

---

# Is the acquisition worth the cost? Surrogate losses for Consistent Two-stage Classifiers

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Recent years have witnessed the emergence of a spectrum of foundation models, covering a broad range of capabilities and costs. Often, we effectively use foundation models as feature generators and train classifiers that use the outputs of these models to make decisions. In this paper, we consider an increasingly relevant setting where we have two classifier stages. The first stage has access to features  $x$  and has the option to make a classification decision or defer, while incurring a cost, to a second classifier that has access to features  $x$  and  $z$ . This is similar to the “learning to defer” setting, with the important difference that we train both classifiers jointly, and the second classifier has access to more information. The natural loss for this setting is an  $\ell_{01c}$  loss, where a penalty is paid for incorrect classification, as in  $\ell_{01}$ , but an additional penalty  $c$  is paid for consulting the second classifier. The  $\ell_{01c}$  loss is unwieldy for training. Our primary contribution in this paper is the derivation of a hinge-based surrogate loss  $\ell_{hinge}^c$  that is much more amenable to training but also satisfies the property that  $\ell_{hinge}^c$ -consistency implies  $\ell_{01c}$ -consistency.

## 1 Introduction

With the emergence of a spectrum of foundation models, covering a broad range of capabilities and costs, we are increasingly faced with a decision as to which model to use. For example, can we make a decision locally, on an edge device, or should we incur the additional communication and computational cost of sending the query to a more powerful remote model? In many cases, we use the pre-trained foundation model essentially as a feature generator, and strive to train a classifier that uses the output of the foundation model as its input. In this setting, we then face a task of training two classifiers, while simultaneously learning when to defer to the more powerful model.

One approach to solve this problem is to 1) train the more powerful classifier first; and 2) train the decision module with the smaller classifier afterwards (either jointly or separately). This strategy has proven successful and benefits from strong theoretical foundations [Keswani et al., 2021, Wilder et al., 2021, Verma et al., 2022, Mao et al., 2023, 2024b], but intuitively, this appears inefficient. Indeed, with this approach, both classifiers expend effort exploring regions of the input space where their predictions will ultimately not be used. Because of this, it is important to consider and formalize the problem where the classifiers and the module deciding which model to use are trained jointly.

While there has been a significant body of work establishing consistent losses for the related problem of classification with learning to defer to experts or with reject [Verma et al., 2022, Herbei and Wegkamp, 2006], these losses do not cover the case where we jointly train multiple inference classifiers along with the decision module. In these well-studied settings, the task is to learn one classifier and to either defer to an oracle (learning to reject) [Chow, 1970], to an expert (learning to defer) [Madras et al., 2018] or to multiple experts (learning to defer to multiple experts) [Verma et al.,

2022]. However, these frameworks assume that the experts are external to the problem setting. They do not address how to train the experts alongside the base classifier.

In this work, we address the problem of jointly training two classifiers with a decision module. The two classifiers incur different costs, with the implication that the more expensive classifier offers better performance. We model this problem by introducing an additional information variable,  $Z$ , which represents the extra information available to the more powerful classifier. The decision module and the base classifier both have access to the same base input variable  $X$ . We refer to this setup as the *two-stage classification* problem.

We provide the optimal solution to this problem, as well as a surrogate loss function that is more suitable for training. The surrogate loss, which is based on the hinge loss, aligns with the standard cost-aware 0-1 loss formulation commonly used in classification tasks. We validate our theoretical findings on synthetic datasets and demonstrate the practical relevance of the problem by presenting results on a standard large language model (LLM) task, where two LLMs of varying sizes are used to answer multi-question math problems. Additionally, we provide a proof that the cross-entropy loss, which is sometimes used heuristically in existing literature, is not Bayes consistent with the natural 0-1 loss, further justifying the need to explore this problem at a theoretical level.

Our main contributions are as follows:

1. We formulate a problem setting for learning a model that integrates two classifiers, where one has access to additional information but comes with a cost  $c$ . The goal is to train the models and simultaneously learn the decision function to determine whether to consult the more powerful classifier for a given sample.
2. We present a surrogate loss function based on the hinge loss, which is suitable for training with cost-aware classification tasks. We show that it is consistent with respect to the 0-1 loss that is natural for the considered problem.
3. We validate the theoretical findings, which are the primary contribution of the work, on synthetic datasets and provide practical insights through experiments on a standard LLM task.

## 2 Related Work

Loss consistency is an important topic that has been widely explored, as it serves as the fundamental link between the loss we optimize in practice and the actual loss we aim to minimize. Foundational results have been established for classical risks [Steinwart, 2007, Tewari and Bartlett, 2007, Bartlett et al., 2006], and the emergence of new target losses has prompted the development of new consistency results. Learning to defer (L2D) is a wide category of settings in which the task is to learn a classifier and a deferral rule, either to reject (learning to abstain) [Chow, 1957, 1970, Herbei and Wegkamp, 2006, Cao et al., 2022, Wiener and El-Yaniv, 2011, Geifman and El-Yaniv, 2019] or to defer to one or more experts of varying costs [Madras et al., 2018, Keswani et al., 2021, Wilder et al., 2021].

Mozannar and Sontag [2020] were the first to provide Bayes-consistency results for their proposed generalized cross entropy loss for learning to defer, followed by Verma et al. [2022], who used a one-vs-all loss. Awasthi et al. [2022] explored stronger guarantees than Bayes-consistency by introducing  $\mathcal{H}$ -consistency bounds [Long and Servedio, 2013]. Mozannar et al. [2023] prove that earlier approaches, such as [Mozannar and Sontag, 2020, Verma et al., 2022], fall short of realizable  $\mathcal{H}$ -consistency, and propose a new algorithm without a Bayes-consistency proof. This shortcoming is addressed by Mao et al. [2024a], who recently published a unifying work, introducing a new family of surrogate losses for the learning to defer problem with a single expert, and providing Bayes-consistency, realizable  $\mathcal{H}$ -consistency, and  $\mathcal{H}$ -consistency bounds. Verma et al. [2022] extended the work of [Mozannar and Sontag, 2020, Verma et al., 2022] to the multi-expert setting with Bayes-consistency guarantees. Mao et al. [2024b] introduced general cost functions and surrogate losses, extending [Mozannar and Sontag, 2020] with  $\mathcal{H}$ -consistency bounds for joint training, and offering stronger guarantees than Bayes-consistency. Mao et al. [2023] also provided  $\mathcal{H}$ -consistency bounds for a slightly different setting where training the classifier and the deferral rule is done separately.

It appears that consistency results have been thoroughly studied in the case where only a single classifier is trained. However, the setting where multiple classifiers are trained jointly with a decision module has been largely neglected. This type of architecture is used in adaptive computation or dynamic networks [Han et al., 2022a], a branch of research focused on developing architectures that

adaptively allocate computation. Since the main objective is to improve average inference efficiency, such dynamic network architectures have attracted significant interest in the development of scalable LLM inference [Liu et al., 2023, Elbayad et al., 2020, Zeng et al., 2024, Xia et al., 2024, Leviathan et al., 2023, Chen et al., 2024].

Although there is growing interest in these types of networks, the losses used to train such models are mostly heuristic and lack strong theoretical foundations. In practice, these models typically train the classifiers separately and rely on threshold-based decisions [Han et al., 2022a, Schuster et al., 2022]. Some theoretical research has been conducted on this separated training approach: Jitkrittum et al. [2023] explored the connection between threshold decisions and risk under a principled 0-1 loss, identifying conditions under which the two coincide. However, this separate training is not guaranteed to be the best approach. In fact, the importance of jointly learning the classifiers and the decision module has been empirically demonstrated [Han et al., 2022b, Yu et al., 2022, Regol et al., 2024, Krzepkowski et al., 2024], motivating the development of joint learning approaches [Regol et al., 2024] and classifier-deferral-aware training methods [Han et al., 2022b, Yu et al., 2022]. These works lack a connection to surrogate losses and a well-defined risk framework, which is the gap we aim to address in this work.

### 3 The two-stage classification problem

We consider a setting of two-stage classification where there are two classifiers:  $f_1$  and  $f_2$ . The second classifier,  $f_2$ , has access to additional information  $z$ , but it also incurs an additional cost, denoted as  $c$ . We therefore have the choice between using the prediction of the first classifier, or to pay the additional cost and then use the more informed second classifier.

In practice,  $z$  can be explicitly modeled as an additional input signal or feature, which may come with higher access costs. For instance, in recommendation systems, different types of user data queries can vary significantly in terms of latency and infrastructure expense. A common approach is to first run a lightweight model for initial inference, and then selectively identify instances that would benefit from a more complex model with access to richer features. This tiered architecture is notably used by Youtube’s recommendation system [Covington et al., 2016], for instance.

Alternatively,  $z$  can conceptually represent the augmented modeling capability of a larger model that has more parameters and/or was trained on a larger data set.

Denote by  $\mathcal{X}$  the feature space,  $\mathcal{Z}$  the additional information space, and  $\mathcal{Y} = \{1, \dots, K\}$  the label space. We are given instance-label-information triples  $\{(x_i, z_i, y_i)\}_{i=1}^n$  independently and identically drawn from an underlying distribution  $\mathcal{D}$  with probability function  $p(X, Z, Y)$ . We additionally introduce the decision module  $r : \mathcal{X} \rightarrow 0, 1$ , which indicates whether we are using, for the final decision, the first classifier  $f_1$  if  $r(x) = 0$  or to defer to the second classifier  $f_2$  if  $r(x) = 1$ .

The goal of two-stage classification is to train a two-stage classifier  $f : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$  that encompasses both the classifiers  $f_1 : \mathcal{X} \rightarrow \mathcal{Y}$ ,  $f_2 : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$ , and the decision module  $r$ . The set  $\mathcal{H}$  of two-stage classifiers is therefore defined as follows;

$$\mathcal{H} = \{f : f(x, z) = \begin{cases} f_1(x), & r(x) = 0 \\ f_2(x, z), & r(x) = 1. \end{cases} \quad (1)$$

The loss associated with such a setting is the zero-one-exit loss  $\ell_{01c}$ , which can be expressed as a variant of the traditional zero-one loss  $\ell_{01}(f(\cdot), y) = \mathbb{1}[f(\cdot) \neq y]$ :

$$\ell_{01c}(f(x, z), y) = \begin{cases} \mathbb{1}[f_1(x) \neq y], & r(x) = 0, \\ \mathbb{1}[f_2(x, z) \neq y] + c, & r(x) = 1, \end{cases} \quad (2)$$

where  $\mathbb{1}[\cdot]$  is the indicator function. The cost  $c$  can be an instance-specific function, i.e.,  $c(x)$ , provided it is known and deterministic. Since the additional information  $z$  is only accessible at a cost  $c$ , the first classifier and the decision function do not have access to it; the classifiers  $f_1(x)$  and  $r(x)$  take only  $x$  as input.

Our task is to train a two-stage classifier  $f \in \mathcal{H}$ , as defined by (1), that can minimize the expectation of  $\ell_{01c}$  over the data distribution. The risk is:

$$R_{01c}(f) = \mathbb{E}_{p(x, z, y)}[\ell_{01c}(f(x, z), y)], \quad (3)$$

135 and its optimal value  $R_{01c}^* = R_{01c}(f^*)$  is obtained by the Bayes-optimal classifier:

$$f^* = \arg \min_{f \in \mathcal{H}} R_{01c}(f). \quad (4)$$

136 The 01c loss is discrete, and thus difficult to work with. We would like to be able to identify a  
 137 surrogate loss  $\ell_\phi$  such that  $\ell_\phi$ -consistency implies  $\ell_{01c}$ -consistency. This is our main contribution in  
 138 this work. We specify a surrogate loss function that satisfy this property, and show that other heuristic  
 139 surrogate losses that are used in the literature for joint training [Regol et al., 2024, Ding et al., 2024]  
 140 do not. Taking a step beyond this, we specify how to construct and train a two-stage classifier using  
 141 the posited surrogate loss and present empirical results to validate our result.

### 142 3.1 The solution

143 We start by providing the solution to the optimization problem specified by (4). We first define a  
 144 compact notation for the posteriors:

$$\eta_y(x) \triangleq p(Y = y|x), \quad \zeta_y(x, z) \triangleq p(Y = y|x, z). \quad (5)$$

145 **Lemma 3.1.** *The optimal solution  $f^* = \arg \min_{f \in \mathcal{H}} R_{01c}(f)$  is the following:*

$$f^* = \begin{cases} \arg \max_y \eta_y(x), & \text{if } \max_y \eta_y(x) \geq \mathbb{E}_{p(Z|x)}[\max_y \zeta_y(x, Z)] - c, \\ \arg \max_y \zeta_y(x, z), & \text{else.} \end{cases} \quad (6)$$

146 See Appendix A.3 for the proof of the lemma.

147 Using our previous definition of a two-stage classifier, this would correspond to:

$$f_1^*(x) = \arg \max_y \eta_y(x) \text{ if } \{x; r^*(x) = 0\} \quad (7)$$

$$f_2^*(x, z) = \arg \max_y \zeta_y(x, z) \text{ if } \{x; r^*(x) = 1\} \quad (8)$$

$$r^*(x) = \begin{cases} 0 & \text{if } \max_y \eta_y(x) \geq \mathbb{E}_{p(Z|x)}[\max_y \zeta_y(x, Z)] - c, \\ 1 & \text{o.w.} \end{cases} \quad (9)$$

148 The optimal solution is interesting. It hints towards a model that is slightly different from most  
 149 existing methods. Yes, the optimal decision should depend on  $p_{\max} = \max \eta(x)$ , but the threshold  
 150 for  $p_{\max}$  should be set based on the *expected future gain*:  $\tau < \mathbb{E}_{p(Z|x)}[\max \zeta(x, Z)] - c$ .  $r(x)$  should  
 151 identify the set on which  $\max \eta(x) \geq \mathbb{E}_{p(Z|x)} \max \zeta(x, z) - c$ , and, for these elements only, it should  
 152 select the class with the largest probability according to the posterior  $\eta(x)$ . This result is similar to  
 153 the solution for the decision rule given fixed classifiers first provided by Jitkrittum et al. [2023], which  
 154 would read as  $\max_y \eta_y(x) \geq \max_y \zeta_y(x, Z) - c$ . However, our explicit modeling of the two-tiered  
 155 information available to  $f_1, r$  and  $f_2$  provides a more practical and detailed solution, as it allows us  
 156 to integrate the constraint that  $r$  cannot fully access the information available to  $f_2$ . This modeling  
 157 choice leads to a decision based on the **expected** future gain.

158 Unsurprisingly, the first and second classifiers simply predict the class with the highest probability  
 159 according to their respective posteriors, but only for the samples assigned to them.

## 160 4 The proposed hinge-based surrogate loss

161 A common strategy to develop a consistent loss for more complex risk functions is to propose a  
 162 surrogate loss and verify its consistency. This strategy was employed by early work for the learning  
 163 to defer problem [Mozannar and Sontag, 2020, Verma and Nalisnick, 2022]).

164 Our proposed surrogate loss is built on a multiclass version of the hinge loss [Tarigan and van de Geer,  
 165 2008]. We chose this version because it is Bayes-consistent, unlike other multiclass hinge losses. We  
 166 use a hinge loss rather than the more popular cross entropy is because of its linear scaling, which  
 167 allows to account for the cost in an additive way as in Eqn. 2. Following the definition of the multiclass  
 168 hinge loss from [Tarigan and van de Geer, 2008], the classifiers are based on  $K$ -dimensional real  
 169 valued outputs  $\mathbf{t}(x), \mathbf{v}(x, z) \in \mathbb{R}^K$  with the constraints that  $\sum_{i=1}^K \mathbf{t}_i(x) = 0, \sum_{i=1}^K \mathbf{v}_i(x, z) = 0$ ,

and the label prediction is obtained by returning the max element of that vector. The decision function  $\tilde{r}(x)$  returns a real value bounded between 0 and 1. For brevity, we omit the dependence on the inputs and only write  $\mathbf{t}, \mathbf{v}$ . We can therefore introduce the link function  $\varphi$  that connects the real valued output and a soft decision function  $\tilde{r}(x)$  to a two-stage classifier function :

$$f = \{f_1, f_2, r\} = \varphi(\mathbf{t}, \mathbf{v}, \tilde{r}) = \{\max_{y \in \mathcal{Y}} \mathbf{t}_y(x), \max_{y \in \mathcal{Y}} \mathbf{v}_y(x, z), \mathbb{1}[\tilde{r}(x) \geq 0.5]\}. \quad (10)$$

Letting  $[x]_+ = \max(x, 0)$ , our proposed hinge-based surrogate loss is given by:

$$\ell_{hinge}^c(\mathbf{t}, \mathbf{v}, \tilde{r}, x, z, y) = (1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ + \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right). \quad (11)$$

We can then define the associated risk as:

$$R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) = \mathbb{E}_{p(x, z, y)} [\ell_{hinge}^c(\mathbf{t}, \mathbf{v}, \tilde{r}, x, z, y)], \quad (12)$$

and consider the triplet of minimizers  $\mathbf{t}^*(x), \mathbf{v}^*(x, z), \tilde{r}^*(x)$  of such a risk:

$$\mathbf{t}^*, \mathbf{v}^*, r^* = \arg \min_{\mathbf{t}, \mathbf{v} \in \mathbb{R}^K, r \in [0, 1]} R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}), \quad (13)$$

$$f_{hinge}^* = \varphi(\mathbf{t}^*, \mathbf{v}^*, r^*). \quad (14)$$

In the following theorem, we establish the consistency of our proposed surrogate loss w.r.t.  $\ell_{01c}$ , meaning that if a learned two-stage classifier  $f$  converges to the optimal surrogate risk  $R_{hinge}^*$ , it also converges to the optimal target risk  $R_{01c}^*$ .

**Theorem 4.1.** *There exists a link function  $\varphi$  s.t. for any distribution  $p(x, z, y)$ , we have that:*

$$R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) \rightarrow R_{hinge}^* \implies R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) \rightarrow R_{01c}^*, \quad (15)$$

i.e., the surrogate loss  $\ell_{hinge}^c(\mathbf{v}, \mathbf{t}, r, x, z, y)$  is consistent with respect to the loss of interest  $\ell_{01c}(\varphi(\mathbf{t}, \mathbf{v}, r), x, z, y)$ .

The proof is provided in Appendix A.4. The proof is built by showing that 1) the minimizers of both risks are unique and coincide:

$$f_{hinge}^* = f^*, \quad (16)$$

(Lemma A.5, with proof provided in Appendix A.4.1); and 2) that for some increasing function  $\Psi$  with  $\Psi(0) = 0$ , the following holds:

$$R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^* \leq \Psi(R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^*), \quad (17)$$

(Lemma A.2, with proof included in Appendix A.4.2). Taken together, these results guarantee consistency. We can actually establish that

$$R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^* \leq \frac{2(K-1)}{K} (R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^*). \quad (18)$$

The bound on the risk gap provided by (18) allows us to further quantify the relationship between the two optimization problems, showing that the consistency is not merely asymptotic. This upper bound is tight and attainable for some cases of  $\eta$  and  $\zeta$ . The  $\frac{K-1}{K}$  term comes from the scaling of the multi-hinge loss, while the factor of 2 accounts for corner cases where the routing decision is uncertain ( $\tilde{r}(x) = 0.5$ ) and the model perfectly estimates the posteriors  $\eta$  and  $\zeta$ .

#### 4.1 Cross entropy version

One might be tempted to build a similar formulation using the widely used cross-entropy loss  $-\log(\mathbf{p}_y)$ . Some heuristics in the literature for training two-stage or early exit models are built around a similar version of this loss [Regol et al., 2024]. Interestingly, we can prove that such a loss is in fact not Bayes consistent with the 0-1closs that we presented. To build a cross entropy version of the proposed loss, we now need to assume that the model outputs predicted class probabilities

$\mathbf{p}^1 \in \Delta^K$  for  $f_1$  and  $\mathbf{p}^2 \in \Delta^K$  for  $f_2$ , where  $\Delta^K$  is the  $K$ -dimensional simplex and  $\varphi$  is the same link function that was previously defined. The cross entropy version of the loss that we consider adds an arbitrary function of the cost  $g(c)$  and is given by:

$$\ell_{ce}^{g(c)}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}, x, z, y) = -((1 - \tilde{r}(x)) \log(\mathbf{p}_y^1) + \tilde{r}(x) (\log(\mathbf{p}_y^2) + g(c))) . \quad (19)$$

We again consider the associated risk:

$$R_{ce}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}) = \mathbb{E}_{p(x,z,y)}[\ell_{ce}^{g(c)}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}, x, z, y)] , \quad (20)$$

and the minimizing function:

$$f_{ce}^* = \arg \min_{\mathbf{p}^1, \mathbf{p}^2, \tilde{r} \in [0,1]} R_{ce}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}). \quad (21)$$

The following lemma shows that this cross-entropy surrogate loss cannot be Bayes-consistent.

**Lemma 4.2.** *There is no function  $g(\cdot)$  for which the solution  $f_{ce}^*$  to the associated problem in Eqn. 21 is equal to the Bayes-classifier  $f^*$  defined in Eqn. 6 for all distributions  $p(X, Z, Y)$ .*

The proof is included in Appendix A.5.

## 5 Experiments

### 5.1 Synthetic Experiments

To validate our findings, we present a synthetic experiment in which the ground-truth posteriors are known. We design a simple  $K$ -class classification task with one-dimensional inputs  $X$  and  $Z$  to enable visualization of the learned functions. Our primary interest lies in visualizing the decision boundary  $\tilde{r}(x)$  of a model  $f$  trained with the proposed surrogate loss. This boundary should closely approximate the optimal decision rule  $r^*(x)$ , as defined in Eqn. 9.

**Task Description** The inputs  $X$  and  $Z$  are drawn uniformly from the interval  $[-1, 1]$ . The label  $Y$  is sampled from a categorical distribution with parameter  $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_K]^T \in [0, 1]^K$ , where  $\sum_{i=1}^K \theta_i = 1$  and  $p(Y = i) = \theta_i$ . The function  $\boldsymbol{\theta}(x, z)$  is defined piecewise by partitioning the domain of  $x, z$  into  $K - 1$  slanted regions. Full details of the construction of the synthetic dataset are provided in Appendix A.1.1. The random variables are distributed as:

$$X \sim \text{Uniform}[-1, 1] = p(X) \quad (22)$$

$$Z \sim \text{Uniform}[-1, 1] = p(Z) \quad (23)$$

$$Y \sim \text{Categorical}(\boldsymbol{\theta}(x, z)) = p(Y|x, z) \quad (24)$$

The constructed task can be visualized in Figure 1, where we show the class distribution in terms of most likely class and the samples  $x_i, y_i, z_i \sim p(X, Z, Y)$  for  $K = 5$ . For this example, we can see that at  $x = 0$ , the value of  $z$  provides essentially no additional information to estimate the correct posterior, which should translate into no deferral to  $f_2$  ( $r^*(x = 0) = 0$ ). At  $x = 0.25$ , the variable  $z$  becomes informative. Therefore, the optimal decision function  $r^*(x)$  will alternate as vertical strips along the  $x$ -axis, with width of size that varies based on the cost parameter  $c$ .

Given this construction, the exact posterior probabilities can be computed in closed form, allowing us to derive the optimal decision rule  $r^*(x)$ . To approximate the expectations  $\mathbb{E}_{p(Z|x)}$ , we use Monte Carlo estimation by sampling from  $P(Z|x)$ :

$$\hat{r}^*(x) \approx \begin{cases} 0 & \text{if } \max_y \frac{1}{M} \sum_{i=1}^M \zeta_y(x, z_i) \geq \frac{1}{M} \sum_{j=1}^M \max_{y \in \mathcal{Y}} \theta_y(x, z_j) - c \\ 1 & \text{otherwise} \end{cases} \quad (25)$$

$$\text{where } z_i, z_j \sim \text{Uniform}[-1, 1] \quad (26)$$

In our experiments, we use  $M = 1000$  samples to approximate the expectations.

**Training details** We build simple 3-layer neural networks (NN) for  $\mathbf{t}$ ,  $\mathbf{v}$ , and  $\tilde{r}$ . Following the requirements for  $\tilde{r}(x)$ , the corresponding network takes  $x$  as input and ends with a sigmoid activation. For  $\mathbf{t}$  and  $\mathbf{v}$ , the NNs take  $x$  and  $(x, z)$  as inputs, respectively, and output a real-valued vector of dimension  $K - 1$ . Appendix A.1.2 provides parameter size and layer details. We use the Adam optimizer with learning rate  $lr = 0.001$ , a batch size of 512 and train for 50 epochs using our surrogate loss defined in Eqn. 11. The training set size is  $N_{tr} = 10,000$  and the test set size is  $N_{te} = 1,000$ .

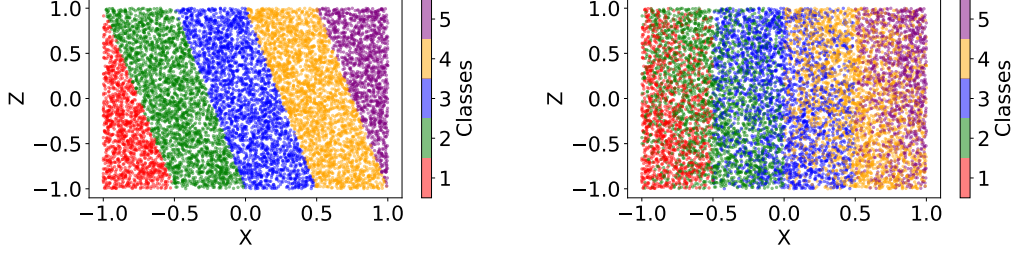


Figure 1: Visualization of multi class synthetic dataset with  $k = 5$ . **Left)** Max probability labels  $\arg \max_y p(Y = y|X, Z)$ . **Right)** Samples  $x, z, y \sim p(X, Z, Y)$  of the synthetic experiment.

**Result and discussion** Figure 2 illustrates the ground truth and predicted decision boundaries for cost values  $c = 0.03, 0.07$  and number of classes  $K = 3, 5$ . We observe that the model trained with the proposed surrogate loss successfully learns the correct decision boundary across different cost values and numbers of classes  $K$ . The learned decision function  $\tilde{r}(x)$  perfectly tracks with the ground truth  $r^*(x)$ . Additionally, although the trained model can output any value in the range  $\tilde{r}(x) \in [0, 1]$  due to the sigmoid activation, it learns to produce sharp values near 0 or 1, which is the optimal behavior.

Now if we turn to a model trained with the additive version of the cross-entropy-based surrogate loss introduced in Eqn 19, using the identity function  $g(c) = c$  with  $K = 5$ , we observe in Figure 3 that the behavior of the learned decision function  $\tilde{r}(x)$  differs significantly. First, we note that since consistency cannot be established for this surrogate loss, it is not possible to precisely target a desired cost level in the  $\ell_{01c}$  loss, unlike with the hinge-based surrogate. Looking at the results, the learned decision boundaries are generally unstable and uneven. The correct pattern of deferral for  $K = 5$  can be observed in the top-right plot of Figure 2, where we see that four regions should be evenly spaced out and deferred to  $f_2$  (regardless of  $c$ ). This pattern is not adequately learned in Figure 3. For instance, we see that the right-most region of  $x$  that should be deferred is slowly erased as the cost increases.

Lastly, we visualize the behavior of the learned model  $f_{\text{hinge}}$  during training in Figure 4. We track the empirical target risk estimated from sampling  $\hat{R}_{01c}(f) = \frac{1}{N} \sum_{i=1}^N \ell_{01c}(f(x_i, z_i), y_i)$  and observe that it converges to the (empirical) optimal risk  $\hat{R}_{01c}^*$  as expected.

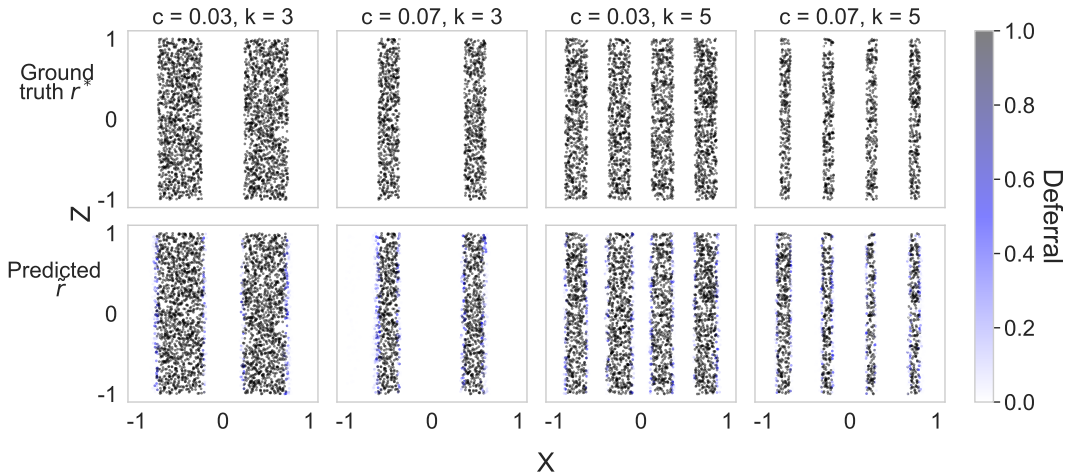


Figure 2: **Top)** Ground truth decision boundary  $\hat{r}^*(x)$  with 2 costs values  $c = 0.03, 0.07$  and number of classes  $K = 3, 5$ . **Bottom)** Learned  $\tilde{r}(x)$  of the model that was trained with our surrogate loss. In all cases, the two decision boundaries are perfectly aligned, confirming our result that the model trained with the proposed surrogate loss successfully learns the optimal decision function. As the cost increases, the green region which represents points deferred to  $f_2$  shrinks.

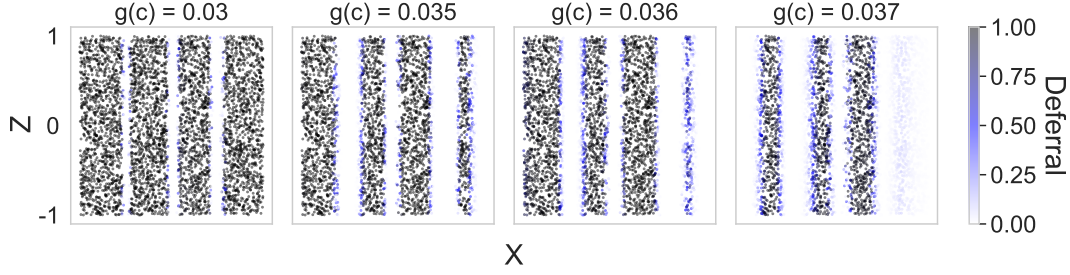


Figure 3: Learned  $\tilde{r}(x)$  of a model trained with the additive cross-entropy surrogate loss with  $g(c) = c$ , for varying  $c$  values. Unlike the model trained with the hinge-based surrogate, the learned decision patterns are generally wrong and not consistent.

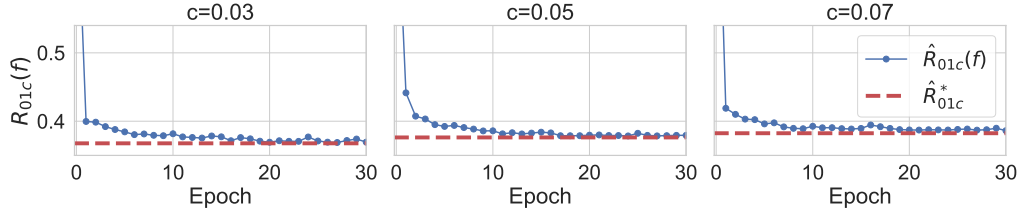


Figure 4: Empirical 0-1 risks of the learned function trained with the surrogate loss and of the optimal solution for  $K = 5$ . We can see that for varying cost values, function trained with the surrogate loss converges to the optimal solution.

## 257 5.2 Large Language Model Experiment

258 To illustrate a practical setting of the problem we consider, we present an experiment based on large  
 259 language models (LLMs). In this experiment, we use two LLMs of different sizes, which correspond  
 260 to different inference costs. The additional inference cost used by the larger model corresponds to  $c$   
 261 in our setup. The task involves solving multi-answer math questions. The intuition behind this setting  
 262 is that some test questions should be more difficult than others. Therefore, it would be desirable to  
 263 efficiently dispatch simpler questions to the smaller LLM and more challenging ones to the larger  
 264 LLM. This allows us to achieve strong performance at a reasonable inference cost.

265 **Task description** We use the Instruction-Tuned Pre-trained models LLaMA 3 8B and LLaMA 3  
 266 70B [Grattafiori et al., 2024] to solve multi-answer math questions from the AQUA dataset [Zhong  
 267 et al., 2024]. The AQUA dataset is composed of multiple-choice math reasoning questions, each with  
 268 5 choices. We frame the task as a 5-class classification problem, where the model must select the  
 269 correct option from a fixed set. The inputs  $x$  and  $z$  are formed by extracting the hidden-states from  
 270 the final tokens of the 8B LLM and the 70B LLM, respectively. We use the first 1000 AQUA [Zhong  
 271 et al., 2024] datapoints from the test split as our dataset, and use a 80/10/10 train/val/test split.

272 **Training details** We use a similar architecture to the one previously presented. The model is trained  
 273 for 1000 epochs using a learning rate of 0.001, a batch size of 32, and early stopping with a patience  
 274 of 20 epochs. Additional details are provided in Appendix A.1.3.

275 **Results and discussion** Although we do not have access to the ground truth decision function  
 276 in this setting, we can examine the accuracy evaluated on the selected samples vs. all samples of  
 277 the model trained with the surrogate loss. The selected samples of  $f_1$  or  $f_2$  are the samples routed  
 278 to these functions by  $\tilde{r}(x)$ . Ideally, the two-stage classifier model  $f$  should learn to route “hard”  
 279 examples to  $f_2$ , and “easier” examples to  $f_1$ . In practice, the surrogate loss can have two effects: 1)  
 280  $f_1$  and  $f_2$  are additionally trained on their respective selected samples; and 2)  $f_1$  and  $f_2$  may receive  
 281 smaller gradient updates depending on the average deferral rates.

282 These two effects can be observed in Figure 5. In the left figure tracking  $f_1$ , we see that the average  
 283 accuracy slightly increases as the cost increases (and consequently the deferral rate decreases), and



the inverse behavior can be seen for  $f_2$  in the right figure.  $f_1$  should, in principle, be given an easier task, so we can expect its selected accuracy to be higher than the average accuracy, which we observe in the left panel of Figure 5. The two values are closest when the selected samples comprise almost all the data (i.e., a deferral rate of 90%). For  $f_2$ , the selected accuracy is closer to the average accuracy. This could suggest that training only on the samples deferred to  $f_2$  does not result in better performance—possibly because these consist of “harder” instances.

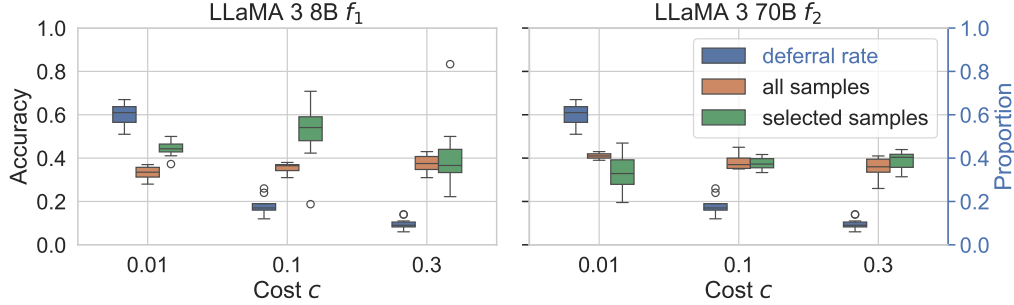


Figure 5: Deferral rate and average accuracy on all samples and on selected samples by **left**) LLaMA 3 8B  $f_1$  and by **right**) LLaMA 3 70B  $f_2$ . The confidence intervals are computed on 10 trials.

In addition to aggregate performance, we can also inspect which types of queries are routed to each model. Figure 6 shows examples of math questions that were consistently routed to the smaller model ( $f_1$ ) and the larger model ( $f_2$ ) across various cost settings. From the presented examples, it appears that the “easy” questions that were consistently routed to the small LLM ( $f_1$ ) generally involve basic arithmetic or proportions. In contrast, the labeled “hard” questions that were consistently routed to the large LLM ( $f_2$ ) seem to require more comprehensive knowledge (such as motion or number theory). This suggests that the routing function aligns with our perceived notion of complexity and the type of reasoning required. See Appendix A.2 for the complete list of questions that were consistently routed to  $f_1$  and  $f_2$ .

Example questions routed to the small LLM ( $f_1$ )	
<b>Question 1:</b>	The cost of 10 kg of mangos is equal to the cost of 24 kg of rice. The cost of 6 kg of flour equals the cost of 2 kg of rice. The cost of each kg of flour is \$22. Find the total cost of 4 kg of mangos, 3 kg of rice and 5 kg of flour?
<b>Question 2:</b>	A man buys an article and sells it at a profit of 20%. If he had bought it at 20% less and sold it for Rs.75 less, he could have gained 25%. What is the cost price?
Example questions routed to the larger LLM ( $f_2$ )	
<b>Question 1:</b>	Two trains 140 m and 160 m long run at the speed of 60 km/hr and 40 km/hr respectively in opposite directions on parallel tracks. The time which they take to cross each other is?
<b>Question 2:</b>	If the product of two numbers is 17820 and their H.C.F. is 12, find their L.C.M.

Figure 6: Sampled questions that are consistently being routed to  $f_1$  or  $f_2$  across different costs.

## 6 Conclusion and Limitations

In conclusion, this work aims to solidify the theoretical foundation behind the design and use of loss functions for the increasingly relevant problem of training multiple models with different costs, while also learning which model to use. We formalized this problem using a principled 0-1-cost-based loss formulation and proposed a surrogate loss based on the hinge loss, showing its consistency. **Limitations** A clear limitation of our work is that we only consider two models in our setup, whereas dynamic networks often require more than two. Extending our approach to the multi-stage setting would be a valuable direction for future research. Moreover, although the theoretical results guarantee loss consistency, the hinge loss is less commonly used in practice. While we have presented a simple proposal of a cross-entropy surrogate loss and shown that it is insufficient for this setting, exploring alternative, more stable losses would be an important next step to ensure the development of practical and principled methods.

## References

- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2022.
- Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie GU, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, 2022.
- Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. Ee-llm: Large-scale training and inference of early-exit large language models with 3d parallelism. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.
- C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
- C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957.
- Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proc. ACM Conf. on Recommender Systems*, page 191–198, 2016.
- Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.
- Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
- Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2019.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira

362 Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain  
 363 Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar  
 364 Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov,  
 365 Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale,  
 366 Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane  
 367 Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha,  
 368 Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal  
 369 Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet,  
 370 Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin  
 371 Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide  
 372 Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei,  
 373 Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan,  
 374 Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey,  
 375 Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma,  
 376 Alex Boesenber, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo,  
 377 Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew  
 378 Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita  
 379 Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh  
 380 Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola,  
 381 Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence,  
 382 Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu,  
 383 Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris  
 384 Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel  
 385 Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich,  
 386 Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine  
 387 Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban  
 388 Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat  
 389 Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella  
 390 Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang,  
 391 Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha,  
 392 Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan  
 393 Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai  
 394 Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya,  
 395 Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica  
 396 Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan  
 397 Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal,  
 398 Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran  
 399 Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A,  
 400 Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca  
 401 Wehrstedt, Madian Khabza, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson,  
 402 Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally,  
 403 Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov,  
 404 Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat,  
 405 Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White,  
 406 Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich  
 407 Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem  
 408 Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager,  
 409 Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang,  
 410 Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra,  
 411 Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ  
 412 Howes, Rutu Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh,  
 413 Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji  
 414 Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin,  
 415 Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang,  
 416 Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe,  
 417 Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny  
 418 Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara  
 419 Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou,  
 420 Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish

421 Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov,  
422 Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian  
423 Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi,  
424 Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao,  
425 Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu  
426 Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

427 Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang. Dynamic neural networks: A survey.  
428 *IEEE Trans. on Pattern Analysis; Mach. Intell.*, 44(11):7436–7456, 2022a.

429 Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfen Cao, Wenhui Huang, Chao  
430 Deng, and Gao Huang. Learning to weight samples for dynamic early-exiting networks. In *Proc.*  
431 *European Conf. on Computer Vision (ECCV)*, 2022b.

432 Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of*  
433 *Statistics / La Revue Canadienne de Statistique*, 34(4):709–721, 2006.

434 Wittawat Jitkrittum, Neha Gupta, Aditya K Menon, Harikrishna Narasimhan, Ankit Rawat, and  
435 Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Proc. Neural Information*  
436 *Processing Systems (NeurIPS)*, 2023.

437 Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral  
438 to multiple experts. In *Proc. AAAI/ACM Conf. on AI, Ethics, and Society*, 2021.

439 Bartłomiej Krzepkowski, Monika Michaluk, Franciszek Szarwacki, Piotr Kubaty, Jary Pomponi,  
440 Tomasz Trzciński, Bartosz Wójcik, and Kamil Adamczewski. Joint or disjoint: Mixing training  
441 regimes for early-exit models, 2024.

442 Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative  
443 decoding. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

444 Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava,  
445 Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. Deja vu: Contextual sparsity for  
446 efficient llms at inference time. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

447 Phil Long and Rocco Servedio. Consistency versus realizable  $h$ -consistency for multiclass classifica-  
448 tion. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2013.

449 David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and  
450 transferable representations. arXiv preprint: arXiv 1802.06309, 2018.

451 Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with  
452 multiple experts. In *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, 2023.

453 Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable  $h$ -consistent and bayes-consistent loss  
454 functions for learning to defer. In *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, 2024a.

455 Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with  
456 multiple experts. In *Artificial Intelligence and Image Analysis*, 2024b.

457 Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In  
458 *Proc. Int. Conf. on Machine Learning (ICML)*, 2020.

459 Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag.  
460 Who should predict? exact algorithms for learning to defer to humans. In *Proc. Int. Conf. on*  
461 *Artificial Intelligence and Statistics (AISTAT)*, 2023.

462 Florence Regol, Joud Chataoui, and Mark Coates. Jointly-learned exit and inference for a dynamic  
463 neural network. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.

464 Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and  
465 Donald Metzler. Confident adaptive language modeling. In *Proc. Adv. in Neural Inf. Proces. Syst.*  
466 *(NeurIPS)*, 2022.

467 Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*,  
468 26(2):225–287, Aug 2007.

469 Bernadetta Tarigan and Sara A. van de Geer. A moment bound for multi-hinge classifiers. *Journal of*  
470 *Machine Learning Research*, 9(71):2171–2185, 2008.

471 Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal*  
472 *of Machine Learning Research*, 8(36):1007–1025, 2007.

473 Rajeev Verma and Eric T. Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *Proc.*  
474 *Int. Conf. on Machine Learning (ICML)*, 2022.

475 Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent  
476 surrogate losses, confidence calibration, and conformal ensembles. In *Proc. Int. Conf. on Artificial*  
477 *Intelligence and Statistics (AISTAT)*, 2022.

478 Yair Wiener and Ran El-Yaniv. Agnostic selective classification. In *Proc. Adv. Neural Info. Process.*  
479 *Syst. (NeurIPS)*, 2011.

480 Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proc. Int. Joint*  
481 *Conf. on Artificial Intelligence*, 2021.

482 Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and  
483 Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey of  
484 speculative decoding. arXiv preprint: arXiv 2401.07851, 2024.

485 Haichao Yu, Haoxiang Li, Gang Hua, Gao Huang, and Humphrey Shi. Boosted dynamic neural  
486 networks. In *Proc. AAAI Conf. on Artif. Intell.*, 2022.

487 Ziqian Zeng, Yihuai Hong, Hongliang Dai, Huiping Zhuang, and Cen Chen. ConsistentEE: A  
488 consistent and hardness-guided early exiting method for accelerating language models inference.  
489 In *Proc. AAAI Conf. Artif. Intell.*, pages 19506–19514, Mar. 2024.

490 Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu  
491 Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In  
492 *Findings of the Association for Computational Linguistics*, June 2024.

## 493 A Appendix

### 494 NeurIPS Paper Checklist

#### 495 1. Claims

496 Question: Do the main claims made in the abstract and introduction accurately reflect the  
497 paper’s contributions and scope?

498 Answer: [Yes]

499 Justification: The contributions are stated in the introduction are supported by theoretical  
500 results and experiments.

#### 501 2. Limitations

502 Question: Does the paper discuss the limitations of the work performed by the authors?

503 Answer: [Yes]

504 Justification: We include a discussion on the limitation of our work in the conclusion.

#### 505 3. Theory assumptions and proofs

506 Question: For each theoretical result, does the paper provide the full set of assumptions and  
507 a complete (and correct) proof?

508 Answer: [Yes]

509 Justification: Proofs are included in the Appendix.

#### 510 4. Experimental result reproducibility

511 Question: Does the paper fully disclose all the information needed to reproduce the main ex-  
512 perimental results of the paper to the extent that it affects the main claims and/or conclusions  
513 of the paper (regardless of whether the code and data are provided or not)?

514 Answer: [Yes]

515 Justification: Implementation details are included in the main text and in the Appendix.

#### 516 5. Open access to data and code

517 Question: Does the paper provide open access to the data and code, with sufficient instruc-  
518 tions to faithfully reproduce the main experimental results, as described in supplemental  
519 material?

520 Answer: [Yes]

521 Justification: n/a

#### 522 6. Experimental setting/details

523 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-  
524 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the  
525 results?

526 Answer: [Yes]

527 Justification: The details are included in the main text and in the Appendix.

#### 528 7. Experiment statistical significance

529 Question: Does the paper report error bars suitably and correctly defined or other appropriate  
530 information about the statistical significance of the experiments?

531 Answer: [Yes]

532 Justification: We report statistical significance on our real data experiment. res or tables in  
533 the text.

#### 534 8. Experiments compute resources

535 Question: For each experiment, does the paper provide sufficient information on the com-  
536 puter resources (type of compute workers, memory, time of execution) needed to reproduce  
537 the experiments?

538 Answer: [NA]

539 Justification: The required computational resources were minimal.

540 **9. Code of ethics**

541 Question: Does the research conducted in the paper conform, in every respect, with the

542 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

543 Answer: [Yes]

544 Justification: This research is inline with NeurIPS Code of Ethics.

545 **10. Broader impacts**

546 Question: Does the paper discuss both potential positive societal impacts and negative

547 societal impacts of the work performed?

548 Answer: [NA]

549 Justification:

550 answerNA

551 **11. Safeguards**

552 Question: Does the paper describe safeguards that have been put in place for responsible

553 release of data or models that have a high risk for misuse (e.g., pretrained language models,

554 image generators, or scraped datasets)?

555 Answer: [NA]

556 Justification: [NA]

557 **12. Licenses for existing assets**

558 Question: Are the creators or original owners of assets (e.g., code, data, models), used in

559 the paper, properly credited and are the license and terms of use explicitly mentioned and

560 properly respected?

561 Answer: [NA]

562 Justification: [NA]

563 **13. New assets**

564 Question: Are new assets introduced in the paper well documented and is the documentation

565 provided alongside the assets?

566 Answer: [NA]

567 Justification: [NA]

568 **14. Crowdsourcing and research with human subjects**

569 Question: For crowdsourcing experiments and research with human subjects, does the paper

570 include the full text of instructions given to participants and screenshots, if applicable, as

571 well as details about compensation (if any)?

572 Answer: [NA]

573 Justification: [NA]

574 **15. Institutional review board (IRB) approvals or equivalent for research with human**

575 **subjects**

576 Question: Does the paper describe potential risks incurred by study participants, whether

577 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

578 approvals (or an equivalent approval/review based on the requirements of your country or

579 institution) were obtained?

580 Answer: [NA]

581 Justification: [NA]

582 **16. Declaration of LLM usage**

583 Question: Does the paper describe the usage of LLMs if it is an important, original, or

584 non-standard component of the core methods in this research? Note that if the LLM is used

585 only for writing, editing, or formatting purposes and does not impact the core methodology,

586 scientific rigorousness, or originality of the research, declaration is not required.

587 Answer: [NA]

588 Justification: [NA]

## 589 A.1 Additional experimental details

### 590 A.1.1 Synthetic task

591 In this section, we provide additional details of the synthetic task. We model  $Y$  by a categorical  
 592 distribution with parameter  $\theta = [\theta_1, \theta_2, \dots, \theta_k]^T \in [0, 1]^k$  satisfying  $\sum_{i=1}^k \theta_i = 1$  and  $p(Y =$   
 593  $i) = \theta_i$ .  $\theta$  is modeled as a piecewise function by partitioning the range of  $X$  into  $k - 1$  equally  
 594 sized bins  $\{B_1, B_2, \dots, B_{k-1}\}$ . Assuming the range of  $X$  is  $[a, b]$ ,  $a, b \in \mathbb{R}$ , we define  $B_i =$   
 595  $\left[a + \frac{(i-1)(b-a)}{k-1}, a + \frac{i(b-a)}{k-1}\right)$ . Within every bin  $B_i$ , only  $\theta_i$  and  $\theta_{i+1}$  take on non-zero values,  
 596 following a scaled and shifted sigmoid:

$$X \sim \text{Uni}[-1, 1] \quad (27)$$

$$Z \sim \text{Uni}[-1, 1] \quad (28)$$

$$Y \sim \text{Categorical}(\theta) \quad (29)$$

$$\theta = [\theta_1, \theta_2, \dots, \theta_k]^T \in [0, 1]^k \quad (30)$$

$$B_i = \left[-1 + \frac{2(i-1)}{k-1}, -1 + \frac{2i}{k-1}\right) \forall i \in [1, 2, \dots, k-1] \quad (31)$$

$$\theta_i = \begin{cases} \sigma(sX + c_i + Z) & \text{if } X \in B_i \\ \sigma(-1 \times (sX + c_{i-1} + Z)) & \text{if } i > 1 \text{ and } X \in B_{i-1} \\ 0 & \text{else} \end{cases} \quad (32)$$

$$s = k - 1 \quad (33)$$

$$c_i = k - 2i \quad (34)$$

$$(35)$$

597 Using this model, we arrive at the closed form for the posteriors:

$$\eta_{y=i}(x) = \int p(Y = i|x, Z)p(Z|x)dZ = \mathbb{E}_{p(Z)}[\theta_i] \quad (36)$$

$$= \mathbb{E}_{p(Z)} \left[ \begin{cases} \sigma(sx + c_i + Z) & \text{if } x \in B_i \\ \sigma(-1 \times (sx + c_{i-1} + Z)) & \text{if } i > 1 \text{ and } x \in B_{i-1} \\ 0 & \text{else} \end{cases} \right], \quad (37)$$

$$\zeta_{y=i}(x, z) = \theta_i = \begin{cases} \sigma(sx + c_i + z) & \text{if } x \in B_i \\ \sigma(-1 \times (sx + c_{i-1} + z)) & \text{if } i > 1 \text{ and } x \in B_{i-1} \\ 0 & \text{else} \end{cases}, \quad (38)$$

598 and we can derive the optimal decision function  $r^*(x)$ :

$$r^*(x) = \begin{cases} 0 & \text{if } \max_i \mathbb{E}_{p(Z)}[\theta_i] \geq \mathbb{E}_{p(Z)}[\max_i \theta_i] - c, \\ 1 & \text{o.w..} \end{cases} \quad (39)$$

### 599 A.1.2 Synthetic Model details

600 We describe the architecture of the model used in the synthetic experiment. The hidden size of all  
 601 networks is 64. Each neural network is defined as:

$$r_\theta(x) = \text{sigmoid}(\text{BatchNorm}(\Theta \text{ReLU}(\text{BatchNorm}(\theta x)))) \quad (40)$$

$$\mathbf{t}_\theta(x) = \Theta \text{ReLU}(\text{BatchNorm}(\Theta \text{ReLU}(\text{BatchNorm}(\theta x)))) \quad (41)$$

$$\mathbf{v}_\theta(x, z) = \Theta \text{ReLU}(\text{BatchNorm}(\Theta \text{ReLU}(\text{BatchNorm}(\Theta[x, z]))))). \quad (42)$$

### 602 A.1.3 LLM Model details

603 We describe the architecture of the model used in the LLM experiment. The hidden size of all  
 604 networks is 128. We performed a grid search for the hidden size across the values  $\{32, 64, \underline{128}, 256\}$   
 605 and for the learning rate across the values  $\{0.01, \underline{0.001}, 0.0001\}$ .



Each neural network is defined as:

$$r_{\theta}(x) = \text{sigmoid}(\text{BatchNorm}(\Theta \text{ReLU}(\text{BatchNorm}(\Theta x)))) \quad (43)$$

$$\mathbf{t}_{\theta}(x) = \Theta \text{ReLU}(\text{BatchNorm}(\Theta \text{ReLU}(\text{BatchNorm}(\Theta x)))) \quad (44)$$

$$\mathbf{v}_{\theta}(x, z) = \Theta \text{ReLU}(\text{BatchNorm}(\Theta \text{ReLU}(\text{BatchNorm}(\Theta [x, z])))). \quad (45)$$

## A.2 Complete list of deferred questions (LLM experiments)

In this section, we provide a comprehensive list of the questions that were consistently deferred to  $f_2$  or sent to  $f_1$  and in the LLM experiment.

### Example questions sent to the small LLM ( $f_1$ ):

1. A man buys an article and sells it at a profit of 20%. If he had bought it at 20% less and sold it for Rs. 75 less, he could have gained 25%. What is the cost price?
2. The cost of 10 kg of mangos is equal to the cost of 24 kg of rice. The cost of 6 kg of flour equals the cost of 2 kg of rice. The cost of each kg of flour is \$22. Find the total cost of 4 kg of mangos, 3 kg of rice and 5 kg of flour?
3. The speed of a boat in upstream is 100 kmph and the speed of the boat downstream is 180 kmph. Find the speed of the boat in still water and the speed of the stream.
4. A and B working together could mow a field in 28 days and with the help of C they could have mowed it in 21 days. How long would C take by himself?
5. Evaluate the expression:  

$$2^2 + 4^2 + 6^2 + \dots + 22^2$$
6. In an examination, 60% failed in Math and 40% failed in French. If 15% failed in both subjects, what percentage of students passed in both?
7. One train crosses a bridge of length 340 m in 42 seconds, and the same train crosses another bridge of length 500 m in 50 seconds. What is the approximate speed of the train in km/hr?
8. Eshan and Mary each wrote two or three poems every day over a period of time. Eshan wrote 43 poems while Mary wrote 61. What is the number of days in this period?
9. Find the value of  $x$  in the sequence of numbers 5, 1, 6, 0, 4, 8,  $x$ , 2 if the sum of the first 7 numbers is 30 and the average is 4.
10. Roja and Pooja start moving in opposite directions from a pole. They are moving at speeds of 7 km/hr and 3 km/hr respectively. After 4 hours, what will be the distance between them?

### Example questions deferred to the larger LLM ( $f_2$ ):

1. If the product of two numbers is 17820 and their H.C.F. is 12, find their L.C.M.
2. Two passenger trains start at the same hour in the day from two different stations and move towards each other at the rate of 14 kmph and 21 kmph respectively. When they meet, it is found that one train has traveled 60 km more than the other one. What is the distance between the two stations?
3. Which is the odd one: 10, 25, 45, 54, 60, 75, 80?
4. Two trains 140 m and 160 m long run at the speed of 60 km/hr and 40 km/hr respectively in opposite directions on parallel tracks. The time which they take to cross each other is?
5. A ladder 100 feet long is leaning against a vertical wall. Its lower end is 60 feet from the bottom of the wall. The side of the largest cubical box that can be placed between the wall and the ladder without disturbing the ladder is (to the nearest foot)?
6. On dividing a certain number by 5, 7 and 8 successively, the remainders obtained are 2, 3 and 4 respectively. When the order of division is reversed and the number is successively divided by 8, 7 and 5, what will be the respective remainders?
7. A tour group of 25 people paid a total of \$670 for entrance to a museum. If this price included a 5% sales tax, and all the tickets cost the same amount, what was the face value of each ticket price without the sales tax?

- 649 8. A rectangular floor is covered by a rug except for a strip 4 meters wide along each of the four  
650 edges. If the floor is 25 meters by 20 meters, what is the area of the rug in square meters?  
651 9. In each of the following questions a number series is given with one term missing. Choose  
652 the correct alternative that will continue the same pattern and fill in the blank space.

2, 7, 14, ?, 34, 47

- 653 10. In a game of 500 points there are three participants A, B, and C. A gives to B 80 points and  
654 to C 101 points. Then how many points can B give to C?  
655 11. When magnified 1,000 times by an electron microscope, the image of a certain circular piece  
656 of tissue has a diameter of 2 centimeters. The actual diameter of the tissue, in centimeters,  
657 is:  
658 12. From the given equation, find the value of  $x$ :

$$2x^2 + 9x - 5 = 0$$

- 659 13. The sum of the non-prime numbers between 50 and 60, non-inclusive, is:  
660 14. Solve the system of equations to find the values of  $c$  and  $d$ :

$$\text{I. } c^3 - 988 = 343 \quad (46)$$

$$\text{II. } d^2 - 72 = 49 \quad (47)$$

- 661 15. How many minutes does Aditya take to cover a distance of 400 meters, if he runs at a speed  
662 of 20 km/hr?  
663 16. An engineer designed a ball so that when it was dropped, it rose with each bounce exactly  
664 one-half as high as it had fallen. The engineer dropped the ball from an 18-meter platform  
665 and caught it after it had traveled 53.4 meters. How many times did the ball bounce?

### 666 A.3 Proof of the solution $f^*$

667 In this section, we provide the proof of Lemma 3.1, which states that the optimal solution

$$f^* = \arg \min_{f \in \mathcal{H}} R_{01c}(f) \quad (48)$$

668 is the following:

$$f^* = \begin{cases} \arg \max_y \eta_y(x), & \text{if } \max_y \eta_y(x) \geq \mathbb{E}_{p(Z|x)}[\max_y \zeta_y(x, Z)] - c, \\ \arg \max_y \zeta_y(x, z), & \text{else.} \end{cases} \quad (49)$$

669 Or alternatively:

$$f^* = \begin{cases} f_1^*(x), & r^*(x) = 0 \\ f_2^*(x, z), & r^*(x) = 1 \end{cases}, \quad (50)$$

$$\text{where } f_1^*(x) = \arg \max_{y \in \mathcal{Y}} \eta_y(x), \quad \forall x \text{ s.t. } r^*(x) = 0 \quad (51)$$

$$f_2^*(x, z) = \arg \max_{y \in \mathcal{Y}} \zeta_y(x, z), \quad \forall x \text{ s.t. } r^*(x) = 1 \quad (52)$$

$$r^*(x) = \mathbb{1} \left[ \max_y \eta_y(x) \leq \mathbb{E}_{p(Z|x)}[\max_y \zeta_y(x, Z)] - c \right]. \quad (53)$$

670 *Proof.* We start by evaluating the risk w.r.t to the function:

$$f^s = \begin{cases} \arg \max_y \eta_y(x) & \text{if } \max_y \eta(x) \geq \mathbb{E}_{p(z|x)} \max \zeta(x, z) - c \\ \arg \max_y \zeta_y(x, z) & \text{else,} \end{cases} \quad (54)$$

$$R_{01c}(f^s) = \mathbb{E}_{p(x,z,y)}[\ell_{01c}(f^s(x, z), y)], \quad (55)$$

671 and prove the result by showing that any function  $f^o \neq f^s$  results in a higher risk, therefore showing  
672 that  $f^s$  is the optimal solution.

673 The risk is given by:

$$R_{01c}(f^s) = \mathbb{E}_{p(x,z,y)}[\ell_{01c}(f^s(x,z),y)] \quad (56)$$

$$R_{01c}(f^s) = \int_x \int_z \sum_y [\ell_{01c}(f^s(x,z),y)] p(y|x,z) p(z|x) p(x) dz dx. \quad (57)$$

674 We can partition  $\mathcal{X}$  in two regions based on the decision function of  $f^s$ , i.e. :  $A = \{x; \max \eta(x) \geq$   
 675  $\mathbb{E}_{p(z|x)} \max \zeta(x,z) - c\}$  and  $B = \{x; \max \eta(x) \leq \mathbb{E}_{p(z|x)} \max \zeta(x,z) - c\}$  and split the expectation  
 676 in two terms:

$$R_{01c}(f^s) = R^A + R^B \quad (58)$$

$$\text{where } R^S \triangleq \int_{x \in S} \int_z \sum_y [\ell_{01c}(f^s(x,z),y)] p(y|x,z) p(z|x) p(x) dx dz. \quad (59)$$

677 Looking at both terms separately, starting with  $R^A$  where  $f^s$  does not use  $z$  (or corresponds to  $f_1$ ):

$$R^A = \int_{x \in A} \int_z \sum_y [\ell_{01c}(f^s(x,z),y)] p(y|x,z) p(z|x) p(x) dx dz \quad (60)$$

$$= \int_{x \in A} \sum_y \left( \int_z \mathbb{1}[f_1(x) \neq y] p(z|x,y) dz \right) p(x,y) dx \text{ by def of } f^s \text{ and } \ell_{01c} \quad (61)$$

$$= \int_{x \in A} \sum_y \mathbb{E}_{z|x,y} [\mathbb{1}[f_1(x) \neq y] p(x,y) dx] \quad (62)$$

$$= \int_{x \in A} \sum_y \mathbb{1}[f_1(x) \neq y] p(x,y) dx \text{ as nothing depends on } z \quad (63)$$

$$= \int_{x \in A} \sum_y \mathbb{1}[\arg \max_{y'} \eta_{y'}(x) \neq y] \eta_y(x) p(x) dx \text{ by def of } f_1^s \text{ and } p(y|x) \quad (64)$$

$$R^A = \int_{x \in A} 1 - \max_y \eta_y(x) p(x) dx \quad (65)$$

678 We can obtain more straightforwardly  $R^B$ :

$$R^B = \int_{x \in B} \int_z (1 + c - \max_y \zeta_y(x,z) p(x,z)) dz dx \quad (66)$$

679 Hence, the risk of  $f^s$  is given by;

$$R_{01c}(f^s) = \int_{x \in A} 1 - \max_y \eta_y(x) p(x) dx + \int_{x \in B} \int_z 1 + c - \max_y \zeta_y(x,z) p(x,z) dz dx. \quad (67)$$

680 Now, we consider a different two-stage classifier  $f^o \in \mathcal{H}$  and  $f^s \neq f^o$ . We show that any  $f^o \in \mathcal{H}$   
 681 will lead to a higher risk  $R_{01c}(f^s) \leq R_{01c}(f^o)$ , therefore proving that  $f^s = f^*$ .

682 We further partition the space  $\mathcal{X}$  by splitting  $A$  and  $B$  where  $f^s \neq f^o$  and  $f^s = f^o$

$$A^s \triangleq \{x; x \in A \text{ and } f^s = f^o\}, \quad A^d \triangleq \{x; x \in A \text{ and } f^s \neq f^o\}, \quad (68)$$

$$B^s \triangleq \{x; x \in B \text{ and } f^s = f^o\}, \quad B^d \triangleq \{x; x \in B \text{ and } f^s \neq f^o\} \quad (69)$$

683 We then use those new partitions to further decompose the risks of  $f^s$  and  $f^o$  in 4 terms:

$$R_{01c}(f^s) = R^{A^s} + R^{A^d} + R^{B^s} + R^{B^d} \quad (70)$$

$$\text{and } R_{01c}(f^o) = R_o^{A^s} + R_o^{A^d} + R_o^{B^s} + R_o^{B^d} \quad (71)$$

684 We can then write the difference between the risks as:

$$R_{01c}(f^s) - R_{01c}(f^o) = R^{A^d} - R_o^{A^d} + R^{B^d} - R_o^{B^d} \quad (72)$$

$$R_{01c}(f^s) - R_{01c}(f^o) = \delta^{A^d} + \delta^{B^d} \quad (73)$$

$$\text{where } \delta^{A^d} \triangleq R^{A^d} - R_o^{A^d} \quad (74)$$

$$\delta^{B^d} \triangleq R^{B^d} - R_o^{B^d} \quad (75)$$

685 In the following, we prove that

$$\delta^{A^d} \leq 0 \text{ and } \delta^{B^d} \leq 0 \quad (76)$$

686 which would imply that

$$R_{01c}(f^s) - R_{01c}(f^o) \leq 0 \quad \forall f \in \mathcal{H} \neq f^s \quad (77)$$

$$\implies R_{01c}(f^o) \geq R_{01c}(f^s) \forall f \in \mathcal{H} \neq f^s \quad (78)$$

$$\implies f^s = f^*. \quad (79)$$

687 **Showing that  $\delta^{A^d} \leq 0$  and  $\delta^{B^d} \leq 0$**  We first consider  $\delta^{A^d}$ :

$$\delta^{A^d} = R^{A^d} - R_o^{A^d} \quad (80)$$

$$= \int_{x \in A^d} 1 - \max_y \eta_y(x) p(x) dx - R_o^{A^d} \text{ using 65} \quad (81)$$

688 We can (once again) further partition the space based on  $f^o$ . We divide  $A^d$  in two based on the  
689 decision function of  $f^o$ :

$$A^{dz} \triangleq \{x; x \in A^d \text{ and } f^o = f_2^o(x, z)\} \quad (82)$$

$$A^{dx} \triangleq \{x; x \in A^d \text{ and } f^o = f_1^o(x)\} \quad (83)$$

690 Continuing our development of  $\delta^{A^d}$ :

$$\delta^{A^d} = \int_{x \in A^d} 1 - \max_y \eta_y(x) p(x) - R_o^{A^d} \quad (84)$$

$$= \int_{x \in A^{dx}} 1 - \max_y \eta_y(x) p(x) dx - R_o^{A^{dx}} + \int_{x \in A^{dz}} 1 - \max_y \eta_y(x) p(x) dx - R_o^{A^{dz}} \quad (85)$$

$$= \int_{x \in A^{dx}} \left( 1 - \max_y \eta_y(x) p(x) - \int_z \sum_y [\ell_{01c}(f^o(x, z), y)] p(y|x, z) p(z|x) p(x) \right) dx \quad (86)$$

$$+ \int_{x \in A^{dz}} \left( 1 - \max_y \eta_y(x) p(x) - \int_z \sum_y [\ell_{01c}(f^o(x, z), y)] p(y|x, z) p(z|x) p(x) \right) dx \text{ using 71} \quad (87)$$

$$= \int_{x \in A^{dx}} \left( 1 - \max_y \eta_y(x) - \int_z \sum_y [\ell_{01c}(f_1^o(x), y)] p(y|x, z) p(z|x) \right) p(x) dx \quad (88)$$

$$+ \int_{x \in A^{dz}} \left( 1 - \max_y \eta_y(x) - \int_z \sum_y [\ell_{01c}(f_2^o(x, z), y)] p(y|x, z) p(z|x) \right) p(x) dz dx \text{ by def. of } A^{dx}, A^{dz} \quad (89)$$

$$= \int_{x \in A^{dx}} \left( 1 - \max_y \eta_y(x) - \sum_y (\mathbb{1}[f_1^o(x) \neq y]) \eta_y(x) \right) p(x) dx \text{ [*]} \quad (90)$$

$$+ \int_{x \in A^{dz}} \left( 1 - \max_y \eta_y(x) - \int_z \sum_y \mathbb{1}[f_2^o(x, z) \neq y] (\zeta_y(x, z) + c) p(z|x) \right) p(x) dz dx \text{ [**]} \quad (91)$$

$$(92)$$

691 At this stage we can focus on one term at the time, starting with the part of  $A^d$  where  $f^o$  is not using  
692  $z$  which is the integral over  $A^{dx}$ :

$$[*] = \int_{x \in A^{dx}} \left( 1 - \max_y \eta_y(x) - \sum_y (\mathbb{1}[f_1^o(x) \neq y]) \eta_y(x) \right) p(x) dx \quad (93)$$

$$= \int_{x \in A^{dx}} \left( - \max_y \eta_y(x) + \eta_{f_1^o(x)}(x) \right) p(x) dx \quad (94)$$

$$\text{no matter what value } f_1^o(x) \text{ takes, } \max_y \eta_y(x) \geq \eta_{f_1^o(x)}. \text{ Therefore:} \quad (95)$$

$$[*] \leq 0 \quad (96)$$

693 Going to the second term  $[**]$ , when  $f^0$  uses  $z$ . We start by restating the definition of the set  $A$ :

$$A = \{x; \max_y \eta_y(x) \geq \mathbb{E}_{p(z|x)}[\max_y \zeta_y(x, z)] - c\} \quad (97)$$

694

$$[**] = \int_{x \in A^{dz}} \left( 1 - \max_y \eta_y(x) - \int_z \sum_y \mathbb{1}[f_2^o(x, z) \neq y] (\zeta_y(x, z) + c) p(z|x) dz \right) p(x) dx \quad (98)$$

$$= \int_{x \in A^{dz}} \left( 1 - \max_y \eta_y(x) - \mathbb{E}_{p(z|x)}[1 - \zeta_{f_2^o(x, z)}(x, z)] \right) p(x) dx \quad (99)$$

$$\leq \int_{x \in A^{dz}} \left( -\max_y \eta_y(x) + \mathbb{E}_{p(z|x)}[\max_y \zeta_y(x, z)] - c \right) p(x) dx \quad (100)$$

$$[**] \leq 0 \text{ by def. of } A \quad (101)$$

695 Combining both results together, we have that

$$\delta^{A^d} = [*] + [**] \leq 0. \quad (102)$$

696 Following similar steps, we also have that

$$\delta^{B^d} \leq 0. \quad (103)$$

697 **Final step** Since we have both that  $\delta^{B^d} \leq 0$  and  $\delta^{A^d} \leq 0$ , we can conclude that

$$R_{01c}(f^s) - R_{01c}(f^o) = \delta^{A^d} + \delta^{B^d} \quad (104)$$

$$R_{01c}(f^s) \leq R_{01c}(f^o) \quad \forall f^o \neq f^s \quad (105)$$

$$\implies f^s = f^* \quad (106)$$

698 This concludes the proof.  $\square$

699

$\square$

#### 700 A.4 Proof of consistency for the multi class surrogate hinge loss

701 We start by restating our surrogate loss:

$$\ell_{hinge}^c(\mathbf{t}, \mathbf{v}, \tilde{r}, x, z, y) = (1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'}(x) + \frac{1}{K-1}]_+ + \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'}(x, z) + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right), \quad (107)$$

702 where we have  $\mathbf{t} \in \mathbb{R}^K$  and  $\mathbf{v} \in \mathbb{R}^K$  as the real-vectorized outputs for  $f_1$  and  $f_2$  respectively with the  
703 constraints that  $\|\mathbf{v}\|_1 = 0$  and  $\|\mathbf{t}\|_1 = 0$ , and  $\tilde{r} \in [0, 1]$  as a soft decision output.

704 Since our surrogate optimization provides us with the triplet  $\mathbf{t}, \mathbf{v}, \tilde{r}$ , we map these to a two-stage  
705 classifier  $f \in \mathcal{H}$  using the following a link function  $\varphi : \mathbb{R}^K \times \mathbb{R}^K \times [0, 1] \rightarrow \mathcal{Y}$ :

$$f = \varphi(\mathbf{t}, \mathbf{v}, r) = \begin{cases} \arg \max_{y \in \mathcal{Y}} \mathbf{t}_y(x) & \text{if } \tilde{r} < 0.5 \\ \arg \max_{y \in \mathcal{Y}} \mathbf{v}_y(x, z) & \text{o.w.} \end{cases} \in \mathcal{H}, \quad (108)$$

$$\text{where } f_1(x) = \arg \max_{y \in \mathcal{Y}} \mathbf{t}_y(x), \quad (109)$$

$$f_2(x, z) = \arg \max_{y \in \mathcal{Y}} \mathbf{v}_y(x, z), \quad (110)$$

$$r(x) = \mathbb{1}[\tilde{r}(x) < 0.5]. \quad (111)$$

706 We can then define our risk as usual:

$$R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) = \mathbb{E}_{p(x, z, y)}[\ell_{hinge}^c(\mathbf{t}, \mathbf{v}, \tilde{r}, x, z, y)]. \quad (112)$$

707 and consider the triplet of minimizers  $\mathbf{t}^*(x), \mathbf{v}^*(x, z), \tilde{r}^*(x)$  of such a risk, which correspond to a  
708 two-stage solution:

$$\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^* = \arg \min_{\mathbf{v}, \mathbf{t} \in \mathbb{R}^K, \tilde{r} \in [0, 1]} R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}). \quad (113)$$

$$f_{hinge}^* = \varphi(\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^*) \in \mathcal{H} \quad (114)$$

709 We prove that our surrogate loss is Bayes-consistent w.r.t to the  $\ell_{01c}$  loss by showing that 1)

$$f^* = f_{hinge}^*, \quad (115)$$

710 and 2)

$$R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^* \leq \Psi(R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^*). \quad (116)$$

711 Taken together, those results guarantee that for any distribution  $p(x, z, y)$ , we have that:

$$R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) \rightarrow R_{hinge}^* \implies R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) \rightarrow R_{01c}^*. \quad (117)$$

712 which defines Bayes-consistency [Steinwart, 2007].

#### 713 A.4.1 The solutions $f^*$ of $R_{01c}(f)$ and $f_{hinge}^*$ coincide

714 Restating the definitions of the solution of the target and surrogate problems;

$$f^* = \arg \min_{f \in \mathcal{H}} R_{01c}(f) \quad (118)$$

$$f_{hinge}^* = \varphi(\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^*) \quad (119)$$

$$\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^* = \arg \min_{\mathbf{t}, \mathbf{v} \in \mathbb{R}^K, \tilde{r} \in [0,1]} R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}). \quad (120)$$

715 In this section, we show that

716 **Lemma A.1.** For any distribution  $p(X, Z, Y)$ ;

$$f^* = f_{hinge}^*. \quad (121)$$

717 *Proof.* We have previously shown in Appendix A.3 that the solution of our targeted problem

$$f^* = \arg \min_{f \in \mathcal{H}} R_{01c}(f) \quad (122)$$

718 is given by;

$$f^* = \begin{cases} f_1^*(x), & r^*(x) = 0 \\ f_2^*(x, z), & r^*(x) = 1 \end{cases}, \quad (123)$$

$$\text{where } f_1^*(x) = \arg \max_{y \in \mathcal{Y}} \eta_y(x), \quad \forall x \text{ s.t. } r^*(x) = 0 \quad (124)$$

$$f_2^*(x, z) = \arg \max_{y \in \mathcal{Y}} \zeta_y(x, z), \quad \forall x \text{ s.t. } r^*(x) = 1 \quad (125)$$

$$r^*(x) = \mathbb{1} \left[ \max_y \eta_y(x) \leq \mathbb{E}_{p(Z|x)}[\max_y \zeta_y(x, Z)] - c \right]. \quad (126)$$

719 Next, we show that the two-stage classifier obtained from the triplet minimizer of our surrogate loss

720  $f_{hinge}^* = \varphi(\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^*)$  corresponds to this solution.

$$\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^* = \arg \min_{\mathbf{t}, \mathbf{v} \in \mathbb{R}^K, \tilde{r} \in [0,1]} R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) \quad (127)$$

$$= \arg \min_{\mathbf{t}, \mathbf{v} \in \mathbb{R}^K, \tilde{r} \in [0,1]} \mathbb{E}_{p(x,z,y)}[\ell_{hinge}^c(\mathbf{t}, \mathbf{v}, \tilde{r}, x, z, y)] \quad (128)$$

$$\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^* = \arg \min_{\mathbf{t}, \mathbf{v} \in \mathbb{R}^K, \tilde{r} \in [0,1]} \mathbb{E}_{p(x,z,y)}[(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ + \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right)]. \quad (129)$$

721 We can push the optimization problem inside the expectation w.r.t  $p(x)$  as  $\mathbf{t}(x), \mathbf{v}(x, z), \tilde{r}(x)$  are all  
722 functions of  $x$  (and the inner expectation term is guaranteed to be bounded):

$$\mathbf{t}^*, \mathbf{v}^*, \tilde{r}^* = \arg \min_{\mathbf{t}, \mathbf{v} \in \mathbb{R}^K, \tilde{r} \in [0,1]} \mathbb{E}_{p(z,y|x)}[(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ + \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right)]. \quad (130)$$

723 Since the loss is a linear combination of two terms that respectively depend on  $\mathbf{t}$  and  $\mathbf{v}$ , we can see  
 724 that for any  $\tilde{r}$ , the minimizers for  $\mathbf{t}$  and  $\mathbf{v}$  will always be equal to the minimizer of the individual  
 725 terms:

$$\mathbf{t}^*(x) = \arg \min_{\mathbf{t} \in \mathbb{R}^K} \mathbb{E}_{p(y|x)} \left[ \sum_{y' \neq y} \left[ \mathbf{t}_{y'} + \frac{1}{K-1} \right]_+ \right] \text{ (for any } \mathbf{v}, \tilde{r}(x) < 1) \quad (131)$$

$$\mathbf{v}^*(x, z) = \arg \min_{\mathbf{v} \in \mathbb{R}^K} \mathbb{E}_{p(y|x, z)} \left[ \sum_{y' \neq y} \left[ \mathbf{v}_{y'} + \frac{1}{K-1} \right]_+ + \frac{Kc}{K-1} \right] \text{ (for any } \mathbf{t}, \tilde{r}(x) \geq 0) \quad (132)$$

726 For the multi class hinge loss that we are considering, it is known that the minimizing functions are  
 727 given by the following [Tarigan and van de Geer, 2008]:

$$\mathbf{t}_y^*(x) = \begin{cases} 1 & \text{if } y = \arg \max_{y \in \mathcal{Y}} \eta_y(x) \\ \frac{-1}{K-1} & \text{o.w.} \end{cases} \quad (133)$$

$$\mathbf{v}_y^*(x, z) = \begin{cases} 1 & \text{if } y = \arg \max_{y \in \mathcal{Y}} \zeta_y(x, z) \\ \frac{-1}{K-1} & \text{o.w.} \end{cases} \quad (134)$$

728 which gets converted into  $f_1^*$  and  $f_2^*$  by the link function  $\varphi$  (see Eqn 10):

$$f_{hinge,1}(x) = \arg \max_{y \in \mathcal{Y}} \mathbf{t}_y^*(x) \quad \forall x \text{ s.t. } \tilde{r}(x) < 1 \quad (135)$$

$$= \arg \max_{y \in \mathcal{Y}} \eta_y(x) \quad \forall x \text{ s.t. } \tilde{r}(x) < 1 \text{ by Eqn. 133} \quad (136)$$

$$f_{hinge,1}(x) = f_1^*(x) \quad \forall x \text{ s.t. } \tilde{r}(x) < 1 \text{ by Eqn. 124} \quad (137)$$

$$f_{hinge,2}(x) = f_2^*(x) \quad \forall x \text{ s.t. } \tilde{r}(x) \geq 0 \text{ by Eqn. 52} \quad (138)$$

729 Next, we turn to the decision function  $\tilde{r}(x)$ . Using  $\tilde{r}$  as shorthands for  $\tilde{r}(x)$ :

$$\tilde{r}^* = \arg \min_{\tilde{r} \in [0,1]} \mathbb{E}_{p(z,y|x)} [\ell_{hinge}^c(x, z, y, \mathbf{t}^*, \mathbf{v}^*, \tilde{r})] \quad (139)$$

$$= \arg \min_{\tilde{r} \in [0,1]} (1 - \tilde{r}(x)) \mathbb{E}_{p(y|x)} \left[ \sum_{y' \neq y} \left[ \mathbf{t}_{y'}^* + \frac{1}{K-1} \right]_+ \right] + \tilde{r}(x) \mathbb{E}_{p(y,z|x)} \left[ \left( \sum_{y' \neq y} \left[ \mathbf{v}_{y'}^* + \frac{1}{K-1} \right]_+ + \frac{Kc}{K-1} \right) \right] \quad (140)$$

730 Since it is a linear combination of  $1 - \tilde{r}(x)$  and  $\tilde{r}(x)$ , it is clear that the minimizer  $\tilde{r}^*(x)$  will either  
 731 be at 0 or 1. We can therefore rewrite the optimization problem as the following:

$$A(x, \tilde{r}) = (1 - \tilde{r}(x)) \mathbb{E}_{p(y|x)} \left[ \sum_{y' \neq y} \left[ \mathbf{t}_{y'}^* + \frac{1}{K-1} \right]_+ \right] + \tilde{r}(x) \mathbb{E}_{p(y,z|x)} \left[ \frac{Kc}{K-1} + \left( \sum_{y' \neq y} \left[ \mathbf{v}_{y'}^* + \frac{1}{K-1} \right]_+ \right) \right] \quad (141)$$

$$= (1 - \tilde{r}(x)) \sum_{y \in \mathcal{Y}} \eta_y(x, z) \sum_{y' \neq y} \left[ \mathbf{t}_{y'}^* + \frac{1}{K-1} \right]_+ + \quad (142)$$

$$+ \tilde{r}(x) \left( \frac{Kc}{K-1} + \mathbb{E}_{p(z|x)} \left[ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \sum_{y' \neq y} \left[ \mathbf{v}_{y'}^* + \frac{1}{K-1} \right]_+ \right] \right) \quad (143)$$

$$\tilde{r}^*(x) = \arg \min_{\tilde{r} \in \{0,1\}} A(x, \tilde{r}). \quad (144)$$

732 We consider the two cases for  $A(x, \tilde{r})$ .

$$A(x, 0) = \sum_{y \in \mathcal{Y}} \eta_y(x) \sum_{y' \neq y} [\mathbf{t}_{y'}^* + \frac{1}{K-1}]_+ \quad (145)$$

$$= \sum_{y \neq \arg \max_y \eta_y(x)} \eta_y(x) \sum_{y' \neq y} [\mathbf{t}_{y'}^* + \frac{1}{K-1}]_+ \quad (146)$$

$$+ \eta_{\arg \max_y \eta_y(x)}(x) \sum_{y' \neq \arg \max_y \eta_y(x)} [\mathbf{t}_{y'}^* + \frac{1}{K-1}]_+ \quad (147)$$

$$= \sum_{y \neq \arg \max_y \eta_y(x)} \eta_y(x) \left( (K-2) \left[ \frac{-1}{K-1} + \frac{1}{K-1} \right]_+ + \left[ 1 + \frac{1}{K-1} \right]_+ \right) \quad (148)$$

$$+ \eta_{\arg \max_y \eta_y(x)}(x) \sum_{y' \neq \arg \max_y \eta_y(x)} \left[ \frac{-1}{K-1} + \frac{1}{K-1} \right]_+ \text{ by def of } \mathbf{t}^* \text{ Eqn.133} \quad (149)$$

$$A(x, 0) = \frac{K}{K-1} \left( 1 - \max_{y \in \mathcal{Y}} \eta_y(x) \right) \quad (150)$$

733 For the second case, using similar steps:

$$A(x, 1) = \frac{Kc}{K-1} + \mathbb{E}_{p(z|x)} \left[ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \sum_{y' \neq y} [\mathbf{v}_{y'}^* + \frac{1}{K-1}]_+ \right] \quad (151)$$

$$= \frac{K}{K-1} (\mathbb{E}_{p(z|x)} [1 - \max_{y \in \mathcal{Y}} \zeta_y(x)] + c) \quad (152)$$

734 This allows us to write the solution as

$$\tilde{r}^*(x) = \begin{cases} 0 & \text{if } A(r=0, x) < A(r=1, x) \\ 1 & \text{o.w.} \end{cases} \quad (153)$$

$$= \begin{cases} 0 & \text{if } \frac{K}{K-1} (1 - \max_{y \in \mathcal{Y}} \eta_y(x)) \leq \frac{K}{K-1} (\mathbb{E}_{p(z|x)} [1 - \max_{y \in \mathcal{Y}} \zeta_y(x)] + c) \\ 1 & \text{o.w.} \end{cases} \quad (154)$$

$$\tilde{r}^*(x) = \begin{cases} 0 & \text{if } \max_y \eta_y(x) \geq \mathbb{E}_{p(z|x)} [\max \zeta_y(x, z)] - c \\ 1 & \text{o.w.} \end{cases} \quad (155)$$

735 We have therefore shown that

$$\tilde{r}^*(x) = r^*(x). \quad (156)$$

736 Since we now have that  $\tilde{r}^*(x)$  is restricted to the binary values  $\tilde{r}^*(x) = \{0, 1\}$ , we can rewrite the  
737 optimal classifiers that we previously obtained:

$$f_{\text{hinge},1}^*(x) = f_1^*(x) \quad \forall x \text{ s.t. } \tilde{r}^*(x) < 1 \quad (157)$$

$$f_{\text{hinge},2}^*(x) = f_2^*(x) \quad \forall x \text{ s.t. } \tilde{r}^*(x) \geq 0 \quad (158)$$

738 as

$$f_{\text{hinge},1}^*(x) = f_1^*(x) \quad \forall x \text{ s.t. } r^*(x) = 0, \quad (159)$$

$$f_{\text{hinge},2}^*(x) = f_2^*(x) \quad \forall x \text{ s.t. } r^*(x) = 1. \quad (160)$$

$$f_{\text{hinge},1}^*(x) = f_1^*(x), \quad (161)$$

$$f_{\text{hinge},2}^*(x) = f_2^*(x). \quad (162)$$

739 Therefore, we can see that the optimal  $\mathbf{t}^*$  and  $\mathbf{v}^*$  leads to the same solution for the internal classifiers  
740 of  $f^*$ . We have therefore shown that:

$$f_{\text{hinge},1}^*(x) = f_1^*(x), \quad (163)$$

$$f_{\text{hinge},2}^*(x) = f_2^*(x), \quad (164)$$

$$\text{and } \tilde{r}^*(x) = r^*(x), \quad (165)$$

$$\implies f^* = f_{\text{hinge}}^* \quad (166)$$

741 This concludes the proof.  $\square$



#### 742 A.4.2 Gap of the hinge loss

743 Next, we aim to show that for some increasing function  $\Psi$  with  $\Psi(0) = 0$ , we can upper bound the  
 744 risk gaps of our loss of interest  $R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^*$  with the risk gap of our surrogate hinge loss  
 745  $R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^*$ .

746 **Lemma A.2.** For any distribution  $p(X, Z, Y)$ :

$$R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^* \leq \Psi\left(R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^*\right). \quad (167)$$

747 *Proof.* We start by developing the hinge risk  $R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r})$ :

$$R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) = \mathbb{E}_{p(x,z,y)}[\ell_{hinge}^c(\mathbf{t}, \mathbf{v}, \tilde{r}, x, z, y)] \quad (168)$$

$$= \mathbb{E}_{p(x,z)}\left[\sum_{y \in \mathcal{Y}} \zeta_y(x, z)(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+\right] \quad (169)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right). \quad (170)$$

748 Next we develop the term associated to the optimal hinge risk  $R_{hinge}^*$ :

$$R_{hinge}^* = \mathbb{E}_{p(x,z)}\left[\sum_{y \in \mathcal{Y}} \zeta_y(x, z)(1 - \tilde{r}^*(x)) \sum_{y' \neq y} [\mathbf{t}_{y'}^* + \frac{1}{K-1}]_+\right] \quad (171)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}^*(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'}^* + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right). \quad (172)$$

$$= \mathbb{E}_{p(x)}[\mathbb{1}[\tilde{r}^*(x) = 0]A(x, 0)] + \mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}^*(x) = 1]A(x, 1)] \text{ reusing Eqn 152, Eqn. 150} \quad (173)$$

$$R_{hinge}^* = \frac{K}{K-1} \mathbb{E}_{p(x)}[\mathbb{1}[\tilde{r}^*(x) = 0]1 - \max_{y \in \mathcal{Y}} \eta_y(x)] \quad (174)$$

$$+ \frac{K}{K-1} \mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}^*(x) = 1](1 - \max_{y \in \mathcal{Y}} \zeta_y(x) + c)] \quad (175)$$

749 Bringing both  $R_{hinge}^*$  and  $R_{hinge}(f)$  to evaluate the gap  $G \triangleq R_{hinge}(f) - R_{hinge}^*$ , we can decompose  
 750 the gap by a sum of 4 terms that are driven by the ground truth decision cases, i.e.  $r^* = 0$  or  $r^* = 1$   
 751 and the decision of the model  $\tilde{r}(x)$ :

$$G \triangleq R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^* \quad (176)$$

$$G = \mathbb{E}_{p(x,z)}\left[\sum_{y \in \mathcal{Y}} \zeta_y(x, z)(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ + \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right)\right] \quad (177)$$

$$- \frac{K}{K-1} \mathbb{E}_{p(x)}[\mathbb{1}[\tilde{r}^*(x) = 0](1 - \max_{y \in \mathcal{Y}} \eta_y(x))] + \mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}^*(x) = 1](c + 1 - \max_{y \in \mathcal{Y}} \zeta_y(x, z))] \quad (178)$$

752 We can define the corresponding gap to each case as follows:

$$G_1 \triangleq \mathbb{1}[\tilde{r}^*(x) = 0] \mathbb{1}[\tilde{r}(x) \leq 0.5] \sum_{y \in \mathcal{Y}} \eta_y(x, z)(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ \quad (179)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right) - \frac{K}{K-1} (1 - \max_{y \in \mathcal{Y}} \eta_y(x)) \quad (180)$$

$$G_2 \triangleq \mathbb{1}[\tilde{r}^*(x) = 0] \mathbb{1}[\tilde{r}(x) \geq 0.5] \sum_{y \in \mathcal{Y}} \eta_y(x, z)(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ \quad (181)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right) - \frac{K}{K-1} (1 - \max_{y \in \mathcal{Y}} \eta_y(x)) \quad (182)$$

$$G_3 \triangleq \mathbb{1}[\tilde{r}^*(x) = 1] \mathbb{1}[\tilde{r}(x) \leq 0.5] \sum_{y \in \mathcal{Y}} \eta_y(x, z)(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ \quad (183)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right) - \frac{K}{K-1} (c + 1 - \max_{y \in \mathcal{Y}} \zeta_y(x, z)), \quad (184)$$

$$G_4 \triangleq \mathbb{1}[\tilde{r}^*(x) = 1] \mathbb{1}[\tilde{r}(x) \geq 0.5] \sum_{y \in \mathcal{Y}} \eta_y(x, z)(1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ \quad (185)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right) - \frac{K}{K-1} (c + 1 - \max_{y \in \mathcal{Y}} \zeta_y(x, z)), \quad (186)$$

753 and rewrite the total gap as:

$$G = \mathbb{E}_{p(x, z)} [G_1 + G_2 + G_3 + G_4] \quad (187)$$

754 We can obtain a similar decomposition for the risk gap of our target risk  $R_{01c}(f)$ . We recall the  
755 definition:

$$R_{01c}(f) = \mathbb{E}_{p(x, z, y)} [\ell_{01c}(f(x, z), y)] \quad (188)$$

$$= \mathbb{E}_{p(x, z, y)} [\mathbb{1}[r(x) = 0] \mathbb{1}[f_1(x) \neq y] + \mathbb{1}[r(x) = 1] [\mathbb{1}[f_2(x, z) \neq y]] + c] \quad (189)$$

$$= \mathbb{E}_{p(x)} [\mathbb{1}[r(x) = 0] \mathbb{E}_{p(y|x)} [\mathbb{1}[f_1(x) \neq y]] + \mathbb{1}[r(x) = 1] \mathbb{E}_{p(y, z|x)} [\mathbb{1}[f_2(x, z) \neq y]] + c] \quad (190)$$

$$R_{01c}(f) = \mathbb{E}_{p(x)} [\mathbb{1}[r(x) = 0] (1 - \eta_{f_1}(x)) + \mathbb{1}[r(x) = 1] \mathbb{E}_{p(z|x)} [1 - \zeta_{f_2}(x, z)] + c] \quad (191)$$

756 and decompose the gap risk with terms based on similar cases:

$$F_1 \triangleq \mathbb{1}[r(x) = 0, r^*(x) = 0] (\eta_{f^{*1}}(x) - \eta_{f_1}(x)) \quad (192)$$

$$F_2 \triangleq \mathbb{1}[r(x) = 0, r^*(x) = 1] (\mathbb{E}_{p(z|x)} [\zeta_{f_2^*}(x, z)] - \eta_{f_1}(x) - c) \quad (193)$$

$$F_3 \triangleq \mathbb{1}[r(x) = 1, r^*(x) = 0] (\eta_{f^{*1}}(x) - \mathbb{E}_{p(z|x)} [\zeta_{f_2}(x, z)] + c) \quad (194)$$

$$F_4 \triangleq \mathbb{1}[r(x) = 1, r^*(x) = 1] (\mathbb{E}_{p(z|x)} [\zeta_{f_2^*}(x, z)] - \zeta_{f_2}(x, z)) \quad (195)$$

$$R_{01c}(f) - R_{01c}^* = \mathbb{E}_{p(x, z)} [F_1 + F_2 + F_3 + F_4] \quad (196)$$

757 This makes sense. If the optimal decision is to delay and  $f$  early-exit, the risk is diminished by the  
758 saved computation  $c$  (the second case). If the optimal decision is to early exit and  $f$  delays, the risk is  
759 increased by the computation cost  $c$  (the third case).

760 To prove the result, we show that :

$$\mathbb{E}_{p(x,z)}[F_1] \leq \Psi(\mathbb{E}_{p(x,z)}[G_1]) \quad (197)$$

$$\mathbb{E}_{p(x,z)}[F_2] \leq \Psi(\mathbb{E}_{p(x,z)}[G_2]) \quad (198)$$

$$\mathbb{E}_{p(x,z)}[F_3] \leq \Psi(\mathbb{E}_{p(x,z)}[G_3]) \quad (199)$$

$$\mathbb{E}_{p(x,z)}[F_4] \leq \Psi(\mathbb{E}_{p(x,z)}[G_4]) \quad (200)$$

761  **$G_1$  inequality** Starting with  $G_1$  and  $F_1$ :

$$G_1 = \mathbb{1}[\tilde{r}^*(x) = 0] \mathbb{1}[\tilde{r}(x) \leq 0.5] \sum_{y \in \mathcal{Y}} \eta_y(x, z) (1 - \tilde{r}(x)) \sum_{y' \neq y} [\mathbf{t}_{y'} + \frac{1}{K-1}]_+ \quad (201)$$

$$+ \sum_{y \in \mathcal{Y}} \zeta_y(x, z) \tilde{r}(x) \left( \sum_{y' \neq y} [\mathbf{v}_{y'} + \frac{1}{K-1}]_+ + \frac{Kc}{K-1} \right) - \frac{K}{K-1} (1 - \max_{y \in \mathcal{Y}} \eta_y(x)) \quad (202)$$

$$\geq \frac{K}{K-1} (1 - \tilde{r}(x)) (1 - \eta_{f1}) + \tilde{r}(x) (1 - \zeta_{f2} + c) - (1 - \eta_*) \text{ using optimal sol of the hinge loss} \quad (203)$$

$$= \frac{K}{K-1} (1 - \eta_{f1} - \tilde{r}(x) + \tilde{r}(x) \eta_{f1} + \tilde{r}(x) - \tilde{r}(x) \zeta_{f2} + \tilde{r}(x) c - 1 + \eta_*) \quad (204)$$

$$G_1 \geq \mathbb{1}[\tilde{r}(x) \leq 0.5] \frac{K}{K-1} (\eta_* - \eta_{f1} + \tilde{r}(x) (\eta_{f1} - \zeta_{f2} + c)) \quad (205)$$

762 We need to find a mapping  $\Psi$  such that the following holds:

$$\mathbb{E}_{p(x,z)}[F_1] \leq \Psi(\mathbb{E}_{p(x,z)}[G_1]) \quad (206)$$

$$\mathbb{E}_{p(x,z)}[(\eta_{f^{*1}} - \eta_{f1})] \leq \Psi(\mathbb{E}_{p(x,z)}[G_1]) \quad (207)$$

763 Using the simple scaling function  $\Psi(x) = \frac{2(K-1)}{K}x$ , we can see that the previous inequality holds:

$$\mathbb{E}_{p(x,z)}[\frac{2(K-1)G_1}{K}] \geq 2\mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}(x) \leq 0.5](\eta_* - \eta_{f1} + \tilde{r}(x)(\eta_{f1} - \zeta_{f2} + c))] \quad (208)$$

$$\geq 2\mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}(x) \leq 0.5](\eta_* - \eta_{f1} + \tilde{r}(x)(\eta_{f1} - \zeta_* + c))] \quad (209)$$

$$\geq 2\mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}(x) \leq 0.5](\eta_* - \eta_{f1} + \tilde{r}(x)(\eta_{f1} - (c + \eta_*) + c))] \text{ as } r^* = 0 \quad (210)$$

$$\geq 2\mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}(x) \leq 0.5](\eta_* - \eta_{f1} - \tilde{r}(x)(\eta_* - \eta_{f1}))] \quad (211)$$

$$\geq 2\mathbb{E}_{p(x,z)}[\mathbb{1}[\tilde{r}(x) \leq 0.5](1 - \tilde{r}(x))(\eta_* - \eta_{f1})] \quad (212)$$

$$\geq 2\mathbb{E}_{p(x,z)}[0.5(\eta_* - \eta_{f1})] \quad (213)$$

$$\Psi(\mathbb{E}_{p(x,z)}[G_1]) \geq \mathbb{E}_{p(x,z)}[(\eta_* - \eta_{f1})] \text{ with } \Psi(x) = \frac{2(K-1)}{K}x \quad \square \quad (214)$$

764  **$G_2$  inequality** For the next inequality with  $G_2$  and  $F_2$ , again using the same function  $\Psi(x) =$   
 765  $\frac{2(K-1)}{K}x$ , the inequality holds:

$$\mathbb{E}_{p(x,z)}\left[\frac{2(K-1)G_2}{K}\right] \geq 2\mathbb{E}_{p(x,z)}\left[\mathbb{1}[\tilde{r}(x) \geq 0.5](\eta_* + \eta_{f1}(\tilde{r}(x) - 1) - \tilde{r}(x)\zeta_{f2} + \tilde{r}(x)c)\right] \quad (215)$$

$$\geq 2\mathbb{E}_{p(x,z)}\left[\mathbb{1}[\tilde{r}(x) \geq 0.5](\eta_* + \eta_*(\tilde{r}(x) - 1) - \tilde{r}(x)\zeta_{f2} + \tilde{r}(x)c)\right] \quad (216)$$

$$= 2\mathbb{E}_{p(x,z)}\left[\mathbb{1}[\tilde{r}(x) \geq 0.5](\tilde{r}(x)(\eta_* - \zeta_{f2} + c))\right] \quad (217)$$

$$\geq 2\mathbb{E}_{p(x,z)}\left[0.5(\eta_* - \zeta_{f2} + c)\right] \quad (218)$$

$$\Psi(\mathbb{E}_{p(x,z)}[G_2]) \geq \mathbb{E}_{p(x,z)}[F_2] \quad \square \quad (219)$$

$$(220)$$

766 Following similar steps, we obtain

$$\Psi(\mathbb{E}_{p(x,z)}[G_3]) \geq \mathbb{E}_{p(x,z)}[F_3] \quad (221)$$

$$\Psi(\mathbb{E}_{p(x,z)}[G_4]) \geq \mathbb{E}_{p(x,z)}[F_4] \quad (222)$$

767 Putting all results together, we obtain

$$R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^* \leq \frac{2(K-1)}{K} \left( R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^* \right) \quad (223)$$

$$R_{01c}(\varphi(\mathbf{t}, \mathbf{v}, \tilde{r})) - R_{01c}^* \leq \Psi \left( R_{hinge}(\mathbf{t}, \mathbf{v}, \tilde{r}) - R_{hinge}^* \right) \quad \square \quad (224)$$

768 with  $\Psi(0) = 0$  and is increasing. This concludes the proof.

769

$\square$

## 770 A.5 Proof of the failure of the cross entropy version

771 In this section, we prove that the cross entropy version of the surrogate loss we presented cannot be  
 772 Bayes-consistent.

773 We recall the entropy version, with  $\mathbf{p}^1 \in \Delta^K$  for  $f_1$  and  $\mathbf{p}^2 \in \Delta^K$  for  $f_2$ , with the same link function  
 774  $\varphi$ . The cross entropy version of the loss that we consider is given by:

$$\ell_{ce}^{g(c)}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}, x, z, y) = -((1 - \tilde{r}(x)) \log(\mathbf{p}_y^1) + \tilde{r}(x)(\log(\mathbf{p}_y^2) + g(c))) , \quad (225)$$

775 where  $g(c)$  is an arbitrary function. We again consider its associated risk:

$$R_{ce}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}) = \mathbb{E}_{p(x,z,y)}[\ell_{ce}^{g(c)}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}, x, z, y)] \quad (226)$$

776 and minimizing function:

$$f_{ce}^* = \arg \min_{\mathbf{p}^1, \mathbf{p}^2, \tilde{r} \in [0,1]} R_{ce}(\mathbf{p}^1, \mathbf{p}^2, \tilde{r}). \quad (227)$$

777 **Lemma 4.2.** Cross-entropy surrogate loss is not Bayes Consistent.

778 There is no  $g(c)$  for which:

$$f^* = f_{ce}^*. \quad (228)$$

779 *Proof.* Following a similar reasoning as the proof of Lemma , we can readily find that the optimal  
 780 predicted probability vectors  $\mathbf{p}^1$  and  $\mathbf{p}^2$  in  $f_{ce}^*$  should match the posteriors;

$$f_{ce}^* = \{\eta(x), \zeta(x, z), \tilde{r}_{ce}^*\} \quad (229)$$

781 Now, to find the optimal decision function of the cross entropy risk  $\tilde{r}_{ce}^*$ , we can again obtain the  
 782 solution as

$$\tilde{r}_{ce}^*(x) = \arg \min_{\tilde{r} \in \{0,1\}} B(\tilde{r}, x) \text{ where} \quad (230)$$

$$B(\tilde{r}, x) = \mathbb{E}_{p(x,z)}[-(1 - \tilde{r}(x)) \log(\eta_*(x)) - \tilde{r}(x)(\log(\zeta_*(x, z)) + g(c))] \quad (231)$$

by following the same steps that were taking to obtain Eqn. 144. We again consider the two cases:

$$B(0, x) = -\log(\eta_*(x)), \quad (232)$$

$$B(1, x) = -\mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] - g(c). \quad (233)$$

The solution for the cross-entropy decision function is hence given by:

$$\tilde{r}_{ce}^*(x) = \begin{cases} 0 & \text{if } B(0, x) < B(1, x), \\ 1 & \text{o.w.} \end{cases} \quad (234)$$

$$\tilde{r}_{ce}^*(x) = \begin{cases} 0 & \text{if } \log(\eta_*(x)) \geq \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] + g(c) \\ 1 & \text{o.w.} \end{cases} \quad (235)$$

If we recall the solution decision of our target problem;

$$r^*(x) = \eta_*(x) < \mathbb{E}_{p(z|x)}[\zeta_*(x, z)] + c, \quad (236)$$

we are searching for a function  $g(c)$  for which

$$\mathbb{1}[\log(\eta_*(x)) < \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] + g(c)] = r^*(x) \quad \forall x, \eta, \zeta. \quad (237)$$

If we define the decision sets of  $x$ :

$$\mathcal{D}_1 := \{x \mid r^*(x) = 1\}, \quad (238)$$

$$\mathcal{D}_0 := \{x \mid r^*(x) = 0\}, \quad (239)$$

we can rewrite the condition Eqn. 237 as:

$$\log(\eta_*(x)) - \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] < g(c) \quad \text{for all } x \in \mathcal{D}_1, \quad (240)$$

$$\log(\eta_*(x)) - \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] \geq g(c) \quad \text{for all } x \in \mathcal{D}_0. \quad (241)$$

This implies that the function  $g(c)$  should satisfy:

$$\sup_{x \in \mathcal{D}_1} \log(\eta_*(x)) - \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] < g(c) \leq \inf_{x \in \mathcal{D}_0} \log(\eta_*(x)) - \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] \quad \forall \eta, \zeta. \quad (242)$$

However, our condition for  $r^*(x)$  is on the absolute scale, not on the log scale  $r^*(x) = \eta_*(x) < \mathbb{E}_{p(z|x)}[\zeta_*(x, z)] + c$ . We will therefore have that for some  $\eta, \zeta$

$$\sup_{x \in \mathcal{D}_1} \log(\eta_*(x)) - \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] > \inf_{x \in \mathcal{D}_0} \log(\eta_*(x)) - \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))]. \quad (243)$$

This implies that there is no  $g(c)$  that can satisfy

$$\mathbb{1}[\log(\eta_*(x)) < \mathbb{E}_{p(z|x)}[\log(\zeta_*(x, z))] + g(c)] = r^*(x). \quad (244)$$

This concludes the proof.  $\square$