size of the difference between the vectors is not large in comparison to the very small probability. Observe that in this case, even with very few coarse states that discretize only a single dimension, we have in effect eliminated the largest eigenvalue $\sqrt{\lambda_2}$, and the rate of convergence of IAD is essentially equal to $\sqrt{\lambda_3}$. Note that $\sin^2(\theta)$ is large for $k = 3$ in this case, so v_3' is not well approximated by a vector that is constant on the coarse states; cf. Figure 6.9.

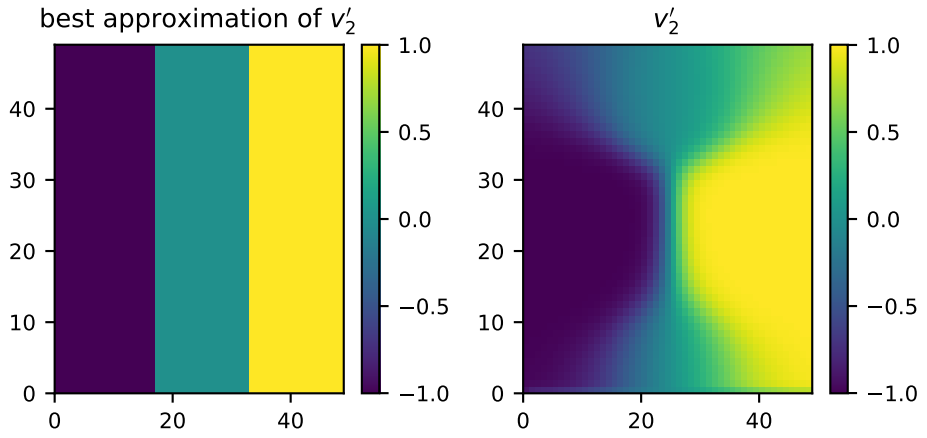

FIG. 6.8. *Right: the second left eigenvector v_2' of $P^{*,1/\mu}P$. Left: the best approximation $\Pi(\mu)^{\mathrm{t}}v_2'$ of v_2' in the $\ell^2(\mu)$-norm by a function that is constant on the one-dimensional grid of coarse states* (6.7). *Note that this figure could also be interpreted as a depiction of the coarse states themselves.*

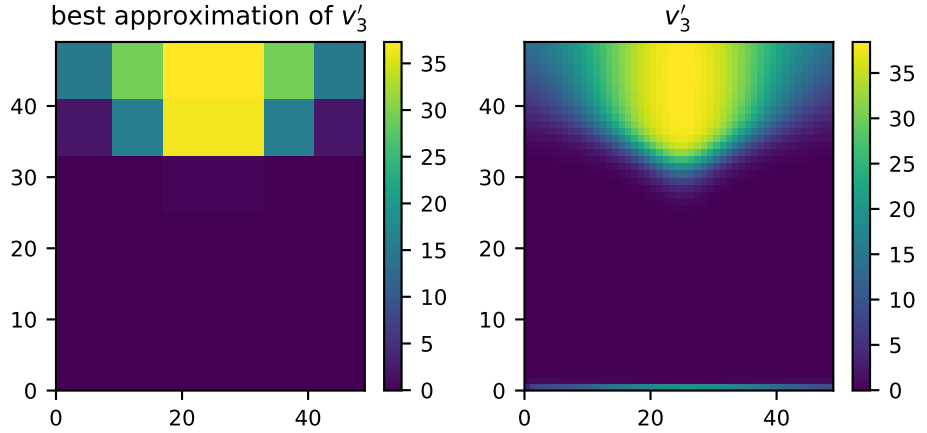

FIG. 6.9. *Right: the third left eigenvector v_3' of $P^{*,1/\mu}P$. Left: the best approximation $\Pi(\mu)^{\mathrm{t}}v_3'$ of v_3' in the $\ell^2(\mu)$-norm by a function that is constant on the two-dimensional grid of coarse states* (6.7).

We now test a six-by-six, two-dimensional grid of coarse states that refines the one-dimensional grid of three coarse states used above. For all $I, J \in \{0, \ldots, 5\}$, we let

$$S_{IJ} = \left\{(x, y) \in \{1, \ldots, N\}^2 : \left\lfloor \frac{x}{6} \right\rfloor = I \text{ and } \left\lfloor \frac{y}{6} \right\rfloor = J \right\}. \tag{6.8}$$

One can see the outlines of some of these coarse states in the left panel of Figure 6.9. We report $\rho(J(\mu))$ and the upper bounds for this choice of coarse states in Table 6.4. Note that both the spectral radius and the upper bounds are smaller, as expected for a refinement of the coarse states given the discussion following Corollary 5.6. For our six-by-six grid, $\sin^2(\theta)$ for $k = 3$ is much smaller than for the one-dimensional grid. However, it is not small enough that the angle upper bound for $k = 3$ is actually lower than for $k = 2$. In fact, the interpolation between $\sqrt{\lambda_{k+1}}$ and $\sqrt{\lambda_2}$ in (5.5) is very steep, and one needs a very small value of $\sin^2(\theta)$ for the upper bound to approximate $\sqrt{\lambda_{k+1}}$. Nonetheless, refining the set of coarse states reduces $\rho(J(\mu))$ significantly, and this is correctly predicted by the norm upper bound (5.4).

TABLE 6.4

The spectral radius $\rho(J(\mu))$, the norm upper bound (5.4)*, and the angle upper bound* (5.5) *together with* $\sin^2(\theta)$ *for* $k = 2, 3$*. The first column is for the one-dimensional grid* (6.7) *and the second column is for the six-by-six, two-dimensional grid* (6.8)*.*

	1-d grid	2-d grid
$\rho(J(\mu))$	0.999410	0.987327
norm bound	0.999410	0.987327
$\sin^2(\theta)$, $k=2$	0.002382	0.000143
angle bound, $k=2$	0.999650	0.999502
$\sin^2(\theta)$, $k=3$	0.854170	0.031745
angle bound, $k=3$	0.999997	0.999920

7. Conclusion. Our work here was motivated by a desire to understand the robustness and efficiency of methods such as nonequilibrium umbrella sampling (NEUS) [36], exact milestoning [1], and injection measures [11] for calculating nonequilibrium (or equilibrium) steady states in statistical physics. We have studied IAD as a simple model of this class of methods. We explain why it may be possible to use methods similar to IAD to efficiently compute steady states of molecular models and how one might choose the coarse states in practice to optimize efficiency. For reversible processes, we conclude that one should choose coarse states so that the leading left eigenvectors of P are well-approximated in the $\ell^2(\mu)$-norm by vectors that are constant on the coarse states. Since error is measured in the $\ell^2(\mu)$-norm, regions of low probability will not have a significant influence on the approximation quality unless some of the leading eigenvectors are concentrated in those regions. This means that in some cases a very naive choice of coarse states can be efficient. For irreversible processes, our conclusions are similar but not so definite. For some very irreversible processes, our upper bounds do not yield much information about the dependence of the asymptotic rate of convergence on the choice of strata. Although our primary interests lie in statistical physics, our results are general, and we hope others will apply them to understand the performance of IAD in other contexts.

Our work does not address all important points. We show only local convergence, not global. We focus primarily on estimates of the asymptotic rate of convergence, ignoring preasymptotic phenomena. We recall that NEUS and similar methods are stochastic evolving particle systems that approximate IAD; we do not consider issues related to the particle approximation. We leave these points for future work.

Appendix A. Computing the coarse steady state. We compute the coarse steady state $z(C(\mu^k))$ by the following algorithm, which was introduced in [10].

Assume that $C(\mu^k)$ is irreducible as guaranteed by Lemma 3.6. The user must specify an error tolerance $\tau > 0$ and an exponent $m \in \mathbb{N}$. In our numerical experiments in sections 6.3 and 6.5, we take $\tau = 10^{-9}$ and $m = 2^{15}$. We compute $z(C(\mu^k))$ by the following procedure:

1. Set $\bar{C} = \frac{1}{2}(I + C(\mu^k))$.
2. Use the algorithm described in section 5 of [13] to compute an initial approximation $\tilde{z}$ to the steady state $z(C(\mu^k))$. That is, compute the QR-factorization of $I - \bar{C}^{\mathrm{t}}$, and let $\tilde{z}$ be the last column of Q renormalized to have a sum eq.
3. Refine the initial approximation $\tilde{z}$ using a version of the power method:
   (a) Set $z^{\mathrm{old}} = \tilde{z}$.

(b) Calculate

$$z^{\text{new}} = \bar{C}^m z^{\text{old}}.$$

(c) If

$$\max_{i=1,\ldots,n} \frac{|z_i^{\text{new}} - \bar{C} z_i^{\text{new}}|}{z_i^{\text{new}}},$$

then return z^{new} to the user. Otherwise, set $z^{\text{old}} = z^{\text{new}}$ and go to step (b) above.

In practice, we find that direct methods (such as the algorithm of [13] based on the QR-factorization of $I - \bar{C}^{\text{t}}$) for calculating the coarse steady state are often not sufficiently accurate due to floating-point error. When direct methods produce an inaccurate result, applying a few power method iterations has usually produced a much better estimate of the steady state in our experience. For efficiency, instead of multiplying by $\bar{C}$ in each step of the power method, we first calculate $\bar{C}^m$ for some large m. We then multiply by $\bar{C}^m$ in each step. If we choose $m = 2^j$ for some $j \in \mathbb{N}$, then computing $\bar{C}^m$ requires only squaring a matrix j times. This may be much less costly than performing m steps of the power method and multiplying by $\bar{C}$ in each step. See [10] for a more detailed explanation, including an example that explains why the power method step of this algorithm can be beneficial.

Appendix B. Well-posedness of IAD. Here, we prove that IAD is well-posed under our assumptions, and we give some examples to illustrate what can go wrong when our assumptions do not hold.

B.1. Proof of Lemma 3.6. We show that IAD is well-posed if P is irreducible and $\mu^0 > 0$.

Proof. First, we show that if P is irreducible and $\nu > 0$, then $C(\nu)$ is irreducible. We prove the contrapositive, showing that if $\nu > 0$ and $C(\nu)$ is reducible, then P is reducible. If $\nu > 0$ and $C(\nu)$ is reducible, then there is a partition of the coarse states

$$\{1, \ldots, n\} = A \cup B$$

into disjoint and nonempty sets A and B so that for all $a \in A$ and $b \in B$

$$C(\nu)_{ba} = APD(\nu)_{ba} = \sum_{\substack{i \in S_a \\ j \in S_b}} \frac{\nu_i}{A\nu_a} P_{ji} = 0.$$

Therefore, since $\nu > 0$, $P_{ji} = 0$ for all $i \in S_a$ and $j \in S_b$. Now define

$$\mathcal{A} = \cup_{a \in A} S_a \quad \text{and} \quad \mathcal{B} = \cup_{b \in B} S_b.$$

The sets $\mathcal{A}$ and $\mathcal{B}$ are a partition of Ω into disjoint and nonempty sets, and we have $P_{ji} = 0$ for any $i \in \mathcal{A}$ and $j \in \mathcal{B}$. It follows that P must be reducible. We conclude that if $\nu > 0$ and P is irreducible, then $C(\nu)$ must be irreducible.

Now we show that if $\mu^k > 0$, then $\mu^{k+1} > 0$. To verify that $\mu^{k+\frac{1}{2}} > 0$, we observe that $C(\mu^k)$ is irreducible when $\mu^k > 0$ by the previous paragraph, and so the steady state $z(C(\mu^k))$ is unique and positive by the Perron–Frobenius theorem. It follows that $\mu^{k+\frac{1}{2}} = D(\mu^k) z(C(\mu_k)) > 0$. Moreover, when P is irreducible, $\nu > 0$ implies

$P\nu > 0$. To see this, observe that if $\nu > 0$, then for any $i \in \Omega$ with $P\nu_i = 0$, $P_{ij} = 0$ for all $j \neq i$. Thus, P is reducible if $P\nu$ is not positive. Therefore, $\mu^{k+1} = P\mu^{k+\frac{1}{2}} > 0$ if $\mu^k > 0$, and by induction $\mu^0 > 0$ implies $\mu^k > 0$ for all $k \in \mathbb{N}$. This concludes the proof that the iterates μ^k are well-defined. ∎

B.2. Examples motivating our assumptions. We now give some pathological examples to motivate our assumptions that $\mu^0 > 0$ and that $P^{\mathrm{t}}P$ is irreducible.

B.2.1. Positive initial condition ($\mu^0 > 0$). We assume $\mu^0 > 0$, since if μ^0 is not positive, then $C(\mu^0)$ may be reducible even when P and $P^{\mathrm{t}}P$ are irreducible and aperiodic. For example, consider the chain on $\Omega = \{1, 2, 3\}$ with transition matrix

$$P = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 1 & \frac{1}{3} & 1 \\ 0 & \frac{1}{3} & 0 \end{pmatrix}.$$

Here, P and $P^{\mathrm{t}}P$ are irreducible and aperiodic. Now define the coarse states

$$S_1 = \{1, 2\} \quad \text{and} \quad S_2 = \{3\}.$$

For $\mu^0 = \frac{1}{2}(\delta_1 + \delta_3) = (\frac{1}{2}, 0, \frac{1}{2})^{\mathrm{t}}$, we have

$$C(\mu^0) = APD(\mu^0) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

which is reducible.

B.2.2. $P^{\mathrm{t}}P$ irreducible. We assume that $P^{\mathrm{t}}P$ is irreducible to prove local convergence of IAD. It was observed in [24, Example 2] that IAD is not locally convergent for

$$P = \begin{pmatrix} 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

with the strata $S_1 = \{1, 2\}$ and $S_2 = \{3, 4\}$. However, P is irreducible and aperiodic, so the Markov chain with transition matrix P is convergent. The reader will note that $P^{\mathrm{t}}P$ is reducible, so this is an example where the power method is convergent but not a strict contraction in $\ell^2(1/\mu)$.

Appendix C. Proofs of results stated in section 4.1.

C.1. Proof of Lemma 4.2. We begin with a proof of Lemma 4.2, which shows that the time reversal of P is its adjoint $P^{*,1/\mu}$.

Proof. For simplicity, we write $\langle , \rangle$ for $\langle , \rangle_{1/\mu}$ and P^* for $P^{*,1/\mu}$. Note that for any $x, y \in \mathbb{R}^N$, we have

$$\begin{aligned} \langle x, Py \rangle &= \langle x, \operatorname{diag}(1/\mu) P y \rangle_{\mathbb{1}} \\ &= \langle \operatorname{diag}(\mu) P^{\mathrm{t}} \operatorname{diag}(1/\mu) x, \operatorname{diag}(1/\mu) y \rangle_{\mathbb{1}} \\ &= \langle \operatorname{diag}(\mu) P^{\mathrm{t}} \operatorname{diag}(1/\mu) x, y \rangle, \end{aligned}$$

so

$$P^* = \operatorname{diag}(\mu) P^{\mathrm{t}} \operatorname{diag}(1/\mu). \tag{C.1}$$

Therefore, $P^*_{ij} = \frac{P_{ji}\mu^j}{\mu_i}$, which is exactly the time reversal of P. See [27, Theorem 1.9.1] for the definition of the time reversal and its properties. (Remember that P is *column* stochastic, so the time reversal here takes a slightly different form than for the row stochastic matrices of [27].) The time reversal P^* is column stochastic and has the same invariant distribution μ as P, since time reversals always have these properties. ∎

C.2. Proof of Lemma 4.4. We now prove convergence of the power method when both P and $P^{\mathrm{t}}P$ are irreducible. Note that if $P^{\mathrm{t}}P$ is irreducible, then the stochastic matrix $P^{*,1/\mu}P$ is irreducible, and this is the essential fact in our proof that the power method is strictly contracting in the $\ell^2(1/\mu)$-norm.

Proof. To simplify notation, for any $M \in \mathbb{R}^{N\times N}$, we write M^* for $M^{*,1/\mu}$ and M^{t} for the transpose. We let $\|\cdot\|$ and $\langle,\rangle$ denote the $\ell^2(1/\mu)$-norm and the inner product. Since P is column stochastic and ν^0 is a probability vector, ν^k is a probability vector for all $k \in \mathbb{N}$, so $\mathbb{1}^{\mathrm{t}}\nu^k = 1$. Therefore,

$$\begin{aligned}\nu^{k+1} - \mu &= P\nu^k - \mu \\ &= (P - \mu\mathbb{1}^{\mathrm{t}})\nu^k \\ &= (P - \mu\mathbb{1}^{\mathrm{t}})(\nu^k - \mu) \\ &= \hat{P}(\nu^k - \mu).\end{aligned}$$

Note that the third equality above follows since $P\mu = \mu$ and μ is a probability vector.

We now observe that for any $M \in \mathbb{R}^{N\times N}$, we have

$$\|M\| = \rho(M^*M)^{\frac{1}{2}} = \|M^*M\|^{\frac{1}{2}}. \tag{C.2}$$

In particular, $\|\hat{P}\| = \rho(\hat{P}^*\hat{P})^{\frac{1}{2}} = \|\hat{P}^*\hat{P}\|^{\frac{1}{2}}$. The analogous result

$$\|M\|_{\mathbb{1}} = \rho(M^{\mathrm{t}}M)^{\frac{1}{2}} = \|M^{\mathrm{t}}M\|_{\mathbb{1}}^{\frac{1}{2}}$$

for the uniformly weighted $\ell^2(\mathbb{1})$-inner product is well known, and the standard proof generalizes to any inner product space. Therefore, it will suffice to show that $\rho(\hat{P}^*\hat{P})^{\frac{1}{2}} < \sqrt{\lambda_2}$. Observe that

$$\begin{aligned}\hat{P}^*\hat{P} &= (P^* - \mu\mathbb{1}^{\mathrm{t}})(P - \mu\mathbb{1}^{\mathrm{t}}) \\ &= P^*P - P^*\mu\mathbb{1}^{\mathrm{t}} - \mu\mathbb{1}^{\mathrm{t}}P + \mu\mathbb{1}^{\mathrm{t}} \\ &= P^*P - \mu\mathbb{1}^{\mathrm{t}},\end{aligned}$$

so $\rho(\hat{P}^*\hat{P}) = \lambda_2$ by the diagonalization (4.2) of P^*P. Thus, $\|\hat{P}\| = \sqrt{\lambda_2}$.

By (C.1), P^*P is irreducible if and only if $P^{\mathrm{t}}P$ is irreducible. Moreover, $\|\hat{P}\| = \sqrt{\lambda_2} < 1$ if P^*P is irreducible by the Perron–Frobenius theorem; cf. [26, section 8.3, p. 673]. Therefore, $P^{\mathrm{t}}P$ irreducible is a sufficient condition for $\|\hat{P}\| < 1$. To see that it is also a necessary condition, we will prove that if $P^{\mathrm{t}}P$ is reducible, then $\|\hat{P}\| \geq 1$. If $P^{\mathrm{t}}P$ is reducible, then there exist nonempty, disjoint subsets A and B of Ω so that

$P^t P_{ij} = 0$ for all $i \in A$ and $j \in B$. Since $P^{\mathrm{t}} P$ is symmetric, we also have $P^{\mathrm{t}} P_{ij} = 0$ for $i \in B$ and $j \in A$. Therefore, $P^* P$ admits a block decomposition of the form

$$P^*P = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \end{array},$$

where P_1 and P_2 are both column stochastic matrices. Let v_1 and v_2 be steady state probability vectors of P_1 and P_2, respectively. Define $V_1 = (v_1, 0)^{\mathrm{t}}$ and $V_2 = (0, v_2)^{\mathrm{t}}$. We have $V_1 - V_2 \neq 0$, and

$$\|\hat{P}(V_1 - V_2)\|^2 = \|P(V_1 - V_2)\|^2 = \langle P^*P(V_1 - V_2), (V_1 - V_2)\rangle = \|V_1 - V_2\|^2,$$

which verifies $\|\hat{P}\| \geq 1$. ∎

Appendix D. Proofs of results stated in section 4.2.

D.1. Proof of Lemma 4.5. We begin with a proof of Lemma 4.5, our reformulation of the steady state eigenproblem $Qz(Q) = z(Q)$ as a linear system.

Proof. Observe that $x = z(Q)$ solves

$$(I - Q + vw^{\mathrm{t}})x = vw^{\mathrm{t}} z(Q), \tag{D.1}$$

since $Qz(Q) = z(Q)$. If we can show that $x = z(Q)$ is the unique solution of (D.1), then $I - Q + vw^{\mathrm{t}}$ is invertible and the result follows.

Suppose to the contrary that $u \neq z(Q)$ also solves (D.1). Then we have

$$\mathbb{1}^{\mathrm{t}}(I - Q + vw^{\mathrm{t}})u = \mathbb{1}^{\mathrm{t}} vw^{\mathrm{t}} u = \mathbb{1}^{\mathrm{t}} vw^{\mathrm{t}} z(Q),$$

since $\mathbb{1}^{\mathrm{t}}(I - Q) = 0$ for any column stochastic Q. Therefore, $w^{\mathrm{t}} u = w^{\mathrm{t}} z(Q)$, since we assume $\mathbb{1}^{\mathrm{t}} v \neq 0$. It follows, again by (D.1), that

$$(I - Q)u = 0.$$

Moreover, since we assume $w^{\mathrm{t}} z(Q) \neq 0$, $u \neq z(Q)$ and $w^{\mathrm{t}} u = w^{\mathrm{t}} z(Q)$ imply

$$u \notin \operatorname{span}\{z(Q)\}.$$

Now recall that since Q is irreducible, $z(Q) > 0$ by the Perron–Frobenius theorem. Thus, for some $\varepsilon > 0$, $z(Q) + \varepsilon u > 0$, $z(Q) + \varepsilon u \notin \operatorname{span}\{z(Q)\}$, and

$$Q(z(Q) + \varepsilon u) = z(Q) + \varepsilon u.$$

This contradicts uniqueness of the stationary distribution of Q, so $z(Q)$ is the unique solution of (D.1), and $I - Q + vw^{\mathrm{t}}$ is invertible. ∎

D.2. Proof of Lemma 4.6. We prove our error propagation formula for the coarse correction step.

Proof. First, note that

$$A(I - P + \mu \mathbb{1}^{\mathrm{t}})D(\nu) = I - C(\nu) + A\mu \mathbb{1}^{\mathrm{t}}$$

is invertible by Lemma 4.5, since $C(\nu)$ is irreducible when $\nu > 0$ by Lemma 3.6.

We prove formula (4.6) for the error after the coarse correction. Using our reformulations (4.4) and (4.5) of the steady state problems, we have

$$\begin{aligned}\mu^{k+\frac{1}{2}} &= D(\mu^k)z(C(\mu^k))\\ &= D(\mu^k)(I - C(\mu^k) + A\mu\mathbb{1}^{\mathrm{t}})^{-1}A\mu\\ &= D(\mu^k)(A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\mu^k))^{-1}A\mu\\ &= D(\mu^k)(A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\mu^k))^{-1}A(I - P + \mu\mathbb{1}^{\mathrm{t}})\mu\\ &= S(\mu^k)\mu.\end{aligned}$$

Now, since $D(\mu^k)\mu^k = \mu^k$, we have

$$\begin{aligned}S(\mu^k)\mu^k &= D(\mu^k)(A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\mu^k))^{-1}A(I - P + \mu\mathbb{1}^{\mathrm{t}})\mu^k\\ &= D(\mu^k)(A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\mu^k))^{-1}A(I - P + \mu\mathbb{1}^{\mathrm{t}})D(\mu^k)\mu^k\\ &= \mu^k.\end{aligned}$$

Therefore,

$$\mu^{k+\frac{1}{2}} - \mu = -(I - S(\mu^k))\mu = (I - S(\mu^k))(\mu^k - \mu),$$

as desired.

It remains to show that $S(\nu)$ is a projection on $\mathrm{Rg}(D(\nu))$. For convenience, we write S for $S(\nu)$, D for $D(\nu)$, and C for $C(\nu)$. To see that S is a projection, observe that

$$\begin{aligned}S^2 &= D[A(I - \hat{P})D]^{-1}(A(I - \hat{P})D)[A(I - \hat{P})D]^{-1}A(I - \hat{P})\\ &= D[A(I - \hat{P})D]^{-1}A(I - \hat{P})\\ &= S.\end{aligned}$$

To see that $\mathrm{Rg}(S) = \mathrm{Rg}(D)$, first observe that

$$\mathrm{Rg}(A(I - \hat{P})) = \mathrm{Rg}(A) = \mathbb{R}^n,$$

since $(I - \hat{P})$ is invertible by Lemma 4.5. Therefore, since $A(I - \hat{P})D = I - C + A\mu\mathbb{1}^{\mathrm{t}}$ is also invertible by Lemma 4.5,

$$\mathrm{Rg}([A(I - \hat{P})D]^{-1}A(I - \hat{P})) = \mathbb{R}^n,$$

and it follows that $\mathrm{Rg}(S) = \mathrm{Rg}(D)$. ∎

D.3. IAD as an algebraic multigrid method. Here, we explain that IAD is more or less an adaptive algebraic multigrid method. Roughly similar observations appear in [16]. Recall that the steady state μ is the unique solution x of the linear system of equations

$$(I - P + \mu\mathbb{1}^{\mathrm{t}})x = \mu;$$

cf. (4.4). Suppose that one were to try to solve this equation by algebraic multigrid with the restriction operator A. Typically, the prolongation operator would be the transpose of restriction. Suppose that instead one were to take an adjoint with respect

to some nonuniform inner product. For example, let $\nu \in \mathbb{R}^N$ be a positive probability vector, and define the prolongation operator to be the operator $A^\dagger \in \mathbb{R}^{N\times n}$ satisfying

$$\langle A\eta, w\rangle_{1/A\nu} = \langle \eta, A^\dagger w\rangle_{1/\nu}$$

for all $w \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^N$. By Lemma D.1, we have $A^\dagger = D(\nu)$.

The coarse grid correction step for a multigrid method with restriction operator A and prolongation $D(\nu)$ is

$$\begin{aligned}\mu^{k+\frac{1}{2}} &= \mu^k + D(\nu)(A(I-P+\mu\mathbb{1}^{\mathrm{t}})D(\nu))^{-1}A(\mu-(I-P+\mu\mathbb{1}^{\mathrm{t}})\mu^k)\\ &= \mu^k + D(\nu)(A(I-P+\mu\mathbb{1}^{\mathrm{t}})D(\nu))^{-1}A(I-P+\mu\mathbb{1}^{\mathrm{t}})(\mu-\mu^k)\\ &= \mu^k - S(\nu)(\mu^k-\mu),\end{aligned}$$

where $S(\nu)$ is the coarse projection defined in Lemma 4.6. Here, the residual is $\mu-(I-P+\mu\mathbb{1}^{\mathrm{t}})\mu^k$ and the coarse system matrix is $A(I-P+\mu\mathbb{1}^{\mathrm{t}})D(\nu)$. Note that if one chooses $\nu=\mu^k$, the coarse grid correction is equivalent with the coarse correction step of IAD. After the coarse grid correction in a multigrid method, one computes several steps of a smoothing iteration, often using some version of the Jacobi or Gauss–Seidel method. In IAD, one performs a step of the power method, which corresponds to using P as the smoothing matrix.

Note that IAD is not a true multigrid method, since the inner product and therefore the prolongation operator depend on the current approximation μ^k of μ. Thus, IAD is nonlinear and the standard theory of multigrid methods does not apply.

D.4. Proof of Lemma 4.9. We prove that $\Pi(\nu)$ is an $\ell^2(\nu)$-orthogonal projection. We begin by showing that A and $D(\nu)$ are adjoint in a certain sense.

LEMMA D.1. *For any $w \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^N$, we have*

$$\langle D(\nu)w, \eta\rangle_{1/\nu} = \langle w, A\eta\rangle_{1/A\nu}.$$

Proof. We have

$$\begin{aligned}\langle D(\nu)w,\eta\rangle_{1/\nu} &= \sum_{j=1}^N\sum_{k=1}^n w_k\nu(j|S_k)\eta_j\frac{1}{\nu_j}\\ &= \sum_{j=1}^N\sum_{k=1}^n w_k\frac{\mathbb{1}_{S_k}(j)}{A\nu_k}\eta j\\ &= \sum_{k=1}^n w_k\left(\sum_{j=1}^N \mathbb{1}_{S_k}(j)\eta_j\right)\frac{1}{A\nu_k}\\ &= \langle w, A\eta\rangle_{1/A\nu}.\end{aligned}$$

∎

We now prove Lemma 4.9, which verifies that $\Pi(\nu)$ is an orthogonal projection.

Proof. First, note that

$$AD(\nu) = I \in \mathbb{R}^{N\times N}.$$

Therefore,

$$\Pi(\nu)^2 = D(\nu)(AD(\nu))A = D(\nu)A = \Pi(\nu),$$

so $\Pi(\nu)$ is a projection.

To see that $\Pi(\nu)$ is orthogonal, note that by Lemma D.1, for any $\eta, \kappa \in \mathbb{R}^N$,

$$\begin{aligned}\langle \Pi(\nu)\eta, \kappa\rangle_{1/\nu} &= \langle D(\nu)A\eta, \kappa\rangle_{1/\nu} \\ &= \langle A\eta, A\kappa\rangle_{1/A\nu} \\ &= \langle \eta, D(\nu)A\kappa\rangle_{1/\nu} \\ &= \langle \eta, \Pi(\nu)\kappa\rangle_{1/\nu}.\end{aligned}$$

Therefore, $\Pi(\nu)^{*,1/\nu} = \Pi(\nu)$ so $\Pi(\nu)$ is an orthogonal projection in $\ell^2(1/\nu)$.

Finally, observe that $\Pi(\nu) = D(\nu)A$ is column stochastic, since both A and $D(\nu)$ map probability vectors to probability vectors. We have $\Pi(\nu)\nu = \nu$ directly from the definition of $\Pi(\nu)$, and $\Pi(\nu)$ is reversible, since it is self-adjoint with respect to the $\ell^2(1/\nu)$-inner product; cf. Lemma 4.2. ∎

D.5. Proof of Lemma 4.11. We begin our proof of Lemma 4.11 by computing the block decomposition of $J(\mu)$ with respect to $\Pi(\mu)$. The essential facts are summarized in the following simple lemma.

LEMMA D.2. *We have*

$$\Pi(\mu)S(\mu) = S(\mu) \quad \textit{and} \quad S(\mu)\Pi(\mu) = \Pi(\mu).$$

Proof. For convenience, we write Π for $\Pi(\mu)$ and S for $S(\mu)$. Since S is a projection with $\mathrm{Rg}(S) = \mathrm{Rg}(\Pi)$ by Lemma 4.6, we have

$$S\Pi = \Pi.$$

Similarly, $\Pi S = S$. ∎

As a consequence of Lemma D.2, we have the following block decomposition of $J(\mu)$.

LEMMA D.3. *We have*

$$J(\mu)\Pi(\mu) = 0$$

and

$$\Pi(\mu)J(\mu) = \Pi(\mu) - S(\mu).$$

Proof. For convenience, we write J for $J(\mu)$, Π for $\Pi(\mu)$, and S for $S(\mu)$. By Lemma D.2, $S\Pi = \Pi$, so

$$J\Pi = \hat{P}(I - S)\Pi = 0. \tag{D.2}$$

We also have

$$\begin{aligned}\Pi J &= \Pi\hat{P}(I - S) \\ &= \Pi\hat{P} - \Pi\hat{P}S \\ &= \Pi\hat{P} - D[A\hat{P}D(I - A\hat{P}D)^{-1}]A(I - \hat{P}) \\ &= \Pi\hat{P} - D[(I - A\hat{P}D)^{-1} - I]A(I - \hat{P}) \\ &= \Pi\hat{P} - D[(A(I - \hat{P})D)^{-1} - I]A(I - \hat{P}) \\ &= \Pi\hat{P} - S + \Pi(I - \hat{P}) \\ &= \Pi - S.\end{aligned}$$

The fifth inequality above follows since $AD = I \in \mathbb{R}^{n\times n}$. ∎

By Lemma D.3,

$$J(\mu) = J(\mu)(I - \Pi(\mu)). \tag{D.3}$$

This will be important in several of our proofs. It implies, for example, that

$$\sigma(J(\mu)) = \sigma(J(\mu)(I - \Pi(\mu)) = \sigma((I - \Pi(\mu))J(\mu)), \tag{D.4}$$

since $\sigma(AB) = \sigma(BA)$ for any matrices $A, B \in \mathbb{R}^{N\times N}$. We now proceed with the proof of Lemma 4.11.

Proof. For convenience, we write J for $J(\mu)$, Π for $\Pi(\mu)$, and S for $S(\mu)$. For $x \in \mathbb{R}^N$,

$$\begin{aligned}
\|Jx\|_\varepsilon^2 &= \|J(I-\Pi)x\|_\varepsilon^2 \\
&= \|(I-\Pi)J(I-\Pi)x\|_{1/\mu}^2 + \varepsilon\|\Pi J(I-\Pi)x\|_{1/\mu}^2 \\
&\le \left(\|(I-\Pi)J\|_{1/\mu}^2 + \varepsilon\|\Pi J\|_{1/\mu}^2\right)\|(I-\Pi)x\|_{1/\mu}^2 \\
&\le \left(\|(I-\Pi)J\|_{1/\mu}^2 + \varepsilon\|\Pi J\|_{1/\mu}^2\right)\|x\|_\varepsilon^2 \\
&= \left(\|(I-\Pi)J\|_{1/\mu}^2 + \varepsilon\|\Pi - S\|_{1/\mu}^2\right)\|x\|_\varepsilon^2.
\end{aligned}$$

The first equality holds by (D.3). The inequality in the second-to-last step holds, since $I - \Pi$ is an orthogonal projection and therefore $\|I - \Pi\|_{1/\mu} = 1$. The last equality holds by Lemma D.3. ∎

D.6. Proof of Theorem 4.12. We derive an upper bound on $\|(I - \Pi(\mu))J(\mu)\|_{1/\mu}$.

Proof. For convenience, we write J for $J(\mu)$, Π for $\Pi(\mu)$, etc. We write $\|\cdot\|$ for $\|\cdot\|_{1/\mu}$, $\langle,\rangle$ for $\langle,\rangle_{1/\mu}$, and $\hat{P}^*$ for $\hat{P}^{*,1/\mu}$. First, observe that since at least one coarse state contains more than one fine state, we have $I - \Pi \neq 0$. Since $J = J(I - \Pi)$ by (D.3), and since $I - \Pi$ is an orthogonal projection, we have

$$\|(I-\Pi)J\| = \|(I-\Pi)J(I-\Pi)\| = \max_{\substack{x\in \mathrm{Rg}(I-\Pi)\\ \|x\|=1}} \|(I-\Pi)Jx\|. \tag{D.5}$$

Now we claim

$$\|(I-\Pi)x\|^2 - \|(I-\Pi)Jx\|^2 = \langle (I-S)x, (I-\hat{P}^*\hat{P})(I-S)x\rangle \tag{D.6}$$

for all $x \in \mathbb{R}^N$. To prove this, observe that

$$\begin{aligned}
\|(I-\Pi)x\|^2 - \|(I-\Pi)Jx\|^2 &= \|(I-\Pi)x\|^2 - (\|Jx\|^2 - \|\Pi Jx\|^2) \\
&= \|(I-\Pi)x\|^2 - \|\hat{P}(I-S)x\|^2 + \|(\Pi - S)x\|^2 \\
&= \|(I-\Pi)x\|^2 - \|\hat{P}(I-S)x\|^2 + \|-S(I-\Pi)x\|^2 \\
&= \|(I-\Pi)x - S(I-\Pi)x\|^2 - \|\hat{P}(I-S)x\|^2 \\
&= \|(I-S)x\|^2 - \|\hat{P}(I-S)x\|^2 \\
&= \langle (I-S)x, (I-\hat{P}^*\hat{P})(I-S)x\rangle.
\end{aligned}$$

The first equality above follows since Π is an orthogonal projection by Lemma 4.9, and therefore $\|\Pi Jx\|^2 + \|(I-\Pi)Jx\|^2 = \|Jx\|^2$ by the Pythagorean theorem. The second follows since $\Pi J = \Pi - S$ by Lemma D.3. The third follows since $S\Pi = \Pi$ by Lemma D.2, so $\Pi - S = -S(I-\Pi)$. The fourth follows from the Pythagorean theorem since $\mathrm{Rg}(S) = \mathrm{Rg}(\Pi)$ and therefore $(I-\Pi)x$ and $-S(I-\Pi)x$ are orthogonal. The fifth follows since $(I-S)(I-\Pi) = I - S$ because $S\Pi = \Pi$ by Lemma D.2.

Combining the results of the last paragraph yields

$$\begin{aligned}\|(I-\Pi)J\|^2 &= \max_{\substack{x\in\mathrm{Rg}(I-\Pi)\\ \|x\|=1}} \|(I-\Pi)Jx\|^2 \\ &= \max_{\substack{x\in\mathrm{Rg}(I-\Pi)\\ \|x\|=1}} \|(I-\Pi)x\|^2 - \langle (I-S)x, (I-\hat{P}^*\hat{P})(I-S)x\rangle \\ &= 1 - \min_{\substack{x\in\mathrm{Rg}(I-\Pi)\\ x\neq 0}} \frac{\langle (I-S)x, (I-\hat{P}^*\hat{P})(I-S)x\rangle}{\|(I-\Pi)x\|^2}.\end{aligned}$$

We now make the change of variables $z = (I-S)x$ in the above minimization problem. By Lemma D.2,

$$(I-\Pi)x = (I-\Pi)(I-S)x = (I-\Pi)z.$$

Moreover, since $(I-S)(I-\Pi) = (I-S)$ by Lemma D.2, we have $(I-S)\,\mathrm{Rg}(I-\Pi) = \mathrm{Rg}(I-S)$. Also, $\mathrm{Rg}(I-S)$ has the same dimension as $\mathrm{Rg}(I-\Pi)$ because S and Π are both projections on $\mathrm{Rg}(D)$. Therefore, $(I-S)(\mathrm{Rg}(I-\Pi)\setminus\{0\}) = \mathrm{Rg}(I-S)\setminus\{0\}$. It follows that

$$\begin{aligned}\|(I-\Pi)J\|^2 &= 1 - \min_{\substack{x\in\mathrm{Rg}(I-\Pi)\\ x\neq 0}} \frac{\langle (I-S)x, (I-\hat{P}^*\hat{P})(I-S)x\rangle}{\|(I-\Pi)x\|^2} \\ &= 1 - \min_{\substack{z\in\mathrm{Rg}(I-S)\\ z\neq 0}} \frac{\langle z, (I-\hat{P}^*\hat{P})z\rangle}{\|(I-\Pi)z\|^2},\end{aligned}$$

which proves the first claim in the statement of this lemma.

To prove the second claim, we make the change of variables $w = (I-\hat{P}^*\hat{P})^{\frac{1}{2}}z$. The square root $(I-\hat{P}^*\hat{P})^{\frac{1}{2}}$ exists since $\rho(\hat{P}^*\hat{P}) < 1$ by Lemma 4.4, and therefore $I-\hat{P}^*\hat{P}$ is symmetric and positive definite with respect to the $\ell^2(1/\mu)$-inner product. See the proof of Theorem 5.4 for a detailed proof of the existence of a similar square root. We have

$$\begin{aligned}\|(I-\Pi)J\|^2 &= 1 - \min_{w\in(I-\hat{P}^*\hat{P})^{\frac{1}{2}}\,\mathrm{Rg}(I-S)} \frac{\|w\|^2}{\|(I-\Pi)(I-\hat{P}^*\hat{P})^{-\frac{1}{2}}w\|^2} \\ &\le 1 - \frac{1}{\|(I-\Pi)(I-\hat{P}^*\hat{P})^{-\frac{1}{2}}\|^2}.\end{aligned}$$

Note that the denominator here is nonzero because $I-\Pi \neq 0$ and $(I-\hat{P}^*\hat{P})^{-\frac{1}{2}}$ is invertible.

Now observe that

$$\begin{aligned}\|(I-\Pi)(I-\hat{P}^*\hat{P})^{-\frac{1}{2}}\|^2 &= \rho(((I-\Pi)(I-\hat{P}^*\hat{P})^{-\frac{1}{2}})^*(I-\Pi)(I-\hat{P}^*\hat{P})^{-\frac{1}{2}})\\ &= \rho((I-\Pi)^*((I-\hat{P}^*\hat{P})^{-\frac{1}{2}})^*(I-\hat{P}^*\hat{P})^{-\frac{1}{2}}(I-\Pi))\\ &= \rho((I-\Pi)(I-\hat{P}^*\hat{P})^{-1}(I-\Pi))\\ &= \|(I-\Pi)(I-\hat{P}^*\hat{P})^{-1}(I-\Pi)\|.\end{aligned}$$

The first equality above follows since for any $M \in \mathbb{R}^{N\times N}$ we have $\|M\|^2 = \rho(M^*M)$ by (C.2). The third equality follows since Π is an orthogonal projection, so $\Pi^* = \Pi$. The fourth follows since $I-\Pi$ and $\hat{P}^*\hat{P}$ are both symmetric in $\ell^2(1/\mu)$ and therefore $(I-\Pi)(I-\hat{P}^*\hat{P})^{-1}(I-\Pi)$ is symmetric, so its norm is equal to its spectral radius. ∎

D.7. Proof of Theorem 4.13. We prove local convergence of IAD.

Proof. By Theorem 4.12, $\|(I-\Pi)J(\mu)\|_{1/\mu} < 1$, and so for $\varepsilon > 0$ sufficiently small

$$\|J(\mu)\|_\varepsilon < \sqrt{\|(I-\Pi(\mu))J(\mu)\|_{1/\mu}^2 + \varepsilon\|\Pi(\mu) - S(\mu)\|_{1/\mu}^2} < 1$$

by Lemma 4.11. We observe that $J(\nu)$ is continuous as a mapping from the set of positive probability vectors to the space of operators on $\mathbb{R}^N$ with any choice of norms. This is because $S(\nu)$ is continuous, and so is the operator product. Therefore, there exist $r > 0$ and $\zeta < 1$ so that if $\|\nu - \mu\|_\varepsilon \le r$, then

$$\|J(\nu)\|_\varepsilon < \zeta.$$

The result follows. ∎

Appendix E. Proofs of results stated in section 5.

E.1. Proof of Lemma 5.1.

Proof. Let $\|\cdot\|$ denote both the norm on $\mathbb{R}^N$ and also the induced operator norm on $\mathbb{R}^{N\times N}$. By Lemma 4.7,

$$\begin{aligned}\limsup_{K\to\infty}\|\mu^K - \mu\|^{\frac{1}{K}} &= \limsup_{K\to\infty}\left\|\prod_{i=0}^{K-1} J(\mu^i)(\mu^0 - \mu)\right\|^{\frac{1}{K}}\\ &\le \limsup_{K\to\infty}\left\|\prod_{i=0}^{K-1} J(\mu^i)\right\|^{\frac{1}{K}}.\end{aligned}$$

By Gelfand's formula,

$$\lim_{n\to\infty}\|J(\mu)^n\|^{\frac{1}{n}} = \rho(J(\mu)).$$

Thus, for any $\delta > 0$ there exists an N so that

$$\|J(\mu)^N\|^{\frac{1}{N}} \le \rho(J(\mu)) + \delta.$$

Now write

$$\begin{aligned}
&\log\left(\left\|\prod_{i=0}^{K-1} J(\mu_i)\right\|^{\frac{1}{K}}\right)\\
&\quad\le \frac{1}{K}\left\{\sum_{i=0}^{\lfloor (K-1)/N\rfloor}\log\left\|\prod_{j=(i-1)N}^{iN-1} J(\mu_j)\right\| + \log\left\|\prod_{j=\lfloor (K-1)/N\rfloor N}^{K-1} J(\mu_j)\right\|\right\}\\
&\quad\le \frac{1}{\lfloor (K-1)/N\rfloor}\sum_{i=1}^{\lfloor (K-1)/N\rfloor}\frac{1}{N}\log\left\|\prod_{j=(i-1)N}^{iN-1} J(\mu_j)\right\| + \frac{1}{K}\sum_{j=\lfloor (K-1)/N\rfloor N}^{K-1}\log\|J(\mu_j)\|\\
&\quad=: A_K + B_K.
\end{aligned}$$

Since $\|J(\mu_n) - J(\mu)\| \to 0$, we have

$$\lim_{M\to\infty}\log\left\|\prod_{i=M+1}^{M+N} J(\mu_i)\right\| = \log\left\|J(\mu)^N\right\| \le N\log(\rho(J(\mu)) + \delta),$$

and so

$$\lim_{K\to\infty} A_K = \frac{1}{N}\log\|J(\mu)^N\| \le \log(\rho(J(\mu)) + \delta).$$

Moreover, $\lim_{K\to\infty} B_K = 0$. To see this, observe that by Appendix D.7, under our assumptions, $\lim_{j\to\infty}\mu_j = \mu$. Therefore, since $J(\nu)$ is continuous as a function of ν (cf. the proof of Appendix D.7), $C := \max\{\|J(\mu_j)\|; j\in\mathbb{N}\}$ is finite. Now the sum defining B_K consists of at most N terms, so $B_K \le \frac{CN}{K}$. The result follows. ∎

E.2. Proof of Theorem 5.4. In the proof of Theorem 5.4, we use that the norms $\|\cdot\|_{1/\mu}$ and $\|\cdot\|_{\mu}$ are dual with respect to the unweighted $\ell^2(\mathbb{1})$-inner product, and therefore $\|M\|_{1/\nu} = \|M^{\mathrm{t}}\|_{\nu}$ for any $M\in\mathbb{R}^{N\times N}$. Similar results are well known, and this might all be obvious to the reader, but we are unable to find a reference, so we offer a proof below.

LEMMA E.1. *For any $M\in\mathbb{R}^{N\times N}$ and any positive $\nu\in\mathbb{R}^N$, we have*

$$\|M\|_{1/\nu} = \|M^{\mathrm{t}}\|_{\nu}.$$

Proof. First, observe that

$$\begin{aligned}
\|y\|_{1/\nu}^2 &= \langle y, \mathrm{diag}(1/\nu)y\rangle_{\mathbb{1}}\\
&= \langle \mathrm{diag}(1/\nu)y, \mathrm{diag}(\nu)\,\mathrm{diag}(1/\nu)y\rangle_{\mathbb{1}}\\
&= \|\mathrm{diag}(1/\nu)y\|_{\nu}^2,
\end{aligned}$$

so $\mathrm{diag}(1/\nu)$ is an isometry mapping $\ell^2(1/\nu)$ to $\ell^2(\nu)$. Therefore, we have

$$\begin{aligned}
\|x\|_{1/\nu} &= \max_{\|y\|_{1/\nu}=1}\langle x, y\rangle_{1/\nu}\\
&= \max_{\|y\|_{1/\nu}=1}\langle x, \mathrm{diag}(1/\nu)y\rangle_{\mathbb{1}}\\
&= \max_{\|z\|_{\nu}=1}\langle x, z\rangle_{\mathbb{1}}.
\end{aligned}$$

The first equality above follows from the Cauchy–Schwarz inequality for $\ell^2(1/\nu)$. The last equality follows by making the change of variables $z = \operatorname{diag}(1/\nu)$ and using that $\operatorname{diag}(1/\nu)$ is an isometry mapping $\ell^2(1/\nu)$ to $\ell^2(\nu)$. We now have

$$\begin{aligned}\|M\|_{1/\nu} &= \max_{\|x\|_{1/\nu}=1} \|Mx\|_{1/\nu} \\ &= \max_{\|z\|_\nu=1} \max_{\|x\|_{1/\nu}=1} \langle Mx, z\rangle_{\mathbb{1}} \\ &= \max_{\|z\|_\nu=1} \max_{\|x\|_{1/\nu}=1} \langle x, M^{\mathrm{t}} z\rangle_{\mathbb{1}} \\ &= \max_{\|z\|_\nu=1} \|M^{\mathrm{t}} z\|_\nu \\ &= \|M^{\mathrm{t}}\|_\nu,\end{aligned}$$

as claimed. ∎

We now prove Theorem 5.4.

Proof. For convenience, we write $\|\cdot\|$ for $\|\cdot\|_{1/\mu}$, Π for $\Pi(\mu)$, and M^* for $M^{*,1/\mu}$. First, we observe that

$$\rho(J) \le \inf_{\varepsilon>0} \|J\|_\varepsilon = \|(I-\Pi)J\| \tag{E.1}$$

by Lemma 4.11. Also, observe that $I - \Pi \neq 0$, since at least one coarse state contains more than one fine state.

We now estimate $\|(I-\Pi)(I-\hat{P}^*\hat{P})^{-1}(I-\Pi)\|$. By Theorem 4.12, an upper bound on this expression implies an upper bound on $\|(I-\Pi)J\|^2$ and therefore on $\rho(J)^2$ by (E.1). Since $\rho(\hat{P}^*\hat{P}) < 1$ by Lemma 4.4, $I - \hat{P}^*\hat{P}$ is symmetric positive definite with respect to the $\ell^2(1/\mu)$-inner product, and therefore so is $(I-\hat{P}^*\hat{P})^{-1}$. It follows that there exists a square root

$$L = (I-\hat{P}^*\hat{P})^{-\frac{1}{2}} = \sum_{k=2}^{N} (1-\lambda_k)^{-\frac{1}{2}} v_i v_i^{\mathrm{t}} \operatorname{diag}(1/\mu),$$

which is also symmetric positive definite. We note that L must commute with the spectral projector Q. Also,

$$\|LQ\| = (1-\lambda_2)^{-\frac{1}{2}} \quad \text{and} \quad \|L(I-Q)\| = (1-\lambda_{k+1})^{-\frac{1}{2}}. \tag{E.2}$$

These facts follow directly from the above formula for L in terms of the diagonalization of P^*P.

Since $(I-\Pi)$ is an orthogonal projection, $(I-\Pi)^* = I - \Pi$, and we have

$$\begin{aligned}\|(I-\Pi)(I-\hat{P}^*\hat{P})^{-1}(I-\Pi)\| &= \|(I-\Pi)L^2(I-\Pi)\| \\ &= \|(L(I-\Pi))^* L(I-\Pi)\| \\ &= \|L(I-\Pi)\|^2.\end{aligned}$$

The last equality above follows since for any $M \in \mathbb{R}^{N\times N}$, $\|M^*M\| = \|M\|^2$ by (C.2). We have

$$\begin{aligned}
\|L(I-\Pi)\|^2 &= \max_{\|x\|=1} \|L(I-\Pi)x\|^2 \\
&= \max_{\|x\|=1} \|QL(I-\Pi)x\|^2 + \|(I-Q)L(I-\Pi)x\|^2 \\
&= \max_{\|x\|=1} \|LQ(I-\Pi)x\|^2 + \|L(I-Q)(I-\Pi)x\|^2 \\
&= \max_{\|x\|=1} \|LQQ(I-\Pi)x\|^2 + \|L(I-Q)(I-Q)(I-\Pi)x\|^2 \\
&\le \max_{\|x\|=1} \|LQ\|^2 \|Q(I-\Pi)x\|^2 + \|L(I-Q)\|^2 \|(I-Q)(I-\Pi)x\|^2 \\
&= \max_{\|x\|=1} \frac{1}{1-\lambda_1} \|Q(I-\Pi)x\|^2 + \frac{1}{1-\lambda_{k+1}} \|(I-Q)(I-\Pi)x\|^2.
\end{aligned}$$

The second equality in the above display follows from the Pythagorean law, since Q is an orthogonal projection. The third equality follows since L and Q commute, as explained in the previous paragraph. The last equality follows from our formulas in (E.2) for the norms of LQ and $L(I-Q)$.

Now observe that

$$\|(I-Q)(I-\Pi)x\|^2 + \|Q(I-\Pi)x\|^2 = \|(I-\Pi)x\|^2 \le 1,$$

again by the Pythagorean law. Therefore,

$$\|(I-Q)(I-\Pi)x\|^2 \le 1 - \|Q(I-\Pi)x\|^2. \tag{E.3}$$

By Lemma E.1,

$$\|Q(I-\Pi)\| = \|(I-\Pi^{\mathrm{t}})Q^{\mathrm{t}}\|_\mu = \sin(\theta). \tag{E.4}$$

Combining (E.3) and (E.4) with the results of the previous paragraph yields

$$\begin{aligned}
&\|(I-\Pi)(I-\hat{P}^*\hat{P})^{-1}(I-\Pi)\| \\
&\quad\le \max_{\|x\|=1} \frac{1}{1-\lambda_1} \|Q(I-\Pi)x\|^2 + \frac{1}{1-\lambda_{k+1}} \|(I-Q)(I-\Pi)x\|^2 \\
&\quad\le \max_{0\le\alpha\le\sin^2(\theta)} \frac{1}{1-\lambda_1}\alpha^2 + \frac{1}{1-\lambda_{k+1}}(1-\alpha^2) \\
&\quad= \frac{1}{1-\lambda_1}\sin^2(\theta) + \frac{1}{1-\lambda_{k+1}}\cos^2(\theta).
\end{aligned}$$

The result now follows by Theorem 4.12. ∎

E.3. Proof of Theorem 5.5.

Proof. For convenience, we write J for $J(\mu)$, Π for $\Pi(\mu)$, etc. Note that $I-\Pi \neq 0$ and $\Pi \neq 0$, since there is more than one coarse state and at least one coarse state contains more than one fine state. Recall that J has the same spectrum as $(I-\Pi)J$ by (D.4). Observe that $0 \in \sigma((I-\Pi)J)$, since we have $J = J(I-\Pi)$ by (D.3), and so $Jx = 0$ for any $x \in \mathrm{Rg}(\Pi)$. Now suppose that

$$(I-\Pi)Jx = \lambda x \tag{E.5}$$

for some $\lambda \neq 0$ and $x \neq 0$. We will show that x is an eigenvector of $(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)$ with eigenvalue $(1-\lambda)^{-1}$. (Note that $\lambda \neq 1$ because $\rho((I-\Pi)J) < 1$ by Theorem 4.12, so $(1-\lambda)^{-1}$ is defined.) Since $\Pi J = \Pi - S$ by Lemma D.3, we have

$$\begin{aligned}(I-\Pi)-(I-\Pi)J &= I-\Pi-J+\Pi-S\\ &= I-S-\hat{P}(I-S)\\ &= (I-\hat{P})(I-S).\end{aligned}$$

Since $\lambda \neq 0$, the eigenvalue equation (E.5) implies that $x \in \mathrm{Rg}(I-\Pi)$. Therefore, since $I-\Pi$ is a projection, we have $(I-\Pi)x = x$, and

$$(I-\hat{P})(I-S)x = (I-\Pi)x - (I-\Pi)Jx = (1-\lambda)(I-\Pi)x.$$

Multiplying on both sides above by $(I-\Pi)(I-\hat{P})^{-1}$ yields

$$\begin{aligned}(1-\lambda)(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)x &= (I-\Pi)(I-S)x\\ &= (I-\Pi)x\\ &= x,\end{aligned}$$

since we have $(I-\Pi)(I-S) = I-\Pi$ by Lemma D.2. Thus, x is an eigenvector of $(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)$ with eigenvalue $(1-\lambda)^{-1}$. We conclude that

$$\sigma(J(\mu)) \subset \left(1 - \frac{1}{\sigma((I-\Pi(\mu))(I-\hat{P})^{-1}(I-\Pi(\mu)))\setminus\{0\}}\right)\cup\{0\}.$$

It remains to prove the opposite inclusion. Consider the operator

$$M = (I-\hat{P})(I-S).$$

We have

$$\begin{aligned}\Pi M &= DA(I-\hat{P})(I-D(A(I-\hat{P})D)^{-1}A(I-\hat{P}))\\ &= DA(I-\hat{P}) - DA(I-\hat{P})D(A(I-\hat{P})D)^{-1}A(I-\hat{P})\\ &= 0,\end{aligned}$$

so $\mathrm{Rg}(M) \subset \mathrm{Rg}(\Pi)$. Moreover, since $I-\hat{P}$ is invertible, $\ker(M) = \mathrm{Rg}(S) = \mathrm{Rg}(\Pi)$. Therefore, when viewed as an operator on the range of $I-\Pi$, M is invertible.

Now suppose that

$$(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)y = \alpha y$$

for some $\alpha \neq 0$ and $y \neq 0$. Then $y \in \mathrm{Rg}(I-\Pi)$, and so we may write $y = Mx$ for some $x \in \mathrm{Rg}(I-\Pi)$. Therefore,

$$\begin{aligned}\alpha Mx &= (I-\Pi)(I-\hat{P})^{-1}(I-\Pi)Mx\\ &= (I-\Pi)(I-\hat{P})^{-1}Mx\\ &= (I-\Pi)(I-S)x\\ &= (I-\Pi)x\\ &= x,\end{aligned}$$

since $(I-\Pi)(I-S) = I-\Pi$ and $x \in \mathrm{Rg}(I-\Pi)$. Therefore, $\frac{1}{\alpha}$ is an eigenvalue of M with eigenvector x. Now

$$\begin{aligned}(I-\Pi)J + M &= \left[(I-\Pi)\hat{P} + (I-\hat{P})\right](I-S)\\ &= \left[(I-\Pi) + \Pi(I-\hat{P})\right](I-S)\\ &= (I-\Pi)(I-S) + \Pi M\\ &= (I-\Pi),\end{aligned}$$

since $\Pi M = 0$ and $(I-\Pi)(I-S) = I-\Pi$ as explained above. Therefore,

$$(I-\Pi)J = (I-\Pi) - M,$$

and $x \in \mathrm{Rg}(I-\Pi)$ is an eigenvector of $(I-\Pi)J$ with eigenvalue $1-\frac{1}{\alpha}$. Finally, 0 is an eigenvalue of J because $Jx = 0$ whenever $x \in \mathrm{Rg}(\Pi)$. Therefore,

$$\sigma(J(\mu)) \supset \left(1 - \frac{1}{\sigma((I-\Pi(\mu))(I-\hat{P})^{-1}(I-\Pi(\mu))) \setminus \{0\}}\right) \cup \{0\}. \qquad \square$$

E.4. Proof of Corollary 5.6.

Proof. For convenience, we write Π for $\Pi(\mu)$, J for $J(\mu)$, and $\|\cdot\|$ for $\|\cdot\|_{1/\mu}$. Observe that $(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)$ is self-adjoint as an operator on $\ell^2(1/\mu)$, since P is reversible and $I-\Pi$ is an orthogonal projection, and so both $\hat{P}$ and $I-\Pi$ are self-adjoint. Therefore, $\sigma((I-\Pi)(I-\hat{P})^{-1}(I-\Pi)) \subset \mathbb{R}$. Since P and $P^{\mathrm{t}}P$ are irreversible, we have $\rho(J) < 1$ by Theorem 4.13. Therefore, by Theorem 5.5 we must have $\sigma((I-\Pi)(I-\hat{P})^{-1}(I-\Pi)) \subset [0,\infty)$ because if any eigenvalue of $(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)$ were negative, then Theorem 5.5 would imply $\rho(J) > 1$. It follows that

$$\begin{aligned}\rho(J) &= \max_{\lambda \in \sigma((I-\Pi)(I-\hat{P})^{-1}(I-\Pi))} \left|1 - \frac{1}{\lambda}\right|\\ &= 1 - \frac{1}{\rho((I-\Pi)(I-\hat{P})^{-1}(I-\Pi))}\\ &= 1 - \frac{1}{\|(I-\Pi)(I-\hat{P})^{-1}(I-\Pi)\|}.\end{aligned}$$

The proof of the inequality is identical with the proof of Theorem 5.4, except with $L = (I-\hat{P})^{-\frac{1}{2}}$ and $\sqrt{\lambda_i}$ in place of λ_i. $\square$

Appendix F. A model of overdamped Langevin dynamics on a two-dimensional grid. We define a family of Markov chains on a two-dimensional grid with properties similar to overdamped Langevin. Let $V : \mathbb{R}^2 \to \mathbb{R}$, $T > 0$, $[a,b] \times [c,d] \subset \mathbb{R}^2$, and $N \in \mathbb{N}$. Define the discrete Boltzmann distribution $\mu \in \mathbb{R}^{N \times N}$ by

$$\mu_{ij} = Z^{-1} \exp\left(-\frac{V\left(a + \frac{b-a}{N} i, c + \frac{d-c}{N} i\right)}{T}\right), \tag{F.1}$$

where $Z = \sum_{i=1}^{N} \exp\big(-\frac{V\left(a+\frac{b-a}{N}i, c+\frac{d-c}{N}i\right)}{T}\big)$. Define a transition probability on $\Omega = \{1,\dots,N\}^2$ by

$$\begin{aligned} P_{(i+k,j+\ell),(i,j)} &:= \frac{1}{4}\frac{\mu(i+k,j+\ell)}{\mu(i+k,j+\ell)+\mu(i,j)} && \text{for } k,\ell \in \{-1,1\} \quad \text{and} \quad (i,j)\in\Omega, \\ P_{(i,j),(i,j)} &:= 1 - \sum_{k,\ell\in\{-1,1\}} P_{(i+k,j+\ell)} && \text{for } (i,j)\in\Omega, \\ P_{(k,\ell),(i,j)} &:= 0 && \text{otherwise.} \end{aligned}$$

In the definition of P, we impose periodic boundary conditions, associating $(0,j)$ with (N,j), $(i,1)$ with $(i,N+1)$, etc. Note that P is in detailed balance with μ.

Acknowledgment. We acknowledge the influence of many deep conversations with David Aristoff, Aaron R. Dinner, Jonathan Mattingly, Gideon Simpson, and Jonathan Weare.

REFERENCES

[1] J. M. Bello-Rivas and R. Elber, *Exact milestoning*, J. Chem. Phys., 142 (2015), 094102.
[2] D. Bhatt, B. W. Zhang, and D. M. Zuckerman, *Steady-state simulations using weighted ensemble path sampling*, J. Chem. Phys., 133 (2010), 014110.
[3] W.-L. Cao and W. J. Stewart, *Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chains*, J. Assoc. Comput. Mach., 32 (1985), pp. 702–719.
[4] F. Chatelin and W. L. Miranker, *Acceleration by aggregation of successive approximation methods*, Linear Algebra Appl., 43 (1982), pp. 17–47.
[5] E. Darve and A. Pohorille, *Calculating free energies using average force*, J. Chem. Phys., 115 (2001), pp. 9169–9183.
[6] H. De Sterck, T. A. Manteuffel, S. F. McCormick, K. Miller, J. Pearson, J. Ruge, and G. Sanders, *Smoothed aggregation multigrid for Markov chains*, SIAM J. Sci. Comput., 32 (2010), pp. 40–61, https://doi.org/10.1137/080719157.
[7] H. De Sterck, T. A. Manteuffel, S. F. McCormick, Q. Nguyen, and J. Ruge, *Multilevel adaptive aggregation for Markov chains, with application to web ranking*, SIAM J. Sci. Comput., 30 (2008), pp. 2235–2262, https://doi.org/10.1137/070685142.
[8] A. Dickson, M. Maienschein-Cline, A. Tovo-Dwyer, J. R. Hammond, and A. R. Dinner, *Flow-dependent unfolding and refolding of an RNA by nonequilibrium umbrella sampling*, J. Chem. Theory Comput., 7 (2011), pp. 2710–2720.
[9] A. R. Dinner, J. C. Mattingly, J. O. B. Tempkin, B. Van Koten, and J. Weare, *Trajectory stratification of stochastic dynamics*, SIAM Rev., 60 (2018), pp. 909–938, https://doi.org/10.1137/16M1104329.
[10] A. R. Dinner, E. H. Thiede, B. Van Koten, and J. Weare, *Stratification as a general variance reduction method for Markov Chain Monte Carlo*, SIAM-ASA J. Uncertain. Quantif., 8 (2020), pp. 1139–1188, https://doi.org/10.1137/18M122964X.
[11] G. Earle and J. Mattingly, *Convergence of Stratified MCMC Sampling of Non-reversible Dynamics*, arXiv:2111.05838, 2022.
[12] C. J. Geyer, *Markov Chain Monte Carlo Maximum Likelihood*, Interface Foundation of North America, Fairfax, VA, 1991.
[13] G. H. Golub and C. D. Meyer, Jr., *Using the QR factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities for Markov chains*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 273–281, https://doi.org/10.1137/0607031.
[14] M. Haviv, *Aggregation/disaggregation methods for computing the stationary distribution of a Markov Chain*, SIAM J. Numer. Anal., 24 (1987), pp. 952–966, https://doi.org/10.1137/0724062.
[15] J. R. Koury, D. F. McAllister, and W. J. Stewart, *Iterative methods for computing stationary distributions of nearly completely decomposable Markov chains*, SIAM J. Algebraic Discrete Methods, 5 (1984), pp. 164–186, https://doi.org/10.1137/0605019.
[16] U. R. Krieger, *On a two-level multigrid solution method for finite Markov chains*, Linear Algebra Appl., 223/224 (1995), pp. 415–438.

[17] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *The weighted histogram analysis method for free-energy calculations on biomolecules.* I. *The method*, J. Comput. Chem., 13 (1992), pp. 1011–1021.
[18] A. Laio and M. Parrinello, *Escaping free-energy minima*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 12562–12566.
[19] T. Lelièvre and G. Stoltz, *Partial differential equations and stochastic methods in molecular dynamics*, Acta Numer., 25 (2016), pp. 681–880.
[20] Y. Li, X. Qu, A. Ma, G. J. Smith, N. F. Scherer, and A. R. Dinner, *Models of single-molecule experiments with periodic perturbations reveal hidden dynamics in RNA folding*, J. Phys. Chem. B, 113 (2009), pp. 7579–7590.
[21] J. Mandel and B. Sekerka, *A local convergence proof for the iterative aggregation method*, Linear Algebra Appl., 51 (1983), pp. 163–172.
[22] I. Marek and P. Mayer, *Convergence analysis of an iterative aggregation/disaggregation method for computing stationary probability vectors of stochastic matrices*, Numer. Linear Algebra Appl., 5 (1998), pp. 253–274.
[23] I. Marek and P. Mayer, *Convergence theory of some classes of iterative aggregation/disaggregation methods for computing stationary probability vectors of stochastic matrices*, Linear Algebra Appl., 363 (2003), pp. 177–200.
[24] I. Marek and I. Pultarová, *A note on local and global convergence analysis of iterative aggregation-disaggregation methods*, Linear Algebra Appl., 413 (2006), pp. 327–341.
[25] I. Marek and D. B. Szyld, *Local convergence of the (exact and inexact) iterative aggregation method for linear systems and Markov operators*, Numer. Math., 69 (1994), pp. 61–82.
[26] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2008.
[27] J. Norris, *Markov Chains*, Cambridge University Press, Camridge, UK, 1998.
[28] S. Park, M. K. Sener, D. Lu, and K. Schulten, *Reaction paths based on mean first-passage times*, J. Chem. Phys., 119 (2003), pp. 1313–1319.
[29] I. Pultarová, *Fourier analysis of the aggregation based algebraic multigrid for stochastic matrices*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1596–1610, https://doi.org/10.1137/130913821.
[30] M. R. Shirts and J. D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states*, J. Chem. Phys., 129 (2008), 124105.
[31] R. H. Swendsen and J.-S. Wang, *Replica Monte Carlo simulation of spin-glasses*, Phys. Rev. Lett., 57 (1986), pp. 2607–2609.
[32] E. Thiede, B. Van Koten, and J. Weare, *Sharp entrywise perturbation bounds for Markov Chains*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 917–941, https://doi.org/10.1137/140987900.
[33] G. Torrie and J. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling*, J. Comput. Phys., 23 (1977), pp. 187–199.
[34] I. Y. Vakhutinsky, L. M. Dudkin, and A. A. Ryvkin, *Iterative aggregation—a new approach to the solution of large-scale problems*, Econometrica, 47 (1979), pp. 821–841.
[35] E. Vanden-Eijnden and M. Venturoli, *Exact rate calculations by trajectory parallelization and tilting*, J. Chem. Phys., 131 (2009), 044120.
[36] A. Warmflash, P. Bhimalapuram, and A. R. Dinner, *Umbrella sampling for nonequilibrium processes*, J. Chem. Phys., 127 (2007), 154112.