

# 793 Appendices

## 794 A Proofs

795 In this section we prove the results in the main paper. For clarity, each theorem is restated and then a  
796 proof is given.

### 797 A.1 Proof of proposition 1

798 **Proposition 1.** *The GLS procedure, as described in section 3, generates samples such that:*

- 799 1.  $\Pr[X^{(k)} = j] = p_j$  for all  $k \in \{1, \dots, K\}$  and  $j \in \{1, \dots, N\}$ .
- 800 2.  $\Pr[Y = j] = q_j$  for all  $j \in \{1, \dots, N\}$ .

801 *Proof.* Since the  $S_i^{(k)}$ 's are i.i.d., the  $X^{(k)}$ 's will be also and it therefore suffices to check the  
802 distribution of  $X^{(1)}$ . Note that

$$\begin{aligned} X^{(1)} = j &\implies \frac{S_j^{(1)}}{p_j} < \frac{S_i^{(1)}}{p_i} \quad \forall i \neq j \\ &\implies S_j^{(1)} < \min_{i \neq j} \frac{S_i^{(1)}}{p_i/p_j}. \end{aligned}$$

803 The left-hand side is an exponential random variable with parameter  $\lambda = 1$ , and the right-hand side is  
804 an independent exponential random variable with parameter  $\lambda = \sum_{i \neq j} p_i/p_j$ . So,

$$\Pr[X^{(1)} = j] = \frac{1}{1 + \sum_{i \neq j} p_i/p_j} = p_j$$

805 as required. Next, we look at the distribution of  $Y$ . Define  $S_j^* = \min_{1 \leq k \leq K} S_j^{(k)}$ . Then,

$$\begin{aligned} Y = j &\implies \frac{S_j^*}{q_j} < \frac{S_i^*}{q_i} \quad \forall i \neq j \\ &\implies S_j^* < \min_{i \neq j} \frac{S_i^*}{q_i/q_j}. \end{aligned}$$

806 On the left-hand side,  $S_j^*$  is an exponential random variable with parameter  $\lambda = K$ , while the  
807 right-hand side is an independent exponential random variable with parameter  $\lambda = K \sum_{i \neq j} q_i/q_j$ .  
808 Finally,

$$\Pr[Y = j] = \frac{K}{K + K \sum_{i \neq j} q_i/q_j} = q_j.$$

809 □

### 810 A.2 Proof of theorem 1

811 **Theorem 1** (List matching lemma). *The matching probability is bounded below as*

$$\Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\}] \geq \sum_{j=1}^N \frac{K}{\sum_{i=1}^N [\max\{q_i/q_j, p_i/p_j\} + (K-1)q_i/q_j]}. \quad (3)$$

812 Furthermore, conditioned on  $Y = j$ , we have

$$\Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j] \geq \left(1 + \frac{q_j}{Kp_j}\right)^{-1}. \quad (4)$$

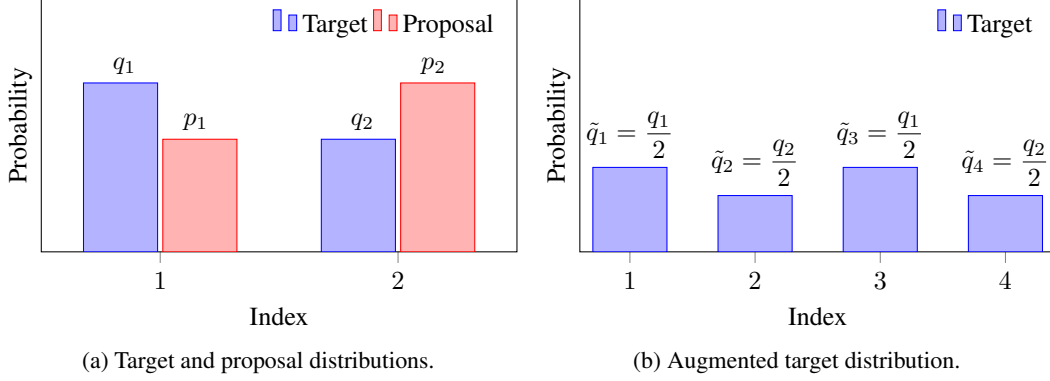


Figure 5: Distributions used in the proof of theorem 1 in the case of  $K = 2$  and  $N = 2$ .

813 *Proof.* To analyze the matching probability, we conceptualize the scheme slightly differently as  
 814 follows. Instead of taking the minimum over  $1 \leq k \leq K$  to obtain  $\{S_i^*\}_{i=1}^N$  like in the proof of  
 815 proposition 1, we form a flattened sequence of the  $S_i^{(k)}$ 's by defining

$$\{T_i\}_{i=1}^{NK} = \{S_1^{(1)}, \dots, S_N^{(1)}, S_1^{(2)}, \dots, S_N^{(2)}, \dots, S_1^{(K)}, \dots, S_N^{(K)}\}.$$

816 We also introduce an augmented target distribution  $\tilde{q}_{\tilde{Y}}$  on the extended alphabet  $\tilde{\Omega} = \{1, \dots, KN\}$   
 817 defined by the probabilities

$$(\tilde{q}_1, \dots, \tilde{q}_{KN}) = \left( \frac{q_1}{K}, \dots, \frac{q_N}{K}, \dots, \frac{q_1}{K}, \dots, \frac{q_N}{K} \right) \quad (6)$$

818 and corresponding output  $\tilde{Y}$ . The setup is visualized in figure 5 for the simplest case of  $N = 2$  and  
 819  $K = 2$ . There,  $Y = 1$  for example corresponds to either  $\tilde{Y} = 1$  or  $\tilde{Y} = 3$ . In general,

$$Y = j \iff \tilde{Y} = j + (k-1)N \text{ for some } k \in \{1, \dots, K\}.$$

820 By symmetry of the construction,

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\}] &= \sum_{j=1}^N \Pr[Y = j, j \in \{X^{(1)}, \dots, X^{(K)}\}] \\ &= \sum_{j=1}^N K \Pr[\tilde{Y} = j, j \in \{X^{(1)}, \dots, X^{(K)}\}] \\ &\geq \sum_{j=1}^N K \Pr[\tilde{Y} = j, X^{(1)} = j]. \end{aligned} \quad (7)$$

821 Since the analysis will be the same for any  $j$ , we now focus on finding the probability of the event  
 $\tilde{Y} = 1$  and  $X^{(1)} = 1$

$$\begin{aligned} \implies \frac{T_1}{\tilde{q}_1} &\leq \min_{2 \leq i \leq KN} \frac{T_i}{\tilde{q}_i} \text{ and } \frac{T_1}{p_1} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i} \\ \implies T_1 &\leq \min \left\{ \frac{T_2}{\max\{\tilde{q}_2/\tilde{q}_1, p_2/p_1\}}, \dots, \frac{T_N}{\max\{\tilde{q}_N/\tilde{q}_1, p_N/p_1\}}, \frac{T_{N+1}}{\tilde{q}_{N+1}/\tilde{q}_1}, \dots, \frac{T_{KN}}{\tilde{q}_{KN}/\tilde{q}_1} \right\}. \end{aligned}$$

822 The right-hand side is an exponential random variable independent of the left-hand side. If its  
 823 parameter is  $\lambda$  then, taking into account the definition of the  $\tilde{q}_i$ 's, we have

$$\begin{aligned} 1 + \lambda &= \sum_{i=1}^N \max \left\{ \frac{q_i}{q_1}, \frac{p_i}{p_1} \right\} + \sum_{k=2}^K \sum_{i=1}^N \frac{q_i}{q_1} \\ &= \sum_{i=1}^N \left[ \max \left\{ \frac{q_i}{q_1}, \frac{p_i}{p_1} \right\} + (K-1) \frac{q_i}{q_1} \right]. \end{aligned}$$

824 Therefore, by the properties of independent exponential random variables,

$$\Pr[\tilde{Y} = 1, X^{(1)} = 1] = \frac{1}{\sum_{i=1}^N [\max\{q_i/q_1, p_i/p_1\} + (K-1)q_i/q_1]}. \quad (8)$$

825 Combining with (7) establishes the bound in (3). We now turn our attention to (4) and find

$$\begin{aligned} \Pr[Y = j, Y \in \{X^{(1)}, \dots, X^{(K)}\}] &= \Pr[Y = j, j \in \{X^{(1)}, \dots, X^{(K)}\}] \\ &\geq K \Pr[\tilde{Y} = j, X^{(1)} = j] \end{aligned} \quad (9)$$

826 by the same symmetry argument used to show (7). Since our choice of  $j = 1$  in establishing (8) was  
827 arbitrary, we can apply that result for each  $j$  to get

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

828 Finally,

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j] &= \Pr[Y = j, Y \in \{X^{(1)}, \dots, X^{(K)}\}] / \Pr[Y = j] \\ &\geq \frac{K}{q_j \sum_{i=1}^N [\max\{q_i/q_j, p_i/p_j\} + (K-1)q_i/q_j]} \\ &= \frac{K}{K + q_j \sum_{i=1}^N \max\{0, p_i/p_j - q_i/q_j\}} \\ &\geq \frac{K}{K + q_j/p_j} \\ &= \left(1 + \frac{q_j}{Kp_j}\right)^{-1} \end{aligned}$$

829

□

830 As an aside, we can now easily find a related relaxed version of (3) by means of the relationship

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(2)}\}] &= \sum_{j=1}^N \Pr[Y = j] \Pr[Y \in \{X^{(1)}, \dots, X^{(2)}\} \mid Y = j] \\ &\geq \sum_{j=1}^N q_j \left(1 + \frac{q_j}{Kp_j}\right)^{-1} \end{aligned}$$

831 since  $\Pr[Y = j] = q_j$  by proposition 1.

### 832 A.3 Extension of proposition 1 to non-identically distributed proposals

833 We briefly consider the case where the proposals are drawn independently from  $K$  different dis-  
834 tributions. Let these distributions be  $p_X^{(1)}, \dots, p_X^{(K)}$  and define  $p_i^{(K)} := p_X^{(K)}(i)$  for convenience.

835 In this setting,  $X^{(k)}$  is sampled from the corresponding  $p_X^{(k)}$ , but the sampling procedure and the  
836 common randomness are otherwise identical to the setup in proposition 1. We then have the following  
837 extension of that result.

838 **Proposition 5.** *The procedure described above generates samples such that:*

- 839 1.  $\Pr[X^{(k)} = j] = p_j^{(k)}$  for all  $k \in \{1, \dots, K\}$  and  $j \in \{1, \dots, N\}$ .
- 840 2.  $\Pr[Y = j] = q_j$  for all  $j \in \{1, \dots, N\}$ .

841 *Proof.* The proof is very similar to that of proposition 1. We start by checking the distribution of any  
842  $X^{(k)}$  for  $1 \leq k \leq K$ . Note that

$$\begin{aligned} X^{(k)} = j &\implies \frac{S_j^{(k)}}{p_j^{(k)}} < \frac{S_i^{(k)}}{p_i^{(k)}} \quad \forall i \neq j \\ &\implies S_j^{(k)} < \min_{i \neq j} \frac{S_i^{(k)}}{p_i^{(k)}/p_j^{(k)}}. \end{aligned}$$

843 The left-hand side is an exponential random variable with parameter  $\lambda = 1$ , and the right-hand side is  
 844 an independent exponential random variable with parameter  $\lambda = \sum_{i \neq j} p_i^{(k)} / p_j^{(k)}$ . So,

$$\Pr[X^{(k)} = j] = \frac{1}{1 + \sum_{i \neq j} p_i^{(k)} / p_j^{(k)}} = p_j^{(k)}$$

845 as required. Next, note that the selection procedure for  $Y$  does not involve the  $p_i^{(k)}$ 's at all. Because  
 846 the introduction of these probabilities is the only deviation from the setting of proposition 1, the  
 847 correctness of  $Y$ 's distribution follows immediately from that result.  $\square$

#### 848 A.4 Proof of proposition 3

849 **Proposition 3.** *For any given  $\tau$  and for all  $1 \leq j \leq \tau$ , the output of algorithm 2 satisfies  $\Pr[Y_{1:j} =$   
 850  $y_{1:j}] = \mathcal{M}_b(y_{1:j} \mid \mathbf{c})$ . Also, algorithm 2 is conditionally drafter invariant in the sense of definition 1.*

851 *Proof.* We start by proving the first part of the proposition, which establishes sequence-level cor-  
 852 rectness. For convenience and to connect the setup more easily to GLS, we first abstract the details  
 853 of Gumbel-max sampling by defining sets of independent random variables  $\{\{S_i^{(j,k)}\}_{i=1}^N\}_{k=1}^K$  for  
 854 all  $1 \leq j \leq L$ , where  $S_i^{(j,k)} = -\ln U_i^{(j,k)}$ . We then have that each  $S_i^{(j,k)} \sim \text{Exp}(1)$ . Further let  
 855  $\mathcal{S}_{j-1}$  be the set of viable candidate drafts immediately before sampling  $Y_j$ , which allows us to keep  
 856 track of changes to  $\mathcal{S}$  in algorithm 2 during the proof. When the maximum draft length is  $L$ , we  
 857 have  $\tau \in \{1, \dots, L+1\}$  due to the possibility of selecting an extra token if the full draft sequence  
 858 is accepted. For any  $\tau$ , we need to verify the distribution of  $Y_1, \dots, Y_\tau$ , which we do by induction.  
 859 First, consider  $Y_1$ . Before the first step,  $\mathcal{S}_0 = \{1, \dots, K\}$ , so algorithm 2 makes the selection

$$Y_1 = \arg \min_{1 \leq i \leq N} \min_{1 \leq k \leq K} \frac{S_i^{(1,k)}}{q_i^{(1,k)}} = \arg \min_{1 \leq i \leq N} \min_{1 \leq k \leq K} \frac{S_i^{(1,k)}}{\mathcal{M}_b(i \mid \mathbf{c})}.$$

860 The second equality follows because the algorithm assigns  $q_i^{(1,k)} = \mathcal{M}_b(i \mid \mathbf{c})$  for all  $k$ . Then,  
 861  $\Pr[Y_1 = y_1] = \mathcal{M}_b(y_1 \mid \mathbf{c})$  for all  $y_1 \in \Omega$ , by proposition 1. Next, the rejection loop keeps only the  
 862 drafts  $k$  satisfying  $X_1^{(k)} = Y_1$ . Hence,  $X_1^{(k)} = Y_1$  for any  $k \in \mathcal{S}_1$ .

863 Next we look at the distribution of  $Y_2$ . Here, we get

$$Y_2 = \arg \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_1} \frac{S_i^{(2,k)}}{q_i^{(2,k)}} = \arg \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_1} \frac{S_i^{(2,k)}}{\mathcal{M}_b(i \mid X_1^{(k)}, \mathbf{c})}.$$

864 However, from the previous step we know that  $X_1^{(k)} = Y_1$  for any  $k \in \mathcal{S}_1$ . Therefore,  $\Pr[Y_2 = y_2 \mid$   
 865  $Y_1 = y_1] = \mathcal{M}_b(y_2 \mid y_1, \mathbf{c})$  by proposition 1 and consequently

$$\Pr[Y_{1:2} = y_{1:2}] = \mathcal{M}_b(y_2 \mid y_1, \mathbf{c}) \mathcal{M}_b(y_1 \mid \mathbf{c}) = \mathcal{M}_b(y_{1:2} \mid \mathbf{c}).$$

866 Moreover, if  $L > 1$ , the subsequent rejection stage removes from the set of viable candidates any  
 867 draft not satisfying  $X_2^{(k)} = Y_2$ , and so  $\mathcal{S}_2 = \{k \mid X_1^{(k)} = Y_1, X_2^{(k)} = Y_2\} = \{k \mid X_{1:2}^{(k)} = Y_{1:2}\}$ .

868 In general, assume that  $\Pr[Y_{1:j} = y_{1:j}] = \mathcal{M}_b(y_{1:j} \mid \mathbf{c})$  and  $X_{1:j}^{(k)} = Y_{1:j}$  for all  $k \in \mathcal{S}_j$ , for some  
 869  $1 \leq j < \tau$ . Implicitly,  $j < L+1$ , otherwise the generation would have already been completed. The  
 870 next token is selected according to

$$Y_{j+1} = \arg \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{q_i^{(j+1,k)}} = \arg \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{\mathcal{M}_b(i \mid X_{1:j}^{(k)}, \mathbf{c})} = \arg \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{\mathcal{M}_b(i \mid Y_{1:j}, \mathbf{c})}$$

871 and hence,  $\Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}] = \mathcal{M}_b(y_{j+1} \mid y_{1:j}, \mathbf{c})$ . Also,

$$\Pr[Y_{1:(j+1)} = y_{1:(j+1)}] = \mathcal{M}_b(y_{j+1} \mid y_{1:j}, \mathbf{c}) \mathcal{M}_b(y_{1:j} \mid \mathbf{c}) = \mathcal{M}_b(y_{1:(j+1)} \mid \mathbf{c}).$$

872 If  $j < L$ , the rejection step ensures that

$$\mathcal{S}_{j+1} = \mathcal{S}_j \cap \{k \mid X_{j+1}^{(k)} = Y_{j+1}\} = \{k \mid X_{1:j}^{(k)} = Y_{1:j}\} \cap \{k \mid X_{j+1}^{(k)} = Y_{j+1}\} = \{k \mid X_{1:j+1}^{(k)} = Y_{1:j+1}\}.$$

873 On the other had, if  $j = L$ , the entire sequence up to  $\tau$  is now complete because  $\tau \leq L + 1$ , and  
 874 there are no more sampling or rejection steps. This concludes the inductive step which, along with  
 875 the case for  $j = 1$ , makes up the proof for any  $1 \leq j \leq \tau$ .

876 We next prove our claim of conditional drafter invariance. The randomness is encapsulated by  
 877  $\mathcal{R} = \{\{S_i^{(j,k)}\}_{i=1}^N\}_{k=1}^K$ . Above, as an intermediate step to proving sequence-level correctness, we  
 878 showed by induction that for any  $0 \leq j < \tau$ , any  $k$  in  $\mathcal{S}_j$  satisfies  $X_{1:j}^{(k)} = Y_{1:j}$ . Then, looking at the  
 879 selection of  $Y_{j+1}$ , we have

$$Y_{j+1} = \arg \min \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{q_i^{(j+1,k)}} = \arg \min \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{\mathcal{M}_b(i \mid X_{1:j}^{(k)}, \mathbf{c})} = \arg \min \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{\mathcal{M}_b(i \mid Y_{1:j}, \mathbf{c})}$$

880 as seen previously. From this, given  $\mathcal{R}$ ,  $\mathbf{c}$  and  $Y_{1:j}$ ,  $Y_{j+1}$  only depends on the draft sequences through  
 881  $\mathcal{S}_j$ . Starting from  $Y_1$ , we see that since  $\mathcal{S}_0 = \{1, \dots, K\}$ ,

$$Y_1 = \arg \min_{1 \leq i \leq N} \min_{1 \leq k \leq K} \frac{S_i^{(1,k)}}{\mathcal{M}_b(i \mid \mathbf{c})}.$$

882 As the draft tokens do not play any role in this expression, we get

$$\Pr[Y_1 = y_1 \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}^{(k)}\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] = \Pr[Y_1 = y_1 \mid \mathcal{R}, \mathbf{c}]$$

883 which proves drafter invariance for  $Y_1$ . Note that we have written  $\{X_{1:L}^{(k)}\}_{k=1}^K$  instead of  
 884  $X_{1:L}^{(1)}, \dots, X_{1:L}^{(K)}$  to simplify the notation somewhat. Now  $Y_2$  is chosen according to

$$Y_2 = \arg \min \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_1} \frac{S_i^{(2,k)}}{q_i^{(2,k)}} = \arg \min \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_1} \frac{S_i^{(2,k)}}{\mathcal{M}_b(i \mid X_1^{(k)}, \mathbf{c})}.$$

885 Explicitly,  $\mathcal{S}_1 = \{k \mid X_1^{(k)} = Y_1\}$ . Hence, the choice of  $Y_2$  depends only on the values of the draft  
 886 tokens  $\{X_{1:L}^{(k)}\}_{k=1}^K$  and not on the language models used to generate them. We can then write

$$\begin{aligned} \Pr[Y_2 = y_2 \mid Y_1 = y_1, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_2 = y_2 \mid Y_1 = y_1, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \end{aligned}$$

887 for any choice of  $\mathcal{M}_s^{(1)}, \dots, \mathcal{M}_s^{(K)}$  and  $\tilde{\mathcal{M}}_s^{(1)}, \dots, \tilde{\mathcal{M}}_s^{(K)}$ . Also,

$$\begin{aligned} \Pr[Y_{1:2} = y_{1:2} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_2 = y_2 \mid Y_1 = y_1, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ \quad \times \Pr[Y_1 = y_1 \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_2 = y_2 \mid Y_1 = y_1, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ \quad \times \Pr[Y_1 = y_1 \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{1:2} = y_{1:2} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \end{aligned}$$

888 Therefore, conditional drafter invariance is satisfied for  $Y_{1:2}$ . To extend the general case and thus  
 889 complete the proof, we assume that  $Y_{1:j}$  satisfies conditional drafter invariance, where  $1 \leq j < \tau$ .  
 890 That is,

$$\begin{aligned} \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \end{aligned}$$

891 for any choice of  $\mathcal{M}_s^{(1)}, \dots, \mathcal{M}_s^{(K)}$  and  $\tilde{\mathcal{M}}_s^{(1)}, \dots, \tilde{\mathcal{M}}_s^{(K)}$ . Since

$$Y_{j+1} = \arg \min \min_{1 \leq i \leq N} \min_{k \in \mathcal{S}_j} \frac{S_i^{(j+1,k)}}{\mathcal{M}_b(i \mid Y_{1:j}, \mathbf{c})}$$

and  $\mathcal{S}_j = \{k \mid X_{1:j}^{(k)} = Y_{1:j}\}$ , we again see that  $Y_{j+1}$  only depends on the values of the draft tokens  $\{X_{1:L}^{(k)}\}_{k=1}^K$  and not on the underlying probability model. Therefore,

$$\begin{aligned} \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \end{aligned}$$

By the induction hypothesis,

$$\begin{aligned} \Pr[Y_{1:(j+1)} = y_{1:(j+1)} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ \times \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ \times \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{1:(j+1)} = y_{1:(j+1)} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\tilde{\mathcal{M}}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \end{aligned}$$

Since we have already shown that conditional drafter invariance holds for  $Y_1$ , it then holds for all sequences  $Y_{1:j}$ , where  $1 \leq j \leq \tau$ .  $\square$

## A.5 Proof of theorem 2

**Theorem 2** (Conditional LML). *Using the strategy above, the error probability satisfies*

$$\Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j, A = a, Z_1^K = z_1^K] \geq \sum_{k=1}^K \left( K + \frac{q_j(a)}{p_j(z_k)} \right)^{-1}.$$

*Proof.* We begin by following a similar approach to the proof of theorem 1, but we condition on  $A$  and the  $Z_k$ 's where necessary. Following the setup introduced in appendix A.2, we define

$$\{T_i\}_{i=1}^{KN} = \{S_1^{(1)}, \dots, S_N^{(1)}, S_1^{(2)}, \dots, S_N^{(2)}, \dots, S_1^{(K)}, \dots, S_N^{(K)}\}$$

and this time use a conditional augmented target distribution  $\tilde{q}_{\tilde{Y}|A}$  on  $\tilde{\Omega} = \{1, \dots, KN\}$  with output  $\tilde{Y}$ . Given  $A = a$ , this distribution is defined by the probabilities

$$(\tilde{q}_1(a), \dots, \tilde{q}_{KN}(a)) = \left( \frac{q_1(a)}{K}, \dots, \frac{q_N(a)}{K}, \dots, \frac{q_1(a)}{K}, \dots, \frac{q_N(a)}{K} \right).$$

With this setup, we find

$$\begin{aligned} \Pr[Y = j, j \in \{X^{(1)}, \dots, X^{(K)}\} \mid A = a, Z_1^K = z_1^K] \\ = \sum_{k=1}^K \Pr[\tilde{Y} = j + (k-1)N, j \in \{X^{(1)}, \dots, X^{(K)}\} \mid A = a, Z_1^K = z_1^K] \\ \geq \sum_{k=1}^K \Pr[\tilde{Y} = j + (k-1)N, X^{(k)} = j \mid A = a, Z_1^K = z_1^K]. \end{aligned} \quad (10)$$

As before, the analysis proceeds in the same manner regardless of the values of  $j$  and  $k$ . For simplicity, we therefore consider the probability

$$\begin{aligned} \Pr[\tilde{Y} = 1, X^{(1)} = 1 \mid A = a, Z_1^K = z_1^K] \\ = \Pr[X^{(1)} = 1 \mid \tilde{Y} = 1, A = a, Z_1^K = z_1^K] \Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K]. \end{aligned} \quad (11)$$

We start by computing

$$\begin{aligned} \Pr[X^{(1)} = 1 \mid \tilde{Y} = 1, A = a, Z_1^K = z_1^K] \\ = \Pr \left[ \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)} \mid \tilde{Y} = 1, A = a, Z_1^K = z_1^K \right] \\ = \Pr \left[ \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)} \mid \tilde{Y} = 1, A = a \right]. \end{aligned} \quad (12)$$

907 To get the second equality above, we note that by construction  $\{T_i\}_{i=1}^{KN} \rightarrow (Y, A) \rightarrow Z_1^K$  forms  
 908 a Markov chain, as noted in section 5.2. Since  $Y$  is a deterministic function of  $\tilde{Y}$ ,  $\{T_i\}_{i=1}^{KN} \rightarrow$   
 909  $(\tilde{Y}, A) \rightarrow Z_1^K$  is also a Markov chain and hence the  $T_i$ 's are conditionally independent of  $Z_1^K$  given  
 910  $\tilde{Y}$  and  $A$ . We can now leverage a similar analysis to that in appendix A.2 to compute the required  
 911 probability by first noting that

$$\begin{aligned} \Pr \left[ \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)} \mid \tilde{Y} = 1, A = a \right] \\ = \Pr \left[ \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)}, \tilde{Y} = 1 \mid A = a \right] / \Pr[\tilde{Y} = 1 \mid A = a]. \end{aligned} \quad (13)$$

912 Now, conditioned on  $A = a$ ,

$$\begin{aligned} \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)} \text{ and } \tilde{Y} = 1 \\ \implies \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)} \text{ and } \frac{T_1}{\tilde{q}_1(a)} \leq \min_{2 \leq i \leq KN} \frac{T_i}{\tilde{q}_i(a)} \\ \implies T_1 \leq \min \left\{ \frac{T_2}{\max\{\tilde{q}_2(a)/\tilde{q}_1(a), p_2(z_1)/p_1(z_1)\}}, \dots, \right. \\ \left. \frac{T_N}{\max\{\tilde{q}_N(a)/\tilde{q}_1(a), p_N(z_1)/p_1(z_1)\}}, \frac{T_{N+1}}{\tilde{q}_{N+1}(a)/\tilde{q}_1(a)}, \dots, \frac{T_{KN}}{\tilde{q}_{KN}(a)/\tilde{q}_1(a)} \right\}. \end{aligned}$$

913 We now note that the  $T_i$ 's are generated independently of the source  $A$ , and therefore remain i.i.d.  
 914  $\text{Exp}(1)$  random variables after conditioning on  $A = a$ . We can then follow our earlier approach  
 915 leveraging the properties of independent exponential random variables to see that the right-hand side  
 916 is an exponential random variable and is independent of the left-hand side. Let its parameter be  $\lambda$ .  
 917 Then, using the definition of the  $\tilde{q}_i(a)$ 's to simplify the result, we have

$$\begin{aligned} 1 + \lambda &= \sum_{i=1}^N \max \left\{ \frac{q_i(a)}{q_1(a)}, \frac{p_i(z_1)}{p_1(z_1)} \right\} + \sum_{k=2}^K \sum_{i=1}^N \frac{q_i(a)}{q_1(a)} \\ &= \sum_{i=1}^N \left[ \max \left\{ \frac{q_i(a)}{q_1(a)}, \frac{p_i(z_1)}{p_1(z_1)} \right\} + (K-1) \frac{q_i(a)}{q_1(a)} \right]. \end{aligned}$$

918 As before, we then get

$$\begin{aligned} \Pr \left[ \frac{T_1}{p_1(z_1)} \leq \min_{2 \leq i \leq N} \frac{T_i}{p_i(z_1)}, \tilde{Y} = 1 \mid A = a \right] \\ = \frac{1}{\sum_{i=1}^N [\max\{q_i(a)/q_1(a), p_i(z_1)/p_1(z_1)\} + (K-1)q_i(a)/q_1(a)]}. \end{aligned} \quad (14)$$

919 Putting together (11)–(14) gives

$$\begin{aligned} \Pr[\tilde{Y} = 1, X^{(1)} = 1 \mid A = a, Z_1^K = z_1^K] \\ = \frac{\Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K] / \Pr[\tilde{Y} = 1 \mid A = a]}{\sum_{i=1}^N [\max\{q_i(a)/q_1(a), p_i(z_1)/p_1(z_1)\} + (K-1)q_i(a)/q_1(a)]} \\ = \frac{q_1(a) \Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K] / \Pr[\tilde{Y} = 1 \mid A = a]}{K + q_1(a) \sum_{i=1}^N \max\{0, p_i(z_1)/p_1(z_1) - q_i(a)/q_1(a)\}} \\ \geq \frac{q_1(a)}{K + q_1(a)/p_1(z_1)} \frac{\Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K]}{\Pr[\tilde{Y} = 1 \mid A = a]} \\ = q_1(a) \left( K + \frac{q_1(a)}{p_1(z_1)} \right)^{-1} \frac{\Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K]}{\Pr[\tilde{Y} = 1 \mid A = a]}. \end{aligned}$$

920 We next note that, given  $A = a$ , the selection procedure for  $Y$  is

$$Y = \arg \min_{1 \leq i \leq N} \min_{1 \leq k \leq K} \frac{S_i^{(k)}}{q_i(a)}$$

921 and the  $S_i^{(k)}$ 's are i.i.d.  $\text{Exp}(1)$  random variables independent of  $A$ . As a result, the distribution  
 922 guarantee of proposition 1 still holds for  $Y$  such that  $Y$  is sampled exactly from  $q_{Y|A}$  and  $\Pr[Y =$   
 923  $1 \mid A = a] = q_1(a)$ . Then,

$$\Pr[\tilde{Y} = 1 \mid A = a] = q_1(a)/K \quad (15)$$

924 by symmetry and so

$$\Pr[\tilde{Y} = 1, X^{(1)} = 1 \mid A = a, Z_1^K = z_1^K] = \left(1 + \frac{q_1(a)}{K p_1(z_1)}\right)^{-1} \Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K]$$

925 or, generalizing to any  $j$  and  $k$ ,

$$\begin{aligned} \Pr[\tilde{Y} = j + (k-1)N, X^{(k)} = j \mid A = a, Z_1^K = z_1^K] \\ = \left(1 + \frac{q_j(a)}{K p_j(z_k)}\right)^{-1} \Pr[\tilde{Y} = j + (k-1)N \mid A = a, Z_1^K = z_1^K]. \end{aligned}$$

926 Going back to (10), we see that

$$\begin{aligned} \Pr[Y = j, j \in \{X^{(1)}, \dots, X^{(K)}\} \mid A = a, Z_1^K = z_1^K] \\ \geq \sum_{k=1}^K \left(1 + \frac{q_j(a)}{K p_j(z_k)}\right)^{-1} \Pr[\tilde{Y} = j + (k-1)N \mid A = a, Z_1^K = z_1^K]. \end{aligned}$$

927 Then, going from the joint to the conditional probability,

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j, A = a, Z_1^K = z_1^K] \\ = \frac{\Pr[Y = j, j \in \{X^{(1)}, \dots, X^{(K)}\} \mid A = a, Z_1^K = z_1^K]}{\Pr[Y = j \mid A = a, Z_1^K = z_1^K]} \\ \geq \sum_{k=1}^K \left(1 + \frac{q_j(a)}{K p_j(z_k)}\right)^{-1} \frac{\Pr[\tilde{Y} = j + (k-1)N \mid A = a, Z_1^K = z_1^K]}{\Pr[Y = j \mid A = a, Z_1^K = z_1^K]}. \end{aligned} \quad (16)$$

928 We now claim that

$$\Pr[\tilde{Y} = j + (k-1)N \mid A = a, Z_1^K = z_1^K] = \Pr[Y = j \mid A = a, Z_1^K = z_1^K]/K. \quad (17)$$

929 Without loss of generality, assume  $j = 1$  and  $k = 1$ . Applying Bayes' rule, we see that

$$\Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K] = \frac{\Pr[Z_1^K = z_1^K \mid \tilde{Y} = 1, A = a] \Pr[\tilde{Y} = 1 \mid A = a]}{\Pr[Z_1^K = z_1^K \mid A = a]}.$$

930 As seen earlier in (15),  $\Pr[\tilde{Y} = 1 \mid A = a] = \Pr[Y = 1 \mid A = a]/K$  by symmetry. Furthermore,  
 931 since  $Y$  is a deterministic function of  $\tilde{Y}$  and  $\tilde{Y} \rightarrow (Y, A) \rightarrow Z_1^K$  forms a Markov chain,

$$\begin{aligned} \Pr[Z_1^K = z_1^K \mid \tilde{Y} = 1, A = a] &= \Pr[Z_1^K = z_1^K \mid Y = 1, \tilde{Y} = 1, A = a] \\ &= \Pr[Z_1^K = z_1^K \mid Y = 1, A = a] \end{aligned}$$

932 and so

$$\begin{aligned} \Pr[\tilde{Y} = 1 \mid A = a, Z_1^K = z_1^K] &= \frac{\Pr[Z_1^K = z_1^K \mid Y = 1, A = a] \Pr[Y = 1 \mid A = a]/K}{\Pr[Z_1^K = z_1^K \mid A = a]} \\ &= \Pr[Y = 1 \mid A = a, Z_1^K = z_1^K]/K. \end{aligned}$$

933 Since the same argument works for arbitrary  $j$  and  $k$ , we have shown (17). Finally, substituting back  
 934 into (16) gives

$$\Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j, A = a, Z_1^K = z_1^K] \geq \sum_{k=1}^K \left(K + \frac{q_j(a)}{p_j(z_k)}\right)^{-1}$$

935

□



## 936 A.6 Proof of proposition 4

937 **Proposition 4.** For the coding scheme in section 5.1, the error probability is bounded above as

$$\Pr[Y \notin \{X^{(1)}, \dots, X^{(K)}\}] \leq 1 - \mathbb{E}_{A,W,T} \left[ \left( 1 + \frac{2^{i(W;A|T)}}{KL_{\max}} \right)^{-1} \right] \quad (5)$$

938 where  $i_{W,A|T}(w; a | t) = \log(p_{W|A}(w | a)/p_{W|T}(w | t))$  is the conditional information density.

939 *Proof.* To obtain the desired result, it is first necessary to identify the target distributions used at the  
 940 encoder and decoder in the coding scheme from section 5.1, then match these to  $q_{Y|A}$  and  $p_{X|Z}$  in  
 941 theorem 2. Doing this, we have  $q_{Y|A}(i | a) = p_{B|A}(B_i | a)$  and  $p_{X|Z}(i | t, \ell) = p_{B|Z}(B_i | t, \ell)$ .  
 942 Recall that we defined  $Z_k = (T_k, \ell_j)$  to encapsulate the side information and message available to  
 943 decoder  $k$  when the selected index is  $Y = j$ . We also defined  $B_i = (i, \ell_i)$ . Given  $T_k = t_k$ , each  $X^{(k)}$   
 944 is then sampled using the target distribution  $p_{X|Z}(\cdot | t_k, \ell_j)$ , and our sampling process generates  
 945  $X^{(1)}, \dots, X^{(K)}$ . Applying theorem 2, we get

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} | Y = j, A = a, Z_1^K = \{(t_k, \ell_j)\}_{k=1}^K] \\ &\geq \sum_{k=1}^K \left( K + \frac{q_{Y|A}(j | a)}{p_{X|Z}(j | t_k, \ell_j)} \right)^{-1} \\ &= \sum_{k=1}^K \left( K + \frac{p_{B|A}(j, \ell_j | a)}{p_{B|Z}(j, \ell_j | t_k, \ell_j)} \right)^{-1} \\ &= \sum_{k=1}^K \left( K + \frac{p_{W|A}(j | a)/L_{\max}}{p_{W|T}(j | t_k)} \right)^{-1} \\ &= \sum_{k=1}^K (K + 2^{i_{W,A|T}(j; a | t_k)}/L_{\max})^{-1} \end{aligned}$$

946 where  $i_{W,A|T}(j; a | t_k) = \log(p_{W|A}(j | a)/p_{W|T}(j | t_k))$  is the conditional information density.  
 947 After removing the conditioning, we get

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\}] &\geq \mathbb{E}_{A,W,T} \left[ \sum_{k=1}^K (K + 2^{i(W;A|T)}/L_{\max})^{-1} \right] \\ &= K \mathbb{E}_{A,W,T} [(K + 2^{i(W;A|T)}/L_{\max})^{-1}] \\ &= \mathbb{E}_{A,W,T} \left[ \left( 1 + \frac{2^{i(W;A|T)}}{KL_{\max}} \right)^{-1} \right]. \end{aligned}$$

948 By taking the complement we finally get a bound on the error probability,

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

949

□

## 950 B Connection to the notion of drafter invariance from Daliri et al. [9]

951 In Daliri et al. [9], an intuitive notion of drafter invariance is proposed for single-draft speculative  
 952 decoding with the following interpretation: given fixed random numbers, the output of the speculative  
 953 decoding algorithm is a function of the context and target model weights only. Our notion as stated in  
 954 definition 1 is somewhat weaker, since we allow the output to depend also on the draft sequences,  
 955 yet we still require that a given set of draft sequences always produce the same conditional output  
 956 distribution. To formalize the notion of Daliri et al. [9] and extend it to the multi-draft case, we  
 957 propose the following more restrictive definition, which we call *strong drafter invariance*.

**Definition 2** (Strong drafter invariance). Let  $\mathcal{R}$  be the source of randomness used to draw samples, e.g. the output of a random number generator. A multi-draft speculative decoding algorithm is strongly drafter invariant if, for all  $1 \leq j \leq \tau$ ,

$$\Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}, X_{1:L}^{(1)} = x_{1:L}^{(1)}, \dots, X_{1:L}^{(K)} = x_{1:L}^{(K)}] = \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}].$$

Unfortunately, satisfying definition 2 incurs a performance penalty in multi-draft implementations, which can be seen empirically from the extended results in appendix D.1. To see why, note that the choice of  $Y_j$  at each step  $j$  in algorithm 2 depends on the current set of active drafts  $\mathcal{S}$ , which itself is a function of the preceding draft tokens. To completely remove any dependence on the draft sequences as required by definition 2, we would need to keep  $\mathcal{S}$  fixed throughout the procedure, wastefully coupling the target model’s output to draft tokens that have already been rejected. Regardless, we now show how our method as described in algorithm 2 can be modified to support strong drafter invariance by way of the following proposition.

**Proposition 6.** If the minimum is taken over all  $k \in \{1, \dots, K\}$  in lines 9 and 13 of algorithm 2 instead of over  $k \in \mathcal{S}$ , strong drafter invariance holds in the sense of definition 2.

*Proof.* We extend the proof of conditional drafter invariance set out in appendix A.4. There, we showed that

$$\Pr[Y_1 = y_1 \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}^{(k)}\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] = \Pr[Y_1 = y_1 \mid \mathcal{R}, \mathbf{c}]$$

which is enough to prove strong drafter invariance for  $Y_1$ . As a result, we only need to modify the inductive step of the proof. Take some  $1 \leq j < \tau$ . With the modification that the minimum is taken over all  $k \in \{1, \dots, K\}$  instead of over  $k \in \mathcal{S}$  in lines 9 and 13 of algorithm 2,  $Y_{j+1}$  is selected as

$$Y_{j+1} = \arg \min_{1 \leq i \leq N} \min_{1 \leq k \leq K} \frac{S_i^{(j+1,k)}}{\mathcal{M}_b(i \mid Y_{1:j}, \mathbf{c})}.$$

From this, given  $\mathcal{R}, \mathbf{c}$  and  $Y_{1:j}$ , the draft tokens are not involved, either directly or indirectly, in the choice of  $Y_{j+1}$ . More precisely,

$$\begin{aligned} \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}]. \end{aligned}$$

Now assume that  $Y_{1:j}$  satisfies strong drafter invariance. As a result,

$$\begin{aligned} \Pr[Y_{1:(j+1)} = y_{1:(j+1)} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ \quad \times \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}, \{X_{1:L}(\mathcal{M}_s^{(k)})\}_{k=1}^K = \{x_{1:L}^{(k)}\}_{k=1}^K] \\ = \Pr[Y_{j+1} = y_{j+1} \mid Y_{1:j} = y_{1:j}, \mathcal{R}, \mathbf{c}] \Pr[Y_{1:j} = y_{1:j} \mid \mathcal{R}, \mathbf{c}] \\ = \Pr[Y_{1:(j+1)} = y_{1:(j+1)} \mid \mathcal{R}, \mathbf{c}] \end{aligned}$$

Therefore, strong drafter invariance holds for all sequences  $Y_{1:j}$ , where  $1 \leq j \leq \tau$ .  $\square$

We can also obtain a lower bound on the token-level acceptance probability of the strongly drafter-invariant scheme. Assume the drafts are i.i.d. from the same model  $\mathcal{M}_s$ , thus using the same setting as proposition 2. Reexamining the steps in the proof of theorem 1 found in appendix A.2, we see that if the number of active drafts at the current step is  $J \leq K$ , one of these  $J$  candidates must match the target for a token to be accepted. On the other hand, rejection now occurs if the target matches any of the remaining  $K - J$  drafts, or none at all. Specifically, assuming without loss of generality that the active drafts are  $X^{(1)}, \dots, X^{(J)}$ , (7) becomes

$$\Pr[Y \in \{X^{(1)}, \dots, X^{(J)}\}] \geq \sum_{j=1}^N J \Pr[\tilde{Y} = j, X^{(1)} = j].$$

However, the augmented target distribution described in (6) remains unchanged, because all  $K$  drafts remain coupled through the common randomness. This includes the  $K - J$  drafts that have already

been rejected during previous steps. Following the same analysis as in appendix A.2 then shows that the probability of accepting at least one token at the current step with context  $c$  is bounded below as

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

where  $p_X := \mathcal{M}_s(\cdot | c)$  and  $q_Y := \mathcal{M}_b(\cdot | c)$ . Conversely, with  $J$  active drafts proposition 2 gives

$$\Pr[\text{accept}] \geq \sum_{j=1}^N \frac{J}{\sum_{i=1}^N [\max\{q_i/q_j, p_i/p_j\} + (J-1)q_i/q_j]}.$$

for the original conditionally drafter-invariant algorithm. Since  $J \leq K$ , we can see that requiring strong drafter invariance reduces the lower bound, providing some theoretical insight into the poor performance observed in appendix D.1 when using this scheme for LLM inference.

## C Extension to continuous distributions via importance sampling

It is not possible to enumerate the entire sample space when dealing with continuous distributions. At first glance, this appears to preclude the use of GLS in such settings. However, we can obtain approximate samples through importance sampling as described in Phan et al. [31]. For clarity, we skip the general case and instead proceed directly to the source coding application set out in section 5.1 of the main paper, removing the assumption that  $W$  is discrete. To start, we choose some sufficiently large  $N$  and generate  $N$  i.i.d. samples of  $W$  following  $p_W$ . Let these samples be  $U_1^N := \{U_i\}_{i=1}^N$ . Recall that  $p_W$  is the marginal target distribution for the decoder’s output. Again, let  $\ell_1, \dots, \ell_N$  be uniform random integers selected from  $\{1, \dots, L_{\max}\}$ . We then have the tuples  $B_i = (U_i, \ell_i)$  forming part of the common randomness, the difference from section 5.1 being that the  $U_i$ ’s are now also random, whereas in the discrete case they were fixed to enumerate the whole sample space.

At the encoder, we want to sample  $Y$  approximately from  $p_{B|A}$  given the observation  $A = a$ . Therefore, we introduce the *unnormalized* importance weights

$$\tilde{\lambda}_{q,i}(U_i) = \frac{p_{B|A}(B_i | a)}{p_B(B_i)} = \frac{p_{W|A}(U_i | a)}{p_W(U_i)}.$$

Similarly, the decoders should use the target distribution  $p_{B|Z}$ , where decoder  $k$  has access to  $Z_k = (T_k, \ell_j)$ . Here,  $j$  is the selected index at the encoder and  $\ell_j$  is the associated random integer. Given  $T_k = t_k$ , we define

$$\tilde{\lambda}_{p,i}^{(k)}(U_i) = \frac{p_{B|Z}(B_i | t_k, \ell_j)}{p_B(B_i)} = \frac{p_{W|T}(U_i | t_k) \mathbb{1}\{\ell_i = \ell_j\}}{p_W(U_i)/L_{\max}}.$$

The final normalized importance weights are

$$\lambda_{q,i}(U_1^N) = \frac{\tilde{\lambda}_{q,i}(U_i)}{\sum_{j=1}^N \tilde{\lambda}_{q,j}(U_j)} \text{ and } \lambda_{p,i}^{(k)}(U_1^N) = \frac{\tilde{\lambda}_{p,i}^{(k)}(U_i)}{\sum_{j=1}^N \tilde{\lambda}_{p,j}^{(k)}(U_j)}.$$

The index selection at the encoder and decoder simply proceeds as

$$Y = \arg \min_{1 \leq i \leq N} \min_{1 \leq k \leq K} \frac{S_i^{(k)}}{\lambda_{q,i}(U_1^N)} \text{ and } X^{(k)} = \arg \min_{1 \leq i \leq N} \frac{S_i^{(k)}}{\lambda_{p,i}^{(k)}(U_1^N)}.$$

The output of decoder  $k$  is then taken to be  $W_k = U_{X^{(k)}}$ . Conditional on  $U_1^N$ , the bound on the index matching probability stated in proposition 4 holds. However, we need a bound that is conditional only on  $U_j$ . To start, we write down the bound when conditioning on  $U_1^N$ , drawing from the development in appendix A.6.

$$\begin{aligned} \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} | Y = j, A = a, Z_1^K = z_1^K, U_1^N = u_1^N] \\ \geq \sum_{k=1}^K \left( K + \frac{\lambda_{q,j}(u_1^N)}{\lambda_{p,j}^{(k)}(u_1^N)} \right)^{-1} \\ = \sum_{k=1}^K \left( K + \frac{\tilde{\lambda}_{q,j}(u_j) \sum_{i=1}^N \tilde{\lambda}_{p,i}^{(k)}(u_1^N)}{\tilde{\lambda}_{p,j}^{(k)}(u_j) \sum_{i=1}^N \tilde{\lambda}_{q,i}(u_1^N)} \right)^{-1}. \end{aligned}$$

For simplicity, let us assume for now without loss of generality that  $j = 1$ . We want to remove the conditioning on  $U_2, \dots, U_N$ , which can be done by taking the expectation to get

$$\begin{aligned}
& \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = 1, A = a, Z_1^K = z_1^K, U_1 = u_1] \\
& \geq \mathbb{E}_{U_2^N} \left[ \sum_{k=1}^K \left( K + \frac{\tilde{\lambda}_{q,1}(u_1)}{\tilde{\lambda}_{p,1}^{(k)}(u_1)} \frac{\sum_{i=1}^N \tilde{\lambda}_{p,i}^{(k)}(U_1^N)}{\sum_{i=1}^N \tilde{\lambda}_{q,i}(U_1^N)} \right)^{-1} \middle| Y = 1, A = a, Z_1^K = z_1^K, U_1 = u_1 \right] \\
& = \sum_{k=1}^K \mathbb{E}_{U_2^N} \left[ \left( K + \frac{\tilde{\lambda}_{q,1}(u_1)}{\tilde{\lambda}_{p,1}^{(k)}(u_1)} \frac{\sum_{i=1}^N \tilde{\lambda}_{p,i}^{(k)}(U_1^N)}{\sum_{i=1}^N \tilde{\lambda}_{q,i}(U_1^N)} \right)^{-1} \middle| Y = 1, A = a, Z_1^K = z_1^K, U_1 = u_1 \right] \\
& \geq \sum_{k=1}^K \left( K + \frac{\tilde{\lambda}_{q,1}(u_1)}{\tilde{\lambda}_{p,1}^{(k)}(u_1)} \mathbb{E}_{U_2^N} \left[ \frac{\sum_{i=1}^N \tilde{\lambda}_{p,i}^{(k)}(U_1^N)}{\sum_{i=1}^N \tilde{\lambda}_{q,i}(U_1^N)} \middle| Y = 1, A = a, Z_1^K = z_1^K, U_1 = u_1 \right] \right)^{-1}
\end{aligned}$$

where the last step uses Jensen's inequality. To simplify the inner expectation, we use the following lemma, stated here without proof, which extracted from the proof of theorem 3 in Phan et al. [31].

**Lemma 1** ([31, p. 23]). *We have that*

$$\mathbb{E}_{U_2^N} \left[ \frac{\sum_{i=1}^N \tilde{\lambda}_{p,i}^{(k)}(U_1^N)}{\sum_{i=1}^N \tilde{\lambda}_{q,i}(U_1^N)} \middle| Y = 1, A = a, Z_1^K = z_1^K, U_1 = u_1 \right] \leq \mu_k(N, u_1) \quad (18)$$

where

$$\mu_k(N, u_1) = \frac{\tilde{\lambda}_{p,1}^{(k)}(u_1) + \bar{N}}{\tilde{\lambda}_{q,1}(u_1) + \bar{N}} + \frac{K(\bar{N})}{\bar{N}} \left( 1 + \frac{\tilde{\lambda}_{q,1}(u_1)}{\bar{N}} \right) + \frac{2\omega L(\bar{N})}{\bar{N}} \left( 1 + \frac{\tilde{\lambda}_{q,1}(u_1)}{\bar{N}} \right).$$

In the equation above, we define  $\bar{N} = N - 1$  and

$$\begin{aligned}
K(\bar{N}) &= \frac{4(\omega - 1)}{(1 + \tilde{\lambda}_{q,1}(u_1)/\bar{N})^2} \left( 1 + \frac{(N + 1)\omega}{\bar{N}} \right) \\
&\quad \times \sqrt{2 + 4 \left( \frac{\tilde{\lambda}_{p,1}^{(k)}(u_1) + \bar{N}}{2\tilde{\lambda}_{q,1}(u_1) + \bar{N}} \right)^2 \left[ \left( 1 + \frac{(N + 1)\omega}{\bar{N}} \right)^2 + \frac{\omega - 1}{\bar{N}} \right]} \\
L(\bar{N}) &= \sqrt{\omega - 1} \sqrt{d_5(p_W \| p_{W|A}(\cdot | a)) - d_3(p_W \| p_{W|A}(\cdot | a))^2} \\
&\quad + (\omega - 1)d_3(p_W \| p_{W|A}(\cdot | a))
\end{aligned}$$

where, for all  $m \geq 1$ ,

$$d_{m+1}(p_W, p_{W|A}(\cdot | a)) = \mathbb{E}_{W \sim p_W} \left[ \frac{p_W(W)^m}{p_{W|A}(W | a)^m} \right]$$

and  $\omega$  is chosen so that, for all  $a$  and  $w$ ,  $p_{W|A}(w | a)/p_W(w) \leq \omega$ .

Using lemma 1, or more specifically (18), and generalizing to arbitrary  $j$ , we see that

$$\begin{aligned}
& \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j, A = a, Z_1^K = z_1^K, U_j = u_j] \\
& \geq \sum_{k=1}^K \left( K + \mu_k(N, u_j) \frac{\tilde{\lambda}_{q,j}(u_j)}{\tilde{\lambda}_{p,j}^{(k)}(u_j)} \right)^{-1} \\
& \geq \sum_{k=1}^K \left( K + \hat{\mu}(N, u_j) \frac{\tilde{\lambda}_{q,j}(u_j)}{\tilde{\lambda}_{p,j}^{(k)}(u_j)} \right)^{-1}
\end{aligned}$$

1028 where  $\hat{\mu}(N, u_j) := \max_{1 \leq k \leq K} \mu_k(N, u_j)$ . Using the definitions of  $\tilde{\lambda}_{q,j}(u_j)$  and  $\tilde{\lambda}_{p,j}^{(k)}(u_j)$ ,

$$\begin{aligned} & \Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\} \mid Y = j, A = a, Z_1^K = z_1^K, U_j = u_j] \\ & \geq \sum_{k=1}^K \left( K + \hat{\mu}(N, u_j) \frac{p_{W|A}(u_j \mid a) / L_{\max}}{p_{W|T}(u_j \mid t_k)} \right)^{-1} \\ & = \sum_{k=1}^K \left( K + \hat{\mu}(N, u_j) 2^{i_{W,A|T}(u_j; a|t_k)} / L_{\max} \right)^{-1}. \end{aligned}$$

1029 To proceed and find a counterpart to the coding theorem in proposition 4, we use the fact that for  
 1030 any  $\varepsilon > 0$ , there exists some  $M_k$  such that  $\mu_k(N, u_j) \leq 1 + \varepsilon$  for all  $N \geq M_k$  and  $1 \leq j \leq N$   
 1031 [31, p. 24]. Hence, with  $M = \max_{1 \leq k \leq K} M_k$ , we see that  $\hat{\mu}(N, u_j) \leq 1 + \varepsilon$  for all  $N \geq M$  and  
 1032  $1 \leq j \leq N$ . As a result, after removing the conditioning and following the steps used in the proof of  
 1033 proposition 4 in appendix A.6, we get

$$\Pr[Y \in \{X^{(1)}, \dots, X^{(K)}\}] \geq \mathbb{E}_{A,W,T} \left[ \left( 1 + (1 + \varepsilon) \frac{2^{i(W;A|T)}}{KL_{\max}} \right)^{-1} \right]$$

1034 and the associated error probability bound is

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

## 1035 D Additional experimental details

### 1036 D.1 Multi-draft speculative decoding

1037 **Proof of concept on toy distributions.** As a simple demonstration of our method with arbitrary  
 1038 discrete distributions, we generate 100 random instances of  $p_X$  and  $q_Y$  each containing  $N =$   
 1039 10 elements, while the number of proposals is varied between 1 and 20. Results are shown in  
 1040 figure 6. As well as showing the token-level matching rate achieved by SpecTr [33], SpecInfer  
 1041 [29] and our algorithm, we also plot the optimal multi-draft acceptance rate *with* communication,  
 1042 which can be computed via a linear programming approach [33], at least for distributions on small  
 1043 alphabets. Note that while this calculation provides a useful upper bound, there is currently no  
 1044 multi-draft token selection scheme that can achieve the optimum in practice. Despite involving no  
 1045 communication between the drafter and the target, our algorithm is competitive with state-of-the-art  
 1046 methods, especially when the number of drafts is large.

1047 **LLM inference.** Implementations of SpecInfer [29], SpecTr [33] and our drafter-invariant schemes  
 1048 can be found in the provided code. To obtain performance measurements, each speculative decoding  
 1049 configuration is tested on 200 prompts from the GSM8K [7], NaturalReasoning [41], MBPP [1] and  
 1050 DROP<sup>1</sup> datasets, and 164 prompts from HumanEval [6]. Our full set of results is given in tables 3  
 1051 and 4; please note that not all datasets and configurations are reported in the main paper. We also  
 1052 include results for the strongly drafter-invariant scheme described in appendix B. The target model is  
 1053 Qwen 2.5-7 B [40] while the drafter is Qwen 2.5-0.5 B, and we use top-K sampling with  $K = 50$ .  
 1054 For our experiments with i.i.d. drafts in table 3, the temperature is 1.0 throughout and the maximum  
 1055 draft length is  $L = 4$ . When we use diverse drafts in table 4, the target temperature is 2.0, the two  
 1056 draft temperatures are varied and  $L = 5$ .

1057 As described in the main text, the block efficiency is equal to the average number of tokens accepted  
 1058 during each iteration of the speculative decoding algorithm, while token rates are calculated from  
 1059 wall-clock measurements and reported as percentage speedups relative to single-draft speculative  
 1060 decoding with the same draft length. We compute the mean across all prompts for each dataset,  
 1061 then repeat the experiments 5 times with different random seeds. For a configuration-dataset pair,

<sup>1</sup>D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–2378, 2019.

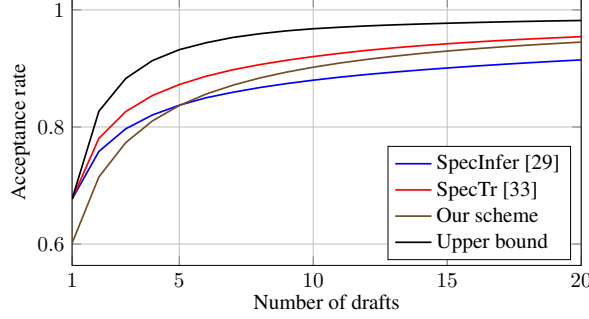


Figure 6: Proof of concept on random toy distributions.

1062 this gives us five measurements each for the average block efficiency and token rate. Let these be  
 1063  $BE_1, \dots, BE_5$  and  $TR_1, \dots, TR_5$ . As our final result, we report

$$BE = \text{mean}(BE_1, \dots, BE_5), \quad TR = \text{mean}(TR_1, \dots, TR_5)$$

1064 while the error bars show one standard error of the mean,

$$\sigma_{BE} = \text{std}(BE_1, \dots, BE_5)/\sqrt{5}, \quad \sigma_{TR} = \text{std}(TR_1, \dots, TR_5)/\sqrt{5}$$

1065 where the std operator is the usual sample standard deviation formula. If  $x$  is a vector of  $M$   
 1066 independent measurements in  $\mathbb{R}$ , this is

$$\text{std}(x) = \sqrt{\frac{\sum_{i=1}^M (x_i - \bar{x})^2}{M-1}}, \quad \bar{x} = \text{mean}(x).$$

1067 On a NVIDIA RTX6000 Ada GPU with 48 GB of memory, each configuration can be tested on a  
 1068 single dataset in around 1 hour.

## 1069 D.2 Synthetic Gaussian source

1070 In this section, we provide some key derivations and details of our experimental procedure for  
 1071 evaluating GLS on the synthetic Gaussian source, and give more numerical results.

1072 **Decoder target distribution.** To determine the decoder target distribution  $p_{W|T}$ , we first recapitulate  
 1073 the problem setting and the random variables involved. The Gaussian source is  $A \sim \mathcal{N}(0, 1)$  while  
 1074 the target distribution at the encoder given  $A = a$  is  $p_{W|A}(\cdot | a) = \mathcal{N}(a, \sigma_{W|A}^2)$ . Meanwhile, the  
 1075 side information at decoder  $k$  is  $T_k = A + \zeta_k$ , where  $\zeta_k \sim \mathcal{N}(0, \sigma_{T|A}^2)$ . Since we are only analyzing  
 1076 one decoder individually, we will drop the  $k$  subscript in what follows and more simply write  $T$  and  
 1077  $\zeta$ . To summarize, we have

$$W = A + \eta \text{ and } T = A + \zeta$$

1078 where  $\eta$  and  $\zeta$  are independent zero-mean Gaussians with variances  $\sigma_\eta^2 = \sigma_{W|A}^2$  and  $\sigma_\zeta^2 = \sigma_{T|A}^2$   
 1079 respectively. From this, we see that  $W$  and  $T$  are jointly distributed as

$$\begin{bmatrix} W \\ T \end{bmatrix} \sim \mathcal{N}(0, \Sigma_{W,T}), \text{ where } \Sigma_{W,T} = \begin{bmatrix} \mathbb{E}[W^2] & \mathbb{E}[WT] \\ \mathbb{E}[TW] & \mathbb{E}[T^2] \end{bmatrix} = \begin{bmatrix} 1 + \sigma_\eta^2 & 1 \\ 1 & 1 + \sigma_\zeta^2 \end{bmatrix}. \quad (19)$$

1080 The variance of  $W$  is  $\sigma_W^2 = 1 + \sigma_\eta^2$  and that of  $T$  is  $\sigma_T^2 = 1 + \sigma_\zeta^2$ . We then know that  $p_{W|T}$  is a  
 1081 Gaussian distribution with mean and variance

$$\mu_{W|T} = \frac{T}{1 + \sigma_\zeta^2} = \frac{T}{\sigma_T^2}, \quad \sigma_{W|T}^2 = 1 + \sigma_\eta^2 - \frac{1}{1 + \sigma_\zeta^2} = \sigma_W^2 - \frac{1}{\sigma_T^2}.$$

1082 That is,  $p_{W|T}(\cdot | t) = \mathcal{N}(t/\sigma_T^2, \sigma_W^2 - 1/\sigma_T^2)$  as asserted in the main paper.

1083 **MMSE estimator.** We now derive the MMSE estimator for the synthetic Gaussian source when side  
 1084 information is available at the decoder. To find the estimator, we assume that the encoder and decoder  
 1085 indices match, i.e.  $W_k = W$ . Recall that proposition 4 in the main paper gives a lower bound for the  
 1086 probability of this event. We proceed by finding the joint distribution of  $A$ ,  $W$  and  $T$ . In fact,

$$\begin{bmatrix} A \\ W \\ T \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \Sigma_{A,(W,T)} \\ \Sigma_{(W,T),A} & \Sigma_{W,T} \end{bmatrix}\right), \text{ where } \Sigma_{(W,T),A} = \Sigma_{A,(W,T)}^T = \begin{bmatrix} \mathbb{E}[AW] \\ \mathbb{E}[AT] \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Strategy	$K$	GSM8K		HumanEval		NaturalReasoning		MBPP		DROP	
		BE	TR (%)	BE	TR (%)	BE	TR (%)	BE	TR (%)	BE	TR (%)
SpecInfer [29]	2	4.47 $\pm$ 0.01	3.75 $\pm$ 0.21	4.11 $\pm$ 0.02	6.07 $\pm$ 0.53	3.75 $\pm$ 0.01	6.26 $\pm$ 0.26	4.04 $\pm$ 0.01	6.51 $\pm$ 0.51	3.33 $\pm$ 0.01	8.08 $\pm$ 0.61
	4	4.64 $\pm$ 0.01	5.20 $\pm$ 0.23	4.35 $\pm$ 0.01	8.69 $\pm$ 0.27	3.99 $\pm$ 0.01	10.63 $\pm$ 0.32	4.30 $\pm$ 0.01	10.92 $\pm$ 0.53	3.58 $\pm$ 0.00	10.46 $\pm$ 0.19
	6	4.72 $\pm$ 0.00	5.39 $\pm$ 0.23	4.46 $\pm$ 0.01	8.60 $\pm$ 0.92	4.12 $\pm$ 0.01	12.35 $\pm$ 0.31	4.41 $\pm$ 0.01	12.65 $\pm$ 0.45	3.72 $\pm$ 0.01	10.98 $\pm$ 0.41
	8	4.75 $\pm$ 0.00	4.56 $\pm$ 0.20	4.54 $\pm$ 0.01	7.89 $\pm$ 0.85	4.19 $\pm$ 0.01	12.49 $\pm$ 0.56	4.49 $\pm$ 0.01	13.81 $\pm$ 0.47	3.82 $\pm$ 0.01	10.60 $\pm$ 0.24
SpecTr [33]	2	4.46 $\pm$ 0.01	2.27 $\pm$ 0.29	4.13 $\pm$ 0.01	4.80 $\pm$ 1.04	3.76 $\pm$ 0.01	5.44 $\pm$ 0.54	4.04 $\pm$ 0.01	5.21 $\pm$ 0.43	3.31 $\pm$ 0.01	6.35 $\pm$ 0.34
	4	4.66 $\pm$ 0.01	4.08 $\pm$ 0.13	4.36 $\pm$ 0.01	7.39 $\pm$ 0.96	4.02 $\pm$ 0.01	10.00 $\pm$ 0.33	4.32 $\pm$ 0.01	10.07 $\pm$ 0.43	3.60 $\pm$ 0.02	10.03 $\pm$ 0.55
	6	4.74 $\pm$ 0.01	4.30 $\pm$ 0.35	4.48 $\pm$ 0.01	7.70 $\pm$ 0.78	4.13 $\pm$ 0.01	11.51 $\pm$ 0.43	4.43 $\pm$ 0.00	11.77 $\pm$ 0.29	3.72 $\pm$ 0.01	9.97 $\pm$ 0.34
	8	4.78 $\pm$ 0.00	3.75 $\pm$ 0.17	4.56 $\pm$ 0.01	7.09 $\pm$ 0.77	4.22 $\pm$ 0.01	12.89 $\pm$ 0.96	4.50 $\pm$ 0.01	12.81 $\pm$ 0.50	3.79 $\pm$ 0.01	8.82 $\pm$ 0.32
Our scheme	2	4.47 $\pm$ 0.00	3.02 $\pm$ 0.28	4.08 $\pm$ 0.01	4.70 $\pm$ 0.78	3.68 $\pm$ 0.02	4.52 $\pm$ 1.42	4.02 $\pm$ 0.00	5.46 $\pm$ 0.27	3.27 $\pm$ 0.01	5.74 $\pm$ 0.57
	4	4.66 $\pm$ 0.01	5.18 $\pm$ 0.26	4.37 $\pm$ 0.01	8.52 $\pm$ 0.77	3.96 $\pm$ 0.01	9.86 $\pm$ 0.96	4.30 $\pm$ 0.01	10.72 $\pm$ 0.28	3.56 $\pm$ 0.00	9.98 $\pm$ 0.30
	6	4.74 $\pm$ 0.01	5.33 $\pm$ 0.25	4.47 $\pm$ 0.01	8.54 $\pm$ 1.02	4.10 $\pm$ 0.01	12.14 $\pm$ 0.85	4.43 $\pm$ 0.01	12.79 $\pm$ 0.46	3.73 $\pm$ 0.01	10.90 $\pm$ 0.26
	8	4.78 $\pm$ 0.00	4.83 $\pm$ 0.24	4.55 $\pm$ 0.01	8.03 $\pm$ 0.97	4.18 $\pm$ 0.00	12.75 $\pm$ 1.14	4.51 $\pm$ 0.01	13.54 $\pm$ 0.41	3.82 $\pm$ 0.01	10.60 $\pm$ 0.34
Strongly invariant	2	4.45 $\pm$ 0.01	3.02 $\pm$ 0.41	4.03 $\pm$ 0.01	3.18 $\pm$ 0.87	3.65 $\pm$ 0.01	3.56 $\pm$ 1.10	4.00 $\pm$ 0.02	4.29 $\pm$ 0.60	3.25 $\pm$ 0.02	3.74 $\pm$ 0.24
	4	4.63 $\pm$ 0.00	4.45 $\pm$ 0.57	4.28 $\pm$ 0.01	6.23 $\pm$ 0.76	3.91 $\pm$ 0.01	8.25 $\pm$ 1.01	4.26 $\pm$ 0.01	8.71 $\pm$ 0.17	3.53 $\pm$ 0.01	7.87 $\pm$ 0.44
	6	4.72 $\pm$ 0.00	5.23 $\pm$ 0.47	4.40 $\pm$ 0.01	6.69 $\pm$ 0.55	4.02 $\pm$ 0.01	9.84 $\pm$ 1.08	4.38 $\pm$ 0.01	10.85 $\pm$ 0.44	3.67 $\pm$ 0.00	7.92 $\pm$ 0.66
	8	4.76 $\pm$ 0.01	4.49 $\pm$ 0.44	4.46 $\pm$ 0.01	5.99 $\pm$ 0.69	4.09 $\pm$ 0.01	10.32 $\pm$ 1.13	4.44 $\pm$ 0.01	11.46 $\pm$ 0.48	3.75 $\pm$ 0.02	7.40 $\pm$ 0.99
Daliri et al. [9]	1	4.16 $\pm$ 0.01	0.63 $\pm$ 0.24	3.69 $\pm$ 0.01	0.04 $\pm$ 0.30	3.32 $\pm$ 0.00	-2.23 $\pm$ 0.29	3.61 $\pm$ 0.01	-0.69 $\pm$ 0.33	2.94 $\pm$ 0.01	0.10 $\pm$ 0.46

Table 3: LLM inference with i.i.d. drafts; we use  $L = 4$ . Token rates (TR) are shown as percentage speedups relative to single-draft speculative decoding, which has mean block efficiency (BE) 4.18, 3.75, 3.43, 3.68 and 3.00 on GSM8K, HumanEval, NaturalReasoning, MBPP and DROP respectively.

Strategy	Temp.	GSM8K		HumanEval		NaturalReasoning		MBPP		DROP	
		BE	TR (%)	BE	TR (%)	BE	TR (%)	BE	TR (%)	BE	TR (%)
SpecInfer [29]	0.5/1.0	4.26 ± 0.02	0.06 ± 1.02	3.57 ± 0.02	-1.96 ± 0.67	3.04 ± 0.02	-6.55 ± 0.61	3.66 ± 0.01	-1.87 ± 0.79	3.21 ± 0.21	4.22 ± 0.99
	1.0/0.5	4.44 ± 0.03	4.57 ± 1.80	3.80 ± 0.03	4.13 ± 1.60	3.29 ± 0.02	1.12 ± 0.99	3.90 ± 0.02	4.77 ± 0.55	3.31 ± 0.02	7.83 ± 1.08
	1.5/1.0	4.63 ± 0.02	8.75 ± 1.70	3.99 ± 0.03	9.67 ± 1.26	3.60 ± 0.02	11.28 ± 0.88	4.05 ± 0.01	10.70 ± 0.59	3.31 ± 0.02	9.11 ± 0.74
	1.0/1.5	4.57 ± 0.03	7.43 ± 1.60	3.95 ± 0.02	8.19 ± 0.82	3.44 ± 0.01	6.06 ± 0.37	3.98 ± 0.02	8.48 ± 0.91	3.30 ± 0.02	9.15 ± 0.83
	2.0/1.0	4.53 ± 0.01	6.38 ± 1.43	3.88 ± 0.01	6.75 ± 0.72	3.51 ± 0.02	9.12 ± 0.98	3.91 ± 0.01	6.83 ± 0.63	3.19 ± 0.01	5.50 ± 0.44
	1.0/2.0	4.55 ± 0.02	6.53 ± 1.79	3.82 ± 0.02	4.73 ± 1.22	3.42 ± 0.02	5.71 ± 0.73	3.91 ± 0.01	6.89 ± 0.33	3.21 ± 0.02	6.28 ± 0.79
	1.0/1.0	4.51 ± 0.02	6.02 ± 1.35	3.87 ± 0.02	6.32 ± 0.36	3.36 ± 0.01	3.37 ± 0.44	3.95 ± 0.02	5.17 ± 1.13	3.30 ± 0.02	9.36 ± 0.93
Our scheme	0.5/1.0	4.75 ± 0.02	11.50 ± 1.78	4.00 ± 0.01	9.80 ± 0.82	3.27 ± 0.01	-0.37 ± 0.15	3.94 ± 0.02	5.64 ± 0.66	3.21 ± 0.01	4.67 ± 0.66
	1.0/0.5	4.75 ± 0.02	11.40 ± 1.58	3.96 ± 0.02	8.77 ± 0.99	3.25 ± 0.01	-0.94 ± 0.57	3.96 ± 0.02	5.99 ± 1.01	3.23 ± 0.01	4.79 ± 0.51
	1.5/1.0	4.77 ± 0.02	11.37 ± 1.81	4.03 ± 0.01	10.42 ± 0.41	3.41 ± 0.01	4.58 ± 0.36	4.02 ± 0.02	9.59 ± 0.83	3.22 ± 0.01	6.26 ± 0.51
	1.0/1.5	4.77 ± 0.02	11.56 ± 1.57	3.99 ± 0.03	9.17 ± 0.98	3.42 ± 0.02	4.71 ± 0.88	4.00 ± 0.01	9.00 ± 0.63	3.23 ± 0.02	6.19 ± 0.43
	2.0/1.0	4.63 ± 0.02	8.15 ± 1.16	3.87 ± 0.02	5.79 ± 0.77	3.31 ± 0.01	1.74 ± 0.69	3.86 ± 0.03	4.95 ± 0.64	3.12 ± 0.01	2.80 ± 0.36
	1.0/2.0	4.62 ± 0.03	8.25 ± 1.29	3.87 ± 0.01	6.08 ± 0.72	3.32 ± 0.01	1.67 ± 0.56	3.88 ± 0.02	5.74 ± 0.43	3.13 ± 0.01	3.17 ± 0.76
	1.0/1.0	4.83 ± 0.02	13.68 ± 1.67	4.08 ± 0.02	12.15 ± 0.83	0.00 ± 0.01	4.79 ± 0.91	4.08 ± 0.01	8.57 ± 0.60	3.27 ± 0.01	7.67 ± 0.29
Strongly invariant	0.5/1.0	4.22 ± 0.03	-1.17 ± 1.21	3.49 ± 0.04	-4.74 ± 1.34	2.93 ± 0.02	-10.09 ± 0.96	3.60 ± 0.03	-3.92 ± 1.27	3.05 ± 0.02	-1.12 ± 0.38
	1.0/0.5	4.27 ± 0.03	-0.07 ± 1.00	3.48 ± 0.03	-5.20 ± 1.31	2.95 ± 0.01	-9.19 ± 0.76	3.57 ± 0.03	-4.49 ± 0.75	3.06 ± 0.02	-0.58 ± 0.96
	1.5/1.0	4.44 ± 0.02	3.98 ± 0.91	3.56 ± 0.03	-2.39 ± 0.60	3.14 ± 0.01	-3.31 ± 0.81	3.69 ± 0.02	-0.83 ± 1.34	3.05 ± 0.00	-0.90 ± 0.71
	1.0/1.5	4.35 ± 0.02	1.71 ± 0.61	3.63 ± 0.03	0.85 ± 1.15	3.10 ± 0.01	-4.35 ± 1.16	3.71 ± 0.01	-0.63 ± 0.88	3.08 ± 0.02	0.22 ± 0.86
	2.0/1.0	4.31 ± 0.02	0.68 ± 0.70	3.47 ± 0.03	-5.00 ± 0.97	3.02 ± 0.02	-6.89 ± 1.07	3.57 ± 0.01	-4.24 ± 0.95	2.97 ± 0.02	-3.72 ± 0.68
	1.0/2.0	4.29 ± 0.01	0.73 ± 0.73	3.50 ± 0.02	-4.17 ± 1.18	3.02 ± 0.02	-6.78 ± 0.77	3.55 ± 0.02	-4.55 ± 1.21	2.98 ± 0.01	-3.25 ± 0.68
	1.0/1.0	4.34 ± 0.02	1.56 ± 0.68	3.61 ± 0.01	-1.30 ± 0.94	3.08 ± 0.01	-4.82 ± 1.14	3.73 ± 0.01	0.06 ± 0.88	3.08 ± 0.02	0.39 ± 0.56

Table 4: LLM inference with diverse drafts; we use  $L = 5$ ,  $K = 2$  and the target temperature is 2.0. The temperatures of drafters 1 and 2 vary and are reported in the second column. Token rates (TR) are shown as percentage speedups relative to single-draft speculative decoding with drafter temperature 1.0, which in this setting has mean block efficiency (BE) 4.28, 3.65, 3.19, 3.71 and 3.06 on GSM8K, HumanEval, NaturalReasoning, MBPP and DROP respectively.



1087 and  $\Sigma_{W,T}$  was found earlier in (19). Then, the MMSE estimate is

$$\hat{A} = \mathbb{E}[A \mid W, T] = \Sigma_{A,(W,T)} \Sigma_{W,T}^{-1} \begin{bmatrix} W \\ T \end{bmatrix} = \frac{\sigma_{\zeta}^2 W + \sigma_{\eta}^2 T}{\sigma_{\eta}^2 + \sigma_{\zeta}^2 + \sigma_{\eta}^2 \sigma_{\zeta}^2}.$$

1088 To conclude, given  $T_k = t_k$  and  $W_k = w_k$ , the reconstruction output by decoder  $k$  is given by

$$g(w_k, t_k) = \frac{\sigma_{\zeta}^2 w_k + \sigma_{\eta}^2 t_k}{\sigma_{\eta}^2 + \sigma_{\zeta}^2 + \sigma_{\eta}^2 \sigma_{\zeta}^2}.$$

1089 **Experiment parameters and further results.** First, we briefly outline the parameters used for  
 1090 the experiment. The number of samples from the prior is  $N = 2^{15}$  for all tests and the source,  
 1091 as mentioned in the main paper, is  $A \sim \mathcal{N}(0, 1)$ . The conditional variance of the side in-  
 1092 formation given  $A$  is fixed at  $\sigma_{T|A}^2 = 0.5$  throughout. We control the rate by varying  $L_{\max}$ ,  
 1093 considering  $L_{\max} \in \{2^1, 2^2, 2^3, 2^4, 2^5, 2^6\}$ . For each  $L_{\max}$ , the resulting distortion is mini-  
 1094 mized over the encoder’s target distribution by exploring different values of  $\sigma_{W|A}^2$ , selecting from  
 1095  $\sigma_{W|A}^2 \in \{0.01, 0.008, 0.006, 0.005, 0.003, 0.002, 0.001\}$  and choosing the best across  $10^4$  trials. The  
 1096 distortion incurred by the best configuration is then further evaluated on  $10^5$  trials. This procedure is  
 1097 carried out for  $K \in \{1, 2, 3, 4\}$ , where  $K$  is the number of decoders. Finally, the entire experiment is  
 1098 repeated 10 times and the results are averaged to obtain those reported in table 5. The error bars show  
 1099 one standard error of the mean, calculated as in appendix D.1 using all 10 trials.

1100 Running one full repetition of the experiment takes around 4 hours when performing the calculations  
 1101 on a Nvidia Tesla T4 GPU with 16 GB of memory. The exact same procedure is used to generate  
 1102 results for the baseline scheme described in the main paper, and these are shown in table 6. We also  
 1103 show the value of  $\sigma_{W|A}^2$  that most often minimizes the distortion in each case.

### 1104 D.3 Distributed image compression

1105 We now give details on our distributed image compression experiments.

1106 **Neural network architectures.** The following notations are used to denote the different layers in  
 1107 our networks:

- 1108 1.  $\text{conv}(a, b, c, d, e)$ : A convolution layer with  $a$  input features,  $b$  output features, kernel size  $c$ ,  
 1109 stride  $d$  and input padding  $e$ .
- 1110 2.  $\text{upconv}(a, b, c, d, e, f)$ : A transposed convolution layer with  $a$  input features,  $b$  output  
 1111 features, kernel size  $c$ , stride  $d$ , input padding  $e$  and output padding  $f$ .
- 1112 3.  $\text{fc}(a, b)$ : A fully-connected layer with input size  $a$  and output size  $b$ .
- 1113 4.  $\text{do}(p)$ : A dropout layer with dropout probability  $p$ .
- 1114 5.  $\text{cat}(a, b)$ : Concatenates two tensors of shapes  $a$  and  $b$ .

$K$	$L_{\max}$	$\sigma_{W A}^2$	Distortion (dB)	$K$	$L_{\max}$	$\sigma_{W A}^2$	Distortion (dB)
1	$2^1$	0.008	$-9.7032 \pm 0.0193$	3	$2^1$	0.005	$-18.3884 \pm 0.0163$
	$2^2$	0.010	$-12.7474 \pm 0.0226$		$2^2$	0.003	$-21.6187 \pm 0.0164$
	$2^3$	0.010	$-16.0116 \pm 0.0369$		$2^3$	0.001	$-25.0515 \pm 0.0213$
	$2^4$	0.003	$-19.5491 \pm 0.0237$		$2^4$	0.001	$-28.5329 \pm 0.0128$
	$2^5$	0.002	$-23.4012 \pm 0.0132$		$2^5$	0.001	$-31.2575 \pm 0.0161$
	$2^6$	0.001	$-27.3470 \pm 0.0183$		$2^6$	0.001	$-33.1515 \pm 0.0108$
2	$2^1$	0.010	$-15.2069 \pm 0.0148$	4	$2^1$	0.005	$-20.6834 \pm 0.0176$
	$2^2$	0.005	$-18.3377 \pm 0.0164$		$2^2$	0.001	$-23.9418 \pm 0.0197$
	$2^3$	0.002	$-21.7032 \pm 0.0101$		$2^3$	0.001	$-27.4313 \pm 0.0106$
	$2^4$	0.001	$-25.3886 \pm 0.0104$		$2^4$	0.001	$-30.4379 \pm 0.0188$
	$2^5$	0.001	$-28.8619 \pm 0.0169$		$2^5$	0.001	$-32.6616 \pm 0.0106$
	$2^6$	0.001	$-31.4737 \pm 0.0152$		$2^6$	0.001	$-34.1082 \pm 0.0102$

Table 5: Results using GLS with a Gaussian source.

$K$	$L_{\max}$	$\sigma_{W A}^2$	Distortion (dB)	$K$	$L_{\max}$	$\sigma_{W A}^2$	Distortion (dB)
1	$2^1$	0.010	$-9.7163 \pm 0.0195$	3	$2^1$	0.010	$-13.8197 \pm 0.0368$
	$2^2$	0.008	$-12.6968 \pm 0.0273$		$2^2$	0.010	$-17.4640 \pm 0.0211$
	$2^3$	0.008	$-16.0124 \pm 0.0153$		$2^3$	0.010	$-21.0096 \pm 0.0185$
	$2^4$	0.003	$-19.5518 \pm 0.0270$		$2^4$	0.003	$-24.6996 \pm 0.0126$
	$2^5$	0.001	$-23.3905 \pm 0.0105$		$2^5$	0.001	$-28.6864 \pm 0.0123$
	$2^6$	0.001	$-27.3705 \pm 0.0135$		$2^6$	0.001	$-32.0109 \pm 0.0156$
2	$2^1$	0.010	$-12.5143 \pm 0.0356$	4	$2^1$	0.010	$-14.6125 \pm 0.0272$
	$2^2$	0.010	$-15.9916 \pm 0.0205$		$2^2$	0.010	$-18.3350 \pm 0.0157$
	$2^3$	0.008	$-19.4843 \pm 0.0137$		$2^3$	0.008	$-21.9300 \pm 0.0238$
	$2^4$	0.003	$-23.1495 \pm 0.0240$		$2^4$	0.002	$-25.5269 \pm 0.0210$
	$2^5$	0.001	$-27.1988 \pm 0.0092$		$2^5$	0.001	$-29.4994 \pm 0.0100$
	$2^6$	0.001	$-30.7772 \pm 0.0169$		$2^6$	0.001	$-32.7141 \pm 0.0208$

Table 6: Results using the baseline decoding scheme with a Gaussian source.

0	Input ( $1 \times 28 \times 28$ )	0	Input (132)
1	conv(1, 128, 3, 1, 1), ReLU	1	fc(132, 512), ReLU
2	conv(128, 128, 3, 2, 1), ReLU	2	fc(132, 6272), ReLU
3	conv(128, 128, 3, 2, 1), ReLU	3	upconv(128, 64, 3, 2, 1, 1), ReLU
4	fc(6272, 512), ReLU	4	upconv(64, 32, 3, 2, 1, 1), do(0.5), ReLU
5	fc(512, 8)	5	upconv(32, 1, 3, 1, 1, 0), tanh
(a) Encoder		(b) Decoder	
0	Input ( $1 \times 14 \times 14$ )	0	Input ( $1 \times 14 \times 14$ )
1	conv(1, 32, 3, 1, 1), ReLU	1	conv(1, 32, 3, 1, 1), ReLU
2	conv(32, 64, 3, 2, 1), ReLU	2	conv(32, 64, 3, 2, 1), ReLU
3	conv(64, 128, 3, 2, 1), ReLU	3	conv(64, 128, 3, 2, 1), ReLU
4	fc(2048, 512), ReLU	4	fc(2048, 512), ReLU
5	fc(512, 128)	5	fc(512, 128), cat(128, 4)
(c) Projection		6	fc(132, 128), LeakyReLU
		7	fc(128, 128), LeakyReLU
		8	fc(128, 128), LeakyReLU
		9	fc(128, 128), LeakyReLU
		10	fc(128, 1), Sigmoid
		(d) Estimator	

Table 7: Neural network architectures.

1115 The network layers are enumerated in table 7 and follow the network constructions in Phan et al. [31].  
 1116 The encoder’s target distribution  $p_{W|A}$  is taken to be a four-dimensional Gaussian with uncorrelated  
 1117 components, where the mean and variance of each component are generated by the encoder network  
 1118 from an input image. More precisely, if we let the image be  $a$ , the encoder network produces two  
 1119 embeddings  $e_1(a)$  and  $e_2(a)$ , each in  $\mathbb{R}^4$ . Then,  $p_{W|A}(\cdot | a) = \mathcal{N}(e_1(a), \text{diag}(e_2(a)))$ , and we  
 1120 arbitrarily choose  $W \sim \mathcal{N}(0, 1)$  as the marginal distribution, which is also the  $\beta$ -VAE’s prior. On the  
 1121 other hand, decoder  $k$  is tasked with generating a reconstruction given the side information  $t_k$  and  
 1122 an embedding  $w_k \in \mathbb{R}^4$ , which is selected depending on the message sent by the encoder. Rather  
 1123 than using the  $14 \times 14$  side information image directly, we employ a projection network to extract a  
 1124 length-128 feature vector  $e(t_k)$  before feeding this representation into the decoder network along  
 1125 with  $w_k$  to get  $\hat{a}^{(k)} = g(w_k, e(t_k))$  for  $1 \leq k \leq K$ . The final estimate  $\hat{a}$  is chosen from among the  
 1126  $\hat{a}^{(k)}$ ’s such that the distortion is minimized.

1127 The estimator network is another important component of our compression protocol, since it acts as  
 1128 a proxy for  $p_{W|T}$ . Recall from section 5.1 and its extension in appendix C that this distribution is  
 1129 used to select the index at the decoder; using this index, decoder  $k$  picks  $W_k$  from the shared list of  
 1130 samples taken from the prior. In practice, the estimator network takes a  $14 \times 14$  side information  
 1131 image as its input and extracts 128-dimensional features, which are then concatenated with a sample  
 1132  $w \in \mathbb{R}^4$ . The final part of the network is classifies whether this joint embedding comes from the joint  
 1133 distribution  $p_{W,T}$  or the product of the marginals  $p_W p_T$ . Its output therefore stands in for  $p_{W|T}$ .

1134 **Network loss functions.** The  $\beta$ -VAE is trained using the rate-distortion loss function

$$\mathcal{L}_{\text{VAE}}(a, \hat{a}) = \beta(a - \hat{a}) - D_{\text{KL}}[p_{W|A}(\cdot | a) \| p_W].$$

1135 Note that since the marginal and conditional distributions  $p_W$  and  $p_{W|A}$  are both Gaussian,  $D_{\text{KL}}$  has  
 1136 a closed form. The neural estimator uses the binary cross-entropy (BCE) loss function. If we let its  
 1137 output as a function of side information  $t$  and given sample  $w$  be  $h(w, t)$ , the loss is

$$\mathcal{L}_{\text{estimator}}(w, t) = \text{BCE}(h(w, t), \mathbb{1}\{w \text{ was sampled from } p_{W|T}(\cdot | t)\})$$

1138 where  $\mathbb{1}$  is the indicator function.

1139 **Training and evaluation procedure.** We use the MNIST dataset [20] with the usual train-test  
 1140 split of 60 000 and 10 000 images respectively and batch size 64. All models are trained for 30  
 1141 epochs on a Nvidia Tesla T4 GPU with 16 GB of memory using the Adam optimizer [19]. The  
 1142 learning rate is  $10^{-3}$  and we set  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ . The encoder, decoder, projection and  
 1143 estimator networks are trained jointly in an end-to-end manner, and we create sets of models for  
 1144  $\beta \in \{0.15, 0.35, 0.55, 0.75, 0.95\}$  to cover a broad range of rate-distortion tradeoffs at the encoder  
 1145 side. Jointly training the networks takes around 45 minutes for each  $\beta$ .

1146 At test time, we vary  $L_{\text{max}}$  to control the rate, considering  $L_{\text{max}} \in \{2^2, 2^3, 2^4, 2^5, 2^6\}$ . For  
 1147 each configuration, we additionally optimize over  $N$ , which is the number of samples from the  
 1148 prior, and the VAE parameter  $\beta$  using a grid search where  $N \in \{2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$  and  
 1149  $\beta \in \{0.15, 0.35, 0.55, 0.75, 0.95\}$ . The experiment is repeated 5 times and the results averaged,  
 1150 with the same procedure also being followed for the baseline scheme described in the main paper.  
 1151 We provide error bars showing one standard error of the mean, which is again calculated as in  
 1152 appendix D.1 with the number of trials now being 5. Each instance of the full experiment takes  
 1153 approximately 6 hours to run, and this is done for  $K \in \{1, 2, 3, 4\}$ . Complete results are given in  
 1154 tables 8 and 9, where we also give the values of  $N$  and  $\beta$  that are most often optimal in each case.

$K$	$L_{\text{max}}$	$N$	$\beta$	MSE	$K$	$L_{\text{max}}$	$N$	$\beta$	MSE
1	$2^2$	$2^7$	0.15	$0.1027 \pm 0.0002$	3	$2^2$	$2^9$	0.15	$0.0791 \pm 0.0000$
	$2^3$	$2^7$	0.15	$0.0942 \pm 0.0001$		$2^3$	$2^7$	0.15	$0.0738 \pm 0.0002$
	$2^4$	$2^7$	0.15	$0.0852 \pm 0.0002$		$2^4$	$2^7$	0.15	$0.0694 \pm 0.0001$
	$2^5$	$2^7$	0.15	$0.0766 \pm 0.0001$		$2^5$	$2^7$	0.15	$0.0660 \pm 0.0001$
	$2^6$	$2^7$	0.15	$0.0693 \pm 0.0002$		$2^6$	$2^8$	0.35	$0.0599 \pm 0.0002$
2	$2^2$	$2^9$	0.15	$0.0860 \pm 0.0001$	4	$2^2$	$2^7$	0.15	$0.0751 \pm 0.0001$
	$2^3$	$2^7$	0.15	$0.0792 \pm 0.0001$		$2^3$	$2^7$	0.15	$0.0710 \pm 0.0001$
	$2^4$	$2^7$	0.15	$0.0734 \pm 0.0001$		$2^4$	$2^7$	0.15	$0.0671 \pm 0.0001$
	$2^5$	$2^7$	0.15	$0.0687 \pm 0.0002$		$2^5$	$2^8$	0.35	$0.0635 \pm 0.0001$
	$2^6$	$2^7$	0.35	$0.0636 \pm 0.0002$		$2^6$	$2^8$	0.35	$0.0564 \pm 0.0001$

Table 8: Results using GLS for distributed image compression on MNIST.

$K$	$L_{\text{max}}$	$N$	$\beta$	MSE	$K$	$L_{\text{max}}$	$N$	$\beta$	MSE
1	$2^2$	$2^7$	0.15	$0.1025 \pm 0.0001$	3	$2^2$	$2^7$	0.15	$0.0906 \pm 0.0002$
	$2^3$	$2^7$	0.15	$0.0937 \pm 0.0002$		$2^3$	$2^7$	0.15	$0.0832 \pm 0.0003$
	$2^4$	$2^7$	0.15	$0.0850 \pm 0.0002$		$2^4$	$2^7$	0.15	$0.0757 \pm 0.0002$
	$2^5$	$2^7$	0.15	$0.0764 \pm 0.0002$		$2^5$	$2^7$	0.15	$0.0704 \pm 0.0001$
	$2^6$	$2^7$	0.15	$0.0693 \pm 0.0002$		$2^6$	$2^7$	0.35	$0.0653 \pm 0.0001$
2	$2^2$	$2^7$	0.15	$0.0941 \pm 0.0002$	4	$2^2$	$2^9$	0.15	$0.0886 \pm 0.0001$
	$2^3$	$2^7$	0.15	$0.0865 \pm 0.0002$		$2^3$	$2^7$	0.15	$0.0815 \pm 0.0001$
	$2^4$	$2^7$	0.15	$0.0783 \pm 0.0002$		$2^4$	$2^7$	0.15	$0.0747 \pm 0.0002$
	$2^5$	$2^7$	0.15	$0.0718 \pm 0.0002$		$2^5$	$2^7$	0.15	$0.0694 \pm 0.0002$
	$2^6$	$2^7$	0.15	$0.0669 \pm 0.0001$		$2^6$	$2^7$	0.35	$0.0639 \pm 0.0002$

Table 9: Results using the baseline decoding scheme for distributed image compression on MNIST.