TOKEN CREDIT ASSIGNMENT AS A BOUNDARY-VALUE PROBLEM IN RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) provides sparse outcome-level feedback, while methods such as GRPO assign the same credit to every token in a response. With binary terminal rewards, exact local hindsight evidence and the initial value determine a Bellman value path whose increments are token advantages and whose endpoints fix their total budget. Motivated by on-policy distillation (OPD), teacher–student likelihood ratios provide surrogate evidence, but their bias and noise can cause the reconstructed path to violate this budget or value boundary. To reconstruct the Bellman value path guided by OPD, we propose three design requirements targeting the objective of RLVR, utilizing the information of OPD. These requirements lead to Endpoint-Locked Value-path Evidence Reconciliation (ELVER). ELVER-Affine projects raw credits onto the budget hyperplane in closed form, while ELVER-Box reconstructs a path with constrained value bounds, and it minimizes squared Euclidean distances to the evidence-induced transition lines, yielding a strictly convex quadratic program. Experiments on five mathematical reasoning benchmarks show that ELVER-Box outperforms standard methods.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) uses a verifier's judgment of completed responses to train large language models (LLMs). Group relative policy optimization (GRPO) estimates advantages by comparing responses to the same prompt, without a learned value model (Shao et al., 2024; Yu et al., 2025; Liu et al., 2025). Yet vanilla GRPO assigns the same advantage to every token in a response, leaving the token-level *credit-assignment problem* unresolved. A strategy choice, a key algebraic step, and a routine continuation receive identical credit despite their different contributions (Wang et al., 2025; Guo et al., 2025b).

Finer credit can come from prefix resampling, tree search, segment-level evaluation, or intermediate supervision (Kazemnejad et al., 2025; Hou et al., 2025; Guo et al., 2025b; Jia et al., 2025; Cui et al., 2026). Recently, on-policy distillation (OPD) instead obtains dense feedback from a teacher on the student's sampled prefixes (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). On-policy self-distillation (OPSD) uses the same model with privileged information, such as a verified solution, to construct a self-teacher (Zhao et al., 2026; Hübotter et al., 2026). Both provide local guidance, but can also introduce biases from the teacher or privileged information. Distillation encourages the student to imitate the teacher, which can limit exploration. RLVR encourages exploration, but coarse credit assignment can make learning inefficient. Recent work combines dense distillation feedback with RLVR outcome advantages (Yang et al., 2026a; Lin et al., 2026; Liu et al., 2026; LLM-Core Xiaomi, 2026).

We study how teacher-derived local evidence can be translated into token credit consistent with the verifier outcome. Hindsight credit assignment (HCA) relates action credit to changes in action probability after conditioning on later information (Harutyunyan et al., 2019; Meulemans et al., 2023). The implied success conditioning policy is the χ^2 metric solution based on TRPO structure(Schulman et al., 2015; Russo, 2026). However, the exact Hindsight credit assignment is hard to reach, for which teacher–student likelihood ratios can serve as surrogates. OPPO recursively constructs a value path from such ratios and anchors token-update directions to the trajectory outcome (Li et al., 2026b). HCAPO similarly uses success-conditioned prompting to estimate normalized hindsight ratios for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

step-level credit in long-horizon agents (Tan et al., 2026). Together, these methods motivate treating teacher likelihoods as noisy surrogates for success-conditioned evidence in RLVR.

We formulate this problem through the Bellman endpoint identity, whose telescoping form is also used in Bellman Policy Optimization (BPO) (Song et al., 2026). The verifier outcome and the prompt's initial value fix a *global credit budget*; intermediate value changes allocate that budget across tokens. Uniform credit (like GRPO) is the minimum-ℓ_2 allocation under this constraint, but it contains no information about differences between tokens. Exact success-conditioned evidence from HCA recovers the Bellman value path through Bayes' rule with binary rewards. Surrogate evidence therefore calls for reconciliation with the observed endpoint, allowing local information to shape credit within a fixed budget. We introduce *Endpoint-Locked Value-Path Evidence Reconciliation* (ELVER) to reconcile noisy teacher evidence with the verifier outcome through value-path reconstruction. Our contributions are threefold:

- We formulate token credit assignment in binary-reward RLVR as value-path reconstruction under a global endpoint budget. This formulation connects success-conditioned evidence to Bellman values and motivates requirements of endpoint consistency, uniform credit under neutral evidence and probability bounds.
- We develop ELVER-Affine, a closed-form projection of raw credits onto the budget hyperplane, and ELVER-Box, a constrained reconstruction that minimizes squared distances to evidence-induced transition lines with fixed endpoints.
- Experiments on mathematical reasoning results show that ELVER-Box consistently surpasses the compared baselines. Supporting reconciling local evidence with the endpoint budget and probability bounds.

2 Related Work

GRPO and token-level credit assignment. GRPO broadcasts a response-level advantage to all tokens, leaving sparse verifier feedback insufficient to distinguish individual contributions (Shao et al., 2024). DAPO and Dr. GRPO refine optimization and normalization (Yu et al., 2025; Liu et al., 2025), while finer-grained methods use prefix resampling, segment-level evaluation, or learned process rewards (Kazemnejad et al., 2025; Guo et al., 2025b; Cui et al., 2026). Token selection and reweighting also exploit entropy or saliency (Wang et al., 2025; Pala et al., 2026). The challenge is to turn outcome supervision into informative local credit.

On-policy distillation and its integration with GRPO. OPD supplies dense teacher feedback on student-generated prefixes (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025), and OPSD/SDPO use privileged-context self-teachers (Zhao et al., 2026; Hübotter et al., 2026). SDPG combines outcome optimization with an auxiliary distillation loss and derives a local policy-gradient interpretation (Liu et al., 2026). RLSD scales response-level advantages by positive weights (Yang et al., 2026a), while OPDVR gates out token signals that oppose the verifier outcome (Lin et al., 2026). CSCR normalizes sensitivity-based weights to preserve both the GRPO credit budget and the sign of the response-level advantage (He et al., 2026a). However, a correct response may contain local mistakes, while an incorrect response may still contain useful reasoning steps. These sign constraints prevent RLSD, OPDVR, and CSCR from assigning both positive and negative token credits within a single response.

Hindsight credit assignment. HCA relates action credit to future-conditioned distributions (Harutyunyan et al., 2019; Meulemans et al., 2023). Success conditioning admits a first-order policy-improvement interpretation (Russo, 2026), while HAPO uses entropy upper bounds on hindsight information to guide sign-preserving reweighting (He et al., 2026b). Exact hindsight distributions are difficult to obtain for LLMs. Practical methods approximate hindsight guidance through outcome or peer conditioning (Tan et al., 2026; Li et al., 2026a), or recursively accumulate likelihood evidence (Li et al., 2026b; Wang et al., 2026). Our method reconstructs a value path from surrogate teacher evidence, with the initial value estimate and verifier outcome fixing the total credit budget. Token credits, which are consecutive value differences, allow their signs to vary within a response.

3 PRELIMINARIES

3.1 REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS (RLVR)

We model autoregressive generation as a Markov decision process $(\mathcal{S}, \mathcal{V}, \gamma, r)$, with prefix states $\mathcal{S}$, vocabulary actions $\mathcal{V}$, discounting factor γ and reward r (Sutton & Barto, 2018; Agarwal et al., 2026). For a prompt $x \sim \mathcal{D}$, the policy $\pi_\theta(\cdot \mid s_t)$ samples token $y_t \in \mathcal{V}$ at state $s_t := (x, y_{<t})$, where $s_1 = (x)$. Appending this token gives the deterministic transition $s_{t+1} = (x, y_{\le t})$; generation ends after T tokens with trajectory $\tau = (x, y_{1:T})$. We consider undiscounted terminal rewards: $\gamma = 1$, $r_t = r(s_t, y_t) = 0$ for $t < T$, and $r_T = R(\tau)$.

For a fixed policy, the Bellman values along the sampled trajectory are

$$V_t^{\pi_\theta} := V^{\pi_\theta}(s_t) = \mathbb{E}_{\pi_\theta}[R \mid s_t], \qquad Q_t^{\pi_\theta} := Q^{\pi_\theta}(s_t, y_t) = \mathbb{E}_{\pi_\theta}[R \mid s_t, y_t]. \tag{1}$$

Straightforwardly, the boundary $V_{T+1}^{\pi_\theta} = R(\tau)$. Deterministic transitions then give $Q_t^{\pi_\theta} = V_{t+1}^{\pi_\theta}$, so the token advantage $A_t^{\pi_\theta} := A^{\pi_\theta}(s_t, y_t) = Q_t^{\pi_\theta} - V_t^{\pi_\theta}$ is the value increment. Maximizing $J(\theta) := \mathbb{E}_{x,\tau}[R(\tau)]$ therefore uses the policy gradient (omitting the importance sampling and KL term)

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, \tau\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{T} \mathrm{sg}(A_t^{\pi_\theta}) \nabla_\theta \log \pi_\theta(y_t \mid s_t) \right]. \tag{2}$$

3.2 GLOBAL CREDIT BUDGET IN RLVR

The value increments telescope to a *global credit budget* fixed by the outcome and prompt value:

$$A_t^{\pi_\theta} = V_{t+1}^{\pi_\theta} - V_t^{\pi_\theta}, \qquad t = 1, \dots, T, \tag{3}$$

$$\sum_{t=1}^{T} A_t^{\pi_\theta} = R(\tau) - V_1^{\pi_\theta} := B^{\pi_\theta}(\tau). \tag{4}$$

The intermediate values determine how this budget is distributed across tokens. Without local evidence, the minimum-ℓ_2-norm allocation is uniform:

$$\left(A_t^{\pi_\theta,\mathrm{unif}}\right)_{t=1}^{T} := \operatorname*{arg\,min}_{\substack{\boldsymbol{a}\in\mathbb{R}^T \\ \sum_{t=1}^{T} a_t = B^{\pi_\theta}(\tau)}} \sum_{t=1}^{T} a_t^2 = \frac{B^{\pi_\theta}(\tau)}{T} \mathbf{1}_T. \tag{5}$$

Connection to GRPO. For each prompt, GRPO samples $G > 1$ responses with rewards R_i and lengths T_i (Shao et al., 2024). Replacing $V_1^{\pi_\theta}$ by the group mean μ_R in equation 5 yields empirical uniform credit $(R_i - \mu_R)/T_i$. With response-length averaging absorbed into the token advantage, GRPO rescales this credit by $1/(\sigma_R + \varepsilon)$, where σ_R is the group standard deviation and $\varepsilon > 0$:

$$\widehat{A}_{i,t}^{\pi_\theta,\mathrm{GRPO}} := \frac{R_i - \mu_R}{T_i(\sigma_R + \varepsilon)}, \qquad t = 1, \dots, T_i, \tag{6}$$

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{T_i} \mathrm{sg}(\widehat{A}_{i,t}^{\pi_\theta,\mathrm{GRPO}}) \nabla_\theta \log \pi_\theta(y_{i,t} \mid s_{i,t}). \tag{7}$$

4 VALUE-PATH RECONSTRUCTION FOR TOKEN CREDIT ASSIGNMENT

When the reward is binary ($R \in \{0, 1\}$), the Bellman values become success probabilities:

$$V_t^{\pi_\theta} = \mathbb{E}_{\pi_\theta}[R \mid s_t] = 1 \cdot \pi_\theta(R = 1 \mid s_t) + 0 \cdot \pi_\theta(R = 0 \mid s_t) = \pi_\theta(R = 1 \mid s_t),$$

so the value path lives in the probability interval $[0, 1]$. Motivated by Hindsight Credit Assignment conditioning on the terminal verifier, we show that the exact Bellman value path can be reconstructed by exact success-conditioned evidence. However, the surrogate evidence motivated by distillation may contain bias and noise. We propose some guidance to target the endpoint budget via local evidence.

4.1 SUCCESS-CONDITIONED LOCAL EVIDENCE

Let $p_t(y) = \pi_\theta(y \mid s_t)$ and $q_{+,t}(y) = \pi_\theta(y_t = y \mid s_t, R = 1)$, where $q_{+,t}(y)$ is the hindsight probability conditioned on the future success. Their likelihood ratio links hindsight credit assignment to Bellman values (Harutyunyan et al., 2019; Russo, 2026).

Proposition 1 (Exact evidence reconstructs the value path)**.** *For $V_t^{\pi_\theta} > 0$ and $p_t(y) > 0$, the local hindsight evidence satisfies*

$$\rho_t^*(y) := \frac{q_{+,t}(y)}{p_t(y)} = \frac{Q^{\pi_\theta}(s_t, y)}{V_t^{\pi_\theta}}, \qquad A^{\pi_\theta}(s_t, y) = V_t^{\pi_\theta}\big(\rho_t^*(y) - 1\big). \tag{8}$$

Along a sampled transition, $V_{t+1}^{\pi_\theta} = V_t^{\pi_\theta}\rho_t^*(y_t)$, recovering the value path and the budget in equation 4.

Proof. By Bayes' rule, the success-conditioned token probability (hindsight) decomposes as

$$q_{+,t}(y) = \pi_\theta(y_t = y \mid s_t, R = 1) = \frac{\pi_\theta(R = 1 \mid s_t, y_t = y)\,\pi_\theta(y_t = y \mid s_t)}{\pi_\theta(R = 1 \mid s_t)} = p_t(y)\,\frac{Q^{\pi_\theta}(s_t, y)}{V_t^{\pi_\theta}},$$

since $\pi_\theta(R = 1 \mid s_t, y_t = y) = Q^{\pi_\theta}(s_t, y)$ and $\pi_\theta(R = 1 \mid s_t) = V_t^{\pi_\theta}$. Dividing by $p_t(y)$ yields $\rho_t^*(y) = Q^{\pi_\theta}(s_t, y)/V_t^{\pi_\theta}$, and $A^{\pi_\theta}(s_t, y) = Q^{\pi_\theta}(s_t, y) - V_t^{\pi_\theta} = V_t^{\pi_\theta}(\rho_t^*(y) - 1)$. The deterministic prefix transition gives $Q^{\pi_\theta}(s_t, y_t) = V_{t+1}^{\pi_\theta}$, so along a sampled transition $V_{t+1}^{\pi_\theta} = V_t^{\pi_\theta}\rho_t^*(y_t)$; telescoping under $V_{T+1}^{\pi_\theta} = R$ recovers the budget in equation 4. □

4.2 SURROGATE EVIDENCE AND RAW RECONSTRUCTION

Estimating the success-conditioned distribution requires additional rollouts or an auxiliary posterior model (He et al., 2026b). Motivated by on policy (self-)distillation, the teacher policy $\pi'(y \mid s_t)$ can serve as a surrogate for the hindsight probability $q_{+,t}(y)$, implicitly relying on its ability to solve the problem successfully (Lu & Thinking Machines Lab, 2025; Hübotter et al., 2026).

$$\widehat{q}_t := \pi'(y_t \mid s_t), \quad \widehat{\rho}_t := \frac{\widehat{q}_t}{\pi_\theta(y_t \mid s_t)}, \quad \widehat{r}_t := \mathrm{clip}(\log \widehat{\rho}_t, C_{\min}, C_{\max}), \tag{9}$$

where $C_{\min} \le 0 \le C_{\max}$ are fixed finite clipping bounds. For fixed-teacher distillation $\pi' = \pi_{\theta'}$ with the teacher parameter θ'. For self-distillation $\pi' = \pi_\theta(\cdot \mid \cdot, c)$ with privileged information c injected.

Raw Value Recursion. Raw value recursion directly starts from the group mean $\widehat{V}_1^{\pi_\theta,\mathrm{Raw}} = \mu_R$. For a positive initial estimate, its raw value path and token credits follow

$$\log \widehat{V}_{t+1}^{\pi_\theta,\mathrm{Raw}} = \log \widehat{V}_t^{\pi_\theta,\mathrm{Raw}} + \widehat{r}_t, \qquad \widehat{A}_t^{\pi_\theta,\mathrm{Raw}} = \widehat{V}_{t+1}^{\pi_\theta,\mathrm{Raw}} - \widehat{V}_t^{\pi_\theta,\mathrm{Raw}}. \tag{10}$$

Exact evidence recovers the true advantages when the initial value is also exact. The teacher model may provide a biased estimate of the success-conditioned distribution due to its preference and style (Yu et al., 2026). Consequently, the raw value recursion may miss the verifier endpoint or leave the probability interval $[0, 1]$.

Design requirements for value-path recursion. We therefore seek an algorithm that reconciles local teacher evidence with the credit budget fixed by the endpoints. When the evidence vanishes, it provides no basis for favoring one token over another, motivating uniform allocation. We additionally consider probability bounds to respect the range of Bellman values under binary rewards. These considerations lead to two core requirements and an optional design goal:

I. **Endpoint consistency.** The reconstructed path $\{v_i\}_{i=1}^{T+1}$ must satisfy $v_1 = \widehat{V}_1^{\pi_\theta}$ and $v_{T+1} = R$. Its token credits $A_t = v_{t+1} - v_t$ then satisfy $\sum_{t=1}^{T} A_t = R - \widehat{V}_1^{\pi_\theta}$. (Goal I)

II. **Uniform credit under neutral evidence.** For fixed T and endpoints, the token credits must satisfy $A_t \to (R - \widehat{V}_1^{\pi_\theta})/T$ as $\|\widehat{r}\|_2 \to 0$. (Goal II)

III. **Probability bounds (optional).** Keeping internal values in $[0, 1]$ preserves their interpretation as success probabilities, i.e. $v_t \in [0, 1]$. (Goal III)

4.3 ENDPOINT-LOCKED VALUE-PATH EVIDENCE RECONCILIATION

Guided by design requirements, we introduce *Endpoint-Locked Value-Path Evidence Reconciliation* (ELVER) to reconstruct token credits from local evidence.

ELVER-Affine. The simplest way is to project the Raw credits onto the budget hyperplane in closed form, which adds an equal correction for each token:

$$\widehat{A}_t^{\pi_\theta,\mathrm{Aff}} = \frac{R - \widehat{V}_1^{\pi_\theta}}{T} + \left(\widehat{A}_t^{\pi_\theta,\mathrm{Raw}} - \frac{1}{T}\sum_{j=1}^{T} \widehat{A}_j^{\pi_\theta,\mathrm{Raw}} \right). \tag{11}$$

The uniform term enforces the budget, while the centered term preserves all tokenwise differences between Raw credits.

ELVER-Box. ELVER-Affine enforces the endpoint budget while preserving all pairwise differences between Raw credits, but its reconstructed value path may leave the probability interval [0,1]. ELVER-Box jointly reconstructs the internal values within this interval, with both endpoints fixed. To reconcile the path with local evidence, we minimize squared orthogonal distances to the evidence-induced transition lines. This geometric normalization prevents large likelihood ratios from disproportionately weighting the residuals.

The transition line $v_{t+1} = e^{\widehat{r}_t} v_t$ has unit normal $(-b_t, a_t)^\top$ and orthogonal residual $\bar{e}_t(v)$, where

$$a_t := \frac{1}{\sqrt{1+e^{2\widehat{r}_t}}}, \qquad b_t := \frac{e^{\widehat{r}_t}}{\sqrt{1+e^{2\widehat{r}_t}}}, \tag{12}$$

$$\bar{e}_t(v) := a_t v_{t+1} - b_t v_t = \frac{v_{t+1} - e^{\widehat{r}_t} v_t}{\sqrt{1+e^{2\widehat{r}_t}}}. \tag{13}$$

Normalization weights transitions equally by distance, while the unnormalized squared residuals $(v_{t+1} - e^{\widehat{r}_t} v_t)^2$ would overweight large likelihood ratios by a factor of $1 + e^{2\widehat{r}_t}$. For observed evidence, define the box optimization problem

$$v^\star = \operatorname*{arg\,min}_{\substack{v_{2:T}\in[0,1]^{T-1} \\ v_1=\widehat{V}_1^{\pi_\theta},\ v_{T+1}=R}} \Phi(v), \qquad \Phi(v) := \frac{1}{2}\sum_{t=1}^{T} \bar{e}_t(v)^2. \tag{14}$$

The reconstructed credits are $A_t^{\pi_\theta,\mathrm{Box}} = v_{t+1}^\star - v_t^\star$ and sum exactly to $R - \widehat{V}_1^{\pi_\theta}$. We show that the induced Hessian is positive definite and tridiagonal, implying the $O(T)$ cost solving second-order optimality, and the spectrum of which is bounded.

Proposition 2 (Well-posedness). *For $T \geq 2$, finite $\widehat{r} \in \mathbb{R}^T$, $\widehat{V}_1^{\pi_\theta} \in [0,1]$, and $R \in \{0,1\}$, the box optimization in equation 14 has a unique minimizer. The Hessian H is tridiagonal, with entries*

$$\begin{aligned} H_{ii} &= a_i^2 + b_{i+1}^2 \quad (1 \leq i \leq T-1), \\ H_{i,i+1} &= H_{i+1,i} = -a_{i+1} b_{i+1} \quad (1 \leq i < T-1), \end{aligned} \tag{15}$$

and its spectrum satisfies

$$0 < \lambda_\star := \lambda_{\min}(H) \leq \lambda_{\max}(H) \leq 2. \tag{16}$$

The proof and detailed discussion of $\lambda_\star$ are provided in Appendix A.2.

Numerical reconstruction. Let $n = T - 1$ and $u = (v_2, \ldots, v_T)^\top$ collect the internal values. Eliminating the endpoints gives $g(u) = \nabla_u \Phi(u) = Hu - d$, where

$$d = [a_1 b_1 \widehat{V}_1^{\pi_\theta}, 0, \ldots, 0, a_T b_T R]^\top \in \mathbb{R}^{T-1}. \tag{17}$$

Let Π be a clip mapping into $\mathcal{B} = [0,1]^n$ and define $G(u) = u - \Pi(u - g(u))$, which vanishes exactly at the optimum. Algorithm 1 uses Newton steps on free coordinates (in $\mathcal{F}$) and diagonal steps near bounds where the negative gradient points outside the box (Bertsekas, 1982) and uses an Armijo methods for efficient line searching.

The complete error analysis between the true advantage and the numerical solutions at iteration k is provided. For simplicity, we use the projected-gradient iteration.

Algorithm 1 ELVER-Box with projected Newton steps

Require: $\widehat{r}_{1:T}, \widehat{V}_1^{\pi_\theta}, R$, tolerance $\epsilon > 0$, iteration threshold $K_{\max}$, line-search threshold $J_{\max}$
Require: active-set thickness $\bar{\delta} \in (0, \frac{1}{2})$, backtracking factor $\beta \in (0,1)$, Armijo constant $\sigma \in (0, \frac{1}{2})$
Get H, d from equation 15 and equation 17; solve $Hu = d$.
if $u \in \mathcal{B}$ **then go to line 14**.
Initialize $u_i = \widehat{V}_1^{\pi_\theta} + (i/T)(R - \widehat{V}_1^{\pi_\theta})$, $1 \le i < T$; $k \leftarrow 0$.
while $k < K_{\max}$ **and** $\|G(u)\|_2 > \epsilon$ **do**
    $g \leftarrow Hu - d$; $\delta \leftarrow \min\{\bar{\delta}, \|G(u)\|_2\}$.
    $\mathcal{A} \leftarrow \{i : u_i \le \delta,\ g_i > 0\} \cup \{i : u_i \ge 1 - \delta,\ g_i < 0\}$; $\mathcal{F} \leftarrow \{1, \dots, n\} \setminus \mathcal{A}$.
    Solve $H_{\mathcal{F}\mathcal{F}} p_{\mathcal{F}} = g_{\mathcal{F}}$; set $p_i = g_i / H_{ii}$ for $i \in \mathcal{A}$.
    $\alpha \leftarrow 1$; $z \leftarrow \Pi(u - \alpha p)$; $j \leftarrow 0$.
    while $j < J_{\max}$ **and** $\Phi(u) - \Phi(z) < \sigma[\alpha g_{\mathcal{F}}^\top p_{\mathcal{F}} + g_{\mathcal{A}}^\top (u_{\mathcal{A}} - z_{\mathcal{A}})]$ **do**
        $\alpha \leftarrow \beta\alpha$; $z \leftarrow \Pi(u - \alpha p)$; $j \leftarrow j + 1$.
    end while
    $u \leftarrow z$; $k \leftarrow k + 1$.
end while
return Advantages $\boldsymbol{A}^{\pi_\theta, \mathrm{Box}}$ which are consecutive differences of $(\widehat{V}_1^{\pi_\theta}, u^\top, R)^\top$.

Theorem 3 (Finite-iteration reconstruction error)**.** *Under the assumptions of Proposition 2, let $u^\star$ be the Box minimizer with evidence $\widehat{r}$ and endpoints $(\widehat{V}_1^{\pi_\theta}, R)$. Consider the projected-gradient iteration $u^{k+1} = \Pi_{\mathcal{B}}(u^k - \frac{1}{2} g(u^k))$ from $u^0 \in \mathcal{B}$. The consecutive differences $\boldsymbol{A}^{(k)}$ of $(\widehat{V}_1^{\pi_\theta}, (u^k)^\top, R)^\top$ satisfy, for every $k \ge 0$,*

$$\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{\pi_\theta}\|_2 \le 2(1 - \frac{\lambda_\star}{2})^k \|u^0 - u^\star\|_2 + \sqrt{1 + \frac{4b_1^2}{\lambda_\star}}\, |\widehat{V}_1^{\pi_\theta} - V_1^{\pi_\theta}| + \frac{2}{\sqrt{\lambda_\star}} \|\bar{e}(\boldsymbol{V}^{\pi_\theta})\|_2.$$

Here $\boldsymbol{V}^{\pi_\theta} = (V_1^{\pi_\theta}, \dots, V_{T+1}^{\pi_\theta})^\top$ is the true Bellman path and $\boldsymbol{A}^{\pi_\theta}$ its exact credits. The residual vector $\bar{e}(v) = (\bar{e}_t(v))_{t=1}^T$ uses $\widehat{r}$ in equation 13.

The three terms measure numerical optimization error, initial-value error, and evidence mismatch, respectively. The proof is provided in Appendices A.4 and A.5.

4.4 SUMMARY: FOUR CREDIT ESTIMATORS

The three reconstruction methods above, together with the uniform credit baseline of Section 3.2, span four ways to turn a verifier outcome and teacher evidence into per-token advantages. Table 1 collects their theoretical value paths and token credits. With $\widehat{V}_1^{\pi_\theta} = \mu_R$, GRPO can be viewed as Uniform credit rescaled by $1/(\sigma_R + \varepsilon)$. The response-length factor $1/T_i$ is already included in the credit, as in equation 6. Figure 1 compares the four value paths for a shared initial value and terminal target.

Table 1: Values and advantages of four credit estimators from the verifier outcome R, the estimated initial value $\widehat{V}_1^{\pi_\theta}$ and evidence $\widehat{r}_t$.

Method	Value path $\widehat{V}_{t+1}$	Token credit
Uniform	$\widehat{V}_1^{\pi_\theta} + \frac{t}{T}(R - \widehat{V}_1^{\pi_\theta})$	$\frac{R - \widehat{V}_1^{\pi_\theta}}{T}$
Raw Value Recursion	$\widehat{V}_1^{\pi_\theta} \exp\big(\sum_{j \le t} \widehat{r}_j\big)$	$\widehat{V}_{t+1}^{\pi_\theta,\mathrm{Raw}} - \widehat{V}_t^{\pi_\theta,\mathrm{Raw}}$
ELVER-Affine	$\widehat{V}_1^{\pi_\theta} + \sum_{j \le t} \widehat{A}_j^{\pi_\theta,\mathrm{Aff}}$	$\frac{R - \widehat{V}_1^{\pi_\theta}}{T} + \big(\widehat{A}_t^{\pi_\theta,\mathrm{Raw}} - \frac{1}{T}\sum_j \widehat{A}_j^{\pi_\theta,\mathrm{Raw}}\big)$
ELVER-Box	$v_{t+1}^\star$ from Alg. 1	$v_{t+1}^\star - v_t^\star$

Uniform Credit Allocation. The uniform credit allocation method (like GRPO) is the minimum ℓ_2 solution satisfying the endpoint budget. Trivially it satisfies Goals I–III. In Figure 1, we could see that it is the straight line connecting the start and end.

Raw Value Recursion. Directly summing up the log ratio induced by teacher-student ratio may lead to drift from the endpoint budget because of biased and noisy observation. So it may not guarantee any goals. Figure 1 directly shows the drift from the endpoint.

ELVER-Affine. ELVER-Affine repairs the endpoint in closed form by adding a common correction $(R - \widehat{V}_{T+1}^{\pi_\theta,\mathrm{Raw}})/T$ to every Raw credit; this enforces the budget while preserving all pairwise Raw differences, yet intermediate values may lie outside of $[0, 1]$. It guarantees Goal I and Goal II since the Raw credits converge to zero as $\|\widehat{r}\|_2 \to 0$. Goal III may not be satisfied, as showed for the Affine method in Figure 1.

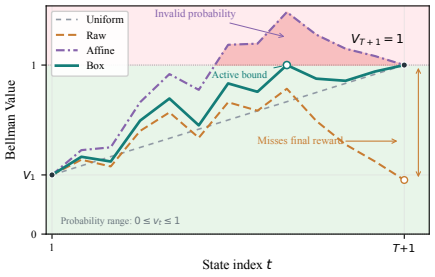


Figure 1: Four value paths with a shared initial value and terminal target.

ELVER-Box. ELVER-Box solves the box-constrained optimization problem in equation 14: it fixes both endpoints and keeps every internal value in $[0, 1]$, satisfying Goals I and III trivially. It reduces to uniform credit as the evidence vanishes ($\|\widehat{r}\|_2 \to 0$) by Hölder inequality:

$$\sum_{t=1}^{T} A_t^2 \geq \frac{1}{T}\left(\sum_{t=1}^{T} A_t\right)^2 = \frac{(R - \widehat{V}_1^{\pi_\theta})^2}{T}.$$

Equality holds when $A_1 = ... = A_T = (R - \widehat{V}_1^{\pi_\theta})/T$, i.e. uniform credit allocation. Therefore, Goal II is satisfied. In Figure 1 we can see that the value path from the Box method stays within the probability bounds with the endpoints fixed.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and data. We evaluate ELVER under three settings for constructing surrogate hindsight evidence:

- We train DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025a) on DAPO-Math-17K (Yu et al., 2025), using JustRL-DeepSeek-1.5B (He et al., 2025a) as a fixed teacher. The models inherit chain-of-thought reasoning from DeepSeek-R1.
- We train Qwen3-4B in non-thinking mode on DeepMath-103K (He et al., 2025b), choosing problems with difficulty at least 6. The fixed teacher is the non-thinking Qwen3-4B math checkpoint obtained after 500 GRPO steps by Yang et al. (2026b).
- We adopt the privileged-context self-distillation setup of RLCSD (Pan et al., 2026), training Qwen3-1.7B on the same DeepMath subset with reference solutions and answers supplied to the teacher. Both the student and self-teacher use thinking mode. The self-teacher shares the current student parameters at each evidence generation.

Evaluation. We report mean@32 accuracy on five mathematical reasoning benchmarks: AMC 2023, AIME 2024, AIME 2025, AIME 2026 and HMMT February 2026.

Comparison and configuration. Across all experiments we set $C_{\min} = -10$, $C_{\max} = 10$. Detailed training and evaluation hyperparameters are provided in Appendix B.

For on policy distillation, we compare results across GRPO (Shao et al., 2024), Raw Value Recursion, ELVER-Affine, ELVER-Box, and sampled-token OPD (Lu & Thinking Machines Lab, 2025). We set the training response length to 8192, while evaluate at 38K.

For self-distillation, we choose GRPO, sampled-token OPSD (Zhao et al., 2026), RLCSD and ELVER-Box. The form of privileged information is shown in Appendix B. The training response length is set 16384 and we evaluate at 38K.

5.2 MAIN RESULTS

Table 2: Mathematical reasoning accuracy (mean@32, 38K response length) across three training settings. Bold marks the best training method in each column within a setting.

	AMC 2023	AIME 2024	AIME 2025	AIME 2026	HMMT Feb-2026	Avg.
DeepSeek-R1-Distill-Qwen-1.5B ← JustRL-1.5B on DAPO-Math-17K						
GRPO	65.85	18.96	21.77	21.04	11.46	27.82
OPD	79.10	34.58	34.06	34.06	17.99	39.96
Raw Value Recursion	66.94	32.08	29.79	28.75	19.03	35.32
ELVER-Affine	67.21	31.35	32.71	31.77	17.99	36.21
ELVER-Box	**80.76**	**37.19**	**35.31**	**36.15**	**19.70**	**41.82**
Qwen3-4B (non-thinking) ← Qwen3-4B-RL-Math-Step500 on DeepMath-103K (difficulty ≥ 6)						
GRPO	81.85	33.23	43.44	45.62	24.81	45.79
OPD	86.03	35.94	52.29	53.75	28.88	51.38
Raw Value Recursion	85.62	37.08	52.92	53.44	30.02	51.82
ELVER-Affine	84.71	38.22	51.45	53.64	29.83	51.57
ELVER-Box	**86.07**	**38.23**	**55.00**	**55.94**	**30.11**	**53.07**
Qwen3-1.7B (thinking) by self-distillation on DeepMath-103K (difficulty ≥ 6)						
GRPO	74.66	46.14	35.20	38.22	**25.28**	43.90
OPSD	72.74	42.29	31.56	37.70	22.35	41.33
RLCSD	75.48	46.45	36.56	34.80	24.52	43.56
ELVER-Box	**76.16**	**48.85**	**37.18**	**39.06**	25.00	**45.25**

Table 2 compares accuracy on five reasoning benchmarks across three training settings. We discuss some findings below.

Local evidence improves reasoning. Raw Value Recursion and ELVER outperform GRPO on all five benchmarks in both fixed-teacher settings, supporting teacher-derived token evidence beyond uniform credit assignment. **Constrained reconstruction improves performance.** ELVER-Box further outperforms Raw and Affine on every benchmark in these settings. Raw accumulates evidence without enforcing the verifier endpoint, while Affine corrects the total credit budget but can violate probability bounds. Box jointly fits local evidence with fixed endpoints and values in $[0, 1]$, producing token credits consistent with both the observed outcome and a valid value path. More careful consideration certainly gains more improvements. **Evidence gains vary across teacher settings.** Box improves average accuracy over GRPO by 14.00 and 7.28 percentage points with fixed teachers, compared with 1.35 points under self-distillation. The gap implies the higher quality of evidence from trained teachers. **Self-distillation benefits from verifier-guided reconstruction.** OPSD underperforms GRPO on every benchmark, whereas Box achieves higher average accuracy than GRPO. RLCSD also incorporates verifier rewards but remains below GRPO on average, suggesting that the benefit depends on how self-teacher evidence is integrated into RLVR.

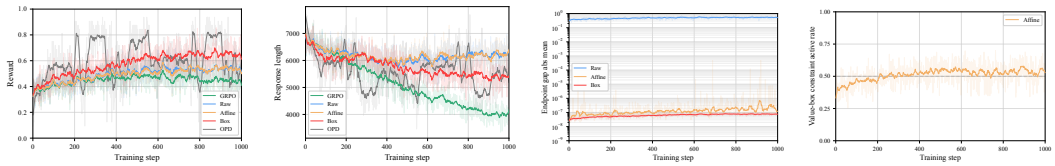
5.3 ANALYSIS

In this part, we show the training dynamics for DeepSeek-R1-Distill-Qwen-1.5B.

Under the 8,192-token response budget, Raw Value Recursion, ELVER-Affine, and ELVER-Box achieve higher mean rewards than GRPO, while producing longer responses on average (Figure 2(a,b)). ELVER-Box also achieves higher mean rewards than Raw Value Recursion and ELVER-Affine with shorter average responses. For evaluation, we use a common 38K-token response budget to reduce truncation of longer solutions.

OPD is a distillation method that matches the teacher's token distribution rather than directly optimizing verifier rewards. In our experiments, its training reward exhibits pronounced fluctuations, whereas ELVER-Box shows a smoother upward trend. By reconciling teacher-derived local evidence with the endpoint budget and probability bounds, ELVER-Box retains the comparatively steady reward dynamics seen in GRPO while attaining higher rewards.

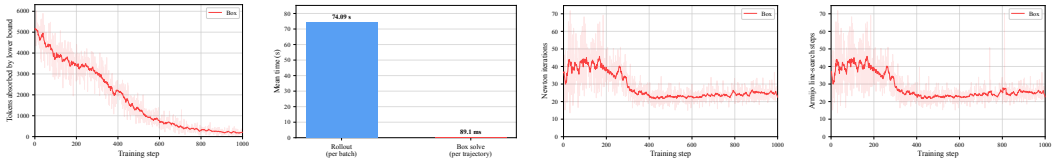
Figure 2(c) shows the mean absolute endpoint residual during training. Raw Value Recursion exhibits persistent budget drift, whereas ELVER-Affine and ELVER-Box keep the residual close to zero. Figure 2(d) further shows that roughly half of the token values of the Affine reconstructions lie outside $[0, 1]$, despite endpoint reconciliation. Together with the higher rewards achieved by ELVER-Box, these observations suggest that enforcing probability bounds can benefit RLVR.



(a) Training reward (b) Response length (c) Endpoint gap (d) Value-box active rate

Figure 2: Training dynamics of the four RLVR methods and value-recursion reconstructions. (a) Training reward. (b) Mean response length. (c) Absolute mean of Endpoint gap ($|\sum_t A_t - (R - \widehat{V}_1^{\pi_\theta})|$). (d) Ratios of tokens whose raw credit shape triggers the value-box constraint for ELVER-Affine.

Focus on ELVER-Box. Figure 3 summarizes the training dynamics and computational cost of ELVER-Box. Panel (a) shows a decrease in the mean number of internal values at the lower bound per trajectory processed by the solver, indicating fewer prefixes with near-zero reconstructed success probabilities. Panel (b) compares the mean Box solve time per trajectory (89.1 ms) with rollout generation time per batch (74.1 s). For $32 \times 8 = 256$ trajectories in one batch, the corresponding serial estimate of solving time is 22.8 s, roughly one-third of the rollout time; the solver time of one batch can be further reduced by parallel execution. Panels (c,d) show that the mean numbers of Newton solver iterations and Armijo line-search trials decrease and settle at lower levels as training progresses, indicating reduced optimization effort per trajectory.



(a) Tokens with value at lower bound (b) Solver vs. rollout time (c) Newton iterations (d) Armijo line-search trials

Figure 3: Training dynamics of ELVER-Box. (a) Mean number of internal values at the lower bound per trajectory processed by the solver. (b) Mean Box solve time per trajectory and rollout time per batch. (c) Mean outer solver iterations per trajectory. (d) Mean Armijo line-search trials per trajectory.

6 CONCLUSION

We formulated token credit assignment in RLVR with binary rewards as value-path reconstruction under a global endpoint budget. With the correct initial value, exact success-conditioned evidence recovers the Bellman path, motivating teacher–student likelihood ratios as noisy surrogates for local value changes. This formulation motivates two core design requirements, endpoint consistency and uniform credit under neutral evidence, together with the optional goal of preserving probability bounds. Guided by these principles, we introduced Endpoint-Locked Value-Path Evidence Reconciliation (ELVER) to reconcile local evidence with the verifier outcome. ELVER-Affine enforces the budget through a closed-form correction, while ELVER-Box jointly reconstructs a path within the probability bounds under budget constraints. For Box, our analysis establishes well-posedness and numerical optimization error on the recovered credits. Across mathematical reasoning benchmarks in fixed-teacher and self-teacher settings, ELVER-Box outperforms GRPO and OPD/OPSD. These findings support using teacher evidence to shape local credit allocation while the verifier outcome and probability structure constrain the entire value path.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, refining hypotheses, providing critical ingredients for proving mathematical claims, helping develop theoretical models or conceptual frameworks or for formulating mathematical claims, designing or providing feedback on research methodology or experiments, implementing methods, and interpreting results. We did not use generative AI tools to generate synthetic datasets, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis. Additionally, we used generative AI tools for creating or modifying scientific figures or images, sourcing or searching for information, editing the paper to improve readability, identifying relevant literature, and formatting references. We have reviewed all AI-assisted work. We checked LLM-generated code and proofs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing, publicly available mathematical datasets and pretrained language models. The experiments do not involve human participants. No personal data are collected as part of this study.

REPRODUCIBILITY STATEMENT

Section 4 specifies the proposed credit reconstructions, and Algorithm 1 describes the Box solver. Appendix A provides the mathematical analysis and proofs. Appendix B identifies the student and teacher models, training datasets, and evaluation benchmarks, with links to the public resources used.

REFERENCES

Alekh Agarwal, Kianté Brantley, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms, 2026. Working draft, June 27, 2026.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.

Dimitri P. Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246, 1982. doi: 10.1137/0320018.

Dimitri P. Bertsekas. *Convex Optimization Theory*. Optimization and Computation Series. Athena Scientific, Belmont, MA, 2009.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026.

A. M. Encinas and M. J. Jiménez. Explicit inverse of nonsingular Jacobi matrices. *Discrete Applied Mathematics*, 263:130–139, 2019. doi: 10.1016/j.dam.2019.03.005.

Stephen Guattery, F. Thomson Leighton, and Gary L. Miller. The path resistance method for bounding the smallest nontrivial eigenvalue of a Laplacian. *Combinatorics, Probability and Computing*, 8 (5):441–460, 1999. doi: 10.1017/S0963548399003958.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025a. doi: 10.1038/s41586-025-09422-z.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-3815.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P. van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. JustRL: Scaling a 1.5b LLM with a simple RL recipe. arXiv preprint arXiv:2512.16649, 2025a.

Qiangqiang He, Zhongheng Wu, and ZiJian Wang. Not all tokens deserve equal credit: Counterfactual sensitivity credit reallocation for long-CoT reasoning. arXiv preprint arXiv:2607.27888, 2026a.

Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. Where hindsight credit can reside: A signed-capacity view of token updates in RLVR. arXiv preprint arXiv:2604.11056, 2026b.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. arXiv preprint arXiv:2504.11456, 2025b.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. TreeRL: LLM reinforcement learning with on-policy tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12355–12369, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.604.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. doi: 10.48550/arXiv.2601.20802.

Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? rethinking process supervision from a theoretical perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27373–27398. PMLR, 2025.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 2025.

Yu Li, Shu Hong, and Tian Lan. Localizing credit at the divergence: Path-conditioned self-distillation for LLM reasoning. arXiv preprint arXiv:2606.15576, 2026a.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. OPPO: Bayesian value recursion for token-level credit assignment in LLM reasoning. arXiv preprint arXiv:2605.21851, 2026b.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward. arXiv preprint arXiv:2608.24696, 2026.

Yifeng Liu, Shiyuan Zhang, Yifan Zhang, and Quanquan Gu. Self-distilled policy gradient. arXiv preprint arXiv:2606.04036, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025.

LLM-Core Xiaomi. MiMo-V2-Flash technical report. arXiv preprint arXiv:2601.02780, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, October 2025. doi: 10.64434/tml.20251026. Published October 27, 2025.

Alexander Meulemans, Simon Schug, Seijin Kobayashi, Nathaniel D. Daw, and Gregory Wayne. Would i have gotten that reward? long-term credit assignment by counterfactual contribution analysis. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3006.

Laurent Miclo. An example of application of discrete Hardy's inequalities. *Markov Processes and Related Fields*, 5(3):319–330, 1999.

Gary L. Miller, Noel J. Walkington, and Alex L. Wang. Hardy–Muckenhoupt bounds for Laplacian eigenvalues. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*, volume 145 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 8:1–8:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.APPROX-RANDOM.2019.8.

Tej Deep Pala, Vernon Toh, and Soujanya Poria. GRAIL: Gradient-reweighted advantages for reinforcement learning with verifiable rewards. arXiv preprint arXiv:2606.04889, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation. arXiv preprint arXiv:2606.11709, 2026.

Daniel Russo. Success-conditioning as policy improvement: The optimization problem solved by imitating success. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Zhuoqing Song, Haotian Xu, Xikun Zhang, and Lidong Bing. Bellman policy optimization. arXiv preprint arXiv:2609.15987, 2026.

Gilbert Strang. *Computational Science and Engineering*. Wellesley-Cambridge Press, Wellesley, MA, 2007. ISBN 9780961408817.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. ISBN 9780262039246.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents. arXiv preprint arXiv:2603.08754, 2026.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3850.

Zi-Han Wang, Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Jie Wu, Zhengzhou Cai, Yueqing Sun, Ziang Ye, Linji Hao, Qi Gu, Xunliang Cai, Yongliang Shen, and Yujiu Yang. AgentOPSD: Recursive self-distillation for agentic reinforcement learning. arXiv preprint arXiv:2608.05987, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. arXiv preprint arXiv:2604.03128, 2026a.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. arXiv preprint arXiv:2602.12125, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3775.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. DOPD: Dual on-policy distillation. arXiv preprint arXiv:2606.30626, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. doi: 10.48550/arXiv.2601.18734.

A Additional Analysis for Elver-Box method

In this appendix, we show the detailed analysis to the Elver-Box method with well-posedness, convergence rates and error analysis.

A.1 Quadratic-program matrix form

For $T \geq 2$, let $u = (v_2, \ldots, v_T)^\top \in \mathbb{R}^{T-1}$. The residual vector in equation 13 can be rewritten as

$$\bar{e}(v) = Mu + c, \qquad \Phi(u) = \frac{1}{2}\|Mu + c\|_2^2, \qquad \nabla^2\Phi(u) = H := M^\top M. \tag{18}$$

Here $c = (-b_1\widehat{V}_1^{\pi_\theta}, 0, \ldots, 0, a_T R)^\top \in \mathbb{R}^T$ collects the contributions of the fixed endpoints $v_1 = \widehat{V}_1^{\pi_\theta}$ and $v_{T+1} = R$, and is independent of the internal values u.

The nonzero entries of $M \in \mathbb{R}^{T\times(T-1)}$ are

$$\begin{aligned} M_{1,1} &= a_1, & M_{T,T-1} &= -b_T, \\ M_{t,t-1} &= -b_t, & M_{t,t} &= a_t \qquad (2 \leq t \leq T-1). \end{aligned} \tag{19}$$

Thus H is tridiagonal, with

$$\begin{aligned} H_{j,j} &= a_j^2 + b_{j+1}^2, & 1 \leq j \leq T-1, \\ H_{j,j+1} &= H_{j+1,j} = -a_{j+1}b_{j+1}, & 1 \leq j \leq T-2. \end{aligned} \tag{20}$$

As in equation 17, we write $d = -M^\top c$ and $g(u) = Hu - d$. The two endpoint contributions to d are $a_1 b_1 \widehat{V}_1^{\pi_\theta} \boldsymbol{e}_1 + a_T b_T R \boldsymbol{e}_{T-1}$, where $\boldsymbol{e}_i$ is the ith coordinate vector. Both contributions enter the same coordinate when $T = 2$.

A.2 Well-posedness and curvature of the box program

Here we show the detailed version of well-posedness from Proposition 2, together with the evidence-dependent curvature bounds used in the convergence analysis.

Proposition 4 (Well-posedness and two-sided path-curvature bounds)**.** *For $T \geq 2$, $\widehat{V}_1^{\pi_\theta} \in [0,1]$, $R \in \{0,1\}$, and every finite evidence vector $\widehat{r} \in \mathbb{R}^T$, the matrix M has full column rank. The* Elver-*Box problem in equation 14 is a strictly convex box-constrained quadratic program with a unique global minimizer. The Hessian H obeys the two-sided spectrum bound*

$$0 < \lambda_\star := \lambda_{\min}(H) \leq \lambda_{\max}(H) \leq 2, \tag{21}$$

while the lower curvature $\lambda_\star$ is evidence-dependent. The following bounds make this dependence explicit through a pathwise curvature certificate. Define the cumulative evidence, edge weights, and edge resistances by

$$S_0 := 0, \qquad S_i := \sum_{t=1}^{i} \widehat{r}_t, \qquad w_t := a_t^2 e^{2S_t}, \qquad \varrho_t := w_t^{-1} = e^{-2S_t} + e^{-2S_{t-1}},$$

and, for $i = 1, \ldots, T-1$, let

$$\mathcal{L}_i := \sum_{t=1}^{i} \varrho_t, \qquad \mathcal{R}_i := \sum_{t=i+1}^{T} \varrho_t.$$

Then the pathwise lower curvature

$$\underline{\lambda}_T(\widehat{r}) := \left[\sum_{i=1}^{T-1} e^{2S_i} \frac{\mathcal{L}_i \mathcal{R}_i}{\mathcal{L}_i + \mathcal{R}_i}\right]^{-1} > 0 \tag{22}$$

satisfies $\underline{\lambda}_T(\widehat{r}) \le \lambda_\star$, and two useful consequences yield more interpretable uniform bounds. First, if $|\widehat{r}_t| \le C$ for every t and some $C > 0$, define

$$\Gamma_m(C) := (1 + e^{2C}) \sum_{d=0}^{m-1} e^{2Cd} = (1 + e^{2C}) \frac{e^{2Cm} - 1}{e^{2C} - 1}, \qquad m = 1, \dots, T-1.$$

Thus $\Gamma_m(C) = \Theta_C(e^{2Cm})$ for fixed $C > 0$. Then

$$\lambda_\star \ge \underline{\lambda}_T(\widehat{r}) \ge \tanh^2(C) e^{-CT}. \tag{23}$$

Moreover, this exponential dependence on the horizon cannot be improved under per-token clipping alone. For even T,

$$\tanh^2(C) e^{-CT} \le \inf_{\|\widehat{r}\|_\infty \le C} \lambda_{\min}(H) \le \frac{2e^{2C}}{1 + e^{2C}} e^{-CT}. \tag{24}$$

Thus, for every fixed $C > 0$, the worst-case lower curvature is $\Theta_C(e^{-CT})$ along even horizons. Second consider more strict cases, if the cumulative-evidence excursion

$$\Omega(\widehat{r}) := \max_{0 \le i \le T} S_i - \min_{0 \le i \le T} S_i$$

satisfies $\Omega(\widehat{r}) \le B$, then

$$\lambda_\star \ge \underline{\lambda}_T(\widehat{r}) \ge \frac{3e^{-2B}}{T^2 - 1}. \tag{25}$$

For bounds independent of T, a cumulative-evidence bound guarantees a polynomial curvature lower bound, whereas a per-token evidence bound guarantees only an exponential lower bound in the worst case.

Proof. Let $h = (h_2, \dots, h_T)^\top$ be a perturbation of the interior values, extended by the fixed-endpoint convention $h_1 = h_{T+1} = 0$. Directly from the unit-normal transition residuals,

$$h^\top H h = \|Mh\|_2^2 = \sum_{t=1}^{T} \left(a_t h_{t+1} - b_t h_t\right)^2.$$

If $Mh = 0$, its first row gives $a_1 h_2 = 0$, hence $h_2 = 0$ because $\widehat{r}_1$ is finite. The subsequent rows recursively give $h_{t+1} = e^{\widehat{r}_t} h_t = 0$, so $h = 0$. Thus M has full column rank and $H \succ 0$. The feasible set $[0, 1]^{T-1}$ is nonempty, compact, and convex, and the quadratic objective is continuous. Hence a minimizer exists by the Weierstrass theorem; since the objective is strictly convex, that minimizer is unique (Bertsekas, 2009).

For the upper-curvature side of equation 21, Cauchy–Schwarz and $a_t^2 + b_t^2 = 1$ give

$$\left(a_t h_{t+1} - b_t h_t\right)^2 \le h_t^2 + h_{t+1}^2. \tag{26}$$

Summing and using $h_1 = h_{T+1} = 0$ gives $h^\top H h \le 2\|h\|_2^2$, so $\lambda_{\max}(H) \le 2$.

We next derive the lower-curvature side of equation 21 while retaining both fixed endpoints. The weighted-Dirichlet formulation connects to path-resistance bounds for Laplacian eigenvalues (Guattery et al., 1999) and discrete Hardy inequalities for one-dimensional spectral gaps (Miclo, 1999). For the present path, we obtain the curvature certificate from the Green function with both endpoints fixed.

Write

$$h_{i+1} = e^{S_i} x_i, \qquad i = 0, \dots, T, \qquad x_0 = x_T = 0.$$

Because $b_t = e^{\widehat{r}_t} a_t$ and $S_t = S_{t-1} + \widehat{r}_t$,

$$a_t h_{t+1} - b_t h_t = a_t e^{S_t} (x_t - x_{t-1}).$$

Hence, with $W = \mathrm{diag}(e^{S_1}, \dots, e^{S_{T-1}})$,

$$h^\top H h = \sum_{t=1}^{T} w_t (x_t - x_{t-1})^2, \qquad H = W^{-1} L_w W^{-1}, \tag{27}$$

where L_w is the weighted path Laplacian with Dirichlet endpoints and edge conductances w_t. The associated Rayleigh quotient is

$$\lambda_\star = \min_{\substack{x \neq 0 \\ x_0 = x_T = 0}} \frac{\sum_{t=1}^{T} w_t (x_t - x_{t-1})^2}{\sum_{i=1}^{T-1} e^{2S_i} x_i^2},$$

which is the weighted Dirichlet eigenvalue problem with vertex masses e^{2S_i} (Miller et al., 2019, Lemma 14). Section 4 of the same work develops Hardy–Muckenhoupt estimates for paths with one fixed endpoint. Its edge resistances are

$$\varrho_t = \frac{1}{w_t} = (1 + e^{2\widehat{r}_t}) e^{-2S_t} = e^{-2S_t} + e^{-2S_{t-1}}.$$

Specializing the Green-kernel representation for tridiagonal matrices (Encinas & Jiménez, 2019) to this Dirichlet path gives

$$(L_w^{-1})_{ij} = \frac{\mathcal{L}_{\min\{i,j\}} \mathcal{R}_{\max\{i,j\}}}{\sum_{t=1}^{T} \varrho_t}, \qquad i, j = 1, \dots, T-1.$$

For fixed j, the displayed column, extended by zero at both endpoints, is harmonic away from j; its flux jump at j is $(\mathcal{L}_j + \mathcal{R}_j)/\sum_{t=1}^{T} \varrho_t = 1$, verifying the inverse formula. Since $\sum_{t=1}^{T} \varrho_t = \mathcal{L}_i + \mathcal{R}_i$ and $H^{-1} = W L_w^{-1} W$, it follows that

$$\operatorname{tr}(H^{-1}) = \sum_{i=1}^{T-1} e^{2S_i} \frac{\mathcal{L}_i \mathcal{R}_i}{\mathcal{L}_i + \mathcal{R}_i}.$$

For a positive-definite matrix, $\operatorname{tr}(H^{-1}) \geq 1/\lambda_{\min}(H)$; taking reciprocals proves the lower bound in equation 22.

For the clipped-evidence consequence, set

$$P_i := e^{2S_i} \mathcal{L}_i, \qquad Q_i := e^{2S_i} \mathcal{R}_i.$$

If $|\widehat{r}_t| \leq C$, then

$$P_i = \sum_{t=1}^{i} \left[e^{2(S_i - S_t)} + e^{2(S_i - S_{t-1})} \right] \leq \Gamma_i(C),$$

$$Q_i = \sum_{t=i+1}^{T} \left[e^{-2(S_t - S_i)} + e^{-2(S_{t-1} - S_i)} \right] \leq \Gamma_{T-i}(C).$$

Moreover,

$$e^{2S_i} \frac{\mathcal{L}_i \mathcal{R}_i}{\mathcal{L}_i + \mathcal{R}_i} = \frac{P_i Q_i}{P_i + Q_i}.$$

The map $(x, y) \mapsto xy/(x + y)$ is increasing in each positive argument and bounded by $\min\{x, y\}$. Substitution into equation 22 and summing the geometric series $\Gamma_m(C)$ gives

$$\operatorname{tr}(H^{-1}) \leq \sum_{i=1}^{T-1} \Gamma_{\min\{i, T-i\}}(C) \leq \coth^2(C) e^{CT},$$

where the last step pairs indices about the midpoint (for odd T, use $2e^C \leq 1 + e^{2C}$). Taking reciprocals proves equation 23.

It remains to verify that the exponential order is sharp. Let $T = 2m$ and choose the evidence $\widehat{r}_t = C$ for $t \leq m$ and $\widehat{r}_t = -C$ for $t > m$. Then $S_i = C \min\{i, T - i\}$. Set $h_1 = h_{T+1} = 0$ and $h_{i+1} = e^{S_i}$ for $i = 1, \dots, T-1$. Every internal residual vanishes, while the two boundary residuals both have magnitude $e^C/\sqrt{1 + e^{2C}}$. Moreover, $\|h\|_2^2 \geq e^{2S_m} = e^{CT}$. The Rayleigh quotient therefore gives

$$\lambda_{\min}(H) \leq \frac{2e^{2C}}{1 + e^{2C}} e^{-CT}.$$

If instead $\Omega(\widehat{r}) \le B$, then every difference between two cumulative evidence values has magnitude at most B, so

$$P_i \le 2ie^{2B}, \qquad Q_i \le 2(T-i)e^{2B}.$$

Therefore

$$\operatorname{tr}(H^{-1}) \le \frac{2e^{2B}}{T} \sum_{i=1}^{T-1} i(T-i) = \frac{e^{2B}(T^2-1)}{3},$$

which proves equation 25. □

A.3 Optimality conditions and projected Newton solver

In this section we give a detailed analysis of numerical Algorithm 1. Let $\mathcal{B} = [0,1]^{T-1}$ and let $\Pi_{\mathcal{B}}$ denote componentwise clipping to this box. The objective $\Phi(u) = \frac{1}{2}\|Mu + c\|_2^2$ couples only adjacent values, so the gradient at internal coordinate i (corresponding to v_{i+1}, which enters edges i and $i+1$) is

$$g_i(u) = [\nabla\Phi(u)]_i = a_i\bar{e}_i(u) - b_{i+1}\bar{e}_{i+1}(u), \qquad i = 1, \dots, T-1.$$

The box KKT conditions balance adjacent transition inconsistencies according to the local unit-normal coefficients, with the sign determined by which probability boundary (if any) is active:

$$\begin{cases} a_i\bar{e}_i - b_{i+1}\bar{e}_{i+1} = 0, & 0 < u_i^\star < 1, \\ a_i\bar{e}_i - b_{i+1}\bar{e}_{i+1} \ge 0, & u_i^\star = 0, \\ a_i\bar{e}_i - b_{i+1}\bar{e}_{i+1} \le 0, & u_i^\star = 1. \end{cases} \tag{28}$$

Strict convexity makes these conditions necessary and sufficient. Equivalently, $u^\star$ is optimal if and only if its projected-gradient residual vanishes:

$$G(u) := u - \Pi_{\mathcal{B}}(u - g(u)) = 0, \tag{29}$$

which gives the stopping criterion $\|G(u)\|_2 \le \epsilon$ in Algorithm 1.

Initialization. We first solve $Hu = d$ using a tridiagonal $LDL^\top$ factorization. If the unconstrained minimizer lies in $\mathcal{B}$, it also solves the constrained problem. Otherwise, we initialize at the uniform-credit path, $u_i^0 = \widehat{V}_1^{\pi_\theta} + (i/T)(R - \widehat{V}_1^{\pi_\theta})$ for $1 \le i < T$.

At iteration k:

1. Compute $g^k = \nabla\Phi(u^k)$ and stop when the projected gradient mapping
$$G(u^k) := u^k - \Pi_{\mathcal{B}}(u^k - g^k)$$
is below tolerance. The equality $G(u) = 0$ is equivalent to Equation 28.
2. Identify coordinates close to a bound whose negative-gradient direction points outside the box:
$$\mathcal{A}_k = \{i : u_i^k \le \delta_k,\ g_i^k > 0\} \cup \{i : u_i^k \ge 1 - \delta_k,\ g_i^k < 0\},$$
and let $\mathcal{F}_k$ be the complementary free set. We use $\delta_k = \min\{\bar{\delta}, \|G(u^k)\|_2\}$ for fixed $\bar{\delta} \in (0, \frac{1}{2})$.
3. Form the scaled Newton vector without constructing an inverse:
$$(H)_{\mathcal{F}_k\mathcal{F}_k} p^k_{\mathcal{F}_k} = g^k_{\mathcal{F}_k}, \qquad p_i^k = \frac{g_i^k}{H_{ii}}, \quad i \in \mathcal{A}_k. \tag{30}$$
Removing active rows and columns splits H into independent tridiagonal positive-definite blocks.
4. Starting from $\alpha = 1$, backtrack $\alpha \leftarrow \beta\alpha$, with $\beta \in (0,1)$. For each trial step apply the Armijo line search
$$\begin{aligned} u^k(\alpha) &:= \Pi_{\mathcal{B}}(u^k - \alpha p^k), \\ \Delta_k(\alpha) &:= \alpha \sum_{i \in \mathcal{F}_k} g_i^k p_i^k + \sum_{i \in \mathcal{A}_k} g_i^k \big(u_i^k - u_i^k(\alpha)\big). \end{aligned} \tag{31}$$
Accept the first step satisfying
$$\Phi(u^k) - \Phi\big(u^k(\alpha)\big) \ge \sigma\Delta_k(\alpha), \qquad u^{k+1} = u^k(\alpha), \qquad \sigma \in (0, 1/2). \tag{32}$$

Nearly active coordinates take diagonal steps toward their bounds. The remaining coordinates form independent contiguous blocks, each with a positive-definite tridiagonal Hessian.

A.4 Newton convergence and projected-gradient rate

We analyze Newton convergence and introduce the following projected-gradient alternative.

Projected-gradient alternative. A simpler solver uses

$$u^{k+1} = \Pi_{\mathcal{B}}\big(u^k - \tfrac{1}{2} g(u^k)\big), \tag{33}$$

where the step $\frac{1}{2}$ comes from the spectrum of H in Proposition 4. The projected gradient iteration admits the convergence considering the initial setting.

Lemma 5 (Difference-operator spectrum)**.** *For $T \geq 2$, let $h = (h_2, \ldots, h_T)^\top$ and set $h_1 = h_{T+1} = 0$. Define $D : \mathbb{R}^{T-1} \to \mathbb{R}^T$ by $(Dh)_t = h_{t+1} - h_t$ for $t = 1, \ldots, T$. Its Euclidean operator norm is $\|D\|_{\mathrm{op}} := \max_{\|h\|_2=1} \|Dh\|_2$. Then*

$$\|D\|_{\mathrm{op}} = 2\cos\frac{\pi}{2T} \leq 2. \tag{34}$$

Proof. The matrix $D^\top D$ is the one-dimensional Dirichlet difference Laplacian, with diagonal entries 2 and off-diagonal entries -1. Its standard sine eigenbasis gives eigenvalues $2 - 2\cos(k\pi/T) = 4\sin^2(k\pi/(2T))$ for $k = 1, \ldots, T-1$ (Strang, 2007, Section 3.5). Taking the square root of the largest eigenvalue yields equation 34. □

We next establish global convergence of projected Newton. Separately, we give a geometric convergence bound for projected gradient in terms of the Hessian spectrum.

Proposition 6 (Convergence of the numerical solvers)**.** ***Global convergence.*** *Fix $T \geq 2$, a finite evidence vector $\widehat{r} \in \mathbb{R}^T$, and endpoints $(\nu, R) \in [0,1]^2$. Let $u^\star$ be the unique minimizer of the Box optimization. Consider the projected Newton iteration in Appendix A.3, initialized at a feasible point and run in exact arithmetic without tolerance-based stopping or caps on the iteration and backtracking counts. Suppose backtracking continues until equation 32 holds at every iteration. Then the iterates converge to $u^\star$.*

Projected-gradient rate. *Separately, let u^k be the iterates generated by equation 33 from any feasible initial point u^0, and let $\boldsymbol{A}^{(k)}$ denote their induced credits. Then, for every $k \geq 0$,*

$$\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \nu, R)\|_2 \leq 2(1 - \lambda_\star/2)^k \|u^0 - u^\star\|_2, \tag{35}$$

where $\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \nu, R)$ denotes the credits induced by the Box minimizer with evidence $\widehat{r}$, initial value ν, and terminal value R.

Proof. After reordering the free and nearly active coordinates, the Newton direction uses the scaling

$$D_k = \mathrm{blkdiag}\big(H_{\mathcal{F}_k\mathcal{F}_k}^{-1}, \mathrm{diag}(H_{ii}^{-1})_{i \in \mathcal{A}_k}\big), \qquad \tfrac{1}{2} I \preceq D_k \preceq \lambda_\star^{-1} I.$$

These bounds follow from $\lambda_\star I \preceq H \preceq 2I$ and also hold when either index set is empty. The near-active set and mixed Armijo rule are the rectangle-constrained form of the method in Bertsekas (1982). For the iteration without stopping limits, these spectral bounds and the Lipschitz gradient ensure finite Armijo backtracking and stationarity of every accumulation point. Compactness of the box and uniqueness of the stationary point then give convergence of the whole sequence.

For projected gradient, the KKT conditions make $u^\star$ a fixed point of equation 33. Nonexpansiveness of projection gives

$$\|u^{k+1} - u^\star\|_2 \leq \|(I - H/2)(u^k - u^\star)\|_2 \leq (1 - \lambda_\star/2)\|u^k - u^\star\|_2.$$

The advantage error is $D(u^k - u^\star)$. Applying Lemma 5 and iterating the contraction proves equation 35. □

Empirical validation of the projected-gradient rate. We illustrate the geometric rate underlying equation 35 on a controlled problem. We run the projected-gradient iteration $u^{k+1} = \Pi_{\mathcal{B}}(u^k - \frac{1}{2}g(u^k))$ on a synthetic Box program with $T = 32$, $V_1 = 0.35$, $R = 1$, and constant clipped evidence $\widehat{r}_t \equiv c$ for four values $c \in \{0.3, 0.5, 0.8, 1.2\}$. Each c fixes the Hessian curvature $\lambda_\star$ and hence the theoretical contraction rate $q = 1 - \lambda_\star/2$. For these instances, projection remains inactive and the relative value and credit errors both decay as q^k. Figure 4 compares the measured relative value error $\|u^k - u^\star\|_2/\|u^0 - u^\star\|_2$ with constant-scaled q^k reference curves; their parallel slopes on the semilogarithmic axes show agreement with the theoretical geometric rate in Proposition 6.

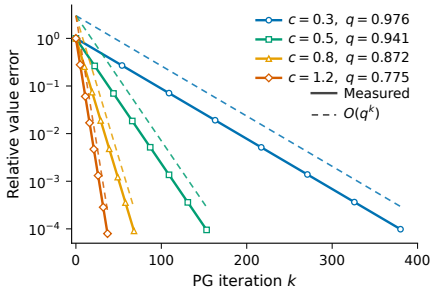


Figure 4: Projected-gradient convergence rate on a synthetic Box program ($T = 32$, $V_1 = 0.35$, $R = 1$, constant evidence $\widehat{r}_t \equiv c$). Solid lines with markers show relative internal-value error; same-colored dashed lines show $O(q^k)$, with $q = 1 - \lambda_\star/2$, as convergence-order guides. Parallel slopes indicate the same geometric rate.

A.5 Advantage reconstruction error

For the exact Box solution, reconstruction error arises from approximate evidence and the estimated initial value. We establish the exact-solution reconstruction bound used in Theorem 3.

Proposition 7 (Advantage reconstruction error)**.** *Under the assumptions of Proposition 4, let $\boldsymbol{A}^{\pi_\theta}$ be the exact Bellman advantages. Let $\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R)$ be the credits induced by the unique minimizer of equation 14 with evidence $\widehat{r}$ and endpoints $(\widehat{V}_1^{\pi_\theta}, R)$. Then*

$$\begin{aligned} &\|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta}\|_2 \\ &\quad \le \sqrt{1 + 4\frac{b_1^2}{\lambda_\star}}|\widehat{V}_1^{\pi_\theta} - V_1^{\pi_\theta}| + \frac{2}{\sqrt{\lambda_\star}}\|\bar{e}(\boldsymbol{V}^{\pi_\theta})\|_2. \end{aligned} \tag{36}$$

where $\boldsymbol{V}^{\pi_\theta} = (V_1^{\pi_\theta}, \ldots, V_{T+1}^{\pi_\theta})^\top$ is the true Bellman path. The residual vector $\bar{e}(v) = (\bar{e}_t(v))_{t=1}^T$ uses the surrogate evidence $\widehat{r}$ in equation 13.

Proof. Throughout, $\widehat{r}$ denotes the surrogate evidence supplied to the Box solver and $r^\dagger$ the exact evidence underlying the true Bellman path. We separate the initial-value and evidence errors in equation 36 by the triangle inequality.

Initial-value term. Fix the surrogate evidence $\widehat{r}$ and compare two Box solutions that share the terminal endpoint R but differ in the initial endpoint: $v = \mathsf{B}(\widehat{r}; V_1^{\pi_\theta}, R)$ with internal values u, and $\widehat{v} = \mathsf{B}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R)$ with internal values $\widehat{u}$. Write $\Phi_{\widehat{r},\nu,R}(w) = \frac{1}{2}\|Mw + c(\nu, R)\|_2^2$ to expose the initial-endpoint dependence. Only the first residual depends on V_1, so

$$c(\widehat{V}_1^{\pi_\theta}, R) - c(V_1^{\pi_\theta}, R) = -b_1\delta_0\, e_1^{(T)},$$

where $\delta_0 = \widehat{V}_1^{\pi_\theta} - V_1^{\pi_\theta}$ and $e_1^{(m)}$ is the first coordinate vector of $\mathbb{R}^m$. Since M has full column rank, $M\mathcal{B}$ is closed and convex, $P_{M\mathcal{B}}$ its projection and the two solutions satisfy $Mu = P_{M\mathcal{B}}[-c(V_1^{\pi_\theta}, R)]$ and $M\widehat{u} = P_{M\mathcal{B}}[-c(\widehat{V}_1^{\pi_\theta}, R)]$. Euclidean projection is nonexpansive, and the two targets differ by $b_1|\delta_0|$ in the first coordinate, so

$$\|M(\widehat{u} - u)\|_2 \le b_1|\delta_0|.$$

The smallest singular value of M is $\sqrt{\lambda_\star}$ (since $H = M^\top M$), hence

$$\|\widehat{u} - u\|_2 \le \frac{b_1}{\sqrt{\lambda_\star}}|\delta_0|.$$

Set $w = \widehat{u} - u$. The variational inequality shows that for any feasible solution $z \in \mathcal{B}$,

$$\langle M^\top(Mu + c(V_1^{\pi_\theta}, R)), z - u\rangle \ge 0, \tag{37}$$

$$\langle M^\top(M\widehat{u} + c(\widehat{V}_1^{\pi_\theta}, R)), z - \widehat{u}\rangle \ge 0, \tag{38}$$

Taking $z = \widehat{u}$ in equation 37 and $z = u$ in equation 38, then adding the two inequalities, gives

$$\begin{aligned}
&\langle M^\top(M(\widehat{u} - u) + (c(\widehat{V}_1^{\pi_\theta}, R) - c(V_1^{\pi_\theta}, R))), u - \widehat{u}\rangle \\
&= \langle M^\top(Mw - b_1\delta_0 e_1^{(T)}), -w\rangle \\
&= -\langle M^\top Mw, w\rangle + b_1\delta_0\langle M^\top e_1^{(T)}, w\rangle \ge 0.
\end{aligned}$$

Since $M_{1,1} = a_1$ from equation 19, then $\|Mw\|_2^2 \le a_1 b_1 \delta_0 w_1$, hence $\delta_0 w_1 \ge 0$. The total difference of value path is $(\delta_0, w^\top, 0)^\top$, implying the credit difference $Dw - \delta_0 e_1^{(T)}$, with $(Dw)_1 = w_1$, so

$$\begin{aligned}
&\|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; V_1^{\pi_\theta}, R)\|_2^2 \\
&= \|Dw - \delta_0 e_1^{(T)}\|_2^2 = \|Dw\|_2^2 + \delta_0^2 - 2\delta_0 w_1 \\
&\le 4\|w\|_2^2 + \delta_0^2 \le \left(1 + \frac{4b_1^2}{\lambda_\star}\right)\delta_0^2.
\end{aligned}$$

Here $\|D\|_{\mathrm{op}} \le 2$ follows from Lemma 5; taking square roots yields the initial-value term.

Evidence term. Fix the true endpoints $(V_1^{\pi_\theta}, R)$ and let $u^\star$ be the Box minimizer under the surrogate evidence $\widehat{r}$. The true internal path $u^\dagger = (V_2^{\pi_\theta}, \ldots, V_T^{\pi_\theta})^\top$, generated by the exact evidence $r^\dagger$, is feasible. The matrix M and vector c are those of the Box problem defined by the surrogate evidence $\widehat{r}$ and the true endpoints, so that $e^\dagger = Mu^\dagger + c = \bar{e}(\boldsymbol{V}^{\pi_\theta})$ is the residual of the true path measured under $\widehat{r}$. Set $h = u^\star - u^\dagger$. Optimality gives $\langle M^\top(Mu^\star + c), u^\dagger - u^\star\rangle \ge 0$, hence

$$\|Mh\|_2^2 \le -\langle e^\dagger, Mh\rangle \le \|e^\dagger\|_2\|Mh\|_2.$$

Using $H = M^\top M \succeq \lambda_\star I$ and $\|D\|_{\mathrm{op}} \le 2$ gives

$$\|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; V_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta}\|_2 = \|Dh\|_2 \le 2\|h\|_2 \le \frac{2}{\sqrt{\lambda_\star}}\|\bar{e}(\boldsymbol{V}^{\pi_\theta})\|_2.$$

The triangle inequality yields

$$\begin{aligned}
&\|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta}\|_2 \\
&\quad\le \|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; V_1^{\pi_\theta}, R)\|_2 \\
&\qquad + \|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; V_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta}\|_2 \\
&\quad\le \sqrt{1 + \frac{4b_1^2}{\lambda_\star}}\,|\widehat{V}_1^{\pi_\theta} - V_1^{\pi_\theta}| + \frac{2}{\sqrt{\lambda_\star}}\|\bar{e}(\boldsymbol{V}^{\pi_\theta})\|_2.
\end{aligned}$$

□

Empirical validation of the two error terms. We examine the two error sources in equation 36 separately, retaining $T = 32$, $V_1 = 0.35$, and $R = 1$ from the preceding experiment. With the initial value fixed exactly, we perturb the reference log evidence by RMS-normalized Gaussian noise of amplitude σ; the empirical log–log exponent of reconstruction error against σ is 0.90 (Figure 5, left). With the evidence fixed exactly, we vary the initial-value deviation δ_0; the credit change relative to the unperturbed Box solution has an empirical exponent of 1.00 against $|\delta_0|$ (Figure 5, right). The numerical results support the linear convergence of error analysis in Theorem 3.

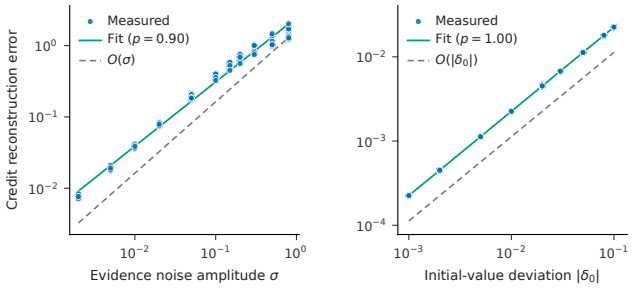


Figure 5: Error convergence to the two reconstruction-error sources ($T = 32$, $V_1 = 0.35$, $R = 1$). Left: Absolute ℓ_2 credit reconstruction error versus evidence-noise amplitude σ, with the initial value fixed exactly. Right: Absolute ℓ_2 credit change from the unperturbed Box solution versus $|\delta_0|$, with exact evidence.

Interpretation of the error terms.

Proof of Theorem 3. Let $\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R)$ denote the credits induced by $u^\star$ for the evidence $\widehat{r}$ and endpoints $(\widehat{V}_1^{\pi_\theta}, R)$. The triangle inequality gives

$$\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{\pi_\theta}\|_2 \le \|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R)\|_2 + \|\boldsymbol{A}^{\pi_\theta,\mathrm{Box}}(\widehat{r}; \widehat{V}_1^{\pi_\theta}, R) - \boldsymbol{A}^{\pi_\theta}\|_2.$$

Apply the projected-gradient bound in Proposition 6 to the first term and Proposition 7 to the second. □

The evidence term vanishes whenever the true Bellman path satisfies the supplied transition relations. With finite positive ratios and a positive initial value, a failed endpoint generally requires nonzero reconciliation residuals, as described in Appendix A.6.

A.6 Boundary and Numerical Handling

Boundary values. The implementation uses the group estimate $\widehat{V}_1^{\pi_\theta} = \mu_R$. All-failure groups have $\mu_R = R = 0$ and receive zero credits, avoiding evaluation of $\log 0$. For Box, the zero path is already the unique solution for these endpoints, so the solver is skipped.

Stable advantage evaluation. For $\widehat{V}_1^{\pi_\theta} > 0$, Raw accumulates evidence additively in log space. Let $L_1 = \log \widehat{V}_1^{\pi_\theta}$, and distinguish the uncapped cumulative log-value $\widetilde{L}_{t+1}$ from the stored value L_{t+1}:

$$\widetilde{L}_{t+1} = L_1 + \sum_{j=1}^{t} \widehat{r}_j, \qquad L_{t+1} = \min\{\widetilde{L}_{t+1}, L_{\max}\}.$$

The evidence $\widehat{r}_j$ is clipped as in equation 9; the separate cumulative-log cap $L_{\max}$ applies only when constructing the Raw path. The capped values do not feed back into the cumulative sum. Without this optional cap, $L_{t+1} = \widetilde{L}_{t+1}$. The resulting Raw credits are evaluated as

$$\widehat{A}_t^{\pi_\theta,\mathrm{Raw}} = \exp(L_t)\,\mathrm{expm1}(L_{t+1} - L_t), \tag{39}$$

where $\mathrm{expm1}(z)$ evaluates $e^z - 1$ accurately near zero. This avoids subtracting two nearly equal exponentiated values when the evidence is small. The log-value difference uses the stored values after any cumulative-log clipping; without that clipping, it equals $\widehat{r}_t$. Affine applies the correction in equation 11 directly to these stable Raw credits, with no additional clipping of the Affine path.

Box coefficients and credits. For finite $\widehat{r}_t$, evaluate the coefficients without forming $e^{2\widehat{r}_t}$:

$$a_t = \exp\left[-\tfrac{1}{2}\,\mathrm{softplus}(2\widehat{r}_t)\right], \qquad b_t = \exp\left[-\tfrac{1}{2}\,\mathrm{softplus}(-2\widehat{r}_t)\right]. \tag{40}$$

Box solves for the bounded value path and computes its consecutive differences in double precision before converting the credits to the training dtype. This computation permits exact zero boundary values and does not take logarithms of the optimized path.

B Experimental Details

B.1 Models and datasets

Models. We consider two fixed-teacher pairings and one self-distillation setting. The first pairing uses DeepSeek-R1-Distill-Qwen-1.5B[1] as the student and JustRL-DeepSeek-1.5B[2] as the teacher. The student inherits chain-of-thought reasoning from DeepSeek-R1 through distillation. The second uses Qwen3-4B in non-thinking mode as the student and a fixed non-thinking Qwen3-4B math teacher trained for 500 GRPO steps by Yang et al. (2026b)[3]. For self-distillation, we use Qwen3-1.7B in thinking mode and follow the privileged-context setup of RLCSD (Pan et al., 2026).

Training data. The DeepSeek pairing is trained on DAPO-Math-17K[4] (Yu et al., 2025). The Qwen3-4B and Qwen3-1.7B settings use DeepMath-103K[5] (He et al., 2025b), retaining problems with difficulty at least 6.

Evaluation benchmarks. We evaluate on five mathematical competition benchmarks: AMC 2023,[6] AIME 2024,[7] AIME 2025,[8] AIME 2026,[9] and HMMT February 2026.[10]. We report mean@32 accuracy on each benchmark.

B.2 Training, Evaluation and Method parameters

OPD setting. In the fixed-teacher setting, students are trained for 1,000 steps with the teacher held fixed. Table 3 lists the training and evaluation parameters for OPD and the corresponding RLVR comparisons. Raw, Affine, and Box use log-evidence bounds $C_{\min} = -10$ and $C_{\max} = 10$; OPD clips its sampled teacher–student log-ratio to the same interval. Prompts longer than 2,048 tokens are filtered before training.

Table 3: Training and evaluation parameters for the fixed-teacher (OPD) setting. The log-evidence bounds apply to the ELVER reconstructions.

Parameter	Value
Total training steps	1,000
Prompt batch size	32
Responses per prompt G	8
PPO mini-batch	32 prompt groups (256 responses)
Learning rate	10^{-6}
Warmup steps	10
Weight decay	0.01
Entropy / reference-KL coefficient	0 / 0
Maximum prompt length	2,048 tokens
Maximum training response length	8,192 tokens
Log-evidence bounds $(C_{\min}, C_{\max})$	$(-10, 10)$
Training temperature / top-p / top-k	1.0 / 1.0 / -1.0
Maximum evaluation response length	38,964 (Deepseek) / 38,900 (Qwen3-4B)
Evaluation temperature / top-p / top-k	0.6 / 0.95 / 20

[1] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
[2] https://huggingface.co/hbx/JustRL-DeepSeek-1.5B
[3] https://huggingface.co/Keven16/Qwen3-4B-Non-Thinking-RL-Math-Step500
[4] https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math-17k
[5] https://huggingface.co/datasets/zwhe99/DeepMath-103K
[6] https://huggingface.co/datasets/math-ai/amc23
[7] https://huggingface.co/datasets/math-ai/aime24
[8] https://huggingface.co/datasets/MathArena/aime_2025
[9] https://huggingface.co/datasets/MathArena/aime_2026
[10] https://huggingface.co/datasets/MathArena/hmmt_feb_2026

Self-distillation setting. Table 4 gives the parameters for the RLCSD-setting-based Qwen3-1.7B experiments with GRPO, OPSD, RLCSD, and ELVER-Box. Training updates all model parameters, and thinking mode is enabled for both training and evaluation.

Table 4: Training and evaluation parameters for the Qwen3-1.7B self-distillation setting. Method-dependent entries are identified explicitly.

Parameter	Value
Total training steps	500
Prompt batch size	32
Responses per prompt G	8
PPO mini-batch	32 prompt groups (256 responses)
Learning rate	10^{-6}
Warmup steps	10
Weight decay	0.01
Entropy coefficient	0
Reference-KL coefficient	GRPO: 10^{-3}; OPSD / RLCSD / Box: 0
Maximum prompt length	2,048 tokens
Maximum training response length	16,384 tokens
Box log-evidence bounds $(C_{\min}, C_{\max})$	$(-10, 10)$
Training temperature / top-p / top-k	1.0 / 1.0 / -1.0
Teacher sync mode	dynamic
Maximum evaluation response length	38,900
Evaluation temperature / top-p / top-k	0.6 / 0.95 / 20

Raw Value Recursion and ELVER-Affine. The cumulative log-value in the Raw reconstruction is capped at $L_{\max} = 80$ before evaluating its credits. Affine then applies the endpoint correction to these Raw credits.

ELVER-Box. Table 5 gives the shared solver settings of Algorithm 1 for all experiments.

Table 5: Box solver parameters in the notation of Algorithm 1, followed by additional implementation settings.

Parameter	Value
Stopping tolerance ϵ	10^{-8}
Maximum outer iterations $K_{\max}$	256
Line-search trial budget $J_{\max}$	24
Active-set thickness cap $\bar{\delta}$	10^{-3}
Backtracking factor β	0.5
Armijo constant σ	10^{-4}

B.3 PRIVILEGED INFORMATION IN SELF-DISTILLATION

Privileged context. For the Qwen3-1.7B self-distillation experiments, RLCSD follows the privileged-context setup in the original paper (Pan et al., 2026). OPSD and ELVER-Box use the same solution-and-answer prompt format. For each training problem x, the privileged information c consists of the dataset-provided reference solution and its ground-truth final answer. The student samples responses from the problem and the answer-format instruction. The teacher receives an augmented message containing the same problem, a delimited reference-solution block, the final answer, and an instruction to understand the reference reasoning before responding. Figure 6 gives the two templates. Evaluation uses the problem and answer-format instruction without the reference solution or answer.

Student prompt

```
Problem: [problem]

Please reason step by step, and put your final answer within
\boxed{}.
```

Teacher prompt

```
Problem: [problem]

Here is a reference solution to this problem:
=== Reference Solution Begin ===
[reference solution]

Correct final answer: [ground-truth answer]
=== Reference Solution End ===

After reading the reference solution above, make sure you understand
the reasoning behind each step.

Please reason step by step, and put your final answer within
\boxed{}.
```

Figure 6: Templates for self-distillation training, using the reference-solution format of RLCSD (Pan et al., 2026).

Teacher sync setting. RLCSD (Pan et al., 2026) uses a periodically updated snapshot teacher. To align with our theoretical formulation, we use a fully synchronized self-teacher that shares the current student parameters, with gradients disabled during teacher scoring.