Importance Weighted Multi-Draft Speculative Sampling

Anonymous Authors[1]

Abstract

We consider multi-draft speculative sampling, where the proposal sequences are sampled independently from the same underlying draft model. At each step, a token-level draft selection scheme takes a list of valid tokens as input and produces an output token whose distribution matches that of the target model. Previous works have demonstrated that the optimal scheme (which maximizes the probability of accepting one of the input tokens) can be cast as a solution to a linear program. In this work we show that the optimal scheme can be decomposed into a two-step solution: in the first step an importance sampling (IS) type scheme is used to select one intermediate token; in the second step (single-draft) speculative sampling is applied to generate the output token. Applying our decomposition result to the case of two drafts we 1) establish a necessary and sufficient condition on the distributions of the target and draft models for the acceptance probability to equal one and 2) provide an explicit expression for the optimal acceptance probability. Our theoretical analysis also motives a new class of token-level selection scheme based on weighted importance sampling. We study the performance of such schemes via experiments involving Llama 2-7B chat model for a natural language task and demonstrate improvements over prior approaches.

1. Introduction

The transformer architecture (Vaswani et al., 2017) has revolutionized the field of natural language processing and deep learning. One of the key factors contributing to the success story of transformers, as opposed to prior recurrent-based architectures (Hochreiter & Schmidhuber, 1997; Chung et al., 2014), is their inherent train-time parallelization due to the attention mechanism. This allows for massive scaling and lead to the development of state-of-the-art Large Language Models (LLMs) (Touvron et al., 2023; Achiam et al., 2023; Brown et al., 2020; Chowdhery et al., 2023) which have demonstrated remarkable performance across a wide range of tasks. Despite their parallelizable training, LLM inference is sequential, owing to their auto-regressive nature. This limits their text-generation to one token per one forward pass, which is known to be memory-bound (Shazeer, 2019).

To alleviate the memory-bound nature of auto-regressive decoding of LLMs, speculative decoding (Chen et al., 2023; Leviathan et al., 2023) leverages an arbitrary smaller language model (draft model) that generates multiple candidate tokens in an auto-regressive manner. The LLM (target model) is then used to score all the tokens in the draft *in parallel*, and the draft tokens are verified through a sequence of token-level rejection sampling which guarantees that the final sequence follows the same distribution as that of the target model. In order for speculative decoding to be beneficial, the combined cost of auto-regressively sampling from the draft model and parallel verification via the target model should be smaller than auto-regressively sampling from the target model. Intuitively, this requires that the draft model distribution resembles that of the target model, which can be measured via the acceptance rate of the speculative decoding process, i.e., the rate at which we accept/reject draft tokens.

A large number of works on speculative decoding (Sun et al., 2024b; Jeon et al., 2024; Miao et al., 2024; Sun et al., 2024a) have emerged recently in an effort to further improve decoding efficiency. The authors in (Sun et al., 2024b) propose SpecTr, a multi-draft extension where the draft model generates K candidate token sequences (which could be sampled in a batch) for each time-step (as opposed to one). The authors consider a token-level selection scheme with the objective of maximizing the probability of accepting some token in the set of available tokens. They demonstrate that this problem can be cast into the framework of optimal transport and solved using a linear program. However due to complexity reasons, the authors instead propose a modified sequential rejection sampling scheme. We provide additional literature survey in the related works section.

1.1. Main Contributions

We revisit the optimal transport framework introduced in (Sun et al., 2024b) and introduce an architectural result. We demonstrate that the optimal acceptance probability can be achieved by a two-step scheme: the first step involves selecting a token from the available list using a type of importance sampling; the second step involves speculative sampling using the selected token and the target distribution. We also provide an analytical expression for the optimal acceptance probability for the case of $K = 2$ drafts, thus generalizing a result known previously for the case when $K = 1$. We also establish a necessary and sufficient condition for the acceptance probability to equal one in the case of $K = 2$ drafts. We propose a new class of token-selection schemes based on weighted importance sampling. To enable a faster implementation, we consider three approaches: 1) truncating the linear program 2) truncating the vocabulary set and 3) hybrid combination with other baseline schemes. We present some experimental results using Llama 2-7B as the target model and a smaller draft model with 115m parameters. We compare the performance of our proposed schemes with baselines on the XSum task.

2. Token-Level Optimal Draft Selection: Theoretical Analysis

We focus on token-level optimal draft selection framework introduced in (Sun et al., 2024b). We assume that $\Omega = \{1, 2, \ldots, n\}$ denotes the vocabulary of tokens and at a given step, say t, $\mathcal{S} = \{X_1, \ldots, X_K\}$, denotes the K valid tokens under consideration. Each of these tokens is generated in an i.i.d. fashion from a distribution $p(\cdot)$ determined by the underlying draft model and the context sequence $u^t \in \Omega^t$ i.e, for each $y \in \Omega$, we have $p(y) = \mathcal{M}_s(y|u^t)$, where $\mathcal{M}_s$ denotes the distribution generated by the small (draft) model. In a similar fashion we let $q(\cdot)$ be the distribution over Ω associated with the large model i.e., $q(y) = \mathcal{M}_b(y|u^t)$ where $\mathcal{M}_b$ denotes the distribution generated by the large model. Note that we do not explicitly indicate the sequence u^t when discussing $p(\cdot)$ and $q(\cdot)$, as it is fixed and common to both models throughout our analysis.

Given an input $\mathcal{S} \sim \prod_{i=1}^{K} p(X_i)$ consisting of K candidate tokens $(X_1, \ldots, X_K)$, a *token-level selection rule* (TLSR) is a conditional distribution $\mathcal{P}(\cdot|\mathcal{S})$ over Ω. A *valid* TLSR must satisfy the constraint that for each $z \in \Omega$, $\sum_{\mathcal{S}} \mathcal{P}(z|\mathcal{S})p(\mathcal{S}) = q(z)$. A natural metric to optimize for TLSR is the probability that one of the tokens is accepted i.e., if $Z \sim \mathcal{P}(\cdot|\mathcal{S})$ denotes the output of the TLSR, then we wish to maximize $\Pr(Z \in \mathcal{S})$.

Problem 1 (Optimal Token Level Selection Rule). *Given distributions $p(\cdot)$ and $q(\cdot)$ find a valid TLSR that maximizes the probability of acceptance: $P(\text{acc}) = \Pr(Z \in \mathcal{S})$ and let $P^\star(\text{acc})$ be the optimal value.*

Problem 1 was studied in (Sun et al., 2024b) and shown to be an instance of optimal transport, which can be cast as a linear program. The authors used this framework to establish the optimality of speculative sampling (Chen et al., 2023; Leviathan et al., 2023) in the case of a single draft i.e., $K = 1$. For $K > 1$ the authors established an information theoretic upper bond on $P^\star(\text{acc})$. In this work, we revisit Problem 1 and develop new insights into the structure of the optimal solution. In fact, we establish that the optimal solution in the case of multiple drafts has a natural connection to importance sampling (Tokdar & Kass, 2010). For the case of $K = 2$ drafts we exactly characterize $P^\star(\text{acc})$ and state necessary and sufficient conditions on $p(\cdot)$ and $q(\cdot)$ for $P^\star(\text{acc})$ to equal 1.

We begin by defining a family of schemes that we will refer to as *importance weighted* sampling.

Definition 1 (Importance Weighted Sampling). *An importance weighted sampling scheme takes as input the set of candidate tokens $\mathcal{S} = \{X_1, \ldots, X_K\}$ and outputs a token $Y_I \in \mathcal{S}$ defined by the conditional distribution:*

$$\Pr(Y_I = y|X_{1:K} = x_{1:K}) = \begin{cases} \beta_y(x_1, \ldots, x_K), & y \in \{x_1, \ldots, x_K\} \\ 0, & y \notin \{x_1, \ldots, x_K\} \end{cases} \tag{1}$$

where $\sum_{y \in \Omega} \beta_y(x_1, \ldots, x_K) = 1$ for each $x_{1:K} \in \Omega^K$ and $0 \leq \beta_y(x_1, \ldots, x_K) \leq 1$

Note that instead of considering the probability over the value of the selected token in (1), one can instead consider the probability of selecting an index i between $\{1, \ldots, K\}$ i.e., $\Pr(I = i|X_{1:K} = x_{1:K})$. Such a distribution maps to (1) by simply summing over all indices where $x_i = y$. We note that the form in (1) will be more convenient in the sequel. Also note that the classical importance sampling scheme (Tokdar & Kass, 2010) corresponds to the case where $\Pr(I = i|X_1^k = x_{1:K}) \propto q(x_i)/p(x_i)$. However the family of schemes in Definition 1 is not restricted to such a choice and we treat $\beta_y(x_1, \ldots, x_K)$ as free parameters that can be optimized. Our first result is a decomposition for the optimal token level selection rule that establishes a connection to the importance weighted sampling in Definition 1. The proof is in Appendix D.

Theorem 1. *Let $P^\star(\text{acc})$ be the acceptance probability for the optimal token level selection rule in Problem 1. Then we have*

$$P^\star(\text{acc}) = \max_{\{\beta_y(x_{1:K})\}} \left\{ \sum_{y \in \Omega} \min \left(q(y), \sum_{x_1, \ldots, x_K \in \Omega} \beta_y(x_{1:K}) \cdot \prod_{i=1}^{K} p(x_i) \right) \right\} \tag{2}$$

where the maximum is over $\beta_y(x_{1:K})$ for each $\{x_1, \ldots, x_K, y\} \in \Omega$ such that $0 \leq \beta_y(x_{1:K}) \leq 1$, *and*

$$\sum_{x_{1:K} \in \Omega^K} \beta_y(x_{1:K}) = 1, \quad \forall y \in \Omega, \tag{3}$$

and furthermore $\beta_y(x_{1:K}) = 0, \quad y \notin \{x_1, \ldots, x_K\}$.. *In addition, if* $\{\beta^\star_y(x_{1:K})\}$ *denotes the parameters that achieve the maximum in* (17)*, then* $P^\star(\mathrm{acc})$ *can be attained by a two step approach as follows: in the first step, given the list of input tokens* $\{x_1, \ldots, x_K\}$*, we apply Importance Weighted Sampling in Definition 1 with parameters* $\beta^\star_y(x_1, \ldots, x_K)$ *to output an intermediate token* $y \in \{x_1, \ldots, x_K\}$*; in the second step we apply a single-draft speculative sampling scheme (Chen et al., 2023; Leviathan et al., 2023) on the selected token* y *to generate the final output token.*

Figure 1 illustrates the proposed two step scheme in Theorem 1, where the first step involves importance weighted sampling to output an intermediate token and the second step involves speculative sampling. This approach requires computing the optimal $\beta^\star_y(x_{1:K})$. In practice one can use sub-optimal choices that are faster to compute, as will be discussed in the sequel. Note that the speculative sampling block in the second step guarantees that the output token Z will follow the target distribution even when such sub-optimal choices for $\beta_y(x_{1:K})$ are used. It is also straightforward to extend Theorem 1 when the distributions of the K tokens are not identical i.e., $\mathcal{S} \sim \prod_{i=1}^K p_i(X_i)$, as discussed in Section D.1 in the supplementary material. We next build upon Theorem 1 to establish new analytical results for the optimal acceptance probability involving $K = 2$ drafts. Our first result is a characterization of the necessary and sufficient condition on the draft and target distributions $p(\cdot)$ and $q(\cdot)$ respectively that leads to $P^\star(\mathrm{accept}) = 1$.

Theorem 2. *With* $K = 2$ *drafts, a necessary and sufficient condition for* $P^\star(\mathrm{acc}) = 1$ *in the Definition 1 is the following:*

$$\sum_{x \in \mathcal{S}} q(x) \geq \left(\sum_{x \in \mathcal{S}} p(x)\right)^2, \qquad \forall \mathcal{S} \subseteq \Omega. \tag{4}$$

Note that the acceptance probability can equal 1 even when $p(\cdot)$ and $q(\cdot)$ are not identical. Thus when the distribution of the draft model is close to the target model but not equal the acceptance probability can equal 1. This is in contrast to the case of $K = 1$, where it is known that the acceptance probability can only equal 1 when $p(\cdot)$ and $q(\cdot)$ are identical distributions (Sun et al., 2024b). Furthermore to the best of our knowledge, previously proposed schemes for the multi-draft setting, such as SpecTr (Sun et al., 2024b) and SpecInfer (Miao et al., 2024) based on modified rejection sampling also require $p(\cdot) = q(\cdot)$ for the acceptance probability to be 1. Theorem 1 is interesting in the context of our two-step architecture in Fig. 1. In this case, the output of importance weighted sampling block Y matches the target distribution $q(\cdot)$ and the second step involving speculative sampling is not needed.

Example 1. *Consider* $\Omega = \{1, 2\}$ *and let the draft and target distributions be given by* $\mathbf{p} = (p_1, p_2)$ *and* $\mathbf{q} = (q_1, q_2)$ *respectively. We assume* $K = 2$ *drafts. In this case* (4) *reduces to* $q_1 \geq p_1^2$ *and* $q_2 \geq p_2^2$*. If* $p_1 = p_2 = 0.5$ *then it follows that* $P^\star(acc) = 1$ *if and only if* $0.25 \leq q_1 \leq 0.75$*. In contrast for the optimal scheme for* $K = 1$ *draft we have* $P^\star(acc) = 1$ *only when* $q_1 = q_2 = 0.5$*.*

The proof of Theorem 2 in Appendix E involves analyzing the output distribution $p_I(\cdot)$ of the Importance Weighted Sampling Scheme in Theorem 1 and demonstrating that a feasible choice of $\beta_y(x_1, x_2)$ exists and sets $p_I(\cdot) = q(\cdot)$ when the condition (4) is satisfied. The proof is based on the Fourier-Motzkin (FM) elimination technique (Ziegler, 2012). However a direct application of such a technique to satisfy the constraints $q(i) = p_I(i)$ for each $i \in \Omega$ becomes intractable. Our key idea is to demonstrate that instead considering a relaxation of the form $q(i) \geq p_I(i)$ leads to the same solution as the equality constraints and is amenable to analysis using Fourier-Motzkin elimination. We explain this further with an example involving $\Omega = \{1, 2, 3\}$ in Appendix E.

The problem of determining whether a system of linear equations has a non-negative solution has been studied previously in the literature, with (Chernikova, 1964; Dines, 1926) providing an algorithm. In Appendix F we also discuss a geometric viewpoint involving polyhedral cones. We explain how the double-description method (Fukuda & Prodon, 1995) for finding dual representations of polyhedral cones can be used to numerically verify the necessary and sufficient condition for the acceptance probability to equal 1. In fact this approach was used to verify analogous conditions to Theorem 2 for up to $K = 6$ drafts and all alphabets of size $|\Omega| \leq 14$ although we only provide an analytical proof of the condition for $K = 2$ drafts in this paper. Our final result is an explicit expression for the optimal acceptance probability for the case of $K = 2$ drafts.

Theorem 3. *For* $K = 2$ *drafts and for a draft distribution* $p(\cdot)$ *and target distribution* $q(\cdot)$ *and arbitrary token alphabet* Ω*, the acceptance probability* $P^\star(\mathrm{acc})$ *for the optimal token level selection rule is given by:*

$$P^\star(\mathrm{acc}) = \min_{\mathcal{S} \subseteq \Omega} \left\{ \sum_{s \in \mathcal{S}} q(s) + \left(\sum_{s \in \mathcal{S}^c} p(s)\right)^2 + 2\left(\sum_{s \in \mathcal{S}} p(s)\right)\left(\sum_{s \in \mathcal{S}^c} p(s)\right) \right\}, \tag{5}$$

where $\mathcal{S}^c = \Omega \setminus \mathcal{S}$ *is the complement of* $\mathcal{S}$*.*

To the best of our knowledge the result in Theorem 3 was

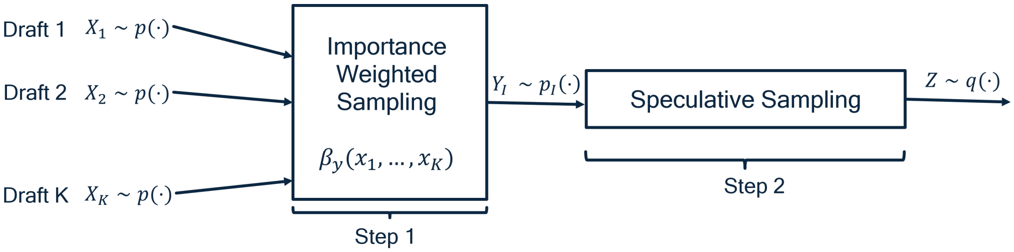
Draft 1 $X_1 \sim p(\cdot)$
Draft 2 $X_2 \sim p(\cdot)$
Draft K $X_K \sim p(\cdot)$
Importance Weighted Sampling
$\beta_y(x_1, \ldots, x_K)$
Step 1
$Y_I \sim p_I(\cdot)$
Speculative Sampling
$Z \sim q(\cdot)$
Step 2

Figure 1. Optimal Approach for Multi-Draft Speculative Sampling

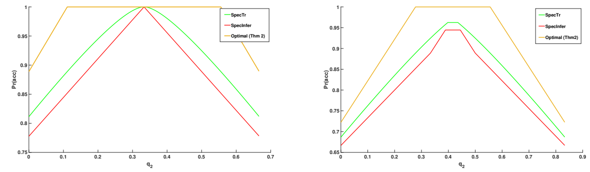

Figure 2. Numerical evaluation of $Pr(\text{accept})$ for the optimal scheme (Theorem 3) as well as two baseline schemes - SpecTr (Sun et al., 2024b) and SpecInfer (Miao et al., 2024). For sake of illustration we select alphabet $\Omega = \{1, 2, 3\}$ and $\mathbf{p} = [1/3, 1/3, 1/3]$. The left plot sets $\mathbf{q} = [1/3, q_2, 2/3 - q_2]$ while the right plot sets $\mathbf{q} = [1/6, q_2, 5/6 - q_2]$ where q_2 is varied on the x-axis.

not known before. Upper bounds on $P^\star(\text{acc})$ are presented in (Sun et al., 2024b), which are not necessarily tight. In contrast (5) provides an exact expression for the acceptance probability for the case of $K = 2$ drafts when X_1 and X_2 are independently sampled from $p(\cdot)$. Also it can be easily verified that the result in Theorem 3 implies the result in Theorem 2. The proof of Theorem 3, presented in Appendix G, applies the Fourier-Motzkin elimination to the linear program presented in Theorem 1 to characterize an analytical solution in the case of $K = 2$ drafts. The proof builds upon the proof of Theorem 2 but requires elimination of additional variables.

We provide numerical evaluation of the optimal acceptance probability in Fig. 2. For sake of illustration we assume that Ω is of size three, and assume $\mathbf{p} = [1/3, 1/3, 1/3]$. We consider $\mathbf{q} = [1/3, q_2, 2/3 - q_2]$ in the left plot and $\mathbf{q} = [1/6, q_2, 5/6 - q_2]$ in the right plot. The value of q_2 is varied on the x-axis. We compare the optimal acceptance probability in Theorem 3 with two baseline schemes SpecTr (Sun et al., 2024b) and SpecInfer (Miao et al., 2024). We observe that the optimal acceptance probability can equal 1 for a wide range of q_2. This is consistent with Theorem 2. In contrast the baseline schemes seem to achieve an acceptance probability of 1 only in the special case when $q_2 = 1/3$ so that $\mathbf{q} = [1/3, 1/3, 1/3]$.

Although we have only focused on the case of $K = 2$ drafts in this section, we believe natural counterparts can be developed for the case of $K > 2$, albeit with more involved notations. We also believe that analogous results can be established when the K drafts have different distributions.

3. Experimental Results

Setup. We conduct experiments using an instance of A100 GPU with 80GB memory and use the Llama2-chat, 7B model as the target model (Touvron et al., 2023), and a custom Llama-chat 115m model as the draft model (trained following (Goel et al., 2024)). Our method and baselines are evaluated on the XSum task (Narayan et al., 2018). The details of the experimental setting can be found at C.

Table 1. Comparison of Block Efficiency, Token Rate and ROUGE-2 for different schemes using the XSUM task dataset, averaged over 5 random seeds.

Scheme	Efficiency	Token Rate	ROUGE-2
Auto-Regressive	1.0	36.26	
SpecTr	2.36	42.70	0.2187
SpecInfer	2.36	42.90	0.2191
Stand-Alone IS	2.38	41.17	0.2210
Hybrid IS + SpecTr	2.37	43.34	0.2178
Hybrid IS + SpecInfer	2.39	43.59	0.2163

In Table 1 we present the results on the block efficiency, token rate and ROUGE-2 scores for different schemes. We consider $K = 2$ draft models and generate $L = 5$ draft tokens in each call. In general our proposed Stand-Alone IS provides competitive performance, achieving better block efficiency than SpecTr and SpecInfer. The Stand-Alone IS is based on a truncated version of linear program (see Appendix B.1) where we set $s = 5$. Furthermore, schemes that are a hybrid between importance sampling and baseline schemes (see Section B.3) achieve superior block efficiency and token rate over the baseline schemes and thus seem to be a promising avenue for improving them. In our implementation if the effective alphabet size of the of either the target or the draft model is at-most 2, we perform weighted IS. As expected, the ROUGE-2 scores are similar between the different schemes, as all methods perform exact sampling of the target model.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bondarenko, Y., Nagel, M., and Blankevoort, T. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36, 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., and Dao, T. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv: 2401.10774*, 2024.

Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Chernikova, N. Algorithm for finding a general formula for the non-negative solutions of system of linear equations. *USSR Computational Mathematics and Mathematical Physics*, 4(4):151–158, 1964.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Conover, Mike and Hayes and others. Free dolly: Introducing the world's first open and commercially viable Instruction-Tuned LLM - the databricks blog. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, April 2023. Accessed: 2024-5-31.

Dantzig, G. B. and Curtis Eaves, B. Fourier-motzkin elimination and its dual. *Journal of Combinatorial Theory, Series A*, 14(3):288–297, 1973.

Dines, L. L. On positive solutions of a system of linear equations. *Annals of Mathematics*, pp. 386–392, 1926.

Frantar, E. and Alistarh, D. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Fukuda, K. and Prodon, A. Double description method revisited. In *Franco-Japanese and Franco-Chinese conference on combinatorics and computer science*, pp. 91–111. Springer, 1995.

Ge, T., Xia, H., Sun, X., Chen, S.-Q., and Wei, F. Lossless acceleration for seq2seq generation with aggressive decoding. *arXiv preprint arXiv:2205.10350*, 2022.

Goel, R., Gagrani, M., Jeon, W., Park, J., Lee, M., and Lott, C. Direct alignment of draft model for speculative decoding with chat-fine-tuned llms. *arXiv preprint arXiv:2403.00858*, 2024.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jaszczur, S., Chowdhery, A., Mohiuddin, A., Kaiser, L., Gajewski, W., Michalewski, H., and Kanerva, J. Sparse is enough in scaling transformers. *Advances in Neural Information Processing Systems*, 34:9895–9907, 2021.

Jeon, W., Gagrani, M., Goel, R., Park, J., Lee, M., and Lott, C. Recursive speculative decoding: Accelerating llm inference via sampling without replacement. *arXiv preprint arXiv:2402.14160*, 2024.

Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Meister, C., Pimentel, T., Wiher, G., and Cotterell, R. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Zhang, Z., Wong, R. Y. Y., Zhu, A., Yang, L., Shi, X., 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, Volume 3*, pp. 932–949, 2024.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Shazeer, N. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Stern, M., Shazeer, N., and Uszkoreit, J. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.

Sun, Z., Ro, J. H., Beirami, A., and Suresh, A. T. Optimal block-level draft verification for accelerating speculative decoding. *arXiv preprint arXiv:2403.10444*, 2024a.

Sun, Z., Suresh, A. T., Ro, J. H., Beirami, A., Jain, H., and Yu, F. Spectr: Fast speculative decoding via optimal transport. *Advances in Neural Information Processing Systems*, 36, 2024b.

Tokdar, S. T. and Kass, R. E. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhou, Y., Lyu, K., Rawat, A. S., Menon, A. K., Rostamizadeh, A., Kumar, S., Kagy, J.-F., and Agarwal, R. Distillspec: Improving speculative decoding via knowledge distillation. *arXiv preprint arXiv:2310.08461*, 2023.

Ziegler, G. M. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.

Zny. Skeleton algorithm, 2018. Available at http://www.uic.unn.ru/ zny/skeleton/.

A. Background and Related Works

Auto-regressive sampling from LLMs is inherently sequential and memory-bound (Shazeer, 2019). Several approaches have been proposed in the literature to accelerate LLM inference (Shazeer, 2019; Jaszczur et al., 2021; Frantar et al., 2022; Frantar & Alistarh, 2023; Stern et al., 2018; Chen et al., 2023; Leviathan et al., 2023; Jeon et al., 2024; Sun et al., 2024b; Miao et al., 2024). Model compression techniques, such as quantization (Frantar et al., 2022; Bondarenko et al., 2024) and sparsification (Jaszczur et al., 2021; Frantar & Alistarh, 2023) have been shown to reduce the overall complexity of LLMs at the expense of some degradation in decoding quality.

For lossless LLM inference acceleration, speculative decoding (Chen et al., 2023; Leviathan et al., 2023; Stern et al., 2018) has emerged as a promising and orthogonal alternative. Earlier works on greedy decoding can draft and predict multiple tokens by augmenting the base LLM (Stern et al., 2018) or aggressive decoding (Ge et al., 2022). However, LLM text-generation often requires sampling with non-zero temperature from the generated logits. To that end, speculative decoding (Chen et al., 2023; Leviathan et al., 2023) was proposed. In speculative decoding, auto-regressive sampling is delegated to a smaller language model (draft model) that generates multiple candidate tokens. The LLM (target model) is then used to score all the tokens in the draft *in parallel*, and the draft tokens are verified through a sequence of token-level rejection sampling. Speculative decoding guarantees that the final sequence follows the same distribution as that of the target model. The performance of speculative methods highly depends on the choice of the draft model. Zhou et al. (2023) use knowledge distillation (Hinton et al., 2015) to better align the draft and target models which results in higher token acceptance rates.

More recently, the works of (Sun et al., 2024b; Miao et al., 2024; Jeon et al., 2024) extend speculative decoding to the multi-draft setting where the draft model(s) generate multiple token sequences per time-step. Specifically, Sun et al. (2024b) formulate the token-level draft selection problem as a discrete optimal transport problem with membership cost and propose SpecTr: a new decoding algorithm that allows for multiple candidates for each token in the draft. A related setting is also studied in (Miao et al., 2024; Jeon et al., 2024) where the authors consider a token tree based construction for improving the draft sequences as well as a token-level selection method different form (Sun et al., 2024b) . Instead of using a dedicated draft model, Cai et al. (2024) propose augmenting the target model with extra decoding heads that can concurrently draft multiple tokens. The extra heads are fine-tuned using parameter-efficient methods, and can be added to any pre-trained target model. Orthogonally, Sun et al. (2024a) study block-level verification in the single-draft setting as a block-level optimal transport problem. They propose a computationally-efficient algorithm that optimally solves the block-level transport problem, and report speedups over prior token-level verification (Leviathan et al., 2023).

B. Faster Importance Weighted Speculative Sampling

In practice the distribution of the target and draft model is often concentrated over a small number of tokens. It has also been observed that sampling from a high probability set (such as the top-k highest probability tokens) or the top-p set (the set of high probability tokens with aggregate probability exceeding a threshold) leads to more coherent outputs (Meister et al., 2023). After such top-p sampling the effective alphabet size, i.e., the number of tokens with non-zero probability is generally small. In Fig. 3 we show the histogram of effective alphabet size for a Llama 7B target model and Llama 115m draft model in an experiment involving 100 random prompts from the XSUM task dataset (Conover, Mike and Hayes and others, 2023). This motivates us to develop some approaches for speeding up our proposed solution by reducing the number of variables required in optimization.

We focus on the case of $K = 2$ drafts and explain how to extend the approach to the general case. Let X_1 and X_2 denote the input tokens and Y denote the selected token. When $X_1 = X_2 = i$ we have that $\beta_i(i, i) = 1$. Furthermore as discussed previously, due to symmetry we have $\beta_y(i, j) = \beta_y(j, i)$. It is more convenient to introduce a new set of variables $w_{i,j}$ which are defined as:

$$w_{i,j} = \Pr(Y = i \mid \{X_1, X_2\} = \{i, j\}), \tag{6}$$

i.e., $w_{i,j}$ denotes the probability that the output token is i given that we see the pair $\{i, j\}$ at the input in any order. For any given choice of $w_{i,j}$ the distribution of Y can be computed as follows:

$$p_I(k) = \Pr(Y = k) = p_k^2 + \sum_{i=1, i \neq k}^{n} 2 p_i p_k w_{i,k} \tag{7}$$

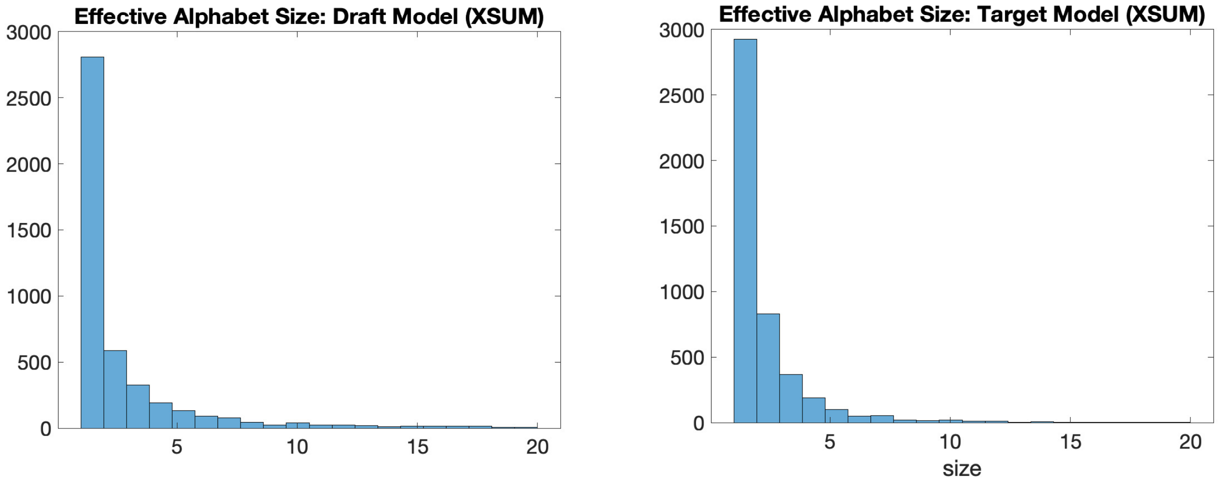

Figure 3. Histogram of Effective Alphabet Size (Number of Tokens with non-zero probability) after top-p sampling with $p = 0.95$. The left plot shows the histogram for the draft model, while the right plot shows the histogram of the target model. We use the XSUM task dataset (Conover, Mike and Hayes and others, 2023) and Llama 2 7B target model (chat version) and a custom Llama 115m draft model following (Goel et al., 2024).

where we use $\mathbf{p} = (p_1, \ldots, p_n)$ as the probability vector for the draft model and $\mathbf{q} = (q_1, \ldots, q_n)$ as the probability vector for the target model. Our aim is to maximize the following objective $\sum_{i=1}^{n} \min\left(p_I(i), q_i\right)$ over the variables $w_{i,j}$ that satisfy $0 \leq w_{i,j} \leq 1$ and $w_{i,j} + w_{j,i} = 1$. This optimization can be cast as a linear programming problem with $O(n^2)$ variables which may be slow in practice. With this formulation as a starting point we discuss two approaches to reduce the number of variables needed to optimize.

B.1. Truncated LP

The idea behind the proposed *truncated LP* scheme is that the linear programming solution will not be sensitive to most choices of $w_{i,j}$ when the target and draft distributions are concentrated over a few values. As a result one can heuristically set most of the variables. Assume that the vocabulary $\Omega = \{1, 2, \ldots, n\}$ has the tokens sorted in decreasing order i.e., $q_1 - p_1^2 \geq q_2 - p_2^2 \ldots \geq q_n - p_n^2$. We partition Ω into two sets $\Omega_1 = \{1, 2, \ldots, s\}$ and $\Omega_2 = \{s+1, \ldots, n\}$, where s is a free parameter to select. We fix a subset of weights as follows:

$$w_{i,j} = \begin{cases} 1, & i \in \Omega_1, j \in \Omega_2 \\ 1, & i \in \Omega_2, j \in \Omega_2, i < j \end{cases} \tag{8}$$

while we leave the weights $w_{i,j}$ for $i < j$ and $i, j \in \Omega_1$ as free parameters. The intuition behind the choice of weights in (8) is that in these cases we prefer token i over token j to increase $p_I(i)$ further so as to decrease the difference between q_i and $p_I(i)$. Note that:

$$p_I(k) = \begin{cases} p_k^2 + \sum_{i=1, i \neq k}^{s} 2 p_i p_k w_{i,k} + \sum_{i=s+1}^{n} 2 p_i p_k, & k \in \Omega_1 \\ p_k^2 + \sum_{i=k+1}^{n} 2 p_i p_k, & k \in \Omega_2 \end{cases} \tag{9}$$

The objective to maximize reduces to $\sum_{k=1}^{s} \min(p_I(k), q_k)$ over the variables $w_{i,j}$. Thus the number of variables is reduced to $O(s^2)$. We further show in Appendix H that if $P^{\star}(\mathrm{acc})$ is the optimal acceptance probability associated by applying the linear program over all $O(n^2)$ weight variables and $\tilde{P}(\mathrm{acc})$ is the acceptance probability for the truncated program then:

$$\tilde{P}(\mathrm{acc}) \geq P^{\star}(\mathrm{acc}) - \sum_{x \in \Omega_2} \left(q(x) - p^2(x)\right)^{+} \tag{10}$$

Thus if Ω_2 is selected so that the penalty term is small then the decrease in the acceptance probability can be kept small. In the experiments we observed that for well-trained target models the drop in accuracy is negligible even for small values of s. Thus by appropriately truncating the number of variables to optimize in the linear program we expect to have a faster implementation.

Algorithm 1 Truncated LP

```
Input: Threshold $s$, Input tokens $X_1, X_2$ sampled independently from $p(\cdot)$
Output: Selected token $Y_I$, output distribution $p_I(\cdot)$.
Order vocabulary $\Omega = \{1, 2, \dots, n\}$, sorted in decreasing order with $(q_i - p_i^2)$.
Set $\Omega_1 = \{1, \dots, s\}$ and $\Omega_2 = \{s+1, \dots, n\}$.
For $i, j \in \Omega_2$ or $i \in \Omega_1$ and $j \in \Omega_2$ set $w_{i,j}$ in (8)
For $i, j \in \Omega_1$, compute $w_{i,j}$ as a solution to a linear program:
    • Maximize: $\sum_{i=1}^{s} \min(q_i, p_I(i))$, where $p_I(i)$ is defined in (9)
    • Constraints: $w_{i,j} \geq 0$, $w_{i,j} + w_{j,i} = 1$
Compute $p_I(\cdot)$ according to (9).
if $X_1 = X_2$ then
    Set $Y_I = X_1$.
else
    Let $\{X_1, X_2\} = \{i, j\}$ and $i < j$.
    Set $Y_I = i$ with probability $w_{i,j}$, and $Y_I = j$ otherwise.
end if
```

B.2. Truncated Vocabulary

Let $\Omega_0 \subseteq \Omega$ be a high probability subset of Ω and let $q(\Omega_0) = 1 - \varepsilon_q$ and $p(\Omega_0) = 1 - \varepsilon_p$. One way of selecting Ω_0 is to select the K most likely tokens of Ω under the $\mathbf{p}$ and $\mathbf{q}$ distributions and taking their union. Let $\tilde{p}(\cdot)$ and $\tilde{q}(\cdot)$ be obtained by truncating $p(\cdot)$ and $q(\cdot)$ to Ω_0 and re-normalizing them.

We modify our proposed importance weighted sampling scheme as follows. Given an input $\mathcal{S} = \{X_1, X_2\}$, we remove tokens that do not belong to Ω_0 and let $\tilde{S} \subseteq \mathcal{S}$ be the resulting set. The input tokens after this filtering operation belong to Ω_0 and have a distribution of $\tilde{p}(\cdot)$. We perform Importance Weighted Sampling over $\tilde{S}$ followed by speculative sampling as in Fig. 1.

The complexity of the proposed scheme is dictated by the size of Ω_0. In practice if the distributions $\mathbf{p}$ and $\mathbf{q}$ are concentrated over a few values we can set Ω_0 to be small. Furthermore we show in the Appendix I that if $\tilde{P}(\mathrm{acc})$ denotes the acceptance probability of the proposed scheme and $P^{\star}(\mathrm{acc})$ denotes the acceptance probability of the optimal scheme over Ω then we have:

$$\tilde{P}(\mathrm{acc}) \geq (1 - 2\varepsilon_p)\left(P^{\star}(\mathrm{acc}) - \varepsilon_q\right). \tag{11}$$

Note that we can combine both the truncated vocabulary and truncated LP schemes. Thus given an alphabet Ω, we first consider a truncated vocabulary Ω_0. We then apply the truncated LP as discussed in Section B.1 and follow with speculative sampling as using the target distribution $q(\cdot)$.

Remark 1. *Although our discussion has focused on the case when both samples are drawn from the same draft distribution $p(\cdot)$, our approach can also be extended to the case when they are sampled from a different distribution, as outlined in the supplementary material.*

Remark 2. *To tackle the case of $K > 2$ drafts we propose to group the input tokens into groups of size 2 and then apply the two-draft importance sampling scheme in a multi-stage manner. For example if $S = \{X_1, X_2, X_3\}$ and $K = 3$ we first apply the fast importance weighted sampling to the group $\{X_1, X_2\}$ to output an intermediate token Y_1 with distribution say $p_1(\cdot)$. Then we apply importance weighted sampling to the input (Y_1, X_3), where the tokens now have non-identical distributions, and produce an output token Y to which speculative sampling is applied.*

B.3. Hybrid Schemes

The weighted importance sampling scheme can be naturally combined with other baselines to combine the strengths of both approaches. In particular, when top-p truncation is applied the effective vocabulary size (i.e., the number of non-zero probability tokens) can be small (see Fig. 3). In such cases, the parameters in our linear programming framework can be

Algorithm 2 Truncated Vocabulary

```
Input: Top-K parameter: $K$ and Input tokens: $X_1$ and $X_2$ sampled independently form $p(\cdot)$
Output: Selected token $Y_I$, output distribution $p_I(\cdot)$.
$\Omega_p \leftarrow \top_K(\Omega)$, $\Omega_q \leftarrow \top_K(\Omega)$ {$\top_K(\cdot)$ denotes the top-K function}
$\Omega_0 = \Omega_p \cup \Omega_q$ denotes high probability set
$\tilde{p}(\cdot) \leftarrow \mathrm{Trunc}_{\Omega_0}(p(\cdot))$, $\tilde{q}(\cdot) \leftarrow \mathrm{Trunc}_{\Omega_0}(q(\cdot))$ {$\mathrm{Trunc}_{\Omega_0}(\cdot)$ denotes truncation operator}
For $i, j \in \Omega_0$, compute $w_{i,j}$ as a solution to a linear program:
    • Maximize: $\sum_{i \in \Omega_0} \min(\tilde{q}_i, \tilde{p}_I(i))$,
    • Constraints: $w_{i,j} \geq 0$, $w_{i,j} + w_{j,i} = 1$
Compute $\tilde{p}_I(\cdot)$ according to (9) with $p(\cdot) \leftarrow \tilde{p}(\cdot)$.
$\tilde{\mathcal{S}} = \mathcal{S} \cap \Omega_0$
Sample $Y$ from $\tilde{\mathcal{S}}$ using the weights $w_{i,j}$
```

efficiently computed and can achieve an improved acceptance probability. On the other hand, in cases when the effective vocabulary size is larger, it is more beneficial to perform fast decoding using one of the existing baselines. We explore the benefits of such approach in the experimental section next.

C. Experimental setup

In our experiments we consider top-p sampling with $p = 0.95$. We use a temperature of 0.9 for the target model and a temperature of 0.3 for the draft model. The different temperatures generate limited misalignment between the logits so that the differences between different schemes become evident. Note that the temperature scaling is performed after top-p selection. In the experiments, our first proposed scheme, dubbed as Stand-Alone IS, is a fast version of importance weighted sampling. We generate a high probability alphabet Ω_0, by selecting the $N = 5$ highest probability tokens for the draft and target distribution and taking the union over these. We also consider a truncated linear program where we set the threshold (see Sec. B) to $s = 5$. Nevertheless we note this implementation has not been optimized for run-time efficiency so the token rate is not a fair metric. We also consider two baseline schemes – SpecTr and SpecInfer (Miao et al., 2024) which have been proposed in the context of multi-draft sampling. In addition we consider a hybrid between our importance weighted sampling and the baselines. In particular our scheme defaults to the baselines if after following top-p sampling, both draft and target models have more than $\ell = 2$ tokens with non-zero probability. Otherwise we use importance sampling and an analytical solution for the importance weights.

D. Proof of Theorem 1

We will consider the case when there are K drafts i.e., $X_1, \ldots, X_K$ are sampled i.i.d. from a draft model with distribution $p(\cdot)$, while the target model has a distribution of $q(\cdot)$. We assume the alphabet $\Omega = \{1, 2, \ldots, M\}$ for some arbitrary M.

Analysis of Importance Weighted Sampling Scheme: We first consider the family of importance sampling schemes followed by speculative sampling and derive the acceptance probability. Assuming Y denotes the selected sample in importance sampling, let[1]:

$$\Pr(Y = y | X_1^K = x_1^K) = \begin{cases} \beta_y(x_1^K), & y \in \{x_1, \ldots, x_K\} \\ 0, & y \notin \{x_1, \ldots, x_K\} \end{cases} \tag{12}$$

where $\sum_y \beta_y(x_1, \ldots, x_K) = 1$ for each $x_1^K \in \Omega^K$ and $0 \leq \beta_y(x_1, \ldots, x_K) \leq 1$.

[1] We use the notation X_1^K as a short hand for $X_{1:K}$. Similarly we use x_1^K as a shrot hand for $x_{1:k}$

It follows that

$$\Pr(Y=y) = \sum_{x_1,\ldots,x_K \in \Omega} \beta_y(x_1^K) \cdot \prod_{i=1}^{K} p(x_i) \tag{13}$$

$$= \sum_{x_1,\ldots,x_K \in \Omega} \beta_y(x_1^K) \cdot \mathbb{I}_y(x_1,\ldots,x_K) \cdot \prod_{i=1}^{K} p(x_i) \tag{14}$$

where $\mathbb{I}_y(x_1,\ldots,x_K)$ denotes the indicator function that equals 1 if $y \in \{x_1,\ldots,x_K\}$ and equals 0 otherwise. Note that (14) follows since $\beta_y(x_1,\ldots,x_K) = 0$ if $y \notin \{x_1,\ldots,x_K\}$.

By applying speculative sampling to the selected sample X_I the probability of acceptance is given by:

$$P^{\mathrm{M-IS}}(\text{accept}=1) = \sum_{y\in\Omega} \min(q(y), \Pr(X_I = y)) \tag{15}$$

$$= \sum_{y\in\Omega} \min\left(q(y), \sum_{x_1,\ldots,x_K \in \Omega} \beta_y(x_1^K) \cdot \mathbb{I}_y(x_1,\ldots,x_K) \cdot \prod_{i=1}^{K} p(x_i)\right) \tag{16}$$

Thus within the proposed class of importance sampling schemes, we can formulate our objective as:

$$\max_{\{\beta_y(x_1^K)\}_{y,x_1,\ldots,x_K}} \left\{ \sum_{y\in\Omega} \min\left(q(y), \sum_{x_1,\ldots,x_K \in \Omega} \beta_y(x_1^K) \cdot \mathbb{I}_y(x_1,\ldots,x_K) \cdot \prod_{i=1}^{K} p(x_i)\right)\right\} \tag{17}$$

such that $0 \le \beta_y(x_1^K) \le 1$ for each $y, x_1, \ldots, x_K \in \Omega$, and

$$\sum_{x_1^K \in \Omega^K} \beta_y(x_1^K) = 1, \quad \forall y \in \Omega, \tag{18}$$

and furthermore

$$\beta_y(x_1^K) = 0, \quad y \notin \{x_1,\ldots,x_K\}. \tag{19}$$

Analysis of Optimal Solution: We now consider the problem of optimizing the acceptance probability for any given $p(\cdot)$ and $q(\cdot)$ in the general setting. Following the framework in (Sun et al., 2024b), we seek to find $p_{Y|X_1,\ldots X_K}(y|x_1,\ldots,x_K)$ for each $y, x_1, \ldots, x_K \in \Omega$ such that we maximize

$$\Pr(\text{accept}=1) = \Pr(Y \in \{X_1,\ldots,X_K\}) \tag{20}$$

subject to the marginal constraints on $P_Y(\cdot)$:

$$q(y) = \Pr(Y=y) = \sum_{x_1^K} \Pr(Y=y, X_1^K = x_1^K) = \sum_{x_1^k} p_{Y|X_1^K}(y|x_1^K) \prod_{i=1}^{K} p(x_i). \tag{21}$$

Next we consider:

$$\begin{aligned} q(y) &= \sum_{x_1^k \in \Omega^k} p_{Y|X_1^K}(y|x_1^K) \prod_{i=1}^{K} p(x_i) \\ &= \sum_{x_1^k \in \Omega^k} p_{Y|X_1^K}(y|x_1^K) \mathbb{I}_y(x_1,\ldots,x_K) \prod_{i=1}^{K} p(x_i) \\ &\quad + \sum_{x_1^k \in \Omega^k} p_{Y|X_1^K}(y|x_1^K) \bar{\mathbb{I}}_y(x_1,\ldots,x_K) \prod_{i=1}^{K} p(x_i) \qquad (22) \\ &\ge \sum_{x_1^k \in \Omega^k} p_{Y|X_1^K}(y|x_1^K) \mathbb{I}_y(x_1,\ldots,x_K) \prod_{i=1}^{K} p(x_i) \qquad (23) \end{aligned}$$

where $\bar{\mathbb{I}}_y(x_1,\ldots,x_K) = 1 - \mathbb{I}_y(x_1,\ldots,x_K)$ denotes the complement of $\mathbb{I}$. Now note that:

$$\begin{aligned}
&\Pr(Y \in \{X_1,\ldots,X_K\}) \\
&= \sum_{x_1^K \in \Omega^K} \Pr(Y \in \{X_1,\ldots,X_K\} \mid X_1^K = x_1^K) p(X_1^K = x_1^K) && (24)\\
&= \sum_{x_1^K \in \Omega_K} \sum_{y \in \Omega} p_{Y|X_1^K}(y|x_1^K)\mathbb{I}_y(x_1,\ldots,x_K)\left(\prod_{i=1}^{K} p(x_i)\right) && (25)\\
&= \sum_{y \in \Omega} \sum_{x_1^K \in \Omega_K} p_{Y|X_1^K}(y|x_1^K)\mathbb{I}_y(x_1,\ldots,x_K)\left(\prod_{i=1}^{K} p(x_i)\right) && (26)\\
&= \sum_{y \in \Omega} \min\left(q(y), \sum_{x_1^K \in \Omega_K} p_{Y|X_1^K}(y|x_1^K)\mathbb{I}_y(x_1,\ldots,x_K)\left(\prod_{i=1}^{K} p(x_i)\right)\right) && (27)
\end{aligned}$$

where we use (23) which implies that for any feasible $p_{Y|X_1^K}(y|x_1^K)$:

$$\sum_{x_1^k \in \Omega^k} p_{Y|X_1^K}(y|x_1^K)\mathbb{I}_y(x_1,\ldots,x_K)\prod_{i=1}^{K} p(x_i) \le q(y) \tag{28}$$

is satisfied.

Upper Bound on the optimal acceptance probability: We now establish an upper bound on (27) and show that it coincides with the acceptance probability optimized in the importance weighted sampling scheme (17).

For each $x_1^K \in \Omega^K$, let us define

$$D(x_1^K) = \sum_{y \in \Omega} p_{Y|X_1^K}(y|x_1^K)\mathbb{I}_y(x_1,\ldots,x_K) \tag{29}$$

and furthermore with $N(x_1,\ldots,x_K)$ denoting the number of unique elements in x_1^K,

$$\tilde{p}_{Y|X_1^K}(y|x_1^K) = \begin{cases} \frac{p_{Y|X_1^K}(y|x_1^K)}{D(x_1^K)}, & y \in \{x_1,\ldots,x_K\}, \qquad D(x_1^K) > 0 \\ \frac{1}{N(x_1\ldots,x_K)} & y \in \{x_1,\ldots,x_K\}, \qquad D(x_1^K) = 0, \\ 0 & y \notin \{x_1,\ldots,x_K\}. \end{cases} \tag{30}$$

Note by construction that for each $x_1^K \in \Omega^K$

$$\sum_{y \in \Omega} \tilde{p}_{Y|X_1^K}(y|x_1^K) = 1 \tag{31}$$

and

$$\tilde{p}_{Y|X_1^K}(y|x_1^K) = 0, \quad y \notin \{x_1,\ldots,x_K\} \tag{32}$$

and furthermore:

$$\tilde{p}_{Y|X_1^K}(y|x_1^K)\cdot\mathbb{I}_y(x_1,\ldots,x_K) \ge p_{Y|X_1^K}(y|x_1^K)\cdot\mathbb{I}_y(x_1,\ldots,x_K), \forall y, x_1,\ldots,x_K \in \Omega \tag{33}$$

Substituting (33) into (27) we have that for any feasible $p_{Y|X_1^K}(\cdot)$ there exists a $\tilde{p}_{Y|X_1^K}(\cdot)$ satisfying (31) and (32) such that:

$$Pr(Y \in \{X_1,\ldots,X_K\}) \le \sum_{y \in \Omega} \min\left(q(y), \sum_{x_1^K \in \Omega_K} \tilde{p}_{Y|X_1^K}(y|x_1^K)\mathbb{I}_y(x_1,\ldots,x_K)\left(\prod_{i=1}^{K} p(x_i)\right)\right) \tag{34}$$

It thus follows that that optimal acceptance probability in the general case is upper bounded by optimizing the (34) over $\tilde{p}_{Y|X_1^K}(y|x_1^K)$ satisfying (31) and (32). But this problem precisely coincides with the optimization in the proposed class of IS schemes as stated in (17)-(19), thus establishing the optimality of the latter.

D.1. Extension to Non-IID Setting

The proof in Theorem 1 assumed that $x_1, \ldots, X_K$ are sampled form the same underlying distribution $p(\cdot)$. Here we provie a natural extension when the sampled are still independently sampled from from a non-i.i.d. distribution i.e $X_i \sim p_i(\cdot)$.

Theorem 4. *Let $P^\star(\mathrm{acc})$ be the acceptance probability for the optimal token level selection rule when $\mathcal{S} \sim \prod_{i=1}^K p_i(X_i)$. Then we have*

$$P^\star(\mathrm{acc}) = \max_{\{\beta_y(x_1^K)\}} \left\{ \sum_{y\in\Omega} \min\left(q(y), \sum_{x_1,\ldots,x_K\in\Omega} \beta_y(x_1^K) \cdot \prod_{i=1}^K p_i(x_i) \right) \right\} \tag{35}$$

where the maximum is over $\beta_y(x_1^K)$ for each $\{x_1, \ldots, x_K, y\} \in \Omega$ such that $0 \leq \beta_y(x_1^K) \leq 1$, and

$$\sum_{x_1^K \in \Omega^K} \beta_y(x_1^K) = 1, \quad \forall y \in \Omega, \tag{36}$$

and furthermore

$$\beta_y(x_1^K) = 0, \quad y \notin \{x_1, \ldots, x_K\}. \tag{37}$$

Furthermore if $\{\beta_y^\star(x_1^K)\}$ denotes the parameters that achieve the maximum in (35), then $P^\star(\mathrm{acc})$ can be attained by a two step approach as follows: in the first step, given the list of input tokens $\{x_1, \ldots, x_K\}$, we apply Importance Weighted Speculative Sampling in Definition 1 with parameters $\beta_y^\star(x_1, \ldots, x_K)$ to output an intermediate token $y \in \{x_1, \ldots, x_K\}$; in the second step we apply a single-draft speculative sampling scheme (Chen et al., 2023; Leviathan et al., 2023) on the selected token y to generate the final output token.

The proof of Theorem 4 is identical to the proof of Theorem 1. We note that replacing the distribution of $\mathcal{S}$ from $\prod_{i=1}^K p(X_i)$ to $\prod_{i=1}^K p_i(X_i)$ does not affect any of the steps in the original proof.

E. Proof of Theorem 2

We first consider the special case of $\Omega = \{1, 2, 3\}$ to illustrate the key ideas. We then proceed with the proof.

Example 2. *Consider the case when $\Omega = \{1, 2, 3\}$ and let $\mathbf{p} = (p_1, p_2, p_3)$ and $\mathbf{q} = (q_1, q_2, q_3)$ denote the draft and target model distribution for the current token of interest. We again assume $K = 2$ drafts. Let $X_1 = i$ and $X_2 = j$ denote the pair of input tokens and Y denote the output of the importance weighted sampling scheme in step 1 in Fig. 1. Since $X_1 \sim p(\cdot)$ and $X_2 \sim p(\cdot)$ it is clear the the optimal TLSR does not depend on the order of X_1 and X_2 but only on the unordered set $\{X_1, X_2\}$ and let $\{i, j\}$ denote the realization. Let $\alpha_{i,j} = \Pr(Y = i, \{X_1, X_2\} = \{i, j\})$ denote the probability of the event that the (unordered) input tokens are $\{i, j\}$ and the output token is $Y = i$. Similarly let $\alpha_{j,i} = \Pr(Y = j, \{X_1, X_2\} = \{i, j\})$. Note that $\alpha_{i,i} = p_i^2$ must hold, as when $X_1 = X_2 = i$, clearly $Y = i$ in the Importance Weighted Sampling scheme. Note that $P^\star(acc) = 1$ requires that $\Pr(Y = i) = q_i$ for each $i \in \Omega$. This results in the following system of linear equations:*

$$q_1 = p_1^2 + \alpha_{1,2} + \alpha_{1,3}, \qquad q_2 = p_2^2 + \alpha_{2,1} + \alpha_{2,3}, \qquad q_3 = p_3^2 + \alpha_{3,1} + \alpha_{3,2} \tag{38}$$

subject to $\alpha_{i,j} + \alpha_{j,i} = 2p_ip_j$ and $0 \leq \alpha_{i,j} \leq 2p_ip_j$. We prove that (4) provides a necessary and sufficient condition that the above system of linear equations has a feasible solution.

Our initial attempt was to directly apply Fourer-Motzkin (FM) elimination technique (Ziegler, 2012; Dantzig & Curtis Eaves, 1973) to (38). However a direct application of FM elimination does not appear to be tractable for arbitrary sized alphabets, as the elimination of each variable introduces a large number of inequalities. Our key observation is that (38) is equivalent to the following relaxed set of inequalities:

$$q_1 \geq p_1^2 + \alpha_{1,2} + \alpha_{1,3}, \qquad q_2 \geq p_2^2 + \alpha_{2,1} + \alpha_{2,3}, \qquad q_3 \geq p_3^2 + \alpha_{3,1} + \alpha_{3,2} \tag{39}$$

with the same conditions on $\alpha_{i,j}$ as before. A solution to (38) exists if and only if a solution to the relaxation (39) exists. Indeed as a contradiction, suppose that a solution to (39) exists with strict inequality in one of conditions. Then summing *over all the inequalities and using $\alpha_{i,j} + \alpha_{j,i} = 2p_ip_j$ gives $q_1 + q_2 + q_3 > (p_1 + p_2 + p_3)^2$. However since $\mathbf{p}$ and $\mathbf{q}$ are probability vectors both sides should sum to 1, leading to a contradiction. Our second key idea is to augment the system of inequalities in (39) with the following additional inequalities:*

$$\begin{aligned} &q_1 + q_2 \geq (p_1 + p_2)^2 + \alpha_{1,3} + \alpha_{2,3}, \\ &q_1 + q_3 \geq (p_1 + p_3)^2 + \alpha_{1,2} + \alpha_{3,2}, \qquad q_2 + q_3 \geq (p_2 + p_3)^2 + \alpha_{2,1} + \alpha_{3,1} \end{aligned} \tag{40}$$

Note that the inequalities in (40) are redundant and follow by simply adding each pair of inequalities in (39) and using $\alpha_{i,j} + \alpha_{j,i} = 2p_ip_j$. However applying FM eliminations simultaneously over the expanded system of inequalities involving (39) and (40) is surprisingly tractable. In fact we show that applying FM elimination for eliminating each $\alpha_{i,j}$ (and by extension $\alpha_{j,i}$) simply involves dropping that variable in the system of inequalities (39) and (40). For example eliminating $\alpha_{1,2}$ (and simultaneously $\alpha_{2,1}$) in the first step is equivalent to:

$$q_1 \geq p_1^2 + \alpha_{1,3},\ q_2 \geq p_2^2 + \alpha_{2,3},\ q_3 \geq p_3^2 + \alpha_{3,1} + \alpha_{3,2} \tag{41}$$

$$q_1 + q_2 \geq (p_1 + p_2)^2 + \alpha_{1,3} + \alpha_{2,3},\ q_1 + q_3 \geq (p_1 + p_3)^2 + \alpha_{3,2},\ q_2 + q_3 \geq (p_2 + p_3)^2 + \alpha_{3,1} \tag{42}$$

Eliminating all $\alpha_{i,j}$ in this fashion establishes that a feasible solution exists if and only if $q_i \geq p_i^2$ and $q_j + q_k \geq (p_j + p_k)^2$ for $i, j, k \in \Omega$ and $j \neq k$. This is precisely the condition in (4) for an alphabet of size $|\Omega| = 3$.

We now proceed with the proof of the result.

Setting of Linear System of Equations and its Relaxation: Following the simplified notation in the main text for the case of $K = 2$ drafts, we let $\mathbf{q} = (q_1, \ldots, q_n)$ be the target model distribution and $\mathbf{p} = (p_1, \ldots, p_n)$ be the draft model distribution. Also recall that we define $\alpha_{i,j} = \Pr(Y = i, \{X_1, X_2\} = \{i, j\})$ as discussed in the main text. In order to match the output distribution $\Pr(Y = i)$ to the target distribution, we need to satisfy the following system of linear equations:

$$q_1 - p_1^2 = \alpha_{1,2} + \ldots + \alpha_{1,n} \tag{43}$$

$$q_2 - p_2^2 = \alpha_{2,1} + \ldots + \alpha_{2,n} \tag{44}$$

$$\vdots \tag{45}$$

$$q_n - p_n^2 = \alpha_{n,1} + \ldots + \alpha_{n,n-1} \tag{46}$$

where $\alpha_{i,j} \geq 0$ and $\alpha_{i,j} + \alpha_{j,i} = 2p_ip_j = 2p_{i,j}$ for each $i \neq j \in \{1, \ldots, n\}$.

We instead consider a relaxed system of inequalities:

$$q_1 - p_1^2 \geq \alpha_{1,2} + \ldots + \alpha_{1,n} \tag{47}$$

$$q_2 - p_2^2 \geq \alpha_{2,1} + \ldots + \alpha_{2,n} \tag{48}$$

$$\vdots \tag{49}$$

$$q_n - p_n^2 \geq \alpha_{n,1} + \ldots + \alpha_{n,n-1} \tag{50}$$

where $\alpha_{i,j} \geq 0$ and $\alpha_{i,j} + \alpha_{j,i} = 2p_ip_j = 2p_{i,j}$ for each $i \neq j \in \{1, \ldots, n\}$. We note that the system of inequalities (43)-(46) has a solution if and only if the system of inequalities (47)-(50) has a solution. Indeed, for contradiction assume that one of the inequalities in (47)-(50) is a strict inequality. Then summing over the left and right hand sides and using $\alpha_{i,j} + \alpha_{j,i} = 2p_ip_j$ we get that

$$\sum_{i=1}^{n} q_i > \left(\sum_{i=1}^{n} p_i\right)^2, \tag{51}$$

which is a contradiction as both sides sum to 1. Thus it suffices to consider the system of linear inequalities.

Augmented System of Inequalities: Instead of the original system of inequalities (47)-(50), we consider an augmented system of inequalities defined as follows.

Lemma 1. *Our original system (47)-(50) has a solution if an only if the following system has a solution:*

$$\sum_{s\in\mathcal{S}} q_s - \left(\sum_{s\in\mathcal{S}} p_s\right)^2 \geq \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} \alpha_{s,t} \quad \forall \mathcal{S}\subseteq\{1,\ldots,n\} \tag{52}$$

for $\alpha_{s,t}\geq 0$ and $\alpha_{s,t}+\alpha_{t,s}=2p_{s,t}$ for $s,t\in\{1,2,\ldots,n\}$ with $s\neq t$.

To establish this, we use (47)-(50) and sum over $s\in\mathcal{S}$:

$$\begin{aligned}
\sum_{s\in\mathcal{S}}\left(q_s-p_s^2\right) &\geq \sum_{s\in\mathcal{S}}\sum_{j=1,j\neq s}^{n}\alpha_{s,j} && (53)\\
&= \sum_{s\in\mathcal{S}}\left(\sum_{t\in\mathcal{S}^c}\alpha_{s,t}+\sum_{t\in\mathcal{S}\setminus\{s\}}\alpha_{s,t}\right) && (54)\\
&= \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c}\alpha_{s,t}+\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}\setminus\{s\}}\alpha_{s,t} && (55)\\
&= \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c}\alpha_{s,t}+\sum_{(s,t)\in\mathcal{S}\times\mathcal{S},t>s}(\alpha_{s,t}+\alpha_{t,s}) && (56)\\
&= \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c}\alpha_{s,t}+\sum_{(s,t)\in\mathcal{S}\times\mathcal{S},t>s}2p_{s,t} && (57)
\end{aligned}$$

It follows that:

$$\begin{aligned}
\sum_{s\in\mathcal{S}}q_s-\sum_{s\in\mathcal{S}}p_s^2-\sum_{(s,t)\in\mathcal{S}\times\mathcal{S},t>s}2p_{s,t} &\geq \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c}\alpha_{s,t} && (58)\\
\Rightarrow \sum_{s\in\mathcal{S}}q_s-\left(\sum_{s\in\mathcal{S}}p_s\right)^2 &\geq \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c}\alpha_{s,t} && (59)
\end{aligned}$$

as required. The other inclusion follows by simply setting $\mathcal{S}=\{i\}$ for each i.

Induction Argument We will prove the following by induction.

Lemma 2. *Let*

$$\mathcal{V}_r=\{(i_1,j_1),(j_1,i_1),\ldots,(i_r,j_r),(j_r,i_r)\} \tag{60}$$

denote the indices (with $i_k<j_k$ for all $k=1,\ldots,r$) of the variables eliminated after r rounds of FM elimination. Then the remaining constraints are given by:

$$\sum_{s\in\mathcal{S}} q_s - \left(\sum_{s\in\mathcal{S}} p_s\right)^2 \geq \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r), \quad \forall \mathcal{S}\subseteq\{1,\ldots,n\} \tag{61}$$

Remark 3. *When all the variables have been eliminated the right hand side in (61) will equal 0 for any choice of $\mathcal{S}\subseteq\{1,2,\ldots,n\}$ and we will recover the result Theorem 2.*

Note that the base case with $\mathcal{V}_r=\{\cdot\}$ immediately follows from (52). We will assume that the variables α_{i_q,j_q} and α_{j_q,i_q} are eliminated for $q\in\{1,\ldots,r-1\}$ and the associated Fourier-Motzkin (FM) conditions are given by:

$$\sum_{s\in\mathcal{S}} q_s - \left(\sum_{s\in\mathcal{S}} p_s\right)^2 \geq \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}), \quad \forall \mathcal{S}\subseteq\{1,\ldots,n\}. \tag{62}$$

At step r we eliminate the variable α_{i_r,j_r} and α_{j_r,i_r} and we will show that (61) is satisfied. In applying the FM elimination, we only need to consider those inequalities in (62) where either α_{i_r,j_r} or α_{j_r,i_r} appears on the right hand side. The remaining equations will not be affected in this step of FM elimination and replacing $\mathcal{V}_{r-1}$ with $\mathcal{V}_r$ will not have any effect there. Any such inequality will be associated with a choice of $\mathcal{S}$ where either both i_r and j_r belong to $\mathcal{S}$ or neither i_r and j_r belong to $\mathcal{S}$. Thus we have:

$$\begin{aligned}\sum_{s\in\mathcal{S}} q_s - \left(\sum_{s\in\mathcal{S}} p_s\right)^2 \geq \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r),\\ \forall\mathcal{S}\subseteq\{1,\ldots,n\} : i_r\in\mathcal{S}\ \&\ j_r\in\mathcal{S} \text{ or } i_r\notin\mathcal{S}\ \&\ j_r\notin\mathcal{S}.\end{aligned}\tag{63}$$

The FM elimination will only consider those inequalities in (62) where either α_{i_r,j_r} or α_{j_r,i_r} appears in the right hand side. The inequalities where α_{i_r,j_r} appears on the right hand side is associated with those subsets $\mathcal{S}_1$ of $\{1,\ldots,n\}$ where $i_r\in\mathcal{S}_1$ and $j_r\notin\mathcal{S}_1$. Likewise the inequalities in (62) where α_{j_r,i_r} is appears on the right hand side are associated those subsets $\mathcal{S}_2\in\{1,2,\ldots,n\}$ where $j_r\in\mathcal{S}_2$ and $i_r\notin\mathcal{S}_2$. Thus the FM elimination applied to variables α_{i_r,j_r} and α_{j_r,i_r} will consider the following system of equations:

$$\begin{aligned}&\sum_{s\in\mathcal{S}_1} q_s - \left(\sum_{s\in\mathcal{S}_1} p_s\right)^2 \geq \sum_{s\in\mathcal{S}_1}\sum_{t\in\mathcal{S}_1^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}),\\ &\qquad\forall\mathcal{S}_1\subseteq\{1,\ldots,n\}, i_r\in\mathcal{S}_1, j_r\notin\mathcal{S}_1, &(64)\\ &\sum_{s\in\mathcal{S}_2} q_s - \left(\sum_{s\in\mathcal{S}_2} p_s\right)^2 \geq \sum_{s\in\mathcal{S}_2}\sum_{t\in\mathcal{S}_2^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}),\\ &\qquad\forall\mathcal{S}_2\subseteq\{1,\ldots,n\}, j_r\in\mathcal{S}_2, i_r\notin\mathcal{S}_2, &(65)\\ &\alpha_{i_r,j_r}+\alpha_{j_r,i_r} = 2p_{i_r,j_r} &(66)\\ &\alpha_{i_r,j_r}\geq 0, \alpha_{j_r,i_r}\geq 0. &(67)\end{aligned}$$

Accounting for (67) and using the fact that $\mathcal{V}_r = \mathcal{V}_{r-1}\cup\{(i_r,j_r),(j_r,i_r)\}$ we immediately have that:

$$\begin{aligned}&\sum_{s\in\mathcal{S}_1} q_s - \left(\sum_{s\in\mathcal{S}_1} p_s\right)^2 \geq \sum_{s\in\mathcal{S}_1}\sum_{t\in\mathcal{S}_1^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r),\\ &\qquad\forall\mathcal{S}_1\subseteq\{1,\ldots,n\}, i_r\in\mathcal{S}_1, j_r\notin\mathcal{S}_1, &(68)\\ &\sum_{s\in\mathcal{S}_2} q_s - \left(\sum_{s\in\mathcal{S}_2} p_s\right)^2 \geq \sum_{s\in\mathcal{S}_2}\sum_{t\in\mathcal{S}_2^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r),\\ &\qquad\forall\mathcal{S}_2\subseteq\{1,\ldots,n\}, j_r\in\mathcal{S}_2, i_r\notin\mathcal{S}_1. &(69)\end{aligned}$$

In addition the FM elimination procedure is required to combine every possible inequality in (64) with every possible inequality in (65) and eliminate α_{i_r,j_r} and α_{j_r,i_r} by applying (66). For a specific choice of $\mathcal{S}_1$ and $\mathcal{S}_2$ the inequality we consider is of the form:

$$\begin{aligned}&\sum_{s\in\mathcal{S}_1} q_S + \sum_{s\in\mathcal{S}_2} q_s - \left(\sum_{s\in\mathcal{S}_1} p_s\right)^2 - \left(\sum_{s\in\mathcal{S}_2} p_s\right)^2\\ &\quad\geq \sum_{s\in\mathcal{S}_1}\sum_{t\in\mathcal{S}_1^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) + \sum_{s\in\mathcal{S}_2}\sum_{t\in\mathcal{S}_2^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}).\end{aligned}\tag{70}$$

We will show that this inequality is redundant as it is dominated by the set of inequalities in (63). Let $\mathcal{R} = \mathcal{S}_1 \cap \mathcal{S}_2$ and $\mathcal{T} = \mathcal{S}_1 \cup \mathcal{S}_2$. Note that $i_r \notin \mathcal{R}$ and $j_r \notin \mathcal{R}$. Now consider the left hand side of (70).

$$\begin{aligned}
&\sum_{s \in \mathcal{S}_1} q_s + \sum_{s \in \mathcal{S}_2} q_s - \left(\sum_{s \in \mathcal{S}_1} p_s\right)^2 - \left(\sum_{s \in \mathcal{S}_2} p_s\right)^2 \\
&= \sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} q_s + \sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} q_s + 2\sum_{s \in \mathcal{R}} q_s - \left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s + \sum_{s \in \mathcal{R}} p_s\right)^2 - \left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s + \sum_{s \in \mathcal{R}} p_s\right)^2 \qquad (71) \\
&= \sum_{s \in \mathcal{T}} q_s + \sum_{s \in \mathcal{R}} q_s - \left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right)^2 - \left(\sum_{s \in \mathcal{R}} p_s\right)^2 - 2\left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{R}} q_s\right) \\
&\quad - \left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)^2 - \left(\sum_{s \in \mathcal{R}} p_s\right)^2 - 2\left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{R}} p_s\right) \qquad (72) \\
&= \sum_{s \in \mathcal{T}} q_s + \sum_{s \in \mathcal{R}} q_s - \left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right)^2 - \left(\sum_{s \in \mathcal{R}} p_s\right)^2 - \left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)^2 - 2\left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{R}} q_s\right) \\
&- 2\left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{R}} p_s\right) - 2\left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right) - 2\left(\sum_{s \in \mathcal{R}} p_s\right)^2 \\
&\quad + 2\left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right) \qquad (73) \\
&= \left\{\sum_{s \in \mathcal{T}} q_s - \left(\sum_{s \in \mathcal{T}} p_s\right)^2\right\} + \left\{\sum_{s \in \mathcal{R}} q_s - \left(\sum_{s \in \mathcal{R}} p_s\right)^2\right\} + 2\left(\sum_{s \in \mathcal{S}_2 \setminus \mathcal{R}} p_s\right)\left(\sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} p_s\right) \qquad (74)
\end{aligned}$$

We now consider the right hand side of (70). We recall that with $\mathcal{T} = \mathcal{S}_1 \cup \mathcal{S}_2$ and $\mathcal{R}. = \mathcal{S}_1 \cap \mathcal{S}_2$ the following relations that can be easily established using Venn diagram of sets $\mathcal{S}_1$ and $\mathcal{S}_2$:

$$\begin{aligned}
\mathcal{S}_1^c &= \mathcal{T}^c \cup (\mathcal{S}_2 \setminus \mathcal{R}), \qquad \mathcal{T}^c \cap (\mathcal{S}_2 \setminus \mathcal{R}) = \{\cdot\} \qquad (75) \\
\mathcal{S}_2^c &= \mathcal{T}^c \cup (\mathcal{S}_1 \setminus \mathcal{R}), \qquad \mathcal{T}^c \cap (\mathcal{S}_1 \setminus \mathcal{R}) = \{\cdot\} \qquad (76) \\
\mathcal{R}^c &= \mathcal{T}^c \cup (\mathcal{S}_1 \setminus R) \cup (\mathcal{S}_2 \setminus R), \qquad (\mathcal{S}_1 \setminus R) \cap (\mathcal{S}_2 \setminus R) = \{\cdot\} \qquad (77) \\
\mathcal{T} &= \mathcal{R} \cup (\mathcal{S}_1 \setminus R) \cup (\mathcal{S}_2 \setminus R) \qquad (78)
\end{aligned}$$

Now consider the following:

$$\begin{aligned}
&\sum_{s \in \mathcal{S}_1} \sum_{t \in \mathcal{S}_1^c} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) \\
&= \sum_{s \in \mathcal{S}_1} \sum_{t \in \mathcal{T}^c} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) + \sum_{s \in \mathcal{S}_1} \sum_{t \in \mathcal{S}_2 \setminus \mathcal{R}} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) \qquad (79) \\
&= \sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} \sum_{t \in \mathcal{T}^c} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) + \sum_{s \in \mathcal{R}} \sum_{t \in \mathcal{T}^c} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) \\
&\quad + \sum_{s \in \mathcal{S}_1 \setminus \mathcal{R}} \sum_{t \in \mathcal{S}_2 \setminus \mathcal{R}} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) + \sum_{s \in \mathcal{R}} \sum_{t \in \mathcal{S}_2 \setminus \mathcal{R}} \alpha_{s,t} \cdot \mathbb{I}((s,t) \notin \mathcal{V}_{r-1}) \qquad (80)
\end{aligned}$$

where we use (75) in (79),

In a similar fashion we can express,

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}_2}\sum_{t\in\mathcal{S}_2^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \\
&=\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \\
&\quad+\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_1\backslash\mathcal{R}}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{S}_1\backslash\mathcal{R}}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})
\end{aligned}
\tag{81}
$$

Combing (80) and (81) and re-arranging terms, we get that:

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}_1}\sum_{t\in\mathcal{S}_1^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{S}_1}\sum_{t\in\mathcal{S}_1^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \\
&=\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \\
&+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{S}_1\backslash\mathcal{R}}\alpha_{s,t}+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \\
&+\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_1\backslash\mathcal{R}}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \qquad (82)\\
&=\sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{R}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \\
&+\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})+\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{t,s}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \qquad (83)\\
&=\sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{R}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r}) \\
&\quad+\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_2\backslash\mathcal{R}}(\alpha_{t,s}+\alpha_{s,t})\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \qquad (84)
\end{aligned}
$$

where we use (77) and (78) in (83) as well as the fact that $\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})=\mathbb{I}((t,s)\notin\mathcal{V}_{r-1})$ as the pair (s,t) and (t,s) is eliminated simultaneously. In (84) we use the fact that $\mathcal{T}$ contains both i_r and j_r while $\mathcal{R}$ contains neither i_{r_1} and j_{r-1} and hence α_{i_r,j_r} or α_{j_r,i_r} do not appear in the first two terms in (84) so that $\mathcal{V}_{r-1}$ can be replaced by $\mathcal{V}_r$. Combining (74) and (84) it follows that the FM elimination for our choice of $\mathcal{S}_1$ and $\mathcal{S}_2$ leads to:

$$
\begin{aligned}
&\left\{\sum_{s\in\mathcal{T}}q_s-\left(\sum_{s\in\mathcal{T}}p_s\right)^2\right\}+\left\{\sum_{s\in\mathcal{R}}q_s-\left(\sum_{s\in\mathcal{R}}p_s\right)^2\right\}+2\left(\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}p_s\right)\left(\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}p_s\right) \\
&\geq\sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r})+\sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{R}^c}\alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r}) \\
&\quad+\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{t\in\mathcal{S}_2\backslash\mathcal{R}}(\alpha_{t,s}+\alpha_{s,t})\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})
\end{aligned}
\tag{85}
$$

Note that this condition is equivalent to:

$$\begin{aligned}&\left\{\sum_{s\in\mathcal{T}} q_s - \left(\sum_{s\in\mathcal{T}} p_s\right)^2 - \sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r)\right\}\\ &+\left\{\sum_{s\in\mathcal{R}} q_s - \left(\sum_{s\in\mathcal{R}} p_s\right)^2 - \sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{R}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r)\right\}\\ &+\left\{2\left(\sum_{s\in\mathcal{S}_2\setminus\mathcal{R}} p_s\right)\left(\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} p_s\right) - \sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{t\in\mathcal{S}_2\setminus\mathcal{R}} (\alpha_{t,s}+\alpha_{s,t})\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1})\right\} \geq 0. \end{aligned} \tag{86}$$

We now show that this condition is redundant as it is implied by other conditions. Since $\mathcal{T}$ and $\mathcal{R}$ satisfy the conditions in (63) we already have that:

$$\sum_{s\in\mathcal{T}} q_s - \left(\sum_{s\in\mathcal{T}} p_s\right)^2 \geq \sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r) \tag{87}$$

$$\sum_{s\in\mathcal{R}} q_s - \left(\sum_{s\in\mathcal{R}} p_s\right)^2 \geq \sum_{s\in\mathcal{R}}\sum_{t\in\mathcal{R}^c} \alpha_{s,t}\cdot\mathbb{I}((s,t)\notin\mathcal{V}_r) \tag{88}$$

Further since the sets $(\mathcal{S}_1\setminus\mathcal{R})$ and $\mathcal{S}_2\setminus\mathcal{R}$ are disjoint it follows that:

$$2\left(\sum_{t\in\mathcal{S}_2\setminus\mathcal{R}} p_t\right)\left(\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} p_s\right) - \sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{t\in\mathcal{S}_2\setminus\mathcal{R}} (\alpha_{t,s}+\alpha_{s,t})\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \tag{89}$$

$$= \sum_{t\in\mathcal{S}_2\setminus\mathcal{R}}\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} 2p_sp_t - \sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{t\in\mathcal{S}_2\setminus\mathcal{R}} (\alpha_{t,s}+\alpha_{s,t})\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \tag{90}$$

$$= \sum_{t\in\mathcal{S}_2\setminus\mathcal{R}}\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} (2p_sp_t - (\alpha_{t,s}+\alpha_{s,t})\cdot\mathbb{I}((s,t)\notin\mathcal{V}_{r-1}) \geq 0 \tag{91}$$

where we use the fact that by construction $\alpha_{s,t}+\alpha_{t,s} = 2p_sp_t$. It thus follows that the condition (86) is implied by other conditions already presented in the FM elimination and is thus redundant. Since our choice $\mathcal{S}_1$ and $\mathcal{S}_2$ is arbitrary it follows that every combination of the form (70) is redundant and the only equations that remain upon elimination of α_{i_r,j_r} and α_{j_r,i_r} are given by (63), (68) and (69). This completes the induction step in Lemma 2 and the proof.

F. Connection between Theorem 2 and Polyhedral Cone Representation

We consider the case of $\Omega = \{1,2,3\}$ for sake of concreteness. We discuss how the characterization of $P^\star(\mathrm{acc}) = 1$ is related to dual representation of a polyhedral cone. Let $\mathbf{p} = (p_1,p_2,p_3)$ denote the draft probability and $\mathbf{q} = (q_1,q_2,q_3)$ denote the target probability vector. As before we define $\alpha_{i,j} = \Pr(Y=i,\{X_1,X_2\}=\{i,j\})$. We need to solve the following system of equations:

$$q_1 - p_1^2 = \alpha_{1,2}+\alpha_{1,3} \tag{92}$$

$$q_2 - p_2^2 = \alpha_{2,1}+\alpha_{2,3} \tag{93}$$

$$q_3 - p_3^2 = \alpha_{3,1}+\alpha_{3,2} \tag{94}$$

subject to the conditions that $\alpha_{i,j}+\alpha_{j,i} = 2p_ip_j$ and $0 \leq \alpha_{i,j} \leq 2p_ip_j$. Using the fact that $q_1+q_2+q_3 = 1$ and $p_1+p_2+p_3 = 1$, it suffices the consider the following system of equations:

$$\alpha_{1,2}+\alpha_{1,3} = q_1 - p_1^2 \tag{95}$$

$$\alpha_{2,1}+\alpha_{2,3} = q_2 - p_2^2 \tag{96}$$

$$\alpha_{1,2}+\alpha_{2,1} = 2p_{1,2} \tag{97}$$

$$\alpha_{1,3}+\alpha_{3,1} = 2p_{1,3} \tag{98}$$

$$\alpha_{2,3}+\alpha_{3,2} = 2p_{2,3} \tag{99}$$

with the additional requirement that $\alpha_{i,j} \geq 0$. We will represent this system of equations in matrix form. Our variables of interest are $\mathbf{x} = [\alpha_{1,2}, \alpha_{1,3}, \alpha_{2,1}, \alpha_{2,3}, \alpha_{3,1}, \alpha_{3,2}]^T \geq 0$. Our equality constraints can be expressed in the following form:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}, \qquad \mathbf{x} \geq 0 \tag{100}$$

where

$$\mathbf{A} = \begin{bmatrix} 1,1,0,0,0,0 \\ 0,0,1,1,0,0 \\ 1,0,1,0,0,0 \\ 0,1,0,0,1,0 \\ 0,0,0,1,0,1 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} \alpha_{1,2} \\ \alpha_{1,3} \\ \alpha_{2,1} \\ \alpha_{2,3} \\ \alpha_{3,1} \\ \alpha_{3,2} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} q_1 - p_1^2 \\ q_2 - p_2^2 \\ 2p_{1,2} \\ 2p_{1,3} \\ 2p_{2,3} \end{bmatrix} \tag{101}$$

Upon application of Farakas' Lemma it follows that the system (100) has a solution if and only if every $\mathbf{y}$ that satisfies $\mathbf{y}^T\mathbf{A} \geq 0$ also satisfies $\mathbf{y}^T\mathbf{b} \geq 0$, where $\mathbf{b}$ depends on $\mathbf{p}$ and $\mathbf{q}$ as in (101). Let us define

$$\mathbf{B} = \mathbf{A}^T = \begin{bmatrix} 1,0,1,0,0 \\ 1,0,0,1,0 \\ 0,1,1,0,0 \\ 0,1,0,0,1 \\ 0,0,0,1,0 \\ 0,0,0,0,1 \end{bmatrix} \tag{102}$$

and note that the set

$$\mathcal{B} = \{\mathbf{y} : \mathbf{B}\mathbf{y} \geq 0\} \tag{103}$$

denotes a polyhedral cone in $\mathbb{R}^5$. We need to show that for each $\mathbf{y} \in \mathcal{B}$ we must have that $\mathbf{y}^T\mathbf{b} \geq 0$. The representation (103) is the so-called hyperplane representation of the code as each row of $\mathbf{B}$ defines a hyperplane. We would like to find an equivalent generator representation of the form:

$$\mathcal{R} = \{\mathbf{z} : \mathbf{z} = \mathbf{R}\lambda, \lambda \geq 0\} \tag{104}$$

The Minikowski-Weyl Theorem (Fukuda & Prodon, 1995) guarantees that for every $\mathbf{B}$ in (103) there exists a $\mathbf{R}$ in (104) of finite dimensions such that $\mathcal{B} = \mathcal{R}$. Furthermore the double-description method is an algorithmic way of computing $\mathbf{R}$ given $\mathbf{B}$ and vice versa. Using the package *skeleton* for double description (Zny, 2018) we could show that for the $\mathbf{B}$ matrix in (102) the associated R matrix is given by:

$$\mathbf{R}^T = \left[\begin{array}{ccccc} \multicolumn{5}{c}{\mathbf{I}_5} \\ \hline 1 & 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 \\ 0 & -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 & 1 \end{array}\right] \tag{105}$$

where $\mathbf{I}_5$ is a 5×5 identity matrix. The generator representation in (104) is convenient as in order to show that (100) has a feasible solution, it suffices to show that $\mathbf{R}^T\mathbf{b} \geq 0$. Indeed substitution of (105) and (101) yields

$$\mathbf{R}^T\mathbf{b} = \begin{bmatrix} \mathbf{b} \\ q_1 + q_2 - (p_1+p_2)^2 \\ -q_1 + p_1^2 + 2p_{1,2} + 2p_{1,3} \\ -q_2 + p_2^2 + 2p_{1,2} + 2p_{2,3} \\ -q_1 - q_2 + p_1^2 + p_2^2 + 2p_{1,2} + 2p_{1,3} + 2p_{2,3} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ q_1 + q_2 - (p_1+p_2)^2 \\ q_2 + q_3 - (p_2+p_3)^2 \\ q_1 + q_3 + (p_1+p_3)^2 \\ q_3 - p_3^2 \end{bmatrix} \tag{106}$$

In the last step we use the fact that $\sum q_i = \sum p_i = 1$. It thus follows that $\mathbf{R}^T\mathbf{b} \geq 0$ if and only if $q_i \geq p_i^2$ and $q_i + q_j \geq (p_i + p_j)^2$ holds as stated in Theorem 2. Thus this approach provides an alternative proof for Theorem 2 for the case of $|\Omega| = 3$. We did not however find a simple approach to analytically compute the generator representation $\mathcal{R}$ from the hyperplane representation $\mathcal{B}$ for arbitrary dimensions. On the other hand we used the numerical implementation of the double description method to compute $\mathbf{B}$ and $\mathbf{R}$ for the case of up-to $K = 6$ drafts and $|\Omega| \leq 14$ and demonstrate that the natural counterpart of our result in Theorem 2 appears to be valid in all these cases

G. Proof of Theorem 3

As in the proof of Theorem 2, we let $\mathbf{p} = [p_1, \ldots, p_n]$ be the distribution of the draft model and $\mathbf{q} = [q_1, \ldots, q_n]$ be the distribution of the target model. Our optimization problem can be expressed as follows:

$$\text{maximize} \sum_{i=1}^{n} t_i, \tag{107}$$

$$t_i \leq \min\left(q_i, p_i^2 + \sum_{j \neq i} \alpha_{i,j}\right), \tag{108}$$

$$\alpha_{i,j} + \alpha_{j,i} = 2p_i p_j, 0 \leq \alpha_{i,j} \leq 1. \tag{109}$$

In order to solve this linear program analytically we introduce an additional variable z satisfying a single inequality $z \leq t_1 + \ldots t_n$. We provide the range of feasible feasible values of z and pick the maximum. Following the techniques used in the proof of Theorem 2 we have the following Lemma:

Lemma 3. *Upon applying Fourier-Motzkin elimination technique to eliminate variables $\alpha_{i,j}$ in (107)-(109), we have the following system of inequalities with $\Omega = \{1, \ldots, n\}$:*

$$t_i \leq q_i, i \in \Omega \tag{110}$$

$$\sum_{i \in \mathcal{S}} t_i \leq \left(\sum_{i \in \mathcal{S}} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S}} p_i\right)\left(\sum_{i \in \mathcal{S}^c} p_i\right), \quad \forall \mathcal{S} \subseteq \Omega, \mathcal{S}^c = \Omega \setminus \mathcal{S} \tag{111}$$

$$z \leq \sum_{i=1}^{n} t_i. \tag{112}$$

We will defer the proof of this lemma after the main proof. We will use (110)-(112) to establish the following step by induction.

Lemma 4. *Suppose that we apply Fourier Motzkin elimination to eliminate variables $t_1, \ldots, t_{j-1}$ in (110)-(112). Let $\Omega_1 = \{1, \ldots, j-1\}$ and $\Omega_2 = \{j, \ldots, n\}$ be partition of Ω. Then we have*

$$\begin{aligned} z \leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}} t_i + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S} \cup \mathcal{V}} p_i\right)\left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i\right), \\ \forall \mathcal{S} \subseteq \Omega_1, \mathcal{V} \subseteq \Omega_2, \mathcal{S}^c = \Omega_1 \setminus \mathcal{S}, \mathcal{V}^c = \Omega_2 \setminus \mathcal{V} \end{aligned} \tag{113}$$

$$\sum_{i \in \mathcal{S}} t_i \leq \left(\sum_{i \in \mathcal{S}} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S}} p_i\right)\left(\sum_{i \in \mathcal{S}^c} p_i\right), \quad \forall \mathcal{S} \subseteq \Omega_2, \quad \mathcal{S}^c = \Omega_2 \setminus \mathcal{S} \tag{114}$$

$$t_i \leq q_i, \qquad \forall i \in \Omega_2 \tag{115}$$

Note that this results implies the main result as by setting $\Omega_1 = \Omega$ and $\Omega_2 = \{\cdot\}$ we have:

$$z \leq \sum_{i \in \mathcal{S}} q_i + \left(\sum_{i \in \mathcal{S}^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S}} p_i\right)\left(\sum_{i \in \mathcal{S}^c} p_i\right) \tag{116}$$

We first consider the base case: $j = 1$. In this case $\Omega_1 = \{\cdot\}$ is the empty set and $\Omega_2 = \Omega$. Thus $\mathcal{S} = \mathcal{S}^c = \{\cdot\}$ and $\mathcal{V} \subseteq \Omega$ and $\mathcal{V}^c = \Omega \setminus \mathcal{V}$. In this case (113) reduces to:

$$z \leq \sum_{i \in \mathcal{V}} t_i + \left(\sum_{i \in \mathcal{V}^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{V}} p_i\right)\left(\sum_{i \in \mathcal{V}^c} p_i\right) \tag{117}$$

and (114) and (115) have $\Omega_2 = \Omega$. Note that (114) and (115) are equivalent to (110) and (111). It thus suffices to show the equivalence between (117) and (112). To show that the condition (117) implies (112) it suffices to set $\mathcal{V} = \Omega$ and $\mathcal{V}^c = \{\cdot\}$. To show that (112) implies (117), for any $\mathcal{V} \subseteq \Omega$, we can express:

$$z \leq \sum_{i \in \mathcal{V}} t_i + \sum_{i \in \mathcal{V}^c} t_i \tag{118}$$

$$\leq \sum_{i \in \mathcal{V}} t_i + \left(\sum_{i \in \mathcal{V}^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{V}} p_i\right)\left(\sum_{i \in \mathcal{V}^c} p_i\right) \tag{119}$$

where we use (111) in the second term. This completes the proof of the base case.

For induction we assume that for some $j > 0$ the application of Fourier-Motzkin elimination on eliminate $t_1, \ldots, t_{j-1}$ leads to (113)-(115) with $\Omega_1 = \{1, \ldots, j-1\}$ and $\Omega_2 = \{j, \ldots, n\}$. We want to show that upon applying Fourier-Motzkin elimination to eliminate t_j, we reduce the system of inequalities again to (113)-(115) with $\Omega_1' = \{1, \ldots, j\}$ and $\Omega_2' = \{j+1, \ldots, n\}$.

Let us consider those $\mathcal{V} \subseteq \Omega_2 = \{j, \ldots, n\}$ where $j \notin \mathcal{V}$ in (113). Each such $\mathcal{V} \subseteq \Omega_2' = \{j+1, \ldots, n\}$ as $j \notin \mathcal{V}$. Since the variable t_j does not appear in the right hand side in (113), the Fourier-Motzkin elimination will not modify the inequality. We can reinterpret (113) as:

$$z \leq \sum_{i \in \mathcal{S}'} q_i + \sum_{i \in \mathcal{V}'} t_i + \left(\sum_{i \in \mathcal{S}'^c \cup \mathcal{V}'^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S}' \cup \mathcal{V}'} p_i\right)\left(\sum_{i \in \mathcal{S}'^c \cup \mathcal{V}'^c} p_i\right),$$
$$\forall \mathcal{S}' \subseteq \Omega_1', j \notin \mathcal{S}', \mathcal{V}' \subseteq \Omega_2', \mathcal{S}'^c = \Omega_1' \setminus \mathcal{S}', \mathcal{V}'^c = \Omega_2' \setminus \mathcal{V}' \tag{120}$$

Next consider the case in (113) where $j \in \mathcal{V}$. In order to apply Fourier-Motzkin elimination, we express $\mathcal{V} = \{j\} \cup \mathcal{V}'$ where $\mathcal{V}' \subseteq \Omega_2' = \{j+1, \ldots, n\}$. We explicitly consider the variable t_j in (113) below.

$$z \leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}'} t_i + t_j + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S} \cup \mathcal{V}} p_i\right)\left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i\right), \tag{121}$$

We first combine (121) with the inequality $t_j \leq q_j$ and introduce $\Omega_1' = \Omega_1 \cup \{j\}$, $\Omega_2' = \Omega_2 \setminus \{j\}$, $\mathcal{S}' = \mathcal{S} \cup \{j\}$ and $\mathcal{S}'^c = \Omega_1' \setminus \mathcal{S}'$, $\mathcal{V}' = \mathcal{V} \setminus \{j\} \subseteq \Omega_2'$ and $\mathcal{V}'^c = \Omega_2' \setminus \mathcal{V}'$ to have:

$$z \leq \sum_{i \in \mathcal{S}'} q_i + \sum_{i \in \mathcal{V}'} t_i + \left(\sum_{i \in \mathcal{S}'^c \cup \mathcal{V}'^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S}' \cup \mathcal{V}'} p_i\right)\left(\sum_{i \in \mathcal{S}'^c \cup \mathcal{V}'^c} p_i\right),$$
$$\forall \mathcal{S}' \subseteq \Omega_1', j \in \mathcal{S}', \mathcal{V}' \subseteq \Omega_2', \mathcal{S}'^c = \Omega_1' \setminus \mathcal{S}', \mathcal{V}'^c = \Omega_2' \setminus \mathcal{V}' \tag{122}$$

Note that (120) and (122) recover all the upper bounds on z in the induction step for (113). We further need to show that the Fourier-Motzkin elimination does not introduce any further inequalities during the elimination of t_j. In particular with $j \in \mathcal{V}$ consider combining:

$$z \leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}} t_i + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{S} \cup \mathcal{V}} p_i\right)\left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i\right), \tag{123}$$

with the inequality:

$$\sum_{i \in \mathcal{W}} t_i \leq \left(\sum_{i \in \mathcal{W}} p_i\right)^2 + 2\left(\sum_{i \in \mathcal{W}} p_i\right)\left(\sum_{i \in \mathcal{W}^c} p_i\right) \tag{124}$$

where $\mathcal{W} \subseteq \Omega_2 = \{j, \ldots, n\}$ and $j \in \mathcal{W}$. Defining $\mathcal{W}_1 = \mathcal{W} \setminus \mathcal{V}$ and $\mathcal{U} = \mathcal{W} \cap \mathcal{V}$ we have:

$$\sum_{i \in \mathcal{W}} t_i = \left(\sum_{i \in \mathcal{W}_1} p_i + \sum_{i \in \mathcal{U}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i + \sum_{i \in \mathcal{U}} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \tag{125}$$

$$\begin{aligned} &= \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + \left(\sum_{i \in \mathcal{U}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \mathcal{U}} p_i \right) \\ &\quad + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) + 2 \left(\sum_{i \in \mathcal{U}} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \end{aligned} \tag{126}$$

$$\begin{aligned} &= \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + \left(\sum_{i \in \mathcal{U}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \mathcal{U}} p_i + \sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \\ &\quad + \left(\sum_{i \in \mathcal{U}} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \end{aligned} \tag{127}$$

$$= \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + \left(\sum_{i \in \mathcal{U}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}_1} p_i \right) + 2 \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \left(\sum_{i \in \mathcal{U}} p_i \right). \tag{128}$$

Next we consider (123):

$$\begin{aligned} z \leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}_1} t_i + \sum_{i \in \mathcal{U}} t_1 + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right)^2 \\ + 2 \left(\sum_{i \in \mathcal{S} \cup \mathcal{V}_1} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right) + 2 \left(\sum_{i \in \mathcal{U}} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right), \end{aligned} \tag{129}$$

where we use the fact that $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{U}$. Adding (129) to (128) and eliminating t_i where $i \in \mathcal{U}$, we get:

$$\begin{aligned} z + \sum_{i \in \mathcal{W}_1} t_i &\leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}_1} t_i + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right)^2 \\ &\quad + 2 \left(\sum_{i \in \mathcal{S} \cup \mathcal{V}_1} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right) + 2 \left(\sum_{i \in \mathcal{U}} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right) \\ &\quad + \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + \left(\sum_{i \in \mathcal{U}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}_1} p_i \right) + 2 \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \left(\sum_{i \in \mathcal{U}} p_i \right) \end{aligned} \tag{130}$$

$$\begin{aligned} &= \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}_1} t_i + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i + \sum_{i \in \mathcal{U}} p_i \right)^2 \\ &\quad + 2 \left(\sum_{i \in \mathcal{S} \cup \mathcal{V}_1} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}^c} p_i \right) + 2 \left(\sum_{i \in \Omega \setminus \mathcal{W}} p_i \right) \left(\sum_{i \in \mathcal{U}} p_i \right) \\ &\quad + \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}_1} p_i \right) \end{aligned} \tag{131}$$

Next note that $\mathcal{S} \cup \mathcal{V}_1 \subseteq \Omega \setminus \mathcal{W}$. This follows since $\Omega = \Omega_1 \cup \Omega_2$ and $\mathcal{S} \subseteq \Omega_2$, $\mathcal{V}_1, \mathcal{W} \subseteq \Omega_2$ and $\mathcal{V}_1 \cup \mathcal{W} = \cdot$ by definition as $\mathcal{V}_1 = \mathcal{V} \setminus \mathcal{W}$. Thus the application of Fourier-Motzkin elimination with $\mathcal{V}_1^c = \mathcal{V}^c \cup \mathcal{U}$ gives:

$$\begin{aligned} z + \sum_{i \in \mathcal{W}_1} t_i \leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}_1} t_i + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}_1^c} p_i \right) + 2 \left(\sum_{i \in \mathcal{S} \cup \mathcal{V}_1} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}_1^c} p_i \right) \\ + \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}_1} p_i \right) \end{aligned} \tag{132}$$

However the above inequality is a consequence of the following:

$$\begin{aligned} z \leq \sum_{i \in \mathcal{S}} q_i + \sum_{i \in \mathcal{V}_1} t_i + \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}_1^c} p_i \right) + 2 \left(\sum_{i \in \mathcal{S} \cup \mathcal{V}_1} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \mathcal{V}_1^c} p_i \right) \\ \sum_{i \in \mathcal{W}_1} t_i \leq \left(\sum_{i \in \mathcal{W}_1} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{W}_1} p_i \right) \left(\sum_{i \in \Omega \setminus \mathcal{W}_1} p_i \right) \end{aligned} \tag{133}$$

where $\mathcal{V}_1 \subseteq \Omega_2$, $\mathcal{V}_1^c = \Omega_2 \setminus \mathcal{V}_1$, $\mathcal{S} \subset \Omega_1$ and $\mathcal{S}^c \subseteq \Omega_1 \setminus \mathcal{S}$ and $\mathcal{W}_1 \subseteq \Omega_2$. which are already implied in the induction step. Thus we conclude that each combination of the form (123) and (124) is redundant and need not be included in the next step of the Fourier-Motzkin elimination. This concludes the analysis of the upper bound on z in (113).

It remains to establish the induction for (114) and (115) i.e., upon elimination of t_j results in

$$\sum_{i \in \mathcal{S}} t_i \leq \left(\sum_{i \in \mathcal{S}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{S}} p_i \right) \left(\sum_{i \in \mathcal{S}^c} p_i \right), \quad \forall \mathcal{S} \subseteq \Omega_2', \quad \mathcal{S}^c = \Omega_2' \setminus \mathcal{S} \tag{134}$$

$$t_i \leq q_i, \qquad \forall i \in \Omega_2' \tag{135}$$

where $\Omega_2' = \{j+1, \ldots, n\}$. Naturally every inequality (134) and (135) is already contained in (114) and (115) where $j \notin \mathcal{S}$. So we only need to show that the application of Fourier-Motzkin elimination to remove any other inequality does not result in any additional inequality. Note that the elimination of t_j simply involves combining each inequality with $t_j > 0$. Thus any inequality in (114) where $\mathcal{S} \subseteq \Omega_2$ with $j \in \mathcal{S}$ reduces to:

$$\sum_{i \in \mathcal{S} \setminus \{j\}} t_i \leq \left(\sum_{i \in \mathcal{S}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{S}} p_i \right) \left(\sum_{i \in \mathcal{S}^c} p_i \right) \tag{136}$$

We show that (136) is weaker than

$$\sum_{i \in \mathcal{S} \setminus \{j\}} t_i \leq \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \{j\}} p_i \right) \tag{137}$$

which is already contained in (114) and hence redundant. In particular consider the right hand side of (136):

$$\left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i + p_j \right)^2 + 2 \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i + p_j \right) \left(\sum_{i \in \mathcal{S}^c} p_i \right) \tag{138}$$

$$\geq p_j^2 + 2p_j \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right) + \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right) \left(\sum_{i \in \mathcal{S}^c} p_i \right) \tag{139}$$

$$\geq \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right)^2 + 2 \left(\sum_{i \in \mathcal{S} \setminus \{j\}} p_i \right) \left(\sum_{i \in \mathcal{S}^c \cup \{j\}} p_i \right), \tag{140}$$

which implies that (136) is indeed weaker.

Thus we have completed the induction step. Continuing the induction to eliminate all variables $t_1, \dots, t_n$ results in

$$z \le \sum_{i\in\mathcal{S}} q_i + \left(\sum_{i\in\mathcal{S}^c} p_i\right)^2 + 2\left(\sum_{i\in\mathcal{S}} p_i\right)\left(\sum_{i\in\mathcal{S}^c} p_i\right), \qquad \forall \mathcal{S} \in \Omega \tag{141}$$

as claimed. It now only remains to establish the proof of Lemma 3 which we will do. As we are considering the elimination of $\alpha_{i,j}$ it suffices to consider the following inequalities:

$$t_i \le p_i^2 + \sum_{j=1, j\neq i}^{n} \alpha_{i,j}, \qquad i = 1, 2, \dots, n \tag{142}$$

$$\alpha_{i,j} + \alpha_{j,i} = 2p_i p_j, \qquad 0 \le \alpha_{i,j} \le 1 \tag{143}$$

We will show the following by induction. Suppose that at step $r >= 1$ let

$$\mathcal{V}_r = \{(i_1, j_1), (j_1, i_1), \dots, (i_{r-1}, j_{r-1}), (j_{r-1}, i_{r-1})\} \tag{144}$$

denotes the indices (with $i_k \le j_k$) of variables that are eliminated using the Fourier-Motzkin elimination. Then the resulting system of inequalities is given by:

$$\sum_{s\in\mathcal{S}} t_s \le \left(\sum_{s\in\mathcal{S}} p_s\right)^2 + \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} \alpha_{s,t}\cdot \mathbb{I}((s,t)\notin\mathcal{V}_r) + \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} 2p_s p_t \cdot \mathbb{I}((s,t)\in\mathcal{V}_r) \tag{145}$$

for all $\mathcal{S} \subseteq \Omega = \{1, 2, \dots, n\}$. For the base case, consider the case when$r = 1$ i.e., $\mathcal{V}_r = \{\cdot\}$. The condition in (145) reduces to:

$$\sum_{s\in\mathcal{S}} t_s \le \left(\sum_{s\in\mathcal{S}} p_s\right)^2 + \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}^c} \alpha_{s,t}\cdot \qquad \forall \mathcal{S} \subseteq \Omega \tag{146}$$

We show that (146) is equivalent to (142). Indeed setting $\mathcal{S} = \{i\}$ in (145) and using $\alpha_{s,t} = 2p_s p_t$ recovers (142) for each $i = 1, 2, \dots, n$. We will show that the conditions (142) and (143) also imply (145). Note that for any $\mathcal{S} \subseteq \Omega$:

$$\sum_{s\in\mathcal{S}} t_s \le \sum_{s\in\mathcal{S}} p_s^2 + \sum_{s\in\mathcal{S}}\sum_{i=1, i\neq s}^{n} \alpha_{s,t} \tag{147}$$

$$= \sum_{s\in\mathcal{S}} p_s^2 + \sum_{s\in\mathcal{S}}\left(\sum_{i\in\mathcal{S}^c} \alpha_{s,i} + \sum_{i\in\mathcal{S}\setminus\{s\}} \alpha_{s,i}\right) \tag{148}$$

$$= \sum_{s\in\mathcal{S}} p_s^2 + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} \alpha_{s,i} + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}\setminus\{s\}} \alpha_{s,i} \tag{149}$$

$$= \sum_{s\in\mathcal{S}} p_s^2 + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} \alpha_{s,i} + \sum_{(s,i)\in\mathcal{S}\times\mathcal{S}, i>s} (\alpha_{s,i} + \alpha_{i,s}) \tag{150}$$

$$= \sum_{s\in\mathcal{S}} p_s^2 + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} \alpha_{s,i} + \sum_{(s,i)\in\mathcal{S}\times\mathcal{S}, i>s} 2p_{s,i} \tag{151}$$

$$= \left(\sum_{s\in\mathcal{S}} p_s\right)^2 + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} \alpha_{s,i} \tag{152}$$

We thus recover (146) from (142). This establishes the base case.

For the induction step, let us assume that we have eliminated all $\alpha_{i,j}$ where the indices (i, j) are in the set $\mathcal{V}_r$ and that (142) is satisfied. We consider elimination of indices (i_r, j_r) and (j_r, i_r) associated with α_{i_r,j_r} and α_{j_r,i_r}:

$$\sum_{s\in\mathcal{S}} t_s \le \left(\sum_{s\in\mathcal{S}} p_s\right)^2 + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} \alpha_{s,i}\cdot \mathbb{I}((s,i)\notin\mathcal{V}_r) + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} 2p_s p_i \cdot \mathbb{I}((s,i)\in\mathcal{V}_r) \tag{153}$$

We need to show that upon elimination of α_{i_r,j_r} and α_{j_r,i_r} using Fourier-Motzkin elimination the resulting system of inequalities is given by:

$$\sum_{s\in\mathcal{S}} t_s \le \left(\sum_{s\in\mathcal{S}} p_s\right)^2 + \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} \alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_{r+1}) \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{S}^c} 2p_s p_i \cdot \mathbb{I}((s,i)\in\mathcal{V}_{r+1}), \tag{154}$$

with

$$\mathcal{V}_{r+1} = \{(i_1,j_1),(j_1,i_1),\ldots,(i_r,j_r),(j_r,i_r)\}\,. \tag{155}$$

We note that in the Fourier-Motzkin elimination step we have to only consider those inequalities where either α_{i_r,j_r} or α_{j_r,i_r} appears on the right hand side of (153). This is equivalent to having $i_r \in \mathcal{S}$ and $j_r \in \mathcal{S}^c$ or $j_r \in \mathcal{S}$ and $i_r \in \mathcal{S}^c$. For those $\mathcal{S}$ that do not satisfy either condition, we immediately have (154). If the selected $\mathcal{S}$ follow either of these cases, combining (153) with $\alpha_{i_r,j_r} \le 2p_{i_r}p_{j_r}$ and $\alpha_{j_r,i_r} \le 2p_{i_r}p_{j_r}$, we reduce to (154). At this point all the equations in (154) have been recovered. Nevertheless Fourier-Motzkin elimination requires us to also consider all pairwise equations where $\mathcal{S}_1,\mathcal{S}_2 \subseteq \Omega$, $i_r \in \mathcal{S}_1$ and $j_r \notin \mathcal{S}_1$ and $i_r \notin \mathcal{S}_2$ and $j_r \in \mathcal{S}_2$:

$$\sum_{s\in\mathcal{S}_1} t_s - \left(\sum_{s\in\mathcal{S}_1} p_s\right)^2 \le \sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_1^c} \alpha_{s,i}\mathbb{I}((s,i)\notin\mathcal{V}_r) + \sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_1^c} 2p_s p_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \tag{156}$$

$$\sum_{s\in\mathcal{S}_2} t_s - \left(\sum_{s\in\mathcal{S}_2} p_s\right)^2 \le \sum_{s\in\mathcal{S}_2}\sum_{i\in\mathcal{S}_2^c} \alpha_{s,i}\mathbb{I}((s,i)\notin\mathcal{V}_r) + \sum_{s\in\mathcal{S}_2}\sum_{i\in\mathcal{S}_2^c} 2p_s p_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \tag{157}$$

The Fourier-Motzkin elimination step requires us to combine (156) and (157) and use $\alpha_{i_r,j_r} + \alpha_{j_r,i_r} = 2p_{i_r}p_{j_r}$, $\alpha_{i_r,j_r}, \alpha_{j_r,i_r} \ge 0$ to eliminate α_{i_r,j_r} and α_{j_r,i_r} in the induction step.

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_1} t_s - \left(\sum_{s\in\mathcal{S}_1} p_s\right)^2 + \sum_{s\in\mathcal{S}_2} t_s - \left(\sum_{s\in\mathcal{S}_2} p_s\right)^2 \\
&\quad\le \sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_1^c} \alpha_{s,i}\mathbb{I}((s,i)\notin\mathcal{V}_r) + \sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_1^c} 2p_s p_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \\
&\qquad + \sum_{s\in\mathcal{S}_2}\sum_{i\in\mathcal{S}_2^c} \alpha_{s,i}\mathbb{I}((s,i)\notin\mathcal{V}_r) + \sum_{s\in\mathcal{S}_2}\sum_{i\in\mathcal{S}_2^c} 2p_s p_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)
\end{aligned} \tag{158}$$

We will show that each such inequality is redundant and already implied by the set of equations already established in (154).

Let $\mathcal{R} = \mathcal{S}_1 \cap \mathcal{S}_2$ and $\mathcal{T} = \mathcal{S}_1 \cup \mathcal{S}_2$. Note that $i_r \notin \mathcal{R}$ and $j_r \notin \mathcal{R}$. First following the same steps leading to (74) we can show that:

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_1} t_s - \left(\sum_{s\in\mathcal{S}_1} p_s\right)^2 + \sum_{s\in\mathcal{S}_2} t_s - \left(\sum_{s\in\mathcal{S}_2} p_s\right)^2 \\
&= \left\{\sum_{s\in\mathcal{T}} t_s - \left(\sum_{s\in\mathcal{T}} p_s\right)^2\right\} + \left\{\sum_{s\in\mathcal{R}} t_s - \left(\sum_{s\in\mathcal{R}} p_s\right)^2\right\} + 2\left(\sum_{s\in\mathcal{S}_2\setminus\mathcal{R}} p_s\right)\left(\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} p_s\right).
\end{aligned} \tag{159}$$

Next, following the steps leading to (80) we have that:

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_1^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&=\sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \qquad (160)\\
&=\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{S}_1\backslash\mathcal{R}}\sum_{i\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{S}_2\backslash\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \qquad (161)
\end{aligned}$$

and, likewise,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_2}\sum_{i\in\mathcal{S}_2^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&=\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{S}_2\backslash\mathcal{R}}\sum_{i\in\mathcal{S}_1\backslash\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r))+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\\
&\quad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{S}_1\backslash\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \qquad (162)
\end{aligned}$$

Next we combine the terms to get:

$$
\begin{align}
&\sum_{s\in\mathcal{S}_1}\sum_{i\in\mathcal{S}_1^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&+\sum_{s\in\mathcal{S}_2}\sum_{i\in\mathcal{S}_2^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&=\Bigg\{\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{S}_2\setminus\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\Bigg\} \nonumber\\
&\qquad+\Bigg\{\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{S}_2\setminus\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{S}_1\setminus\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)\Bigg\} \tag{163}\\
&\qquad+\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{i\in\mathcal{S}_2\setminus\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{S}_2\setminus\mathcal{R}}\sum_{i\in\mathcal{S}_1\setminus\mathcal{R}}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r))+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \tag{164}\\
&=\sum_{s\in\mathcal{T}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{R}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{i\in\mathcal{S}_2\setminus\mathcal{R}}(\alpha_{s,i}+\alpha_{i,s})\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+4p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \tag{165}\\
&=\sum_{s\in\mathcal{T}}\sum_{i\in\mathcal{T}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{R}^c}\alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r)+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \nonumber\\
&\qquad+\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{i\in\mathcal{S}_2\setminus\mathcal{R}}2p_sp_i+2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \tag{166}\\
& \tag{167}
\end{align}
$$

Thus the resulting inequality from Fourier-Motzkin elimination is given by:

$$\begin{aligned}
&\left\{\sum_{s\in\mathcal{T}} t_s - \left(\sum_{s\in\mathcal{T}} p_s\right)^2\right\} + \left\{\sum_{s\in\mathcal{R}} t_s - \left(\sum_{s\in\mathcal{R}} p_s\right)^2\right\} + 2\left(\sum_{s\in\mathcal{S}_2\setminus\mathcal{R}} p_s\right)\left(\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} p_s\right) \\
&\le \sum_{s\in\mathcal{T}}\sum_{i\in\mathcal{T}^c} \alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r) + 2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \\
&\quad + \sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{R}^c} \alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r) + 2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \\
&\quad + \sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{i\in\mathcal{S}_2\setminus\mathcal{R}} 2p_sp_i + 2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)
\end{aligned} \tag{168}$$

Next note that since $(i_r, j_r) \in \mathcal{T}$ and $(i_r, j_r) \notin \mathcal{R}$, the inequalities:

$$\begin{aligned}
\sum_{s\in\mathcal{T}} t_s &\le \left(\sum_{s\in\mathcal{T}} p_s\right)^2 + \sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{R}^c} \alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r) + 2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r) \\
\sum_{s\in\mathcal{R}} t_s &\le \left(\sum_{s\in\mathcal{R}} p_s\right)^2 + \sum_{s\in\mathcal{R}}\sum_{i\in\mathcal{R}^c} \alpha_{s,i}\cdot\mathbb{I}((s,i)\notin\mathcal{V}_r) + 2p_sp_i\cdot\mathbb{I}((s,i)\in\mathcal{V}_r)
\end{aligned} \tag{169}$$

are already constructed in the induction step. Also clearly (168) is implied by these since:

$$2\left(\sum_{s\in\mathcal{S}_2\setminus\mathcal{R}} p_s\right)\left(\sum_{s\in\mathcal{S}_1\setminus\mathcal{R}} p_s\right) = \sum_{s\in\mathcal{S}_1\setminus\mathcal{R}}\sum_{i\in\mathcal{S}_2\setminus\mathcal{R}} 2p_sp_i \tag{170}$$

is an identity since $\mathcal{S}_1 \setminus cR$ and $\mathcal{S}_2 \setminus \mathcal{R}$ are disjoint. Thus each such inequality form the Fourier-Motzkin elimination is redundant and we have completed the induction step and in turn established Lemma 3.

H. Proof of Equation (10)

First we consider the non-truncated program and let $w_{i,j}$ be the variables for $i, j \in \Omega$ with $i < j$ that maximize the objective:

$$\sum_{i=1}^{n} \min(q_i, p_I(i)) \tag{171}$$

where

$$p_I(i) = p_i^2 + \sum_{j=1, j\neq i}^{n} 2p_ip_jw_{i,j} \tag{172}$$

Note that we have

$$P^\star(\mathrm{acc}) = \sum_{i=1}^{n} \min(q_i, p_I(i)) \le \sum_{i=1}^{s} \min(q_i, p_I(i)) + \sum_{i=s+1}^{n} q_i \tag{173}$$

For the truncated linear program we have for each $i \in \{1, 2, \ldots, s\}$:

$$\tilde{p}_I(i) = p_i^2 + \sum_{j=1, j\neq i}^{s} 2p_ip_j\tilde{w}_{i,j} \sum_{j=s+1}^{n} 2p_ip_j \tag{174}$$

and for $i > s$:

$$\tilde{p}_I(i) = p_i^2 + \sum_{j=i+1}^{n} 2p_ip_j \tag{175}$$

We consider a potentially sub-optimal choice of weights $\tilde{w}_{i,j} = w_{i,j}$ for $i, j \in \{1, \ldots, s\}$ for the truncated linear program. Note that

$$\tilde{p}_I(i) \geq p_I(i), \qquad \forall i \leq s \tag{176}$$

and

$$\tilde{p}_I(i) \geq p_i^2, \qquad \forall i > s. \tag{177}$$

As a result, using (173) we have:

$$\tilde{P}(\text{acc}) \geq \sum_{i=1}^{n} \min(q_i, \tilde{p}_I(i)) \tag{178}$$

$$\geq \sum_{i=1}^{s} \min(q_i, p_I(i)) + \sum_{i=s+1}^{n} \min(q_i, p_i^2) \tag{179}$$

$$\geq P^\star(\text{acc}) - \sum_{i=s+1}^{n} q_i + \sum_{i=s+1}^{n} \min(q_i, p_i^2) \tag{180}$$

$$= P^\star(acc) - \sum_{i=s+1}^{n} (q_i - p_i^2)^+ \tag{181}$$

I. Proof of Equation (11)

Recall that $\Omega = \{1, \ldots, n\}$ denotes the full vocabulary and $\Omega_0 \subseteq \Omega$ is a high probability set of tokens selected so that $q(\Omega_0) = 1 - q_\varepsilon$ and $p(\Omega_0) = 1 - p_\varepsilon$. Recall that $\tilde{p}(\cdot)$ and $\tilde{q}(\cdot)$ denote the distributions over Ω_0 obtained by truncating $p(\cdot)$ and $q(\cdot)$ to Ω_0 and re-normalizing them.

Given the set of input tokens $\mathcal{S} = \{X_1, \ldots, X_K\}$ we select a subset $\tilde{\mathcal{S}} \subseteq \mathcal{S}$ by discarding any tokens in $\mathcal{S}$ that do not belong to Ω. Let $\{\tilde{X}_1, \ldots, \tilde{X}_{K'}\}$ denote the tokens in $\tilde{S}$, which are effectively sampled from the distribution $\tilde{p}(\cdot)$ i.e., $\Pr(\tilde{X}_i = x) = \tilde{p}(x)$ if $x \in \Omega_0$.

We then perform importance weighted sampling over the input tokens $\tilde{\mathcal{S}}$ to generate output token $Y \sim p_I(\cdot)$. We then perform speculative sampling using target distribution $q(\cdot)$ and output the resulting token $Z \sim q(\cdot)$.

Let $\tilde{P}_{\Omega_0}(\text{acc})$ be the acceptance probability of the proposed truncated alphabet scheme and $\tilde{P}^\star(\text{acc})$ be the optimal acceptance probability without truncation of the alphabet.

Without loss of generality assume $\Omega_0 = \{1, 2, \ldots, \delta\}$ and $\Omega = \{1, 2, \ldots, n\}$. Also assume $p(\Omega_0) = 1 - p_\varepsilon$ and $q(\Omega_0) = 1 - q_\varepsilon$. Let $\mathcal{S} = \{X_1, X_2\}$ denote the input tokens and $\tilde{\mathcal{S}}$ denote the tokens that are in Ω_0.

We consider the following:

$$\tilde{P}_{\Omega_0}(\text{acc} \mid \tilde{\mathcal{S}} = \mathcal{S}) = \sum_{i \in \Omega_0} \min(q_i, \tilde{p}_I(i)) \tag{182}$$

where

$$\tilde{p}_I(i) = \tilde{p}_i^2 + \sum_{j=1, j\neq i}^{\delta} 2\tilde{p}_i\tilde{p}_j\tilde{w}_{i,j} \tag{183}$$

where $\tilde{w}_{m,n}$ are the associated variables as discussed in our formulation.

When truncation is not used, let $w_{i,j}$ be the associated variables for $i, j \in \Omega$ and $i \neq j$ such that

$$p_I(i) = p_i^2 + \sum_{j=1, j\neq i}^{n} 2p_ip_jw_{i,j} \tag{184}$$

achieves the optimal acceptance probability:

$$P^\star(\text{acc}) = \sum_{i=1}^{n} \min(p_I(i), q_i) \leq \sum_{i=1}^{\delta} \min(p_I(i), q_i) + q_\epsilon \tag{185}$$

In the optimization program (183), we consider a potentially sub-optimal choice:: $\tilde{w}_{i,j} = w_{i,j}$ for $i < j$. Also note that due to truncation, $\tilde{p}_i = p_i / p(\Omega_0) \geq p_i$ for each $i = 1, 2, \ldots, \delta$. Thus it follows that

$$\tilde{p}_I(i) \geq p_I(i), \quad \forall i \in \Omega_0 \tag{186}$$

It thus follows that:

$$\begin{align} P^\star(\text{acc}) &\leq \sum_{i=1}^{\delta} \min(\tilde{p}_I(i), q_i) + q_\epsilon \tag{187} \\ &\leq \tilde{P}^\star(\text{acc}|\tilde{\mathcal{S}} = \mathcal{S}) + q_\epsilon \tag{188} \end{align}$$

and:

$$\begin{align} &\tilde{P}^\star(\text{acc}) \geq \Pr(\tilde{\mathcal{S}} = \mathcal{S}) \cdot \tilde{P}^\star(\text{acc}|\tilde{\mathcal{S}} = \mathcal{S}) \tag{189} \\ &\prod_{i=1}^{2} P(X_i \in \mathcal{S}) \Pr(\tilde{\mathcal{S}} = \mathcal{S}) \cdot \tilde{P}^\star(\text{acc}|\tilde{\mathcal{S}} = \mathcal{S}) \tag{190} \\ &= (1 - p_\varepsilon)^2 \left(P^\star(\text{acc}) - q_\epsilon\right) \tag{191} \\ &\geq (1 - 2p_\varepsilon) \left(P^\star(\text{acc}) - q_\varepsilon)\right) \tag{192} \end{align}$$

J. LP and Fast LP version for non-identical draft distributions, $K = 2$

For the case of $K = 2$ drafts we explain how the importance weighted sampling scheme and its faster variants can be extended when the two tokens are sampled independently but from different distribution i.e. $X_1 \sim p_1(\cdot)$ and $X_2 \sim p_2(\cdot)$. We let $\mathbf{p}_1 = (p_{1,1}, \ldots, p_{1,n})$ and $\mathbf{p}_2 = (p_{2,1}, \ldots, p_{2,n})$ denote the distributions of the draft models to sample X_1 and X_2. We let $\mathbf{q} = (q_1, \ldots, q_n)$ denote the target distribution.

The order of the tokens matters and accordingly for $i < j$, we define:

$$\begin{align} w_{i,j} = \Pr(Y = i | X_1 = i, X_2 = j), \bar{w}_{i,j} = 1 - w_{i,j} = \Pr(Y = j | X_1 = i, X_2 = j) \tag{193} \\ w_{j,i} = \Pr(Y = i | X_1 = j, X_2 = i), \bar{w}_{j,i} = 1 - w_{j,i} = \Pr(Y = j | X_1 = j, X_2 = i) \tag{194} \end{align}$$

If Y denotes the selected token, then considering all cases where token i appears as one of the input tokens, we have:

$$p_I(i) = p_{1,i} p_{2,i} + \sum_{j=1}^{i-1} p_{1,i} p_{2,j} \bar{w}_{i,j} + \sum_{j=i+1}^{n} p_{1,i} p_{2,j} w_{i,j} + \sum_{j=1}^{i-1} p_{1,j} p_{2,i} \bar{w}_{j,i} + \sum_{j=i+1}^{n} p_{1,j} p_{2,i} w_{j,i} \tag{195}$$

We need to find $w_{i,j}$ and $w_{j,i}$ that maximizes $\sum_{i=1}^{n} \min(q_i, p_I(i))$. This is a linear program in variables $w_{i,j}$ satisfying $0 \leq w_{i,j} \leq 1$. The truncated version of LP is obtained by sorting the tokens in Ω based on $q_i - p_{1,i} p_{2,i}$ again considering sets $\Omega_1 = \{1, 2, \ldots, s\}$ and $\Omega_2 = \{s + 1, \ldots, n\}$. We treat $w_{i,j}$ as variables that need to be optimized if $i, j \in \Omega_1$. If $i \in \Omega_1$ and $j \in \Omega_2$ we set $w_{i,j} = 1$. If both $i, j \in \Omega_2$ we set $w_{i,j} = 1$ if $i < j$ and 0 if $i > j$. The resulting distribution is given as follows. For $i \in \{1, \ldots, s\}$:

$$\begin{align} \tilde{p}_I(i) = p_{1,i} p_{2,i} + \sum_{j=1}^{i-1} p_{1,i} p_{2,j} \bar{w}_{i,j} + \sum_{j=i+1}^{n} p_{1,i} p_{2,j} w_{i,j} + \sum_{j=1}^{i-1} p_{1,j} p_{2,i} \bar{w}_{j,i} \\ + \sum_{j=i+1}^{n} p_{1,j} p_{2,i} w_{j,i} + \sum_{j=s+1}^{n} (p_{1,j} p_{2,i} + p_{1,i} p_{2,j}) \tag{196} \end{align}$$

and for $i = s+1, \ldots, n$, we have:

$$\tilde{p}_I(i) = p_{1,i}p_{2,i} + \sum_{j=i+1}^{n} (p_{1,j}p_{2,i} + p_{1,i}p_{2,j}) \tag{197}$$

Upon following the sequence of steps leading to (181) we can show that

$$\tilde{P}(\text{acc}) \geq P^{\star}(\text{acc}) - \sum_{i \in \Omega_2} (q_i - p_{1,i}p_{2,i})^{+}. \tag{198}$$

The truncated alphabet scheme can be applied in a similar fashion by considering a high probability subset $\Omega_0 \subseteq \Omega$ and only keeping those input tokens that belong to Ω_0. We generate truncated distributions $\tilde{p}_1(\cdot)$ and $\tilde{p}_2(\cdot)$ and apply the linear program on these followed by speculative sampling using the target distribution $q(\cdot)$.