Bounding the Integrality Distance of LP Relaxations for Structured Prediction

Ben London
blondon@amazon.com

Ofer Meshi
meshi@google.com

Adrian Weller
aw665@cam.ac.uk

1 Introduction

In structured prediction, a predictor optimizes an objective function over a combinatorial search space, such as the set of all image segmentations, or the set of all part-of-speech taggings. Unfortunately, finding the optimal structured labeling—sometimes referred to as *maximum a posteriori* (MAP) inference—is, in general, NP-hard [12], due to the combinatorial structure of the problem. Many inference approximations have been proposed, some of which are based on *linear programming (LP) relaxations* [e.g., 5, 1, 15, 16, 4, 13], which "relax" the combinatorial search space to a convex polytope with a polynomial number of constraints. These LP relaxations can be solved efficiently, but may result in *fractional* solutions, i.e., non-integral labelings. The approximation quality of an LP relaxation is traditionally measured by a quantity known as the *integrality gap*, defined as the difference in the objective values obtained at the optima of the relaxed and exact problems. When the integrality gap is zero, the solution is said to be *tight*.

While studying the integrality gap is useful from an optimization perspective, it is arguably less important in structured prediction, wherein the solution to the optimization, the inferred labeling, is more important than its objective value. We do not really care whether the optimum of the relaxed problem equals that of the integral one; we just want relaxed inference to yield the optimal integral assignment—or, lacking that, an assignment that is "close to" the optimal integral one. If we assume that the relaxed problem has a unique solution, then tightness implies that the assignments are the same. However, lacking this assumption, there may be multiple, disparate solutions, so the assignments may differ. Further, when the integrality gap is nonzero, we know nothing about the distance between the relaxed and exact assignments.

We therefore propose an alternate measure of approximation quality based on the Manhattan distance between the maximizers of the relaxed and exact optimizations, which we refer to as the *integrality distance*. The integrality distance is conceptually similar to *persistence* (see [14] for definition) in that a persistent fractional solution will have a subset of variables with zero integrality distance. The integrality distance is also related to the integrality gap, although the distance is arguably more intuitive and informative: when the integrality distance is small, the relaxed solution is close to the exact solution, which is what we ultimately care about; moreover, when the integrality distance is zero, the integrality gap must also be zero, regardless of whether we assume uniqueness.

In this paper, we examine the integrality distance in the context of learning, asking the question: if a predictor is trained to use relaxed inference, what is its expected integrality distance? We begin by relating the integrality distance to several structured loss functions that are commonly analyzed in the literature on structured prediction. We then show that the integrality distance generalizes from an empirical sample to the population average. This result builds on recent work by Meshi et al. [9], who showed that the probability of tightness generalizes. We take our analysis one step further than Meshi et al.[1] and show that the integrality distance is upper-bounded by a constant multiple of the structured hinge loss—a convex loss function that is commonly used in practice. Combining these results, we obtain a high-probability bound on the expected integrality distance that can be evaluated

[1] Though Meshi et al. related the integrality gap to the hinge loss, they did not directly relate tightness to a tractable metric.

NIPS workshop on Optimization for Machine Learning. Barcelona, Spain, 2016.

from the training data, and whose additive error decreases with the number of examples. Further, a simple argument shows how this bound applies to an integral rounding of a fractional solution. Thus, our novel theory proves that max-margin training not only minimizes prediction error, but also the approximation error of relaxed inference.

2 Preliminaries

Suppose we wish to learn the weights, $\mathbf{w} \in \mathbb{R}^d$, of a structured predictor,

$$\arg\max_{\mathbf{y} \in \mathcal{Y}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}), \tag{1}$$

where $\mathbf{x} \in \mathcal{X}$ is a vector of observations, $\mathbf{y} \in \mathcal{Y}$, is a vector of discrete labels, for some $\mathcal{Y} \triangleq \prod_{i=1}^{n} \mathcal{Y}_i$, and $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a vector-valued feature mapping that decomposes over a set of factors, $\mathcal{F}$, such that $\mathbf{f}(\mathbf{x}, \mathbf{y}) = (f(\mathbf{x}, \mathbf{y}))_{f \in \mathcal{F}}$. We assume that $\mathcal{F}$ contains a *unary* factor for each $i = 1, \ldots, n$. The optimization in Equation 1 is equivalent to MAP inference in a Markov network. Using the *marginal polytope*, $\mathcal{M}$ (see Wainwright and Jordan [14] for a definition), we can reformulate Equation 1 as a linear program,

$$\arg\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \boldsymbol{\mu}, \tag{2}$$

where $\boldsymbol{\theta}(\mathbf{x}; \mathbf{w})$ is a vector of *potential functions*. Note that $\mathcal{M}$ contains non-integral vectors, but a solution to Equation 2, denoted $\boldsymbol{\mu}_{\text{I}}(\mathbf{x}; \mathbf{w})$, is always a vertex of $\mathcal{M}$, and is therefore always integral. Thus, a straightforward decoding of $\boldsymbol{\mu}_{\text{I}}(\mathbf{x}; \mathbf{w})$ produces a discrete labeling. With a slight abuse of notation, we assume that $\mathbf{f}(\mathbf{x}, \boldsymbol{\mu})$ exists and $\boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \boldsymbol{\mu} = \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \boldsymbol{\mu})$. When $\mathbf{x}$ and $\mathbf{w}$ are clear from context, we will omit them from our notation.

If describing $\mathcal{M}$ is intractable, we can replace $\mathcal{M}$ with the *local marginal polytope*, $\mathcal{M}_{\text{L}}$, which is described by a polynomial number of local constraints. For example, the local marginal polytope of a pairwise model, containing unary (node) and pairwise (edge) factors, indexed by $\mathcal{V}$ and $\mathcal{E}$ respectively, is given by:

$$\mathcal{M}_{\text{L}}^{\text{pairwise}} \triangleq \left\{ \boldsymbol{\mu} \ : \begin{array}{cc} \forall v \in \mathcal{V}, & \sum_{j=1}^{|\mathcal{Y}_v|} \mu_v^j = 1 \ ; \\ \forall e = \{u, v\} \in \mathcal{E}, & \sum_{i=1}^{|\mathcal{Y}_u|} \mu_e^{ij} = \mu_v^j \ , \\ & \sum_{j=1}^{|\mathcal{Y}_v|} \mu_e^{ij} = \mu_u^i \ . \end{array} \right\} \tag{3}$$

Maximizing over $\mathcal{M}_{\text{L}}$ is an LP relaxation of the original combinatorial optimization. Since $\mathcal{M}_{\text{L}}$ is an outer bound on $\mathcal{M}$, with no new integer vertices, the relaxed solution, denoted $\boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w})$, may be fractional. The *integrality gap* is $\boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w}) - \boldsymbol{\mu}_{\text{I}}(\mathbf{x}; \mathbf{w}))$. We will not assume that either the integral or relaxed optimizations have unique solutions, but we will assume that there exists some deterministic tie-breaking mechanism to select a single optimum. Thus, $\boldsymbol{\mu}_{\text{I}}(\mathbf{x}; \mathbf{w})$ and $\boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w})$ always output a single solution (not a set), and we define the *integrality distance* as

$$\left\| \boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w}) - \boldsymbol{\mu}_{\text{I}}(\mathbf{x}; \mathbf{w}) \right\|_1 . \tag{4}$$

3 Structured Loss Functions

We will focus on the integrality distance of the unary factors, denoted $\boldsymbol{\mu}_{\text{u}} \triangleq (\mu_i)_{i=1}^{n}$, since they are sufficient for decoding a labeling, $\mathbf{y}$. Let

$$D_1(\boldsymbol{\mu}, \boldsymbol{\mu}') \triangleq \frac{1}{2n} \left\| \boldsymbol{\mu}_{\text{u}} - \boldsymbol{\mu}'_{\text{u}} \right\|_1 \tag{5}$$

denote the normalized Manhattan distance between unary factors. Note that when one of the inputs—say, $\boldsymbol{\mu}$—is integral, $D_1(\boldsymbol{\mu}, \boldsymbol{\mu}')$ can be expressed as $\delta(\boldsymbol{\mu}) \cdot \boldsymbol{\mu}'$, where

$$\delta(\boldsymbol{\mu}) \triangleq \frac{1}{n} \begin{bmatrix} \mathbf{1} - \boldsymbol{\mu}_{\text{u}} \\ \mathbf{0} \end{bmatrix} . \tag{6}$$

(Padding this vector with zeros makes its dot product with $\boldsymbol{\mu}$ discard higher-order factors.) Further, when both inputs are integral, D_1 is equivalent to the normalized Hamming distance.

Given a model, $\mathbf{w}$, an input, $\mathbf{x}$, and an assignment, $\boldsymbol{\mu}$, let

$$L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}) \triangleq D_1\left(\boldsymbol{\mu}, \boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w})\right) \tag{7}$$

denote the *L1 loss*. This loss function is a relaxation of the the Hamming loss, which is commonly used to measure the prediction error of exact inference. If the third argument is a reference (i.e., "ground truth") labeling, $\boldsymbol{\mu}_{\text{T}}$, then $L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{T}})$ measures the prediction error of approximate inference. However, if the third argument is the exact, integral MAP state, $\boldsymbol{\mu}_{\text{I}}$, then $L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}})$ is the normalized integrality distance. This latter quantity is what we will focus on upper-bounding.

Let

$$L_{\text{h}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}) \triangleq \max_{\boldsymbol{\mu}' \in \mathcal{M}_{\text{L}}} D_1(\boldsymbol{\mu}, \boldsymbol{\mu}') + \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot (\boldsymbol{\mu}' - \boldsymbol{\mu}) \,. \tag{8}$$

denote a loss function commonly referred to as the (relaxed) *structured hinge loss*. The structured hinge loss is minimized when $\boldsymbol{\mu}$ scores higher than all alternate assignments, $\boldsymbol{\mu}'$, by a margin that is at least $D_1(\boldsymbol{\mu}, \boldsymbol{\mu}')$. Note that when $\boldsymbol{\mu}$ is the exact MAP state, the hinge loss computes a *loss-augmented* integrality gap, using Manhattan distance for loss augmentation.

A related loss function is the (relaxed) *structured ramp loss*,

$$L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}) \triangleq \max_{\boldsymbol{\mu}' \in \mathcal{M}_{\text{L}}} D_1(\boldsymbol{\mu}, \boldsymbol{\mu}') + \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot (\boldsymbol{\mu}' - \boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w})) \,, \tag{9}$$

which can be considered a normalized version of the hinge loss. L_{r} is bounded by $[0, 1]$, whereas L_{h} might be unbounded (depending on the features and weights).

The hinge loss is often used in max-margin training, since it is convex in $\mathbf{w}$. The ramp loss is not convex in $\mathbf{w}$, but it is bounded, Lipschitz, and has a convenient relationship to the L1 (or Hamming) and hinge losses:

$$L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}) \leq L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}) \leq L_{\text{h}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}). \tag{10}$$

Thus, the ramp loss is often used as an analytical tool to derive generalization bounds, such as those that follow.

4 Generalization Bound

Generalization analysis bounds the maximum discrepancy between the expected loss on a random example and an empirical estimate of the loss from a random set of (training) examples. Typically, one is interested in upper-bounding the expected prediction error, but the loss function can in fact measure any quantity—such as the integrality distance. The following theorem states that the average integrality distance on a training sample generalizes to the population average. The interested reader will note that the proof (deferred to Appendix A.1) uses a PAC-Bayesian analysis, similar to London et al. [8], though the main result is stated for a deterministic predictor.

Theorem 1. *Let $\mathbb{D}$ denote a distribution over $\mathcal{X}$. Let $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ denote a feature mapping such that $\sup_{\mathbf{x},\mathbf{y}} \|\mathbf{f}(\mathbf{x}, \mathbf{y})\|_2 \leq B$, for some finite constant, $B < \infty$. Then, for any $\delta \in (0, 1)$ and $m \geq 1$, with probability at least $1 - \delta$ over draws of $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}) \in \mathcal{X}^m$, according to $\mathbb{D}^m$, every weight vector, $\mathbf{w}$, with $\|\mathbf{w}\|_2 \leq R < \infty$, satisfies*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{D}}[L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] \leq \frac{1}{m} \sum_{i=1}^{m} L_r(\mathbf{w}, \mathbf{x}^{(i)}, \boldsymbol{\mu}_{\text{I}}^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d \ln(mBR) + \ln \frac{2}{\delta}}{2m}}. \tag{11}$$

Further, Equation 11 holds when L_r is replaced with L_h.

Theorem 1 says that the expected integrality distance on a random instance is upper-bounded by the average integrality ramp (or hinge) loss on the training set, plus two terms that vanish as the number of training examples grows. Thus, the more training data we have, the better we can estimate the expected integrality distance at test time.

Remark 1. There is nothing special about $\boldsymbol{\mu}_{\text{I}}$ to Theorem 1. Indeed, we could use any integral assignment as a reference labeling for the loss functions and the proof would be the same. For example, we could replace $\boldsymbol{\mu}_{\text{I}}$ with $\boldsymbol{\mu}_{\text{T}}$ (a reference labeling) and obtain a risk bound for learning with approximate inference, which is a well-studied topic [e.g., 7, 8].

5 Relationship to Max-margin Training

In practice, computing the integrality loss is infeasible, since it requires exact inference. Therefore, the upper bound in Equation 11 cannot be evaluated. However, the empirical hinge loss with respect to the true labels *can* be evaluated efficiently. In this section, we show how minimizing this quantity actually minimizes the integrality distance. That is, max-margin training with approximate inference—which is something people do anyway to learn graphical models—reduces not only the prediction error, but also the inference approximation error.

The key insight that enables this result comes from the following technical lemma.

Lemma 1. *For any* $\mathbf{w}$ *and* $\mathbf{x}$*, if* $\boldsymbol{\mu}_{\text{T}}$ *is the reference (ground truth) labeling of* $\mathbf{x}$ *and* $\boldsymbol{\mu}_{\text{I}}$ *is the exact MAP state under* $\mathbf{w}$*, then*

$$L_h(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}}) \;\leq\; 2\, L_h(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{T}}), \tag{12}$$

meaning the integrality hinge loss is at most twice the hinge loss with respect to the true labeling.

The proof (given in Appendix A.2) relies on the optimality of the fractional solution and the triangle inequality. Note that Lemma 1 also yields an upper bound on the integrality ramp loss, since it is upper-bounded by the integrality hinge loss.

We can now prove the following corollary of Theorem 1.

Corollary 1. *Let* $\mathbb{D}$ *denote a distribution over a labeled example space,* $\mathcal{X} \times \mathcal{Y}$*. For a reference labeling,* $\mathbf{y}$*, denote its corresponding marginal vector by* $\boldsymbol{\mu}_{\text{T}}$*. Let* $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ *denote a feature mapping such that* $\sup_{\mathbf{x},\mathbf{y}} \|\mathbf{f}(\mathbf{x}, \mathbf{y})\|_2 \leq B$*, for some finite constant,* $B < \infty$*. Then, for any* $\delta \in (0, 1)$ *and* $m \geq 1$*, with probability at least* $1 - \delta$ *over draws of* $((\mathbf{x}^{(1)}, \boldsymbol{\mu}_{\text{T}}^{(1)}), \ldots, (\mathbf{x}^{(m)}, \boldsymbol{\mu}_{\text{T}}^{(m)}))$ *according to* $\mathbb{D}^m$*, every weight vector,* $\mathbf{w}$*, with* $\|\mathbf{w}\|_2 \leq R < \infty$*, satisfies*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{D}}[L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] \leq \frac{2}{m} \sum_{i=1}^{m} L_h(\mathbf{w}, \mathbf{x}^{(i)}, \boldsymbol{\mu}_{\text{T}}^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d \ln(mBR) + \ln \frac{2}{\delta}}{2m}}. \tag{13}$$

Corollary 1 says that max-margin training with relaxed inference directly minimizes the integrality distance on future examples. If the constants B and R are known, then this bound can be efficiently evaluated from training data.

6 Rounding a Fractional Solution

When the solution to an LP relaxation is fractional, we often round the solution to an integral assignment. Rounding schemes have been studied extensively [e.g., 10, 6, 2, 11]. Arguably, the simplest method is to select the local assignments with the highest values, which is equivalent to decoding the approximate *max-marginals*. One question that arises is how far the rounding, denoted $\boldsymbol{\mu}_{\text{R}}(\mathbf{x}; \mathbf{w})$, is from the exact solution; once this relationship is determined, one can apply our prior generalization analysis to the rounding. It turns out that the distance from $\boldsymbol{\mu}_{\text{R}}$ to $\boldsymbol{\mu}_{\text{I}}$ can be upper-bounded by a multiple of the integrality distance.

Lemma 2. *Suppose that every output variable has the same domain—i.e.,* $\mathcal{Y}_1 = \mathcal{Y}_2 = \ldots = \mathcal{Y}_n$*—and that each domain has size* k*. If* $\boldsymbol{\mu}_{\text{R}}(\mathbf{x}; \mathbf{w})$ *is the rounding of the fractional solution,* $\boldsymbol{\mu}_{\text{L}}(\mathbf{x}; \mathbf{w})$*, then*

$$D_1(\boldsymbol{\mu}_{\text{R}}, \boldsymbol{\mu}_{\text{I}}) \;\leq\; k\, D_1(\boldsymbol{\mu}_{\text{L}}, \boldsymbol{\mu}_{\text{I}}). \tag{14}$$

The proof is given in Appendix A.3. Lemma 2 can be combined with Corollary 1 to generate bounds on the expected integrality distance of rounding; the bound simply scales by k.

7 Discussion

We have introduced a new measure of approximation quality for LP-relaxed inference, which we call the integrality distance. We have shown that the average integrality distance generalizes from an empirical sample to the population, and that it is minimized by performing max-margin training. Interestingly, our results hold for any outer bound on the marginal polytope, and any LP solver. Thus, as the number of training examples grows, all LP relaxations that minimize the empirical hinge loss are, in terms of expected integrality distance, equally accurate.

References

[1] C. Chekuri, S. Khanna, J. Naor, and L. Zosin. Approximation algorithms for the metric labeling problem via a new linear programming formulation. In *Symposium on Discrete Algorithms*, 2001.

[2] C. Chekuri, S. Khanna, J. Naor, and L. Zosin. A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM Journal on Discrete Mathematics*, 18(3):608–625, 2005.

[3] M. Donsker and S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[4] A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Neural Information Processing Systems*, 2008.

[5] E. Santos Jr. On the generation of alternative explanations with implications for belief revision. In *Uncertainty in Artificial Intelligence*, 1991.

[6] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *Journal of the ACM*, 49(5): 616–639, 2002.

[7] A. Kulesza and F. Pereira. Structured learning with approximate inference. In *Neural Information Processing Systems*, 2007.

[8] B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *Journal of Machine Learning Research*, 17, 2016.

[9] O. Meshi, M. Mahdavi, A. Weller, and D. Sontag. Train and test tightness of LP relaxations in structured prediction. In *International Conference on Machine Learning*, 2016.

[10] P. Raghavan and C. Tompson. Randomized rounding: A technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374, 1987.

[11] P. Ravikumar, A. Agarwal, and M. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *Journal of Machine Learning Research*, 11:1043–1080, 2010.

[12] S. Shimony. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, 68(2): 399–410, 1994.

[13] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *Uncertainty in Artificial Intelligence*, 2008.

[14] M. Wainwright and M. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.

[15] M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory*, 51(11):3697–3717, 2005.

[16] Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Uncertainty in Artificial Intelligence*, 2007.

A Supplemental Material

The following appendices are provided to supplement the paper.

A.1 Proof of Theorem 1

To prove Theorem 1, we will use the following PAC-Bayes bound. (There are other PAC-Bayes bounds in the literature on structured prediction, but we prefer the following for its simplicity.)

Lemma 3. *Let $\mathbb{D}$ denote a distribution over an instance space, $\mathcal{Z}$. Let $\mathcal{H}$ denote a hypothesis class. Let $L : \mathcal{H} \times \mathcal{Z} \to [0,1]$ denote a bounded loss function. Let $\mathbb{P}$ denote a fixed prior distribution over $\mathcal{H}$. Then, for any $\delta \in (0,1)$ and $m \geq 1$, with probability at least $1-\delta$ over draws of $(Z^{(1)}, \ldots, Z^{(m)}) \in \mathcal{Z}^m$, according to $\mathbb{D}^m$, every posterior distribution, $\mathbb{Q}$, over $\mathcal{H}$, satisfies*

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z)] \leq \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z^{(i)})] \;+\; 2\sqrt{\frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}}{2m}} \tag{15}$$

Proof. To simplify notation, let

$$\varphi(h, Z^{(i)}) \triangleq \frac{1}{m}\left(\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}[L(h,Z)] - L(h,Z^{(i)})\right).$$

For any free parameter, $\epsilon \in \mathbb{R}$, observe that

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z)] - \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z^{(i)})] = \frac{1}{\epsilon} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\sum_{i=1}^{m} \epsilon\,\varphi(h,Z^{(i)})\right]. \tag{16}$$

The next step uses Donsker and Varadhan's [1975] *change of measure* inequality, which states that, if X is a random variable taking values in Ω, then for any two distributions, $\mathbb{P}$ and $\mathbb{Q}$, on Ω,

$$\mathop{\mathbb{E}}_{X\sim\mathbb{Q}}[X] \leq D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{X\sim\mathbb{P}}\left[e^X\right].$$

Applying change of measure to the righthand side of Equation 16, we have

$$\frac{1}{\epsilon} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\sum_{i=1}^{m} \epsilon\,\varphi(h,Z^{(i)})\right] \leq \frac{1}{\epsilon}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\exp\left(\sum_{i=1}^{m} \epsilon\,\varphi(h,Z^{(i)})\right)\right]\right). \tag{17}$$

By Markov's inequality, with probability $1-\delta$ over draws of a training set, $\mathbf{Z} \triangleq (Z^{(1)}, \ldots, Z^{(m)})$, according to $\mathbb{D}^m$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\exp\left(\sum_{i=1}^{m} \epsilon\,\varphi(h,Z^{(i)})\right)\right] &\leq \frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}^m} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\exp\left(\sum_{i=1}^{m} \epsilon\,\varphi(h,Z^{(i)})\right)\right] \\
&= \frac{1}{\delta} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}^m}\left[\prod_{i=1}^{m} \exp\left(\epsilon\,\varphi(h,Z^{(i)})\right)\right] \\
&= \frac{1}{\delta} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \prod_{i=1}^{m} \mathop{\mathbb{E}}_{Z^{(i)}\sim\mathbb{D}}\left[\exp\left(\epsilon\,\varphi(h,Z^{(i)})\right)\right].
\end{aligned} \tag{18}$$

In the last line, we leveraged the fact that the expectation of a product of i.i.d. random variables (in this case, $\varphi(h, Z^{(i)})$) is the product of their expectations. To upper-bound each expectation, we use Hoeffding's inequality, which states that if X is a zero-mean random variable, such that $a \leq X \leq b$ almost surely, then, for all $\epsilon \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\epsilon X}\right] \leq \exp\left(\frac{\epsilon^2(b-a)^2}{8}\right).$$

Note that $\varphi(h, Z^{(i)})$ has mean zero, and

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}[L(h,Z)] - \frac{1}{m} \leq \varphi(h,Z^{(i)}) \leq \mathop{\mathbb{E}}_{Z\sim\mathbb{D}}[L(h,Z)] - 0.$$

Therefore,

$$\mathop{\mathbb{E}}_{Z^{(i)}\sim\mathbb{D}}\left[\exp\left(\epsilon\,\varphi(h,Z^{(i)})\right)\right] \leq \exp\left(\frac{\epsilon^2}{8m^2}\right). \tag{19}$$

Combining Equations 16 to 19, we have for any $\epsilon \in \mathbb{R}$, with probability at least $1-\delta$,

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z)] - \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z^{(i)})] \leq \frac{1}{\epsilon}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta}\right) + \frac{\epsilon}{8m}. \tag{20}$$

What remains is to optimize ϵ for all posteriors simultaneously. To do so, we define an infinite sequence of values,

$$\forall j = 0,1,2,\ldots,\ \epsilon_j \triangleq 2^j\sqrt{8m\ln\frac{2}{\delta}}. \tag{21}$$

For each ϵ_j, we assign $\delta_j \triangleq \delta 2^{-(j+1)}$ probability to the probability that Equation 20 does not hold, substituting (ϵ_j,δ_j) for (ϵ,δ). Thus, with probability at least $1-\sum_{j=0}^{\infty}\delta_j = 1-\delta\sum_{j=0}^{\infty}2^{-(j+1)} = 1-\delta$, all $j=0,1,2,\ldots$ satisfy

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z)] - \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z^{(i)})] \leq \frac{1}{\epsilon_j}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_j}\right) + \frac{\epsilon_j}{8m}$$

For any given posterior, $\mathbb{Q}$, we choose an index, $j^\star$, by taking

$$j^\star \triangleq \left\lfloor \frac{1}{2\ln 2}\ln\left(\frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)}+1\right)\right\rfloor,$$

which implies

$$\sqrt{2m\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)} \leq \epsilon_{j^\star} \leq \sqrt{8m\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)}.$$

We further have (from London et al. [8]) that

$$D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}} \leq \frac{3}{2}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right).$$

Thus, with probability at least $1-\delta$, all posteriors satisfy

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z)] - \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[L(h,Z^{(i)})] \\
&\quad\leq \frac{1}{\epsilon_{j^\star}}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}}\right) + \frac{\epsilon_{j^\star}}{8m} \\
&\quad\leq \frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(1/\delta_{j^\star})}{\sqrt{2m(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}} + \frac{\sqrt{8m(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}}{8m} \\
&\quad\leq \frac{3\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\ln(2/\delta)\right)}{2\sqrt{2m(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}} + \frac{\sqrt{8m(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}}{8m} \\
&\quad= 2\sqrt{\frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta)}{2m}},
\end{aligned}$$

which completes the proof. □

We are now ready to prove Theorem 1. Let $\mathbb{P}$ denote a uniform prior over $\{\mathbf{w}\in\mathbb{R}^d : \|\mathbf{w}\|_2 \leq R\}$. Given a learned weight vector, $\mathbf{w}$, we construct a posterior, $\mathbb{Q}$, as a uniform distribution over $\{\mathbf{w}'\in\mathbb{R}^d : \|\mathbf{w}'\|_2 \leq R,\ \|\mathbf{w}'-\mathbf{w}\|_2 \leq 2/(mB)\}$. It can easily be shown that $D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) \leq d\ln(mBR)$.

The next part of the proof "derandomizes" the loss functions in Equation 15, replacing the randomized hypothesis, with weights $\mathbf{w}'\sim\mathbb{Q}$, with the deterministic predictor based on $\mathbf{w}$. To do so, we will bound the difference $|L_{\text{r}}(\mathbf{w},\mathbf{x},\boldsymbol{\mu}_{\text{I}}) - L_{\text{r}}(\mathbf{w}',\mathbf{x},\boldsymbol{\mu}_{\text{I}})|$. Note that $\mathbf{w}'$ is being evaluated with

respect to the MAP assignment under $\mathbf{w}$, i.e., $\boldsymbol{\mu}_{\text{I}} \in \arg\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \boldsymbol{\mu}$. To simplify notation, we will use the following shorthand:

$$\boldsymbol{\theta} \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \quad \text{and} \quad \boldsymbol{\theta}' \triangleq \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}')\,;$$

$$\boldsymbol{\mu}_{\text{L}} \in \operatorname*{arg\,max}_{\boldsymbol{\mu} \in \mathcal{M}_{\text{L}}} \boldsymbol{\theta} \cdot \boldsymbol{\mu} \quad \text{and} \quad \boldsymbol{\mu}'_{\text{L}} \in \operatorname*{arg\,max}_{\boldsymbol{\mu} \in \mathcal{M}_{\text{L}}} \boldsymbol{\theta}' \cdot \boldsymbol{\mu}\,;$$

$$\tilde{\boldsymbol{\mu}}_{\text{L}} \in \operatorname*{arg\,max}_{\boldsymbol{\mu} \in \mathcal{M}_{\text{L}}} (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \boldsymbol{\mu} \quad \text{and} \quad \tilde{\boldsymbol{\mu}}'_{\text{L}} \in \operatorname*{arg\,max}_{\boldsymbol{\mu} \in \mathcal{M}_{\text{L}}} (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \boldsymbol{\mu}.$$

We then have that the difference of ramp losses decomposes as

$$\begin{aligned}
&|L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}}) - L_{\text{r}}(\mathbf{w}', \mathbf{x}, \boldsymbol{\mu}_{\text{I}})| \\
&\quad = \left|((\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} - \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\text{L}}) \;-\; \left((\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}'_{\text{L}} - \boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}}\right)\right| \\
&\quad \leq \left|(\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} - (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}'_{\text{L}}\right| \qquad (22) \\
&\qquad + \left|\boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}} - \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\text{L}}\right|. \qquad (23)
\end{aligned}$$

We will upper-bound Equations 22 and 23 separately.

Starting with Equation 23, assume that $\boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}} \geq \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\text{L}}$. (If the inequality goes in the other direction, we simply swap the left and right terms, which is equivalent inside the absolute value.) We then have that

$$\left|\boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}} - \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\text{L}}\right| = \boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}} - \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\text{L}} \leq \boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}} - \boldsymbol{\theta} \cdot \boldsymbol{\mu}'_{\text{L}} = (\mathbf{w}' - \mathbf{w}) \cdot \mathbf{f}(\mathbf{x}, \boldsymbol{\mu}'_{\text{L}}),$$

due to the optimality of $\boldsymbol{\mu}_{\text{L}}$ for $\boldsymbol{\theta}$. Then, using Cauchy-Schwarz,

$$(\mathbf{w}' - \mathbf{w}) \cdot \mathbf{f}(\mathbf{x}, \boldsymbol{\mu}'_{\text{L}}) \leq \|\mathbf{w}' - \mathbf{w}\|_2 \, \|\mathbf{f}(\mathbf{x}, \boldsymbol{\mu}'_{\text{L}})\|_2 \leq \|\mathbf{w}' - \mathbf{w}\|_2 \, B.$$

By construction, every $\mathbf{w}' \sim \mathbb{Q}$ has distance at most $2/(mB)$ from $\mathbf{w}$. Therefore, combining the previous inequalities,

$$\left|\boldsymbol{\theta}' \cdot \boldsymbol{\mu}'_{\text{L}} - \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\text{L}}\right| \leq \|\mathbf{w}' - \mathbf{w}\|_2 \, B \leq \frac{2}{mB} \cdot B = \frac{2}{m}. \qquad (24)$$

Using the same approach to upper-bound Equation 22, we assume, without loss of generality, that $(\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} \geq (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}'_{\text{L}}$. Then,

$$\begin{aligned}
\left|(\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} - (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}'_{\text{L}}\right| &= (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} - (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}'_{\text{L}} \\
&\leq (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} - (\delta(\boldsymbol{\mu}_{\text{I}}) + \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}') \cdot \tilde{\boldsymbol{\mu}}_{\text{L}} \\
&= (\mathbf{w} - \mathbf{w}') \cdot \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\mu}}_{\text{L}}) \\
&\leq \|\mathbf{w} - \mathbf{w}'\|_2 \, B \leq \frac{2}{m}. \qquad (25)
\end{aligned}$$

Using Equation 25 to upper-bound 22, and Equation 24 to upper-bound 23, we have that

$$|L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}}) - L_{\text{r}}(\mathbf{w}', \mathbf{x}, \boldsymbol{\mu}_{\text{I}})| \leq \frac{4}{m}.$$

Thus, the loss of any random weight vector, $\mathbf{w}'$, is at most $4/m$ above or below that of the learned weights, $\mathbf{w}$. We can therefore derandomize the randomized loss by bounding its distance to the deterministic loss:

$$\left|L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}}) - \mathop{\mathbb{E}}_{\mathbf{w}' \sim \mathbb{Q}}[L_{\text{r}}(\mathbf{w}', \mathbf{x}, \boldsymbol{\mu}_{\text{I}})]\right| \leq \mathop{\mathbb{E}}_{\mathbf{w}' \sim \mathbb{Q}}[|L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}}) - L_{\text{r}}(\mathbf{w}', \mathbf{x}, \boldsymbol{\mu}_{\text{I}})|] \leq \frac{4}{m}.$$

Combining this bound with the inequalities in Equation 10, we then have that

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{D}}[L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{D}}[L_{\text{r}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{D}} \mathop{\mathbb{E}}_{\mathbf{w}' \sim \mathbb{Q}}[L_{\text{r}}(\mathbf{w}', \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] + \frac{4}{m}, \qquad (26)$$

and

$$\frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{w}' \sim \mathbb{Q}}[L_{\text{r}}(\mathbf{w}', \mathbf{x}^{(i)}, \boldsymbol{\mu}_{\text{I}}^{(i)})] \leq \frac{1}{m} \sum_{i=1}^{m} L_{\text{r}}(\mathbf{w}, \mathbf{x}^{(i)}, \boldsymbol{\mu}_{\text{I}}^{(i)}) + \frac{4}{m}. \qquad (27)$$

All that remains now is to apply Lemma 3, using Equations 26 and 27 to lower- and upper-bound the randomized losses. With probability at least $1 - \delta$ over draws of the training set, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{D}}[L_1(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] &\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{D}} \mathop{\mathbb{E}}_{\mathbf{w}'\sim\mathbb{Q}}[L_{\text{r}}(\mathbf{w}', \mathbf{x}, \boldsymbol{\mu}_{\text{I}})] + \frac{4}{m} \\
&\leq \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{w}'\sim\mathbb{Q}}[L_{\text{r}}(\mathbf{w}', \mathbf{x}^{(i)}, \boldsymbol{\mu}_{\text{I}}^{(i)})] + \frac{4}{m} + 2\sqrt{\frac{d\ln(mBR) + \ln\frac{2}{\delta}}{2m}} \\
&\leq \frac{1}{m}\sum_{i=1}^{m} L_{\text{r}}(\mathbf{w}, \mathbf{x}^{(i)}, \boldsymbol{\mu}_{\text{I}}^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d\ln(mBR) + \ln\frac{2}{\delta}}{2m}}.
\end{aligned}$$

Noting that the hinge loss uniformly upper-bounds the ramp loss completes the proof.

A.2 Proof of Lemma 1

First, we decompose the integrality hinge loss as follows:

$$\begin{aligned}
L_{\text{h}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{I}}) &= \max_{\boldsymbol{\mu}\in\mathcal{M}_{\text{L}}} D_1(\boldsymbol{\mu}_{\text{I}}, \boldsymbol{\mu}) + \boldsymbol{\theta}\cdot(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{I}}) \\
&\leq \max_{\boldsymbol{\mu}\in\mathcal{M}_{\text{L}}} D_1(\boldsymbol{\mu}_{\text{I}}, \boldsymbol{\mu}) + \boldsymbol{\theta}\cdot(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{T}}) \\
&\leq D_1(\boldsymbol{\mu}_{\text{I}}, \boldsymbol{\mu}_{\text{T}}) + \max_{\boldsymbol{\mu}\in\mathcal{M}_{\text{L}}} D_1(\boldsymbol{\mu}_{\text{T}}, \boldsymbol{\mu}) + \boldsymbol{\theta}\cdot(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{T}}) \\
&= D_1(\boldsymbol{\mu}_{\text{I}}, \boldsymbol{\mu}_{\text{T}}) + L_{\text{h}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{T}}).
\end{aligned} \tag{28}$$

The second term on the right-hand side is the hinge loss of the approximate predictor with respect to the true labeling, which can be evaluated efficiently. The first term on the right-hand side is the Hamming loss of exact inference, which cannot be evaluated efficiently. However, this latter quantity can be upper-bounded as follows:

$$\begin{aligned}
D_1(\boldsymbol{\mu}_{\text{I}}, \boldsymbol{\mu}_{\text{T}}) &\leq \max_{\boldsymbol{\mu}\in\mathcal{M}} D_1(\boldsymbol{\mu}, \boldsymbol{\mu}_{\text{T}}) + \boldsymbol{\theta}\cdot(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{T}}) \\
&\leq \max_{\boldsymbol{\mu}\in\mathcal{M}_{\text{L}}} D_1(\boldsymbol{\mu}, \boldsymbol{\mu}_{\text{T}}) + \boldsymbol{\theta}\cdot(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{T}}) \\
&= L_{\text{h}}(\mathbf{w}, \mathbf{x}, \boldsymbol{\mu}_{\text{T}}).
\end{aligned} \tag{29}$$

Combining Equations 28 and 29 completes the proof.

A.3 Proof of Lemma 2

Consider any output variable. If the fractional solution assigns the majority of the local belief to the "correct" label (i.e., the label chosen by exact inference), then the rounding of that variable will be exact. However, if the fractional solution puts most of the local belief on an "incorrect" label, then the rounding of that variable will have D_1 distance 1 from the correct label. Since the incorrect label must have had a fractional value of at least $1/k$, it follows that the fractional solution has D_1 distance at least $1/k$, which is no less than $(1/k)^{\text{th}}$ that of the rounding. Applying this logic to every variable completes the proof.