List-Level Distribution Coupling with Applications to Speculative Decoding and Lossy Compression

Joseph Rowan, Buu Phan and Ashish Khisti
University of Toronto

Coordinated sampling from probability distributions

Say Alice wants to sample X from a distribution p_X, and Bob wants to sample Y from q_Y.

How should they sample so that $\Pr[X = Y]$ is maximized?

- A *maximal coupling* achieves $\Pr[X = Y] = 1 - d_{\mathrm{TV}}(p_X, q_Y)$ [9]. But, this requires p_X and q_Y be shared between the parties.
- If communication is *not* allowed, Alice and Bob can apply the Gumbel-max trick to shared random numbers, achieving $\Pr[X = Y] \geq (1 - d_{\mathrm{TV}}(p_X, q_Y))/(1 + d_{\mathrm{TV}}(p_X, q_Y))$ [1, 3].

Coupling with multiple samples

We extend the communication-free problem to a setting where Alice generates K independent samples from p_X.

The new matching probability is $\Pr[Y \in \{X^{(1)}, \ldots, X^{(K)}\}]$. What sampling strategy should Alice and Bob use now?

- To answer this question, we propose *Gumbel-max list sampling* (GLS) for generating the coupled samples.
- Our new algorithm can be applied to multi-draft speculative decoding for accelerated LLM inference [4], and to distrbuted lossy compression with side information [7].

The GLS algorithm

With N being the alphabet size:

1. Generate exponential random variables $\{\{S_i^{(k)}\}_{i=1}^N\}_{k=1}^K$
2. Select $Y = \arg\min_{1\leq i\leq N} \min_{1\leq k\leq K}\{S_i^{(k)}/q_i\}$ to sample q_Y.
3. Select $X^{(k)} = \arg\min_{1\leq i\leq N} S_i^k/p_i$ to sample p_X K times.

In code:

```
def gls_sample(p, q, K)
    S = -np.log(np.random.rand(len(p), K))
    S_ = np.min(S, axis=-1)
    X = np.argmin(S / p[:, None], axis=0)  # X has K elements
    Y = np.argmin(S_ / q)
    return X, Y
```

The list matching lemma

Our main theoretical result, which we call the *list matching lemma* (LML) concerns the matching probability of GLS.

Theorem (List matching lemma)

The matching probability of GLS satisfies

$$\Pr[Y \in \{X^{(1)}, \ldots, X^{(K)}\}] \geq \sum_{j=1}^{N} \frac{K}{\sum_{i=1}^{N}[\max\{q_i/q_j, p_i/p_j\} + (K-1)q_i/q_j]}.$$

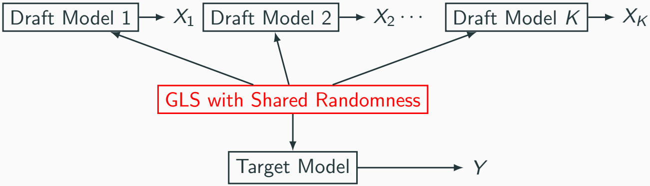
Furthermore, conditioned on $Y = j$,

$$\Pr[Y \in \{X^{(1)}, \ldots, X^{(K)}\} \mid Y = j] \geq \left(1 + \frac{q_j}{Kp_j}\right)^{-1}.$$

Application to multi-draft speculative decoding

Speculative decoding can accelerate LLM inference by drafting tokens using a small, efficient model before verifying them in parallel [5, 2].

In multi-draft variants, the draft model suggests several tokens at once [8, 6]. GLS functions as a drop-in alternative solution for this extended speculative decoding problem.



Results: Multi-draft speculative decoding

GLS compares favorably to existing multi-draft methods like SpecTr [8] and SpecInfer [6] on common langauage tasks, using Qwen2.5 LLMs, especially when the drafts are non-identically distributed.

We also offer a degree of *invariance* with respect to the choice of draft model, which previous algorithms do not provide.

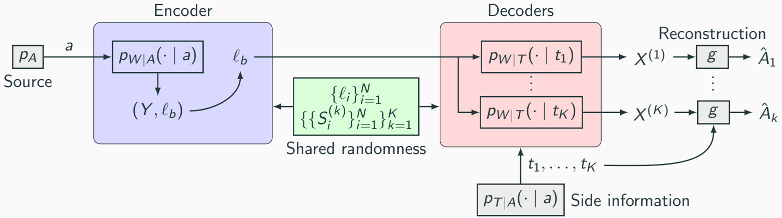
		GSM8K		HumanEval		MBPP	
Strategy	Tmp. 1/2	BE	TR (%)	BE	TR (%)	BE	TR (%)
SpecInfer	0.5/1.0	4.26 ± 0.02	0.06 ± 1.02	3.57 ± 0.02	-1.96 ± 0.67	3.66 ± 0.01	-1.87 ± 0.79
	1.0/0.5	4.44 ± 0.03	4.57 ± 1.80	3.80 ± 0.03	4.13 ± 1.60	3.90 ± 0.02	4.77 ± 0.55
	1.0/1.0	4.51 ± 0.02	6.02 ± 1.35	3.87 ± 0.02	6.32 ± 0.36	3.95 ± 0.02	5.17 ± 1.13
Our scheme	0.5/1.0	4.75 ± 0.02	11.50 ± 1.78	4.00 ± 0.01	9.80 ± 0.82	3.94 ± 0.02	5.64 ± 0.66
	1.0/0.5	4.75 ± 0.02	11.40 ± 1.58	3.96 ± 0.02	8.77 ± 0.99	3.96 ± 0.02	5.99 ± 1.01
	1.0/1.0	4.83 ± 0.02	13.68 ± 1.67	4.08 ± 0.02	12.15 ± 0.83	4.08 ± 0.01	8.57 ± 0.60

Application to distributed lossy compression

Suppose there is one encoder and K decoders, each having access to independent side information. GLS offers an efficient communication scheme with bounded error probability

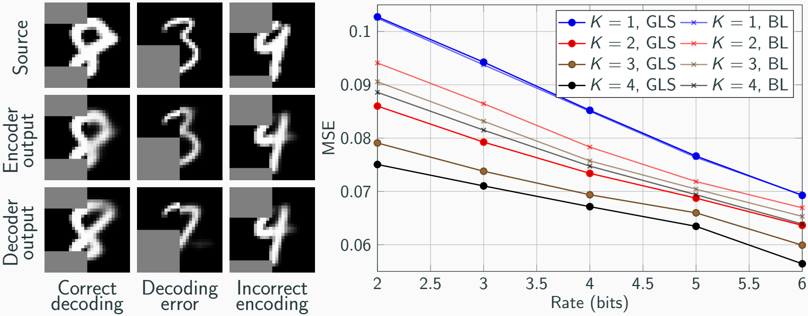
$$\Pr[Y \notin \{X^{(1)}, \ldots, X^{(K)}\}] \leq 1 - \mathrm{E}_{A,W,T}\left[\left(1 + \frac{2^{i(W;A|T)}}{KL_{\max}}\right)^{-1}\right]$$

at rate $\log L_{\max}$, where W follows a fixed target distribution, A is the source and T is the side information.



Results: Distributed lossy compression

Compared to a baseline list-decoding scheme, GLS gives better rate-distortion performance on MNIST and CIFAR-10, where the side information is a randomly selected segment from the left-hand side of the image.



References

[1] Mohammad Bavarian et al. "Optimality of correlated sampling strategies". In: 16.12 (2020), pp. 1–18.

[2] Charlie Chen et al. *Accelerating large language model decoding with speculative sampling*. 2023. arXiv: `2302.01318 [cs.CL]`. URL: `https://arxiv.org/abs/2302.01318`.

[3] Majid Daliri, Christopher Musco, and Ananda Theertha Suresh. "Coupling without Communication and Drafter-Invariant Speculative Decoding". In: *IEEE International Symposium on Information Theory*. 2025.

[4] Ashish J. Khisti et al. "Multi-draft speculative sampling: canonical decomposition and theoretical limits". In: *13th International Conference on Learning Representations*. 2025.

References

[5] Yaniv Leviathan, Matan Kalman, and Yossi Matias. “Fast inference from transformers via speculative decoding”. In: *Proceedings of the 40th International Conference on Machine Learning*. 2023, pp. 19274–19286.

[6] Xupeng Miao et al. “SpecInfer: accelerating large language model serving with tree-based speculative inference and verification”. In: *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*. 2024, pp. 932–949.

[7] Buu Phan, Ashish Khisti, and Christos Louizos. “Importance matching lemma for lossy compression with side information”. In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*. 2024, pp. 1387–1395.

References

[8] Ziteng Sun et al. "SpecTr: fast speculative decoding via optimal transport". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 30222–30242.

[9] Hermann Thorisson. *Coupling, stationarity, and regeneration*. Springer, 2000.