On Regularization and Inference with Label Constraints

Kaifu Wang [1] **Hangfeng He** [2] **Tin D. Nguyen** [3] **Piyush Kumar** [4] **Dan Roth** [1]

Abstract

Prior knowledge and symbolic rules in machine learning are often expressed in the form of label constraints, especially in structured prediction problems. In this work, we compare two common strategies for encoding label constraints in a machine learning pipeline, *regularization with constraints* and *constrained inference*, by quantifying their impact on model performance. For regularization, we show that it narrows the generalization gap by precluding models that are inconsistent with the constraints. However, its preference for small violations introduces a bias toward a suboptimal model. For constrained inference, we show that it reduces the population risk by correcting a model's violation, and hence turns the violation into an advantage. Given these differences, we further explore the use of two approaches together and propose conditions for constrained inference to compensate for the bias introduced by regularization, aiming to improve both the model complexity and optimal risk.

1. Introduction

Domain knowledge in machine learning is often framed as constraints on the output label space. Such label constraints have been widely identified in natural language processing tasks (Roth & Yih, 2004; Martins et al., 2009b; Ning et al., 2018a; Lu et al., 2022) and studied in the context of structured prediction (Punyakanok et al., 2005; Chang et al., 2007; 2008; Ganchev et al., 2010; Chang et al., 2012; Pan et al., 2020). For example, in temporal reasoning (Ning et al., 2018a) where the model is asked to label the relations ("before" or "after") among a set of events, the assigned labels will need to satisfy a transitivity constraint which means, for example, the facts that an event E_1 is after E_2 and that E_2 is after E_3 imply that E_1 is after E_3. The central question is how to encode such a constraint into a learning algorithm to ensure better performance and generalization of the learned model.

Practitioners have developed two techniques to encode a label constraint in a machine learning pipeline. The first, called *regularization with constraints*, penalizes a model for its violation of the constraint in addition to the classification loss (Ganchev et al., 2010; Li et al., 2019). The second, called *inference with constraints*, modifies prediction rules directly by enforcing strictly constrained inference (Roth & Yih, 2004; Scholak et al., 2021) or balancing the original model's output with the constraint in a soft way (Chang et al., 2008; 2012).

Although these two learning algorithms have been shown to be empirically successful, we are not aware of theoretical analyses that elucidate each algorithm's advantages or disadvantages in comparison with the other one. Natural questions include, how do these two differ in their impact on the learned model? Moreover, in practice, the constraints could be noisy i.e. (Hu et al., 2016; Wang et al., 2021). In such cases, do they still improve the model performance? If so, by how much?

Focusing on multiclass classification with label constraints, we compare regularization with constraints and constrained inference. For each algorithm, we quantify its optimal risk (aka approximation error) and its generalization gap (aka estimation error). Specifically, in Section 3, we show that regularization with constraints achieves a smaller generalization error by reducing the model complexity but will introduce a bias towards a suboptimal model if the risk minimizer and violation minimizer does not coincide. In Section 4, we study a broad family of constrained inference model called Constrained Conditional Model (CCM) (Chang et al., 2008; 2012) and point out that the constrained inference could reduce the risk of a model if and only if the model violates the constraint more than the true data distribution. This further suggests finding models with higher violation, which contrasts the learning objective used in regularization that discourages violation. Given these contrasts, we further

[1]University of Pennsylvania, Philadelphia, PA, USA [2]University of Rochester, Rochester, NY, USA (Part of the work done while at the University of Pennsylvania.) [3]Massachusetts Institute of Technology, Cambridge, MA, USA [4]Systems and Technology Research, Woburn, MA USA. Correspondence to: Piyush Kumar <piyush.kumar@str.us>, Dan Roth <danroth@seas.upenn.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

study the combination and interaction of the two methods in Section 5 and describe how constrained inference could compensate for the bias introduced by regularization.

To the best of our knowledge, our analysis is the first to provide a theoretical view on comparing the two approaches. We believe in the importance of this comparison and hope to bring this problem to the attention of the machine learning community. In summary, our contributions include:

1. We provide an error bound (Theorem 3.6) that describes the tradeoff between the generalization gap and the optimal risk when performing regularization with constraints.
2. We propose a sufficient and necessary condition (Theorem 4.3) for constrained inference to improve a model by quantifying its reduction in risk. Based on this, we further argue that constrained inference, when used at training time, implicitly modifies the training objective in an opposite direction as in the regularization approach (Proposition 4.6).
3. We study the combination of regularization and constrained inference, and propose sufficient (Theorem 5.1) as well as necessary (Theorem 5.2) conditions for the combined algorithm to achieve improvement in both optimal risk and model complexity.

Proofs of all the theoretical results are in the appendix.

2. Preliminaries

Our goal is to learn a mapping from the instance space $\mathcal{X}$ to the output space $\mathcal{Y}$. The learner has access to a set of labeled training data S_{L} of size m_{L}, which contains i.i.d. samples of a distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. The marginal distribution of X is denoted as $\mathbb{P}_X$. In this work, we assume the ground truth label associated with $x \in \mathcal{X}$ is generated by a deterministic mapping $y_{\text{ora}} : \mathcal{X} \to \mathcal{Y}$ ($_{\text{ora}}$ is short for oracle). We also denote the true label as y_{ora} when the context is clear.

Model. The scoring class $\mathcal{F}$ contains scoring functions $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We will also call a $f \in \mathcal{F}$ a classifier. Let $\Delta_{\mathcal{Y}}$ be the $|\mathcal{Y}|$-dimensional probability simplex. Each scoring function induces a probabilistic prediction $\mathbb{P}_f(\cdot|x) \in \Delta_{\mathcal{Y}}$ by performing *softmax inference* as $\mathbb{P}(y|x) \propto \exp(f(x, y))$.

Loss Function. The prediction of f at x is evaluated by the *classification error* (or ℓ^1 loss) $L(x, y_{\text{ora}}, f) := 1 - \mathbb{P}_f(y_{\text{ora}}|x)$, which is half the ℓ^1 distance the between the one-hot distribution $e_{y_{\text{ora}}}$ and $\mathbb{P}_f$ on $\Delta_{\mathcal{Y}}$. It can also be viewed as a smoothed version of the standard zero-one loss in the sense that $\lim_{t\to\infty} L(x, y_{\text{ora}}, tf) = \mathbb{1}\{\text{argmax}_{y\in\mathcal{Y}} f(x, y) \neq y_{\text{ora}}\}$. More background on the definition of the ℓ^1 loss are provided in Appendix A. A scoring function f is evaluated by its *risk* $R(f) := \mathbb{E}[L(x, y_{\text{ora}}, f)]$. The empirical estimate of the risk using the labeled examples in S_{L} is denoted as $\widehat{R}(f, S_{\mathsf{L}})$. We also consider the cross-entropy surrogate loss defined as $L_{\text{CE}}(x, y_{\text{ora}}, f) := -\log \mathbb{P}_f(y_{\text{ora}}|x)$ and refer its expectation $R_{\text{CE}}(f) = \mathbb{E}[L_{\text{CE}}(x, y_{\text{ora}}, f)]$ as cross-entropy risk.

Label constraint. A *label constraint* (or *constraint* for short) is a deterministic mapping $C : \mathcal{X} \to 2^{\mathcal{Y}} - \{\varnothing\}$. Namely, C maps an instance x to a nonempty subset of $\mathcal{Y}$, which may or may not contain the true label $y_{\text{ora}}(x)$. In particular, we say a constraint C is *noise-free* if $\mathbb{P}(y_{\text{ora}} \in C(x)) = 1$. Otherwise, C is said to be a *noisy* constraint and its noise rate is denoted as $V_{\text{ora}} := \mathbb{P}(y_{\text{ora}}(x) \notin C(x))$.

Violation. A constraint C is equipped a *violation function*, which is an indicator function $v_C(x, y) = \mathbb{1}\{y \notin C(x)\}$. We also overload the notation v and define the violation of a classifier f at an instance x as $v_C(x, f) := 1 - \mathbb{P}_f(C(x)|x) = \sum_{y\notin C(x)} \mathbb{P}_f(y|x)$. Its expectation is $V_C(f) := \mathbb{E}[v_C(x, f)]$. We elide the subscript C and write them as $v(x, y), v(x, f)$ and $V(f)$ when the context is clear. Similar to the classification error, we consider a cross-entropy surrogate of the violation function defined as $v_{\text{CE}}(x, f) := -\log \mathbb{P}_f(C(x))$ and its expectation $V_{\text{CE}}(f) = \mathbb{E}[v_{\text{CE}}(x, f)]$.

Rademacher complexity. We use the following version of Rademacher complexity that is adopted from Cortes et al. (2016) to characterize the generalization ability of the scoring space of multiclass classifiers $\mathcal{F}$:

Definition 2.1 (Rademacher complexity of $\mathcal{F}$)**.** *The* empirical Rademacher complexity *of scoring class $\mathcal{F}$ with respect to a set $S = \{x_i\}_{i=1}^m$ that contains m samples of the instance is defined as*

$$\widehat{\mathfrak{R}}_m(\mathcal{F}; S) := \frac{1}{m}\mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \sum_{y\in\mathcal{Y}} \epsilon_{i,y} f(x_i, y)\right] \tag{1}$$

where $\epsilon = (\epsilon_{i,y})_{i\in[m],y\in\mathcal{Y}}$ are independent Rademacher random variables, each of which is uniformly distributed over $\{-1, +1\}$. The Rademacher complexity *of scoring class $\mathcal{F}$ is the expectation of the empirical version:*

$$\mathfrak{R}_m(\mathcal{F}) := \mathbb{E}_{S\sim\mathbb{P}_X^m}[\widehat{\mathfrak{R}}_m(\mathcal{F}; S)] \tag{2}$$

This definition of Rademacher complexity is a special case of the *factor graph complexity* proposed by Cortes et al. (2016), which is defined for more general structured prediction models. It is hence possible to extend our results of the generalization bounds to structured models by replacing the Rademacher complexity with factor graph complexity. In this work, we focus on multiclass classifiers for the simplicity of presentation.

3. Regularization with Constraints

In a standard machine learning algorithm, the learner receives a set of labeled data $S_{\mathsf{L}} \in \cup_{m=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^m$ and finds the *empirical risk minimizer*, which is defined as $\operatorname{argmin}_{f\in\mathcal{F}} \widehat{R}(f; S_{\mathsf{L}})$. In this section, we consider a method that modifies this learning objective by adding a regularization term defined with the constraint C. Precisely, we consider minimizing an augmented objective defined as $\mathcal{L}_\rho(f) := R(f) + \rho V(f)$ where $\rho \geq 0$ is a fixed tradeoff parameter.

The idea of regularizing the model by adding a penalty for the violation of the constraints on an unlabeled dataset is widely adopted in the literature. In particular, the cross entropy violation is known as the *semantic loss* (Xu et al., 2018) in the context of logical constraints. Other designs of the regularization term include using the KL-divergence on the probability space in the *posterior regularization* algorithm (Ganchev et al., 2010) and using the t-norms from fuzzy logic (Li et al., 2019).

We will show this algorithm improves the generalization error by reducing the complexity of the scoring space (Theorem 3.6), but in general leads to a larger classification risk in the long run (Proposition 3.2), thus resulting in a tradeoff between estimation and approximation errors.

3.1. Semi-supervised Regularization with Constraints

We consider a semi-supervised approach where the learner has access to an unlabeled dataset S_{U} that contains m_{U} independent samples of the instance X, resulting in the following definition.

Definition 3.1 (ERVM)**.** *Given a labeled dataset S_{L} of size m_{L} and an unlabeled dataset S_{U} of size m_{U}, a scoring space $\mathcal{F}$ and a tradeoff parameter $\rho \geq 0$, we define and denote the* empirical risk and violation minimizer *(ERVM) as:*

$$\widehat{f}_\rho(S_{\mathsf{L}}, S_{\mathsf{U}}) := \operatorname*{argmin}_{f\in\mathcal{F}} \left(\frac{1}{m_{\mathsf{L}}} \sum_{(x,y)\in S_{\mathsf{L}}} L(x, y, f) + \frac{\rho}{m_{\mathsf{U}}} \sum_{x\in S_{\mathsf{U}}} v_C(x, f) \right). \tag{3}$$

We also denote the expected version as:

$$f_\rho := \operatorname*{argmin}_{f\in\mathcal{F}} R(f) + \rho V_C(f). \tag{4}$$

For example, with our notation, $\widehat{f}_0$ is the ERM and f_∞ is the minimizer of the expected violation function. Notice that the minimizer in general is non-unique. Therefore, when we state any proposition that is related to f_ρ or $\widehat{f}_\rho$, we mean the proposition will hold for any of the minimizers.

3.2. Deviation from The Optimal Risk

In this section, we study how the risk of the minimizer f_ρ will deviate from the optimal risk in $\mathcal{F}$. The reason that we are interested in bounding $R(f_\rho)$ is that in general the minimizer $R(f_\rho)$ is non-unique and may have different values of risks. Therefore, to describe the risk of ERVM in the long run (in Theorem 3.4), we provide an upper bound for all the possible risks of f_ρ.

Proposition 3.2 (Deviation from the optimal risk)**.** *For any constraint C and $\rho \geq 0$, the following holds.*

$$R(f_0) \leq R(f_\rho) \leq R(f_0) + \rho(V(f_0) - V(f_\infty)). \tag{5}$$

The same relation also holds for the empirical estimates $\widehat{R}$ and $\widehat{V}$. Moreover, for any $\rho > 0$, there exists a scoring space and data distribution so that the RHS can be reached even with a noise-free constraint C.

This result shows the minimizer of the regularized objective in general has a suboptimal risk over $\mathcal{F}$. On the other hand, if the risk minimizer is simultaneously a violation minimizer, i.e., $V(f_0) = V(f_\infty)$, this relation implies consistency, i.e., $R(f_\rho) = R(f_0)$. This quantity $V(f_0)$ can be small when the noise rate V_{ora} is small and the model is expressive enough (e.g., a deep neural net) to approximate the true model.

3.3. Generalization Bounds

Now we discuss how regularization could reduce the complexity of the hypothesis class. The first step is to show that the violation of the target hypothesis is not too large. In particular, the following bound is a direct consequence of minimizing the regularized objective:

Lemma 3.3 (Regularization implies small violation)**.** *Let f_ρ be the regularized learning objective defined as in (4). If the minimum violation in $\mathcal{F}$ is upper bounded by a known constant $u \geq 0$, i.e., $V(f_\infty) \leq u$, then $V(f_\rho) \leq 1/\rho + u$.*

The upper bound u can be set to arbitrarily small by adding a baseline model defined as $f_t(x, y) = t \cdot \mathbb{1}\{y \in C(x)\}$ and driving t to infinite. This construction is possible due to the fact that the mapping C is known to the learner. The benefits of knowing C will be further explored in Section 4 when we discuss inference with constraints.

For any $B \geq 0$, we let $\mathcal{F}_B := \{f \in \mathcal{F} | V(f) \leq B\}$ be the set of classifiers with small violation. From the above discussion, we know that the target hypothesis f_ρ will lie in a smaller space $\mathcal{F}_{u+1/\rho}$, which is characterized by the violation function and hence can be identified only with unlabeled data. To this end, we describe how the violation as well as the risk can be estimated with data.

Lemma 3.4 (Generalization gap of ℓ^1 loss and violation)**.** *Given a labeled dataset S_{L} of size m_{L}, for any $\delta > 0$, with*

probabilistic at least $1-\delta$, the following inequality holds uniformly for $f \in \mathcal{F}$:

$$R(f) \leq \widehat{R}(f; S_{\mathsf{L}}) + \Re_{m_{\mathsf{L}}}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{L}}}} \tag{6}$$

Given a unlabeled dataset S_{U} of size m_{U}, for any $\delta > 0$, with probabilistic at least $1-\delta$, the following inequality holds uniformly for $f \in \mathcal{F}$:

$$V(f) \leq \widehat{V}(f; S_{\mathsf{U}}) + \Re_{m_{\mathsf{U}}}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}} \tag{7}$$

The proof of this result relies on a contraction lemma established in Cortes et al. (2016), which was used to analyze the argmax inference with margin losses. Our analysis extends their results to softmax inference, which may be of independent interest.

Furthermore, if the size of the constrained set $C(x)$ is a constant, namely $|C(x)| = c_0 < c = |\mathcal{Y}|$ for all $x \in \mathcal{X}$, then the Rademacher complexity term of equation (7) can be improved to $\frac{\sqrt{2}}{2}\sqrt{\frac{1}{c-c_0} + \frac{1}{c_0}}\Re_{m_{\mathsf{U}}}(\mathcal{F})$ (see the discussion in the proof). This term is symmetric with the transformation $c_0 \mapsto c - c_0$, due to the fact that estimating the violation V_C of a constraint C is equivalent to estimating $V_{\mathcal{Y}-C}$. In particular, when $c_0 < c/2$, if the constraint is more restrictive and informative (so that c_0 is small), it can be more difficult to estimate the violation.

Remark 3.5 (Relation to cross-entropy loss)**.** *Assuming $\lim_{m\to\infty} \Re_m(\mathcal{F}) = 0$, this result implies $\mathcal{L}_\rho$ can be approximated by its empirical version $\widehat{\mathcal{L}}_\rho$ with sufficient amount of data. On the other hand, since $\widehat{\mathcal{L}}_\rho$ is upper bounded by its cross-entropy surrogate $\widehat{R}_{\mathrm{CE}} + \rho\widehat{V}_{\mathrm{CE}}$, we further have that*

$$\mathcal{L}_\rho(f) \leq \widehat{R}_{\mathrm{CE}}(f, S_{\mathsf{L}}) + \rho\widehat{V}_{\mathrm{CE}}(f, S_{\mathsf{U}}) + o_{m_{\mathsf{L}},m_{\mathsf{U}}}(1) \tag{8}$$

where $o_{m_{\mathsf{L}},m_{\mathsf{U}}}(1)$ converges to 0 as $m_{\mathsf{L}}, m_{\mathsf{U}} \to \infty$. Therefore, in practice one can minimize this upper bound by solving the convex surrogate problem

$$\min_{f\in\mathcal{F}} \widehat{R}_{\mathrm{CE}}(f, S_{\mathsf{L}}) + \rho\widehat{V}_{\mathrm{CE}}(f, S_{\mathsf{U}}). \tag{9}$$

where $\widehat{R}_{\mathrm{CE}}(f, S_{\mathsf{L}})$ and $V_{\mathrm{CE}}(f, S_{\mathsf{U}})$ are the empirical average of the cross-entropy loss and violation.

Finally, using these results, we bound the risk of the classifier learned by ERVM. For simplicity, we will denote the generalization gap $B(\delta, m, \mathcal{F}) := \Re_m(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2m}}$.

Theorem 3.6 (Error bound)**.** *We have with probability at least $1-6\delta$ that*

$$\begin{aligned} R(\widehat{f}_\rho) \leq\ & R(f_0) + \rho V(f_0) - \rho V(f_\infty) \\ & + \Re_{m_{\mathsf{L}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) \\ & + \rho\Re_{m_{\mathsf{U}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) \\ & + 2\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{L}}}} + 2\rho\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{U}}}} \end{aligned} \tag{10}$$

where $\Re(\cdot)$ is the Rademacher complexity defined in (2).

Proof sktech. First, we show $\widehat{f}_\rho$ and f_ρ both lie in the subspace $\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}$ with high probability since the violation can be well-approximated, according to Lemma 3.4. Then, the gap between the objective $\mathcal{L}(f_\rho)$ and $\mathcal{L}(\widehat{f}_\rho)$ is controlled by the Rademacher complexity of $\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}$. Finally, we use the inequalities established in Lemma 3.2 to further upper bound the term $\mathcal{L}(f_\rho)$ using the risk and violation of f_0.

Using the same proof technique, this result can be extended to other choices of loss function as long as: (a) The loss is bounded so that the optimal regularized model has a small violation, as in Lemma 3.3. (b) The loss is Lipschitz with the model scores so that a generalization bound associated with the loss holds, as in Lemma 3.4. □

Reducing the generalization gap. The bound (10) contains three parts: the first line is the worst risk that can be achieved by f_ρ as we described in Proposition 3.2, the second and the third line is the complexity of the classifiers that have a small violation, and the last line is the errors that are independent of the model. This bound (10) is most preferable when a large set of unlabeled data is available so that the approximation errors of violations (i.e., term $B(\delta/2, m_{\mathsf{U}}, \mathcal{F})$, $\Re_{m_{\mathsf{U}}}(\mathcal{F}_{1/\rho+u+B(\delta/2,m_{\mathsf{U}},\mathcal{F})})$ and $\sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}}$) are all small. Then, the model complexity is mainly described by the term $\Re_{m_{\mathsf{L}}}(\mathcal{F}_{1/\rho+u})$, which is the Rademacher complexity of a proper subset of $\mathcal{F}$. In this sense, the regularization method reduces the generalization gap by reducing the model complexity of the scoring space.

Tradeoff in regularization. In situations where m_{U} is large, the tradeoff parameter ρ balances two quantities: a larger ρ leads to a smaller scoring space $\mathcal{F}_{1/\rho+u}$, but brings more bias depending on the suboptimality of f_0 in violation, measured by $V(f_0) - V(f_\infty)$. The benefit of regularization is greater if fewer classifiers can achieve a violation that is close to the optimal value $V(f_\infty)$.

We provide the following example to illustrate how the Rademacher complexity can be reduced in linear models.

Example 3.7 (Logistic Regression)**.** *Consider a linear model for multiclass classification where $\mathcal{Y} = [c]$ and*

$f(x,j) = w_j^\intercal x$ with $\sum_{j=1}^c \|w_j\|_2^2 \le 1$. Suppose $x \in \mathbb{R}^p$ is distributed in the unit sphere $\|x\|_2 \le 1$ with expectation $\mathbb{E}[x] = \alpha \in \mathbb{R}^p$ and covariance matrix $\sigma^2 I_{p\times p}$. Without constraint, the Rademacher complexity is upper bounded as $\Re_m(\mathcal{F}) \le \sqrt{c/m}$ as in Cortes et al. (2016) (Theorem 2). Now, consider a constraint that removes exactly one label so that $C(x) \equiv [c-1]$. With regularization, for sufficient small $t < 1/(c+2)$, we have the following bound

$$\Re_m(\mathcal{F}_t) \le \frac{1}{2}\left(\sqrt{\frac{c}{m}} + \sqrt{\frac{c-\sigma^2-\|\alpha\|_2^2}{m}}\right) \tag{11}$$

which is strictly tighter than the standard bound. Intuitively, if x is concentrated around the origin $\mathbf{0}$, the prediction by any classifier will tend to be a uniform distribution. Therefore, a large bias and variance in x (captured by $\sigma^2 + \|\alpha\|_2^2$) help to distinguish models with different levels of violation.

Compare to existing results. Previous works mostly consider a zero-one loss for both classification and violation under the assumption that the risk minimizer also achieves zero violation. Then, one can simply preclude all the classifiers $f \in \mathcal{F}$ that have nonzero empirical violations on the unlabeled dataset and find the ERM among the remaining classifiers. This approach has been theoretically studied in Balcan & Blum (2005; 2010) for binary classification and Tulab et al. (2014) in a similar manner for regression by characterizing the complexity of the reduced set of hypotheses that achieve zero violation. Conceptually, we can regard this algorithm as a special case of problem (4) when $\rho = \infty$. Our study, therefore, extends previous works with a soft learning objective to multiclass classification problems.

4. Inference with Constraints

An *inference algorithm* is a mapping $\mathcal{F} \times \mathcal{X} \to \Delta_\mathcal{Y}$. By default, we define it as the softmax inference: $(f,x) \mapsto \mathbb{P}_f(\cdot|x)$. When performing *inference with constraints* (or constrained inference), we modify this softmax mapping for the given function f using the additional information of C.

In this section, we study the Constrained Conditional Model (CCM) (Chang et al., 2008; 2012), a broad family of models that perform inference with constraints. We show at testing time, whether CCM reduces the risk depends on whether the model's expected violation is larger than the noise rate of the constraint V_{ora} (Theorem 4.3). In particular, when the constraint is noise-free, CCM always achieves a smaller or equal risk. Furthermore, we show better risks are achieved if the constrained inference is also performed at training time, and pursuing this optimal risk leads to a learning objective that contrasts with the one used in the regularization approach (Proposition 4.6).

To make distinguishment, we will refer to a model in the original spaces $\mathcal{F}$ as a *base model* and refer to an augmented model as a *constrained model*.

4.1. Constrained Conditional Model

CCM augments existing scoring functions using a linear combination with the violation function. Precisely, given a vanilla scoring space $\mathcal{F}$, the scoring space of CCM is defined as follows.

Definition 4.1 (Constrained conditional model (Chang et al., 2008; 2012))**.** *Given a scoring space $\mathcal{F}$, a constraint C and a fixed tradeoff parameter $\mu \in [0,\infty]$, the scoring space of the Constrained Conditional Model (CCM) is defined as:*

$$\mathcal{F}^\mu := \{(x,y) \mapsto f(x,y) - \mu v_C(x,y) | f \in \mathcal{F}\} \tag{12}$$

We will also denote

$$f^\mu(x,y) := f(x,y) - \mu v_C(x,y) \tag{13}$$

to be the augmented scoring function for a given $f \in \mathcal{F}$. In particular, setting $\mu = \infty$ will assign a score $-\infty$ to any $y \notin C(x)$, which implies $\mathbb{P}_{f^\infty}(y|x) = 0$, namely forcing strictly-constrained inference.

The tradeoff parameter μ allows CCM to improve the base model f despite noisy constraints, as we will discuss in detail in the following sections. Otherwise, if the noise rate is large, performing strictly-constrained inference can be harmful because it assigns 0 probability mass to any label y that is outside $C(x)$ and hence has a classification loss $L(x, y_{\text{ora}}, f^\infty) = 1$ at any x where $y_{\text{ora}} \notin C(x)$.

The learner can choose whether or not to perform the constrained inference either at the training time. This choice leads to the following two different approaches:

1. On-training approach: perform constrained inference both at training and testing time, and directly find the ERM over $\mathcal{F}^\mu$ using labeled data (also known as *(Inference Based Training* in (Punyakanok et al., 2005))

2. Post-training approach: first find the ERM over the vanilla $\mathcal{F}$ using labeled data, and then perform constrained inference at the testing time (also known as *Learning Plus Inference* in (Punyakanok et al., 2005)).

For both approaches, the generalization ability of CCM is characterized by the complexity of $\mathcal{F}^\mu$. So, we first point out that CCM does not increase the Rademacher complexity.

Proposition 4.2 (Rademacher Complexity of CCM)**.** *For any fixed $\mu \ge 0$ and $m \in \mathbb{N}$, we have the following identity:*

$$\Re_m(\mathcal{F}^\mu) = \Re_m(\mathcal{F}) \tag{14}$$

4.2. Post-training Constrained Inference

For a given and fixed classifier f (presumably trained with data), how does performing constrained inference impact the model performance? In this section, we study the change in risk when the learner chooses to augment f as a CCM f^μ defined in (13). It is most convenient to characterize the risk of a CCM using the cross-entropy loss, although we will also conduct the same analysis for the hinge and ℓ^1 losses, as we will point out later. To start with, for any f and $\mu \in [0, \infty]$, we let

$$\Delta^\mu_{\mathrm{CE}}(f) := R_{\mathrm{CE}}(f) - R_{\mathrm{CE}}(f^\mu) \tag{15}$$

be the difference in the risk between the base model and the CCM (the larger the better).

Theorem 4.3 (Change in cross-entropy risk)**.** *We have:*

(a) For any fixed model f, there exists an $\mu_0 > 0$ such that $R_{\mathrm{CE}}(f^{\mu_0}) < R_{\mathrm{CE}}(f)$ if and only if

$$V(f) > V_{\mathrm{ora}} \tag{16}$$

(b) The change in risk can be lower bounded as

$$\Delta^\mu_{\mathrm{CE}}(f) \geq V(f)(1 - \mathrm{e}^{-\mu}) - \mu V_{\mathrm{ora}} \tag{17}$$

(c) In particular, if the constraint is noise-free, we have

$$\Delta^\infty_{\mathrm{CE}}(f) = V_{\mathrm{CE}}(f) \tag{18}$$

The first result describes the sufficient and necessary condition for constrained inference to be helpful. It requires f to have a larger violation (measured by ℓ^1 violation) than the true data on average so that it has the potential to be improved. This condition is easier to satisfy when the constraint is less noisy.

The second result further quantifies the risk reduction as an explicit function of μ. The last result shows that in the noise-free case, the maximum risk reduction is exactly the expected violation measured by cross-entropy. Its consequences will be further discussed in the next section.

Remark 4.4 (CCM with alternative losses)**.** *We present the counterparts of Theorem 4.3 for hinge loss and ℓ^1 loss in the Appendix D. The information delivered by those results is consistent with Theorem 4.3 in the sense that (1) whether CCM can reduce the risk depends on the comparison between the violation of the original model and the oracle. (2) the reduction can be described or lower bounded by some measures of the violation.*

The drawback of the hinge loss is its non-smoothness due to the discontinuity of the argmax inference. The drawback of the ℓ^1 loss is that the range of μ such that $R(f^\mu) \leq R(f)$ can be disconnected and difficult to describe. Therefore, we provide weaker results by deriving only sufficient or necessary conditions for CCM to reduce the risks.

As an application of Theorem 4.3, we derive a sufficient condition under which CCM achieves smaller risks.

Corollary 4.5 (Choice of μ)**.** *Assuming $V(f) \geq V_{\mathrm{ora}}$, then $R_{\mathrm{CE}}(f^\mu) \leq R_{\mathrm{CE}}(f)$ if the following condition holds:*

$$\mu \leq W(-\eta/\mathrm{e}^\eta) + \eta \tag{19}$$

where $\eta := V(f)/V_{\mathrm{ora}}$ is the relative violation rate and W is the Lambert W function whose value $W(t)$ is defined to be the solution to the equation $w\mathrm{e}^w = t$ of w.

The RHS of (19) increases with η and vanishes as $\eta \to 1$. In particular, when the constraint is noise-free, one should encourage strictly-constrained inference and set $\mu = \infty$. We also provide a plot of the RHS in the proof in the appendix.

4.3. On-training Constrained Inference

In this subsection, we study the on-training approach where we perform constrained inference both at the training and testing time. We use the results we established in the last subsection to describe the learning objective of the on-training approach, and argue that it achieves better risks than the post-training approach. Based on this, we further show that minimizing the cross entropy over CCM encourages a large violation of the base model, which contrasts the learning objective (9) that is used in regularization.

We provide a simplified analysis for the noise-free setting where we choose $\mu = \infty$ and perform strictly-constrained inference. Then, the on-training approach aims to find the optimal (in terms of cross entropy) base model as follows:

$$f_{\mathrm{on}} := \operatorname*{argmin}_{f \in \mathcal{F}} R_{\mathrm{CE}}(f^\infty) \tag{20}$$

(recall f^∞ means performing strictly-constrained inference with f) We characterize the behavior of f_{on} with the following results, which are direct corollaries of Theorem 4.3.

Proposition 4.6 (Learning Optimal CCM; Post-training vs On-training)**.** *Assuming C is noise-free, we can reformulate the learning objective (20) as*

$$f_{\mathrm{on}} = \operatorname*{argmin}_{f \in \mathcal{F}} R_{\mathrm{CE}}(f) - V_{\mathrm{CE}}(f) \tag{21}$$

A fundamental difference. Surprisingly, the reformulated learning objective (21) is opposite to the surrogate regularized objective defined in (9) in their attitudes towards violations. This contrast suggests a fundamental difference between regularization and constrained inference: the regularization method views violation as a bad thing and it precludes classifiers with substantial violations. But constrained inference corrects a model from its violation, so a large violation means a great potential to be improved.

On-training vs post-training. Loosely speaking, this result also suggests that in general, the best constrained model is not the constrained best model. To be more precise, suppose we perform post-training constrained inference for the cross-entropy risk minimizer in the vanilla model, i.e., $f_{\text{post}} := \operatorname{argmin}_{f\in\mathcal{F}} R_{\text{CE}}(f)$. Then, we can reformulate the definition of f_{post} as

$$f_{\text{post}} := \operatorname*{argmin}_{f\in\mathcal{F}} \underbrace{(R_{\text{CE}}(f) - V_{\text{CE}}(f))}_{\text{objective in (21), post-training risk}} + V_{\text{CE}}(f) \tag{22}$$

which can be regarded as a "regularized" version of (21). Therefore, similar to Proposition 3.2, we can argue that the risk minimizer f_{post} over $\mathcal{F}$, as a base model of CCM, contains a bias towards a higher risk than the on-training method's as follows:

$$R_{\text{CE}}(f_{\text{on}}^{\infty}) \leq R_{\text{CE}}(f_{\text{post}}^{\infty}) \leq R_{\text{CE}}(f_{\text{on}}) - \min_{f\in\mathcal{F}} V_{\text{CE}}(f) \tag{23}$$

The proof is included in the proof of Proposition 4.6.

Computational considerations. In practical structured prediction problems where the output is sequential or graphical, performing constrained inference during training time is typically expensive due to the complexity of the constraints. For example, as pointed out by Xu et al. (2018), when the constraint is defined by a logical expression over several output variables, computing the probability of constraint being satisfied corresponds to the problem of weighted model counting (WMC) and is #P-complete (Roth, 1996). Therefore, to implement the on-training approach in practice, one can alternatively use approximate inference to ensure tractability. For example, strictly constrained inference, formulated as Integer Linear Programming (Roth & Yih, 2004), can be further relaxed as Linear Programming (Martins et al., 2009a). Another example is *amortized inference* (Chang et al., 2015), which accelerates the convergence to the optimal model while only performing exact inference in every $\tau > 1$ iterations.

Compare to existing results. There has been limited theoretical work discussing the impact of performing constrained inference. The most related one is Punyakanok et al. (2005), which derives VC-style generalization bounds for linear structured models to argue that (1) performing strictly constrained inference in a post-training manner (*Learning Plus Inference* in the paper) improves the model performance and (2) the on-training approach (*Inference Based Training* in the paper) further reduces the error in the long run. Our approach directly analyses the classification risk and extends the comparison to noisy constraints and soft-constrained inference with CCM.

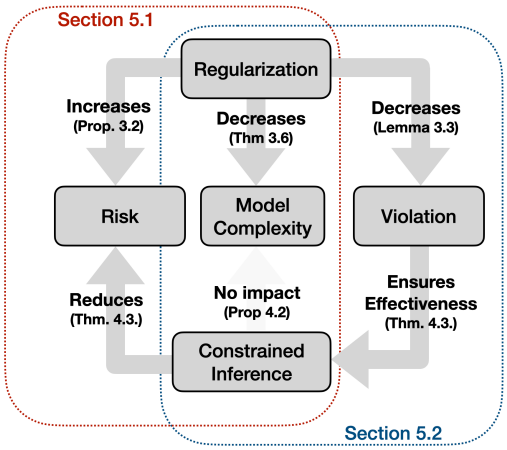

Figure 1. A summary of the established results, as motivations to the problems of this section. In Section 5.1, we describe how CCM can reduce the additional risk introduced by regularization. In section 5.2, we claim that the decrease of violation with regularization can make post-training constrained inference unnecessary.

5. Regularization with Constrained Inference

We have seen that regularization and constrained inference have different impacts on the generalization gap and the risk. On one hand, CCM has an equal Rademacher complexity (Proposition 4.2) as the original model $\Re(\mathcal{F})$, which can be reduced by regularization. So, performing regularized algorithm to CCM also reduces the generalization gap. On the other hand, their impacts on the risks are contradicting, as summarized in figure 1. In this section, we aim to describe how these impacts can interact with each other by applying our established results to explore the usage of these two methods together.

We show both positive and negative results for the combination. On one hand, we propose sufficient conditions under which the bias introduced by regularization can be compensated by performing constrained inference (Proposition 5.1). On the other hand, we study if post-training constrained inference can reduce the risk of the optimal classifier f_ρ. We show with a noisy constraint, choosing a large value of ρ in the regularized objective (4) will make CCM incapable to reduce the risk (Proposition 5.2).

5.1. CCM Compensates for Regularization Bias

As the red part of Figure 1 summarizes, we have shown that the regularization and constrained inference have contradicting influences on the risk. Moreover, the regularization bias is controlled by the violation of the risk minimizer (Proposition 3.2), which can be reduced by constrained inference. This suggests the possibility for CCM to reduce the additional risk introduced by regularization.

We formally describe this phenomenon by considering the following combination: an on-training approach that aims to find the minimizer of the following regularized surrogate objective over the CCM $\mathcal{F}^\mu$:

$$f_\star^\mu := \operatorname*{argmin}_{g \in \mathcal{F}^\mu} R_{\mathrm{CE}}(g) + \rho V_{\mathrm{CE}}(g) \tag{24}$$

Recall that $R_{\mathrm{CE}}(f_{\mathrm{post}})$ is the minimum cross-entropy risk that can be achieved in $\mathcal{F}$. We show that unlike the vanilla regularized objective (4), it is possible for this algorithm to achieve a smaller risk than $R_{\mathrm{CE}}(f_{\mathrm{post}})$ as follows.

Theorem 5.1 (Regularization with on-training constrained inference)**.** *If CCM improves f_{post} so that $\Delta_{\mathrm{CE}}^\mu(f_{\mathrm{post}}) > 0$, then letting*

$$\rho < \frac{V_{\mathrm{CE}}(f_{\mathrm{post}}) - \mu V_{\mathrm{ora}}}{V_{\mathrm{CE}}(f_{\mathrm{post}}^\mu)} - 1 \tag{25}$$

will imply $R_{\mathrm{CE}}(f_\star^\mu) < R_{\mathrm{CE}}(f_{\mathrm{post}})$.

This result shows a small choice of ρ allows the regularized optimizer $f_\star^\mu$ to achieve better cross-entropy. A less noisy constraint allows more choices of ρ to make this happen. In particular, when the constraint is noise-free, since $V_{\mathrm{CE}}(f_{\mathrm{post}}^\mu) \to 0$ as $\mu \to \infty$, driving μ to ∞ will make $R(f_\star^\mu) < R(f_{\mathrm{post}})$ for all $\rho > 0$. As a cost, regularization will be less effective in reducing the Rademacher complexity with a large value of μ. In the extreme case, all the classifiers in $\mathcal{F}^\infty$ make zero violation, and hence cannot be distinguished by the regularization objective.

5.2. Post-regularized-training Constrained Inference

Finally, as the blue part of Figure 1 summarizes, we have shown that post-training inference is beneficial only if the average violation of f is larger than V_{ora} (Theorem 4.3). However, the minimizer of the regularized objective f_ρ tends to have a small violation (Proposition 3.3) scaled with $1/\rho$. Therefore, it is possible that choosing a large value of ρ will make post-training incapable to reduce the risk with a noisy constraint. Formally, assuming a model is already trained with the vanilla regularized ℓ^1 objective as in (4), we have the following holds.

Theorem 5.2 (When post-training inference is not helpful for regularized model)**.** *Recall $V(f_\infty)$ is the minimal expected violation that can be achieved by $\mathcal{F}$. If $V_{\mathrm{ora}} \geq V(f_\infty)$ and*

$$\rho \geq \frac{1}{V_{\mathrm{ora}} - V(f_\infty)} \tag{26}$$

then the minimizer f_ρ of the regularized objective (4) will not *be improved by post-training constrained inference for any $\mu \in (0, \infty]$ in the sense that $R_{\mathrm{CE}}(f_\rho) \leq R_{\mathrm{CE}}((f_\rho)^\mu)$.*

The RHS of (26) shrinks with a larger noise rate V_{ora} and smaller $V(f_\infty)$. Intuitively, a more noisy constraint is less helpful (Theorem 4.3), while a small value of $V(f_\infty)$ allows f_ρ to violate less (Proposition 3.3) and hence gains fewer benefits from constrained inference (Theorem 4.3). As a consequence, with a noisy constraint, choosing a large ρ in the regularized objective will make post-training constrained inference unnecessary or even harmful.

6. Related Works

Regularization with constraints. In the context of structured prediction, the Posterior Regularization (PR) framework (Ganchev et al., 2010) proposed to regularize the log-likelihood by adding a distance of the probabilistic prediction to the constrained subspace of distributions. The CoDL algorithm (Chang et al., 2007; 2012) is a semi-supervised algorithm that repetitively assigns constrained pseudo-labels to the unlabeled dataset and uses pseudo-labels to retrain the model. CoDL and PR are further unified in Samdani et al. (2012) as special cases of a parameterized EM algorithm. More recent works have proposed injecting logical constraints into deep models by augmenting the training objective with explicitly defined violation functions, such as the semantic loss (Xu et al., 2018), the DL2 loss (Fischer et al., 2019) and the inconsistency loss (Li et al., 2019), which motivate our theoretical formulation in (4).

Inference with constraints. The idea of injecting prior knowledge directly into a predictive model dates back to Roth & Yih (2004), which formulates the problem of inference with hard constraints as Integer Linear Programming (ILP). The idea of constrained inference has been followed and developed by NLP researchers and empirically shown to be effective in various problems such as summarization (Clarke & Lapata, 2008), temporal reasoning (Ning et al., 2018b), semantic parsing (Scholak et al., 2021) and text generation (Lu et al., 2022). Chang et al. (2008; 2012) further defines the CCM to incorporate soft constraints into linear models. Another related work is Enrique Sucar et al. (2014), which uses Bayesian networks to model the label correlations and define an order to the labels. The order information is then taken as extended features at inference time. Theoretically, Punyakanok et al. (2005) provides a comparison between the on-training and post-training constrained inference using VC-style error bounds.

Semi-supervised learning theory. Several theoretical semi-supervised learning frameworks such as Balcan & Blum (2005; 2010) and Tulab et al. (2014) illustrate how hard constraints on the hypothesis space could reduce the generalization error. A detailed comparison can be seen in the discussion at the end of Section 3.

Learning with partial labels. The problem of learning with constraints is closely related to the problem of learning from partial labels (also known as superset labels) (Cour et al., 2011; Liu & Dietterich, 2014; Cid-Sueiro, 2012; Cabannes et al., 2020) where each instance x in the dataset is assigned with a partial label s which also takes value in $2^{\mathcal{Y}}$. The difference is that the constraint mapping itself is known to the learner and hence can be encoded in the inference algorithm directly, for example, via the CCM. Another difference is that the partial labels are typically more informative and can guarantee learnability alone (Liu & Dietterich, 2014; Wang et al., 2020). In contrast, the constraints that appear in practice typically provide only side information and need to be used with gold labels together.

7. Conclusion and Future Works

In this paper, we presented a theoretical study of two methods to encode label constraints into a learning system: regularization and constrained inference. We compared these two approaches by quantifying their impact on the optimal risk as well as the generalization error. Our study revealed that the success of these two approaches replies on different data assumptions: the regularization method requires the optimal classifier in the model to have a small violation while constrained inference requires the true data to have a small violation. We further elucidated the detrimental consequences that arise when these assumptions fail to hold. Finally, we demonstrate how their impacts on the model can interact when used together.

We have focused on multiclass classification, aiming to provide a starting point for understanding the different mechanisms of the two methods. For future work, we will extend the discussion to structured prediction problems where complex constraints are naturally defined. In particular, while the presence of constraints can improve the model performance, it also suggests a strong dependency inside the structure, which may hurt the generalization performance, as pointed out by London et al. (2016).

Acknowledgements

This work was partially supported by Contract FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

This work was also partially sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0080. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

This work was also partially funded by ONR Contract N00014-19-1-2620.

References

Balcan, M.-F. and Blum, A. A PAC-Style Model for Learning from Labeled and Unlabeled Data. In *Proc. of the ACM Conference on Computational Learning Theory (COLT)*, 2005.

Balcan, M.-F. and Blum, A. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57:1 – 46, 2010.

Cabannes, V., Rudi, A., and Bach, F. Structured prediction with partial labelling through the infimum loss. In *ICML*, pp. 1230–1239, 2020.

Chang, K.-W., Upadhyay, S., Kundu, G., and Roth, D. Structural Learning with Amortized Inference. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, 2015. URL http://cogcomp.org/papers/CUKR15.pdf.

Chang, M.-W., Ratinov, L., and Roth, D. Guiding Semi-Supervision with Constraint-Driven Learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 280–287, Prague, Czech Republic, 6 2007. Association for Computational Linguistics. URL http://cogcomp.org/papers/ChangRaRo07.pdf.

Chang, M.-W., Ratinov, L., Rizzolo, N., and Roth, D. Learning and Inference with Constraints. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, 7 2008. URL http://cogcomp.org/papers/CRRR08.pdf.

Chang, M.-W., Ratinov, L., and Roth, D. Structured Learning with Constrained Conditional Models. *Machine Learning*, 88(3):399–431, 6 2012. URL http://cogcomp.org/papers/ChangRaRo12.pdf.

Cid-Sueiro, J. Proper losses for learning from partial labels. In *NeurIPS*, pp. 1574–1582, 2012.

Clarke, J. and Lapata, M. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research (JAIR)*, 2008.

Cortes, C., Kuznetsov, V., Mohri, M., and Yang, S. Structured Prediction Theory Based on Factor Graph Complexity. In *NIPS*, 2016.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *Journal of Machine Learning Research*, 12: 1501–1536, 2011. ISSN 1532-4435.

Enrique Sucar, L., Bielza, C., Morales, E. F., Hernandez-Leal, P., Zaragoza, J. H., and Larrañaga, P. Multi-label classification with bayesian network-based chain classifiers. *Pattern Recognition Letters*, 41:14–22, 2014. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2013.11.007. URL https://www.sciencedirect.com/science/article/pii/S0167865513004303. Supervised and Unsupervised Classification Techniques and their Applications.

Fischer, M., Balunovic, M., Drachsler-Cohen, D., Gehr, T., Zhang, C., and Vechev, M. T. DL2: Training and Querying Neural Networks with Logic. In *International Conference on Machine Learning*, 2019.

Ganchev, K., Graça, J., Gillenwater, J., and Taskar, B. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research (JMLR)*, 2010.

Hu, Z., Ma, X., Liu, Z., Hovy, E., and Xing, E. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2410–2420, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1228. URL https://aclanthology.org/P16-1228.

Li, T., Gupta, V., Mehta, M., and Srikumar, V. A logic-driven framework for consistency of neural models. *ArXiv*, abs/1909.00126, 2019.

Liu, L.-P. and Dietterich, T. G. Learnability of the superset label learning problem. In *ICML*, pp. 1629—-1637, 2014.

London, B., Huang, B., and Getoor, L. Stability and Generalization in Structured Prediction. *Journal of Machine Learning Research*, 17(221):1–52, 2016. URL http://jmlr.org/papers/v17/15-501.html.

Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., Le Bras, R., Qin, L., Yu, Y., Zellers, R., Smith, N. A., and Choi, Y. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 780–799, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.57. URL https://aclanthology.org/2022.naacl-main.57.

Martins, A., Smith, N., and Xing, E. Concise integer linear programming formulations for dependency parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 342–350, Suntec, Singapore, August 2009a. Association for Computational Linguistics. URL https://aclanthology.org/P09-1039.

Martins, A., Smith, N. A., and Xing, E. Concise integer linear programming formulations for dependency parsing. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, 2009b.

Meshi, O., London, B., Weller, A., and Sontag, D. Train and Test Tightness of LP Relaxations in Structured Prediction. *Journal of Machine Learning Research*, 20(13):1–34, 2019. URL http://jmlr.org/papers/v20/17-535.html.

Ning, Q., Feng, Z., Wu, H., and Roth, D. Joint Reasoning for Temporal and Causal Relations. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2278–2288, Melbourne, Australia, 7 2018a. Association for Computational Linguistics. URL http://cogcomp.org/papers/NingFeWuRo18.pdf.

Ning, Q., Wu, H., and Roth, D. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1318–1328, Melbourne, Australia, 7 2018b. Association for Computational Linguistics. URL http://cogcomp.org/papers/NingWuRo18.pdf.

Pan, X., Mehta, M., and Srikumar, V. Learning Constraints for Structured Prediction Using Rectifier Networks. *ArXiv*, abs/2006.01209, 2020.

Punyakanok, V., Roth, D., au Yih, W., and Zimak, D. Learning and Inference over Constrained Output. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1124–1129, 2005. URL http://cogcomp.org/papers/PRYZ05.pdf.

Roth, D. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2):273–302, 4 1996. URL http://cogcomp.org/papers/hardJ.pdf.

Roth, D. and Yih, W. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In Ng, H. T. and Riloff, E. (eds.), *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*, pp. 1–8. Association for Computational Linguistics, 2004. URL http://cogcomp.org/papers/RothYi04.pdf.

Samdani, R., Chang, M.-W., and Roth, D. Unified Expectation Maximization. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 6 2012. URL http://cogcomp.org/papers/SamdaniChRo12.pdf.

Scholak, T., Schucher, N., and Bahdanau, D. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9895–9901, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.779. URL https://aclanthology.org/2021.emnlp-main.779.

Tulab, T., hula, and Rudin, C. Generalization bounds for learning with linear, polygonal, quadratic and conic side knowledge. *Machine Learning*, 100:183–216, 2014.

Wang, H., Zhang, H., Chen, M., and Roth, D. Learning Constraints and Descriptive Segmentation for Subevent Detection. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://cogcomp.seas.upenn.edu/papers/WZCR21.pdf.

Wang, K., Ning, Q., and Roth, D. Learnability with Indirect Supervision Signals. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020. URL https://cogcomp.seas.upenn.edu/papers/WangNiRo20.pdf.

Xu, J., Zhang, Z., Friedman, T., Liang, Y., and Van den Broeck, G. A semantic loss function for deep learning with symbolic knowledge. In *ICML*, pp. 5502–5511, 2018.

Appendix

A. Details on Loss Function

The ℓ^1 loss is a smoothed alternative to the zero-one loss and has been used in the theoretical analysis for the generalization error, see, for example, in London et al. (2016) (Section 6.2). It can be related to other common loss functions as follows.

As distances on the probability simplex. Let $e_y \in \mathbb{R}^{|\mathcal{Y}|}$ be a one-hot vector with the yth coordinate be 1 and all others be 0. We then have that

$$L(x, y_{\text{ora}}, f) := 1 - \mathbb{P}_f(y_{\text{ora}}|x) = \frac{1}{2}\|e_{y_{\text{ora}}} - \mathbb{P}_f\|_1$$

Moreover, since our label space $\mathcal{Y}$ is of finite cardinality, we further have that $\frac{1}{2}\|e_{y_{\text{ora}}} - \mathbb{P}_f\|_1 = \text{TV}(e_{y_{\text{ora}}}, \mathbb{P}_f)$, the total variation distance.

Relation to zero-one loss. By introducing a temperature parameter $t \in \mathbb{R}_{\geq 0}$ to the softmax function, it is well known that $\lim_{t\to\infty} \text{softmax}(tu) = \text{argmax}(u)$ for a vector u. This implies

$$\lim_{t\to\infty} L(x, y_{\text{ora}}, tf) = 1 - \mathbb{1}\{\operatorname*{argmax}_{y\in\mathcal{Y}} f(x, y) = y_{\text{ora}}\} = \mathbb{1}\{\operatorname*{argmax}_{y\in\mathcal{Y}} f(x, y) \neq y_{\text{ora}}\}$$

which is the zero-one loss. Since performing softmax inference with temperature t can be equivalently regarded as performing softmax inference for the scoring space $t\mathcal{F}$, for the simplicity of our presentation, we omit the temperature parameter in the softmax inference.

Relation to cross-entropy. The total variation distance to a one-hot probability can be lower bounded by cross-entropy due to Pinsker's inequality. More directly, in our case, we have $1 - p \leq -\log(p)$ for any $p \in [0, 1]$ from basic inequality. This implies $L(x, y, f) \leq L_{\text{CE}}(x, y, f)$.

In conclusion, the ℓ^1 loss is a ℓ^1 and total variation distance on the probability space, is a smoothed version of the zero-one loss, and is upper bounded by cross-entropy. It is differentiable and bounded so that we can derive generalization bounds with Rademacher complexity. Another reason that we are interested in softmax inference will be clearer in the discussion for constrained inference, where in Theorem 4.3, D.1 and D.4, the change of expected cross entropy and ℓ^1 loss can be lower bounded by a smooth function. But with the argmax inference, the risk is in general not continuous and needs to be assumed to be Lipschitz to obtain similar results.

B. Proofs from Section 3

B.1. Proof of Proposition 3.2

The first inequality is straightforward. For the second inequality, by definition (4) we have

$$R(f_\rho) + \rho V(f_\rho) \leq R(f_0) + \rho V(f_0)$$

and

$$V(f_\rho) \geq V(f_\infty).$$

Combining the two above inequalities yields

$$R(f_\rho) + \rho V(f_\infty) \leq R(f_0) + \rho V(f_0).$$

The desired inequality follows by rearranging these terms. This argument also holds if we replace the expectations with empirical estimates.

To see how the RHS bound can be reached, consider the following scoring space that contains two classifiers, f_0 and f_∞, and an instance space $\mathcal{X}$ that only contains one point x. Let $C(x) = \{y_{\text{ora}}, y'\}$. Let f_0 be such that $\mathbb{P}_{f_0}(y_{\text{ora}}) = a \in (0, 1)$ and $\mathbb{P}_{f_0}(y') = b$. Let f_∞ be such that $\mathbb{P}_{f_\infty}(y_{\text{ora}}) = a - \epsilon_1$ and $\mathbb{P}_{f_\infty}(y') = b + \epsilon_2$ so that $\epsilon_1 < \rho\epsilon_2$. Then

$$R(f_\infty) + \rho V(f_\infty) \leq 1 - (a - \epsilon_1) + \rho(b - \epsilon_2) < 1 - a + \rho b = R(f_0) + \rho V(f_0) \tag{27}$$

which means f_∞ will be preferred to f_0 by the regularized objective.

B.2. Proof of Lemma 3.3

By definitions, we have

$$\begin{aligned}\rho V(f_\rho) &\le R(f_\rho) + \rho V(f_\rho)\\ &\le R(f_\infty) + \rho V(f_\infty)\\ &\le 1 + \rho V(f_\infty)\\ &\le 1 + \rho u\end{aligned} \tag{28}$$

Therefore, we have that $V(f_\rho) \le u + 1/\rho$.

B.3. Proof of Lemma 3.4

To prove this theorem, we need the following lemmas. The first one is a contraction inequality established in (Cortes et al., 2016).

Lemma B.1 (Lemma 5 from (Cortes et al., 2016))**.** *Let $\mathcal{H}$ be a set of functions mapping $\mathcal{X}$ to $\mathbb{R}^N$. Suppose Φ_i is μ_i-Lipschtz with the 2-norm, i.e.,*

$$|\Phi_i(v') - \Phi_i(v)| \le \mu_i \|v' - v\|_2 \quad \forall v, v' \in \mathbb{R}^N \tag{29}$$

Then for any set of m points $x_1, \dots, x_m \in \mathcal{X}$, the following inequality holds

$$\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^m \sigma_i \Phi_i(h(x_i))\right] \le \frac{\sqrt{2}}{m}\mathbb{E}_\epsilon\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^m \sum_{j=1}^N \epsilon_{ij}\mu_i h_j(x_i)\right] \tag{30}$$

where σ_is and ϵ_{ij}s are independent Rademacher variables uniformly distributed over $\{-1, +1\}$.

The second one computes the Lipschitz constants of the ℓ^1 losses by bounding its gradient's 2-norm.

Lemma B.2 (Lipschitzness)**.** *Given a scoring function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, let $f(x) = [f(x, y)]_{y\in\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}|}$ be the vector of scores for each label. For any two scoring functions f, f' and data (x, y), we have that*

$$|\mathbb{P}_f(y|x) - \mathbb{P}_{f'}(y|x)| \le \frac{\sqrt{2}}{4}\|f(x) - f'(x)\|_2 \tag{31}$$

Furthermore, for any constraint C, we have

$$|\mathbb{P}_f(C|x) - \mathbb{P}_{f'}(C|x)| \le \frac{1}{4}\sqrt{1 + \frac{1}{|C(x)|}}\|f(x) - f'(x)\|_2 \tag{32}$$

where $\mathbb{P}_f(C|x) = \mathbb{P}_f(C(x)|x) = \sum_{y\in C(x)} \mathbb{P}_f(y|x)$.

Proof. We start with the second claim. Suppose $C(x) = \mathcal{Y}$, then $\mathbb{P}_f(C|x) = 0$ for any scoring function f, so the inequality trivially holds. Next, we assume $C(x) \subset \mathcal{Y}$. Given a constraint $C : \mathcal{X} \to 2^{\mathcal{Y}}$, the derivative of its violation function with respect to the score for a label y is

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}_f(C|x)}{\mathrm{d}f(x,y)} &= \sum_{y'\in C(x)} \frac{\mathrm{d}\mathbb{P}_f(y'|x)}{\mathrm{d}f(x,y)}\\ &= \sum_{y'\in C(x)} \mathbb{P}_f(y|x)\mathbb{1}\{y' = y\} - \mathbb{P}_f(y|x)\mathbb{P}_f(y'|x)\end{aligned} \tag{33}$$

The 2-norm of the gradient of the mapping $f(x) \mapsto \mathbb{P}_f(y|x)$ is then

$$\left(\sum_{y\in\mathcal{Y}}\left(\sum_{y'\in C(x)} \mathbb{P}_f(y|x)\mathbb{1}\{y' = y\} - \mathbb{P}_f(y|x)\mathbb{P}_f(y'|x)\right)^2\right)^{1/2} \tag{34}$$

which is maximized when $\mathbb{P}_f(y|x) = \frac{1}{2|C(x)|}$ for all $y \in C(x)$ and $\mathbb{P}_f(y|x) = \frac{1}{2(\mathcal{Y}-|C(x)|)}$ for all $y \notin C(x)$ (so that $\mathbb{P}_f(C|x) = 1/2$). The maximum is then

$$\begin{aligned}
&\left(\sum_{y\in C(x)}\left(\sum_{y'\in C(x)}\mathbb{P}_f(y|x)\mathbb{1}\{y'=y\} - \mathbb{P}_f(y|x)\mathbb{P}_f(y'|x)\right)^2 + \sum_{y\notin C(x)}\left(\sum_{y'\in C(x)}\mathbb{P}_f(y|x)\mathbb{P}_f(y'|x)\right)^2\right)^{1/2} \\
&= \sqrt{|C(x)|\left(\frac{1}{4|C(x)|}\right)^2 + |\mathcal{Y}-C(x)|\left(\frac{1}{2|\mathcal{Y}-C(x)|}\right)^2} \\
&= \sqrt{\frac{1}{16|C(x)|} + \frac{1}{16|\mathcal{Y}-C(x)|}} \\
&\leq \sqrt{\frac{1}{16|C(x)|} + \frac{1}{16}} \\
&= \frac{1}{4}\sqrt{1+\frac{1}{|C(x)|}}
\end{aligned} \tag{35}$$

The boundedness of the gradient implies that the function $f(x) \mapsto \mathbb{P}_f(C|x)$ is Lipschitz with a Lipschitz constant $\frac{1}{4}\sqrt{1+\frac{1}{|C(x)|}}$.

The first claim then follows by considering the special constraint $C(x) := \{y_{\text{ora}}(x)\}$ so that $|C(x)| = 1$. □

Next, we present the proof of the theorem. By standard Rademacher complexity bounds, given a labeled dataset S of size m, for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds uniformly for $f \in \mathcal{F}$:

$$R(f) \leq \widehat{R}(f; S_{\mathsf{L}}) + 2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}} \tag{36}$$

where

$$\mathcal{H} := \{(x,y) \mapsto 1 - \mathbb{P}_f(y|x) : f \in \mathcal{F}\} \tag{37}$$

By the contraction lemma and Lipschitzness, we have

$$\begin{aligned}
\mathfrak{R}_m(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^m \sigma_i\left(1-\mathbb{P}_f(y_i|x_i)\right)\right] \\
&\leq \frac{\sqrt{2}}{m}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sum_{y\in\mathcal{Y}}\epsilon_{iy}\frac{\sqrt{2}}{4}f(x,y)\right] \\
&= \frac{1}{2m}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sum_{y\in\mathcal{Y}}\epsilon_{iy}f(x,y)\right]
\end{aligned} \tag{38}$$

This implies

$$R(f) \leq \widehat{R}(f; S_{\mathsf{L}}) + \mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m}} \tag{39}$$

The proof for the generalization bound of violation follows from the same argument. In particular, if the size of the constrained set $C(x)$ is a constant, namely $|C(x)| = c_0 < c = |\mathcal{Y}|$ for all $x \in \mathcal{X}$, then from Equation (35), we know that the mapping $x \mapsto 1 - \mathbb{P}_f(y|x)$ is Lipschitz with a Lipschitz constant $\frac{1}{4}\sqrt{\frac{1}{c_0}+\frac{1}{c-c_0}}$. So in this case, the generalization bound for the violation function can be improved as

$$V(f) \leq \widehat{V}(f; S_{\mathsf{U}}) + \frac{\sqrt{2}}{2}\sqrt{\frac{1}{c_0}+\frac{1}{c-c_0}}\mathfrak{R}_{m_{\mathsf{U}}}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}} \tag{40}$$

B.4. Proof of Theorem 3.6

Step 1. Showing the expected violation of $\widehat{f}_\rho$ is bounded.

First, we have with probability $1-\delta$,

$$\begin{aligned}\rho\widehat{V}(\widehat{f}_\rho) &\le \widehat{R}(\widehat{f}_\rho) + \rho\widehat{V}(\widehat{f}_\rho)\\ &\le \widehat{R}(f_\infty) + \rho\widehat{V}(f_\infty)\\ &\le 1 + \rho\widehat{V}(f_\infty)\\ &\le 1 + \rho\left(u + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}}\right)\end{aligned} \tag{41}$$

where the last step follows by applying Hoeffding's inequality to $\widehat{V}(f_\infty)$. This result implies $\widehat{V}(\widehat{f}_\rho) \le \frac{1}{\rho} + u + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}}$.

Second, Theorem 3.4 claims that with probability $1-\delta$, the following inequality holds:

$$V(\widehat{f}_\rho) - \widehat{V}(\widehat{f}_\rho) \le \Re_{m_{\mathsf{U}}}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}} \tag{42}$$

Putting these two inequalities together using union bound, we know with probability $1-2\delta$,

$$\begin{aligned}V(\widehat{f}_\rho) &\le \frac{1}{\rho} + u + \Re_{m_{\mathsf{U}}}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}} + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}}\\ &= \frac{1}{\rho} + u + B(\delta, m_{\mathsf{U}}, \mathcal{F})\end{aligned} \tag{43}$$

Namely, with probability no less than $1-2\delta$, $\widehat{f}_\rho$ lies in $\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}$, which is a fixed hypothesis class.

Step 2. Bounding the generalization gap of $\mathcal{L}_\rho$.

Since $\widehat{f}_\rho \in \mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}$, we can bound the generalization gap of $\mathcal{L}_\rho$ using the uniform convergence property of $\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}$. By standard decomposition,

$$\mathcal{L}_\rho(\widehat{f}_\rho) - \mathcal{L}_\rho(f_\rho) = \underbrace{\mathcal{L}_\rho(\widehat{f}_\rho) - \widehat{\mathcal{L}}_\rho(\widehat{f}_\rho)}_{(*)} + \underbrace{\widehat{\mathcal{L}}_\rho(\widehat{f}_\rho) - \widehat{\mathcal{L}}_\rho(f_\rho)}_{\le 0} + \underbrace{\widehat{\mathcal{L}}_\rho(f_\rho) - \mathcal{L}_\rho(f_\rho)}_{(**)} \tag{44}$$

For term $(*)$, combining the two inequalities in Lemma 3.4 and Step 1 via union bound, we know with probability $1-4\delta$,

$$(*) \le \Re_{m_{\mathsf{L}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{L}}}} + \rho\left(\Re_{m_{\mathsf{U}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{U}}}}\right) \tag{45}$$

For term $(**)$, using Hoeffding's inequality for the risk and violation separately, we have with probability $1-2\delta$,

$$(**) \le \sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{L}}}} + \rho\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{U}}}} \tag{46}$$

By union bound, with probability $1-6\delta$,

$$\mathcal{L}_\rho(\widehat{f}_\rho) - \mathcal{L}_\rho(f_\rho) \le \underbrace{\Re_{m_{\mathsf{L}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) + \rho\Re_{m_{\mathsf{U}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) + 2\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{L}}}} + 2\rho\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{U}}}}}_{\text{for convenience, denote these terms as } B'} \tag{47}$$

Step 3. Bounding the risk of f_ρ.

By Step 2, we have with probability $1-6\delta$,

$$\begin{aligned} R(\widehat{f}_\rho) &\le R(f_\rho) + \rho V(f_\rho) - \rho V(\widehat{f}_\rho) + B' \\ &\le R(f_0) + \rho V(f_0) - \rho V(\widehat{f}_\rho) + B' \\ &\le R(f_0) + \rho V(f_0) - \rho V(f_\infty) + B' \end{aligned} \tag{48}$$

We conclude that with probability $1-6\delta$,

$$\begin{aligned} R(\widehat{f}_\rho) &\le R(f_0) + \rho V(f_0) - \rho V(f_\infty) \\ &\quad + \Re_{m_{\mathsf{L}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) + \rho \Re_{m_{\mathsf{U}}}(\mathcal{F}_{1/\rho+u+B(\delta,m_{\mathsf{U}},\mathcal{F})}) + 2\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{L}}}} + 2\rho\sqrt{\frac{\log(2/\delta)}{2m_{\mathsf{U}}}} \end{aligned} \tag{49}$$

as claimed.

B.5. Proof of Example 3.7

The normalizing factor $\sum_{j=1}^c \mathrm{e}^{w_j^\mathsf{T} x}$ is maximized at $w_1 = x = [1,0,0,\dots,0]$ and $w_2 = \cdots = w_c = 0$ so that

$$\sum_{j=1}^c \mathrm{e}^{w_j^\mathsf{T} x} \le \mathrm{e} + (c-1) \le c+2 \tag{50}$$

This implies $\mathbb{P}_w(y_c) \ge (\mathrm{e}^{w_c^\mathsf{T} x})/(c+2)$. Therefore, $\mathbb{E}[\mathbb{P}_w(y_c)] \le t$ implies $t(c+2) \ge \mathbb{E}[\mathrm{e}^{w_c^\mathsf{T} x}] \ge \mathrm{e}^{\mathbb{E}[w_c^\mathsf{T} x]} = \mathrm{e}^{\alpha^\mathsf{T} w_c}$, or equivalently $\alpha^\mathsf{T} w_c \le \log(t(c+2))$. Therefore, given a set of data $S = \{x_i\}_{i=1}^m$ and Rademacher random variables ϵ, the inner supremum in the definition of Rademacher complexity can be upper bounded by solving the following program

$$\begin{aligned} \max \quad & \sum_{i=1}^m \sum_{j=1}^c \epsilon_{i,j} w_j^\mathsf{T} x_i \\ \text{s.t.} \quad & \sum_{j=1}^c w_j^\mathsf{T} w_j \le 1 \\ & \alpha^\mathsf{T} w_c \le \log(t(c+2)) \end{aligned} \tag{51}$$

Consider its Lagrangian

$$L(w,\lambda,\mu) = \sum_{i=1}^m \sum_{j=1}^c \epsilon_{i,j} w_j^\mathsf{T} x_i + \lambda\left(1 - \sum_{j=1}^n w_j^\mathsf{T} w_j\right) + \nu\left(\log(t(c+2)) - \alpha^\mathsf{T} w_c\right) \tag{52}$$

Denote $\xi_j := \sum_{i=1}^m \epsilon_{i,j} x_i$. The Lagrangian is then maximized at $w_j = \xi_j/(2\lambda)$ for $j < c$ and $w_c = (\xi_c - \nu\alpha)/(2\lambda)$. The dual function then writes:

$$g(\lambda,\nu) = \nu\log(t(c+2)) + \lambda + \sum_{j=1}^{c-1} \frac{\|\xi_j\|_2^2}{4\lambda} + \frac{\|\xi_c - \nu\alpha\|_2^2}{4\lambda} \ge \nu\log(t(c+2)) + \sqrt{\sum_{j=1}^{c-1} \|\xi_j\|_2^2 + \|\xi_c - \nu\alpha\|_2^2} \tag{53}$$

By weak duality, we have that

$$\widehat{\Re}_m(\mathcal{F}_t) \le \frac{1}{m}\mathbb{E}_\epsilon\left[\min_{\nu\ge 0}\left(\nu\log(t(c+2)) + \sqrt{\sum_{j=1}^{c-1} \|\xi_j\|_2^2 + \|\xi_c - \nu\alpha\|_2^2}\right)\right] \tag{54}$$

Assuming $t < 1/(c+2)$ so that $\log(t(c+2)) < 0$. We can upper bound (54) as

$$\frac{1}{m}\mathbb{E}_\epsilon\left[\min_{\nu\ge 0}\left(\sqrt{\sum_{j=1}^{c-1} \|\xi_j\|_2^2 + \|\xi_c - \nu\alpha\|_2^2}\right)\right] \tag{55}$$

The function $\sum_{j=1}^{c-1} \|\xi_j\|_2^2 + \|\xi_c - \nu\alpha\|_2^2$ is minimized at $\nu = 0$ if $\xi_c^\mathsf{T}\alpha \leq 0$ and $\nu = \xi_c^\mathsf{T}\alpha / \|\alpha\|_2^2$ otherwise. Denote the event $\xi_c^\mathsf{T}\alpha \leq 0$ as $\mathcal{E}$. By symmetry, we have that $\mathbb{P}(\mathcal{E}) = 1/2$ so that

$$\frac{1}{m}\mathbb{E}_\epsilon\left[\min_{\nu\geq 0}\left(\sqrt{\sum_{j=1}^{c-1}\|\xi_j\|_2^2 + \|\xi_c - \nu\alpha\|_2^2}\right)\right] = \frac{1}{2}\mathbb{E}_\epsilon\left[\sqrt{\sum_{j=1}^{c}\|\xi_j\|_2^2}\middle|\mathcal{E}\right] + \frac{1}{2}\mathbb{E}_\epsilon\left[\sqrt{\sum_{j=1}^{c}\|\xi_j\|_2^2 - \frac{(\xi_c^\mathsf{T}\alpha)^2}{\|\alpha\|_2^2}}\middle|\overline{\mathcal{E}}\right] \tag{56}$$

Again by symmetry, the quantity $(\xi_c^\mathsf{T}\alpha)^2$ is independent of $\mathcal{E}$. Therefore, by Jensen's inequality, we have that

$$\begin{aligned}
\mathbb{E}_{S,\epsilon}\left[\sqrt{\sum_{j=1}^{c}\|\xi_j\|_2^2 - \frac{(\xi_c^\mathsf{T}\alpha)^2}{\|\alpha\|_2^2}}\middle|\overline{\mathcal{E}}\right] &\leq \sqrt{\mathbb{E}_{S,\epsilon}\left[\sum_{j=1}^{c}\|\xi_j\|_2^2 - \frac{(\xi_c^\mathsf{T}\alpha)^2}{\|\alpha\|_2^2}\right]} \\
&\leq \sqrt{cm - \mathbb{E}_{S,\epsilon}\left[\frac{(\xi_c^\mathsf{T}\alpha)^2}{\|\alpha\|_2^2}\right]} \\
&= \sqrt{cm - \frac{\mathrm{Var}(\xi_c^\mathsf{T}\alpha)}{\|\alpha\|_2^2}} \\
&= \sqrt{cm - m\frac{\sigma^2\|\alpha\|_2^2 + \|\alpha\|_2^4}{\|\alpha\|_2^2}} \\
&= \sqrt{(c - \sigma^2 - \|\alpha\|_2^2)m}
\end{aligned} \tag{57}$$

Similarly, we can use Jensen's inequality to bound $\mathbb{E}_{S,\epsilon}\left[\sqrt{\sum_{j=1}^{c}\|\xi_j\|_2^2}\middle|\mathcal{E}\right] \leq \sqrt{cm}$. Putting these together, we have that

$$\Re_m(\mathcal{F}_t) = \mathbb{E}_x[\widehat{\Re}_m(\mathcal{F}_t)] \leq \frac{1}{2}\sqrt{\frac{c}{m}} + \frac{1}{2}\sqrt{\frac{c - \sigma^2 - \|\alpha\|_2^2}{m}} \tag{58}$$

C. Proofs from Section 4

C.1. Proof of Propostion 4.2

First, we show the Rademacher complexity of the singleton mapping is zero:

$$\begin{aligned}
\Re_m(\{(x,y) \mapsto -\mu v(x,y)\}) &= \frac{1}{m}\mathbb{E}_{x,\epsilon}\left[\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}} -\epsilon_{i,y}\mu v(x_i,y)\right] \\
&= \frac{1}{m}\mathbb{E}_{x}\left[\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}} -\mathbb{E}[\epsilon_{i,y}]\mu v(x_i,y)\right] \\
&= 0
\end{aligned} \tag{59}$$

Second, we use the linearity of Rademacher complexity to obtain the desired result.

$$\begin{aligned}
\Re_m(\mathcal{F}^\mu) &= \frac{1}{m}\mathbb{E}_{x,\epsilon}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}}\epsilon_{i,y}(f(x_i,y) - \mu v(x_i,y))\right] \\
&= \frac{1}{m}\mathbb{E}_{x,\epsilon}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}}\epsilon_{i,y}f(x_i,y)\right] + \frac{1}{m}\mathbb{E}_{x,\epsilon}\left[\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}} -\epsilon_{i,y}\mu v(x_i,y)\right] \\
&= \Re_m(\mathcal{F}) + \Re_m(\{(x,y)\mapsto -\mu v(x,y)\}) = \Re_m(\mathcal{F})
\end{aligned} \tag{60}$$

C.2. Proof of Proposition 4.3

(a) Given any scoring function f, let $Z_f^C(x) := \sum_{y \in C(x)} \exp(f(x,y))$ and $Z_f^{-C}(x) := \sum_{y \notin C(x)} \exp(f(x,y))$. We have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu}\Delta_{\mathrm{CE}}^{\mu}(f) &= \frac{\mathrm{d}}{\mathrm{d}\mu}\mathbb{E}\left[\log \frac{\exp(f(x,y_{\mathrm{ora}}) - \mu v(x,y_{\mathrm{ora}}))}{Z_f^C(x) + Z_f^{-C}(x)/\mathrm{e}^{\mu}}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\mu}\log \frac{\exp(f(x,y_{\mathrm{ora}}) - \mu v(x,y_{\mathrm{ora}}))}{Z_f^C(x) + Z_f^{-C}(x)/\mathrm{e}^{\mu}}\right] \\
&= \mathbb{E}\left[\frac{Z_f^{-C}(x)/\mathrm{e}^{\mu}}{Z_f^C(x) + Z_f^{-C}(x)/\mathrm{e}^{\mu}} - v(x,y_{\mathrm{ora}})\right] \\
&= V(f^{\mu}) - V_{\mathrm{ora}}
\end{aligned} \tag{61}$$

Moreover,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu}V(f^{\mu}) =& \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\mu}\frac{Z_{f^{\mu}}^{-C}(x)}{Z_{f^{\mu}}(x)}\right] \\
=& \mathbb{E}\left[\frac{Z_{f^{\mu}}(x)(-Z_{f^{\mu}}^{C}(x)) + (Z_{f^{\mu}}^{C}(x))^2}{(Z_{f^{\mu}}(x))^2}\right] \\
=& \mathbb{E}\left[\mathbb{P}_{f^{\mu}}^2(-C) - \mathbb{P}_{f^{\mu}}(-C)\right]
\end{aligned} \tag{62}$$

which is negative and bounded, implying $V(f^{\mu}) - V_{\mathrm{ora}}$ is decreasing and Lipschitz with μ. Therefore, there is a $\mu > 0$ such that $R_{\mathrm{CE}}(f^{\mu}) < R_{\mathrm{CE}}(f)$ if and only if the derivative is positive at $\mu = 0$, i.e., $V(f) > V_{\mathrm{ora}}$.

(b) By (61),

$$\begin{aligned}
\Delta_{\mathrm{CE}}^{\mu}(f) &= \int_0^{\mu} \left(V(f^t) - V_{\mathrm{ora}}\right)\mathrm{d}t \\
&= \mathbb{E}\left[\int_0^{\mu} \frac{Z_f^{-C}(x)/\mathrm{e}^t}{Z_f^C(x) + Z_f^{-C}(x)/\mathrm{e}^t}\mathrm{d}t\right] - \mu V_{\mathrm{ora}} \\
&\geq \mathbb{E}\left[\int_0^{\mu} \frac{Z_f^{-C}(x)/\mathrm{e}^t}{Z_f^C(x) + Z_f^{-C}(x)}\mathrm{d}t\right] - \mu V_{\mathrm{ora}} \\
&= (1 - \mathrm{e}^{-\mu})\mathbb{E}\left[\frac{Z_f^{-C}(x)}{Z_f^C(x) + Z_f^{-C}(x)}\right] - \mu V_{\mathrm{ora}} \\
&= (1 - \mathrm{e}^{-\mu})V(f) - \mu V_{\mathrm{ora}}
\end{aligned} \tag{63}$$

(c) If $V_{\mathrm{ora}} = 0$, we have

$$\begin{aligned}
\Delta_{\mathrm{CE}}^{\infty}(f) &= \int_0^{\infty} \mathbb{E}\left[\frac{Z_f^{-C}(x)/\mathrm{e}^t}{Z_f^C(x) + Z_f^{-C}(x)/\mathrm{e}^t}\right]\mathrm{d}t \\
&= \mathbb{E}\left[\int_0^{\infty} \frac{Z_f^{-C}(x)/\mathrm{e}^t}{Z_f^C(x) + Z_f^{-C}(x)/\mathrm{e}^t}\mathrm{d}t\right] \\
&= \mathbb{E}\left[\log\left(\frac{Z_f^C(x) + Z_f^{-C}}{Z_f^C}\right)\right] \\
&= V_{\mathrm{CE}}(f)
\end{aligned} \tag{64}$$

C.3. Proof of Corollary 4.5

Using Proposition 4.3 (b), this result follows by solving the following equation

$$(1 - \mathrm{e}^{-\mu})V(f) - \mu V_{\mathrm{ora}} \geq 0 \tag{65}$$

It is known that the solution to the inequaltiy $u \leq a + b\mathrm{e}^{cu}$ of u is $u \leq a - \frac{1}{c}W(-bc\mathrm{e}^{ac})$. Substituting $a = \eta = V(f)/V_{\mathrm{ora}} = -b$ and $c = -1$ yields the desired result:

$$\mu \leq W(-\eta/\mathrm{e}^{\eta}) + \eta \tag{66}$$

where the RHS is positive only when $\eta > 1$. A plot of this solution as a function of η is presented below in Figure 2.

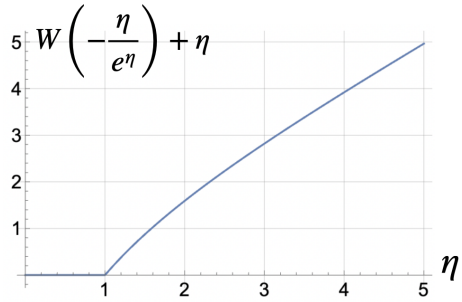

Figure 2. Choice of μ as a function of $\eta = V(f)/V_{\mathrm{ora}}$.

C.4. Proof of Proposition 4.6

This claim follows from the fact that $R_{\mathrm{CE}}(f^{\infty}) = R_{\mathrm{CE}}(f) - V_{\mathrm{CE}}(f)$ from Proposition 4.3 (c).

For equation (23), the first inequality follows from the optimality of f_{on}. For the second inequality, by definition we have

$$\begin{aligned} &R_{\mathrm{CE}}(f^{\infty}_{\mathrm{post}}) + V_{\mathrm{CE}}(f_{\mathrm{post}}) = R_{\mathrm{CE}}(f_{\mathrm{post}}) \leq R_{\mathrm{CE}}(f_{\mathrm{on}}) \\ \Rightarrow &R_{\mathrm{CE}}(f^{\infty}_{\mathrm{post}}) \leq R_{\mathrm{CE}}(f_{\mathrm{on}}) - V_{\mathrm{CE}}(f_{\mathrm{post}}) \leq R_{\mathrm{CE}}(f_{\mathrm{on}}) - \min_{f \in \mathcal{F}} V_{\mathrm{CE}}(f) \end{aligned} \tag{67}$$

D. Analysis for Hinge Loss and ℓ^1 Loss

D.1. Hinge Loss

The *margin* of a scoring function f at a sample (x, y_{ora}) is defined as

$$m(x, y_{\mathrm{ora}}, f) := \max_{y \in \mathcal{Y}} \{f(x, y)\} - f(x, y_{\mathrm{ora}}) \tag{68}$$

We denote its expectation as $M(f) = \mathbb{E}[m(x, y_{\mathrm{ora}}, f)]$.

Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the structured hinge loss (London et al., 2016; Meshi et al., 2019) is defined as the margin of the *loss augmented* scoring function $f + \ell : (x, y) \mapsto f(x, y) + \ell(y, y_{\mathrm{ora}})$. Namely,

$$L_{\text{hinge}}(x, y_{\mathrm{ora}}, f) := m(x, y_{\mathrm{ora}}, f + \ell) \tag{69}$$

Therefore, we can study the impact of constrained inference on the hinge loss via the impact on the margin. Let $\Delta^{\mu}_{\mathrm{margin}}(f) = M(f) - M(f^{\mu})$. We present the following result.

Theorem D.1 (Change of Margin)**.** *The following results hold:*

(a) For any fixed model f, there exists an $\mu_0 > 0$ such that $M(f^{\mu}) \leq M(f)$ only if

$$V_{01}(f) > V_{\mathrm{ora}} \tag{70}$$

where $V_{01}(f)$ is the zero-one style violation defined as $\mathbb{E}[\mathbb{1}\{\mathrm{argmax}_{y \in \mathcal{Y}} f(x, y) \neq y_{\mathrm{ora}}\}]$.

(b) In particular, if the constraint is noise-free, we have

$$\begin{aligned} \Delta^{\infty}_{\mathrm{margin}}(f) &= \mathbb{E}\left[\max_{y \in \mathcal{Y}} f(x, y) - \max_{y \in C(x)} f(x, y)\right] \\ &= \mathbb{E}\left[\left(\max_{y \notin C(x)} f(x, y) - \max_{y \in C(x)} f(x, y)\right)_+\right] \end{aligned} \tag{71}$$

Proof. (a) The derivative of the change of the margin is

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\Delta^{\mu}_{\mathrm{margin}}(f) = -\frac{\mathrm{d}}{\mathrm{d}\mu}M(f^{\mu}) = -\frac{\mathrm{d}}{\mathrm{d}\mu}\mathbb{E}\left[\max_{y\in\mathcal{Y}}\{f(x,y)-\mu v(x,y)\} - f(x,y_{\mathrm{ora}}) + \mu v(x,y_{\mathrm{ora}})\right] \\ = \mathbb{E}[v(x,y_{f^{\mu}}) - v(x,y_{\mathrm{ora}})] \tag{72}$$

where $y_{f^{\mu}} := \mathrm{argmax}_{y\in\mathcal{Y}}\{f(x,y)-\mu v(x,y)\}$ is the argmax inference output of CCM. Moreover, this derivative is non-increasing with μ. Therefore, a necessary condition for CCM to reduce the margin is

$$\mathbb{E}[v(x,y_f)] = V_{01}(f) > V_{\mathrm{ora}} \tag{73}$$

(b) This follows directly by taking the difference between $M(f)$ and $M(f^{\infty})$.

□

Remark D.2. *Due to the discontinuous nature of the argmax inference, the function $v(x,y_{f^{\mu}})$ is in general not continuous with μ. On the other hand, if we assume $\mu \mapsto \mathbb{E}[v(x,y_{f^{\mu}})]$ is Lipschitz continuous, the condition proposed in (a) is also sufficient, as in the analysis for cross-entropy.*

The impact of constrained inference on the hinge loss can be investigated by substituting f by $f+\ell$. For example, a sufficient for improving the average hinge loss will be $V_{01}(f+\ell) > V_{\mathrm{ora}}$.

The quantity $\left(\max_{y\notin C(x)} f(x,y) - \max_{y\in C(x)} f(x,y)\right)_+$ is closely related to the integrality loss *defined in Meshi et al. (2019). It is a hinge-stye surrogate loss function for the zero-one style violation function of f with argmax inference:*

$$\mathbb{P}\left\{\max_{y\notin C(x)} f(x,y) - \max_{y\in C(x)} f(x,y) \geq 0\right\} = V_{01}(f) \tag{74}$$

D.2. ℓ^1 Loss

To facilitate our discussion, we first present the following lemmas that will be useful in this section.

Lemma D.3 (Gradients of CCM)**.** *For any constraint C we have the following:*

1. *The derivative of the predicted probability is*

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\mathbb{P}_{f^{\mu}}(y|x) = \mathbb{P}_{f^{\mu}}(y)\left(\mathbb{P}_{f^{\mu}}(-C|x) - v(x,y)\right) \tag{75}$$

2. *The second order derivative of the probability is*

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\mathbb{P}_{f^{\mu}}(-C|x) = \mathbb{P}_{f^{\mu}}(y|x)\left(\left(\mathbb{P}_{f^{\mu}}(-C|x) - v(x,y)\right)^2 + \mathbb{P}^2_{f^{\mu}}(-C|x) - \mathbb{P}_{f^{\mu}}(-C|x)\right) \tag{76}$$

Proof. Recall that given any scoring function f, we denote

$$Z^{C}_{f}(x) := \sum_{y\in C(x)} \exp(f(x,y))$$

and

$$Z^{-C}_{f}(x) := \sum_{y\notin C(x)} \exp(f(x,y))$$

We also let $Z_f(x) = Z^{C}_{f}(x) + Z^{-C}_{f}(x)$.

(a) The pointwise derivative of CCM's l^1 risk with respect to μ is then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mu}\mathbb{P}_{f^{\mu}}(y|x) =& \frac{\mathrm{d}}{\mathrm{d}\mu}\frac{\mathrm{e}^{f(x,y)-\mu v(x,y)}}{Z_{f^{\mu}}(x)} \\ =& \frac{1}{\left(Z_{f^{\mu}}(x)\right)^2}\left(Z_{f^{\mu}}(x)(-v(x,y)\mathrm{e}^{f(x,y)-\mu v(x,y)}) + Z^{-C}_{f^{\mu}}(x)\mathrm{e}^{f(x,y)-\mu v(x,y)}\right) \\ =& \mathbb{P}_{f^{\mu}}(y)\left(\mathbb{P}_{f^{\mu}}(-C) - v(x,y)\right)\end{aligned} \tag{77}$$

where the second equality follows from the fact that $\frac{\mathrm{d}}{\mathrm{d}\mu}Z_{f^{\mu}}(x) = -Z^{-C}_{f^{\mu}}(x)$.

(b) Based on (a),

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}^2\mu}\mathbb{P}_{f^\mu}(y|x) &= \left(\mathbb{P}_{f^\mu}(y)\left(\mathbb{P}_{f^\mu}(-C)-v(x,y)\right)\right)\left(\mathbb{P}_{f^\mu}(-C)-v(x,y)\right)\\ &\quad+\mathbb{P}_{f^\mu}(y)\left(\mathbb{P}^2_{f^\mu}(-C)-\mathbb{P}_{f^\mu}(-C)\right)\\ &= \mathbb{P}_{f^\mu}(y|x)\left(\left(\mathbb{P}_{f^\mu}(-C|x)-v(x,y)\right)^2+\mathbb{P}^2_{f^\mu}(-C|x)-\mathbb{P}_{f^\mu}(-C|x)\right)\end{aligned} \tag{78}$$

□

Now we discuss the change in ℓ^1 risk that is defined as $\Delta^\mu(f) := R(f) - R(f^\mu)$.

Theorem D.4 (Change of ℓ^1 Risk)**.** *The following results hold:*

(a) For any fixed model f, there exists an $\mu_0 > 0$ such that $R(f^\mu) < R(f)$ if

$$\mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)] > \mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})v(x,y_{\mathrm{ora}})] \tag{79}$$

(b) The change of risk can be lower bounded by

$$\Delta^\mu(f) \geq \frac{1-\mathrm{e}^{-2\mu}}{2}\mathbb{E}_x[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)] - \mu V_{\mathrm{ora}} \tag{80}$$

(c) In particular, if the constraint is noise-free, we have

$$\Delta^\infty(f) \geq \mathbb{E}_x[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)] \tag{81}$$

Proof. (a) From Lemma D.3 (a) we know the derivative of the risk with respect to μ at $\mu = 0$ is

$$\mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)] - \mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})v(x,y_{\mathrm{ora}})] \tag{82}$$

Further, Lemma D.3 (b) implies this derivative is Lipschitz with respect to μ since for any μ,

$$\left|\mathbb{P}_{f^\mu}(y|x)\left(\left(\mathbb{P}_f(-C|x)-v(x,y)\right)^2+\mathbb{P}^2_{f^\mu}(-C|x)-\mathbb{P}_{f^\mu}(-C|x)\right)\right| \leq 1 \tag{83}$$

Therefore, a sufficient condition for the existence of an $\mu_0 > 0$ such that $R(f^\mu) < R(f)$ is that $\mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)] > \mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})v(x,y_{\mathrm{ora}})]$.

(b) First, we note for any y and μ that

$$\begin{aligned}\mathbb{P}_{f^\mu}(y)\mathbb{P}_{f^\mu}(-C) &= \frac{\mathrm{e}^{f(x,y)-\mu v(x,y)}Z_f^{-C}(x)/\mathrm{e}^\mu}{\left(Z_{f^\mu}(x)\right)^2}\\ &\geq \frac{\mathrm{e}^{f(x,y)-\mu v(x,y)}Z_f^{-C}(x)/\mathrm{e}^\mu}{\left(Z_f(x)\right)^2}\\ &\geq \frac{\mathrm{e}^{f(x,y)-\mu}Z_f^{-C}(x)/\mathrm{e}^\mu}{\left(Z_f(x)\right)^2}\\ &= \mathbb{P}_f(y)\mathbb{P}_f(-C)\mathrm{e}^{-2\mu}\end{aligned} \tag{84}$$

Also,

$$\mathbb{E}[\mathbb{P}_f(y_{\mathrm{ora}})v(x,y_{\mathrm{ora}})] \leq \mathbb{E}[v(x,y_{\mathrm{ora}})] = V_{\mathrm{ora}} \tag{85}$$

Integrating the derivative gives

$$\begin{aligned}\Delta^\mu(f) &\geq \int_0^\mu \mathbb{E}\left[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)\mathrm{e}^{-2t} - V_{\mathrm{ora}}\right]\mathrm{d}t\\ &= \frac{1-\mathrm{e}^{-2\mu}}{2}\mathbb{E}_x[\mathbb{P}_f(y_{\mathrm{ora}})\mathbb{P}_f(-C)] - \mu V_{\mathrm{ora}}\end{aligned} \tag{86}$$

(c) With noise-free constraints,

$$\begin{aligned}\mathbb{P}_{f^\mu}(y_{\text{ora}})\mathbb{P}_{f^\mu}(-C) &= \frac{\mathrm{e}^{f(x,y_{\text{ora}})}Z_f^{-C}(x)/\mathrm{e}^\mu}{\left(Z_{f^\mu}(x)\right)^2} \\ &\geq \frac{\mathrm{e}^{f(x,y_{\text{ora}})}Z_f^{-C}(x)/\mathrm{e}^\mu}{\left(Z_f(x)\right)^2} \\ &= \mathbb{P}_f(y_{\text{ora}})\mathbb{P}_f(-C)\mathrm{e}^{-\mu}\end{aligned} \tag{87}$$

Integrating both sides gives

$$\begin{aligned}\Delta^\mu(f) &\geq \int_0^\mu \mathbb{E}\left[\mathbb{P}_f(y_{\text{ora}})\mathbb{P}_f(-C)\mathrm{e}^{-t}\right]\mathrm{d}t \\ &= \mathbb{E}_x[\mathbb{P}_f(y_{\text{ora}})\mathbb{P}_f(-C)]\end{aligned} \tag{88}$$

□

The term $\mathbb{E}_x[\mathbb{P}_f(y_{\text{ora}})\mathbb{P}_f(-C)]$ plays a key role in these results, and it measures the average violation of the model f, weighted by the model's confidence of the true label. The first result shows that if this weighted average violation is larger than that of the true data distribution, then CCM is helpful. The last result shows that a model with a larger weighted violation obtains more benefits from strictly constrained inference.

E. Proofs from Section 5

E.1. Proof of Theorem 5.1

Recall $f_\star^\mu = \operatorname{argmin}_{g\in\mathcal{F}^\mu} R_{\text{CE}}(g) + \rho V_{\text{CE}}(g)$ is the optimal CCM for the regularized surrogate objective and f_{post} is the cross entropy risk minimizer in $\mathcal{F}$. According to our notation, f_{post}^μ is the constrained model with base model f_{post}. By this definition, we have

$$R_{\text{CE}}(f_\star^\mu) + \rho V_{\text{CE}}(f_\star^\mu) \leq R_{\text{CE}}(f_{\text{post}}^\mu) + \rho V_{\text{CE}}(f_{\text{post}}^\mu) \tag{89}$$

Therefore,

$$\begin{aligned}R_{\text{CE}}(f_\star^\mu) &\leq R_{\text{CE}}(f_{\text{post}}^\mu) + \rho(V_{\text{CE}}(f_{\text{post}}^\mu) - V_{\text{CE}}(f_\infty^\mu)) \\ &\leq R_{\text{CE}}(f_{\text{post}}^\mu) + \rho V_{\text{CE}}(f_{\text{post}}^\mu) \\ &\leq R_{\text{CE}}(f_{\text{post}}) - \Delta_{\text{CE}}^\mu(f_{\text{post}}) + \rho V_{\text{CE}}(f_{\text{post}}^\mu)\end{aligned} \tag{90}$$

Therefore, a sufficient condition for $R_{\text{CE}}(f_\star^\mu) \leq R_{\text{CE}}(f_{\text{post}})$ is that $\rho V_{\text{CE}}(f_{\text{post}}^\mu) < \Delta_{\text{CE}}^\mu(f_{\text{post}})$. Furthermore, recall for any scoring function f, we define $Z_f^C(x) := \sum_{y\in C(x)}\exp(f(x,y))$ and $Z_f^{-C}(x) := \sum_{y\notin C(x)}\exp(f(x,y))$. We then have

$$\begin{aligned}V_{\text{CE}}(f) - V_{\text{CE}}(f^\mu) &= \mathbb{E}\left[-\log\left(\frac{Z_f^C(x)}{Z_f^C(x)+Z_f^{-C}(x)}\right)\right] - \mathbb{E}\left[-\log\left(\frac{Z_f^C(x)}{Z_f^C(x)+Z_f^{-C}(x)/\mathrm{e}^\mu}\right)\right] \\ &= \mathbb{E}\left[-\log\left(\frac{Z_f^C(x)+Z_f^{-C}(x)/\mathrm{e}^\mu}{Z_f^C(x)+Z_f^{-C}(x)}\right)\right] \\ &= \int_0^\mu \mathbb{E}\left[\frac{Z_f^{-C}(x)/\mathrm{e}^t}{Z_f^C(x)+Z_f^{-C}(x)/\mathrm{e}^t}\right]\mathrm{d}t \\ &= \Delta_{\text{CE}}^\mu(f) + \mu V_{\text{ora}} \qquad \text{(compare to equation (61))}\end{aligned} \tag{91}$$

Therefore, $\Delta_{\text{CE}}^\mu(f_{\text{post}}) = V_{\text{CE}}(f_{\text{post}}) - V_{\text{CE}}(f_{\text{post}}^\mu) - \mu V_{\text{ora}}$. So, the sufficient condition can be reformulated as

$$\rho < \frac{V_{\text{CE}}(f_{\text{post}}) - V_{\text{CE}}(f_{\text{post}}^\mu) - \mu V_{\text{ora}}}{V_{\text{CE}}(f_{\text{post}}^\mu)} \tag{92}$$

E.2. Proof of Theorem 5.2

We have seen in Theorem 4.3 that for any scoring function f, there is a $\mu > 0$ such that $R_{\mathrm{CE}}(f^{\mu}) < R_{\mathrm{CE}}(f)$ if and only if $V(f) \geq V_{\mathrm{ora}}$. On the other hand, we know from Lemma 3.3 that

$$V(f_{\rho}) \leq V(f_{\infty}) + \frac{1}{\rho} \tag{93}$$

Therefore, if

$$\rho \geq \frac{1}{V_{\mathrm{ora}} - V(f_{\infty})} \tag{94}$$

we must have $V(f_{\rho}) \leq V_{\mathrm{ora}}$, which implies there is no $\mu > 0$ such that $R_{\mathrm{CE}}((f_{\rho})^{\mu}) < R_{\mathrm{CE}}(f_{\rho})$.