
Machine Learning Theory and Mathematics in the Era of Infinite GPUs

Yiping Lu

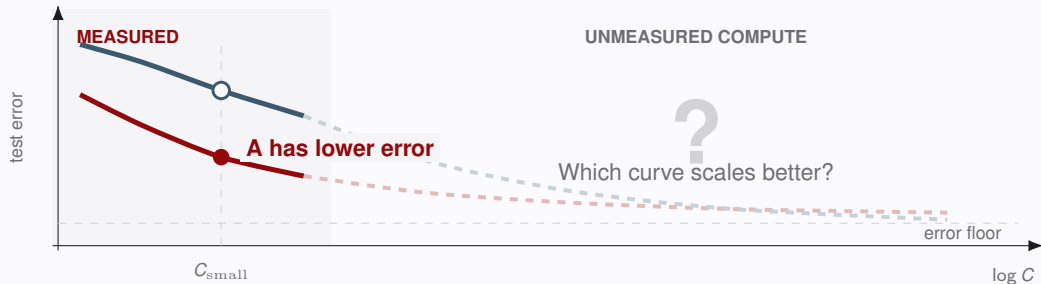
Beijing International Center for Mathematical Research

Peking University

Industrial Engineering & Management Sciences, Northwestern University



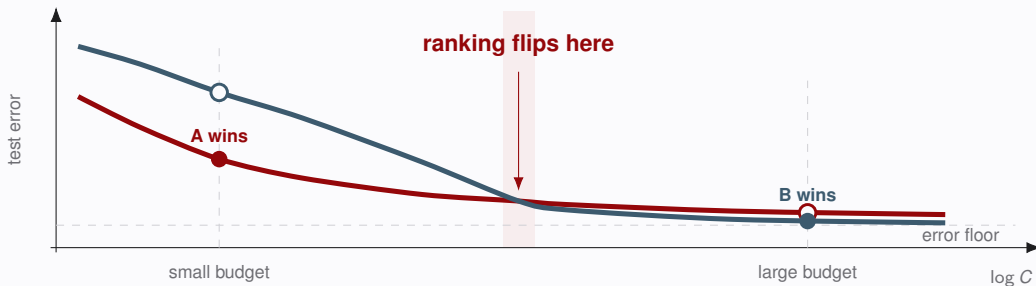
One budget reveals only the local winner



A wins this benchmark. A single budget ranks methods here, but it does not reveal which method improves faster as compute grows.

Source: Schematic curves for illustration; no empirical measurements are shown.

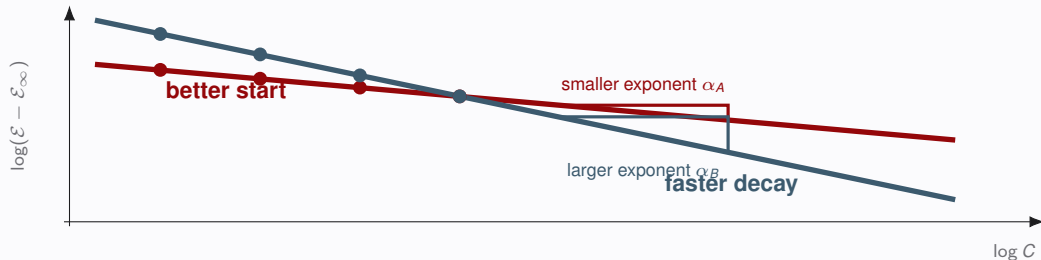
Scale can reverse the ranking



The winner is budget dependent. A can start with lower error while B converts additional compute into faster improvement.

Source: Schematic ranking reversal; no empirical claim is made about either algorithm.

The exponent can overturn the first point ranking



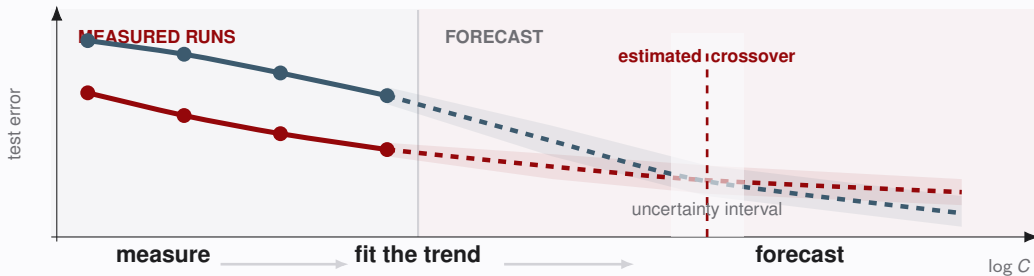
$$\varepsilon_j(C) - \varepsilon_\infty = A_j \left(\frac{C}{C_0} \right)^{-\alpha_j}$$

α_j measures return on compute.

The picture assumes a shared floor and an exact power law. With different floors, a crossing can move or disappear.

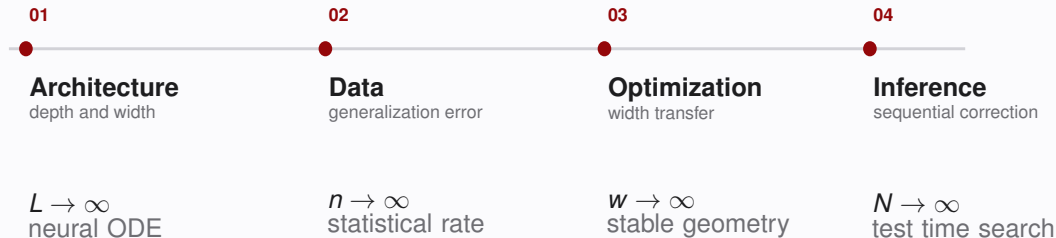
Source: Stylized shared floor power law; curves are schematic rather than empirical.

Scaling laws turn measured runs into a forecast



If we have infinite token, to prove RH we need
**hire more clever researchers, or merely Shakespeare's
monkey?**

My research arc follows four scaling limits



One question across four limits

How should architecture, data, optimization, and inference scale with compute?

Architecture scaling: depth and width limits

DEPTH LIMIT

$$x_{k+1} = x_k + \frac{1}{L} f_{\theta_k}(x_k)$$

With $t = k/L$ and a stable residual scale,

$$\dot{x}(t) = f_{\theta(t)}(x(t)), \quad x(0) = x_0.$$

Neural ODE viewpoint

Depth scaling is meaningful only after choosing the step size.

WIDTH LIMIT

$$f_m(x) = \frac{1}{m} \sum_{j=1}^m a_j \sigma(w_j^\top x),$$

$$\rho_m = \frac{1}{m} \sum_{j=1}^m \delta_{(a_j, w_j)}.$$

As $m \rightarrow \infty$,

$$f_\rho(x) = \int a \sigma(w^\top x) d\rho(a, w),$$

and training becomes a flow of measures.

Source: Lu et al., ICML 2018, *Beyond Finite Layer Neural Networks*; Lu et al., ICML 2020, *A Mean-field Analysis of Deep ResNet and Beyond*.

Data scaling: three minimax laws

PINNS

ICLR 2022

Machine Learning for Elliptic PDEs

Fast Rate Generalization Bound, Neural Scaling Law and Minimax Optimality

For a prototype elliptic PDE, fast rate bounds prove PINNs and a modified Deep Ritz method minimax optimal up to logarithmic factors; data scaling is observed empirically.

OPERATOR

LEARNING

ICLR 2023

Minimax Optimal Kernel Operator Learning via Multilevel Training

A multilevel kernel estimator matches the information theoretic minimax rate for Hilbert–Schmidt operator learning between infinite dimensional Sobolev RKHSs.

QUADRATURE

NeurIPS 2023

When Can Regression-Adjusted Control Variates Help?

Rare Events, Sobolev Embedding and Minimax Optimality

For Sobolev function moments, control variates are minimax optimal when smoothness rules out rare extremes; otherwise truncation is optimal and the rate gain vanishes.

Source: Lu et al., ICLR 2022, arXiv:2110.06897; Jin et al., ICLR 2023, arXiv:2209.14430; Blanchet et al., NeurIPS 2023, arXiv:2305.16527.

Today: optimization and inference scaling

EARLIER WORK

01

ARCHITECTURE

$$L, w \rightarrow \infty$$

depth and width limits

02

DATA

$$n \rightarrow \infty$$

generalization rates

TODAY

03

OPTIMIZATION

$$W \rightarrow \infty$$

learning rate transfer

04

INFERENCE

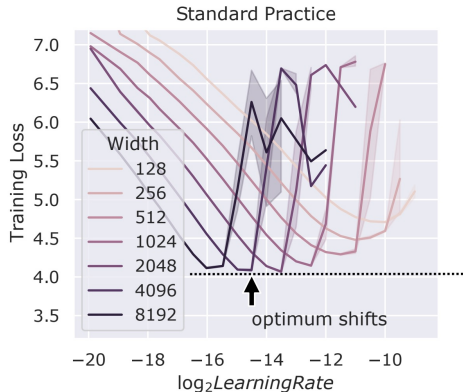
$$N \rightarrow \infty$$

correction and
search

Pretraining time scaling

Optimizer geometry should make learning rates transferable across width.

Hyperparameter transfer is an optimizer problem



Source: Yang et al., hyperparameter transfer experiments;

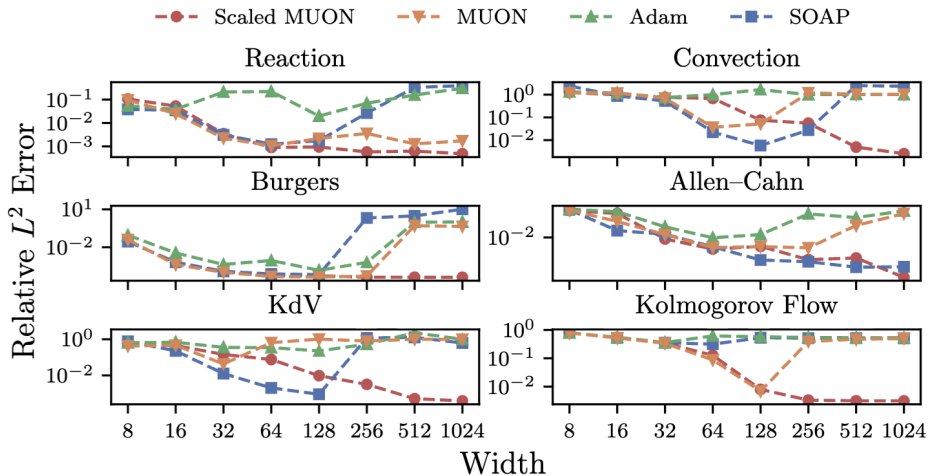
OBSERVED FAILURE MODE

Under the baseline parameterization shown in this sweep, the best learning rate shifts downward with width. A rate tuned at width 512 can diverge or slow dramatically at width 2048.

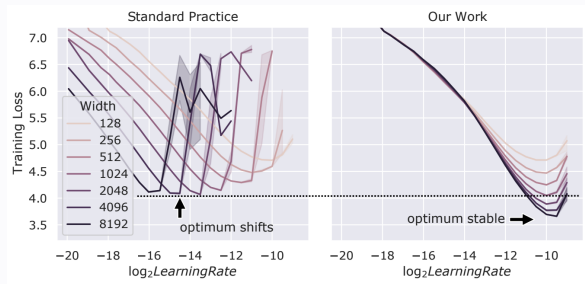
Why this matters at LLM scale

Large pretraining runs are too expensive for a full hyperparameter sweep at every model size. Transfer is therefore part of the optimizer design problem, not an afterthought.

Scaling PINN



Existing transfer rules



Source: Yang et al., maximal update parameterization; MOGA motivation.

ADAM LEARNING RATE SCALING

For a weight matrix with effective input dimension d_{in} ,

$$\eta_W = \frac{1}{d_{\text{in}}} \eta_{\text{base}}.$$

Embedding matrices fix the learning rate.

The theoretical gap

Existing derivations are architecture specific, rely on detailed random matrix arguments, and are typically justified near initialization.

Problem setup: width and parameter geometry

We study

$$\min_{\Theta \in \mathcal{C}} f(\Theta), \quad f(\Theta) = \mathcal{L}(y_\ell(x)), \quad \Theta = \{W_{1:\ell}, b_{1:\ell}\},$$

for the feedforward recursion

$$y_i(x) = \sigma(W_i y_{i-1}(x) + b_i), \quad i = 1, \dots, \ell,$$

with

$$W_1 \in \mathbb{R}^{w \times d}, \quad W_i \in \mathbb{R}^{w \times w} \ (2 \leq i \leq \ell - 1), \quad W_\ell \in \mathbb{R}^{1 \times w}.$$

Scaling regime

Depth, input dimension, and output dimension are fixed; hidden width $w \rightarrow \infty$.

Source: MOGA paper, neural network optimization setup.

Central question

Which norm on Θ keeps both forward sensitivity and curvature controlled as w grows?

Decoupled weight decay is an implicit constraint

Let d^t be the normalized update direction. Consider

$$\theta^{t+1} = (1 - \eta\lambda)\theta^t - \eta d^t, \quad 0 \leq \eta\lambda \leq 1.$$

Invariant norm ball

If $\|d^t\| \leq \lambda R$ and $\|\theta^0\| \leq R$, then

$$\|\theta^{t+1}\| \leq (1 - \eta\lambda)\|\theta^t\| + \eta\|d^t\| \leq R,$$

so $\|\theta^t\| \leq R$ for every $t \geq 0$.

INTERPRETATION This is a sufficient invariance condition: weight decay suggests analyzing optimization inside a norm ball.

Source: Loshchilov and Hutter, AdamW; invariant ball interpretation used in the MOGA analysis.

AdamW reveals sign geometry

$$\begin{aligned}
 g^t &= \nabla f(\theta^{t-1}; \xi_t), \\
 m^t &= \beta_1 m^{t-1} + (1 - \beta_1) g^t, \\
 v^t &= \beta_2 v^{t-1} + (1 - \beta_2) g^t \odot g^t, \\
 \hat{m}^t &= m^t / (1 - \beta_1^t), \quad \hat{v}^t = v^t / (1 - \beta_2^t), \\
 \theta^t &= \theta^{t-1} - \eta \frac{\hat{m}^t}{\sqrt{\hat{v}^t} + \epsilon} - \eta \lambda \theta^{t-1}.
 \end{aligned}$$

Idealized direction

$$\frac{g^t}{\sqrt{g^t \odot g^t}} = \text{sign}(g^t).$$

At $\beta_1 = \beta_2 = 0$ and $\epsilon \rightarrow 0$, Adam's normalized direction is elementwise sign.

Source: Kingma and Ba, Adam; Loshchilov and Hutter, AdamW; Bernstein et al., signSGD.

Lion keeps sign geometry and changes the momentum rule

Sign update with two momentum coefficients and weight decay

$$\begin{aligned}g^t &= \nabla f(\theta^{t-1}; \xi_t), \\c^t &= \beta_1 m^{t-1} + (1 - \beta_1) g^t, \\ \theta^t &= (1 - \eta \lambda) \theta^{t-1} - \eta \operatorname{sign}(c^t), \\m^t &= \beta_2 m^{t-1} + (1 - \beta_2) g^t.\end{aligned}$$

GEOMETRY

The update remains an elementwise sign direction.

Source: Chen et al., Symbolic Discovery of Optimization Algorithms, 2023.

PRACTICAL MOTIVATION

Lion was discovered through program search and uses less optimizer state than AdamW.

Muon replaces elementwise sign with matrix sign

For a matrix parameter block,

$$G^t = \nabla f(\Theta^{t-1}; \xi_t),$$

$$M^t = \beta_1 M^{t-1} + (1 - \beta_1) G^t,$$

$$\Theta^t = \Theta^{t-1} - \eta \text{matrixsign}(M^t) - \eta \lambda \Theta^{t-1}.$$

Matrix sign

If $G = U\Sigma V^\top$ is a compact singular value decomposition, then

$$\text{matrixsign}(G) = UV^\top.$$

It is the spectral geometry analogue of the elementwise sign and can be approximated by Newton–Schulz iterations.

Muon currently has competing scaling rules

μ P variant

$$\eta_{\text{Muon}} = \sqrt{\frac{d_{\text{out}}}{d_{\text{in}}}} \eta_{\text{base}}.$$

The grafting is expressed in spectral geometry; η_{base} is tuned on a small model.

Moonlight variant

$$\eta_{\text{Muon}} = \sqrt{\max\{d_{\text{out}}, d_{\text{in}}\}} \eta_{\text{Adam}}.$$

This uses Euclidean grafting and implies an $O(w^{-1/2})$ peak learning rate when the Adam rate scales as $O(w^{-1})$.

Which rule follows from the geometry of the network itself?

Source: Yang et al., spectral condition; Liu et al., Moonlight; Agarwal et al., grafting.

Steepest descent unifies normalized update maps

For norms on the input and output spaces, define

$$\|D\|_{\text{in} \rightarrow \text{out}} = \sup_{\|x\|_{\text{in}}=1} \|Dx\|_{\text{out}}.$$

Given a gradient matrix G , define the normalized update map

$$\Psi_{\|\cdot\|}(G) \in \arg \max_{\|D\| \leq 1} \langle G, D \rangle, \quad W^+ = W - \eta \Psi_{\|\cdot\|}(G).$$

Then

$$\langle G, \Psi_{\|\cdot\|}(G) \rangle = \|G\|_*.$$

Matrix thinking

The optimizer is determined by the parameter norm, its dual gradient norm, and a tractable oracle for $\Psi(G)$.

Sign updates and Muon are two operator geometries

Sign family

SignSGD and Lion use this oracle;
AdamW reaches it only in the idealized
limit on the preceding slide.

$$\|D\|_{1 \rightarrow \infty} = \max_{i,j} |D_{ij}|,$$

$$\Psi_{1 \rightarrow \infty}(G) = \text{sign}(G).$$

The unit ball constrains every matrix
entry.

Muon

$$\|D\|_{2 \rightarrow 2} = \|D\|_{\text{spectral}},$$

$$\Psi_{2 \rightarrow 2}(G) = UV^{\top}, \quad G = U\Sigma V^{\top}.$$

The unit ball constrains singular values.

The same steepest descent template produces very different normalized directions because the matrix geometry changes.

Column normalization is steepest descent in $\ell_1 \rightarrow \ell_q$

For $2 \leq q < \infty$, let q^* satisfy $1/q + 1/q^* = 1$. Then

$$\|D\|_{1 \rightarrow q} = \max_c \|D_{:,c}\|_q.$$

The normalized update is computed independently for each column:

$$\Psi_{1 \rightarrow q}(G)_{:,c} = \text{colnorm}_q(G)_{:,c} := \frac{\text{sign}(G_{:,c}) \odot |G_{:,c}|^{q^*-1}}{\|G_{:,c}\|_{q^*}^{q^*-1}}.$$

Set the oracle to zero on a zero gradient column. At $q = \infty$, use the sign endpoint $\Psi_{1 \rightarrow \infty}(G) = \text{sign}(G)$ from the preceding slide.

Interpretation

Every nonzero gradient column saturates its ℓ_q constraint in the direction selected by ℓ_{q^*} duality.

Row normalization is steepest descent in $\ell_p \rightarrow \ell_\infty$

For $1 < p < \infty$, let p^* satisfy $1/p + 1/p^* = 1$. Then

$$\|D\|_{p \rightarrow \infty} = \max_r \|D_{r,:}\|_{p^*}.$$

The normalized update is computed independently for each row:

$$\Psi_{p \rightarrow \infty}(G)_{r,:} = \text{rownorm}_p(G)_{r,:} := \frac{\text{sign}(G_{r,:}) \odot |G_{r,:}|^{p-1}}{\|G_{r,:}\|_p^{p-1}}.$$

Set the oracle to zero on a zero gradient row.

Interpretation

Every nonzero gradient row saturates its ℓ_{p^*} constraint in the direction selected by ℓ_p duality.

Matrix thinking: the optimizer is a norm choice

Update family	Parameter geometry	Normalized update map
Sign family	$\ \cdot\ _{1 \rightarrow \infty}$	SignSGD and Lion; idealized AdamW
Column norm.	$\ \cdot\ _{1 \rightarrow q}$	columnwise dual normalization
Row norm.	$\ \cdot\ _{p \rightarrow \infty}$	rowwise dual normalization
Muon	$\ \cdot\ _{2 \rightarrow 2}$	matrix sign
ORACLE Can $\Psi(G)$ be computed cheaply?	COMPOSITION Do adjacent layer norms interface without width factors?	CURVATURE Is the loss smoothness stable as width grows?

Classical norms pay a width factor across layers

One layer gives

$$\begin{aligned}\|\delta y_{i+1}\|_q &\leq \|W_{i+1}\|_{p \rightarrow q} \|\delta y_i\|_p. \\ \|\delta y_{i+2}\|_q &\leq \|W_{i+2}\|_{p \rightarrow q} \|\delta y_{i+1}\|_p.\end{aligned}$$

To apply the next operator bound, the same intermediate feature must be reinterpreted from ℓ_q to ℓ_p :

$$\|\delta y_{i+1}\|_p \leq \underbrace{w^{1/p-1/q}}_{\text{embedding gap}} \|\delta y_{i+1}\|_q, \quad p \leq q.$$

Hence two layers already produce

$$\|\delta y_{i+2}\|_q \leq w^{1/p-1/q} \|W_{i+2}\|_{p \rightarrow q} \|W_{i+1}\|_{p \rightarrow q} \|\delta y_i\|_p.$$

The obstruction

The optimizer geometry may be well behaved for one matrix while still accumulating powers of width across a network.

Mean normalized norms repair the layer interface

MEAN NORMALIZED VECTOR NORM

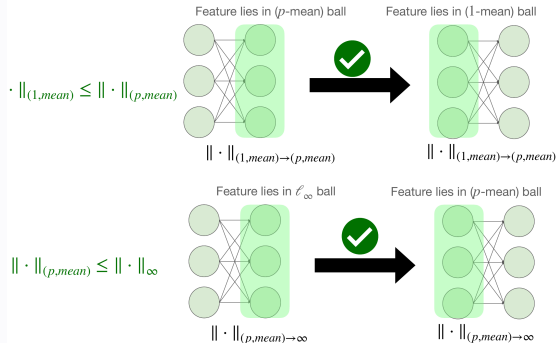
For $1 \leq p < \infty$,

$$\|x\|_{(p,\text{mean})} = \left(\frac{1}{w} \sum_{j=1}^w |x_j|^p \right)^{1/p} = w^{-1/p} \|x\|_p.$$

For $p = \infty$, set $\|x\|_{(\infty,\text{mean})} := \|x\|_\infty$.

For $1 \leq p \leq q \leq \infty$,

$$\|x\|_{(p,\text{mean})} \leq \|x\|_{(q,\text{mean})}.$$



The normalization cancels the finite dimensional embedding penalty at each layer interface.

Source: MOGA paper, mean normalized norms and width stable composition.

The forward map is width independent Lipschitz

Define the block norm

$$\|\Theta\|_{\text{block}} = \max \left\{ \|W_1\|_{1 \rightarrow (q, \text{mean})}, \max_{2 \leq i \leq \ell-1} \|W_i\|_{(p, \text{mean}) \rightarrow (q, \text{mean})}, \|W_\ell\|_{(p, \text{mean}) \rightarrow \infty}, \max_i \|b_i\|_\infty \right\}.$$

Width independent Lipschitz bound

Fix the depth and suppose $1 \leq p \leq q < \infty$, inputs are bounded, and the first derivatives of σ and $\mathcal{L}$ are uniformly bounded. On

$$\Omega_C = \{\Theta : \|\Theta\|_{\text{block}} \leq C\},$$

the loss satisfies

$$|f(\Theta) - f(\Theta')| \leq M_{\text{net}} \|\Theta - \Theta'\|_{\text{block}},$$

where M_{net} is independent of every hidden width.

When the smoothness certificate is width stable

With respect to a norm $\|\cdot\|$, L smoothness means

$$\|\nabla f(\Theta) - \nabla f(\Theta')\|_* \leq L\|\Theta - \Theta'\|.$$

Width dependence of smoothness

Fix $1 \leq p \leq q \leq \infty$. Under the same fixed-depth, bounded-input, and $\Theta, \Theta' \in \Omega_C$ assumptions, with uniformly bounded first and second derivatives of σ and $\mathcal{L}$,

$$\|\nabla f(\Theta) - \nabla f(\Theta')\|_{\text{block},*} \leq L_{\text{net}} w^{\max\{0, 2/q-1/p\}} \|\Theta - \Theta'\|_{\text{block}},$$

where L_{net} is independent of width.

Transferable regime

$q \geq 2p$ removes the stated worst case width factor from the smoothness bound.

The condition $q \geq 2p$ comes from nonlinear feature interactions

Featurewise nonlinearities create Hadamard products in second order perturbations.

Mean normalized Hadamard inequality

For $x, y \in \mathbb{R}^w$,

$$\|x \odot y\|_{(p, \text{mean})} \leq w^{\max\{0, 2/q - 1/p\}} \|x\|_{(q, \text{mean})} \|y\|_{(q, \text{mean})}.$$

Consequently,

$$\begin{aligned} |\nabla^2 f(W)[\Delta W^1, \Delta W^2]| &\lesssim \|(\Delta W^1 x) \odot (\Delta W^2 x)\|_{(p, \text{mean})} \\ &\lesssim w^{\max\{0, 2/q - 1/p\}} \|\Delta W^1 x\|_{(q, \text{mean})} \|\Delta W^2 x\|_{(q, \text{mean})}. \end{aligned}$$

Source: MOGA paper, Hadamard product lemma and Hessian reduction.

Width aware learning rate scaling

For $D \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$,

$$\|D\|_{(p,\text{mean}) \rightarrow (q,\text{mean})} = \frac{d_{\text{in}}^{1/p}}{d_{\text{out}}^{1/q}} \|D\|_{p \rightarrow q},$$

Adam endpoint $(p, q) = (1, \infty)$

Muon shape $(p, q) = (2, 2)$

Curvature $L(w) = O(\sqrt{w})$

$s_{1,\infty} = 1/d_{\text{in}}$, the algebraic μP factor.

$s_{2,2} = \sqrt{d_{\text{out}}/d_{\text{in}}}$, the layer shape factor.

Holding the dual descent signal comparable, it suggests a conservative normalized base step of order $w^{-1/2}$ for square layers.

Recover...

- **Width Transfer:** recover μP limit
- **Depth Transfer:** recover complete-P limit (on going)

MOGA optimizer: momentum and column geometry

PART I

Algorithm 1 MOGA: Matrix-Operator-Geometry-Aware Steepest Descent

Require: Learning rates $\{\eta^t\}_{t \geq 0}$, momentum parameters $\beta_1, \beta_2 \in (0, 1)$, initial parameters Θ^0

- 1: $M^0 \leftarrow \mathbf{0}$ ▷ M^t has the same block structure as Θ^t
- 2: **for** $t = 1, 2, \dots$ **do**
- 3: Sample stochastic gradient: $G \leftarrow \nabla F(\Theta^{t-1}; \xi)$, $\xi \sim \mathbb{P}$
- 4: Exponential moving average: $M^t \leftarrow \beta_1 M^{t-1} + (1 - \beta_1)G$
- 5: Lookahead (Nesterov-type) momentum: $\tilde{M}^t \leftarrow \beta_2 M^{t-1} + (1 - \beta_2)G$
- 6: **if** Descent under $(1, \text{mean}) \rightarrow (q, \text{mean})$ **then** ▷ $q \geq 2$
- 7: **for** $i = 1$ to ℓ **do**
- 8:
$$W_i^t \leftarrow W_i^{t-1} - \eta^t \cdot \frac{(d_{\text{out}}^i)^{1/q}}{d_{\text{in}}^i} \cdot \text{colnorm}_q \left(\tilde{M}_{W_i^{t-1}}^t \right)$$
- 9:
$$b_i^t \leftarrow b_i^{t-1} - \eta^t \text{sign} \left(\tilde{M}_{b_i}^t \right)$$
 ▷ or steepest descent under (q, mean) norm
- 10: **end for**

The boxed prefactor converts the column oracle into mean geometry; bias blocks use a sign update.

Source: MOGA paper, Algorithm 1, lines 1 to 10.

MOGA optimizer: row geometry and loop closure

PART II

```

11:   else if Descent under  $(p, \text{mean}) \rightarrow \infty$  then
12:     for  $i = 1$  to  $\ell$  do
13:        $\mathbf{W}_i^t \leftarrow \mathbf{W}_i^{t-1} - \eta^t \cdot \frac{1}{(d_{\text{in}}^i)^{1/p}} \cdot \text{rownorm}_p \left( \tilde{\mathbf{M}}_{\mathbf{W}_i^{t-1}}^t \right)$ 
14:        $\mathbf{b}_i^t \leftarrow \mathbf{b}_i^{t-1} - \eta^t \text{sign} \left( \tilde{\mathbf{M}}_{\mathbf{b}_i}^t \right)$ 
15:     end for
16:   end if
17: end for

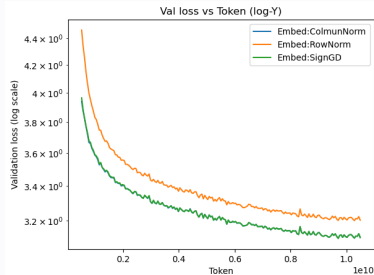
```

COMPUTATIONAL PROFILE Row and column normalization use closed form vector operations; unlike Muon, they do not require an SVD or Newton–Schulz iteration.

Source: MOGA paper, Algorithm 1, lines 11 to 17.

Adapting the geometry to Transformers

- **MLP and attention matrices:** MOGA row or column updates with the mean geometry scaling factor.
- **Attention logits:** row MOGA uses $QK^\top / d_v$, rather than $QK^\top / \sqrt{d_v}$, to keep logits width stable.
- **Bias and LayerNorm parameters:** sign gradient updates.
- **Embeddings:** SignGD or scaled column normalization $d_{\text{out}}^{1/q} \text{colnorm}_q$; row normalization is not used in the reported setup.
- **Unembedding:** tied to the token embedding matrix.

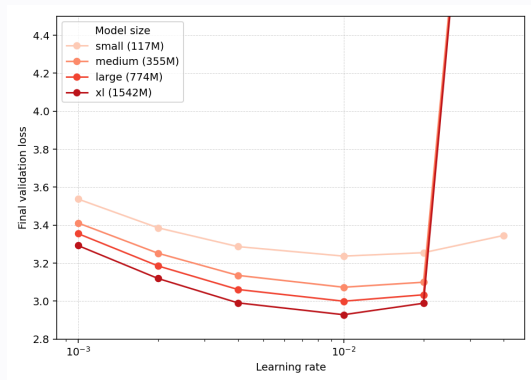


Embedding update comparison.

Source: MOGA paper, Transformer parameter block rules and embedding experiment.

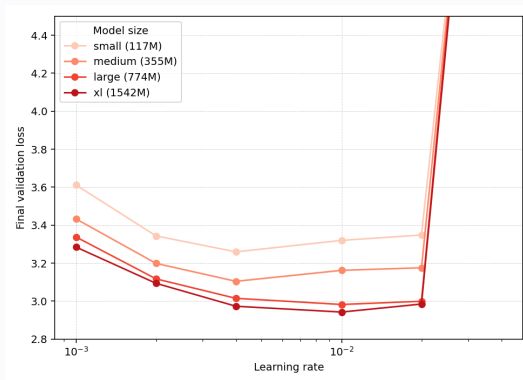
Learning rate transfer from GPT 2 Small to XL

WIDTHS AND DEPTHS CHANGE; ONE PEAK LEARNING RATE REMAINS COMPETITIVE



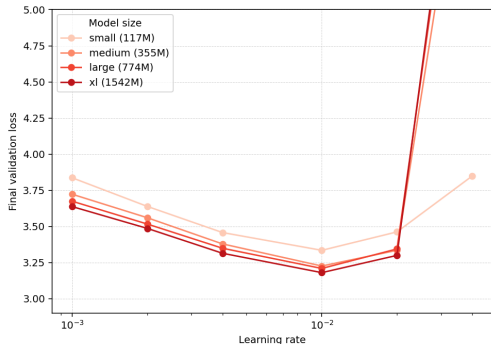
row MOGA, $p = 1.5$

Source: MOGA experiments, GPT 2 Small through GPT 2 XL.



row MOGA, $p = 2$

The same transfer persists for $p = 3$



Source: MOGA experiments, row MOGA learning rate transfer for $p = 3$.

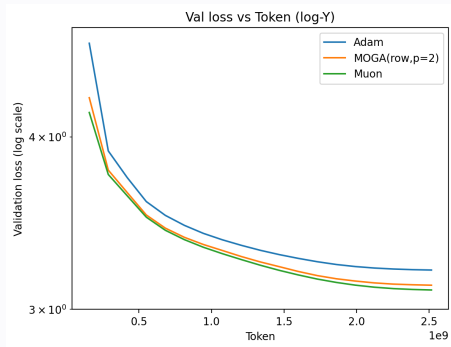
ROW MOGA WITH $p = 3$

The peak learning rate remains approximately width invariant from GPT 2 Small to XL.

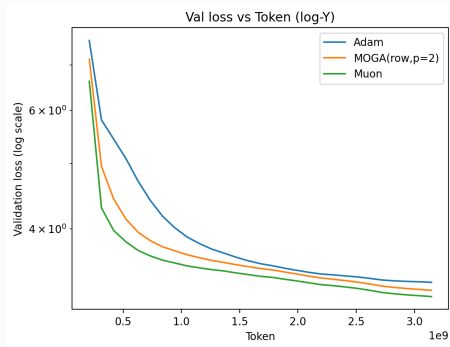
Empirical scope

The minimizing learning rate remains in the same neighborhood for this reported architecture family and sweep; this is not a universal guarantee.

Standard budget: MOGA is close to Muon



LLaMA 130M

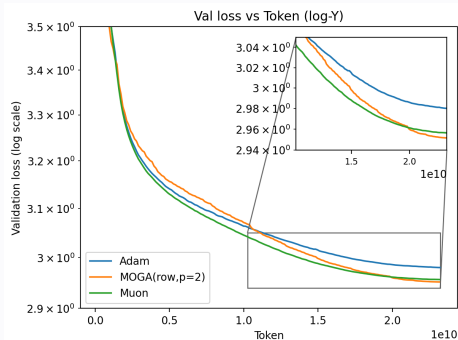


GPT 2 Small

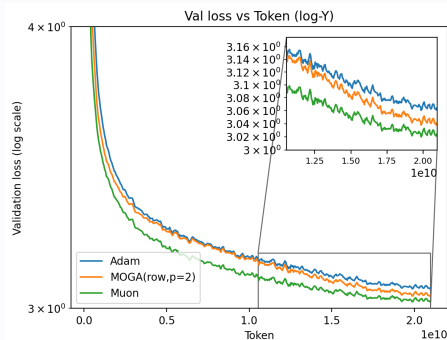
EMPIRICAL RESULT On these reported runs, row MOGA is close to Muon and improves on Adam in the shown validation loss curves.

Source: MOGA experiments on LLaMA 130M and GPT 2 Small.

Long budgets separate the late time trajectories



LLaMA 130M



GPT 2 Small

EMPIRICAL RESULT MOGA is competitive late in these fixed scale runs. This comparison is separate from the cross-width transfer experiment and the smoothness certificate.

Source: MOGA long budget pretraining experiments.

Inference time scaling

Write a law for what the model missed, then simulate that law.

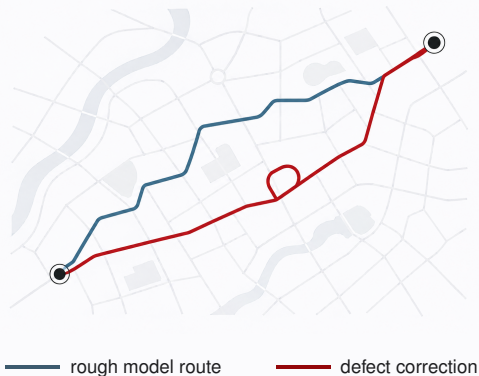
The equation checks the model's route

MY AI PHILOSOPHY

Check every answer against known structure.

Like double checking a map after initial navigation by a language model, the residual marks where the proposed route violates known local rules.

The model proposes a route; mathematics measures the defect.



Defect correction has exactly two steps

Write the Law of Defect

Derive the equation obeyed by $\check{u} := u - \hat{u}$.

1

known equation + autodiff

Simulate the Law of Defect

Spend new compute on stochastic paths for $\check{u}$.

2

linear: Feynman Kac + Monte Carlo
semilinear: multilevel Picard

$$u_{\text{corr}} = \hat{u} + \hat{\check{u}}_N$$

Write analytically. Spend inference compute stochastically.

No retraining: additional compute targets only what the model missed.

Source: SCaSML paper, structure preserving Law of Defect and stochastic correction.

Linear case: subtraction gives the defect law

EQUATION 1 GOVERNING EQUATION: THE BELLMAN EQUATION

$$\partial_t u + \mathcal{L}u = f, \quad u(T, \cdot) = g,$$

EQUATION 2 ANOTHER BELLMAN EQUATION THAT THE SURROGATE SATISFIED

$$\partial_t \hat{u} + \mathcal{L}\hat{u} = \hat{f}, \quad \hat{u}(T, \cdot) = \hat{g}.$$

EQUATION 1 MINUS EQUATION 2

$$\underbrace{(\partial_t + \mathcal{L})(u - \hat{u})}_{(\partial_t + \mathcal{L})\check{u}} = \underbrace{f - \hat{f}}_{\text{Bellman error } \delta_B}$$

Law of Defect: Value evaluation with Bellman Error

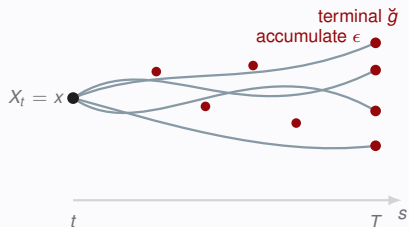
$$\check{u} := u - \hat{u}, \quad \check{g} := g - \hat{g}, \quad \delta_B := f - \hat{f} = -\epsilon,$$

$$(\partial_t + \mathcal{L})\check{u} = \delta_B, \quad \check{u}(T, \cdot) = \check{g}.$$

Step 2: Simulate the Law of Defect

SAMPLE THE GENERATOR

$$X_t = x, \quad dX_s = \mu(s, X_s) ds + \sigma(s, X_s) dW_s.$$



FEYNMAN KAC REPRESENTATION

$$\begin{aligned} \Delta &:= \check{g}(X_T) - \int_t^T \delta_B(s, X_s) ds \\ &= \check{g}(X_T) + \int_t^T \epsilon(s, X_s) ds, \\ \check{u}(t, x) &= \mathbb{E}_{t,x}[\Delta]. \end{aligned}$$

MONTE CARLO REALIZATION

$$\hat{u}_{\text{CORR}}^{(N)}(t, x) = \hat{u}(t, x) + \frac{1}{N} \sum_{n=1}^N \Delta^{(n)}.$$

Defect size controls the Monte Carlo variance

Monte Carlo Factor

$$\text{Var}(\hat{u}_{\text{corr}}^{(N)} | \hat{u}) = \frac{\text{Var}(\Delta | \hat{u})}{N}.$$

Defect Factor Condition on $\hat{u}$ and assume $\|\check{g}\|_{\infty}, \|\epsilon\|_{\infty} < \infty$.

$$|\Delta| \leq \|\check{g}\|_{\infty} + (T - t)\|\epsilon\|_{\infty},$$

The errors multiply

$$\|\check{g}\|_{\infty} \leq C_T e(\hat{u}), \|\epsilon\|_{\infty} \leq C_{\epsilon} e(\hat{u}) \implies \boxed{\text{RMSE}_{\hat{u}} \leq (C_T + (T - t)C_{\epsilon}) \frac{e(\hat{u})}{\sqrt{N}}}.$$

Use surrogate model as Control Variate!

Better surrogate model $\rightarrow$ Better Variance

Nonlinear case: subtraction gives the defect law

Semi-Linear Equation

$$\mathcal{N}[v] := \partial_t v + \mathcal{L}v + F(v, \sigma^\top \nabla v).$$

EQUATION 1

$$\mathcal{N}[u] = f, \quad u(T) = g.$$

EQUATION 1 MINUS EQUATION 2

$$\begin{aligned} \partial_t \check{u} + \mathcal{L}\check{u} + F(\hat{u} + \check{u}, \hat{z} + \check{z}) - F(\hat{u}, \hat{z}) &= \underbrace{f - \hat{f}}_{\text{Bellman error } \delta_B} = -\epsilon, \\ \check{u}(T, \cdot) &= g - \hat{g}. \end{aligned}$$

EQUATION 2

$$\mathcal{N}[\hat{u}] = \hat{f}, \quad \hat{u}(T) = \hat{g}.$$

$$\text{Law of Defect: } \partial_t \check{u} + \mathcal{L}\check{u} + \check{F}(\check{u}, \check{z}) = 0.$$

Key Observation: Closeness

$$\check{F}(v, z) := F(\hat{u} + v, \hat{z} + z) - F(\hat{u}, \hat{z}) + \epsilon$$

$$\check{u} := u - \hat{u}, \hat{z} := \sigma^\top \nabla \hat{u}, \check{z} := \sigma^\top \nabla \check{u}.$$

The correction error is a product, not a sum

SCaSML

Solving Semi-linear **Law of Defect** ($\partial_t \check{u} + \mathcal{L}\check{u} + \check{F}(\check{u}, \check{z}) = 0$) using Multi-level Picard Iteration!

Assume the surrogate controls both the residual and defect magnitude:

$$\sup_{t,x} |\epsilon(t, x)| \leq C_{F,1} e(\hat{u}), \quad \sup_t \|\check{u}(t, \cdot)\|_{W^{1,\infty}} \leq C_{F,2} e(\hat{u}).$$

Let $\check{\mathbf{u}} = (\check{u}, \sigma^\top \nabla \check{u})$ and let $\check{\mathbf{U}}_{N,M}$ be the full history multilevel Picard estimator.

SCaSML product error bound

$$\sup_{t,x} \max_j \left\| (\check{\mathbf{U}}_{N,M} - \check{\mathbf{u}})_j \right\|_{L^2} \leq E(M, N) C_F e(\hat{u}).$$

$E(M, N)$ is the error of MLP Algorithm.

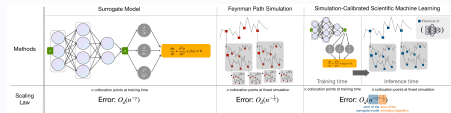
Training and inference exponents add

Under the same heat equation assumptions and the paper's matched training and inference allocation, if the pretrained surrogate obeys

$$e(\hat{u}) = O(m^{-\gamma}),$$

and inference receives a comparable budget of m additional Monte Carlo samples, the correction theorem yields

$$\text{Err}_{\text{SCaSML}} = O\left(m^{-\gamma-1/2+o(1)}\right).$$



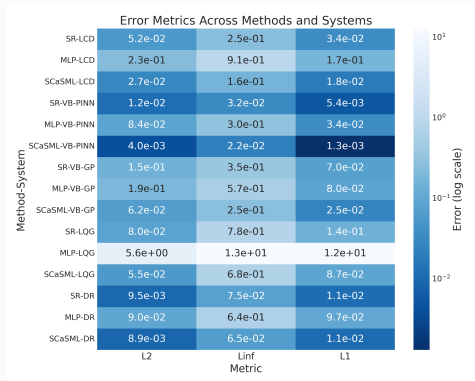
Interpretation

Monte Carlo does not replace the pretrained rate. It multiplies the surrogate error by an additional sampling factor, so the exponents add.

$$\text{pretraining rate} + \text{inference rate} = \gamma + \frac{1}{2}.$$

Source: SCaSML scaling corollary; statement uses m training points and m additional inference samples.

One defect corrector works across PDE families



REPRESENTATIVE SETTINGS

LCD $d = 10$

Burgers PINN $d = 20$

Burgers GP $d = 20$

LQG $d = 100$

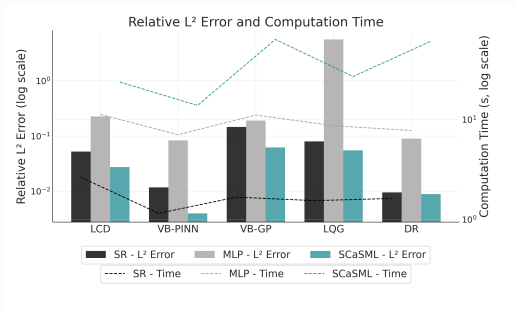
Diffusion reaction $d = 100$

Across three error norms

SCaSML is lower in 14 of 15 displayed comparisons and tied at the plotted precision in the last one.

Source: Fan et al., arXiv:2504.16172v3, Sections 5.1 to 5.4, Table 1 and Figure 4c.

SCaSML spends inference compute to buy accuracy



The correction costs more inference time, but lowers relative L^2 error on every representative benchmark.

Source: Fan et al., arXiv:2504.16172v3, Section 5, Table 1 and Figure 4c.

Guidance turns “at work” into a sampling direction

Suppose a diffusion model already knows how José looks.

BASE DISTRIBUTION $X_t \sim p_t$, public images of José



REWARD

Is José at work?

The hard reward verifies the final image.

extra drift follows predicted reward

$$V(t)^2 \nabla G_t(X_t)$$



GUIDED SAMPLE synthetic target



synthetic illustration • José at work

The base model supplies identity; the reward specifies the task; guidance changes the proposal.

Guidance changes the proposal, not the target

Base dynamics and reward tilted marginal:

$$dX_t = v(X_t, t)dt + V(t)dW_t, \quad q_t(x) \propto p_t(x)e^{r(x,t)}.$$

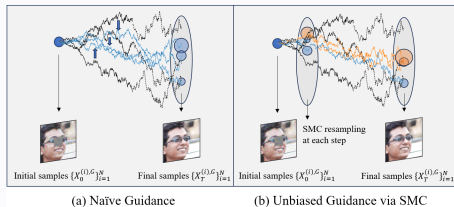
The ideal Doob value is

$$h(x, t) = \mathbb{E}_{\mathbb{P}}[e^{r(X_T)} \mid X_t = x],$$

while a practical guidance potential G produces

$$dX_t^G = (v(X_t^G, t) + V(t)^2 \nabla G(X_t^G, t))dt + V(t)dW_t.$$

$$G \neq \log h \implies \mathbb{P}^G \neq \mathbb{Q} \text{ in general.}$$



Exact guidance requires the intractable Doob value.

Girsanov gives the exact path correction

The change from guided to base dynamics is

$$\frac{d\mathbb{P}}{d\mathbb{P}^G}(X_{0:t}^G) = \exp\left(-\int_0^t V \nabla G^\top dW - \frac{1}{2} \int_0^t V^2 \|\nabla G\|_2^2 ds\right).$$

The reward tilted path law satisfies

$$\frac{d\mathbb{Q}}{d\mathbb{P}}(X_{0:t}) \propto e^{r(X_t, t) - r(X_0, 0)}.$$

Therefore

$$\frac{d\mathbb{Q}}{d\mathbb{P}^G}(X_{0:t}^G) \propto \exp\left[r(X_t^G, t) - r(X_0^G, 0) - \int_0^t V \nabla G^\top dW - \frac{1}{2} \int_0^t V^2 \|\nabla G\|_2^2 ds\right].$$

Key advantage

The correction weight needs no reward gradient, reward Laplacian, or score.

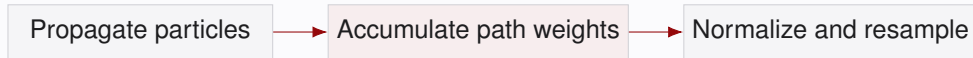
URGE corrects diffusion paths sequentially

For particle i and step k ,

$$X_{k+1}^{i,G} = X_k^{i,G} + (v_k^i + V_k^2 \nabla G_k^i) \Delta t_k + V_k \sqrt{\Delta t_k} \xi_k^i, \quad \xi_k^i \sim \mathcal{N}(0, I).$$

The incremental log weight is

$$\log \beta_k^i = r(X_{k+1}^{i,G}, t_{k+1}) - r(X_k^{i,G}, t_k) - V_k (\nabla G_k^i)^\top \sqrt{\Delta t_k} \xi_k^i - \frac{1}{2} V_k^2 \|\nabla G_k^i\|_2^2 \Delta t_k.$$



Source: URGE paper, Algorithm 1 and incremental Girsanov reward weight.

URGE is exact in an iterated limit

Iterated consistency

Under the paper's regularity assumptions and correct initial law,

$$\lim_{N \rightarrow \infty} \lim_{\Delta t \rightarrow 0} \mathbb{E}_{\mathbb{P}}[M_T^{N, \Delta t}(\varphi)] = \frac{\mathbb{E}_{\mathbb{P}}[e^{r(X_T) - r(X_0, 0)} \varphi(X_T, T)]}{\mathbb{E}_{\mathbb{P}}[e^{r(X_T) - r(X_0, 0)}]}.$$

Scope of the theorem

Asymptotic consistency only: the result gives no finite N total variation or sample complexity bound.

Source: URGE paper, continuous approximation freeness theorem.

Path and particle corrections agree infinitesimally

$$\lambda(x, t) = \lim_{s \uparrow t} \frac{\mathbb{E}_{\mathbb{P}^G}[w_{s,t}^{\text{URGE}} \mid X_t^G = x] - 1}{t - s}.$$

If the guided marginal density is sufficiently smooth, then

$$\lambda(x, t) = w_{\text{particle}}(x, t).$$

What agrees

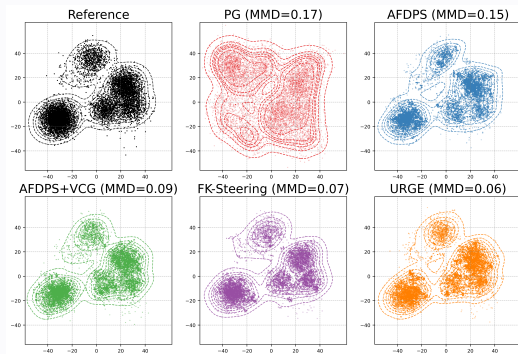
After marginalization, the infinitesimal path weight growth equals the particle space potential.

What does not

This is not equality of the finite grid, finite particle algorithms.

Source: URGE paper, reweighted generator lemma and path to particle equivalence theorem.

Pathwise correction recovers the target modes

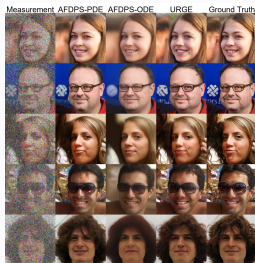


On the reported 30 dimensional Gaussian mixture, URGE gives the closest sample cloud to the analytical target, with **MMD 0.06**.

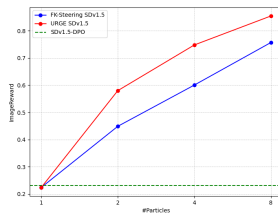
Source: Wang et al., arXiv:2605.17850v1, Section 4.1, Table 2, Appendix D and Figure 4.

From inverse problems to neural rewards

FFHQ 256 SUPER RESOLUTION



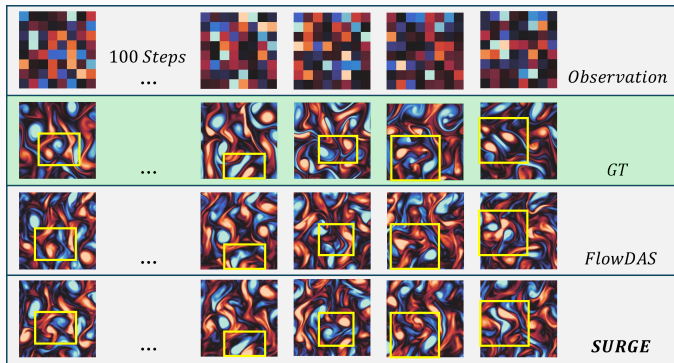
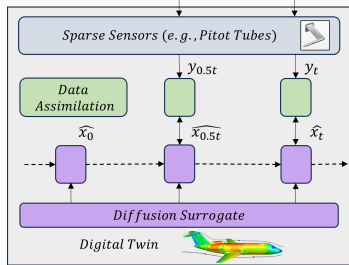
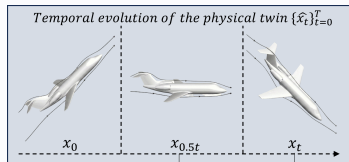
SD V1.5 PARTICLE SCALING



URGE matches derivative based restoration; ImageReward exceeds FK Steering for $N = 2, 4, 8$.

Source: Wang et al., arXiv:2605.17850v1, Sections 4.2 and 4.3, Table 8, Figures 7 and 11.

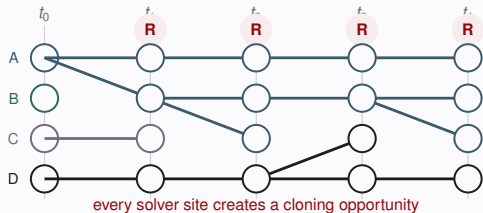
Data Assimilation



Every microstep can become a statistical generation

FULL MICRO RESAMPLING

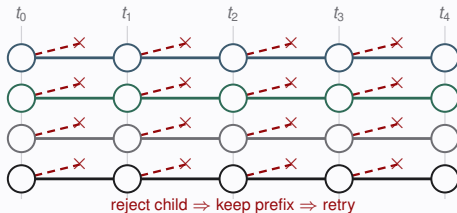
one parent pool



$$A_k^i \sim \text{Cat}(W_{k-1}^{1:N}), \quad X_k^i \sim \mathcal{M}_{k,n}(X_{k-1}^{A_k^i}, \cdot)$$

EXACT CHILD REJECTION

independent proposal streams



$$Y \sim \mathcal{M}_{k,n}(x, \cdot), \quad U \sim \text{Unif}(0, 1)$$

$$\text{accept if } U \leq \frac{G_{k,n}(x, Y)}{\exp\{m_{k,n}(x) + a_{k,n}\}}$$

Separate finite N result for discretized path targets. Pure child rejection is shown; parent normalizers require exact outer correction.

Resampling moves mass **across lineages**. Rejection spends serial trials **within a lineage**.

Source: Supplied diffusion imaging paper, full micro resampling and sequential child rejection.

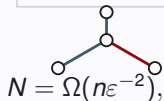
Shared ancestry creates the variance gap

$$\text{Var}\left(\frac{1}{N} \sum_{i=1}^N f(X_n^i)\right) = \underbrace{\frac{1}{N} \text{Var}(f(X_n^1))}_{\text{one path}} + \underbrace{\frac{N-1}{N} \text{Cov}(f(X_n^1), f(X_n^2))}_{\text{genealogy}}$$

FULL MICRO RESAMPLING · CAN BE THIS BAD

Neutral: $G_{k,n} \equiv 1$, $\text{ESS} = N$, $g = \sin(\cdot)/2$, $\gamma = \mathcal{N}(0, 1)$, $n, N \geq 2$.

$$\mathbb{E}[(\eta_n^N(g) - \gamma(g))^2] \geq \frac{1}{32e^2} \min\left\{\frac{n}{N}, 1\right\}$$



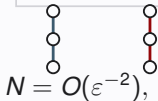
shared ancestor
creates covariance

$$\text{work} = \Omega(n^2 \varepsilon^{-2})$$

EXACT REJECTION · UNIFORM UPPER BOUND

Independent accepted paths; $\text{osc}(f) \leq 1$.

$$\text{MSE} \leq \frac{1}{4N}, \quad \text{Cov}(f(X_n^1), f(X_n^2)) = 0$$



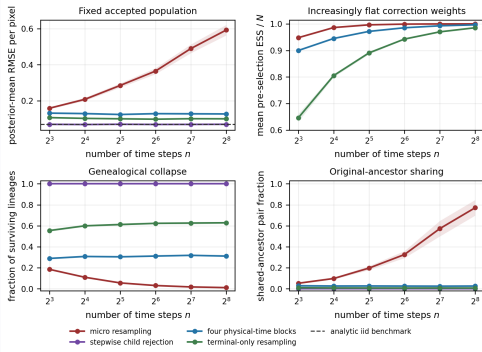
no shared parent pool
so covariance is zero

$$\mathbb{E}[\text{work}] = O(n \varepsilon^{-2})$$

Rejection removes the avoidable factor n from **statistical width**, not from path depth.

Source: Supplied diffusion imaging paper, variance anatomy, neutral population barrier, and population separation.

A finer mesh can improve ESS while destroying genealogy



Local ESS improves while genealogy collapses.

Source: Supplied diffusion imaging manuscript, controlled linear imaging experiment and scaling figure.

CONTROLLED LINEAR IMAGING

$N = 160$ accepted paths $\cdot$ 36 runs
 $n = 8, 16, 32, 64, 128, 256$

Stepwise rejection

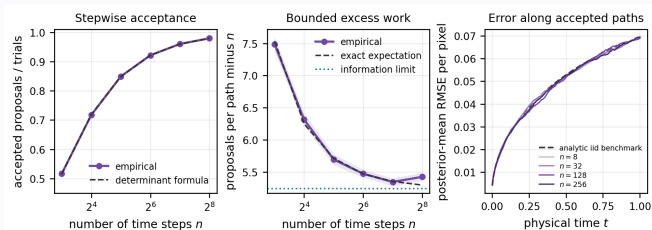
RMSE : 0.0683 to 0.0696

Micro resampling

RMSE : 0.159 to 0.593

At the finest mesh, ESS is $0.99994N$, yet only 1.02% of lineages survive and 77.2% of terminal pairs share an ancestor.

Rejection moves mesh cost into serial proposals



$$W_n = \sum_{k=1}^n \alpha_{k,n}^{-1} = n + I(X_1; Y) + O(n^{-1})$$

Information controls retry overhead; population controls independent sampling variance.

Acceptance: 0.517 $\rightarrow$ 0.979

Excess proposals: 7.54 $\rightarrow$ 5.30

Information limit: $I(X_1; Y) = 5.239$ nats

Source: Supplied diffusion imaging manuscript, rejection diagnostics and exact Gaussian work formula.

LLM reasoning is reward tilted path sampling

The reference model defines a trajectory law

$$\pi_{\text{ref}}(\mathbf{s}_{1:T} \mid \mathbf{s}_0) = \prod_{t=1}^T \pi_{\text{ref}}(s_t \mid \mathbf{s}_0, \mathbf{s}_{1:t-1}).$$

A terminal verifier or reward ϕ defines the target

$$\tilde{\pi}_T(\mathbf{s}_{1:T} \mid \mathbf{s}_0) \propto \pi_{\text{ref}}(\mathbf{s}_{1:T} \mid \mathbf{s}_0) \phi(\mathbf{s}_{1:T}).$$

Intermediate targets use a positive value approximation:

$$\tilde{\pi}_t(\mathbf{s}_{1:t} \mid \mathbf{s}_0) \propto \pi_{\text{ref}}(\mathbf{s}_{1:t} \mid \mathbf{s}_0) \hat{V}(\mathbf{s}_0, \mathbf{s}_{1:t}), \quad \hat{V}(\mathbf{s}_0, \mathbf{s}_{1:T}) = \phi(\mathbf{s}_{1:T}).$$

Inference compute allocation

The value model is not just a final ranker. It allocates particles among partial reasoning trajectories.

Source: Reward guided SMC for LLMs paper, target and intermediate Feynman Kac path measures.

Bellman consistency measures value quality

EXACT VALUE AND BELLMAN RECURSION

$$V_t^*(s_{1:t}) := \mathbb{E}_{\pi_{\text{ref}}}[\phi(S_{1:T}) \mid s_{1:t}] = \mathbb{E}_{\pi_{\text{ref}}}[V_{t+1}^*(S_{1:t+1}) \mid s_{1:t}].$$

LOCAL MULTIPLICATIVE DEFECT FOR $\hat{V} > 0$

$$\frac{1}{1 + \varepsilon} \leq \frac{\hat{V}(s_{1:t})}{\mathbb{E}_{\pi_{\text{ref}}}[\hat{V}(S_{1:t+1}) \mid s_{1:t}]} \leq 1 + \varepsilon,$$

Also assume the ratio bound $L^{-1} \leq \hat{V}(s_{1:t})/\hat{V}(s_{1:t-1}) \leq L$.

Source: Reward guided SMC for LLMs paper, optimal twist and local Bellman error assumption.

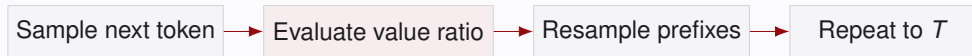
SMC turns a value approximation into local weights

For proposal π_p , the Feynman Kac potential is

$$G_t(s_{1:t}) = \frac{\pi_{\text{ref}}(s_t \mid s_{1:t-1})}{\pi_p(s_t \mid s_{1:t-1})} \frac{\hat{V}(s_{1:t})}{\hat{V}(s_{1:t-1})}.$$

With the naive proposal $\pi_p = \pi_{\text{ref}}$,

$$G_t(s_{1:t}) = \frac{\hat{V}(s_{1:t})}{\hat{V}(s_{1:t-1})}.$$



Output law

The theorem controls the averaged law of one uniformly selected final particle, not concentration of a single realized empirical measure.

Small Bellman error makes particle correction polynomial

ONE PARTICLE

$$\|\tilde{\pi}_t - \hat{\pi}_t\|_{\text{TV}} \leq 2t\varepsilon.$$

A sufficient certificate, not a converse.

Naive particle SMC

$$N \geq \frac{L^6 T (1 + \varepsilon)^{6(T-1)}}{2\delta_{\text{TV}}} \implies \|\bar{\eta}_{T+1}^N - \tilde{\pi}_T\|_{\text{TV}} \leq \delta_{\text{TV}}.$$

In the clean regime $\varepsilon = O(1/T)$,

$$\text{TC}_{\text{naive SMC}} = O(L^6 T^2 / \delta_{\text{TV}}).$$

Source: Reward guided SMC for LLMs paper, single particle TV theorem and particle bias theorem.

Fixed Bellman error can make search exponential

NO GUESS ORACLE MODEL Every output prefix must have been visited through oracle queries.

Worst case lower bounds in the no guess oracle model

To achieve total variation error at most $1/3$ on every admissible instance,

$$TC = \Omega(L^{2T/3})$$

under the ratio bound alone. With $L > 1$ and $\varepsilon \in (0, L - 1]$ under the local Bellman bound,

$$TC = \Omega\left((1 + \varepsilon)^{2T/3}\right).$$

Interpretation. Fixed ε permits exponential worst case cost; $\varepsilon = O(1/T)$ gives a clean polynomial upper bound.

Source: Reward guided SMC for LLMs paper, lower bound theorem and Bellman error corollary.

MH correction yields geometric mixing

DISTORTION AND CONTRACTION

$$b = \left((1 + \varepsilon) \frac{1 + \xi}{1 - \xi} \right)^{T-1},$$

$$\|\mu K_{\mathcal{G}}^H - \nu K_{\mathcal{G}}^H\|_{\text{TV}} \leq (1 - b^{-2})^H \|\mu - \nu\|_{\text{TV}}.$$

Exactness. The augmented target is invariant; the token path marginal is within δ_{TV} of $\tilde{\pi}_T$.

HIGH PROBABILITY COMPLEXITY If $\varepsilon, \xi = O(1/T)$ and $0 < \delta \lesssim \delta_{\text{TV}}$, then with probability at least $1 - \delta$,

$$\text{TC}_{\text{pool MH}} = \tilde{O}\left(LT^3 \log \frac{1}{\delta} \log \frac{1}{\delta_{\text{TV}}}\right).$$

Source: Reward guided SMC for LLMs paper, candidate pool MH algorithm and mixing theorem.

One correction principle, three languages

PDE

 $\hat{u}(t, x)$

PDE residual

Feynman Kac or multilevel Picard

DIFFUSION

 $\mathbb{P}^G$

Doob or Bellman residual

Girsanov, SMC, or exact rejection

LLM

 $\hat{V}(s_{1:t})$

local Bellman defect

SMC averaging or MH mixing

Inference time scaling law

General URGE gives asymptotic consistency. Under exact rejection assumptions, $N = O(\varepsilon^{-2})$ accepted paths give range one RMSE ε , uniformly in the mesh.

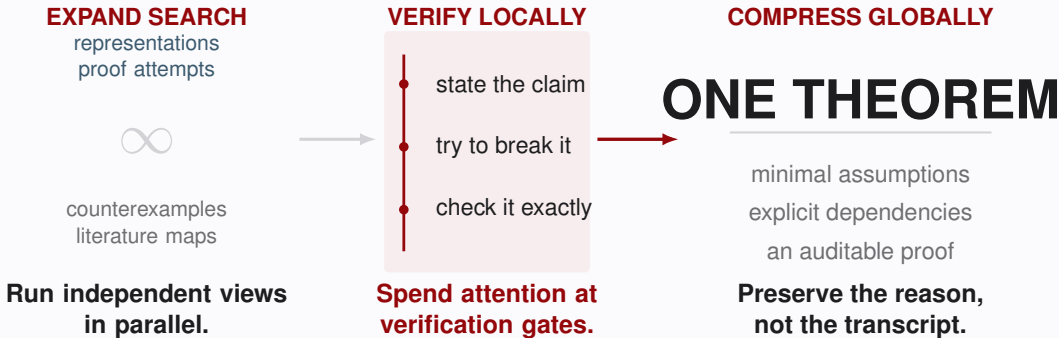
Source: SCaSML, URGE, diffusion rejection, and reward guided SMC papers supplied for this talk.

AI for mathematics

Inference time search is useful only when representation and verification scale with it.

Doing mathematics with infinite tokens

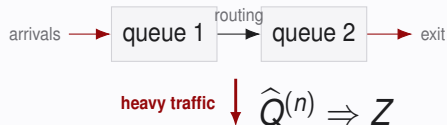
Generation becomes abundant. Trustworthy compression becomes the bottleneck.



AI should make the search tree **wide**.

Mathematics should make the final argument **small**.

Heavy traffic yields a reflected diffusion



For many queueing networks, a centered and diffusion-scaled queue-length or workload process converges to

$$Z_t = Z_0 + \mu t + \Sigma^{1/2} W_t + RY_t, \quad Z_t \in E := \mathbb{R}_+^d.$$

On $F_i = \{x \in E : x_i = 0\}$, the column R_i is the reflection direction and its regulator satisfies

$$Y_i(0) = 0, \quad Y_i \text{ continuous and nondecreasing,}$$

$$\int_0^t \mathbf{1}_{\{Z_i(s) > 0\}} dY_i(s) = 0.$$

Why reflection matters. The term RY_t keeps Z_t in the orthant: queues and workloads cannot become negative.

Source: Harrison and Reiman, 1981; Harrison and Williams, 1987;

The BAR balances interior and boundary motion

Let $E = \mathbb{R}_+^d$ and $F_i = \{x \in E : x_i = 0\}$. For $f \in C_b^2(E)$, define

$$Lf = \mu \cdot \nabla f + \frac{1}{2} \Sigma : D^2 f, \quad D_i f = R_i \cdot \nabla f.$$

The finite boundary occupation measure is

$$\nu_i^0(B) = \mathbb{E}_{\pi_0} \left[\int_0^1 \mathbf{1}_{\{Z_s \in B\}} dY_i(s) \right], \quad B \in \mathcal{B}(F_i).$$

Itô's formula and stationarity give the Basic Adjoint Relationship:

$$0 = \mathbb{E}_{\pi_0} [f(Z_1) - f(Z_0)] = \int_E Lf d\pi_0 + \sum_{i=1}^d \int_{F_i} D_i f d\nu_i^0.$$

Question. Can the BAR admit hidden signed solutions beyond the stationary tuple?

The signed problem asks for the entire kernel

A finite signed BAR tuple

$$(\bar{\pi}, \bar{\nu}_1, \dots, \bar{\nu}_d) \in \mathcal{M}(E) \times \prod_{i=1}^d \mathcal{M}(F_i)$$

satisfies

$$\mathcal{A}_{\text{BAR}}(\bar{\pi}, \bar{\nu})[f] := \int_E Lf \, d\bar{\pi} + \sum_{i=1}^d \int_{F_i} D_i f \, d\bar{\nu}_i = 0$$

for every $f \in C_b^2(E)$. The stationary tuple $\mathbf{s}_0 = (\pi_0, \nu_1^0, \dots, \nu_d^0)$ is one element of the kernel. The conjecture asks whether

$$\ker \mathcal{A}_{\text{BAR}} = \text{span}\{\mathbf{s}_0\}.$$

Why signed measures are difficult

Positivity rules out cancellation by sign within a measure; signed interior and boundary measures can cancel across strata.

Uniqueness holds in the stable Harrison Reiman class

The positive theorem and the obstruction

Signed uniqueness holds for stable Harrison Reiman data. In the broader stable completely $\mathcal{S}$ class, a singular proper principal block creates counterexamples.

Assume $\Sigma \succ 0$ and R is a nonsingular M matrix: $R_{ii} > 0, R_{ij} \leq 0$ ($i \neq j$), together with the stability condition $R^{-1}\mu < 0$ componentwise. Equivalently, after column normalization,

$$\hat{R} = I - P^\top, \quad P \geq 0, \quad \rho(P) < 1.$$

Signed BAR uniqueness

Every finite signed BAR tuple satisfies $\bar{\pi} = c\pi_0, \bar{\nu}_i = c\nu_i^0, i = 1, \dots, d$, for some $c \in \mathbb{R}$. Hence the signed BAR space is one dimensional.

Source: Signed BAR manuscript, positive uniqueness theorem and Harrison Reiman assumptions.

PROPOSITION 2.3 · RESOLVENT IDENTITY CRITERION

Resolvent identity forces signed BAR uniqueness

ASSUME · THE RESOLVENT IDENTITY

$$\int_E (\lambda R_\lambda h - h) d\bar{\pi} = 0$$

every finite signed BAR tuple · $h \in C_0(E)$ · $\lambda > 0$

FORCES



CONCLUSION · SIGNED BAR UNIQUENESS

$$\bar{\pi} = c\pi_0, \quad \bar{\nu}_i = c\nu_i^0 \quad (i = 1, \dots, d)$$

The signed BAR space is one dimensional.

01 · RI $\implies$ INVARIANCE

$$F_h(t) := \bar{\pi}(P_t h)$$

$$0 = \lambda \int_0^\infty e^{-\lambda t} [F_h(t) - F_h(0)] dt$$

Laplace uniqueness + C_0 Feller continuity

$$\bar{\pi} P_t = \bar{\pi} \quad (t \geq 0)$$

02 · JORDAN DECOMPOSITION

$$\bar{\pi} P_t = \bar{\pi}, \quad |\bar{\pi} P_t| \leq |\bar{\pi}| P_t$$

equal total mass $\implies |\bar{\pi}| P_t = |\bar{\pi}|$

$$\bar{\pi}^\pm P_t = \bar{\pi}^\pm$$

unique invariant probability

$$\bar{\pi} = c\pi_0, \quad c = \bar{\pi}(E)$$

03 · BOUNDARY INJECTIVITY

$$\eta_i := \bar{\nu}_i - c\nu_i^0$$

$$\sum_{i=1}^d \int_{F_i} D_i f d\eta_i = 0 \quad \forall f \in C_b^2(E)$$

pure boundary injectivity
invertible principal blocks

$$\eta_i = 0 \implies \bar{\nu}_i = c\nu_i^0$$

The hard analysis proves RI; Proposition 2.3 converts RI into full signed uniqueness.

$\Sigma \succ 0$ · R nonsingular M matrix · $R^{-1}\mu < 0$

Why the C^2 resolvent idea is right

$$\mathcal{A}_{\text{BAR}}(\bar{\pi}, \bar{\nu})[f] = 0, \quad f \in C_b^2(E).$$

For any probe $h \in C_0(E)$ and every $\lambda > 0$, the resolvent is

$$g = R_\lambda h := \int_0^\infty e^{-\lambda t} P_t h \, dt.$$

If g were an admissible C^2 test, its two defining identities would give

$$\left. \begin{aligned} (\lambda - L)g &= h, \\ D_i g &= 0 \quad \text{on } F_i \end{aligned} \right\} \implies \boxed{\int_E (\lambda R_\lambda h - h) \, d\bar{\pi} = 0.} \quad (\text{RI})$$

The resolvent turns an arbitrary observable into the exact Poisson test for stationarity.

Source: Lu and Zhu, signed BAR manuscript, resolvent insertion strategy.

Resolvent insertion without corner regularity

Smooth from the interior, insert only legal tests, then pass to a measure level limit.

FORMAL MOVE

$$g = R_\lambda h, \quad (\lambda - \mathcal{L})g = h$$

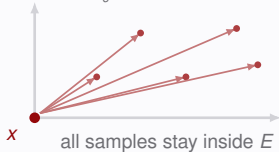
CORNER OBSTRUCTION

$R_\lambda h \notin C_b^2(E)$ may occur
The resolvent is not a legal BAR test at the corner.

01 · SMOOTH INWARD

$$g_\varepsilon(x) = \int \rho(w) g(x + \varepsilon w) dw$$

$\rho \in C_c^\infty((1, 2)^d)$
first $h \in C_c^\infty$; then extend by density



02 · INSERT A LEGAL TEST

$$g_\varepsilon \in C_b^2(E), \quad (\lambda - \mathcal{L})g_\varepsilon = h_\varepsilon$$
$$0 = \int_E \mathcal{L}g_\varepsilon d\bar{\pi} + \sum_{i=1}^d \int_{F_i} D_i g_\varepsilon d\bar{\nu}_i$$

Every $\varepsilon > 0$ produces an admissible test in the signed BAR.

03 · PASS TO THE BAR LIMIT

INTERIOR · UNIFORM

$$g_\varepsilon \rightarrow R_\lambda h, \quad h_\varepsilon \rightarrow h$$

BOUNDARY · POINTWISE + DOMINATION

$$D_i g_\varepsilon(x) \rightarrow \ell_x(R_i) = 0 \quad (x \in F_i)$$

$$|D_i g_\varepsilon| \leq C_{\rho, R} \text{Lip}(g)$$
$$\int_{F_i} D_i g_\varepsilon d\bar{\nu}_i \rightarrow 0$$

INSERT g_ε , THEN LET $\varepsilon \downarrow 0$

$$\bar{\pi}(\lambda R_\lambda h - h) = 0$$

stable Harrison–Reiman data · nonsingular M matrix

every finite signed BAR tuple · $h \in C_0(E)$ · $\lambda > 0$

Corners break regularity, not the idea

FORMAL GAP · $\mathcal{G}$ IS THE C_0 FELLER GENERATOR

$$R_\lambda h \in \mathcal{D}(\mathcal{G}) \quad \not\Rightarrow \quad R_\lambda h \in C_b^2(E).$$

WHAT SURVIVES AT A CORNER · $A = I(x) := \{i : x_i = 0\}$

$$\begin{aligned} L_A v &:= v - R_A R_{AA}^{-1} v_A, & L_A R_i &= 0 \quad (i \in A), \\ \partial_w^+ R_\lambda h(x) &= \Lambda_x(L_A w), & w &\text{feasible.} \end{aligned}$$

The projection L_A kills every active reflection direction, even when the classical corner gradient does not exist.

The proof idea is right; the test function is not yet legal.

Inward smoothing makes the resolvent a legal test

$$g_\varepsilon(x) = \int \rho(w) g(x + \varepsilon w) dw, \quad \rho \in C_c^\infty((1, 2)^d), \quad \int \rho = 1.$$

$$\begin{aligned} g_\varepsilon &\in C_b^2(E), \\ (\lambda - L)g_\varepsilon &= h_\varepsilon \longrightarrow h, \\ D_i g_\varepsilon(x) &\longrightarrow 0 \quad (x \in F_i), \quad \sup_\varepsilon \|D_i g_\varepsilon\|_\infty < \infty. \end{aligned}$$

Pointwise convergence plus the uniform bound gives, for every finite signed boundary measure,

$$\int_{F_i} D_i g_\varepsilon d\bar{\nu}_i \longrightarrow 0.$$

Shift inward, smooth, insert into the BAR, and let $\varepsilon \downarrow 0$.

Source: Lu and Zhu, one sided smoothing and measure Neumann approximation.

Resolvent invariance fixes the interior measure

FROM LAPLACE SPACE TO TIME

$$\lambda \bar{\pi} R_\lambda = \bar{\pi} \quad (\forall \lambda > 0) \xrightarrow[\text{C}_0 \text{ Feller continuity}]{\text{Laplace uniqueness}} \boxed{\bar{\pi} P_t = \bar{\pi} \quad \forall t \geq 0.}$$

THE SIGNED STEP

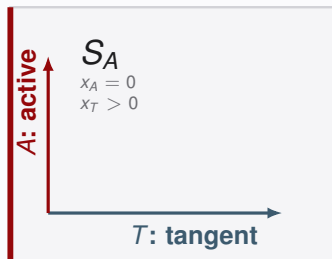
$$\left. \begin{array}{l} \bar{\pi} P_t = \bar{\pi}, \\ (|\bar{\pi} P_t| \leq |\bar{\pi}| P_t, \\ (|\bar{\pi}| P_t)(E) = |\bar{\pi}|(E) \end{array} \right\} \implies |\bar{\pi}| P_t = |\bar{\pi}| \implies \bar{\pi}^\pm P_t = \bar{\pi}^\pm \implies \boxed{\bar{\pi} = c \pi_0.}$$

WHAT REMAINS AFTER SUBTRACTING THE STATIONARY TUPLE

$$\boxed{\sum_{i=1}^d \int_{F_i} D_i f \, d\eta_i = 0, \quad \eta_i := \bar{\nu}_i - c \nu_i^0.}$$

Source: Lu and Zhu, resolvent invariance and signed invariant measure argument.

One stratum fixes the notation



LOCAL BOUNDARY SYMBOL

COORDINATE SPLIT

$$J = A \sqcup T, \quad T = J \setminus A, \quad S_A = \{x_A = 0, x_T > 0\}.$$

REFLECTION BLOCKS

$$R = \left[\begin{array}{c|c} R_{AA} & R_{AT} \\ \hline R_{TA} & R_{TT} \end{array} \right], \quad \nabla f = \begin{bmatrix} \nabla_A f \\ \nabla_T f \end{bmatrix}.$$

R_{BC} means rows in B and columns in C .

$$D_A f := (D_i f)_{i \in A} = R_A^\top \nabla f = R_{AA}^\top \nabla_A f + R_{TA}^\top \nabla_T f.$$

Source: Lu and Zhu, Section 7, local boundary algebra on the stratum S_A .

Uniqueness splits into two obstructions

INTERIOR COORDINATES OF SIGNED BAR SOLUTIONS

$\mathcal{I}_{\text{BAR}} := \{\pi \in \mathcal{M}(E) : \exists \nu \in \mathcal{M}_{\partial}, (\pi, \nu) \in \ker \mathcal{A}_{\text{BAR}}\}.$
RESOLVENT INSERTION DEFECT

STATIONARY DIRECTION

$s_0 := (\pi_0, \nu^0) \in \ker \mathcal{A}_{\text{BAR}}.$

$$\tau(m)(\lambda, h) := \int_E (\lambda R_{\lambda} h - h) dm, \quad h \in C_0(E), \quad \lambda > 0.$$

$\mathcal{I}_{\text{RI}} := \mathcal{I}_{\text{BAR}} \cap \ker \tau$, $\mathcal{Q}_{\text{RI}} := \mathcal{I}_{\text{BAR}} / \mathcal{I}_{\text{RI}}$, and $\partial_R \nu[f] := \sum_i \int_{F_i} D_i f d\nu_i$.

The algebraic split

$$0 \longrightarrow \underbrace{\ker \partial_R}_{\text{pure boundary}} \longrightarrow \underbrace{\ker \mathcal{A}_{\text{BAR}} / \mathbb{R} s_0}_{\text{signed BAR modulo stationarity}} \longrightarrow \underbrace{\mathcal{Q}_{\text{RI}}}_{\text{interior RI defect}} \longrightarrow 0.$$

Uniqueness $\iff \ker \partial_R = \{0\}$ and $\mathcal{Q}_{\text{RI}} = \{0\}.$

Source: Lu and Zhu, arXiv:2607.03639, Section 7, algebraic decomposition of the signed BAR kernel.

Invertible blocks kill both obstructions

RI OBSTRUCTION

$$L_A z = z - R_A R_{AA}^{-1} z_A, \quad L_A R_i = 0 \quad (i \in A).$$

Projected derivatives plus inward smoothing
give

$$\mathcal{Q}_{\text{RI}} = 0.$$

R_{AA}^{-1} removes active reflection directions; $R_{AA}^{-\top}$ prescribes active oblique jets.

$$\ker \mathcal{A} = \mathbb{R} s_0.$$

BOUNDARY OBSTRUCTION

$$\begin{aligned} \nabla_T f|_{S_A} &= 0, & a &:= \nabla_A f|_{S_A}, \\ a = R_{AA}^{-\top} \psi &\implies & D_A f|_{S_A} &= \psi. \end{aligned}$$

Localized tests prescribe every active jet,
hence

$$\ker \partial_R = \{0\}.$$

Source: Lu and Zhu, Section 7, positive Harrison Reiman argument.

A singular block creates a boundary source

ASSUMPTIONS AND NULL DIRECTION

R nonsingular, $\emptyset \neq A \subsetneq J$, R_{AA} singular, $0 \neq v \in \ker R_{AA}$, $w := R_{TA}v \neq 0$.

THE ACTIVE NORMAL PART DISAPPEARS

$$\sum_{i \in A} v_i D_i f = \underbrace{(R_{AA}v) \cdot \nabla_A f}_0 + \underbrace{(R_{TA}v) \cdot \nabla_T f}_{w \cdot \nabla_T f}.$$

Let $\iota_A(y) := (0_A, y_T) \in S_A$. Choose $\phi \in C_c^\infty((0, \infty)^T)$ with $w \cdot \nabla_T \phi \not\equiv 0$ and define

$$\zeta_i = v_i(\iota_A)_\#(\phi(y) dy) \quad (i \in A), \quad \zeta_i = 0 \quad (i \notin A).$$

Tangential integration by parts gives

$$\partial_R \zeta(f) = - \int_{(0, \infty)^T} (w \cdot \nabla_T \phi)(y) f(\iota_A(y)) dy =: \chi(f), \quad \chi \neq 0, \quad \chi(E) = 0.$$

Source: Lu and Zhu, Proposition 6.1 and Section 7, singular boundary symbol.

Zero potential lifts the source into an RI defect

ZERO POTENTIAL

$$Q_\chi = (U_\chi, \Theta_\chi), \quad U_\chi := \int_0^\infty \chi P_t dt.$$

Under the paper's Feller, recurrence, ergodicity, and regulator moment assumptions,

$$\mathcal{A}(Q_\chi) = -\chi, \quad \mathcal{A}(0, \zeta) = +\chi.$$

$$Q_\chi + (0, \zeta) = (U_\chi, \Theta_\chi + \zeta) \in \ker \mathcal{A}. \quad (7.5)$$

The source survives as an interior resolvent defect:

$$\mathfrak{r}(U_\chi)(\lambda, h) = - \int_E R_\lambda h d\chi, \quad \chi \neq 0 \implies 0 \neq [U_\chi] \in \mathcal{Q}_{\text{RI}}. \quad (7.6)$$

One block decides: collapse or a nonzero class.

Source: Lu and Zhu, Proposition 7.3 and the completely $\mathcal{S}$ obstruction.

Three weeks of AI assisted mathematical search

PHASE I · STRESS TEST

Boundary layer expansions
Sign and cancellation
checks
Search for a failure regime

PHASE II · STRUCTURE

Kato type route
Stratum bookkeeping
150 exploratory pages

PHASE III · VERIFICATION

Human checking of every
condition
Literature connection
Shorter resolvent proof

Independent review

Ten model reviewers were used only as an initial screening filter. Model agreement was evidence about where to inspect, never a proof certificate.

What the AI did, and what it did not do

AI assistance

- Exploratory boundary expansions
- Counterexample stress testing
- Proof organization
- Literature discovery
- Independent screening

Human responsibility

- Reframing the mathematical question
- Choosing the regime split
- Constructing the counterexample
- Checking every assumption and proof step
- Taking sole responsibility for correctness

The crucial disclosure

The final proof was not generated in one pass. One shot attempts by ChatGPT 5.5 Pro extended and Claude Opus 4.8 max both failed. Additional attempts supplied with the relevant literature also remained unsuccessful. Model consensus was treated as a search heuristic, never as proof.

Source: Signed BAR manuscript, full AI assistance and responsibility disclosure.

In the new Era of AI4Math

- Thinking more on the formulation of the problem (connection to the application) rather than the technical side.
- Working on much harder problems.

Example

We are starting working on LDP for random matrix theory.

Zhu, Youheng, and Yiping Lu. "High-Dimensional Interpolators Can Be Fragile: Heavy Tails and High-Dimensional Large Deviations." arXiv preprint arXiv:2607.09547 (2026).

High-dimensional ridgeless interpolators suffer from heavy-tailed risk and exhibit a substantially slower large deviation principle (LDP) rate (scaling as $n \log n$ rather than n^2) compared to ridge regression.

AI4Math needs a new scorecard

Measure failure probability, memory state, and human understanding.

01

Benchmark the FAILURE Case

Preselect the problem, then count every attempt.
Solved examples alone hide the denominator.

$$p_{\text{solve}}(\mathcal{P} \mid M, \mu, C)$$

problem $\mathcal{P}$ · model M · memory μ · compute C

02

MEMORY is part of the experiment

Same prompt + different memory = different output.
Report and ablate the memory profile.

memory state changes the output distribution

Distill recurring behavior into a memory profile

03

CREDIT shifts when Proof gets cheap

Problem choice, explanation, verification, writing,
and provenance become part of the contribution.

annotate why the proof works

human proof understanding becomes training data
for future models

Record the **failure log**, **memory state**, **contribution ledger**, and **human explanation**.

SCALE

CORRECT

VERIFY



COMPUTE

WITH STRUCTURE

Thank you

Questions?

Yiping Lu

Beijing International Center for Mathematical Research, Peking
University

Assistant Professor, Northwestern University