A TWO-STATE LOWER BOUND FOR CONSTANT-STEP NATURAL POLICY GRADIENT IN FINITE-HORIZON MARKOV DECISION PROCESSES

KOREA SUPERINTELLIGENCE LABS

Abstract. Barua and Khodadadian have recently shown that exact natural policy gradient (NPG) with a constant step size $\eta > 0$ and uniform initialization, run on a tabular finite-horizon Markov decision process with horizon H and rewards in $[0,1]$, satisfies $V^{\star,h}(s) - V^{\pi_T,h}(s) \le (H-h+1)\log|\mathcal{A}|/(\eta T) + (H-h+1)^2/T$ for every $T \ge 1$, and they ask, in their own words, whether "this quadratic dependence is unavoidable for constant step size NPG or is an artifact of our proof technique", noting that "a matching lower bound for the finite-horizon setting remains an open problem". We give a lower bound in that setting; we located no earlier one. For every horizon $H \ge 2$, every step size $\eta \ge 0$ and every $T \ge 1$ we exhibit a finite-horizon Markov decision process with two states, two actions and rewards in $[0,1]$ — $H-1$ independent two-armed bandits placed behind a single delay amplifier whose flat alternative is worse by a parameter tuned to (H,η,T) — on which the optimality gap of the T-th NPG iterate at the start state is known *exactly*: $V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) = \big((H-1)/(1+e^{\eta T}) + \Delta_T\big)/2$ with $\Delta_T = (H-1)S_T(\eta)/T$ and $S_T(\eta) = \sum_{t<T}(1+e^{\eta t})^{-1}$. Two consequences follow. The gap is at least $(H-1)/(4T)$ *for every step size, however large*, so the η-free second term of the upper bound cannot be deleted. That uniformity in the step size is what distinguishes this bound from the constant-step lower bounds we located in the literature, each of which decays to 0 as $\eta \to \infty$. And the gap is at least $(H-1)/(8\eta T)$ whenever $\eta T \ge 1$, so the first term is tight up to an absolute constant when $|\mathcal{A}| = 2$. Together these bracket the worst-case rate between $\Omega(H/T)$ and the published $O(H^2/T)$: exactly one factor of H remains open, and we do not close it. All results hold for all H, η, T in the stated ranges, and every theorem below has been formally verified in Lean 4 against *Mathlib*; no theorem here rests on a finite computation, and the computations we do report support no claim and are labelled as such.

1. Introduction

1.1. **The objects.** A *finite-horizon tabular Markov decision process* is a tuple $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$ with $\mathcal{S}$ and $\mathcal{A}$ finite, horizon $H \in \mathbb{N}$, transition kernels $\mathcal{P} = \{P^h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S}) \mid h = 1,\dots,H\}$ and reward functions $\mathcal{R} = \{R^h : \mathcal{S}\times\mathcal{A} \to [0,1] \mid h = 1,\dots,H\}$. Policies are *nonstationary*, $\pi = (\pi^1,\dots,\pi^H)$ with $\pi^h(\cdot \mid s) \in \Delta(\mathcal{A})$, and with the convention $V^{\pi,H+1} \equiv 0$ the value and action-value functions are

$$V^{\pi,h}(s) = \sum_a \pi^h(a \mid s)\, Q^{\pi,h}(s,a), \qquad Q^{\pi,h}(s,a) = R^h(s,a) + \sum_{s'} P^h(s' \mid s,a)\, V^{\pi,h+1}(s'). \tag{1}$$

Since $R^h \in [0,1]$ one has $0 \le V^{\pi,h}(s) \le H-h+1$. We write $V^{\star,h}$ for the optimal value function, obtained by backward induction with a maximum, and $\pi^\star$ for a policy attaining it.

Exact natural policy gradient (NPG) is the multiplicative-weights iteration on this object. Given a constant step size $\eta > 0$ and the uniform initialization $\pi_0^h(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$, it updates

$$\pi_{t+1}^h(a \mid s) = \frac{\pi_t^h(a \mid s)\exp\big(\eta\, Q^{\pi_t,h}(s,a)\big)}{Z_t^h(s)}, \qquad Z_t^h(s) = \sum_{a'} \pi_t^h(a' \mid s)\exp\big(\eta\, Q^{\pi_t,h}(s,a')\big), \tag{2}$$

for every $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$. The model is known and the action-value functions are evaluated exactly; there is no sampling and no approximation anywhere in this paper.

1.2. **The source, and the question it leaves open.** Barua and Khodadadian [1] prove the following finite-time guarantee for (2); it is their Theorem 1. Every number we attribute to [1] is read off a compilation of its arXiv version 1: in that numbering their equation (1) is our (2), while their

equation (2) is the variational (Kullback–Leibler policy-mirror-descent) form of the same update, used only in their analysis of increasing step sizes and playing no role here.

> "Consider a finite-horizon MDP and the NPG update (1) with constant step size $\eta > 0$ and uniform initialization $\pi_0^h(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$ for every $h \in [H]$ and $s \in \mathcal{S}$. Then, for every integer $T \geq 1$, horizon $h \in [H]$, and state $s \in \mathcal{S}$,
>
> $$V^{\pi^\star,h}(s) - V^{\pi_T,h}(s) \leq \frac{(H-h+1)\log|\mathcal{A}|}{\eta T} + \frac{(H-h+1)^2}{T}.$$"

Immediately after it they write, again verbatim:

> "The H^2 factor in Theorem 1 arises from Lemma 15, where the value potential is bounded by $\sum_{i=H-j}^{H}(H-i+1) = \frac{(j+1)(j+2)}{2} \leq (j+1)^2$. *We do not know whether this quadratic dependence is unavoidable for constant step size NPG or is an artifact of our proof technique.* In infinite-horizon discounted MDPs, [Johnson et al.] proved matching upper and lower bounds for exact PMD; *a matching lower bound for the finite-horizon setting remains an open problem.*"

and, in their concluding section: "Another direction is to establish lower bounds that determine whether the H^2/t rate in Theorem 1 and the H-factor losses in the multi-start reduction are necessary."

So the upper bound is published and the matching lower bound is stated by its authors to be open. This paper is about the lower-bound side.

1.3. **Results.** Write $\sigma(x) = 1/(1+e^x)$ — the logistic function reflected, so that $\sigma(0) = 1/2$ and σ is decreasing — and

$$S_T(\eta) \;=\; \sum_{t=0}^{T-1} \sigma(\eta t) \;=\; \sum_{t=0}^{T-1} \frac{1}{1+e^{\eta t}}, \qquad \Delta_T \;=\; \frac{(H-1)\,S_T(\eta)}{T}. \tag{3}$$

For $H \geq 2$ and $\Delta \in [0, H-1]$ we build in Section 3 an explicit process $A(H,\Delta)$ with two states, two actions, horizon H and rewards in $[0,1]$; the start state is written s_0. Our main result computes its optimality gap at the canonical parameter $\Delta = \Delta_T$ exactly.

Theorem 1.1 (Theorem A)**.** *Let $H \geq 2$, let $\eta \geq 0$ and let $T \geq 1$. In the process $A(H,\Delta_T)$, the T-th iterate π_T of constant-step-η NPG (2) from the uniform initialization satisfies*

$$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \;=\; \frac{(H-1)\,\sigma(\eta T) + \Delta_T}{2} \;=\; \frac{1}{2}\left(\frac{H-1}{1+e^{\eta T}} + \frac{(H-1)S_T(\eta)}{T}\right),$$

and in particular

$$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \;\geq\; \frac{(H-1)\,S_T(\eta)}{2T}.$$

Two corollaries follow by bounding $S_T(\eta)$ from below in two different ways, and they are the two comparisons with [1]'s Theorem 1 that we can make.

Corollary 1.2 (Corollary B: the η-free term cannot be deleted)**.** *Let $H \geq 2$, $\eta \geq 0$, $T \geq 1$. Then in $A(H,\Delta_T)$,*

$$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \;\geq\; \frac{H-1}{4T}.$$

The right-hand side does not depend on η and does not decay as $\eta \to \infty$. Consequently no upper bound of the shape $(H-h+1)\log|\mathcal{A}|/(\eta T)$ alone — that is, Theorem 1 of [1] *with its second term removed — can hold.*

Corollary 1.3 (Corollary C: the first term is tight up to a constant at $|\mathcal{A}| = 2$)**.** *Let $H \geq 2$, $\eta > 0$, $T \geq 1$ with $\eta T \geq 1$. Then in $A(H,\Delta_T)$,*

$$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \;\geq\; \frac{H-1}{8\,\eta\,T}.$$

The constant 8 is deliberately loose: the proof produces $2(1+e) = 7.437\ldots$, and 8 is kept because it makes the statement legible. Corollary 1.3 concerns the first term of Theorem 1 of [1] at $h = 1$, namely $H \log|\mathcal{A}|/(\eta T)$, evaluated at $|\mathcal{A}| = 2$; the $\log|\mathcal{A}|$ *growth* is untested by it, since $|\mathcal{A}| = 2$ throughout this paper.

It is worth recording the two-sided statement the three results add up to. For $k \geq 2$ let

$$\mathcal{G}_k(H, \eta, T) \;=\; \sup_{M,\, s} \big(V_M^{\star,1}(s) - V_M^{\pi_T,1}(s)\big) \tag{4}$$

be the supremum, over all finite-horizon processes M with horizon H, $|\mathcal{A}| = k$ actions, any finite state space and rewards in $[0,1]$, and over all states s of M, of the optimality gap of the T-th constant-step-η NPG iterate from the uniform initialization. Then, for $H \geq 2$, $\eta > 0$ and $T \geq 1$,

$$\frac{H-1}{4T} \;\leq\; \mathcal{G}_2(H,\eta,T) \;\leq\; \frac{H\log 2}{\eta T} + \frac{H^2}{T}, \qquad \text{and additionally} \quad \mathcal{G}_2(H,\eta,T) \;\geq\; \frac{H-1}{8\eta T} \;\text{ if } \eta T \geq 1, \tag{5}$$

the upper bound being Theorem 1 of [1] at $h = 1$ and the two lower bounds being Corollaries 1.2 and 1.3. The two ends of (5) differ by exactly one factor of H in the η-free term, and by an absolute constant in the $1/(\eta T)$ term.

Remark 1.4 (how loose Theorem 1 is on this process). On $A(H, \Delta_T)$ the ratio of the published bound to the exact gap of Theorem 1.1 can be written down. As $\eta \to \infty$ one has $\sigma(\eta T) \to 0$ and $S_T(\eta) \to \sigma(0) = 1/2$, so the exact gap tends to $(H-1)/(4T)$ while the bound of Theorem 1 of [1] at $h = 1$ tends to H^2/T; hence

$$\lim_{\eta\to\infty} \frac{H\log 2/(\eta T) + H^2/T}{V^{\star,1}(s_0) - V^{\pi_T,1}(s_0)} \;=\; \frac{4H^2}{H-1},$$

which is 16 at $H = 2$ and 136.125 at $H = 33$: linear in H, which is the one factor that (5) leaves open. This limit is an elementary consequence of Theorem 1.1 and is not part of the formal development.

1.4. **Why the construction is a family and not a single process.** The reward parameter Δ_T of Theorem 1.1 depends on *both* T and η, and it has to. For a *fixed* process, constant-step NPG converges at a linear (geometric) rate — Liu, Li and Wei [3] state "global linear convergence of softmax natural policy gradient for any constant step size" as one of the conclusions of their analysis in the discounted setting — so the gap of any one process decays faster than $1/T$ eventually, and no single process can witness a $\Omega(H/T)$ rate for all T. The same point is made structurally by Johnson, Pike-Burke and Rebeschini [2], verbatim: "As the step-size tends to infinity, any PMD update recovers a PI update. This implies that general PMD can be arbitrarily close to PI's exact convergence for the same finite number of iterations. Thus, any lower-bound on the convergence of PMD must be limited to finite iterations." A lower bound in this setting is therefore necessarily a *family* indexed by (H, η, T), and the work is in tuning one reward so that the amplifier's accumulated evidence cancels *exactly* at time T. That is what Δ_T does.

1.5. **What is not claimed.** We do not settle whether the H^2 of [1] is necessary. Nothing in this paper implies either that $\mathcal{G}_2 = \Theta(H^2/T)$ or that $\mathcal{G}_2 = O(H/T)$; the bracket (5) is what is proved, and Section 6 says what we know about the remaining factor and why we assert none of it. In particular the process $A(H, \Delta_T)$ is provably *not* a witness for $\Omega(H^2/T)$: its gap never exceeds $(H-1)/2$ (Proposition 4.1(3)). We also do not test the $\log|\mathcal{A}|$ factor, our processes having two actions throughout. Finally, the statement of Section 5 about policy iteration is an observation that follows from Theorem 1.1 and standard facts, and we present it as such.

1.6. **Related work, and where this was looked for.** The reference [1] names for the matching bounds in the discounted setting is Johnson, Pike-Burke and Rebeschini [2]. Their result concerns a different quantity: they determine the linear γ-rate of exact policy mirror descent under an *adaptive* step size, and their lower bound — their Theorem 4.2 in the numbering of arXiv:2302.11381v3 — reads, verbatim, "Fix $n > 0$. There exists a class of MDPs with $|\mathcal{S}| = 2n+1$ and a policy $\pi^0 \in \operatorname{rint}\Pi$ such that running iterative updates of (4) for any positive step-size regime, we have for $k < n$: $\|V^\star - V^k\|_\infty \geq \frac{1}{2}\gamma^k\|V^\star - V^0\|_\infty$" — where their equation (4) is the policy-mirror-descent update. This is a statement

about the geometric rate on a fixed instance, not about the constant in front of $1/T$ under a constant step size. We record this neutrally: [2] is not a lower bound of the kind [1] asks for, and [1] does not claim it is.

The only prior lower bound for *constant-step* NPG that we located is Khodadadian, Jhunjhunwala, Varma and Maguluri [4], whose Proposition 3.1 reads, verbatim, "For an MDP with a single state, the convergence of NPG with constant step size is lower bounded as follows: $V^\star - V^{\pi_K} \geq \frac{\Delta}{(1-\gamma)|\mathcal{A}|} e^{-\eta\Delta K}$", where the instance is discounted with factor γ and $\Delta = r^\star - \max_{a \notin \mathcal{A}^\star} r(a)$ is its reward gap. This is the statement [2] describes, verbatim, as one that "only applies to MDPs with a single-state, which can be solved in a single iteration with exact policy evaluation as the step-size goes to infinity (for which the lower-bound goes to 0)": its right-hand side is $e^{-\eta\Delta K}$, which vanishes as $\eta \to \infty$. The nearest constant-step *sublinear* lower bound is for softmax policy gradient rather than NPG, in the discounted setting: Liu, Li and Wei [3] prove that for any constant step size $\eta > 0$ and any $\varsigma \in (0, 1)$ there is a time $T(\varsigma)$ beyond which the softmax policy-gradient iterates satisfy a bound whose constant carries the factor $\big(\exp(2\eta/(1-\gamma)^2) - 1\big)^{-1}$. Both of these decay to 0 as the step size grows. Corollary 1.2 does not, and that uniformity in η is the distinguishing feature of the bound proved here; it is what the second term of Theorem 1 of [1] is about. Two further lines concern the *exponential* rate rather than the sublinear one, both in the discounted setting: Li, Wei, Chi and Chen [5] on the exponential time that softmax policy gradient can take, and Müller and Cayci [6], who state that they "provide an anytime analysis of unregularized tabular natural policy gradients and show that the exponential convergence rate $O(e^{-\Delta\eta k})$ is tight up to polynomial terms". (This last is the work Müller, Çaycı and Montúfar have in mind when they write that "an essentially matching lower bound has been established for Kakade's NPG"; their bibliography key for it resolves to the same e-print, arXiv:2406.04163, which we cite directly as [6].) None of these is finite-horizon.

The searches behind the sentence "we located no finite-horizon lower bound for constant-step NPG" were run on 3 September 2026 and were these. A phrase pass over a local corpus of 369 shards (86 GB, roughly 883,000 arXiv full texts, announcements through early September 2026) for "natural policy gradient" or "policy mirror descent" isolated exactly 500 e-prints in the whole corpus containing either phrase; all 500 were extracted in full (1,480,966 lines) and searched. Thirty of them place "lower bound" on a line with either phrase; reading all thirty hit lines, every lower bound found is of one of three kinds — the exponential or linear rate on a *fixed* discounted instance, the adaptive-step γ-rate of [2], or a statement about softmax policy gradient rather than NPG — and not one is finite-horizon or uniform in η. A second pass over the same 500 for "H^2" on a line with either phrase returns [1] and nothing else. Live arXiv queries (`all:"natural policy gradient" AND all:"finite-horizon"`, four results; `all:"policy mirror descent" AND all:"lower bound"`, two results; `all:"natural policy gradient" AND all:"lower bound"`, ten results) added nothing on point. A citation lookup for [1] in OpenAlex returns count 0, but the same query for [2] — a NeurIPS 2023 paper — also returns 0, so that source is *no signal* here rather than a negative; Semantic Scholar likewise reports no citations of the e-print. Every one of these negatives should be read as "not found", never as "new". What makes this a settled open question rather than folklore is not the searches but the source: [1] states it, in its own words, twice.

2. Preliminaries

We record the algorithm in the coordinates the proof uses. Fix a process $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$ and a step size $\eta \in \mathbb{R}$, and define logit vectors $L_t^h(s, \cdot) : \mathcal{A} \to \mathbb{R}$ by

$$L_0 \;=\; 0, \qquad L_{t+1}^h(s,a) \;=\; L_t^h(s,a) + \eta\, Q^{\pi_t,h}(s,a), \qquad \pi_t^h(a \mid s) \;=\; \frac{e^{L_t^h(s,a)}}{\sum_{a'} e^{L_t^h(s,a')}}. \tag{6}$$

Lemma 2.1. *The iteration* (6) *is the NPG iteration* (2)*. Precisely:* $\pi_0^h(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$*; each* π_t *is a policy; the pair* (π_t, π_{t+1}) *satisfies* (2) *with its normalizer* $Z_t^h(s)$*; and the logits are the scaled cumulative*

action-values,

$$L_t^h(s,a) \;=\; \eta \sum_{k=0}^{t-1} Q^{\pi_k,h}(s,a).$$

Moreover $V^{\pi,h}(s) = \sum_a \pi^h(a \mid s) Q^{\pi,h}(s,a)$ *for* $h \le H$*, and* $V^{\pi,h}(s) \le V^{\star,h}(s)$ *for every policy* π*, every* h *and every* s*, where* $V^\star$ *is defined by backward induction with a maximum.*

Proof sketch. Uniform initialization is $L_0 = 0$; that each π_t is a policy is positivity and normalization of the softmax. For (2), substitute $L_{t+1} = L_t + g$ with $g = \eta Q^{\pi_t}$ into the softmax and divide numerator and denominator by $\sum_{a'} e^{L_t(a')}$; this is the identity

$$\operatorname{softmax}(f+g)(a) \;=\; \frac{\operatorname{softmax}(f)(a)\, e^{g(a)}}{\sum_b \operatorname{softmax}(f)(b)\, e^{g(b)}},$$

whose denominator is exactly $Z_t^h(s)$. The cumulative form is a one-line induction on t. The one-step identity is the definition (1) once the recursion has fuel left, and $V^{\pi,h} \le V^{\star,h}$ is the usual induction on the number of remaining layers, using $\sum_a \pi^h(a \mid s) = 1$ and nonnegativity of the kernel. □

The cumulative form is not stated in [1] and it is not deep, but it is the fact that organizes everything below: each pair (h,s) runs an independent exponential-weights process on the *cumulative* action-value vector $\sum_{k<t} Q^{\pi_k,h}(s,\cdot)$. In particular, at two actions, if $\delta_t = Q^{\pi_t,h}(s,a_0) - Q^{\pi_t,h}(s,a_1)$ and $D_t = \sum_{k<t} \delta_k$ then

$$\pi_t^h(a_1 \mid s) \;=\; \frac{1}{1+e^{\eta D_t}} \;=\; \sigma(\eta D_t), \qquad \pi_t^h(a_0 \mid s) \;=\; 1 - \sigma(\eta D_t). \tag{7}$$

Everything in Section 4 is an exact evaluation of D_t for the two decisions of the process $A(H,\Delta)$, followed by (7) at $t = T$.

We shall use three elementary facts about $S_T(\eta)$ of (3), valid for $\eta \ge 0$ and $T \ge 1$:

$$0 \le S_T(\eta) \le \frac{T}{2}, \qquad S_T(\eta) \ge \sigma(0) = \frac{1}{2}, \qquad S_T(\eta) \ge \frac{1}{(1+e)\,\eta} \;\text{ when } \eta > 0,\ \eta T \ge 1. \tag{8}$$

The first two are termwise (σ is decreasing and $\sigma(0) = 1/2$) and the third counts terms: with $m = \lceil 1/\eta \rceil$ one has $m \le T$ because $1/\eta \le T$, and for $t < m$ one has $\eta t \le \eta(m-1) < 1$, so each of the first m terms is at least $\sigma(1) = 1/(1+e)$, while $m \ge 1/\eta$.

3. The amplified product family

Definition 3.1. Let $H \ge 1$ and $\Delta \in \mathbb{R}$, and for $H \ge 2$ set $z = 1 - \Delta/(H-1)$. The process $A(H,\Delta)$ has $\mathcal{S} = \{0,1\}$, $\mathcal{A} = \{a_0, a_1\}$, horizon H, deterministic transitions, start state $s_0 = 0$, and the following rewards and successors.

layer	state	a_0	a_1
1	$0\ (= s_0)$	reward 0, $\to 0$	reward 0, $\to 1$
$2,\dots,H$	0	reward 1, $\to 0$	reward 0, $\to 0$
$1,\dots,H$	1	reward z, $\to 1$	reward z, $\to 1$

State 1 is unreachable at layer 1, so its row there is recorded only for completeness. Everything in this paper concerns $H \ge 2$, with the single exception of Proposition 4.1(1), which uses the degenerate instance $H = 1$: there the only live layer is layer 1, the deep block is empty, and the value of z never enters.

The three blocks play three different roles.

The bandits. At every layer $h \ge 2$ in state 0, both actions lead to state 0, so the continuation value is the same for both and, by (1),

$$Q^{\pi,h}(0,a_0) - Q^{\pi,h}(0,a_1) \;=\; R^h(0,a_0) - R^h(0,a_1) \;=\; 1 \qquad \text{for } \textit{every} \text{ policy } \pi. \tag{9}$$

The $H-1$ deep layers are therefore $H-1$ independent copies of one two-armed bandit with action-value gap exactly 1, and no policy can change what they look like. This is what makes their NPG behaviour computable in closed form for all t at once.

The flat branch. State 1 is absorbing and pays z at every layer regardless of the action, so it carries no information: under every policy its value over k remaining layers is exactly kz, and $V^{\pi,2}(1) = (H-1)z = (H-1) - \Delta$.

The amplifier. Layer 1 is a single decision between the two: a_0 enters the bandit block, a_1 commits to the flat branch, and both pay 0 immediately. Its two action-values are the two continuation values, so the amplifier's evidence is a running comparison between what the bandits have learned so far and the fixed alternative $(H-1) - \Delta$.

Proposition 3.2 (admissibility and non-vacuity)**.** *Let $H \geq 2$.*

1. *If $0 \leq \Delta \leq H-1$ then every reward of $A(H,\Delta)$ lies in $[0,1]$, so $A(H,\Delta)$ is a process of the class considered in* [1].
2. *For every $\eta \geq 0$ and $T \geq 1$, the canonical parameter satisfies $0 \leq \Delta_T \leq H-1$; in particular $z \in [0,1]$ and part (1) applies to $A(H,\Delta_T)$.*
3. *If $0 \leq \Delta$ then $V^{\star,1}(s_0) = H-1$, and this value is attained: the deterministic policy that plays a_0 at every layer and state achieves it.*

Proof sketch. (1) The only reward other than 0 and 1 is $z = 1 - \Delta/(H-1)$, which lies in $[0,1]$ exactly when $0 \leq \Delta \leq H-1$. (2) is $0 \leq S_T(\eta) \leq T/2$ from (8); the same two facts in fact give the sharper $\Delta_T \leq (H-1)/2$ and $z \in [1/2, 1]$, which is never needed below and is not what is formalised. (3) Backward induction: over k layers from state 0 at a deep layer the maximum is k (take a_0 each time), and from state 1 it is kz; at layer 1 the maximum of $(H-1)$ and $(H-1)z$ is $H-1$ since $z \leq 1$. Attainment is the direct evaluation of (1) along the always-a_0 policy. Part (3) is what makes $V^{\star,1}(s_0) - V^{\pi_T,1}(s_0)$ the optimality gap of [1] and not an over-estimate obtained from a maximum that no policy realises. □

4. The exact gap, and the two lower bounds

Proof of Theorem 1.1, in five steps. Throughout, π_t is the NPG iterate of (2) in $A(H,\Delta)$ at step size $\eta \geq 0$, and we use the logit form (6) and the two-action identity (7).

Step 1 (the deep logits). By (9) the action-value gap at every layer $h \geq 2$ in state 0 is 1 under every policy, so by (6) the logit gap there after t steps is exactly ηt, by induction on t. Hence, by (7),

$$\pi_t^h(a_1 \mid 0) = \sigma(\eta t), \qquad \pi_t^h(a_0 \mid 0) = 1 - \sigma(\eta t) \qquad (h \geq 2),$$

the same for every deep layer. Note the value is exact for all t and needs no numerical bound on the exponential.

Step 2 (the two continuation values). Under any policy that plays a_0 with one fixed probability c at every layer $h \geq 2$ in state 0, the value of state 0 over k remaining layers is exactly kc: each layer contributes $c \cdot 1 + (1-c) \cdot 0$ and both actions return to state 0. With $c = 1 - \sigma(\eta t)$ from Step 1, $V^{\pi_t,2}(0) = (H-1)\big(1 - \sigma(\eta t)\big)$. Under every policy the value of state 1 over k layers is kz, so $V^{\pi_t,2}(1) = (H-1)z = (H-1) - \Delta$.

Step 3 (the amplifier's evidence). Layer 1 pays 0 and its successors are state 0 under a_0 and state 1 under a_1, so by (1) and Step 2,

$$Q^{\pi_t,1}(s_0,a_0) = (H-1)\big(1 - \sigma(\eta t)\big), \qquad Q^{\pi_t,1}(s_0,a_1) = (H-1) - \Delta,$$

hence $\delta_t = \Delta - (H-1)\sigma(\eta t)$ and, summing over $t < T$,

$$D_T \;=\; \sum_{t<T} \delta_t \;=\; T\Delta - (H-1)S_T(\eta).$$

Step 4 (the cancellation). Take $\Delta = \Delta_T = (H-1)S_T(\eta)/T$. This is admissible by Proposition 3.2, and it makes $D_T = 0$ *exactly*. By (7) the amplifier is then an exact coin flip at time T,

$$\pi_T^1(a_0 \mid s_0) = \pi_T^1(a_1 \mid s_0) = \tfrac{1}{2},$$

whatever η is — the cancellation is in the exponent, so a large step size does not help.

Step 5 (evaluation). By (1) at $h = 1$ with the values of Steps 2–4,

$$V^{\pi_T,1}(s_0) = \tfrac{1}{2}(H-1)\big(1-\sigma(\eta T)\big) + \tfrac{1}{2}\big((H-1)-\Delta_T\big),$$

and $V^{\star,1}(s_0) = H - 1$ by Proposition 3.2(3). Subtracting gives the stated equality. The displayed inequality follows by dropping the nonnegative term $(H-1)\sigma(\eta T)/2$ and substituting Δ_T. □

Proof of Corollaries 1.2 and 1.3. Both combine the inequality of Theorem 1.1, namely

$$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \;\geq\; \frac{(H-1)S_T(\eta)}{2T},$$

with a lower bound on $S_T(\eta)$ taken from (8). For Corollary 1.2, $S_T(\eta) \geq 1/2$ — the $t = 0$ term alone, which is $\sigma(0) = 1/2$ for every η — gives $(H-1)/(4T)$. For Corollary 1.3, the bound $S_T(\eta) \geq 1/((1+e)\eta)$ gives $(H-1)/\big(2(1+e)\eta T\big)$, and $2(1+e) < 8$ because $e < 3$. □

Two remarks on what the proof does and does not use. It uses no numerical evaluation of the exponential function beyond $e^0 = 1$, monotonicity, and the single inequality $e < 3$ in Corollary 1.3; $S_T(\eta)$ is carried as a symbol throughout and is never evaluated. And it uses no exhaustive computation of any kind: Theorem 1.1 and both corollaries are proved for all $H \geq 2$, all real $\eta \geq 0$ and all $T \geq 1$ simultaneously, by induction on the step index and on the layer index. The step size is allowed to be 0, which is outside the range $\eta > 0$ of Theorem 1 of [1]; at $\eta = 0$ the iterates never move and the identity of Theorem 1.1 returns the gap $(H-1)/2$ of the uniform policy.

The mechanism deserves one sentence in isolation, because it is the only idea in the construction. The bandit block accumulates evidence at a rate that *decays*: after t steps the deep policy is still wrong with probability $\sigma(\eta t)$, and the total wrongness over T steps is $S_T(\eta)$, which by (8) is at least $1/2$ however large the step size, and at least of order $1/\eta$ once $\eta T \geq 1$. The amplifier is charged that entire history at once: its evidence D_T is the accumulated shortfall of the bandit block against a flat alternative, and setting the flat alternative to be worse by exactly the average shortfall $\Delta_T = (H-1)S_T(\eta)/T$ freezes the amplifier at the uniform policy at time T. The factor $H - 1$ enters because the amplifier sits above $H - 1$ bandits at once, so the same per-bandit shortfall is multiplied by their number. The one thing this gadget provably does not do is multiply regret — see Section 6.

Proposition 4.1 (controls)**.** (1) *At $H = 1$ the process $A(1, \Delta)$ has optimality gap exactly 0 at $(h, s) = (1, s_0)$, for every Δ, every η and every T. Consequently the $H - 1$ of Corollary 1.2 is not a disguised H: the statement obtained from it by replacing $(H-1)/(4T)$ with $H/(4T)$ and $H \geq 2$ with $H \geq 1$ is false.*

(2) *The bound of Corollary 1.2 cannot be weakened by a further factor of two: the statement "$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \leq (H-1)/(8T)$ for all $H \geq 2$, $\eta \geq 0$, $T \geq 1$" is false, already at $H = 2$, $\eta = 1$, $T = 1$.*

(3) *For all $H \geq 2$, $\eta \geq 0$, $T \geq 1$ one has $V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \leq (H-1)/2$ in $A(H, \Delta_T)$. In particular the statement "$V^{\star,1}(s_0) - V^{\pi_T,1}(s_0) \geq H - 1$" is false for this family, so $A(H, \Delta_T)$ is not a witness for a bound of order H^2 at fixed T.*

Proof sketch. (1) At $H = 1$ the only live layer is layer 1, whose rewards are all 0 and whose successors are outside the horizon; both $V^{\star,1}(s_0)$ and $V^{\pi,1}(s_0)$ are 0. (2) is Corollary 1.2 at $H = 2$, $\eta = 1$, $T = 1$, which gives $\geq 1/4 > 1/8$. (3) is the exact identity of Theorem 1.1 together with $\sigma(\eta T) \leq 1/2$ and $S_T(\eta) \leq T/2$. □

Part (3) is the reason Section 1.5 is worded as it is. The construction here reaches $\Omega(H/T)$ and is proved not to reach $\Omega(H^2/T)$; the question of [1] is about the supremum over all processes, and nothing about that supremum follows from a family whose gaps are all at most $(H-1)/2$.

5. Constant-step NPG is not iterated policy iteration

Discussing the role of the step size, [1] writes, verbatim: "Because the tabular model is known, its action-value functions can be evaluated exactly. As $\eta \to \infty$, the NPG update approaches the greedy policy-improvement step of exact policy iteration with respect to $Q^{\pi_t,h}$, whereas a finite η yields a

smooth policy-improvement step." That is correct as a statement about *one* update applied to a *fixed* π_t, and we do not dispute it. It does not, however, extend to the iterate sequence, because π_t itself carries the step size: by Lemma 2.1 the logit is $\eta \sum_{k<t} Q^{\pi_k}$, so the large-$\eta$ iterate follows the leader on the *cumulative* action-values, not on $Q^{\pi_{t-1}}$, and the two differ from $t = 2$ on. The family of Section 3 makes the difference quantitative. Let π_t^{PI} denote exact policy iteration from the uniform policy: π_{t+1}^{PI} is greedy with respect to $Q^{\pi_t^{\mathrm{PI}}}$, ties broken toward a_0.

Proposition 5.1. *Let $H \geq 2$, $\eta \geq 0$ and $T \geq 2$, and work in $A(H, \Delta_T)$. Then*

$$V^{\pi_T^{\mathrm{PI}},1}(s_0) \;=\; V^{\star,1}(s_0) \qquad \text{and} \qquad V^{\pi_T,1}(s_0) \;<\; V^{\star,1}(s_0).$$

More precisely $V^{\pi_{t+2}^{\mathrm{PI}},1}(s_0) = V^{\star,1}(s_0)$ for every $t \geq 0$ and every $\Delta \geq 0$: exact policy iteration is optimal here from step 2 on, while the NPG gap at step T is at least $(H-1)/(4T) > 0$ for every step size, however large.

Proof sketch. One greedy step fixes the deep layers, because by (9) their action-value gap is 1 under every policy, so π_1^{PI} already plays a_0 there. Given that, $Q^{\pi_1^{\mathrm{PI}},1}(s_0, a_0) = H - 1 \geq (H-1)z = Q^{\pi_1^{\mathrm{PI}},1}(s_0, a_1)$, so π_2^{PI} plays a_0 at the amplifier as well, and by Proposition 3.2(3) it is optimal. The second half is Corollary 1.2. On this family the greedy maximiser is a singleton at every step the statement uses — the gap is 1 at the deep layers and $\Delta_T > 0$ at the amplifier, since $S_T(\eta) \geq 1/2$ — so the conclusion does not depend on the tie-breaking convention. □

The mathematics of Proposition 5.1 is routine once Theorem 1.1 is available: exact policy iteration terminates in a finite-horizon process after at most H sweeps, and here after two. We state it because the contrast is easy to mis-transfer, and because the same structural point is on record for the discounted setting in [2], quoted in Section 1.4.

6. What remains

6.1. **The factor of H.** The bracket (5) leaves exactly one factor of H in the η-free term, and we believe the following is the right way to phrase the remaining question. Write $\mathrm{gap}_t^h(s) = V^{\star,h}(s) - V^{\pi_t,h}(s)$ and $C_T^h(s) = \sum_{t<T} \mathrm{gap}_t^h(s)$, the *cumulative regret* of the iterate sequence. Both ingredients of the reformulation are already in [1]. Their Lemma 2 gives $V^{\pi_{t+1},h}(s) \geq V^{\pi_t,h}(s)$ and their Lemma 10 turns that into $\mathrm{gap}_t^h(s) \geq \mathrm{gap}_T^h(s)$ for $t \leq T$, so $\mathrm{gap}_T^h(s) \leq C_T^h(s)/T$; and the last line of the telescoping chain in their proof of Theorem 1, immediately before that final appeal to Lemma 10, is exactly $C_T^h(s)/T$ bounded by

$$\frac{1}{T}\left(\frac{(H-h+1)\log|\mathcal{A}|}{\eta} + (H-h+1)^2\right), \qquad \text{that is,} \qquad C_T^h(s) \;\leq\; \frac{(H-h+1)\log|\mathcal{A}|}{\eta} + (H-h+1)^2,$$

the two summands being their Lemma 14 (the initial Kullback–Leibler potential $m_0^{h\to H}(s) \leq (H-h+1)\log|\mathcal{A}|$) and their Lemma 15 (the value potential $n_T^{h\to H}(s) \leq (H-h+1)^2$) respectively, with $j = H - h$. So the cumulative-regret bound is theirs, not ours; what is ours is only the decision to read the question through it. We do not formalise any of it. On that reading, the source's question about the constant in front of $1/T$ becomes a question about the largest cumulative regret of constant-step NPG: if $\sup C_T^1 = O\big(H(1+\log|\mathcal{A}|/\eta)\big)$ then Theorem 1 of [1] would improve to $O(H/T)$ with no new argument and the answer to their question would be "artifact". Conversely, a proof that $\sup C_T^1 = \Theta(H^2)$ would *not* settle the question, because $\mathrm{gap}_T \leq C_T/T$ is itself lossy: equality needs gap_t to be essentially constant on $[0, T]$, which is exactly what the geometric convergence of a fixed process forbids. Nothing in this subsection is asserted as a result of this paper.

The obstruction we ran into is worth stating, since it is where a reader might otherwise expect the present construction to keep going. The amplifier of Section 3 is *regret-conserving*, not regret-multiplying: it converts the early regret of the block below it into a late gap at rate one, and this is precisely why $A(H, \Delta_T)$ attains $\Omega(H/T)$ and, by Proposition 4.1(3), cannot attain $\Omega(H^2/T)$. Whether amplifiers composed in series, in parallel, or with overlapping branches can be made to multiply regret is open; we have no general statement in either direction, and every cascade model we wrote down that

predicted super-quadratic growth was refuted by Theorem 1 of [1] itself, in each case because it double-counted regret that the trajectory does not pay — a wrong action at one layer shields the mistakes below it exactly when it creates the perturbation.

6.2. **Computations, and what they do not support.** The following computations were run and are reported for reproducibility. *No statement in this paper depends on any of them*; Theorem 1.1, its corollaries and Propositions 3.2, 4.1 and 5.1 are proved for all (H, η, T) in their stated ranges, and Section 7 says in what sense.

Remark 6.1 (numerical reproduction; outside the formal development). An independently written simulator implementing (2) in its multiplicative form — with no logits, so that it shares no structure with the proof — was run in 200-digit decimal arithmetic. In 116 cells spanning $H \in \{2, \dots, 32\}$, η from 10^{-4} to 10^{6} and T from 1 to 200, the exact identity of Theorem 1.1, the value $V^{\star,1}(s_0) = H - 1$, the coin flip $\pi_T^1(a_1 \mid s_0) = 1/2$ and the admissibility $z \in [0, 1]$ all held to within 10^{-150}. A further 210 cells spanning $H \le 33$, η from 10^{-6} to 10^{6} and $T \le 120$ gave 1.000000000000 for the smallest ratio of the gap to the bound of Corollary 1.2. Twelve digits is all this check can ever show: the corollary's slack is $S_T(\eta) - \frac{1}{2}$, which is of order $e^{-\eta}$, so beyond $\eta \approx 10^4$ the true margin sinks below any working precision — there the computed ratio drifts below 1 by as much as 10^{-149}, which is the arithmetic's own error and not a counterexample. Another 84 cells gave 2.604557 for the smallest ratio of the gap to the bound of Corollary 1.3 as stated, and 2.421119 for the smallest ratio to the sharper $(H-1)/\big(2(1+e)\eta T\big)$ that its proof actually yields; both are comfortably above 1, the corollary's constant 8 being deliberately loose. As a control on our reading of [1], its Theorem 1 was itself checked on 125 cells of this family and on 270 randomly generated processes with $H \le 5$, $|\mathcal{S}| \le 3$, $|\mathcal{A}| \in \{2, 3\}$, with no violation and with a tightest observed ratio of bound to gap of 2.95, so the control is not vacuous. One methodological point is worth passing on: a first version of the same simulator in double precision *failed* at $\eta = 1000$, reporting *exactly twice* the correct gap. That is catastrophic cancellation and not a mathematical error — the identity $D_T = 0$ of Step 4 must hold to within $o(1/\eta)$ for the coin flip to be visible, and when it does not the amplifier collapses onto one action and contributes the whole $(H-1)/(2T)$ instead of half of it. How many cells are hit depends on how the double-precision update is arranged: 11 of the 40 cells $H \in \{2, 4, 8, 16, 32\}$, $\eta \in \{0.3, 1, 10, 1000\}$, $T \in \{50, 200\}$ in that first version, and 8 of the same 40 in a re-implementation that subtracts the largest exponent before normalising, all of them at $\eta = 1000$ and all of them off by the same factor 2. The lesson is not the count: it is that a double-precision check of this family at large step size can fail in either direction.

Remark 6.2 (an exact reduction, verified but not proved here). For a deterministic optimal policy with actions $a^\star(i, \tilde{s})$, the per-step recursion $\mathrm{gap}_t^h(s) = \mathbb{E}_{s' \sim P^h(\cdot \mid s, a^\star)}[\mathrm{gap}_t^{h+1}(s')] + A^{\pi_t, h}(s, a^\star)$ integrates against the $\pi^\star$-visitation measure to $C_T^h(s) = \sum_{i \ge h} \sum_{\tilde{s}} d_s^{h \to i, \pi^\star}(\tilde{s}) R_T^i(\tilde{s})$ with $R_T^i(\tilde{s}) = \sum_{t<T} A^{\pi_t, i}(\tilde{s}, a^\star)$ the regret of a single exponential-weights process; at two actions $A^{\pi_t}(a^\star) = \pi_t(b)\delta_t$ and $\pi_t(b) = (1 + e^{\eta D_t})^{-1}$ as in (7). These identities were verified in 60-digit arithmetic on 648 randomly generated processes ($H \le 6$, $|\mathcal{S}| \le 4$, $|\mathcal{A}| \le 3$) to a maximum residual of $1.4 \cdot 10^{-57}$. They are elementary and we expect them to be provable by the performance-difference lemma, but they are *not* part of the formal development and no statement above uses them; they are recorded because they are the objects the question of Section 6 is about.

6.3. **Two further directions.** Two questions we did not pursue seem worth naming. First, the $\log|\mathcal{A}|$ factor of Theorem 1 of [1]: Corollary 1.3 matches the first term only at $|\mathcal{A}| = 2$, and the construction here does not obviously extend to a lower bound that grows with the number of actions. Second, the step-size-free limit: at $\eta = \infty$ the NPG iteration degenerates, by Lemma 2.1, to follow-the-leader on $\sum_{k<t} Q^{\pi_k}$, a dynamical system with rational data on processes with rational rewards; the largest cumulative regret of that system is a finite, exactly computable question for each $(|\mathcal{S}|, |\mathcal{A}|, H)$, and a definite answer there would decide whether an η-free H^2 can live in that regime at all.

7. Verification

Every lemma, proposition, theorem and corollary stated above — Lemma 2.1, Theorem 1.1, Corollaries 1.2 and 1.3, and Propositions 3.2, 4.1 and 5.1 — has been formally verified in Lean 4 against *Mathlib* [8, 9]. Since *Mathlib* contains no Markov decision process library, the whole finite-horizon skeleton is built from scratch over $\mathbb{R}$: the tuple with its two kernel axioms, the layer-indexed value recursion carrying $V^{\pi,H+1} = 0$ as the base case, backward induction with a maximum together with the proof that no policy beats it, and the iteration (6); the four parts of Lemma 2.1 are what certify that the object formalised is the algorithm (2) of [1] and not a variant of it, and the attainment clause of Proposition 3.2(3) is what certifies that $V^\star$ is the optimal value and not an over-estimate. The development is three modules — the model, the family with its theorems, and the controls — of about 1140 lines in total, containing no `sorry` and declaring no axiom by hand; a print of the axiom set of each of the 24 declarations the results above rest on returns `propext`, `Classical.choice` and `Quot.sound` and nothing else. In particular there is no `sorryAx` and no compiled-evaluation axiom: every statement above is a statement about real numbers quantified over all $H \geq 2$, all η in a real interval and all $T \geq 1$, so no finite search space exists anywhere in this problem to decide, no decision procedure over one is invoked, and no external solver or certificate is used. The three modules compile in about 84 seconds. The computations of Remarks 6.1 and 6.2, and the limit of Remark 1.4, are outside the formal development and are labelled where they appear. The source of the development is currently held in a private repository: repository reference to be added at submission.

On the reference list. Each entry records how far the reading went. Every number we attribute to another paper was read off a compilation of that paper's own LaTeX e-print, at the version cited, rather than taken from a third paper's description of it. For [1] that compilation gives: the multiplicative NPG update is its equation (1) and its equation (2) is the variational form; the upper bound is its Theorem 1; the performance-difference lemma is its Lemma 1, the one-step improvement bound its Lemma 2, the monotonicity transfer its Lemma 10, the Kullback–Leibler potential bound its Lemma 14 and the value potential bound its Lemma 15; and the sentence about lower bounds quoted in Section 1 is in its Section 6.

References

[1] A. Barua and S. Khodadadian, *Finite-Time Analysis of the Natural Policy Gradient in Finite-Horizon Markov Decision Processes*, arXiv:2607.22982v1 [cs.LG] (cross-listed math.OC, stat.ML), submitted 25 July 2026. The source of the problem and of everything attributed above: the setting of Section 1, the update (2), the upper bound quoted in Section 1, the open question, the proof whose cumulative-regret form Section 6 uses, and the remark on $\eta \to \infty$ quoted in Section 5. Read here in full, and compiled from its own e-print source to fix the numbering quoted above. arXiv record checked live on 3 September 2026: a single version (submitted 25 July 2026), no journal reference and no publisher DOI (only the arXiv-issued DataCite DOI `10.48550/arXiv.2607.22982`), and a comment field reading "Accepted at the Reinforcement Learning Conference (RLC 2026)". No proceedings version was found; DBLP indexed only the e-print on that date. A later version could of course answer the question quoted in Section 1; what this paper reports as open is what version 1 states to be open.

[2] E. Johnson, C. Pike-Burke and P. Rebeschini, *Optimal Convergence Rate for Exact Policy Mirror Descent in Discounted Markov Decision Processes*, arXiv:2302.11381v3 [math.OC]; Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Cited for the adaptive-step γ-rate and its matching lower bound (their Theorem 4.2 in the numbering of v3), for the remark that any lower bound on policy mirror descent must be limited to finitely many iterations, and for their description of [4]. Full text read here; the three passages quoted above are verbatim from it. The venue is recorded by the e-print's own comment field and by DBLP; neither the DBLP record nor the online proceedings carries a page range, so none is quoted here.

[3] J. Liu, W. Li and K. Wei, *Elementary Analysis of Policy Gradient Methods*, arXiv:2404.03372v2 [math.OC]. Cited for the linear convergence of softmax natural policy gradient at any constant step size on a fixed discounted process, and for their sublinear lower bound for softmax policy gradient, whose constant carries the factor $(\exp(2\eta/(1-\gamma)^2)-1)^{-1}$. The theorem statement and the abstract were read here.

[4] S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma and S. T. Maguluri, *On the Linear Convergence of Natural Policy Gradient Algorithm*, in: 2021 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, 3794–3799, doi:10.1109/CDC45484.2021.9682908; e-print arXiv:2105.01424v1 [cs.LG]. Cited as the only prior lower bound for constant-step NPG that we located. Its Proposition 3.1, quoted in Section 1.6, was read in the e-print itself, not through [2]; the proposition number is that of v1. This is the work [2] cites as `Khodadadian2021OnAlgorithm`: that

key resolves, in [2]'s own bibliography, to these four authors, this title, and this CDC volume and page range, all of which agree with the CROSSREF record for the DOI above and with the arXiv record of the e-print.

[5] G. Li, Y. Wei, Y. Chi and Y. Chen, *Softmax policy gradient methods can take exponential time to converge*, Mathematical Programming **201** (2023), 707–802, doi:10.1007/s10107-022-01920-6; e-print arXiv:2102.11270v3 [cs.LG]. Cited in Section 1.6 as an exponential lower bound for softmax policy gradient in the discounted setting. Title, authors and identifier confirmed against the arXiv record on 3 September 2026 and the journal reference against CROSSREF; the abstract was read here.

[6] J. Müller and S. Cayci, *Optimal Rates of Convergence for Entropy Regularization in Discounted Markov Decision Processes*, arXiv:2406.04163v5 [math.OC]. Cited for the tightness of the exponential rate of unregularized tabular natural policy gradient in the discounted setting; the sentence quoted in Section 1.6 is verbatim from it. Title, authors and identifier confirmed against the arXiv record on 3 September 2026; the author names are spelled as on the e-print itself. Version 1 of this e-print (June 2024) carries the title *Essentially Sharp Estimates on the Entropy Regularization Error in Discrete Discounted Markov Decision Processes*, which is how [7] cites it; the sentence quoted above was read in version 5.

[7] J. Müller, S. Çaycı and G. Montúfar, *Fisher–Rao Gradient Flows of Linear Programs and State-Action Natural Policy Gradients*, SIAM Journal on Optimization **35** (2025), 1060–1088, doi:10.1137/24M1653422; e-print arXiv:2403.19448v2 [math.OC]. Cited only in Section 1.6, for the sentence quoted there; its bibliography key for the "essentially matching lower bound" was resolved, in its own `.bib` and `.bbl` files, to arXiv:2406.04163, i.e. to [6].

[8] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE 28, Lecture Notes in Computer Science **12699**, Springer, 2021, 625–635, doi:10.1007/978-3-030-79876-5_37.

[9] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, 367–381, doi:10.1145/3372885.3373824.