Global Resolution

Optimal Multi-Draft Speculative Sampling via Convex Minimization

Rahul Thomas[1,2] Arka Pal[1]

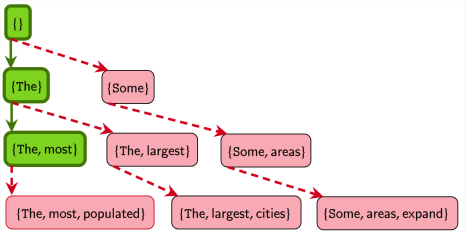
[1]Ritual [2]Columbia University

ICLR

speculative decoding optimal transport

What is multi-draft speculative decoding?

- Autoregressive decoding is usually memory-bound.
- Speculative decoding improves utilization.
- A cheaper draft model generates n i.i.d. tokens.
- In verification, the target model uses these to output one token with its usual distribution **[lossless]**.
- The goal is to maximize **acceptance**:



$$\max_{\text{Verifier}} \Pr[x \in \{y_1, \ldots, y_n\} \mid y_1, \ldots, y_n \sim \text{Draft},\ x \sim \text{Verifier}(\cdot|y_1, \ldots, y_n)]\,.$$

The OT formulation is exact but impractical

$$
\begin{aligned}
\max_{\mathbf{C} \succeq 0} \quad & \sum_{i \in \mathcal{V}} \sum_{\mathbf{i} \in A_i} C_{i,\mathbf{i}} \\
\text{s.t.} \quad & \sum_{\mathbf{i} \in \mathcal{V}^n} C_{i,\mathbf{i}} = p(i) \quad \forall i \in \mathcal{V}, \\
& \sum_{i \in \mathcal{V}} C_{i,\mathbf{i}} = p_{\mathrm{draft}}(\mathbf{i}) \quad \forall \mathbf{i} \in \mathcal{V}^n.
\end{aligned}
$$

- **Incidence structure** $A_i = \{\mathbf{i} \in \mathcal{V}^n : i \in \mathbf{i}\}$.
- Objective value is **optimal acceptance** α^*.
- Verifier is **optimal transport** $\pi(i \mid \mathbf{i})$.

Bottleneck

The LP has $|\mathcal{V}|^{n+1}$ variables, which is not practical for most LLM vocabularies.

Previous work

- **Sun et al. (2023)**: **OTLP** gives optimal transport and optimal acceptance.
- **Khisti et al. (2025)**: **canonical decomposition** into importance sampling.
- **Hu et al. (2025)**: **OTLP relaxation** and **submodular minimization** for optimal acceptance.

$$\alpha^* = 1 + \min_{H \subseteq \mathcal{V}} \psi(H), \qquad \psi(H) = \sum_{i \in H} p(i) - \sum_{\mathbf{i} \in H^n} p_{\mathrm{draft}}(\mathbf{i}).$$

Gap

These results do not give an **exact** and **efficient** way to recover the OT verifier. They cannot be used to implement **optimal** multi-draft speculative decoding.

Global resolution roadmap

- **Obstacle:** Khisti's decomposition is computationally equivalent to Hu's relaxed OTLP.
- **Solution:** Simplify the relaxed OTLP to max-flow and reduce with complementary slackness.
- **Algorithm:** Use polymatroid theory to obtain low-dimensional convex problem for max-flow.

Relaxed OTLP

- Hu's relaxed OTLP is $\approx V\times$ sparser: variables $S_{i,\mathbf{i}} \geq 0$ for $\mathbf{i} \in A_i$.
- **Dimension reduction is the same as canonical decomposition.**
- Excluding degenerate cases, the optimal importance-sampling rule in Khisti is a normalization of the relaxed OTLP solution in Hu.

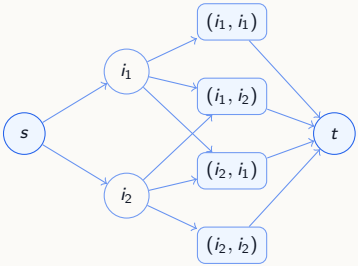
$$\beta(i \mid \mathbf{i}) = \frac{S^*_{i,\mathbf{i}}}{\sum_{j\in\mathcal{V}} S^*_{j,\mathbf{i}}}.$$

Gap

Canonical decomposition is conceptually useful, but it is not an algorithmic breakthrough.

Max-flow

- **Relaxed OTLP is a max-flow problem.**
- Source s to token nodes $i \in \mathcal{V}$ with capacity $p(i)$.
- Each token node i to all tuple nodes $\mathbf{i} \in A_i \subseteq \mathcal{V}^n$.
- Each tuple node $\mathbf{i}$ to sink t with capacity $p_{\text{draft}}(\mathbf{i})$.
- The flows through edges $i \to \mathbf{i}$ are $S_{i,\mathbf{i}}$.

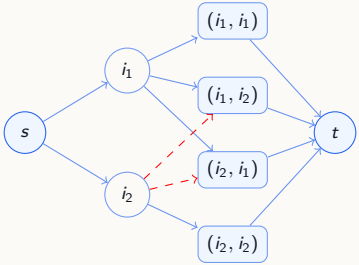


Gap

Max-flow is faster than generic LP solvers but the problem size is still V^n.

Complementary slackness: overview

- **Reverse engineer Hu's work to split the max-flow network into two disjoint networks.**
- Let H^* minimize the subset-selection objective $\psi(H)$ in Hu's submodular problem.
- **Inner system** handles $i \in H^*, \mathbf{i} \in (H^*)^n$; **outer system** handles $i \in \mathcal{V} \setminus H^*, \mathbf{i} \in \mathcal{V}^n \setminus (H^*)^n$.



Complementary slackness: i.i.d. assumption

- In Hu et al., the computation of H^* assumes i.i.d. draft tokens from one draft q.
- We show ψ is submodular even when draft tokens are not i.i.d., as long as they are independent.
- In the i.i.d. case, Hu finds H^* by sorting tokens in decreasing $q(i)/p(i)$ and checking prefixes:

$$\{\} = H_0 \subset H_1 \subset H_2 \subset \cdots \subset H_{|\mathcal{V}|} = \mathcal{V}, \qquad H^* = \arg\min_j \psi(H_j).$$

- In the independent case with $n = 2$, there is a reduction to QPBO.
- In the independent case with $n > 2$, fall back to generic submodular minimization.

Complexity
Finding H^* is $O(V \log V)$ in the i.i.d. case, but not practical for $n > 2$ independent tokens.

Complementary slackness: derivation

- The dualization of the relaxed OTLP in Hu et al. has a solution closely related to H^*.
- Complementary slackness recovers the primal solution from this dual solution.
- Not all edges can be solved, but some $s \to i$ and $\mathbf{i} \to t$ flows are zero, splitting the network.
- Some other $s \to i$ and $\mathbf{i} \to t$ flows can be set to $p(i)$ and $p_{\text{draft}}(\mathbf{i})$, respectively.

Gap

Running max-flow on the split network yields **optimized max-flow**, as we only need to solve the inner system online if $\mathbf{i} \in (H^*)^n$. Unfortunately, the network size is still asymptotically $|\mathcal{V}|^n$.

Convex solvers: outer and inner structural differences

- We split the relaxed OTLP into independent **outer** and **inner** systems.
- **Inner system** has $|H^*|$ vocabulary-side equalities and $|H^*|^n$ tuple-side upper bounds:

$$\sum_{\mathbf{i}\in(H^*)^n} S^*_{i,\mathbf{i}} = p(i) \quad \forall i \in H^*, \qquad \sum_{i\in H^*} S^*_{i,\mathbf{i}} \le p_{\mathrm{draft}}(\mathbf{i}) \quad \forall \mathbf{i} \in (H^*)^n.$$

- **Outer system** has $|\mathcal{V}^n \setminus (H^*)^n|$ tuple-side equalities and $|\mathcal{V} \setminus H^*|$ vocabulary-side upper bounds:

$$\sum_{\mathbf{i}\in\mathcal{V}^n\setminus(H^*)^n} S^*_{i,\mathbf{i}} \le p(i) \quad \forall i \in \mathcal{V} \setminus H^*, \qquad \sum_{i\in\mathcal{V}\setminus H^*} S^*_{i,\mathbf{i}} = p_{\mathrm{draft}}(\mathbf{i}) \quad \forall \mathbf{i} \in \mathcal{V}^n \setminus (H^*)^n.$$

Gap

Standard techniques turn the inner system to a $O(V)$-variable convex problem. But the outer system has vocabulary-side **inequalities**; these must be equalities to do the same reduction.

Convex solvers: outer residual LP

- **The outer residual LP turns vocabulary-side outer system inequalities into equalities.**
- Replacing $\leq p(i)$ with $= p_i$ keeps the outer system feasible if and only if

$$0 \leq p_i \leq p(i), \qquad \forall i \in \mathcal{V} \setminus H^*,$$

$$\sum_{i \in S} p_i \leq \sum_{i \in S} p(i) + \psi(\mathcal{V} \setminus S) \qquad \forall S \subseteq \mathcal{V} \setminus H^*,$$

  and the last inequality is an equality at $S = V \setminus H^*$.

Gap

This LP has exponentially many constraints, so we require additional (i.i.d.) structure to solve it.

Convex solvers: polymatroid algorithm

- **When ψ is submodular, the greedy polymatroid algorithm solves the outer residual LP.**
- Let $V \setminus H^* = \{v_1, \ldots, v_k\}$ in increasing order of $q(i)/p(i)$, and define

$$p_{v_i} = p(v_i) + \min_{T \supseteq H^* \cup \{v_{i+1}, \ldots, v_k\}} \psi(T) - \min_{T \supseteq H^* \cup \{v_i, \ldots, v_k\}} \psi(T), \qquad i \in [k].$$

- The minima above can be recovered from cumulative minimums over prefixes.

Complexity

The outer residual LP is solved in $O(V \log V)$ time, with cost dominated by sorting.

Convex solvers: outer and inner convex solvers

- Both systems now reduce to $O(V)$-size convex problems, with efficient truncated versions:
- **Outer:** For truncation subset $T \subseteq V \setminus H^*$, minimize Φ_T and substitute α into $S_{i,\mathbf{i}}$:

$$\Phi_T(\alpha) = \sum_{\mathbf{i} \in (H^* \cup T)^n \setminus (H^*)^n} p_{\text{draft}}(\mathbf{i}) \log \left(\sum_{i \in \text{set}(\mathbf{i}) \setminus H^*} e^{\alpha_i} \right) - \sum_{i \in T} p_i \alpha_i.$$

$$S_{i,\mathbf{i}} = \frac{e^{\alpha_i}}{\sum_{j \in \text{set}(\mathbf{i}) \setminus H^*} e^{\alpha_j}} p_{\text{draft}}(\mathbf{i}).$$

- **Inner:** For truncation subset $T \subseteq H^*$, minimize Θ_T and substitute α into $S_{i,\mathbf{i}}$:

$$\Theta_T(\alpha) = \sum_{\mathbf{i} \in T^n} p_{\text{draft}}(\mathbf{i}) \log \left(1 + \sum_{i \in \text{set}(\mathbf{i})} e^{\alpha_i} \right) - \sum_{i \in T} p(i) \alpha_i.$$

$$S_{i,\mathbf{i}} = \frac{e^{\alpha_i}}{1 + \sum_{j \in \text{set}(\mathbf{i})} e^{\alpha_j}} p_{\text{draft}}(\mathbf{i}).$$

Global resolution

Inputs: p, q, n, and drafted n-tuple $\omega \in \mathcal{V}^n$.

1. Compute H^* using the sorted-prefix rule.
2. Compute p_i for $i \in V \setminus H^*$ using the polymatroid algorithm.
3. If $\omega \in (H^*)^n$, choose $T \subseteq H^*$, minimize the inner objective Θ_T, and recover the slice $S_{\cdot,\omega}$.
4. In $\omega \notin (H^*)^n$, choose $T \subseteq V \setminus H^*$, minimize the outer objective Φ_T, and recover $S_{\cdot,\omega}$.
5. Post-process $S_{\cdot,\omega}$ with the p_i values to obtain the optimal transport $\pi(\cdot \mid \omega)$.

Disclaimer

We choose T based on a specified tolerance threshold τ. Increasing τ allows one to decrease $|T|$ and solve the truncated convex problem faster, while still maintaining the desired error tolerance.

Practical deployment

- Truncation introduces errors, and T must be large enough to ensure the error is $O(\tau)$.
- If T is too large to quickly solve the inner system, we terminate and use a **fallback**.
- If T is too large to quickly solve the outer system, we sample from a reweighted p.
- Any multi-draft verifier (e.g. K-SEQ, SpecInfer, etc.), can be used as a fallback.

Solver error

If outer and inner solves incur errors α and β:

$$\text{constraint error} \leq \alpha + 2\beta,$$
$$\text{acceptance loss} \leq \alpha + \beta.$$

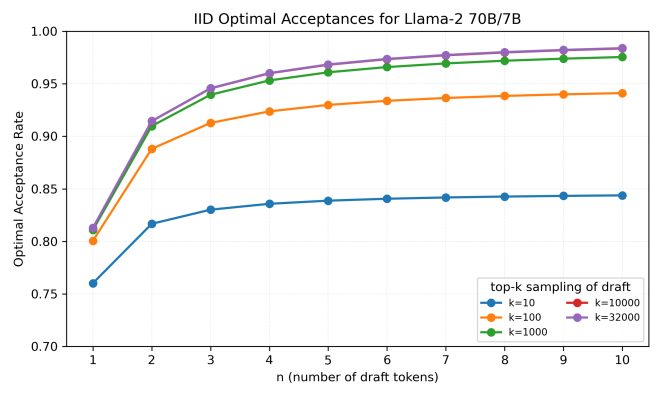
Truncation threshold

With truncation threshold τ:

$$\text{target L1 error} \leq 15\tau,$$
$$\text{acceptance loss} \leq 10\tau.$$

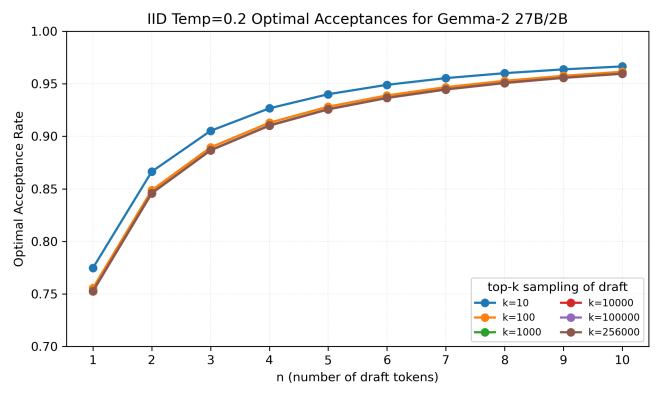
Top-k draft sampling

- Top-k sampling reduces the fallback rate by reducing the minimum $|T|$ for a fixed τ tolerance.
- For small k, optimal acceptances are low, but there are diminishing gains after $k = 1000$.



Temperature variations

- At lower temperatures, top-k draft sampling yields diminishing returns at smaller k.
- Global resolution fallback rates and solve times decrease, improving acceptance.



Inference results

	10 ms/token		100 ms/token	
Model	Max-flow	G.R.	Max-flow	G.R.
Llama-3	83.94%	85.65%	89.01%	90.04%
Gemma-2	80.69%	81.90%	85.83%	86.60%

- At fixed latency budgets, global resolution gives the best acceptance over other approaches.
- Even with the worst fallback [NSS], global resolution doubles throughput on Gemma-27B/2B.

Guidance

To optimize global resolution, one should experiment with various top-k drafting, fallbacks, and truncation thresholds. Different fallback OTLP solvers may perform better in different settings.

Future work

- **Main limitations are fallbacks and restriction to i.i.d. drafting.**
- Inner system fallbacks are resolved, but outer system fallbacks require non-i.i.d. drafting.

Drafting	**Find H^***	**Find p_i's**	**Convex Solvers**
Sampling without replacement	Yes	Yes	—
$n = 2$ independent and distinct	Yes	—	—
$n \geq 3$ independent and distinct	**No**	**No**	—

Thank you

Questions?