SOLVING BURGERS' EQUATION USING

approximation methods for solving partial differential and integral equations in higher dimensions, where the ability to construct near optimal rational (or exponential) approximations to functions of one variable is a key component.

Since the seminal result in [21], it has been known that functions with singularities may be efficiently approximated in the $\mathbf{L}^{\infty}$ norm using proper rational functions. Indeed, the number of poles required to approximate a function with singularities is directly related to the sparsity of the function's wavelet coefficients (see [16, Theorem 11.1]). However, in contrast to more traditional $\mathbf{L}$ -type methods (using e.g., wavelet bases as in [2]), the use of such optimal $\mathbf{L}^{\infty}$-type approximations in numerical analysis has been limited due to a lack of efficient and robust algorithms.

Given a proper rational function $\mathbf{f}$, we present an algorithm—which we refer to as the reduction algorithm—to compute, for a fixed number 20.8801030e many times. For example, in the context of solving Burgers' equation with viscosity $\nu = 10^{-}$ and approximation tolerance $\epsilon = 10^{-}$, on the order of a million applications of the reduction algorithm are performed.

For functions with n poles resulting from intermediate computations, the reduction algorithm requires only $O(m\ n)$ operations to find an optimal approximation with m poles. In our numerical experiments with the reduction algorithm, we find that an approximation error of $\epsilon \approx 10^{-}$ may be reliably obtained within double precision arithmetic, even when the number of poles n is large and their spatial distribution is highly clustered.

There is a significant literature devoted to applications of the AAK approach in control theory (cf. [23]), signal processing (cf. [8]), and numerical analysis (cf. [25, 27, 29, 5]), to mention just a select few. The reformulation of the AAK theory given here could be related to the approaches taken in [28], [20], and [10]. However, as far as we know, all of the AAK-type algorithms discussed in the literature require $O(n\ )$ operations when applied to a rational function with n poles, and may require extended precision arithmetic if high accuracy of the result is desired. In contrast, our reduction algorithm requires only $O(m\ n)$ operations to find an optimal approximation with m poles and achieves high accuracy ($\epsilon \approx 10^{-}$) using only double precision arithmetic.

We show in this paper that solutions of Burgers' equation with viscosity ν require only $O(\log \nu^{-}\ ) + O(\log \epsilon^{-}\ )$ poles for its rational approximation with an L^{∞} error of size ϵ. Burgers' equation has been traditionally used to test the limits of new numerical methods since the solution develops sharp transition regions that need to be captured adaptively. Conceptually, the two closest adaptive methods are those in [24] and [2]. While in [2] adaptivity is achieved by adding wavelet scales when needed, the algorithm in [24] achieves spectral accuracy by a [illegible] 3000.69b8.6614.1602o)T0Td90Td(resuTj32 smaller number of poles. As mentioned earlier, our reduction algorithm is based on a theorem of Adamyan, Arov, and Krein ([1]), which concerns the approximation of a periodic function $\mathbf{f}$, essentially bounded on the unit circle $\partial\mathbb{D}$, by a meromorphic function $\mathbf{r(z)}$ ($\mathbf{z = e^{\pi i x}}$) containing a specified number of poles in the unit disk. We limit our presentation to rational functions $\mathbf{f}$ taking real values on $\partial\mathbb{D}$. This case turns out to be particularly important, as it allows us to develop a practical algorithm based on approximating the Fourier series coefficients of $\mathbf{f}$ with positive index. More general functions $\mathbf{f}$ may be dealt with by using the techniques in

The fact that there are exactly m zeros in the unit disk, corresponding to the index m of the con-eigenvalue λ_m, is a consequence of the AAK theory. As shown in Section 4.1 (see equations (4.8)), the key to the high accuracy of evaluating the function $v(z)$ is the relationship

$$v(\gamma_i) = \overline{u_i}/\sqrt{\overline{\alpha_i}}, \quad i = 1,\ldots,n, \tag{2.4}$$

which, together with the n poles $1/\overline{\gamma_i}$, uniquely determines $v(z)$.

Step 3: Find the coefficients β_i of $g(z)$ by solving the $m \times m$ linear system,

$$\sum_{i=}^{m} \frac{1}{1-\eta_i\overline{\eta_j}}\beta_i = \sum_{i=}^{n} \frac{\alpha_i}{1-\gamma_i\overline{\eta_j}}, \quad j = 1,\ldots,m. \tag{2.5}$$

Denoting $□f - g□_\infty = \sup_{x\in[\ ,\ ]}|f(e^{-\ \pi ix}) - g(e^{-\ \pi ix})|$, the resulting rational approximation $g(e^{\ \pi ix})$ satisfies $□f - g□_\infty \approx □$ and, thus, is close to the best L^∞-error achievable by rational functions with no more than m poles in the unit disk (see also [25] for a discussion of optimal rational approximations).

Remark 1. In Step 3, we solve for the coefficients β_i in $O(m\ )$ operations by exploiting the structure of Cauchy matrices (see [11, 7]). We note that such a solver may require quadruple precision if the overall desired approximation error □ is smaller than $\approx 10^{-}$. However, since $m = \log(□^{-}\ )$ is small, Step 3 for finding coefficients β_i does not impact the overall speed of the algorithm even if performed in quadruple precision.

Remark 2. In applications where the function $f(e^{\ \pi ix})$ has singularities or sharp transitions, the poles γ_j in the rational representation of $f(e^{\ \pi ix})$ may be located very close to the unit circle (and/or to each other). In such cases, it is advantageous to maintain the poles in the form $\gamma_j = \exp(-\tau_j)$, since they are well separated on a logarithmic scale. Importantly, the reduction algorithm computes the new poles

α of p(sitv)-2999327e-definit(d)]T863.15980Td[(Cauc)2998.66(h)3000.69(y)]T43.08010Td(matrex)Tj/153

$(r < 1)$. This estimate shows that, for accuracy □, we may reasonably expect $O(\log □^{-})$ terms in our approximation. In fact, we have observed this behavior in our numerical experiments.

Let us now briefly discuss the algorithmic aspects behind efficiency and accuracy of solving steps 1-3 above.

2.2. Accurate computation of con-eigenvalues/eigenvectors.

For Step 1, we use a recent algorithm developed and analyzed in [?] for computing con-eigenvalues of Cauchy matrices with high relative accuracy, which we briefly describe in this section.

It is well-known that standard eigenvalue algorithms compute an approximate con-eigenvalue $\widehat{\ }$

using (2.4) to rewrite (2.3) as

$$\sum_{i=}^{n} \frac{\overline{\alpha_i}\, v(\gamma_i)}{1 - \overline{\gamma_i} z} = \lambda_m v(z),$$

we see that

to the known eigenvalues $\boldsymbol{\beta}\ ,\ldots,\boldsymbol{\beta}_{\mathbf{m}-}$, one by one. We then orthogonalize these $\mathbf{m}-1$ vectors using the stabilized Gram-Schmidt procedure, thus yielding a basis $\hat{\mathbf{q}}\ ,\ldots,\hat{\mathbf{q}}_{\mathbf{m}-}$ for the invariant subspace span $\{\mathbf{q}\ ,\ldots,\mathbf{q}_{\mathbf{m}-}\ \} =$ span $\{\hat{\mathbf{q}}\ ,\ldots,\hat{\mathbf{q}}_{\mathbf{m}-}\ \}$. Finally, we use simultaneous inverse iteration applied to $\hat{\mathbf{q}}\ ,\ldots,\hat{\mathbf{q}}_{\mathbf{m}-}\ ,\mathbf{q}$, where $\mathbf{q}$ is chosen randomly. Notice that each step of this process

8ess

depend on the timestep $\mathbf{t}$, the number $\mathbf{M_t}$ of quadrature nodes in time, and the number of quadrature nodes $\mathbf{M_x}$ used in space to discretize the convolution kernels. From the rapid decay of the periodic heat kernel,

$$\mathbf{K_\nu(x,t)} = \frac{1}{\sqrt{4\pi\nu t}} \sum_{k\in} e^{-(x+k)^2/(\ \nu t)},$$

where ν is the viscosity parameter in (3.1), it follows that φ_p^l and ψ_p^l are localized to a $\mathbf{O}\left(\sqrt{\ \nu t}\right)$ neighborhood of $\mathbf{x = 0}$ (see Section 4.2 for details).

We assume that the initial function $\mathbf{u(x,0) = u\ \ (x)}$ is given as a periodic rational function of the form

$$\mathbf{u\ \ (x)} = \sum_{j=}^{M_0} \frac{\alpha_j}{e^{-\ \pi i x} - y_j} + \sum_{i=}^{M_0} \frac{\overline{\alpha_j}}{e^{\ \pi i x} - \overline{y_j}} + \alpha\ ,$$

and that this representation is nearly optimal. We then solve the system of equations (3.2) by approximating each function $\mathbf{u_l}$ using the reduction algorithm. We obtain, via fixed point iteration applied to (3.2) and the reduction algorithm, rational functions $\mathbf{u_l(x)}$ of the form,

$$\mathbf{u_l(x)} = \sum_{j=}^{M_l} \frac{\alpha_{j,l}}{e^{-\ \pi i x} - y_{j,l}} + \sum_{j=}^{M_l} \frac{\overline{\alpha_{j,l}}}{e^{\ \pi i x} - \overline{y_{j,l}}} + \alpha\ , \tag{3.3}$$

which solve (3.2) to a specified level of precision, and have a (near) optimal number of poles.

More specifically, given $\mathbf{u_j^{(m)} \approx u_j(x)}$, $\mathbf{1 \le j \le M_t}$, at iteration $\mathbf{m}$, we use (3.2) to define the next iterates $\mathbf{u_l^{(m+\ )}(x)}$ for $\mathbf{l = 1, \ldots, M_t}$,

$$\mathbf{u_l^{(m+\ )}(x)} = \sum_{p=}^{M_x} \lambda_p^l u\ \ \left(x - \varphi_p^l\right)$$

(3.4)

$$_j\ ^{)}\ 0 - \psi_p{}^{\eta} \qquad + \sum_{p=}^{M_x} \lambda_{p,j}^l \sum_{j=}^{l-l} \left\{ \left(u_j^{(m+\ )} \left(x - \psi_p^l\right) \right) + \sum_{j=l}^{M_t} \right.$$

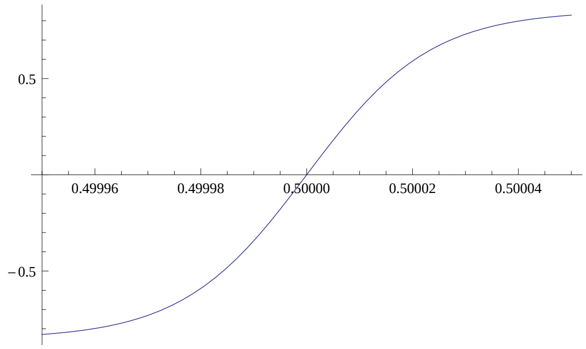
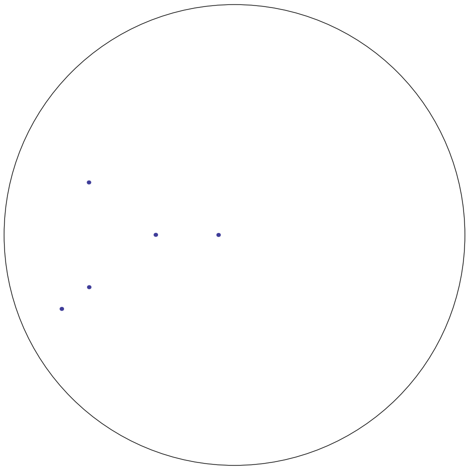


FIGURE 3.3. Solution $u(x,t)$ at time $t = .4$, localized about the transition region $(1/2 - 10^{-}, 1/2 + 10^{-})$. Note the absence of any Gibbs-type phenomena.



Suppose $\mathbf{f} \in \mathbf{L}^\infty$ has the we calculate from (4.4)

$$\sum_{j=}^{\infty}\left(\sum_{m=}^{M}\alpha_m\gamma_m^{i+j-}\right)v_j = \sum_{m=}^{M}\alpha_m\gamma_m^{i-}\sum_{j=}^{\infty}\gamma_m^{j-}\,v_j$$
$$= \sum_{m=}^{M}\alpha_m\gamma_m^{i-}\,v\,(\gamma_m) = \sigma w_i.$$

Now multiplying both sides of the last equation by z^{i-} and summing, we obtain

$$\sum_{m=}^{M}\frac{\alpha_m}{1-\gamma_m z}v\,(\gamma_m) = \sigma z^{-}\,w(z^{-}\,). \tag{4.6}$$

Similarly, from (4.5), we have

$$\sum_{j=}^{\infty}\left(\sum_{m=}^{M}\overline{\alpha_m}\,\overline{\gamma_m}^{i+j-}\right)w_j = \sum_{m=}^{M}\overline{\alpha_m}\,\overline{\gamma_m}^{i-}\sum_{j=}^{\infty}\overline{\gamma_m}^{j-}\,w_j$$
$$= \sum_{m=}^{M}\overline{\alpha_m}\,\overline{\gamma_m}^{i-}\left(\overline{\gamma_m}^{-}\,w\left(\overline{\gamma_m}^{-}\right)\right) = \sigma v_i.$$

Finally, multiplying by z^{i-} and summing, we arrive at

$$\sum_{m=}^{M}\frac{\overline{\alpha_m}}{1-\overline{\gamma_m}z}\overline{\gamma_m}^{-}\,w\left(\overline{\gamma_m}^{-}\,\right) = \sigma v(z). \tag{4.7}$$

Hence, for a function f of the form (4.3), the functions v and w in (4.2) turn out to be rational and fully determined by their values at the poles of f. Taking $z = \overline{\gamma_n}$ and $z = \gamma_n$ in equations (4.6) and (4.7), respectively, we obtain

$$\sum_{m=}^{M}\quad\alpha_m$$

Let us define the vectors $\mathbf{p}$ and $\mathbf{q}$ with entries $p_m = \alpha_m^{\frac{1}{2}} v(\gamma_m)$, $q_m = \overline{\alpha_n}\ /\ \overline{\gamma_m}^{-}\ w(\overline{\gamma_m}^{-}\ )$, and the positive definite matrix $\mathbf{C}$ with entries

$$C_{mn} = \frac{\alpha_m^{\frac{1}{2}} \overline{\alpha_n}^{\frac{1}{2}}}{1 - \gamma_m \overline{\gamma_n}}.$$

Then the above equations are equivalent to

$$\begin{aligned} C\,p &= \sigma q, \\ C\,\overline{q} &= \sigma \overline{p}, \end{aligned}$$

which may be reduced to a con-eigenvalue problem for $\sigma > 0$, see [15, Section 4.6]. One simple way to see this and obtain an equation of the form (2.2) is by defining $x = p + \overline{q}.$ If $x = 0,$ then $iq = \overline{ip}$ and hence

$$C(ip) = \sigma \overline{ip}.$$

If $x \boxminus 0$, we have

$$Cx = \sigma \overline{x}$$

and, in both cases, we obtain a con-eigenvalue problem for the matrix $\mathbf{C}$.

4.2. **Discretization of Burgers' equation.** We rewrite the equation (3.1) in semi-group form (see, e.g., [14, 17, 18, 3])

$$u(t) = e^{\nu t L} u(0) + \int^{t} e^{\nu(t-\tau)L} N(u(\tau))d\tau, \tag{4.9}$$

where $u(t)$ denotes the function $u(\cdot, t)$. The operator L, $Lu(x) = u_{xx}$, represents the linear part of (3.1) while the operator N, $N(u) = 1/2\,(u\ )_x$, represents the nonlinear part. The action of the operator $e^{\nu t L}$ on a function f is given by

$$\left(e^{\nu t L} f\right)(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} K_\nu(y,t) f(x-y)dy, \text{ with } K_\nu(y,t) = \frac{1}{\sqrt{4\pi\nu t}} \sum_{k\in} e^{-(y+k)^2/(\ \nu t)}.$$

To discretize equation (4.9) in time, we use the approximation

$$N(u(\tau)) \approx \sum_{j=}^{M_t} R_j(\tau) N(u(\tau_j)), \quad \tau \ \square\ [0,t]$$

where $\{\tau_j\}_{j=}^{M_t}$ denote the Gauss-Legendre nodes on the interval $(0,t)$, and $R_j(\tau)$ denote the Legendre interpolating polynomials for these nodes, i.e.,

$$R_j(\tau_m) = \delta_{jm}, \quad \text{for } j, m = 1, \ldots, M_t.$$

Taking $t = \tau_l$ in (4.9), we obtain the semi-discrete system of equations

$$u_l = e^{\nu \tau_l L} u\ \ + \sum_{j=}^{M_t} \left(\int^{\tau_l} e^{\nu(\tau_l - \tau)L} R_j(\tau) d\tau \right) N(u_j), \quad 1 \leq l \leq M_t, \tag{4.10}$$

where $u_l = u_l(x)$ denote the computed values of u at time $t = \tau_l$ and $u\ \ = u(x, 0)$.

For the spatial discretization, using $N(u) = 1/2\,(u\ )_x$ and integrating by parts, equation (4.10) may be written as

$$u_l(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} K_\nu(y,\tau_l)u\ (x-y)dy + \sum_{j=}^{M_t} \int_{-\frac{1}{2}}^{\frac{1}{2}} L_{\nu,j}(y,\tau_l)u_j(x-y)dy, \tag{4.11}$$

where the kernel $L_{\nu,j}(y,t)$ is given by

$$L_{\nu,j}(y,t) = -\frac{1}{2}\int^{t} (\partial_y K_\nu)(y,t-s)\,R_j(s)ds.$$

For small ν

[9] G. Dahlquist and Å. Björck. *Numerical methods in scientific computing. Vol. I.* Society for Industrial and Applied Mathematics SIAM), Philadelphia, PA, 2008.

[10] Ph. Delsarte, Y. Genin, and Y. Kamp. On the role of the Nevanlinna-Pick problem in circuit and system theory. *International Journal of Circuit Theory and Applications*, 9 2):177–187, 1981.

[11] J. Demmel. Accurate singular value decompositions of structured matrices. *SIAM J. Matrix Anal. Appl.*, 21 2):562–580, 1999.

[12] J. Demmel, M. Gu, S. Eisenstat, I. Slapnicar, K. Veselic, and Z. Drmac. Computing the singular value decomposition with high relative accuracy. *LAPACK Working Note*, 119 CS-97- 48), 1997.

[1] J. Demmel and K. Veselic. Jacobi's method is more accurate than QR. *SIAM. J. Matrix Anal. and Appl.*, 1 4):1204–1245, 1992.

[14] E. Hille and R. S. Phillips. *Functional Analysis and Semi-groups.* American Mathematical Society, Providence, RI, 1957.