Set $g'(x) = 0$:

$$\ln\Big[\big(1 - \tfrac{T}{Mx}\big)\big(1 - \tfrac{T}{C}\big)\Big] \;+\; \frac{T}{Mx\big(1 - \tfrac{T}{Mx}\big)} = 0.$$

Let $A \;=\; \frac{1}{1-\frac{T}{Mx}}$, then we have

$$\ln\Big[\big(1 - \tfrac{T}{C}\big)\Big] \;+\; A - 1 = \ln(A).$$

Let $z := 1 - T/C$. (Since $T/C < 1$, $z = 1 - T/C > 0$.) By moving terms, we have:

$$-\frac{z}{e} \;=\; -A\exp(-A).$$

Therefore,

$$A = -W^{-1}(-\frac{z}{e}) = -Z,$$

Finally, we have

$$N(M,T) = x = \frac{TZ}{M(Z+1)}$$

Here $W(\cdot)$ is the **Lambert W function**, and for $0 < 1 - \frac{T}{C} < 1$, the argument $\alpha = -\frac{1-T/C}{e}$ lies in the interval $\big(-\frac{1}{e}, 0\big)$. This means there are two real branches W_0 and W_{-1} in that domain, but since $\frac{Z}{Z+1} > 0$,we have $Z < -1$. Therefore, we only take the solution on branch W_{-1}. □

G.3 Proof of Corollary 4.4

Corollary 4.4 (Scaling laws of Optimal CoT Length)**.** *Based on Theorem 4.3, one can derive:*

- $N^*(M,T)$ *increases monotonically with* T*, i.e., harder tasks require more reasoning steps to attain the optimal performance.*
- *The optimal number of operators per step* $t^* = T/N^*(M,T) = M(1 + 1/Z)$ *increases monotonically with* T*. This aligns with the envelope curve result (Figure 3a).*
- $N^*(M,T)$ *decreases monotonically with* M*, i.e., more capable models require fewer reasoning steps to attain the optimal performance, reflecting the simplicity bias.*

Proof. The second and third conclusions can be easily derived through monotonic composition, so we primarily focus on proving the first point. We begin the proof by incorporating the notation from G.2. We have

$$g'(x) = \Big[\ln\big(1 - \tfrac{T}{Mx}\big) \;+\; \frac{T}{Mx\,\big(1 - \tfrac{T}{Mx}\big)}\Big] \;+\; \ln\big(1 - \tfrac{T}{C}\big),$$

and $x^*(T)$ such that $g'(x^*(T)) = 0$.

Let $F(x^*(T), T) = g'(x^*(T)) = 0$ We want to see how $x^*(T)$ changes as T changes, therefore we take total derivative w.r.t. T. By the chain rule,

$$0 \;=\; \frac{d}{dT}\,F\big(x^*(T),\,T\big) \;=\; \underbrace{\frac{\partial F}{\partial x}\big(x^*(T), T\big)}_{\text{call this } F_x} \cdot \frac{\partial x^*}{\partial T}\big(T\big) \;+\; \underbrace{\frac{\partial F}{\partial T}\big(x^*(T), T\big)}_{\text{call this } F_T}.$$

Hence

$$\frac{\partial x^*}{\partial T}\big(T\big) = \;-\;\frac{F_T\Big(x^*(T),\,T\Big)}{F_x\Big(x^*(T),\,T\Big)}.$$

So the sign of $x'^*(T)$ is the opposite of the sign of F_T, provided $F_x \neq 0$.

Since

$$F_x\left(x,\,T\right) = -\frac{T^2}{x(Mx-T)^2} < 0, \forall x > 0, \tag{16}$$

all we need to prove is

$$F_T\left(x^*(T),\,T\right) = \frac{T}{(Mx^*(T)-T)^2} - \frac{1}{C-T} > 0. \tag{17}$$

That is

$$\frac{\sqrt{T(C-T)}+T}{M} > x^*(T). \tag{18}$$

Let $x_0(T) = \frac{\sqrt{T(C-T)}+T}{M}$ be the test point.

According to Lemma G.1, $F(x_0(T),T) < 0$. Since $F(x^*(T),T) = 0$, and $F_x\left(x^*(T),\,T\right) < 0$, we have $x_0(T) > x^*(T)$.

Thus, $F_T\left(x^*(T),\,T\right) > 0$ holds and we have proved our corollary with $\frac{\partial x^*}{\partial T}(T) > 0$.

□

G.4 Proof of Theorem F.3

Theorem F.3. *For a noise function $0 < \sigma(T) < 1$ and a subtask error rate function $0 < E(N,M,T) < 1$ satisfying Assumptions F.1 and F.2, the general final accuracy function $A(N)$ from Proposition 4.2 has the following properties:*

- $\lim_{N\to+\infty} A(N) = 0$. *(Excessively long chains always fail.)*
- *If $A(N)$ has a maximum at $N^* > 1$, then N^* has a lower bound related to M and T:*

$$N^* \geq N_{LB}(M,T) = E_N^{-1}\left(1 - \frac{1}{e^2(1-\sigma(T))}; M, T\right), \tag{6}$$

where $E_N^{-1}(\cdot; M, T)$ is the inverse of $E(N,M,T)$ with respect to N.

Proof. (1) Since $0 < A(N) < (1-\sigma(T))^N$, and $\lim_{N\to+\infty}(1-\sigma(T))^N = 0$, $\lim_{N\to+\infty} A(N,M,T) = 0$

(2) Let $g(x)$ denote $E(x,M,T)$ and define $f(x) = \ln A(x)$. Then,

$$f'(x) = \ln(1-\sigma(T)(1-g(x))) - \frac{xE'(x)}{1-E(x)} \tag{19}$$

$$< \ln(1-\sigma(T)(1-g(x))) + 2, \quad \text{(since } E \text{ is convex and } x = N \geq 1) \tag{20}$$

If $A(N)$ attains its maximum at some point $N^* > 1$, then $\ln(1-\sigma(T)) + 2 > 0$. Otherwise, we would have $f'(x) < \ln(1-\sigma(T)) + 2 \leq 0\ \forall x > 1$, leading to a contradiction.

Thus, it follows that $e^2(1-\sigma(T)) > 1$.

Now, define $N(M,T) = E^{-1}\left(1 - \frac{1}{e^2(1-\sigma(T))}\right)$, which satisfies

$$\ln(1-\sigma(T)(1-g(N(M,T)))) + 2 = 0.$$

If there exists $x^* < N(M,T)$ such that $f'(x^*) = 0$, then we obtain

$$0 = f'(x^*) < \ln(1-\sigma(T)(1-E(x))) + 2 < 0,$$

which is a contradiction. Hence, the assumption that $x^* < N(M,T)$ must be false.

Therefore, we conclude that $x^* = N^* > N(M,T)$.

□

G.5 Proof of Theorem F.6

Theorem F.6. *Let $\alpha_1 = T$, $\beta_1 = C - T$, $\alpha_2 = T$, and $\beta_2 = NM - T$. Then the expected error rates for sub-questions and sub-answers are given by $\mathbb{E}[\sigma_i] = \frac{T}{C}$ and $\mathbb{E}[e_i] = \frac{T}{MN}$, respectively. Based on these estimates, we can derive an upper bound $\hat{A}(N)$ on the final accuracy*

$$\mathbb{E}\left[\prod_{i=1}^{N}(1-e_i)(1-\sigma_i)\right] \leq \hat{A}(N) = \left[\left(1-\frac{T}{C+2N-1}\right)\left(1-\frac{T}{NM+2N-1}\right)\right]^N,$$

which initially increases and then decreases as the number of CoT steps N grows.

Proof. According to the multidimensional version of Hölder's inequality,

$$\mathbb{E}\left[\prod_{i=1}^{N}(1-e_i)(1-\sigma_i)\right] \leq \prod_{i=1}^{N}\left(\mathbb{E}[(1-e_i)^{2N}]\mathbb{E}[(1-\sigma_i)^{2N}]\right)^{\frac{1}{2N}} \tag{21}$$

$$\text{(Lemma G.2)} \leq \prod_{i=1}^{N}\left(1-\frac{T}{C+2N-1}\right)\left(1-\frac{T}{NM+2N-1}\right) \tag{22}$$

$$= \left[\left(1-\frac{T}{C+2N-1}\right)\left(1-\frac{T}{NM+2N-1}\right)\right]^N \tag{23}$$

□

G.6 Proof of Corollary 4.5

Corollary 4.5 (RL Converges to Optimal CoT Length)**.** *For gradient ascent on $J(\theta)$ with sufficiently small step size, the policy converges to a deterministic policy $\pi_\theta(N_i) = 1$ iff $i = \arg\max_j A(N_j)$. Thus, RL training converges to the optimal CoT length $N^* = \arg\max_{N\in\mathcal{A}} A(N)$.*

Proof. We treat the choice of CoT length as a k-armed stochastic bandit with action set $\mathcal{A} = \{N_1, \ldots, N_k\}$ and unknown success probabilities[7] $A(N_i) \in (0,1)$. Without loss of generality, relabel the arms so that

$$A(N_1) = \max_j A(N_j) =: A^*, \qquad A(N_1) \geq A(N_2) \geq \cdots \geq A(N_k).$$

The agent uses a softmax (Gibbs) policy

$$\pi_\theta(N_i) \;=\; \frac{e^{\theta_i}}{\sum_{j=1}^{k} e^{\theta_j}}, \qquad \theta \in \mathbb{R}^k, \tag{24}$$

and maximises the expected reward

$$J(\theta) \;=\; \sum_{i=1}^{k} \pi_\theta(N_i)\, A(N_i). \tag{25}$$

Because π_θ is C^∞ in θ and $A(N_i)$ are constants, J is smooth.

Under the REINFORCE estimator with sufficiently small, fixed step size $\eta > 0$, gradient ascent updates take the form

$$\theta^{(t+1)} \;=\; \theta^{(t)} + \eta\, \nabla_\theta J\big(\theta^{(t)}\big), \tag{26}$$

where

$$\frac{\partial J}{\partial \theta_i} \;=\; \pi_\theta(N_i)\Big(A(N_i) - J(\theta)\Big). \tag{27}$$

[7] By Proposition 4.2, $A(N_i)$ is the probability that the final answer is correct when a chain of length N_i is used. The bandit is *stationary* because $A(N_i)$ does not depend on time or the agent's past actions.