Gauge-equivariant multigrid networks

based on Lehner, Wettig 2302.05419, 2304.10438

Christoph Lehner

(University of Regensburg)

November 25, 2023 – Tsukuba

Motivation of research program

- Goal: approximate propagators D^{-1}, $\det(D)$, and hadronic correlation functions
- Deep networks work (cf. Krylov solvers)
- Multigrid paradigm makes much shallower models perform as well as deep ones (cf. multigrid solvers)
- By casting it in language of neural networks it is easy to investigate non-Krylov methods.
- Focus on explicitly gauge-equivariant models such that gauge-equivariance does not have to be learned. Helps with transfer learning.

Preconditioning

- First study Dirac equation

$$Du = b$$

- Time to solution is determined by condition number of Dirac matrix
  - Condition number increases dramatically in physical quark-mass and continuum limit
- Can be addressed by Preconditioning
  - Find a preconditioner M such that $M \approx D^{-1}$
  - Define $v = M^{-1}u$ and use

    $$DMM^{-1}u = (DM)v = b$$

    to solve for v with preconditioned matrix DM (smaller condition number)
  - Then $u = Mv$

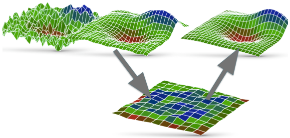
Low and high modes

- Consider the eigendecomposition of D

$$D = \sum_n \lambda_n |n\rangle\langle n|$$

Preconditioner should approximate low-mode and high-mode components of D^{-1}. Needs to be adaptive but would be nice to have geometric version (see later).

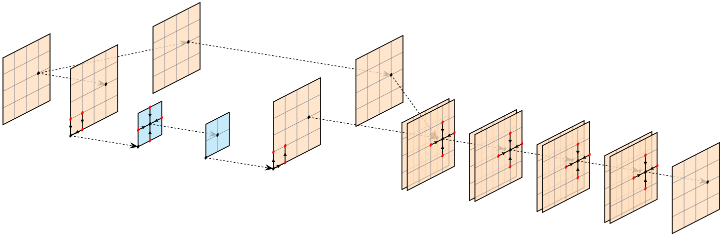
- State-of-the-art algorithms (multigrid) are designed to do this



- We will follow this paradigm, but here we learn the preconditioner

Source: https://summerofhpc.prace-ri.eu/multithreading-the-multigrid-solver-for-lattice-qcd

A model to implement a multigrid preconditioner



Gauge-equivariant layers

Parallel transport

- Consider a field $\phi(x)$ with $x \in S$ (space-time lattice, dim $= d$) and $\phi \in V_I = V_G \otimes V_{\bar{G}}$
(gauge space: $V_G = \mathbb{C}^N$, non-gauge space: $V_{\bar{G}} = \mathbb{C}^{\bar{N}}$)

- Also consider an SU(N) gauge field $U_\mu(x)$ acting on V_G

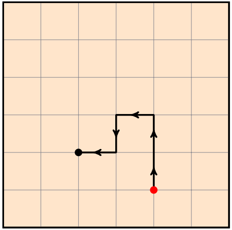
- Define the parallel-transport operator for a path $p = p_1, \ldots, p_{n_p}$ with $p_i \in \{\pm 1, \ldots, \pm d\}$

$$T_p = H_{p_{n_p}} \cdots H_{p_2} H_{p_1} \qquad \text{with} \qquad H_\mu \phi(x) = U^\dagger_\mu(x - \hat{\mu})\phi(x - \hat{\mu})$$

- H_μ transports information by a single hop in direction $\hat{\mu}$

- H_μ acts on field; new field $H_\mu \phi$ is evaluated at x

- Example: $T_p = H_{-1} H_{-2} H_{-1} H_2 H_2$



Gauge equivariance

- A gauge transformation by $\Omega(x) \in \mathrm{SU}(N)$ acts in the usual way

$$\begin{aligned} \phi(x) &\to \Omega(x)\phi(x) \\ U_\mu(x) &\to \Omega(x)U_\mu(x)\Omega^\dagger(x+\hat{\mu}) \end{aligned}$$

- Such gauge transformations commute with T_p for any path p

$$T_p\phi(x) \to \Omega(x)T_p\phi(x)$$

Parallel-transport convolutions

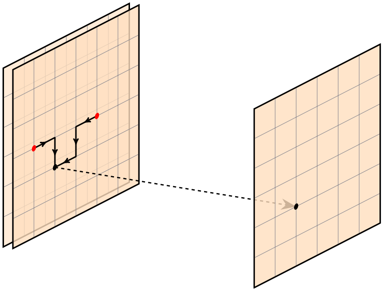
- Parallel-transport convolution layer and local parallel-transport convolution layer

$$\psi_a(x) \stackrel{\text{PTC}}{=} \sum_b \sum_{p\in P} W_a^{bp} T_p \phi_b(x) \qquad \psi_a(x) \stackrel{\text{LPTC}}{=} \sum_b \sum_{p\in P} W_a^{bp}(x) T_p \phi_b(x)$$

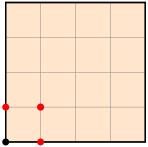
  - $a =$ output feature index
  - $b =$ input feature index
  - $P =$ set of paths
  - W_a^{bp} acts in $V_{\bar{G}}$ (here: 4×4 spin matrix)
  - Elements of W: "layer weights"

- Layers are gauge-equivariant

- No activation function since we want to learn a linear preconditioner; will be different for correlators.



Explicit gauge degree of freedom on coarse grid

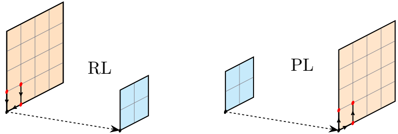


$B(y) = \{\bullet, \bullet\}$

$B_r(y) = \bullet$

- Field on fine grid: $\phi : S \to V_G \otimes V_{\bar{G}}, x \mapsto \phi(x)$ with local gauge space (V_G), non-gauge space ($V_{\bar{G}}$), and set of fine-grid sites S
- Gauge transformation: $\phi(x) \to \Omega(x)\phi(x)$
- Set of coarse sites $\tilde{S}$ and block map $B : \tilde{S} \to \mathcal{P}(S), y \mapsto B(y)$ (sites $B(y)$ on fine grid correspond to y on coarse grid)
- A reference site $B_r : \tilde{S} \to S, y \mapsto B_r(y)$ such that $B_r(y) \subset B(y)$
- Field on coarse grid: $\tilde{\phi} : \tilde{S} \to V_G \otimes \tilde{V}_{\bar{G}}, y \mapsto \tilde{\phi}(y)$ (note: same local gauge space as on fine grid)
- Find restriction and prolongation layers such that $\tilde{\phi}(y) \to \tilde{\Omega}(y)\tilde{\phi}(y)$ under gauge transformation Ω with

$$\tilde{\Omega}(y) = \Omega(B_r(y)) .$$



- Define RL/PL by pooling and subsampling layers:

$$\mathrm{RL} = \mathrm{SubSample} \circ \mathrm{Pool}\,, \tag{1}$$

$$\mathrm{PL} = \mathrm{Pool}^\dagger \circ \mathrm{SubSample}^\dagger\,. \tag{2}$$

(weights in RL and PL can differ, so not necessarily $\mathrm{RL}^\dagger = \mathrm{PL}$)

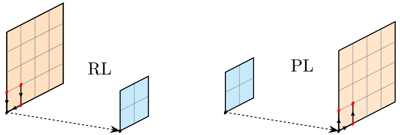
- The pooling layer Pool: $\mathcal{F}_\phi \to \mathcal{F}_\phi$, $\phi \mapsto \mathrm{Pool}\phi$ is given by

$$\mathrm{Pool}\phi(x) = \sum_{q\in Q} W_q(x)\, T_q \phi(x) \tag{3}$$

with $q = (p, \bar{U})$, path p, gauge field $\bar{U}$, and $T_q = T_p(\bar{U})$. Weights $W_q(x)$ are spin matrices, separated gauge DOF.

- The subsampling layer is given by

$$\mathrm{SubSample}\phi(y) = \phi(B_r(y))\,. \tag{4}$$



- Gauge field $\bar{U}$ in $T_p(\bar{U})$ needs to satisfy

$$\bar{U}_\mu(x) \to \Omega(x)\bar{U}_\mu(x)\Omega^\dagger(x+\hat{\mu})\,. \tag{5}$$

  In practice, we use a variety of differently smeared links.

- Complete set of paths P transports every element of $B(y)$ exactly once to $B_r(y)$ $\Rightarrow |P| = |B(y)|$

- Efficient implementation for each complete set of path possible: GPT

- $\tilde{\phi} = \mathrm{RL}\phi$ yields $\tilde{\phi}(y) \to \tilde{\Omega}(y)\tilde{\phi}(y)$ under gauge transformations $\phi(x) \to \Omega(x)\phi(x)$

Model details I

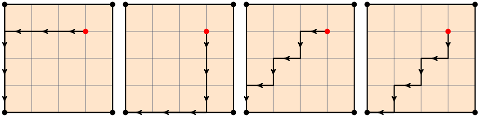
- Need prescription for q in

$$\mathrm{Pool}\phi(x) = \sum_{q \in Q} W_q(x) T_q \phi(x)$$

with $q = (p, \bar{U})$, path p, gauge field $\bar{U}$, and $T_q = T_p(\bar{U})$.

- For fixed i, we define paths $p^{(ij)}$ that connect all elements of $B(y)$, enumerated by $j = 1, \dots, |B(y)|$, to the reference site $B_r(y)$. For different i we use different prescriptions for the paths $p^{(ij)}$, and then use the couples $q_{ij} = (p^{(ij)}, \bar{U}^{(i)})$.

- We define four different prescriptions $\hat{p}_1, \dots, \hat{p}_4$ (depth first/breadth first lexicographic/reverse lexicographic)



and set $p^{(ij)} = \hat{p}^{(j)}_{i \bmod 4}$.

Model details II

- Concretely, we use 9 different gauge fields $\bar{U}^{(i)}$ with $i = 1, \ldots, 9$. We construct the $\bar{U}^{(i)}$ by applying $i(i-1)/2$ steps of $\rho = 0.1$ stout smearing to the unsmeared gauge fields U. Smearing radius proportional to $\sqrt{i(i-1)}$.

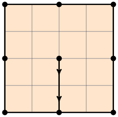
- So we have 9 different spin-matrix fields $W_1(x), \ldots, W_9(x)$.

- In practice, sufficient to use same weights in PL and RL such that $\mathrm{PL} = \mathrm{RL}^\dagger$. Found no benefits from general case.

- Coarse-grid size $2^3 \times 4$

Explicit gauge-equivariant coarse layers need coarse gauge field

- Plain coarse gauge field construction:



$$B_r(y') - B_r(y) = b\hat{\mu}$$

with unit vector $\hat{\mu}$ in direction μ and $b \in \mathbb{N}^+$. The coarse-grid gauge field $\tilde{U}_\mu(y)$ corresponding to this pair of reference points is then simply

$$\tilde{U}_\mu(y) = U_\mu(B_r(y)) \cdots U_\mu(B_r(y) + (b-1)\hat{\mu}) \,. \tag{6}$$

- Galerkin coarse gauge field construction:

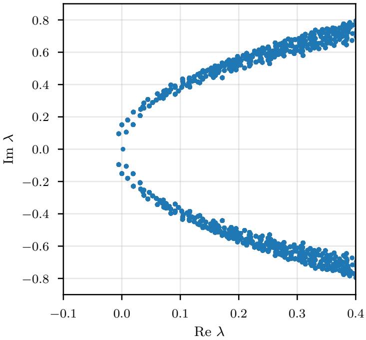
$$\tilde{U}_\mu(y) = \tilde{D}(y, y + \hat{\mu}) \tag{7}$$

with

$$\tilde{D} = \mathrm{RL} \circ D \circ \mathrm{PL} \tag{8}$$

for Wilson-clover D.

Spectrum of Wilson-clover Dirac operator



- $\beta = 6$ pure Wilson gauge field with topological charge $Q = 1$
- $8^3 \times 16$ lattice sites
- Wilson-clover operator with $m = -0.5645$ and $c_{\mathrm{sw}} = 1$

Training setup – How to train RL/PL?

- Obvious approach: train

$$\mathrm{PL} \circ \mathrm{RL} \tag{9}$$

as an autoencoder with training vectors from the near-null space.

- This could be done with a cost function

$$C = |\mathrm{PL} \circ \mathrm{RL} v_\ell - v_\ell|^2 \tag{10}$$

with fine-grid vectors v_ℓ. For each training step we select a random element of $v_\ell \in \{u_1, \ldots, u_s\}$ of the near-null space vectors u_i defined above.

- Use Adam optimizer.

- Result: did not perform well in MG preconditioner!

Training setup – How to train RL/PL?

- What was missing: $\mathrm{PL} \circ \mathrm{RL}$ should also project high eigenmodes to zero (if not could overload smoother layers)

- Found also additional benefit from encouraging $\mathrm{RL} \circ \mathrm{PL} = \mathbb{1}$ such that we have a proper projection operator $P = \mathrm{PL} \circ \mathrm{RL}$ with $P^2 = P$.

- We implement this strategy by using the cost function

$$C = |\mathrm{PL} \circ \mathrm{RL} v_\ell - v_\ell|^2 + |\mathrm{PL} \circ \mathrm{RL} v_h - P_\ell v_h|^2 + |\mathrm{RL} \circ \mathrm{PL} v_c - v_c|^2 \qquad (11)$$

  with additional fine-grid vector v_h and coarse-grid vector v_c. For each training step v_h and v_c are random vectors with elements normally distributed about zero.
  P_ℓ is the blocked low-mode projector

$$P_\ell = W^\dagger W \,, \qquad\qquad W(y,x)^\dagger = \sum_{i=1}^{s} \bar{u}_i^y(x) \hat{e}_i^\dagger \qquad (12)$$

  with block-orthonormalized $\bar{u}_i$ from u_i.

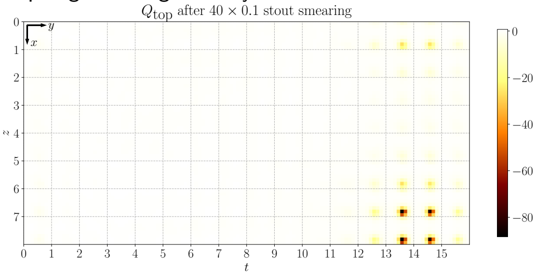
- All vectors v_ℓ, v_h, and v_c are normalized to unit length before being used in the cost function.

Training setup – How to train RL/PL?

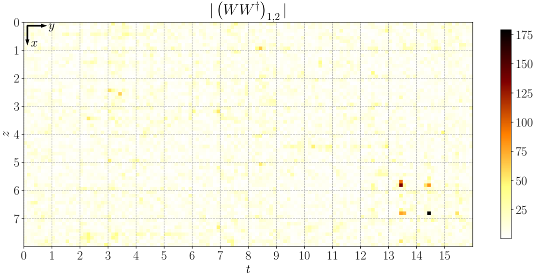
- Train with $s = 4$.
- Training converged after O(1000) steps.
- Yields $W_1(x), \dots, W_9(x)$ but still costly since we first need near-null space vectors.
- In future work: obtain $W_i(x)$ as output of gauge-invariant models based on energy density $E(x)$, topology density $Q(x)$, plaquette $P(x)$ and other Wilson loops. At this point the u_i are no longer needed. (In a sense we generate training data for the next step in this work.)

Constructing the architecture for the W model

Topological charge density:

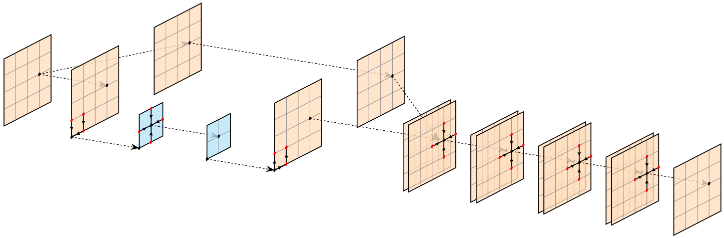


Density of W_8:



Plots by Daniel Knüttel

Training setup – combined preconditioner model

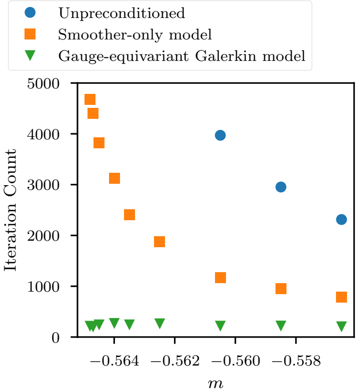


- First train RL/PL as described above.
- Then train combined model with frozen RL/PL using cost function

$$C = |Mb_h - u_h|^2 + |Mb_\ell - u_\ell|^2 \tag{13}$$

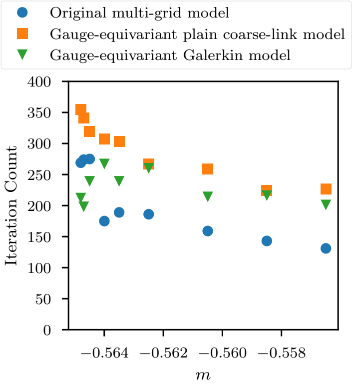
  with $b_h = Dv_1$, $u_h = v_1$, $b_\ell = v_2$, and $u_\ell = D^{-1}v_2$.
- Further training with unfrozen RL/PL leads to no notable improvement.

Results – critical slowing down



- Show outer iteration count in GMRES to 10^{-8} precision with and without model as preconditioner.
- Model with Galerkin gauge fields removes critical slowing down.

Results – critical slowing down



- Original multi-grid model also removes critical slowing down.
- Model with plain gauge fields shows small remnants of critical slowing down.

The gauge-equivariant multigrid neural network research program

- Future work:
  - Relate RL/PL spin matrices to energy density, topology density, Wilson loops via gauge-invariant models. This would eliminate most of the typical multigrid setup cost. Useful for ensemble generation.
  - Address fermions with more complex spectrum (such as DWF)
  - Do not just approximate D^{-1} but directly complex hadronic correlation functions to be used in AMA.

Related work

1. Neural networks for multigrid (but not for gauge theories), e.g.,
   - Katrutsa, Daulbaev, Oseledets arXiv:1711.03825 [math.NA]
   - He & Xu arXiv:1901.10415 [cs.CV]
   - Greenfeld, Galun, Basri, Yavneh, Kimmel arXiv:1902.10248 [cs.LG]
   - Eliasof, Ephrath, Ruthotto, Treister arXiv:2011.09128 [cs.CV]
   - Huang, Li, Xi arXiv:2102.12071 [math.NA]
2. Gauge-equivariant neural networks (but not for solving Dirac equation), e.g.,
   - Cohen, Weiler, Kicanaoglu, Welling arXiv:1902.04615 [cs.LG]
   - Finzi, Stanton, Izmailov, Wilson arXiv:2002.12880 [stat.ML]
   - Luo, Carleo, Clark, Stokes arXiv:2012.05232 [cond-mat.str-el]
   - Kanwar et al. arXiv:2003.06413 [hep-lat]
   - Boyda et al. arXiv:2008.05456 [hep-lat]
   - Favoni, Ipp, Müller, Schuh arXiv:2012.12901 [hep-lat]
   - Abbott et al. arXiv:2207.08945 [hep-lat]

Related work

3. Multigrid algorithms in lattice QCD
   - Brannick, Brower, Clark, Osborn, Rebbi arXiv:0707.4018 [hep-lat]
   - R. Babich et al. arXiv:1005.3043 [hep-lat]
   - Frommer et al. arXiv:1303.1377 [hep-lat]
   - Boyle arXiv:1402.2585 [hep-lat]
   - Brannick et al. arXiv:1410.7170 [hep-lat]
   - Yamaguchi & Boyle arXiv:1611.06944 [hep-lat]
   - Brower, Clark, Strelchenko, Weinberg arXiv:1801.07823 [hep-lat]
   - Brower, Clark, Howarth, Weinberg arXiv:2004.07732 [hep-lat]
   - Boyle & Yamaguchi arXiv:2103.05034 [hep-lat]
4. Neural networks as preconditioners in gauge theories without multigrid and gauge equivariance
   - Calì, Hackett, Lin, Shanahan, Xiao arXiv:2208.02728 [hep-lat]