A Constant-Competitive Algorithm for Dynamic Mixture-of-Experts Serving

Ian D'Ambrosio
Nth Research Collective
ian@nthresearch.org

11 August 2026

Abstract

Huang, Lou, and Xiao introduced Dynamic Mixture-of-Experts Serving and gave an $O(\sqrt{\log k})$-competitive randomized algorithm for its integral primal problem, where k is the number of replica GPUs beyond the mandatory copy of each expert. Their matching lower barrier applies to an auxiliary dual and leaves the primal order open. We prove that the randomized primal competitive ratio is in fact $\Theta(1)$ for arbitrary numbers of experts. The upper bound reduces reciprocal-max service costs to chasing positive bodies with covering row sparsity two. A finite tangent envelope approximates each reciprocal epigraph within a constant factor, summable positive resets convert accumulated service into movement, and a nonexpansive balanced projection removes the positive-body algorithm's resource augmentation. Combining the resulting fractional path with Lazy Threshold Rounding gives

$$\mathbb{E}[\mathrm{ALG}] \leq 10C_{\mathrm{PB}} \cdot \mathrm{OPT} + (5C_{\mathrm{PB}} + 2)k + 16,$$

where C_{PB} is the absolute constant from Chasing Positive Bodies at resource augmentation one and covering sparsity two. The full reduction, rounding composition, and quantified main theorem are machine-checked in Lean 4 relative to exact formal interfaces for the two cited source theorems. Deterministic rational controls and a fresh independent replay accompany the formal proof.

1 Introduction

Dynamic Mixture-of-Experts (MoE) inference must decide how many replica GPUs to assign to each expert as workloads change. Moving replicas has a switching cost, while insufficient replication increases bottleneck latency. Huang, Lou, and Xiao formalized this tradeoff as an online problem and proved an $O(\sqrt{\log k})$ randomized competitive ratio for the integral primal problem [2]. Their $\Omega(\sqrt{\log k})$ construction concerns an offset-Fenchel-dual maximization problem. Weak duality points in the wrong direction for transferring it to a primal lower bound, and the asymptotic primal ratio remained open.

Our main result is that the primal problem has a constant randomized competitive ratio. The central observation is that the epigraph of $r/(1+u)$ admits a finite positive-polyhedral inner approximation with constant distortion. This permits a reduction to the positive-body chasing theorem of Bhattacharya, Buchbinder, Levin, and Saranurak [1]. Two obstacles must be handled explicitly. First, the reciprocal epigraph has a continuum of tangents, whereas the cited theorem takes finite matrices. Second, charging a soft service cost to movement suggests resetting an auxiliary height to zero, but a zero right-hand-side packing row lies outside the normalized source theorem. We use a finite tangent grid and geometrically shrinking positive resets. Both repairs preserve a horizon-independent constant.

The reduction yields a deterministic fractional path. Applying the Lazy Threshold Rounding lemma of Huang, Lou, and Xiao converts it online to an exact integral allocation with no expected movement loss and at most a factor-three service loss.

Verification. The new reduction is formalized in Lean 4. The formal theorem treats the two published ingredients as explicit source interfaces and proves every new MoE bridge, including the finite tangent construction, event/reset feasibility, the comparator movement bound, exact-budget projection, service charging, per-step-to-path rounding telescope, and final constant assembly. The source interfaces were audited against pinned source archives. A fresh independent referee reproduced 24 tests and both canonical artifact hashes and returned correctness, openness, and major-significance verdicts.

2 Model and main result

Fix $m \geq 1$ experts and $k \geq 1$ replica GPUs. The integral state space is

$$X_{m,k} = \left\{ x \in \mathbb{Z}_{\geq 0}^m : \sum_{i=1}^m x_i = k \right\}.$$

At round t, a nonnegative workload vector $r_t \in \mathbb{R}_{\geq 0}^m$ is revealed. The online algorithm chooses $x_t \in X_{m,k}$ and pays

$$f_t(x_t) + \|x_t - x_{t-1}\|_1, \qquad f_t(x) = \max_{i \in [m]} \frac{r_{t,i}}{1 + x_i},$$

where the integral initial state x_0 is part of the instance. For a fixed finite workload sequence, OPT is the minimum total cost over all integral paths. We use the fixed-sequence, or oblivious-adversary, convention underlying the expected-cost rounding result in [2].

For fixed k, write $\mathrm{CR}_{\mathrm{rand}}(k)$ for the infimum of factors Γ with the following property: for every $m \geq 1$ there is a constant $A_{m,k} \geq 0$ such that, for every integral initial state, some causal randomized integral algorithm satisfies $\mathbb{E}[\mathrm{ALG}] \leq \Gamma \cdot \mathrm{OPT} + A_{m,k}$ on every fixed finite workload sequence. The additive term may depend on m and k, but not on the horizon.

Let C_{PB} denote an absolute competitive constant supplied by Theorem 1.1 of [1] at resource augmentation $\epsilon = 1$ and maximum covering-row support $d = 2$. We absorb into C_{PB} the constant from reducing a positive body to positive halfspaces and the standard conversion between upward movement and full ℓ_1 movement.

Two properties of the service cost. The proof uses two elementary features of f_t. First, its reciprocal terms are smooth under capacity scaling: for every $r, u \geq 0$,

$$\frac{r}{1 + u/2} \leq 2 \frac{r}{1 + u}.$$

Thus halving capacity from $2k$ to k while retaining at least half of every coordinate increases service by at most a factor two. Second, f_t is the maximum of univariate coordinate functions. Its epigraph therefore separates into constraints involving only the height and one allocation coordinate, and the tangent discretization preserves covering-row support $d = 2$. Equation (4) makes the sparsity explicit, while (9) applies the capacity-scaling inequality.

Theorem 1 (Constant upper bound)**.** *For every $m, k \geq 1$ and every integral initial state, there is a causal randomized integral online algorithm such that every fixed finite nonnegative workload sequence satisfies*

$$\mathbb{E}[\mathrm{ALG}] \leq 10 C_{\mathrm{PB}} \cdot \mathrm{OPT} + (5 C_{\mathrm{PB}} + 2)k + 16.$$

The multiplicative constant is independent of m, k, and the horizon.

Every realized online path is also a feasible offline path. Scaling an equal-workload one-round instance therefore makes the additive constant negligible, so no competitive factor is below one and $\mathrm{CR}_{\mathrm{rand}}(k) = \Theta(1)$.

3 Two source theorems

We use the following exact consequences of published results.

Positive-body chasing. A positive body has the form

$$K_t = \{z \in \mathbb{R}^D_{\geq 0} : C^t z \geq \mathbf{1},\ P^t z \leq \mathbf{1}\},$$

where both matrices are nonnegative. Theorem 1.1 of [1] gives, for $\epsilon \in (0, 1]$, an $O(\epsilon^{-1} \log(d/\epsilon))$-competitive online chaser whose covering constraints remain exact and whose packing constraints may be violated by a factor $1 + \epsilon$; d is the largest covering-row support. Its movement is compared with an exact offline path beginning at zero. Thus, when $\epsilon = 1$ and $d = 2$, the movement factor C_{PB} is absolute, covering rows remain exact, and packing right-hand sides may be doubled. Strictly positive right-hand sides normalize to one without changing row support. A covering row with zero coefficients and zero right-hand side is vacuous and may be deleted.

Lazy Threshold Rounding. Lemma 4.4 of [2] applies online to any exact-budget fractional MoE path, using one shared set of random thresholds over all rounds. Starting from the given integral initial state, it produces integral exact-budget states $\bar{x}_t$ satisfying

$$f_t(\bar{x}_t) \leq 3 f_t(x_t), \tag{1}$$

$$\mathbb{E}\|\bar{x}_t - \bar{x}_{t-1}\|_1 \leq \|x_t - x_{t-1}\|_1. \tag{2}$$

The same thresholds are used in consecutive prefixes, so summing (2) gives the corresponding expected path-movement bound. The source implements each step greedily in polynomial time.

4 A finite positive approximation to reciprocal service

Fix a workload coordinate $r \geq 0$ and augmented allocation $u \geq 0$. Write $q = 1 + u$. For a scale $p \geq 1$, define the tangent

$$L_{r,p}(u) = \frac{r(2p - q)}{p^2} = \frac{r(2p - 1 - u)}{p^2}.$$

Lemma 2 (Tangent envelope)**.** *For all $r \geq 0$, $q > 0$, and $p > 0$,*

$$\frac{r}{q} - L_{r,p}(u) = \frac{r(q - p)^2}{qp^2} \geq 0.$$

If $0 \leq u \leq 2k$, some $p \in \{1, \ldots, 2k + 1\}$ also satisfies

$$L_{r,p}(u) \geq \frac{3}{4} \frac{r}{1 + u}.$$

Proof. The first identity follows by putting the two terms over qp^2. For the second claim, take $p = \lceil q \rceil$. Then $1 \leq p \leq 2k + 1$ and $q \leq p < q + 1 \leq 2q$. Hence $a = q/p \in (1/2, 1]$, and

$$\frac{L_{r,p}(u)}{r/q} = a(2 - a) \geq \frac{3}{4}.$$

The case $r = 0$ is immediate. □

Define the finite envelope

$$H_r(u) = \max_{1 \le p \le 2k+1} L_{r,p}(u).$$

Lemma 2 gives

$$\frac{3}{4}\frac{r}{1+u} \le H_r(u) \le \frac{r}{1+u} \qquad (0 \le u \le 2k). \tag{3}$$

For $r > 0$, the inequality $s \ge L_{r,p}(u)$ is equivalent to the positive covering row

$$\frac{p^2}{r(2p-1)}s + \frac{1}{2p-1}u \ge 1. \tag{4}$$

Its support has size at most two. This is the $d = 2$ sparsity supplied by the maximum-of-univariate form of f_t. Coordinates with $r = 0$ require no row; equivalently, one may insert and delete a vacuous zero-coefficient covering row with zero right-hand side, as in the formal reduction.

Remark 3. A geometric grid $p \in \{(3/2)^j : (3/2)^j \le 1+2k\}$ gives the same 3/4 factor with $O(\log(k+1))$ rows per expert. The concrete Lean reduction uses the elementary integer grid above; both envelope lemmas are machine-checked.

5 The event/reset reduction

Use augmented coordinates $(u_1, \dots, u_m, s) \in \mathbb{R}^{m+1}_{\ge 0}$. When request r_t arrives, present the positive body

$$E_t = \left\{ (u, s) : \sum_i u_i \le k, \quad s \ge L_{r_{t,i},p}(u_i) \text{ for all } i, p \right\}. \tag{5}$$

After taking the round-t MoE action, present the reset body

$$R_t = \left\{ (u, s) : \sum_i u_i \le k, \quad s \le \delta_t \right\}, \qquad \delta_t = 2^{-t} \quad (t \ge 1), \tag{6}$$

and set $\delta_0 := 0$ for the initial accounting convention. Both packing rows have positive right-hand side after normalization. The event covering rows have support at most two. Both bodies are nonempty. The reset is processed after the action at time t and before the next request, so the construction remains causal.

Run the positive-body chaser with $\epsilon = 1$. At an event, write its point as (u_t, s_t). At the following reset, write its height as b_t. Then

$$\sum_i u_{t,i} \le 2k, \qquad s_t \ge H_{r_{t,i}}(u_{t,i}) \quad \forall i, \qquad 0 \le b_t \le 2\delta_t. \tag{7}$$

5.1 Removing resource augmentation

Map every $u \ge 0$ with $\sum_i u_i \le 2k$ to the exact-budget fractional allocation

$$\rho(u) = \frac{k - \frac{1}{2}\sum_i u_i}{m}, \qquad x_i(u) = \frac{u_i}{2} + \rho(u). \tag{8}$$

Lemma 4 (Balanced projection)**.** *The map in* (8) *satisfies* $x(u) \ge 0$, $\sum_i x_i(u) = k$, $x_i(u) \ge u_i/2$, *and*

$$\|x(u) - x(v)\|_1 \le \|u - v\|_1.$$

Proof. The budget bound gives $\rho(u) \geq 0$, and summing (8) gives total mass k. Moreover,

$$\begin{aligned}\|x(u) - x(v)\|_1 &\leq \frac{1}{2}\|u - v\|_1 + m|\rho(u) - \rho(v)| \\ &= \frac{1}{2}\|u - v\|_1 + \frac{1}{2}\left|\sum_i (u_i - v_i)\right| \\ &\leq \|u - v\|_1.\end{aligned}$$

□

Combining (3), (7), and the capacity-halving bound $1 + x_i(u_t) \geq (1 + u_{t,i})/2$ gives

$$f_t(x(u_t)) \leq 2 \max_i \frac{r_{t,i}}{1 + u_{t,i}} \leq \frac{8}{3} s_t. \tag{9}$$

5.2 Charging service to vertical movement

Let a_{t-1} be the chaser height after the preceding reset, with $a_0 = 0$. Thus $a_{t-1} \leq 2\delta_{t-1}$ for every $t \geq 1$, including the initial case. Let V_t be the vertical movement while processing E_t, R_t. Endpoint distances imply

$$V_t \geq |s_t - a_{t-1}| + |s_t - b_t| \geq 2s_t - a_{t-1} - b_t.$$

Using (9), $a_{t-1} \leq 2\delta_{t-1}$, and $b_t \leq 2\delta_t$, we obtain

$$f_t(x(u_t)) \leq \frac{4}{3} V_t + \frac{8}{3}(\delta_{t-1} + \delta_t).$$

Because $\delta_0 = 0$ and $\sum_{t \geq 1} \delta_t = 1$, summing the error terms contributes at most $16/3$. If M denotes the chaser's complete movement, then

$$\sum_t f_t(x(u_t)) \leq \frac{4}{3} M + \frac{16}{3}. \tag{10}$$

The balanced projection is nonexpansive. Write $x_t := x(u_t)$ for $t \geq 1$ and identify the supplied integral x_0 with its real vector. The first state can be at distance at most $2k$ from x_0, the diameter of the exact simplex. Hence

$$\|x_1 - x_0\|_1 + \sum_{t=2}^{T} \|x_t - x_{t-1}\|_1 \leq M + 2k. \tag{11}$$

6 Comparison with the MoE optimum

Take any integral comparator path $y_1, \ldots, y_T$ and let $h_t = f_t(y_t)$. At event time use the body point (y_t, h_t), and at reset time use $(y_t, 0)$. The tangent-underestimator identity makes the event point feasible; the reset point is feasible as well. Starting from zero, this lifted body path has movement

$$\begin{aligned}k + \sum_{t=2}^{T} \|y_t - y_{t-1}\|_1 + 2\sum_{t=1}^{T} h_t \\ \leq k + 2\,\mathrm{cost}(y_1, \ldots, y_T).\end{aligned} \tag{12}$$

Applying the positive-body theorem to an optimal MoE path gives

$$M \leq C_{\mathrm{PB}}(k + 2\,\mathrm{OPT}). \tag{13}$$

We now prove the upper bound.

Proof of Theorem 1. Apply Lazy Threshold Rounding to the fractional path constructed above. Its two guarantees, together with (10) and (11), give

$$\begin{aligned}
\mathbb{E}[\mathrm{ALG}] &\le (M+2k) + 3\left(\frac{4}{3}M + \frac{16}{3}\right) \\
&= 5M + 2k + 16 \\
&\le 10C_{\mathrm{PB}} \cdot \mathrm{OPT} \\
&\quad + (5C_{\mathrm{PB}} + 2)k + 16,
\end{aligned}$$

where the last step uses (13). The additive term is independent of the horizon, and every step depends only on revealed requests and the one initial rounding seed. □

7 Formal and executable verification

The formal development uses Lean 4.32.2 and Mathlib 4.32.2. Its main upper-bound declaration is

- `DynamicMoeMain.dynamicMoe_explicit_constant_upper_of_source_theorems`.

It states the explicit factor $10C_{\mathrm{PB}}$. The proof files also contain the CPB-compatible body condition and the theorem telescoping HLX's per-step expected movement inequality. Lean reports only the standard axioms `propext`, `Classical.choice`, and `Quot.sound` for every audited declaration.

The source-relative boundary is explicit: the development does not reprove the complete CPB or HLX papers. It formalizes their exact consequences as primitive premises and proves all new Dynamic MoE implications from them. Pinned source archives and theorem-bearing members are checked by SHA-256 before replay.

From the repository root, the complete replay is:

```
python3 scripts/bootstrap_dynamic_moe_sources.py
python3 -m pytest tests/test_bootstrap_dynamic_moe_sources.py \
  tests/test_dynamic_moe_positive_body_eval.py \
  tests/test_dynamic_moe_positive_body_formal.py \
  tests/test_dynamic_moe_semantics_formal.py \
  tests/test_dynamic_moe_lower_bound_formal.py \
  tests/test_dynamic_moe_formal_eval.py -q
python3 -m evaluators.dynamic_moe_positive_body_eval \
  --output output/dynamic-moe-serving/positive-body-reduction-control.json
python3 -m evaluators.dynamic_moe_formal_eval \
  --output output/dynamic-moe-serving/formal-reduction-audit.json
```

The run passes 24 tests. The canonical artifact hashes are

exact control: `53c7ac33689b4cef22057fd1f9c040dab4e00a8aa13c5277b6a1b6e91e041a66`,
formal packet: `adcb13dad31d6fc688982068396e9733cf3d6f6444285f015cd0e6e1c1fde3f4`.

A fresh independent process reproduced both hashes exactly. The formal packet normalizes machine-local path and architecture metadata. The exact control includes exhaustive rational tests of the tangent envelope, reset service charge, exact-budget maps, movement normalization, and a small offline dynamic program, together with a deliberately corrupted movement negative control.

8 Scope and open directions

The theorem resolves the asymptotic randomized integral ratio under the fixed-sequence convention. It does not determine the best numerical constant, the deterministic integral competitive ratio, or the ratio against an adaptive adversary. The concrete integer tangent grid has $O(mk)$ covering rows per event. The geometric grid noted after Lemma 2 reduces this to $O(m \log(k+1))$ without changing the proof. The reset schedule uses exact real numbers; replacing it with another positive summable schedule can be useful for an explicit bit-complexity implementation.

Research-system disclosure. Beyond, the research system operated by Nth Research Collective, materially assisted with literature retrieval, hypothesis generation, formalization, executable controls, and adversarial review. Ian D'Ambrosio selected the target, directed the investigation, reconciled the evidence, verified the final claims, and accepts full responsibility for the manuscript.

References

[1] S. Bhattacharya, N. Buchbinder, R. Levin, and T. Saranurak, Chasing positive bodies, arXiv:2304.01889v2, 2023 (revised 2024). `https://arxiv.org/abs/2304.01889v2`.

[2] Z. Huang, Q. Lou, and T. Xiao, Mixture-of-Experts Serving, arXiv:2607.17880v1, 2026. `https://arxiv.org/abs/2607.17880v1`.