Novelty Assessment Report

Paper: Global Resolution: Optimal Multi-Draft Speculative Sampling via Convex Optimization

PDF URL: https://openreview.net/pdf?id=gpsczXOsHn

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-04

Abstract

Speculative sampling reduces the latency of autoregressive decoding for target model LLMs without sacrificing inference quality, by using a cheap draft model to suggest a candidate token and a verification criterion to accept or resample this token. To improve acceptance and decoding efficiency, recent work has explored the multi-draft extension, where at each step N draft tokens are generated, and the verification criterion is a distribution conditioned on these. When this criterion maximizes the probability of accepting some draft token, it is called the optimal transport (OT). However, finding the OT is difficult, as it is the solution of a linear program (OTLP) in over V^n variables, with V being the vocabulary size. Two recent theoretical works have reframed the OTLP in terms of importance sampling or subset selection. In this work, we prove that these formulations are equivalent to an exponentially large relaxed OTLP, so it remains infeasible to solve. Then, we reverse engineer subset selection to formulate the OTLP as a max-flow problem. With a novel application of polymatroid theory, we reduce the exponentially large OTLP to a convex optimization problem in at most V variables. This allows us to devise an algorithm for optimal N-draft speculative sampling when the N tokens are chosen i.i.d. from a single draft model, which can be tuned to arbitrary accuracy. Finally, we measure acceptance rates and algorithm runtimes for various N and top-K draft sampling settings. Our findings give the first multi-draft algorithm with 90\% acceptance and under 100 ms of overhead per generated token with negligible deviation from the target model distribution.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Optimal Multi-Draft Speculative Sampling via Convex Optimization**

A total of **10 papers** were analyzed and organized into a taxonomy with **8 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Optimal Verification and Acceptance Criteria**
- **Adaptive Draft Structure and Policy Optimization**
- **Multi-Model and Batch Inference Optimization**
- **Randomized Drafting Strategies**

Complete Taxonomy Tree

- Optimal Multi-Draft Speculative Sampling via Convex Optimization Survey Taxonomy
- Optimal Verification and Acceptance Criteria
  - Convex Optimization Formulations ★ (2 papers)
  - [0] Global Resolution: Optimal Multi-Draft Speculative Sampling via Convex Optimization (Anon et al., 2026) View paper
  - [1] Global Resolution: Optimal Multi-Draft Speculative Sampling via Convex Minimization (Rahul Krishna Thomas, 2025) View paper
  - Optimal Transport and Linear Programming (2 papers)
  - [2] Towards optimal multi-draft speculative decoding (Hu, 2025) View paper
  - [3] SpecHub: Provable Acceleration to Multi-Draft Speculative Decoding (Ryan Sun, 2024) View paper
  - Canonical Decomposition and Theoretical Limits (2 papers)
  - [5] Multi-Draft Speculative Sampling: Canonical Decomposition and Theoretical Limits (Khisti, 2024) View paper
- Adaptive Draft Structure and Policy Optimization
  - Adaptive Tree Structure Design (1 papers)
  - [4] OPT-Tree: Speculative Decoding with Adaptive Draft Tree Structure (Jikai Wang, 2025) View paper
  - Draft Policy Alignment via Training (1 papers)
  - [8] Bridging Draft Policy Misalignment: Group Tree Optimization for Speculative Decoding (Hu Shijing, 2025) View paper
- Multi-Model and Batch Inference Optimization
  - Hierarchical Multi-Draft Models (1 papers)
  - [7] Fast Inference via Hierarchical Speculative Decoding (Mohri, 2025) View paper
  - Batch-Level Draft Token Selection (1 papers)
  - [9] TETRIS: Optimal Draft Token Selection for Batch Speculative Decoding (Wu Zhaoxuan, 2025) View paper
- Randomized Drafting Strategies (1 papers)
  - [10] Higher Acceptance Rates for Speculative Decoding with Randomised Drafting (W Toner, n.d.) View paper

Narrative

Core task: optimal multi-draft speculative sampling via convex optimization. The field addresses how to accelerate large language model inference by generating multiple candidate token sequences (drafts) from cheaper models and then verifying them against a target model in parallel. The taxonomy organizes research into several main branches. Optimal Verification and Acceptance Criteria focuses on formulating the acceptance decision as a mathematical optimization problem, often using convex methods to determine which drafts to accept. Adaptive Draft Structure and Policy Optimization explores how to dynamically adjust the number, length, or arrangement of drafts based on runtime feedback or learned policies. Multi-Model and Batch Inference Optimization examines scenarios involving multiple draft models or batched requests, seeking to balance computational resources across heterogeneous workloads. Randomized Drafting Strategies investigates stochastic approaches to draft generation, trading deterministic guarantees for potential efficiency gains. Together, these branches reflect a progression from foundational acceptance rules to more flexible, context-aware sampling schemes.

A particularly active line of work centers on convex formulations for acceptance, where Global Resolution Convex Optimization[0] and its close neighbor Global Resolution Convex Minimization[1] rigorously frame the verification step as a convex program, enabling provably optimal token acceptance under distributional constraints. Nearby, Optimal Multi-Draft Decoding[2] and SpecHub[3] explore related optimization perspectives, while works like OPT-Tree[4] and Multi-Draft Canonical Decomposition[5] investigate structured draft arrangements that can be optimized jointly. In contrast, Hierarchical Speculative Decoding[7] and Group Tree Optimization[8] emphasize adaptive, tree-based policies that adjust draft topology on the fly, and Randomised Drafting[10] introduces stochastic elements to escape local optima. The original paper, Global Resolution Convex Optimization[0], sits squarely within the convex optimization branch, offering a principled mathematical framework for acceptance decisions. Compared to Global Resolution Convex Minimization[1], which shares a similar optimization foundation, and SpecHub[3], which also targets verification efficiency, the original work emphasizes a global resolution perspective that unifies multiple drafts under a single convex objective, distinguishing it from more heuristic or tree-centric approaches.

Related Works in Same Category

No comparison data available.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Equivalence of canonical decomposition and relaxed OTLP

Description: The authors prove that the canonical decomposition approach by Khisti et al. (2025) and the subset selection formulation by Hu et al. (2025) are mathematically equivalent to solving a relaxed optimal transport linear program. This unifies two recent theoretical frameworks for multi-draft speculative sampling.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Max-flow and convex optimization reduction of OTLP

Description: The authors reformulate the optimal transport linear program as a max-flow problem and then apply polymatroid theory to reduce it to a tractable convex optimization problem with at most V variables, enabling efficient computation of the optimal verification criterion.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Global resolution algorithm for optimal multi-draft speculative sampling

Description: The authors develop the global resolution algorithm that computes the optimal transport solution for i.i.d. multi-draft speculative sampling to arbitrary accuracy. The algorithm achieves over 90% acceptance rates with under 100 ms overhead per token, representing the first practical multi-draft method with such performance.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 20 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. ImportanceWeighted Multi-Draft Speculative Sampling

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Global Resolution: Optimal Multi-Draft Speculative Sampling via Convex Optimization View paper
- [1] Global Resolution: Optimal Multi-Draft Speculative Sampling via Convex Minimization View paper
- [2] Towards optimal multi-draft speculative decoding View paper
- [3] SpecHub: Provable Acceleration to Multi-Draft Speculative Decoding View paper
- [4] OPT-Tree: Speculative Decoding with Adaptive Draft Tree Structure View paper
- [5] Multi-Draft Speculative Sampling: Canonical Decomposition and Theoretical Limits View paper
- [6] Multi-Draft Speculative Sampling: Canonical Architectures and Theoretical Limits View paper
- [7] Fast Inference via Hierarchical Speculative Decoding View paper
- [8] Bridging Draft Policy Misalignment: Group Tree Optimization for Speculative Decoding View paper
- [9] TETRIS: Optimal Draft Token Selection for Batch Speculative Decoding View paper
- [10] Higher Acceptance Rates for Speculative Decoding with Randomised Drafting View paper
- [11] Canonical supermartingale couplings View paper
- [12] Multiuser gesture recognition using sEMG signals via canonical correlation analysis and optimal transport. View paper
- [13] A tale of two minimization problems: Semimartingale transportation and rough paths lifts View paper
- [14] PURE and APPLIED View paper
- [15] Make every token count: A systematic survey on decoding methods for foundation models View paper
- [16] A theoretical perspective for speculative decoding algorithm View paper
- [17] Cautious next token prediction View paper
- [18] Beyond tokens: A survey on decoding methods for large language models and large vision-language models View paper
- [19] ImportanceWeighted Multi-Draft Speculative Sampling View paper

- [20] A Multi-Model Adaptation of Speculative Decoding for Classification View paper
- [21] : Faster Test-Time Scaling through Speculative Drafts View paper
- [22] List-Level Distribution Coupling with Applications to Speculative Decoding and Lossy Compression View paper
- [23] Speculative Decoding: Exploiting Speculative Execution for Accelerating Seq2seq Generation View paper
- [24] Maximum Flow and Minimum-Cost Flow in Almost-Linear Time View paper
- [25] A universal network strategy for lightspeed computation of entropy-regularized optimal transport. View paper
- [26] Immiscible color flows in optimal transport networks for image classification View paper
- [27] Optimal Transport Flows for Distributed Production Networks View paper
- [28] A Study of Performance of Optimal Transport View paper
- [29] Author Archive View paper