Trust Region Masking for Long-Horizon LLM Reinforcement Learning

Yingru Li

December 20, 2025

Outline

I. Motivation: Off-Policy Mismatch in LLM-RL

Why $\pi_{\text{roll}} \neq \pi_\theta$ is unavoidable in modern LLM-RL (prior work)

II. Tighter Error Bounds

Classical $O(T^2)$ vs. new $O(T^{3/2})$ and $O(T)$ bounds

Key: All bounds depend on $D_{\text{KL}}^{\text{tok,max}}$—the maximum token-level divergence across all positions in a sequence (a sequence-level quantity)

III. Why Token-Level Methods Fail

Token-level methods (PPO clipping, token masking) cannot control this sequence-level quantity—they operate independently at each position

IV. Solution: Trust Region Masking

Mask entire sequences $\Rightarrow$ ensures $D_{\text{KL}}^{\text{tok,max}} \leq \delta \Rightarrow$ non-vacuous guarantees

V. Conclusion

Summary and future directions

Contributions

Motivation: Prior work shows off-policy mismatch ($\pi_{\text{roll}} \neq \pi_\theta$) is unavoidable in modern LLM-RL.

Our Contributions:

1. **Tighter Error Bounds:** Derive $O(T^{3/2})$ Pinsker-Marginal and $O(T)$ Mixed bounds, improving over classical $O(T^2)$ by $O(\sqrt{T})$ to $O(T)$ factors
2. **Key Insight:** Both bounds depend on $D_{\text{KL}}^{\text{tok,max}}$—the maximum token-level divergence across all positions in the sequence. This is inherently a *sequence-level* quantity.
3. **Failure of Token-Level Methods:** Token-independent methods (PPO clipping, token masking) cannot control this sequence-level quantity
4. **TRM Algorithm:** Mask entire sequences violating trust region $\Rightarrow$ ensures $D_{\text{KL}}^{\text{tok,max}} \leq \delta$ for accepted sequences $\Rightarrow$ first non-vacuous guarantees

Section I

Motivation: Off-Policy Mismatch in LLM-RL

Why the rollout policy π_{roll} differs from the training policy π_θ

(Background from prior work)

The Assumption: Identical parameters θ produce identical distributions:

$$\text{Logits}_{\text{inference}}(x, y_{<t}; \theta) \equiv \text{Logits}_{\text{train}}(x, y_{<t}; \theta)$$

where x is the prompt and $y_{<t}$ denotes tokens generated before position t.

The Reality: Modern LLM stacks use different implementations for inference vs. training.

Inference (vLLM/SGLang):

- PagedAttention
- FP8/INT8 KV-cache quantization
- Aggressive operator fusion

Training (Megatron/FSDP):

- FlashAttention-2
- BF16/FP32 accumulation
- Tensor parallelism

Result: Even with identical weights, logits differ systematically.

1.1.2 Floating-Point Non-Associativity

Root Cause: Floating-point arithmetic is non-associative.

$$(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$$

In Attention:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

The softmax denominator involves summing over context length. Different reduction orders yield different results.

Autoregressive Amplification:

1. Token y_1: Small logit difference δ_1
2. Token y_1 sampled and fed back
3. Token y_2: Difference compounds to $\delta_2 > \delta_1$
4. ... continues for T steps

Conclusion: The rollout distribution π_{roll} differs from the training distribution π_θ.

1.2.1 The Top-K Discontinuity

In Mixture-of-Experts models (Mixtral, DeepSeek-V2):

$$y = \sum_{i \in \mathcal{K}} g_i(x) \cdot E_i(x), \quad \mathcal{K} = \text{Top-}K(h(x))$$

The Problem: Top-K is a **discontinuous** function of router logits $h(x)$.
Combined with Precision Drift:

$$h_{\text{inf}} = h_{\text{train}} + \epsilon_{\text{drift}}$$

If $|h_{(K)} - h_{(K+1)}| < \|\epsilon_{\text{drift}}\|$, different experts are selected:

$$\mathcal{K}_{\text{train}} \neq \mathcal{K}_{\text{inf}}$$

Result: Completely different token distributions from the same weights.

1.2.2 Support Collapse

When different experts are selected, token probabilities can differ drastically.

Example:

- Rollout (Expert A): $\pi_{\text{roll}}(\text{“apple”}) = 0.9$
- Train (Expert B): $\pi_\theta(\text{“apple”}) = 0.001$

Importance Ratio (ratio of training to rollout probability):

$$\rho = \frac{\pi_\theta(y)}{\pi_{\text{roll}}(y)} = \frac{0.001}{0.9} \approx 0.001 \quad \text{or} \quad \frac{0.9}{0.001} = 900$$

This is **impulse noise**—not Gaussian, but discrete jumps that corrupt gradient estimates.
(We will formally define ρ_t in Section 2.)

1.3.1 The Staleness Gap

Large-scale LLM-RL uses decoupled architectures:

- **Actors:** Generate rollouts with $\pi_{\theta_{\textbf{old}}}$
- **Learner:** Updates to $\pi_{\theta_{\textbf{new}}}$
- **Latency:** k gradient steps between generation and consumption

$$\theta_{\text{train}} = \theta_{\text{rollout}} + \sum_{i=1}^{k} \Delta\theta_i$$

Effect: Even with identical implementations, $\pi_{\text{roll}} \neq \pi_\theta$ due to parameter drift.
Compound Effect: Staleness shifts expert routing boundaries, amplifying MoE discontinuities.

1.4 Section Summary

Prior Work Finding: In modern LLM-RL, off-policy mismatch is **systemic**, not incidental.

Source	**Mechanism**
Implementation	Different kernels, precision, reduction order
MoE Routing	Discontinuous Top-K selection
Staleness	Parameter drift in distributed training

Implication: We cannot assume $\pi_{\text{roll}} = \pi_\theta$. Theory must account for distribution mismatch.

Next: What theoretical guarantees do we need for safe optimization?

Section II

Tighter Error Bounds

Classical $O(T^2)$ vs. new $O(T^{3/2})$ and $O(T)$ bounds

2.1.1 Autoregressive Generation

Setup:

- Prompt: $x \sim P(x)$
- Response: $y = (y_1, \dots, y_T)$, each $y_t \in \mathcal{V}$ (vocabulary)
- Context at step t: $c = (x, y_{<t})$
- Policy (parameterized by θ): $\pi_\theta(y_t|x, y_{<t})$
- Terminal reward: $R(x, y) \in [0, 1]$

Two Key Policies:

- π_{roll}: **Rollout policy** — generates training data
- π_θ: **Training policy** — policy being optimized

Trajectory Distribution:

$$P^\pi(y|x) = \prod_{t=1}^{T} \pi(y_t|x, y_{<t})$$

Objective: $J(\pi_\theta) = \mathbb{E}_{x \sim P(x)} \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} [R(x, y)]$

Definition: The probability of reaching context $(x, y_{<t})$ under policy π:

$$d_t^{\pi}(x, y_{<t}) = P(x) \prod_{s=1}^{t-1} \pi(y_s | x, y_{<s})$$

Key Property: Different policies induce different context distributions:

$$\pi_\theta \neq \pi_{\text{roll}} \implies d_t^{\pi_\theta} \neq d_t^{\pi_{\text{roll}}} \quad \text{for } t \geq 2$$

Small per-token differences compound into large distributional shifts over the generation.

2.2.1 The Optimization Problem

Goal: Maximize $J(\pi_\theta) = \mathbb{E}_{\pi_\theta}[R(x, y)]$
Constraint: We only have samples from π_{roll}, not π_θ.
From Section I: $\pi_{\text{roll}} \neq \pi_\theta$ always (systemic mismatch).

Need: An objective $L(\pi_\theta)$ such that:

1. Computable from π_{roll} samples
2. Optimizing L also improves J (at least locally)

2.2.2 Why Trajectory Importance Sampling Fails

Naive Approach: Use importance sampling on the full trajectory.

$$J(\pi_\theta) = \mathbb{E}_{\pi_{\mathbf{roll}}}\left[\frac{P^{\pi_\theta}(y|x)}{P^{\pi_{\mathbf{roll}}}(y|x)} \cdot R(x,y)\right] = \mathbb{E}_{\pi_{\mathbf{roll}}}\left[\prod_{t=1}^{T} \rho_t \cdot R\right]$$

where $\rho_t = \pi_\theta(y_t|x, y_{<t})/\pi_{\text{roll}}(y_t|x, y_{<t})$.

The Problem:

$$\text{Var}\left(\prod_{t=1}^{T} \rho_t\right) = O\left(e^T\right)$$

For $T = 1000$: estimator is **useless**.

Key Question: Can we avoid the product $\prod_t \rho_t$?

2.2.3 The Policy Gradient Has Sum Structure

Key Insight: The gradient ∇J decomposes as a **sum**!
REINFORCE [Williams, 1992]:

$$\nabla J = \mathbb{E}_{\pi_\theta}\left[R \cdot \nabla \log P^{\pi_\theta}(y|x)\right] = \mathbb{E}_{\pi_\theta}\left[R \cdot \sum_{t=1}^{T} \nabla \log \pi_\theta(y_t|x, y_{<t})\right]$$

With Baseline (reduces variance, same expectation):

$$\nabla J = \mathbb{E}_{\pi_\theta}\left[A \cdot \sum_{t=1}^{T} \nabla \log \pi_\theta(y_t|x, y_{<t})\right]$$

where $A = R(x, y) - b$ is the **trajectory advantage** and b is a baseline (e.g., batch mean reward).
This is the key: Sum structure enables per-token importance sampling!

2.2.4 The Surrogate Objective

Idea: Apply IS to each term in the sum, not the whole trajectory.
The Surrogate [Kakade & Langford, 2002; Schulman et al., 2015]:

$$L_{\pi_{\textbf{roll}}}(\pi_\theta) = \mathbb{E}_{\pi_{\textbf{roll}}}\left[A \cdot \sum_{t=1}^{T} \rho_t\right]$$

Equivalently (distributing the advantage):

$$L_{\pi_{\textbf{roll}}}(\pi_\theta) = \mathbb{E}_{\pi_{\textbf{roll}}}\left[\sum_{t=1}^{T} \rho_t \cdot A\right]$$

Critical Difference:

- Trajectory IS: $\prod_t \rho_t$ — variance $O(e^T)$
- Per-token IS: $\sum_t \rho_t$ — variance $O(T)$

2.2.5 Why the Surrogate Works

Claim: At $\pi_\theta = \pi_{\text{roll}}$, we have $\nabla L = \nabla J$.
Proof:

$$\nabla_\theta L = \mathbb{E}_{\pi_{\text{roll}}}\left[A \cdot \sum_{t=1}^{T} \nabla_\theta \rho_t\right] = \mathbb{E}_{\pi_{\text{roll}}}\left[A \cdot \sum_{t=1}^{T} \rho_t \nabla_\theta \log \pi_\theta(y_t|x, y_{<t})\right]$$

At $\pi_\theta = \pi_{\text{roll}}$ (so $\rho_t = 1$):

$$\begin{aligned} \nabla_\theta L\Big|_{\pi_{\text{roll}}} &= \mathbb{E}_{\pi_{\text{roll}}}\left[A \cdot \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t|x, y_{<t})\right] \\ &= \mathbb{E}_{\pi_{\text{roll}}}\left[A \cdot \nabla_\theta \log P^{\pi_\theta}(y|x)\right] = \nabla_\theta J\Big|_{\pi_{\text{roll}}} \quad \checkmark \end{aligned}$$

Also: $L(\pi_{\text{roll}}) = \mathbb{E}[A \cdot T] = 0$ when $b = \mathbb{E}[R]$.

2.2.6 From Local to Global: The Optimization Gap

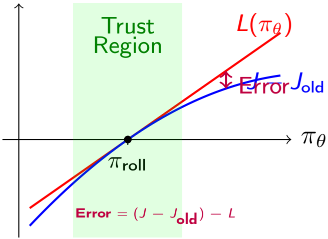
What we've shown:

- $L(\pi_{\text{roll}}) = 0$ and $\nabla L|_{\pi_{\textbf{roll}}} = \nabla J|_{\pi_{\textbf{roll}}}$
- So L and $J - J(\pi_{\text{roll}})$ are **tangent** at π_{roll}

The optimization scenario:

- We can only compute L (from π_{roll} samples)
- We maximize L, hoping this improves J
- After update: $\pi_\theta \neq \pi_{\text{roll}}$

The key question:
Does $L(\pi_\theta) > 0$ guarantee $J(\pi_\theta) > J(\pi_{\text{roll}})$?



2.2.7 Defining the Approximation Error

From the figure: The gap between L and $J - J(\pi_{\text{roll}})$ grows as π_θ moves from π_{roll}.

Define the **approximation error**:

$$\text{Error}(\pi_\theta) = J(\pi_\theta) - J(\pi_{\text{roll}}) - L(\pi_\theta)$$

Rearranging: $J(\pi_\theta) - J(\pi_{\text{roll}}) = L(\pi_\theta) + \text{Error}(\pi_\theta)$
Guarantee Condition:

- If $L(\pi_\theta) > |\text{Error}(\pi_\theta)|$, then $J(\pi_\theta) > J(\pi_{\text{roll}})$ (guaranteed improvement!)

We know: At $\pi_\theta = \pi_{\text{roll}}$: Error $= 0$; as π_θ diverges: Error grows.
Goal: Derive an exact expression for this error, then bound it.

2.3.1 The Performance Difference Identity

Theorem (PDI) [Kakade & Langford, 2002]:
Define per-step advantage: $A_t^{\pi_{\text{roll}}}(x, y_{\le t}) = Q^{\pi_{\text{roll}}}(x, y_{\le t}) - V^{\pi_{\text{roll}}}(x, y_{<t})$
where $Q^{\pi_{\text{roll}}}(x, y_{\le t}) = \mathbb{E}_{\pi_{\text{roll}}}[R|x, y_{\le t}]$ and $V^{\pi_{\text{roll}}}(x, y_{<t}) = \mathbb{E}_{\pi_{\text{roll}}}[R|x, y_{<t}]$.
Let $g_t(x, y_{<t}) = \mathbb{E}_{y_t \sim \pi_\theta}[A_t^{\pi_{\text{roll}}}(x, y_{\le t})]$.
True improvement:

$$J(\pi_\theta) - J(\pi_{\text{roll}}) = \sum_{t=1}^{T} \mathbb{E}_{d_t^{\pi_\theta}}[g_t]$$

Surrogate value:

$$L(\pi_\theta) = \sum_{t=1}^{T} \mathbb{E}_{d_t^{\pi_{\text{roll}}}}[g_t]$$

The error:

$$\boxed{\text{Error} = \sum_{t=1}^{T} \left(\mathbb{E}_{d_t^{\pi_\theta}}[g_t] - \mathbb{E}_{d_t^{\pi_{\text{roll}}}}[g_t] \right)}$$

2.3.3 Interpretation

From the PDI:

$$\text{Error} = \sum_{t=1}^{T} \underbrace{\left(\mathbb{E}_{d_t^{\pi_\theta}}[g_t] - \mathbb{E}_{d_t^{\pi_{\textbf{roll}}}}[g_t]\right)}_{\text{Expectation under wrong context distribution}}$$

Expanding the expectation: Let $c_t = (x, y_{<t})$ denote a context at step t.

$$= \sum_{t=1}^{T} \sum_{c_t} \underbrace{\left(d_t^{\pi_\theta}(c_t) - d_t^{\pi_{\textbf{roll}}}(c_t)\right)}_{\text{Context probability shift}} \cdot g_t(c_t)$$

Key Insights:

1. **At** $\pi_\theta = \pi_{\textbf{roll}}$**:** $d_t^{\pi_\theta} = d_t^{\pi_{\textbf{roll}}}$, so error $= 0$. ✓
2. **As** π_θ **diverges:** Context distributions diverge, error grows.
3. **Accumulation:** Errors compound over T steps.

See Appendix A.4–A.5 for connection to trajectory advantage $A = R - b$.

Next: How do we **bound** this error?

2.4.1 The Error Structure

From Performance Difference Identity:

$$\text{Error} = \sum_{t=1}^{T} \left(\mathbb{E}_{c_t \sim d_t^{\pi_\theta}}[g_t(c_t)] - \mathbb{E}_{c_t \sim d_t^{\pi_{\text{roll}}}}[g_t(c_t)] \right)$$

where $g_t(c_t) := \mathbb{E}_{y_t \sim \pi_\theta}[A_t(c_t, y_t)]$ is the expected advantage at context $c_t = (x, y_{<t})$.

Bounding via Total Variation:

$$\begin{aligned} |\text{Error}| &\leq \sum_{t=1}^{T} \left| \mathbb{E}_{d_t^{\pi_\theta}}[g_t] - \mathbb{E}_{d_t^{\pi_{\text{roll}}}}[g_t] \right| \\ &\leq 2 \sum_{t=1}^{T} \|g_t\|_\infty \cdot \|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{\text{TV}} \end{aligned}$$

Two Quantities to Bound:

1. $\|g_t\|_\infty = \max_{c_t} |g_t(c_t)|$: How much can expected advantage vary?
2. $\|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{\text{TV}}$: How different are the context distributions?

[Source: Performance Difference Lemma, Kakade & Langford 2002]

2.4.2 The Martingale Property

Claim: For **any** reward structure, the advantage satisfies:

$$\mathbb{E}_{y_t \sim \pi_{\textbf{roll}}(\cdot|c)}[A_t(c, y_t)] = 0 \quad \text{for all contexts } c$$

Proof: By definition of value function:

$$\begin{aligned}
\mathbb{E}_{y_t \sim \pi_{\textbf{roll}}}[A_t(c, y_t)] &= \mathbb{E}_{\pi_{\textbf{roll}}}[Q^{\pi_{\textbf{roll}}}(c, y_t) - V^{\pi_{\textbf{roll}}}(c)] \\
&= \mathbb{E}_{\pi_{\textbf{roll}}}[Q^{\pi_{\textbf{roll}}}(c, y_t)] - V^{\pi_{\textbf{roll}}}(c) \\
&= V^{\pi_{\textbf{roll}}}(c) - V^{\pi_{\textbf{roll}}}(c) = 0 \quad \checkmark
\end{aligned}$$

The last step uses $V(c) = \mathbb{E}_{y_t \sim \pi}[Q(c, y_t)]$ **by definition**.

Key Point: This is NOT an assumption — it follows from the definitions of V and Q. It holds for **all** reward structures (dense, sparse, discounted, undiscounted).

2.4.3 Bounding $|g_t(c_t)|$ via the Martingale

Step 1: Rewrite using martingale property

$$g_t(c_t) = \mathbb{E}_{y_t \sim \pi_\theta}[A_t(c_t, y_t)] - \underbrace{\mathbb{E}_{y_t \sim \pi_{\textbf{roll}}}[A_t(c_t, y_t)]}_{=0}$$

$$= \sum_{y_t} (\pi_\theta(y_t|c_t) - \pi_{\text{roll}}(y_t|c_t)) \cdot A_t(c_t, y_t)$$

Step 2: Bound via Total Variation

Define $D_{\text{TV}}^{\text{tok}}(c_t) := \frac{1}{2} \sum_y |\pi_\theta(y|c_t) - \pi_{\text{roll}}(y|c_t)|$ (token-level TV at context c_t).
For rewards $R \in [0, 1]$: $|A_t| \le 1$ (since $Q, V \in [0, 1]$).

$$|g_t(c_t)| \le \sum_{y_t} |\pi_\theta(y_t|c_t) - \pi_{\text{roll}}(y_t|c_t)| \cdot |A_t(c_t, y_t)| \le 2D_{\text{TV}}^{\text{tok}}(c_t)$$

Tightest Bound:

$$\boxed{\|g_t\|_\infty \le 2D_{\text{TV}}^{\text{tok,max}}} \quad \text{where } D_{\text{TV}}^{\text{tok,max}} := \max_{t, c_t} D_{\text{TV}}^{\text{tok}}(c_t)$$

2.5.0 Notation Summary: Divergence Measures

Context at timestep t: $c_t = (x, y_{<t})$ where $y_{<t} = (y_1, \ldots, y_{t-1})$.
Token-level divergences (at context c_t):

$$D_{\mathrm{TV}}^{\mathrm{tok}}(c_t) := \frac{1}{2} \sum_y |\pi_\theta(y|c_t) - \pi_{\mathrm{roll}}(y|c_t)|$$

$$D_{\mathrm{KL}}^{\mathrm{tok}}(c_t) := D_{\mathrm{KL}}(\pi_{\mathrm{roll}}(\cdot|c_t) \| \pi_\theta(\cdot|c_t)) = \sum_y \pi_{\mathrm{roll}}(y|c_t) \log \frac{\pi_{\mathrm{roll}}(y|c_t)}{\pi_\theta(y|c_t)}$$

Following TRPO, we use $D_{\mathrm{KL}}(\pi_{\mathrm{roll}} \| \pi_\theta)$ — computable exactly from stored logits.
Maximum token-level divergences (over all timesteps and contexts):

$$D_{\mathrm{TV}}^{\mathrm{tok,max}} := \max_{t, c_t} D_{\mathrm{TV}}^{\mathrm{tok}}(c_t), \quad D_{\mathrm{KL}}^{\mathrm{tok,max}} := \max_{t, c_t} D_{\mathrm{KL}}^{\mathrm{tok}}(c_t)$$

Sequence-level KL (via chain rule):

$$D_{\mathrm{KL}}^{\mathrm{seq}} := \sum_{t=1}^{T} \mathbb{E}_{c_t \sim d_t^{\pi_{\mathrm{roll}}}} \left[D_{\mathrm{KL}}^{\mathrm{tok}}(c_t) \right] \le T \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}$$

Pinsker's Inequality: $D_{\mathrm{TV}}(P, Q) \le \sqrt{D_{\mathrm{KL}}(P \| Q)/2}$ (holds for either direction!)

2.5.1 TV Bound: Simulation Lemma (TRPO)

Simulation Lemma [Kakade & Langford 2002]:

$$\|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{\text{TV}} \leq \sum_{s=1}^{t-1} D_{\text{TV}}^{\text{tok,max}} = (t-1) \cdot D_{\text{TV}}^{\text{tok,max}}$$

(Each step contributes at most $D_{\text{TV}}^{\text{tok,max}}$ to the context distribution shift.)

Summing over t:

$$\sum_{t=1}^{T} \|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{\text{TV}} \leq \sum_{t=1}^{T} (t-1) \cdot D_{\text{TV}}^{\text{tok,max}} = \frac{T(T-1)}{2} \cdot D_{\text{TV}}^{\text{tok,max}}$$

TRPO-Style Error Bound (combining with $\|g_t\|_\infty \leq 2D_{\text{TV}}^{\text{tok,max}}$):

$$|\text{Error}| \leq 2 \sum_{t=1}^{T} \|g_t\|_\infty \cdot \|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{\text{TV}}$$

$$\leq 2 \cdot (2D_{\text{TV}}^{\text{tok,max}}) \cdot \frac{T(T-1)}{2} \cdot D_{\text{TV}}^{\text{tok,max}} = \boxed{2T(T-1) \cdot (D_{\text{TV}}^{\text{tok,max}})^2}$$

Via Pinsker $(D_{\text{TV}})^2 \leq D_{\text{KL}}/2$: $\boxed{|\text{Error}| \leq T(T-1) \cdot D_{\text{KL}}^{\text{tok,max}}}$ Scaling: $O(T^2)$

2.5.2 The Problem with TRPO Bound

For Long-Horizon Reasoning ($T = 4096$, $D_{\mathbf{KL}}^{\mathbf{tok,max}} = 10^{-4}$):
Pure KL form:

$$|\text{Error}| \leq T(T-1) \cdot D_{\text{KL}}^{\text{tok,max}} = 4096 \times 4095 \times 10^{-4} \approx 1677$$

The Vacuous Bound Problem:

- For any practical step size, the error bound exceeds any possible gain
- No theoretical guarantee of improvement!

Root Cause:

- TV sub-additivity: $\|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{\text{TV}} \leq (t-1) \cdot D_{\text{TV}}^{\text{tok,max}}$
- Final bound has $(D_{TV})^2$, so Pinsker's $\sqrt{\cdot}$ cancels with square
- Result: $O(T^2)$ scaling persists even in pure KL form

Can we do better? Yes — by using KL's chain rule! (See Appendix A.7)

2.5.3 New Bound: Pinsker on Marginal KL (Key Insight)

Key Insight: Apply Pinsker to the **accumulated marginal KL**, not per-step TV.
Step 1: KL of marginal context distributions (chain rule)

$$D_{\mathrm{KL}}(d_t^{\pi_{\mathbf{roll}}} \| d_t^{\pi_\theta}) = \sum_{s=1}^{t-1} \mathbb{E}_{c_s \sim d_s^{\pi_{\mathbf{roll}}}} \left[D_{\mathrm{KL}}^{\mathrm{tok}}(c_s) \right] \le (t-1) \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}$$

Step 2: Apply Pinsker to the MARGINAL KL

$$\|d_t^{\pi_\theta} - d_t^{\pi_{\mathbf{roll}}}\|_{\mathrm{TV}} \le \sqrt{\frac{D_{\mathrm{KL}}(d_t^{\pi_{\mathbf{roll}}} \| d_t^{\pi_\theta})}{2}} \le \sqrt{\frac{(t-1) \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}}{2}}$$

Crucial difference:

- TRPO: $\|d_t^{\pi_\theta} - d_t^{\pi_{\mathbf{roll}}}\|_{\mathrm{TV}} \le (t-1) \cdot D_{\mathrm{TV}}^{\mathrm{tok,max}}$ (linear in t)
- **New:** $\|d_t^{\pi_\theta} - d_t^{\pi_{\mathbf{roll}}}\|_{\mathrm{TV}} \le \sqrt{(t-1) \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}/2}$ ($\sqrt{t}$ growth!)

The $\sqrt{\cdot}$ in Pinsker converts linear KL accumulation to $\sqrt{t}$ TV growth!

2.5.4 Main Result: Pinsker-Marginal Bound

Step 3: Sum over t

$$\sum_{t=1}^{T} \|d_t^{\pi_\theta} - d_t^{\pi_{\textbf{roll}}}\|_{\text{TV}} \le \sqrt{\frac{D_{\text{KL}}^{\text{tok,max}}}{2}} \sum_{k=0}^{T-1} \sqrt{k} \le \frac{2}{3} T^{3/2} \sqrt{\frac{D_{\text{KL}}^{\text{tok,max}}}{2}}$$

Step 4: Combine with $\|g_t\|_\infty \le \sqrt{2 \cdot D_{\textbf{KL}}^{\textbf{tok,max}}}$
Pure KL Form (via Pinsker on both bounds):

$$\boxed{|\text{Error}| \le \frac{4}{3} T^{3/2} \cdot D_{\text{KL}}^{\text{tok,max}}}$$

where $D_{\text{KL}}^{\text{tok,max}} = \max_{t,c_t} D_{\text{KL}}(\pi_{\text{roll}}(\cdot|c_t)\|\pi_\theta(\cdot|c_t))$ (TRPO convention).
Scaling: $O(T^{3/2})$ — $\sqrt{T}$ improvement over TRPO!
[NEW — Our contribution]

2.5.5 Alternative: Mixed DPI Bound

Alternative approach: Use sequence-level KL for uniform bound.
Step 1: Marginal KL is a partial sum

$$D_{\mathrm{KL}}(d_t^{\pi_{\mathbf{roll}}} \| d_t^{\pi_\theta}) \leq D_{\mathrm{KL}}^{\mathrm{seq}} \quad \text{(uniform in } t\text{)}$$

Step 2: Apply Pinsker (gives constant bound in t!)

$$\|d_t^{\pi_\theta} - d_t^{\pi_{\mathbf{roll}}}\|_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}^{\mathrm{seq}}/2}$$

Step 3: Sum and combine with $\|g_t\|_\infty \leq \sqrt{2 \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}}$
Pure KL Form (via Pinsker):

$$\boxed{|\mathrm{Error}| \leq 2T \cdot \sqrt{D_{\mathrm{KL}}^{\mathrm{tok,max}} \cdot D_{\mathrm{KL}}^{\mathrm{seq}}}}$$

Scaling: $O(T)$ — linear in T, requires both max and seq divergences.
[NEW — Our contribution]

2.5.6 Summary: Complete Error Bound Hierarchy

All bounds use $D_{\mathrm{KL}}(\pi_{\mathrm{roll}} \| \pi_\theta)$ (TRPO convention, exactly computable).

Method	Error Bound	Scaling
TV Chain (TRPO)	$T(T-1)D_{\mathrm{KL}}^{\mathrm{tok,max}}$	$O(T^2)$
Pinsker-Marginal	$\frac{4}{3}T^{3/2} \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}$	$O(T^{3/2})$
Mixed (DPI)	$2T\sqrt{D_{\mathrm{KL}}^{\mathrm{tok,max}} \cdot D_{\mathrm{KL}}^{\mathrm{seq}}}$	$O(T)$

Key Insight: Apply Pinsker to *marginal* KL (from chain rule), not per-step TV.
Note: No pure $D_{\mathrm{KL}}^{\mathrm{seq}}$ bound exists (see Appendix A.8 for counterexample).

2.5.7 Numerical Comparison

Setting: $T = 4096$, $D_{\mathrm{KL}}^{\mathrm{tok,max}} = 10^{-4}$

Scenario 1: Uniform KL ($D_{\mathrm{KL}}^{\mathrm{seq}} = T \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}} = 0.41$)

Bound	**Value**	**Source**	**Tighter?**
TV Chain (TRPO)	1677	TRPO	
Pinsker-Marginal	35.0	New	✓
Mixed (DPI)	52.4	New	

Improvement: Pinsker-Marginal is 48× tighter than TRPO!

Scenario 2: Sparse high-KL ($D_{\mathrm{KL}}^{\mathrm{seq}} = 0.01$)

Bound	**Value**	**Tighter?**
Pinsker-Marginal	35.0	
Mixed (DPI)	8.2	✓

2.6.1 Constructing the Lower Bound

From: $J(\pi_\theta) - J(\pi_{\text{roll}}) = L(\pi_\theta) + \text{Error}$
Lower Bound (Minorizer):

$$\mathcal{M}(\pi_\theta) := L(\pi_\theta) - |Error|_{\text{bound}}$$
$$= L - \min\left\{\frac{4}{3}T^{3/2} \cdot D_{\text{KL}}^{\text{tok,max}},\ 2T \cdot \sqrt{D_{\text{KL}}^{\text{tok,max}} \cdot D_{\text{KL}}^{\text{seq}}}\right\}$$

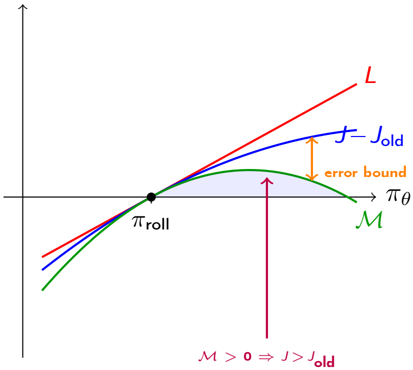
Properties:

1. $J(\pi_\theta) - J(\pi_{\text{roll}}) \geq \mathcal{M}(\pi_\theta)$ (valid lower bound)
2. $\mathcal{M}(\pi_{\text{roll}}) = 0$ (tight at reference)

Monotonic Improvement Guarantee:

$$\boxed{\mathcal{M}(\pi_{\text{new}}) > 0 \implies J(\pi_{\text{new}}) > J(\pi_{\text{roll}})}$$

2.6.3 Visualization



Key: The lower bound $\mathcal{M}$ uses the **tighter** of two error bounds. Shaded region = guaranteed improvement.

2.6.4 Comparison with TRPO

TRPO Lower Bound [Schulman et al. 2015]:

$$\mathcal{M}_{\text{TRPO}} = L - C \cdot T^2 \cdot (D_{\text{TV}}^{\text{tok,max}})^2$$

Our Lower Bound:

$$\mathcal{M}_{\text{new}} = L - \min\left\{\frac{4}{3}T^{3/2}D_{\text{KL}}^{\text{tok,max}},\ 2T\sqrt{D_{\text{KL}}^{\text{tok,max}} \cdot D_{\text{KL}}^{\text{seq}}}\right\}$$

Key Improvements:

1. **Better T-scaling:** $O(T^{3/2})$ or $O(T)$ vs $O(T^2)$
2. **Linear in KL:** Not quadratic in TV
3. **Adaptive:** Uses tighter of two bounds based on $D_{\text{KL}}^{\text{seq}}$

Result ($T = 4096$, $D_{\text{KL}}^{\text{tok,max}} = 10^{-4}$): Error bound 35.0 (PM) or 8.2 (Mixed) vs 1677 (TRPO) — up to 200$\times$ tighter!

Maximizing $\mathcal{M}$ is equivalent to:

$$\max_{\pi_\theta} L_{\pi_{\textbf{roll}}}(\pi_\theta) \quad \text{s.t.} \quad D_{\text{KL}}^{\text{tok,max}}(\pi_\theta \| \pi_{\text{roll}}) \leq \delta$$

With constraint $D_{\text{KL}}^{\text{tok,max}} \leq \delta$:

$$\mathcal{M} \geq L - \min\left\{\frac{4}{3}T^{3/2} \cdot \delta,\ 2T\sqrt{\delta \cdot D_{\text{KL}}^{\text{seq}}}\right\}$$

Key Insights for LLM-RL:

- Trust region must constrain $D_{\text{KL}}^{\text{tok,max}}$ (worst-case token KL)
- This is a **sequence-level** constraint (cannot be enforced token-by-token)
- Token-level constraints (PPO clipping, token masking) are NOT sufficient

Section III

The Failure of Token-Level Constraints

Why PPO violates the trust region

3.1.1 The PPO Objective

PPO [Schulman et al., 2017] approximates trust regions via **ratio clipping**:

$$L^{\text{CLIP}} = \mathbb{E}\left[\sum_{t=1}^{T} \min\left(\rho_t A_t,\ \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_t\right)\right]$$

Intuition: Clipping ρ_t at each token should limit policy change.
The Assumption: Token-level clipping $\Rightarrow$ sequence-level trust region.

Problem: This assumption fails under systemic mismatch.

The min operator creates asymmetric behavior:

ρ_t	A_t	Clipped Term	**Selected**
$> 1 + \epsilon$	> 0	$(1 + \epsilon)A_t$	Clipped (smaller)
$< 1 - \epsilon$	< 0	$(1 - \epsilon)A_t$	Clipped (less negative)
$> 1 + \epsilon$	< 0	$(1 + \epsilon)A_t$	**Unclipped!**
$< 1 - \epsilon$	> 0	$(1 - \epsilon)A_t$	**Unclipped!**

Problem: When $\rho_t \gg 1$ and $A_t < 0$, gradient is **not bounded**.

3.2.2 Gradient Leakage Example

Scenario: MoE routing flip causes $\rho_t = 100$, noisy reward gives $A_t = -1$.

1. Unclipped: $\rho_t A_t = 100 \times (-1) = -100$
2. Clipped: $(1 + \epsilon) A_t = 1.2 \times (-1) = -1.2$
3. PPO selects: $\min(-100, -1.2) = -100$

Result: Gradient magnitude $\propto 100$, completely uncontrolled.

Under Systemic Mismatch:
MoE artifacts routinely produce $\rho \gg 1$. Combined with noisy advantages, this injects massive erroneous gradients.

3.3.1 The Token Masking Proposal

Attempted Fix: Mask tokens with excessive divergence.

$$\nabla \approx \sum_{t=1}^{T} M_t \cdot \rho_t \nabla \log \pi_\theta(y_t | x, y_{<t}) \cdot A$$

where $M_t = 0$ if $|\log \rho_t| > \delta$.
Intuition: Remove "bad" tokens from gradient $\Rightarrow$ safe update?
Problem: This does NOT satisfy Section 2's requirements.

3.3.2 The Theoretical Problem

Recall Section 2's Error Bound:

$$|\text{Error}| \leq \frac{4}{3} T^{3/2} \cdot D_{\text{KL}}^{\text{tok,max}}$$

This bound depends on $D_{\text{KL}}^{\text{tok,max}} = \max_{t,c_t} D_{\text{KL}}^{\text{tok}}(c_t)$ — the worst-case over all contexts in the **sequence**.

If token k has large divergence ($|\log \rho_k| \gg \delta$):

- This indicates π_θ and π_{roll} differ significantly at context c_k
- The sequence's $D_{\text{KL}}^{\text{tok,max}}$ is large
- Masking token k changes the **gradient** we compute
- But $D_{\text{KL}}^{\text{tok,max}}$ is **unchanged** — the divergence still exists!
- Error bound remains **vacuous**

Key Insight: Token masking changes *what we optimize*, not *the error bound*. The theory requires $D_{\text{KL}}^{\text{tok,max}} \leq \delta$ for the **sequence**, not just for tokens we include.

3.4.1 Token-Level Methods Cannot Win

Summary of Failure Modes:

Method	Problem	Theory Satisfied?
Include bad tokens	Gradient Leakage	No
Mask bad tokens	$D_{\mathrm{KL}}^{\mathrm{tok,max}}$ unchanged	No

Root Cause:

- The error bound requires $D_{\mathrm{KL}}^{\mathrm{tok,max}} \leq \delta$ for the **sequence**
- Token-level operations cannot control **sequence-level** divergence

The Only Solution:
If ANY token violates the trust region, reject the **entire sequence**.

3.4.2 The Necessity of Sequence Masking

The Theory Requires: $D_{\text{KL}}^{\text{tok,max}} \leq \delta$ for the bound to hold.
The Only Solution: Mask entire sequences where ANY token violates.

$$M(x, y) = \mathbb{I}\left[\max_t |\log \rho_t| \leq \delta\right]$$

Why This Works:

- **Masked sequences:** Contribute 0 to gradient (valid, just zero)
- **Accepted sequences:** Have $D_{\text{KL}}^{\text{tok,max}} \leq \delta$ (bound applies!)

Contrast with Token Masking:

- Token masking changes the **gradient**
- But error bound depends on **sequence's** $D_{\text{KL}}^{\text{tok,max}}$
- Masking tokens doesn't reduce $D_{\text{KL}}^{\text{tok,max}}$ — the divergence still exists!

Conclusion: Sequence masking is **directly required** by Section 2's theory.

Section IV

Solution: Trust Region Masking

Sequence-level masking with valid gradient estimation

4.1.1 The Hard Trust Region Mask

Definition: Binary sequence mask

$$M(x, y) = \mathbb{I}\left[(x, y) \in \text{Trust Region}\right]$$

Modified Objective: We optimize a **masked surrogate**:

$$L_{\text{masked}} = \mathbb{E}_{\pi_{\text{roll}}}\left[M \cdot A \cdot \sum_{t=1}^{T} \rho_t\right]$$

Gradient Estimator:

$$\nabla L_{\text{masked}} \approx \frac{1}{N} \sum_{i=1}^{N} M_i \sum_{t=1}^{T} \rho_t \nabla \log \pi_\theta(y_t | x, y_{<t}) \cdot A$$

where $M_i = M(x^{(i)}, y^{(i)})$ and N is batch size.
Key: Divide by N (total batch), not $|\{i : M_i = 1\}|$ (accepted only).
This estimates ∇L_{masked}, which equals ∇L restricted to the trust region.

4.1.2 Why Sequence Masking Works

What happens to each sequence:

- **Masked sequences** ($M_i = 0$):
  - Contribute 0 to gradient
  - This is valid: we choose not to learn from these
  - The bound doesn't need to hold for masked sequences
- **Accepted sequences** ($M_i = 1$):
  - Have $D_{\mathrm{KL}}^{\mathrm{tok,max}} \leq \delta$ by construction
  - Error bound applies with penalty $\propto \delta$
  - Monotonic improvement guarantee holds!

Contrast with Token Masking:

- Token masking: keeps sequence, removes tokens $\Rightarrow$ invalid gradient target
- Sequence masking: removes entire sequence $\Rightarrow$ valid (just zero contribution)

4.2.1 Two Possible Criteria

Criterion A (Max-based): Mask if $\max_t f(\rho_t) > \delta$

- Directly bounds $D_{\text{KL}}^{\text{tok,max}}$ (what theory requires)
- Length-invariant: max doesn't grow with T

Criterion B (Total-based): Mask if $\sum_t f(\rho_t) > \delta$

- Bounds total divergence
- Length-biased: sum grows linearly with T

Example (per-token $f(\rho) = 0.01$):

	$T = 100$	$T = 4000$	Grows with T?
Max	0.01	0.01	No
Total	1	40	Yes ($40\times$)

Conclusion: Max-based is length-invariant! But what if max is too strict?

4.3.1 Practical Considerations

Max-based masking ($\max_t f(\rho_t) > \delta$) is theoretically optimal but:

- Single outlier token masks entire trajectory
- Under MoE noise, may mask too many sequences

Practical Alternative: Average-based criterion

$$\frac{1}{T}\sum_t f(\rho_t) > \delta$$

Also length-invariant; more tolerant of occasional outliers.

Divergence Estimators:

- Max: $\hat{D}_{\max} = \max_t f(\rho_t)$ — detects worst-case divergence
- Avg: $\hat{D}_{\text{avg}} = \frac{1}{T}\sum_t f(\rho_t)$ — estimates $D_{\text{KL}}^{\text{seq}}/T$

Choice of $f(\rho)$ depends on criterion:

- **For max:** Use $|\log \rho|$ (symmetric: detects both $\rho \gg 1$ and $\rho \ll 1$)
- **For avg:** Use $\rho - 1 - \log \rho$ (unbiased AND non-negative)

Note: Sample-based methods are approximate detectors, not rigorous bounds.

4.3.2 Recommended Masking Criteria

For Full Theoretical Guarantee (bounds $D_{\mathrm{KL}}^{\mathrm{tok,max}}$):

$$M(x,y) = \mathbb{I}\left[\hat{D}_{\mathrm{max}}(x,y) \leq \delta_{\mathrm{max}}\right]$$

For Practical Average-Based Filter:

$$M(x,y) = \mathbb{I}\left[\hat{D}_{\mathrm{avg}}(x,y) \leq \delta_{\mathrm{avg}}\right]$$

Combined (Recommended):

$$\boxed{M(x,y) = \mathbb{I}\left[\hat{D}_{\mathrm{max}} \leq \delta_{\mathrm{max}} \textbf{ AND } \hat{D}_{\mathrm{avg}} \leq \delta_{\mathrm{avg}}\right]}$$

- Max criterion: ensures theoretical bound applies
- Average criterion: additional robustness for overall divergence

4.4.1 Connection to Theory

Error bound requires $D_{\text{KL}}^{\text{tok,max}}$:

$$|\text{Error}| \leq \min\left\{\frac{4}{3}T^{3/2}\cdot D_{\text{KL}}^{\text{tok,max}},\ 2T\sqrt{D_{\text{KL}}^{\text{tok,max}}\cdot D_{\text{KL}}^{\text{seq}}}\right\}$$

Max-based criterion directly bounds this: $\max_t D_{\text{KL}}(c_t) \leq \delta \Rightarrow D_{\text{KL}}^{\text{tok,max}} \leq \delta$

Exact vs Sample-Based Divergence:

- **Exact:** $D_{\text{KL}}(c_t) = \sum_v \pi_{\text{roll}}(v|c_t)\log\frac{\pi_{\textbf{roll}}(v|c_t)}{\pi_\theta(v|c_t)}$ — requires stored logits, **rigorous guarantee**
- **Sample-based:** $f(\rho_t) = \rho_t - 1 - \log\rho_t$ — unbiased estimator, **approximate guarantee**

Key Properties:

- Threshold δ is **length-invariant**
- Bounds hold for reverse KL via Pinsker symmetry

4.5.1 TRM Algorithm

Trust Region Masking (TRM):

Require: Divergence function f; thresholds δ_{max}, δ_{avg}; batch $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$

for each $(x, y) \in \mathcal{D}$ **do**

    Compute $\rho_t = \pi_\theta(y_t|x, y_{<t})/\pi_{\text{roll}}(y_t|x, y_{<t})$ for all t

    Compute max divergence: $\hat{D}_{\text{max}} = \max_t f(\rho_t)$

    (Optional) Compute average: $\hat{D}_{\text{avg}} = \frac{1}{T}\sum_t f(\rho_t)$

    Set mask: $M_i = \mathbb{I}[\hat{D}_{\text{max}} \leq \delta_{\text{max}}]$ (and optionally $\hat{D}_{\text{avg}} \leq \delta_{\text{avg}}$)

end for

Compute gradient (divide by N, not $|\{i : M_i = 1\}|$):

$$\nabla L_{\text{masked}} = \frac{1}{N}\sum_{i=1}^{N} M_i \cdot A^{(i)} \cdot \sum_{t=1}^{T} \rho_t^{(i)} \nabla \log \pi_\theta(y_t^{(i)}|x^{(i)}, y_{<t}^{(i)})$$

Update: $\theta \leftarrow \theta + \alpha \nabla L_{\text{masked}}$

Note: For $f(\rho)$ choices, see slide 4.3.3 and Appendix A.1–A.2.

Key: $D_{\mathrm{KL}}(\pi_{\mathrm{roll}} \| \pi_\theta)$ is Exactly Computable

- Following TRPO, we use $D_{\mathrm{KL}}(\pi_{\mathrm{roll}}(\cdot|c_t) \| \pi_\theta(\cdot|c_t))$
- Computed as: `KL(softmax(roll), log_softmax(`θ`))`
- This is the natural choice: expectation over π_{roll} (from which we sample)

TRM Guarantee:

1. **Bounded Divergence:** $\max_t D_{\mathrm{KL}}(\pi_{\mathrm{roll}}(\cdot|c_t) \| \pi_\theta(\cdot|c_t)) \leq \delta$ *(exactly verifiable)*
2. **Improvement Bound:**

$$\boxed{\mathcal{M} = L - \min\left\{\frac{4}{3}T^{3/2} \cdot \delta,\ 2T\sqrt{\delta \cdot D_{\mathrm{KL}}^{\mathrm{seq}}}\right\}}$$

Numerical Example ($T = 4096$, $\delta = 10^{-4}$, $D_{\mathrm{KL}}^{\mathrm{seq}} = 0.01$):

- Error bounds: 35.0 (PM), 8.2 (Mixed) vs 1677 (classical) — **non-vacuous!**

References

- Williams, R. J. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*.
- Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (2000). Policy Gradient Methods for RL with Function Approximation. *NeurIPS*.
- Kakade, S. & Langford, J. (2002). Approximately Optimal Approximate Reinforcement Learning. *ICML*.
- Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley.
- Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust Region Policy Optimization. *ICML*.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv:1707.06347*.

Appendix

A.1 The k_3 Estimator: Why It's Ideal for Averaging

Problem with $k_1 = -\log \rho$: Can be negative when $\rho > 1$.
Extreme ratios cancel: $-\log(0.01) + (-\log(100)) = 4.61 - 4.61 = 0$

Solution: Use $k_3(\rho) = \rho - 1 - \log \rho$

Three Key Properties for Averaging:

1. **Unbiased:** $\mathbb{E}_{y \sim \pi_{\text{roll}}}[k_3(\rho)] = D_{\text{KL}}(c_t)$ exactly
2. **Non-negative:** $k_3(\rho) \geq 0$ for all $\rho > 0 \Rightarrow$ no cancellation
3. **Calibrated asymmetry:** High k_3 when $\rho \gg 1$ is compensated by low $P(\rho \gg 1)$ under π_{roll}

Result: $(1/T) \sum_t k_3(\rho_t) \to D_{\text{KL}}^{\text{seq}}/T$ by law of large numbers

Contrast: $-\log \rho$ is unbiased but cancels; $|\log \rho|$ is non-negative but biased.

A.2 Why $|\log \rho|$ for Max, k_3 for Average

ρ	k_3	$\lvert\log \rho\rvert$	**Interpretation**
0.01	3.6	4.61	π_θ assigns low prob
100	94.4	4.61	π_θ assigns high prob

For MAX criterion: Need symmetric detection!

- Both $\rho \ll 1$ and $\rho \gg 1$ indicate large $D_{\mathrm{KL}}^{\mathrm{tok}}$
- $|\log \rho|$: Detects both equally $(4.61 = 4.61)$ ✓
- k_3: Misses $\rho \ll 1$ if threshold set for $\rho \gg 1$ $\times$

For AVERAGE criterion: Need unbiased + non-negative!

- k_3: Unbiased $(\mathbb{E}[k_3] = D_{\mathrm{KL}})$ AND non-negative ✓
- $-\log \rho$: Unbiased but cancellation when summing $\times$
- $|\log \rho|$: Non-negative but biased $\times$

Caveat: Sample-based methods are approximate detectors, not rigorous bounds.

A.3 Proof Sketch: Performance Difference Identity

Goal: Show $J(\pi_\theta) - J(\pi_{\text{roll}}) = \sum_t \mathbb{E}_{d_t^{\pi_\theta}}[g_t]$.
Step 1: Telescope over timesteps.

$$
\begin{aligned}
J(\pi_\theta) &= \mathbb{E}_{d_1^{\pi_\theta}}[V_1^{\pi_{\text{roll}}}] + \sum_{t=1}^{T} \mathbb{E}_{d_t^{\pi_\theta}} \left[\mathbb{E}_{\pi_\theta}[Q_t^{\pi_{\text{roll}}} - V_t^{\pi_{\text{roll}}}]\right] \\
&= V^{\pi_{\text{roll}}}(x) + \sum_{t=1}^{T} \mathbb{E}_{d_t^{\pi_\theta}}[g_t]
\end{aligned}
$$

Step 2: Note $J(\pi_{\text{roll}}) = V^{\pi_{\text{roll}}}(x)$.
Step 3: Subtract to get:

$$
J(\pi_\theta) - J(\pi_{\text{roll}}) = \sum_{t=1}^{T} \mathbb{E}_{d_t^{\pi_\theta}}[g_t]
$$

See Kakade & Langford (2002) for complete proof.

A.4 Connecting Trajectory and Per-Step Advantages

Per-step advantage (used in PDI): $A_t^{\pi_{\textbf{roll}}}(x, y_{\leq t}) = Q^{\pi_{\textbf{roll}}}(x, y_{\leq t}) - V^{\pi_{\textbf{roll}}}(x, y_{<t})$
Trajectory advantage (used in surrogate): $A = R(x, y) - b$ (same for all t)
Key Identity: For terminal reward: $\sum_{t=1}^{T} A_t = R - \mathbb{E}_{\pi_{\textbf{roll}}}[R]$ (telescope)
Surrogate Equivalence: $L = \mathbb{E}_{\pi_{\textbf{roll}}}[\sum_t \rho_t A_t] = \sum_t \mathbb{E}_{d_t^{\pi_{\textbf{roll}}}}[g_t]$
Proof sketch:

$$L = \sum_{t=1}^{T} \mathbb{E}_{c_t \sim d_t^{\pi_{\textbf{roll}}}} \left[\mathbb{E}_{y_t \sim \pi_{\textbf{roll}}} \left[\frac{\pi_\theta(y_t | c_t)}{\pi_{\text{roll}}(y_t | c_t)} A_t \right] \right]$$
$$= \sum_{t=1}^{T} \mathbb{E}_{c_t \sim d_t^{\pi_{\textbf{roll}}}} \left[\mathbb{E}_{y_t \sim \pi_\theta} [A_t] \right] = \sum_{t=1}^{T} \mathbb{E}_{d_t^{\pi_{\textbf{roll}}}} [g_t]$$

Using trajectory advantage $A = R - b$ is a practical simplification maintaining first-order validity.

A.6 Derivation: Simulation Lemma Bound

Claim: $\|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{TV} \leq (t-1) \cdot D_{\text{TV}}^{\text{tok,max}}$
Proof by Induction:
Base: $t = 1$: $d_1^{\pi_\theta} = d_1^{\pi_{\text{roll}}} = P(x)$, so $\|d_1^{\pi_\theta} - d_1^{\pi_{\text{roll}}}\|_{TV} = 0$. ✓
Inductive Step:

$$d_{t+1}^{\pi}(x, y_{\leq t}) = d_t^{\pi}(x, y_{<t}) \cdot \pi(y_t | x, y_{<t})$$

Using the coupling bound $\|PQ - P'Q'\|_{TV} \leq \|P - P'\|_{TV} + \|Q - Q'\|_{TV}$:

$$\begin{aligned} \|d_{t+1}^{\pi_\theta} - d_{t+1}^{\pi_{\text{roll}}}\|_{TV} &\leq \|d_t^{\pi_\theta} - d_t^{\pi_{\text{roll}}}\|_{TV} + D_{\text{TV}}^{\text{tok,max}} \\ &\leq (t-1) \cdot D_{\text{TV}}^{\text{tok,max}} + D_{\text{TV}}^{\text{tok,max}} = t \cdot D_{\text{TV}}^{\text{tok,max}} \end{aligned}$$

Hence $\|d_{t+1}\|_{TV} \leq t \cdot D_{\text{TV}}^{\text{tok,max}}$, completing the induction.

A.7 Why KL Has a Chain Rule But TV Doesn't

KL Chain Rule (EQUALITY):

$$D_{\mathrm{KL}}(d_t^{\pi_{\mathbf{roll}}} \| d_t^{\pi_\theta}) = \sum_{s=1}^{t-1} \mathbb{E}_{c_s \sim d_s^{\pi_{\mathbf{roll}}}} \left[D_{\mathrm{KL}}^{\mathrm{tok}}(c_s) \right]$$

This is an **exact equality** due to the logarithmic structure of KL.

TV Simulation Lemma (INEQUALITY):

$$\| d_t^{\pi_\theta} - d_t^{\pi_{\mathbf{roll}}} \|_{\mathrm{TV}} \leq \sum_{s=1}^{t-1} D_{\mathrm{TV}}^{\mathrm{tok,max}} = (t-1) \cdot D_{\mathrm{TV}}^{\mathrm{tok,max}}$$

This is only an **upper bound** — TV has no equality chain rule.

Key Insight:

- KL accumulates: $D_{\mathrm{KL}}(d_t^{\pi_{\mathbf{roll}}} \| d_t^{\pi_\theta}) \leq (t-1) \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}$
- Apply Pinsker: $\| d_t^{\pi_\theta} - d_t^{\pi_{\mathbf{roll}}} \|_{\mathrm{TV}} \leq \sqrt{(t-1) \cdot D_{\mathrm{KL}}^{\mathrm{tok,max}}/2}$
- The $\sqrt{\cdot}$ converts **linear** KL to $\sqrt{t}$ TV growth!

This trick is **impossible with TV alone** — we need KL's equality chain rule.

A.8 Why No Pure $D_{\mathrm{KL}}^{\mathrm{seq}}$ Bound Exists

The Issue: The bound $|g_t(c_t)| \le 2D_{\mathrm{TV}}^{\mathrm{tok}}(c_t)$ is **context-dependent**.
To get $\|g_t\|_\infty$, we must take max: $\|g_t\|_\infty \le 2D_{\mathrm{TV}}^{\mathrm{tok,max}}$
Can we bound $D_{\mathrm{TV}}^{\mathrm{tok,max}}$ (or $D_{\mathrm{KL}}^{\mathrm{tok,max}}$) in terms of $D_{\mathrm{KL}}^{\mathrm{seq}}$?

Counterexample: At one rare context c^*:

- $D_{\mathrm{KL}}^{\mathrm{tok}}(c^*) = 1$, and $D_{\mathrm{KL}}^{\mathrm{tok}}(c_t) = 0$ for all $c_t \neq c^*$
- Probability $\Pr(c^*) = \epsilon$ under $d_t^{\pi_{\mathrm{roll}}}$

Then:

- $D_{\mathrm{KL}}^{\mathrm{tok,max}} = 1$ (fixed, regardless of ϵ)
- $D_{\mathrm{KL}}^{\mathrm{seq}} \approx \epsilon$ (can be arbitrarily small!)

Conclusion: There is NO function f such that $D_{\mathrm{KL}}^{\mathrm{tok,max}} \le f(D_{\mathrm{KL}}^{\mathrm{seq}})$.
This is why our bounds **must** involve $D_{\mathrm{KL}}^{\mathrm{tok,max}}$.