On the accuracy of self-normalized log-linear models

Jacob Andreas,* **Maxim Rabinovich**,* **Dan Klein**, **Michael I. Jordan**
Computer Science Division, University of California, Berkeley
{jda,rabinovich,klein,jordan}@cs.berkeley.edu

Abstract

Calculation of the log-normalizer is a major computational obstacle in applications of log-linear models with large output spaces. The problem of fast normalizer computation has therefore attracted significant attention in the theoretical and applied machine learning literature. In this paper, we analyze a recently proposed technique known as "self-normalization", which introduces a regularization term in training to penalize log normalizers for deviating from zero. This makes it possible to use unnormalized model scores as approximate probabilities. Empirical evidence suggests that self-normalization is extremely effective, but a theoretical understanding of why it should work, and how generally it can be applied, is largely lacking.

We prove generalization bounds on the estimated variance of normalizers and upper bounds on the loss in accuracy due to self-normalization, describe classes of input distributions that self-normalize easily, and construct explicit examples of high-variance input distributions. Our theoretical results make predictions about the difficulty of fitting self-normalized models to several classes of distributions, and we conclude with empirical validation of these predictions.

1   Introduction

Log-linear models, a general class that includes conditional random fields (CRFs) and generalized linear models (GLMs), offer a flexible yet tractable approach modeling conditional probability distributions $p(x|y)$ [1, 2]. When the set of possible y values is large, however, the computational cost of computing a normalizing constant for each x can be prohibitive—involving a summation with many terms, a high-dimensional integral or an expensive dynamic program.

The machine translation community has recently described several procedures for training "self-normalized" log-linear models [3, 4]. The goal of self-normalization is to choose model parameters that simultaneously yield accurate predictions and produce normalizers clustered around unity. Model scores can then be used as approximate surrogates for probabilities, obviating the computation normalizer computation.

In particular, given a model of the form

$$p_\eta(y \mid x) = e^{\eta^T T(y,\, x) - A(\eta,\, x)} \tag{1}$$

with

$$A(\eta,\, x) = \sum_{y \in \mathcal{Y}} e^{\eta^T T(y,\, x)}, \tag{2}$$

we seek a setting of η such that $A(x, \eta)$ is close enough to zero (with high probability under $p(x)$) to be ignored.

—————————————
*Authors contributed equally.

This paper aims to understand the theoretical properties of self-normalization. Empirical results have already demonstrated the efficacy of this approach—for discrete models with many output classes, it appears that normalizer values can be made nearly constant without sacrificing too much predictive accuracy, providing dramatic efficiency increases at minimal performance cost.

The broad applicability of self-normalization makes it likely to spread to other large-scale applications of log-linear models, including structured prediction (with combinatorially many output classes) and regression (with continuous output spaces). But it is not obvious that we should expect such approaches to be successful: the number of inputs (if finite) can be on the order of millions, the geometry of the resulting input vectors x highly complex, and the class of functions $A(\eta, x)$ associated with different inputs quite rich. To find to find a nontrivial parameter setting with $A(\eta, x)$ roughly constant seems challenging enough; to require that the corresponding η also lead to good classification results seems too much. And yet for many input distributions that arise in practice, it appears possible to choose η to make $A(\eta, x)$ nearly constant without having to sacrifice classification accuracy.

Our goal is to bridge the gap between theoretical intuition and practical experience. Previous work [5] bounds the sample complexity of self-normalizing training procedures for a restricted class of models, but leaves open the question of how self-normalization interacts with the predictive power of the learned model. This paper seeks to answer that question. We begin by generalizing the previously-studied model to a much more general class of distributions, including distributions with continuous support (Section 3). Next, we provide what we believe to be the first characterization of the interaction between self-normalization and model accuracy Section 4. This characterization is given from two perspectives:

- a bound on the "likelihood gap" between self-normalized and unconstrained models
- a conditional distribution provably hard to represent with a self-normalized model

In Figure 5, we present empirical evidence that these bounds correctly characterize the difficulty of self-normalization, and in the conclusion we survey a set of open problems that we believe merit further investigation.

2   Problem background

The immediate motivation for this work is a procedure proposed to speed up decoding in a machine translation system with a neural-network language model [3]. The language model used is a standard feed-forward neural network, with a "softmax" output layer that turns the network's predictions into a distribution over the vocabulary, where each probability is log-proportional to its output activation. It is observed that with a sufficiently large vocabulary, it becomes prohibitive to obtain probabilities from this model (which must be queried millions of times during decoding). To fix this, the language model is trained with the following objective:

$$\max_{W} \sum_{i} \left[N(y_i|x_i; W) - \log \sum_{y'} e^{N(y'|x_i; W)} - \alpha \left(\log \sum_{y'} e^{N(y'|x_i; W)} \right)^2 \right]$$

where $N(y|x; W)$ is the response of output y in the neural net with weights W given an input x. From a Lagrangian perspective, the extra penalty term simply confines the W to the set of "empirically normalizing" parameters, for which all log-normalizers are close (in squared error) to the origin. For a suitable choice of α, it is observed that the trained network is simultaneously accurate enough to produce good translations, and close enough to self-normalized that the raw scores $N(y_i|x_i)$ can be used in place of log-probabilities without substantial further degradation in quality.

We seek to understand the observed success of these models in finding accurate, normalizing parameter settings. While it is possible to derive bounds of the kind we are interested in for general neural networks [6], in this paper we work with a simpler linear parameterization that we believe captures the interesting aspects of this problem. [1]

---

[1] It is possible to view a log-linear model as a single-layer network with a softmax output. More usefully, all of the results presented here apply directly to trained neural nets in which the last layer only is *retrained* to self-normalize [7].

Related work

The approach described at the beginning of this section is closely related to an alternative self-normalization trick described based on noise-contrastive estimation (NCE) [8]. NCE is an alternative to direct optimization of likelihood, instead training a classifier to distinguish between true samples from the model, and "noise" samples from some other distribution. The structure of the training objective makes it possible to replace explicit computation of each log-normalizer with an estimate. In traditional NCE, these values are treated as part of the parameter space, and estimated simultaneously with the model parameters; there exist guarantees that the normalizer estimates will eventually converge to their true values. It is instead possible to fix all of these estimates to one. In this case, empirical evidence suggests that the resulting model will also exhibit self-normalizing behavior [4].

A host of other techniques exist for solving the computational problem posed by the log-normalizer. Many of these involve approximating the associated sum or integral using quadrature [9], herding [10], or Monte Carlo methods [11]. For the special case of discrete, finite output spaces, an alternative approach—the hierarchical softmax—is to replace the large sum in the normalizer with a series of binary decisions [12]. The output classes are arranged in a binary tree, and the probability of generating a particular output is the product of probabilities along the edges leading to it. This reduces the cost of computing the normalizer from $\mathcal{O}(k)$ to $\mathcal{O}(\log k)$. While this limits the set of distributions that can be learned, and still requires greater-than-constant time to compute normalizers, it appears to work well in practice. It cannot, however, be applied to problems with continuous output spaces.

3 Self-normalizable distributions

We begin by providing a slightly more formal characterization of a general log-linear model:

Definition 1 (**Log-linear models**). Given a space of *inputs* $\mathcal{X}$, a space of *outputs* $\mathcal{Y}$, a measure μ on $\mathcal{Y}$, a nonnegative function $h : \mathcal{Y} \to \mathbb{R}$, and a function $T : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ that is μ-measurable with respect to its second argument, we can define a *log-linear model* indexed by parameters $\eta \in \mathbb{R}^d$, with the form

$$p_\eta(y|x) = h(y)e^{\eta^\top T(x,y) - A(x,\eta)} , \tag{3}$$

where

$$A(x,\eta) \triangleq \log \int_{\mathcal{Y}} h(y)e^{\eta^\top T(x,y)} \, \mathrm{d}\mu(y) . \tag{4}$$

If $A(x,\eta) \leq \infty$, then $\int_y p_\eta(y|x) \, \mathrm{d}\mu(y) = 1$, and $p_\eta(y|x)$ is a probability density over $\mathcal{Y}$.[2]

We next formalize our notion of a self-normalized model.

Definition 2 (**Self-normalized models**). The log-linear model $p_\eta(y|x)$ is *self-normalized with respect to* a set $\mathcal{S} \subset \mathcal{X}$ if for all $x \in \mathcal{S}$, $A(x,\eta) = 0$. In this case we say that $\mathcal{S}$ is *self-normalizable*, and η is *self-normalizing* w.r.t. $\mathcal{S}$.

An example of a normalizable set is shown in Figure 1a, and we provide additional examples below:

---

[2]Some readers may be more familiar with generalized linear models, which also describe exponential family distributions with a linear dependence on input. The presentation here is strictly more general, and has a few notational advantages: it makes explicit the dependence of A on x and η but not y, and lets us avoid tedious bookkeeping involving natural and mean parameterizations. [13]
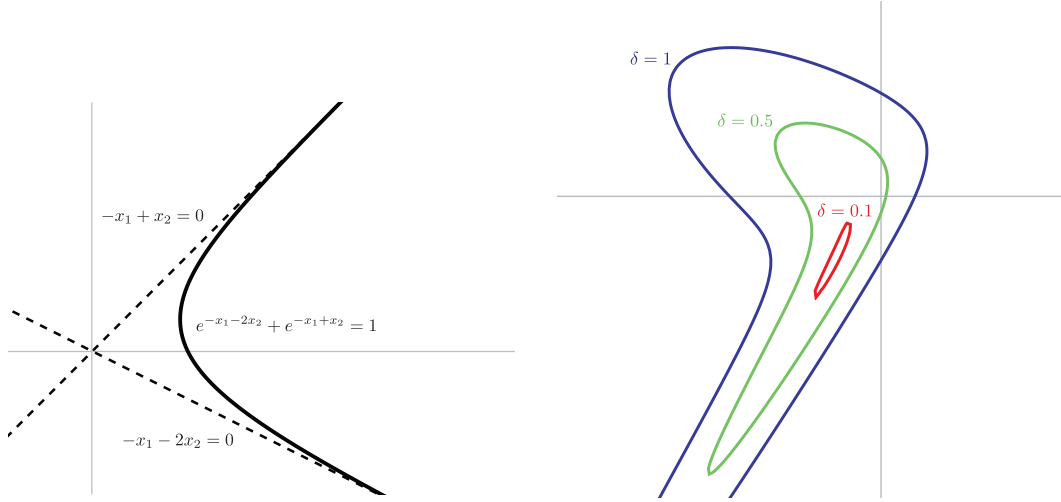
(a) A *self-normalizable* set $\mathcal{S}$ for fixed η: the solutions (x_1, x_2) to $A(x, \eta) = 0$ with $\eta^\top T(x, y) = \eta_y^\top x$ and $\eta = \{(-1, 1), (-1, -2)\}$. The set forms a smooth one-dimensional manifold bounded on either side by hyperplanes normal to $(-1, 1)$ and $(-1, -2)$.

(b) Sets of *approximately normalizing* parameters η for fixed $p(x)$: solutions (η_1, η_2) to $\mathbb{E}[A(x, \eta)^2] = \delta^2$ with $T(x, y) = (x + y, -xy)$, $y \in \{-1, 1\}$ and $p(x)$ uniform on $\{1, 2\}$. For a given upper bound on normalizer variance, the feasible set of parameters is nonconvex, and grows as δ increases.

Figure 1: Self-normalizable data distributions and parameter sets.

Example. *Suppose*

$$\mathcal{S} = \{\log 2, -\log 2\},$$
$$\mathcal{Y} = \{-1, 1\}$$
$$T(x, y) = [xy, 1]$$
$$\eta = (1, \log(2/5)).$$

Then for either $x \in \mathcal{S}$,

$$A(x, \eta) = \log(e^{\log 2 + \log(2/5)} + e^{-\log 2 + \log(2/5)})$$
$$= \log((2/5)(2 + 1/2))$$
$$= 0,$$

and η *is self-normalizing with respect to* $\mathcal{S}$.

It is also easy to choose parameters that do not result in a self-normalized distribution, and in fact to construct a target distribution which cannot be self-normalized:

Example. *Suppose*

$$\mathcal{X} = \{(1, 0), (0, 1), (1, 1)\}$$
$$\mathcal{Y} = \{-1, 1\}$$
$$T(x, y) = (x_1 y, x_2 y, 1)$$

Then there is no η *such that* $A(x, \eta) = 0$ *for all* x, *and* $A(x, \eta)$ *is constant if and only if* $\eta = \mathbf{0}$.

As previously motivated, downstream uses of these models may be robust to small errors resulting from improper normalization, so it would be useful to generalize this definition of normalizable distributions to distributions that are only approximately normalizable. Exact normalizability of the conditional distribution is a deterministic statement—there either does or does not exist some x that violates the constraint. In Figure 1a, for example, it suffices to have a single x off of the indicated surface to make a set non-normalizable. Approximate normalizability, by contrast, is inherently a *probabilistic* statement, involving a distribution $p(x)$ over inputs. Note carefully that we are attempting to represent $p(y|x)$ but have no representation of (or control over) $p(x)$, and that approximate normalizability depends on $p(x)$ but not $p(y|x)$.

4

Informally, if some input violates the self-normalization constraint by a large margin, but occurs only very infrequently, there is no problem; instead we are concerned with *expected* deviation. It is also at this stage that the distinction between penalization of the normalizer vs. log-normalizer becomes important. The normalizer is necessarily bounded below by zero (so overestimates might appear much worse than underestimates), while the log-normalizer is unbounded in both directions. For most applications we are concerned with log probabilities and log-odds ratios, for which an expected normalizer close to zero is just as bad as one close to infinity. Thus the log-normalizer is the natural choice of quantity to penalize.

Definition 3 (**Approximately self-normalized models**). The log-linear distribution $p_\eta(y|x)$ is δ-*approximately normalized with respect to* a distribution $p(x)$ over $\mathcal{X}$ if $\mathbb{E}[A(X, \eta)^2] < \delta^2$. In this case we say that $p(x)$ is δ-*approximately self-normalizable*, and η is δ-*approximately self-normalizing*.

The sets of δ-approximately self-normalizing parameters for a fixed input distribution and feature function are depicted in Figure 1b. Unlike self-normalizable sets of *inputs*, self-normalizing and approximately self-normalizing sets of *parameters* may have complex geometry.

Throughout this paper, we will assume that vectors of sufficient statistics $T(x, y)$ have bounded ℓ_2 norm at most R, natural parameter vectors η have ℓ_2 norm at most B (that is, they are Ivanov-regularized), and that vectors of both kinds lie in $\mathbb{R}^d$. Finally, we assume that all input vectors have a constant feature—in particular, that $x_0 = 1$ for every x (with corresponding weight η_0). [3]

The first question we must answer is whether the problem of training self-normalized models is feasible at all—that is, whether there exist any exactly self-normalizable data distributions $p(x)$, or at least δ-approximately self-normalizable distributions for small δ. Section 3 already gave an example of an exactly normalizable distribution. In fact, there are large classes of both exactly and approximately normalizable distributions.

Observation. *Given some fixed* η*, consider the set* $S_\eta = \{x \in \mathcal{X} : A(x, \eta) = 0\}$*. Any distribution* $p(x)$ *supported on* S_η *is normalizable. Additionally, every self-normalizable distribution is characterized by at least one such* η*.*

This definition provides a simple geometric characterization of self-normalizable distributions. An example solution set is shown in Figure 1a. More generally, if y is discrete and $T(x, y)$ consists of $|\mathcal{Y}|$ repetitions of a fixed feature function $t(x)$ (as in Figure 1a), then we can write

$$A(x, \eta) = \log \sum_{y \in \mathcal{Y}} e^{\eta_y^\top t(x)}. \tag{5}$$

Provided $\eta_y^\top t(x)$ is convex in x for each η_y, the level sets of A as a function of x form the boundaries of convex sets. In particular, exactly normalizable sets are always the boundaries of convex regions, as in the simple example Figure 1a.

We do not, in general, expect real-world datasets to be supported on the precise class of self-normalizable surfaces. Nevertheless, it is very often observed that data of practical interest lie on other low-dimensional manifolds within their embedding feature spaces. Thus we can ask whether it is sufficient for a target distribution to be well-approximated by a self-normalizing one. We begin by constructing an appropriate measurement of the quality of this approximation.

Definition 4 (**Closeness**). An input distribution $p(x)$ is *D-close* to a set $\mathcal{S}$ if

$$\mathbb{E}\left[\inf_{x^* \in \mathcal{S}} \sup_{y \in \mathcal{Y}} ||T(X, y) - T(x^*, y)||_2\right] \leq D \tag{6}$$

In other words, $p(x)$ is D-close to $\mathcal{S}$ if a random sample from p is no more than a distance D from $\mathcal{S}$ in expectation. Now we can relate the quality of this approximation to the level of self-normalization achieved. Generalizing a result from [5], we have:

Proposition 1. *Suppose* $p(x)$ *is* D*-close to* $\{x : A(x, \eta) = 1\}$*. Then* $p(x)$ *is* BD*-approximately self-normalizable (recalling that* $||x||_2 \leq B$*).*

---

[3]It will occasionally be instructive to consider the special case where $\mathcal{X}$ is the Boolean hypercube, and we will explicitly note where this assumption is made. Otherwise all results apply to general distributions, both continuous and discrete.

(Proofs for this section may be found in Appendix A.)

The intuition here is that data distributions that place most of their mass in feature space close to normalizable sets are approximately normalizable on the same scale.

4 Normalization and model accuracy

So far our discussion has concerned the problem of finding conditional distributions that self-normalize, without any concern for how well they actually perform at modeling the data. Here the relationship between the approximately self-normalized distribution and the true distribution $p(y|x)$ (which we have so far ignored) is essential. Indeed, if we are not concerned with making a good model it is always trivial to make a normalized one—simply take $\eta = \mathbf{0}$ and then scale η_0 appropriately! We ultimately desire both good self-normalization and good data likelihood, and in this section we characterize the tradeoff between maximizing data likelihood and satisfying a self-normalization constraint.

We achieve this characterization by measuring the *likelihood gap* between the classical maximum likelihood estimator, and the MLE subject to a self-normalization constraint. Specifically, given pairs $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, let $\ell(\eta|x, y) = \sum_i \log p_\eta(y_i|x_i)$. Then define

$$\hat{\eta} = \arg\max_\eta \ell(\eta|x, y) \tag{7}$$

$$\hat{\eta}_\delta = \arg\max_{\eta:V(\eta)\leq\delta} \ell(\eta|x, y) \tag{8}$$

(where $V(\eta) = \frac{1}{n} \sum_i A(x_i, \eta)^2$).

We would like to obtain a bound on the *likelihood gap*, which we define as the quantity

$$\Delta_\ell(\hat{\eta}, \hat{\eta}_\delta) = \frac{1}{n}(\ell(\hat{\eta}|x, y) - \ell(\hat{\eta}_\delta|x, y)) . \tag{9}$$

We claim:

Theorem 2. *Suppose $\mathcal{Y}$ has finite measure. Then asymptotically as $n \to \infty$*

$$\Delta_\ell(\hat{\eta}, \hat{\eta}_\delta) \leq \left(1 - \frac{\delta}{R||\hat{\eta}||_2}\right) \mathbb{E}\, \mathrm{KL}(p_\eta(\cdot|X) \,||\, \mathrm{Unif}) . \tag{10}$$

(Proofs for this section may be found in Appendix B.)

This result lower-bounds the likelihood at $\hat{\eta}_\delta$ by explicitly constructing a scaled version of $\hat{\eta}$ that satisfies the self-normalization constraint. Specifically, if η is chosen so that normalizers are penalized for distance from $\log \mu(\mathcal{Y})$ (e.g. the logarithm of the number of classes in the finite case), then any increase in η along the span of the data is guaranteed to increase the penalty. From here it is possible to choose an $\alpha \in (0, 1)$ such that $\alpha\hat{\eta}$ satisfies the constraint. The likelihood at $\alpha\hat{\eta}$ is necessarily less than $\ell(\hat{\eta}_\delta|x, y)$, and can be used to obtain the desired lower bound.

Thus at one extreme, distributions close to uniform can be self-normalized with little loss of likelihood. What about the other extreme—distributions "as far from uniform as possible"? With suitable assumptions about the form of $p_{\hat{\eta}}(y|x)$, we can use the same construction of a self-normalizing parameter to achieve an alternative characterization for distributions that are close to deterministic:

Proposition 3. *Suppose that $\mathcal{X}$ is a subset of the Boolean hypercube, $\mathcal{Y}$ is finite, and $T(x, y)$ is the conjunction of each element of x with an indicator on the output class. Suppose additionally that in* every *input x, $p_{\hat{\eta}}(y|x)$ makes a unique best prediction—that is, for each $x \in \mathcal{X}$, there exists a unique $y^* \in \mathcal{Y}$ such that whenever $y \neq y^*$, $\eta^\top T(x, y^*) > \eta^\top T(x, y)$. Then*

$$\Delta_\ell(\hat{\eta}, \hat{\eta}_\delta) \leq b \left(||\eta||_2 - \frac{\delta}{R}\right)^2 e^{-c\delta/R} \tag{11}$$

for distribution-dependent constants b and c.

This result is obtained by representing the constrained likelihood with a second-order Taylor expansion about the true MLE. All terms in the likelihood gap vanish except for the remainder; this can be

upper-bounded by the $||\hat{\eta}_\delta||_2^2$ times the largest eigenvalue the feature covariance matrix at $\hat{\eta}_\delta$, which in turn is bounded by $e^{-c\delta/R}$.

The favorable rate we obtain for this case indicates that "all-nonuniform" distributions are also an easy class for self-normalization. Together with Theorem 2, this suggests that hard distributions must have some mixture of uniform and nonuniform predictions for different inputs. This is supported by the results in Section 4.

The next question is whether there is a corresponding lower bound; that is, whether there exist any conditional distributions for which all nearby distributions are provably hard to self-normalize. The existence of a direct analog of Theorem 2 remains an open problem, but we make progress by developing a general framework for analyzing normalizer variance.

One key issue is that while likelihoods are invariant to certain changes in the natural parameters, the log normalizers (and therefore their variance) is far from invariant. We therefore focus on equivalence classes of natural parameters, as defined below. Throughout, we will assume a fixed distribution $p(x)$ on the inputs x.

Definition 5 (**Equivalence of parameterizations**). Two natural parameter values η and η' are said to be *equivalent* (with respect to an input distribution $p(x)$), denoted $\eta \sim \eta'$ if

$$p_\eta(y|X) = p_{\eta'}(y|X) \quad \text{a.s. } p(x)$$

We can then define the optimal log normalizer variance for the distribution associated with a natural parameter value.

Definition 6 (**Optimal variance**). We define the *optimal log normalizer variance* of the log-linear model associated with a natural parameter value η by

$$V^*(\eta) = \inf_{\eta' \sim \eta} \text{Var}_{p(x)} \left[A(X, \eta) \right].$$

We now specialize to the case where $\mathcal{Y}$ is finite with $|\mathcal{Y}| = K$ and where $T : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^{Kd}$ satisfies

$$T(k, x)_{k'j} = \delta_{kk'} x_j.$$

This is an important special case that arises, for example, in multi-way logistic regression. In this setting, we can show that despite the fundamental non-identifiability of the model, the variance can still be shown to be high under *any* parameterization of the distribution.

Theorem 4. *Let $\mathcal{X} = \{0, 1\}^d$ and let the input distribution $p(x)$ be uniform on $\mathcal{X}$. There exists an $\eta^0 \in \mathbb{R}^{Kd}$ such that for $\eta = \alpha\eta^0$, $\alpha > 0$,*

$$V^*(\eta) \geq \frac{||\eta||_2^2}{32d(d-1)} - 4Ke^{-\frac{\sqrt{1-\frac{1}{d}}||\eta||_2}{2(d-1)}} ||\eta||_2.$$

5 Experiments

The high-level intuition behind the results in the preceding section can be summarized as follows: 1) for predictive distributions that are in expectation high-entropy or low-entropy, self-normalization results in a relatively small likelihood gap; 2) for mixtures of high- and low-entropy distributions, self-normalization may result in a large likelihood gap. More generally, we expect that an increased tolerance for normalizer variance will be associated with a decreased likelihood gap.

In this section we provide experimental confirmation of these predictions. We begin by generating a set of random sparse feature vectors, and an initial weight vector η_0. In order to produce a sequence of label distributions that smoothly interpolate between low-entropy and high-entropy, we introduce a temperature parameter τ, and for various settings of τ draw labels from $p_{\tau\eta}$. We then fit a self-normalized model to these training pairs. In addition to the synthetic data, we compare our results to empirical data [3] from a self-normalized language model.

Figure 2a plots the tradeoff between the likelihood gap and the error in the normalizer, under various distributions (characterized by their KL from uniform). Here the tradeoff between self-normalization and model accuracy can be seen—as the normalization constraint is relaxed, the likelihood gap decreases.
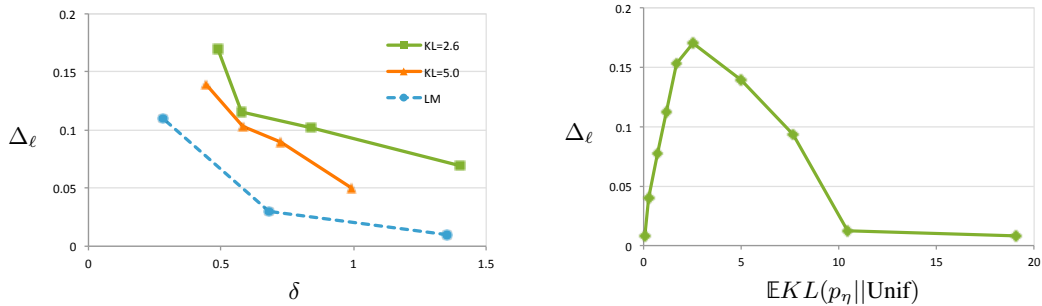
(a) Normalization / likelihood tradeoff. As the normalization constraint δ is relaxed, the likelihood gap Δ_ℓ decreases. Lines marked "KL=" are from synthetic data; the line marked "LM" is from [3].

(b) Likelihood gap as a function of expected divergence from the uniform distribution. As predicted by theory, the likelihood gap increases, then decreases, as predictive distributions become more peaked.

Figure 2: Experimental results

Figure 2b shows how the likelihood gap varies as a function of the quantity $\mathbb{E}KL(p_\eta(\cdot|X)||\text{Unif})$. As predicted, it can be seen that both extremes of this quantity result in small likelihood gaps, while intermediate values result in large likelihood gaps.

6 Conclusions

Motivated by the empirical success of self-normalizing parameter estimation procedures for log-linear models, we have attempted to establish a theoretical basis for the understanding of such procedures. We have characterized both self-normalizable *distributions*, by constructing provably easy examples, and self-normalizing *training procedures*, by bounding the loss of likelihood associated with self-normalization.

While we have addressed many of the important first-line theoretical questions around self-normalization, this study of the problem is by no means complete. We hope this family of problems will attract further study in the larger machine learning community; toward that end, we provide the following list of open questions:

1. **How else can the approximately self-normalizable distributions be characterized?** The class of approximately normalizable distributions we have described is unlikely to correspond perfectly to real-world data. We expect that Proposition 1 can be generalized to other parametric classes, and relaxed to accommodate spectral or sparsity conditions.

2. **Are the upper bounds in Theorem 2 or Proposition 3 tight?** Our constructions involve relating the normalization constraint to the ℓ_2 norm of η, but in general some parameters can have very large norm and still give rise to almost-normalized distributions.

3. **Do corresponding lower bounds exist?** While it is easy to construct of exactly self-normalizable distributions (which suffer no loss of likelihood), we have empirical evidence that hard distributions also exist. It would be useful to lower-bound the loss of likelihood in terms of some simple property of the target distribution.

4. **Is the hard distribution in Theorem 4 stable?** This is related to the previous question. The existence of high-variance distributions is less worrisome if such distributions are comparatively rare. If the variance lower bound falls off quickly as the given construction is perturbed, then the associated distribution may still be approximately self-normalizable with a good rate.

We have already seen that new theoretical insights in this domain can translate directly into practical applications. Thus, in addition to their inherent theoretical interest, answers to each of these questions might be applied directly to the training of approximately self-normalized models in practice. We expect that self-normalization will find increasingly many applications, and we hope the results in this paper provide a first step toward a complete theoretical and empirical understanding of self-normalization in log-linear models.

References

[1] Lafferty, J. D.; McCallum, A.; Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001; pp 282–289.

[2] McCullagh, P.; Nelder, J. A. *Generalized linear models*; Chapman and Hall, 1989.

[3] Devlin, J.; Zbib, R.; Huang, Z.; Lamar, T.; Schwartz, R.; Makhoul, J. Fast and robust neural network joint models for statistical machine translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2014.

[4] Vaswani, A.; Zhao, Y.; Fossum, V.; Chiang, D. Decoding with large-scale neural language models improves translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2013.

[5] Andreas, J.; Klein, D. When and why are log-linear models self-normalizing? Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics. 2014.

[6] Bartlett, P. L. *IEEE Transactions on Information Theory* **1998**, *44*, 525–536.

[7] Anthony, M.; Bartlett, P. *Neural network learning: theoretical foundations*; Cambridge University Press, 2009.

[8] Gutmann, M.; Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Proceedings of the International Conference on Artificial Intelligence and Statistics. 2010; pp 297–304.

[9] O'Hagan, A. *Journal of statistical planning and inference* **1991**, *29*, 245–260.

[10] Chen, Y.; Welling, M.; Smola, A. *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **2010**, 109–116.

[11] Doucet, A.; De Freitas, N.; Gordon, N. *An introduction to sequential Monte Carlo methods*; Springer, 2001.

[12] Morin, F.; Bengio, Y. *Proceedings of the International Conference on Artificial Intelligence and Statistics* **2005**, 246.

[13] Yang, E.; Allen, G.; Liu, Z.; Ravikumar, P. K. Graphical models via generalized linear models. Advances in Neural Information Processing Systems. 2012; pp 1358–1366.

A Normalizable distributions

Proof of Proposition 1 (distributions close to normalizable sets are approximately normalizable).

Let $T(x, y) = T^*(x, y) + T^-(x, y)$, where $T^*(x, y) = \underset{T(x,y):x\in\mathcal{S}}{\arg\min} \, ||T(X, y) - T(x, y)||_2$.

Then,

$$
\mathbb{E}\left(\log\left(\int e^{\eta^\top T(X,y)} \, \mathrm{d}y\right)\right)^2 = \mathbb{E}\left(\log\left(\int e^{\eta^\top (T^*(X,y)+T^-(X,y))} \, \mathrm{d}y\right)\right)^2
$$

$$
\leq \mathbb{E}\left(\log\left(e^{\eta^\top \tilde{T}}\int e^{\eta^\top T^*(X,y)} \, \mathrm{d}y\right)\right)^2
$$

for $\tilde{T} = \underset{T(X,y)}{\arg\max} \, ||\eta^\top T(X, y)||_2$,

$$
\leq \mathbb{E}\left(\log\left(e^{\eta^\top \tilde{T}}\right)\right)^2
$$

$$
= (DB)^2 \qquad\qquad \square
$$

B Normalization and likelihood

B.1 General bound

Lemma 5. *If $||\eta||_2 \leq \delta/R$, then $p_\eta(y|x)$ is δ-approximately normalized about $\log\mu(\mathcal{Y})$.*

Proof. If $\int e^{\eta^\top T(X,y)} \, \mathrm{d}\mu(y) \geq \log\mu(\mathcal{Y})$,

$$
\left(\log\int_{\mathcal{Y}} e^{\eta^\top T(X,y)} \, \mathrm{d}\mu(y) - \log\mu(\mathcal{Y})\right)^2 \leq \left(\log\int_{\mathcal{Y}} e^{||\eta||_2 R} \, \mathrm{d}\mu(y) - \log\mu(\mathcal{Y})\right)^2
$$

$$
= ||\eta||_2^2 R^2
$$

$$
\leq \delta^2
$$

The case where $\int e^{\eta^\top T(X,y)} \, \mathrm{d}\mu(y) \leq \log\mu(\mathcal{Y})$ is analogous, instead replacing $\eta^\top T(x, y)$ with $-||\eta||_2 R$. The variance result follows from the fact that every log-partition is within δ of the mean. $\square$

Proof of Theorem 2 (loss of likelihood is bounded in terms of distance from uniform). Consider the likelihood evaluated at $\alpha\hat{\eta}$, where $\alpha = \delta/R||\hat{\eta}||_2$. We know that $0 \leq \alpha \leq 1$ (if $\delta > R\eta$, then the MLE already satisfying the normalizing constraint). Additionally, $p_{\alpha\hat{\eta}}(y|x)$ is δ-approximately normalized. (Both follow from Lemma 5.)

Then,

$$
\Delta_\ell = \frac{1}{n}\sum_i \left[(\hat{\eta}^\top T(x_i, y_i) - A(x_i, \hat{\eta})) - (\alpha\hat{\eta}^\top T(x_i, y_i) - A(x_i, \alpha\hat{\eta}))\right]
$$

$$
= \frac{1}{n}\sum_i \left[(1-\alpha)\hat{\eta}^\top T(x_i, y_i) - A(x_i, \hat{\eta}) + A(x_i, \alpha\hat{\eta})\right]
$$

Because $A(x, \alpha\eta)$ is convex in α,

$$
A(x_i, \alpha\hat{\eta}) \leq (1-\alpha)A(x_i, \mathbf{0}) + \alpha A(x_i, \hat{\eta})
$$

$$
= (1-\alpha)\mu(\mathcal{Y}) + \alpha A(x_i, \hat{\eta})
$$

Thus,

$$\Delta_\ell = \frac{1}{n} \sum_i \left[(1-\alpha)\hat{\eta}^\top T(x_i, y_i) - A(x_i, \hat{\eta}) + (1-\alpha)\log\mu(\mathcal{Y}) + \alpha A(x_i, \hat{\eta}) \right]$$

$$= (1-\alpha)\frac{1}{n} \sum_i \left[\hat{\eta}^\top T(x_i, y_i) - A(x_i, \hat{\eta}) + \log\mu(\mathcal{Y}) \right]$$

$$= (1-\alpha)\frac{1}{n} \sum_i \left[\log p_\eta(y|x) - \log\mathrm{Unif}(y) \right]$$

$$\asymp (1-\alpha)\,\mathbb{E}\,\mathrm{KL}(p_\eta(\cdot|X) \,||\, \mathrm{Unif})$$

$$\leq \left(1 - \frac{\delta}{R||\hat{\eta}||_2}\right) \mathbb{E}\,\mathrm{KL}(p_\eta(\cdot|X) \,||\, \mathrm{Unif}) \qquad \qquad \square$$

B.2   All-nonuniform bound

We make the following assumptions:

- Labels y are discrete. That is, $\mathcal{Y} = \{1, 2, \ldots, k\}$ for some k.

- $x \in \mathcal{H}(d)$. That is, each x is a $\{0, 1\}$ indicator vector drawn from the Boolean hypercube in q dimensions.

- Joint feature vectors $T(x, y)$ are just the features of x conjoined with the label y. Then it is possible to think of η as a sequence of vectors, one per class, and we can write $\eta^\top T(x, y) = \eta_y^\top x$.

- As in the body text, let all MLE predictions be nonuniform, and in particular let each $\hat{\eta}_{y^*}^\top x - \hat{\eta}_y^\top x > c||\hat{\eta}||$ for $y \neq y^*$.

Lemma 6. *For a fixed x, the maximum covariance between any two features x_i and x_j under the model evaluated at some η in the direction of the MLE:*

$$Cov[T(X, Y)_i, T(X, Y)_j | X = x] \leq 2(k-1)e^{-c\delta} \qquad (12)$$

Proof. If either i or j is not associated with the class y, or associated with a zero element of x, then the associated feature (and thus the covariance at (i, j)) is identically zero. Thus we assume that i and j are both associated with y and correspond to nonzero elements of x.

$$\mathrm{Cov}[T_i, T_j | X = x] = \sum_y p_\eta(y|x) - p_\eta(y|x)^2$$

Suppose y is the majority class. Then,

$$p_\eta(y|x) - p_\eta(y|x)^2 = \frac{e^{\eta_y^\top x}}{\sum_{y'} e^{\eta_{y'}^\top x}} - \frac{e^{2\eta_y^\top x}}{\left(\sum_{y'} e^{\eta_{y'}^\top x}\right)^2}$$

$$= \frac{e^{\eta_y^\top x}\left(\sum_{y'} e^{\eta_{y'}^\top x}\right) - e^{2\eta_y^\top x}}{\left(\sum_{y'} e^{\eta_{y'}^\top x}\right)^2}$$

$$\leq \frac{e^{\eta_y^\top x}\left(\sum_{y'} e^{\eta_{y'}^\top x}\right) - e^{2\eta_y^\top x}}{e^{2\eta_y^\top x}}$$

$$= \sum_{y' \neq y} e^{(\eta_{y'}' - \eta_y)^\top x}$$

$$\leq (k-1)e^{-c||\eta||}$$

11

Now suppose y is not in the majority class. Then,

$$p_\eta(y|x) - p_\eta(y|x)^2 \leq p(y|x)$$

$$= \frac{e^{\eta_y^\top x}}{\sum_{y'} e^{\eta_{y'}^\top x}}$$

$$\leq e^{-c||\eta||}$$

Thus the covariance

$$\sum_y p_\eta(y|x) - p_\eta(y|x)^2 \leq 2(k-1)e^{-c||\eta|||}$$

$\square$

Lemma 7. *Suppose $\eta = \beta\hat{\eta}$ for some $\beta < 1$. Then for a sequence of observations $(x_1, \ldots, x_n)$, under the model evaluated at ξ, the largest eigenvalue of the feature covariance matrix*

$$\frac{1}{n}\sum_i \left[\mathbb{E}_\xi[TT^\top | X = x_i] - (\mathbb{E}_\theta[T|X = x_i])(\mathbb{E}_\xi[T|X = x_i])^\top \right] \tag{13}$$

is at most

$$q(k-1)e^{-c\beta||\hat{\eta}||} \tag{14}$$

Proof. From Lemma 6, each entry in the covariance matrix is at most $(k-1)e^{-c||\eta||} = (k-1)e^{-c\beta||\hat{\eta}||}$. At most q features are nonzero active in any row of the matrix. Thus by Gershgorin's theorem, the maximum eigenvalue of each term in Equation 13 is $q(k-1)e^{-c\beta||\hat{\eta}||}$, which is also an upper bound on the sum. $\square$

Proof of Proposition 3 (loss of likelihood goes as $e^{-\delta}$). As before, let us choose $\hat{\eta}_\delta = \alpha\hat{\eta}$, with $\alpha = \delta/R||\hat{\eta}||_2$. We have already seen that this choice of parameter is normalizing.

Taking a second-order Taylor expansion about η, we have

$$\log p_{\hat{\eta}_\delta}(y|x) = \log p_\eta(y|x) + (\hat{\eta}_\delta - \hat{\eta})^\top \nabla \log p_{\hat{\eta}}(y|x) + (\hat{\eta}_\delta - \hat{\eta})^\top \nabla\nabla^\top \log p_\xi(y|x)(\hat{\eta}_\delta - \hat{\eta})$$

$$= \log p_{\hat{\eta}}(y|x) + (\hat{\eta}_\delta - \hat{\eta})^\top \nabla\nabla^\top \log p_\xi(y|x)(\hat{\eta}_\delta - \hat{\eta})$$

where the first-order term vanishes because $\hat{\eta}$ is the MLE. It is a standard result for exponential families that the Hessian in the second-order term is just Equation 13. Thus we can write

$$\geq \log p_{\hat{\eta}}(y|x) - ||\hat{\eta}_\delta - \hat{\eta}||^2 q(k-1)e^{-c\beta||\eta||}$$

$$\geq \log p_{\hat{\eta}}(y|x) - (1-\alpha)^2||\hat{\eta}||^2 q(k-1)e^{-c\alpha||\eta||}$$

$$= \log p_{\hat{\eta}}(y|x) - (||\hat{\eta}|| - \delta/R)^2 q(k-1)e^{-c\delta/R}$$

The proposition follows. $\square$

C   Variance lower bound

Let

$$U_0 = \{\beta \in \mathbb{R}^{Kd} : \exists \tilde{\beta} \in \mathbb{R}^d, \beta_{kj} = \tilde{\beta}_j, \ 1 \leq k \leq K, \ 1 \leq j \leq d\}.$$

Lemma 8. *If $\text{span}(\mathcal{X}) = \mathbb{R}^d$, then equivalence of natural parameters is characterized by*

$$\eta \sim \eta' \iff \eta - \eta' \in U_0.$$

Proof. For $x \in \mathcal{X}$, denote by $P_\eta(x) \in \Delta_K$ the distribution over $\mathcal{Y}$. Now, suppose that $\eta \sim \eta'$ and fix $x \in \mathcal{X}$. By the definition of equivalence, we have

$$\frac{P_\eta(x)_k}{P_\eta(x)_{k'}} = \frac{P_{\eta'}(x)_k}{P_{\eta'}(x)_{k'}},$$

which immediately implies
$$(\eta_k - \eta_{k'})^T x = (\eta_k' - \eta_{k'}')^T x,$$

whence
$$[(\eta_k - \eta_k') - (\eta_{k'} - \eta_{k'}')]^T x = 0.$$

Since this holds for all $x \in \mathcal{X}$ and $\mathrm{span}(\mathcal{X}) = \mathbb{R}^d$, we get
$$\eta_k - \eta_k' = \eta_{k'} - \eta_{k'}'.$$

That is, if we define
$$\tilde{\beta}_j = \eta_{1j} - \eta_{1j}',$$

we get
$$\eta_{kj} - \eta_{kj}' = \eta_{1d} - \eta_{1j}' = \tilde{\beta}_j,$$

and $\eta - \eta' \in U_0$, as required.

Conversely, if $\eta - \eta' \in U_0$, choose an appropriate $\tilde{\beta}$. We then get
$$\eta_k^T x = (\eta')^T x + \tilde{\beta}^T x.$$

It follows that
$$A(\eta', x) = A(\eta, x) + \tilde{\beta}^T x,$$

so that
$$\eta^T T(k, x) - A(\eta, x) = (\eta')^T x + \tilde{\beta}^T x - \left[A(\eta', x) + \tilde{\beta}^T x \right] = (\eta')^T x - A(\eta', x)$$

and the claim follows. $\qquad\square$

The key tool we use to prove the theorem reinterprets $V^*(\eta)$ as the norm of an orthogonal projection. We believe this may be of independent interest. To set it up, let $\mathcal{S} = L^2\left(Q, \mathbb{R}^D\right)$ be the Hilbert space of square-integrable functions with respect to the input distribution $p(x)$, define
$$w_j(x) = x_j - \mathbb{E}_{p(x)}\left[X_j\right]$$

and
$$\mathcal{C} = \mathrm{span}\left(w_j\right)_{1 \le j \le d}.$$

We then have

Lemma 9. *Let* $\tilde{A}(\eta, x) = A(\eta, x) - \mathbb{E}_{p(x)}\left[A(\eta, X)\right]$. *Then*
$$V^*(\eta) = \left\| \tilde{A}(\eta, \cdot) - \Pi_\mathcal{C} \tilde{A}(\eta, \cdot) \right\|_2^2 .$$

The second key observation, which we again believe is of independent interest, is that under certain circumstances, we can completely replace the normalizer $A(\eta, \cdot)$ by $\max_{y \in \mathcal{Y}} \eta^T T(y, x)$. For this, we define
$$E_\infty(\eta)(x) = \max_k \eta^T T(k, x) = \max_k \eta_k^T x$$

and correspondingly let $\bar{E}_\infty(\eta) = \mathbb{E}_{p(x)}\left[E_\infty(\eta)(x)\right]$.

Proof. By Lemma 8, we have
$$V^*(\eta) = \inf_{\beta \in \mathbb{R}^d} \int_{\mathbb{R}^{Kd}} \left[A(\eta, x) - \bar{A}(\eta) - \left(\beta^T x - \beta^T \mathbb{E}_{p(x)}\left[X\right] \right) \right]^2 \mathrm{d}p(x).$$

But now, we observe that this can be rewritten with the aid of the isomorphism $\mathbb{R}^d \simeq \mathcal{C}$ defined by the identity
$$\beta^T x - \beta^T \mathbb{E}_{p(x)}\left[X\right] = \sum_j \beta_j w_j(x)$$

to read
$$V^*(\eta) = \inf_{f \in \mathcal{C}} \int_{\mathbb{R}^d} \left[A(\eta, x) - \bar{A}(\eta) - f \right]^2 \mathrm{d}p(x) = \left\| \tilde{A}(\eta, \cdot) - \Pi_\mathcal{C} \tilde{A}(\eta, \cdot) \right\|_2^2 ,$$

as required. $\qquad\square$

Lemma 10. *Suppose for each $x \in \mathcal{X}$, there is a unique $k^* = k^*(x)$ such that $k^*(x) = \arg\max_k \eta_k^T x$ and such that for $k \neq k^*$, $\eta_k^T x \leq \eta_{k^*}^T x - \Delta$ for some $\Delta > 0$. Then*

$$\sup_{x \in \mathcal{X}} \left| A(\eta, x) - \bar{A}(\eta) - \left[E_\infty(\eta)(x) - \bar{E}_\infty(\eta) \right] \right| \leq K e^{-\Delta \alpha}.$$

Proof. Denote by $\tilde{E}_\infty$ the centered version of E_∞. Using the identity $1 + t \leq e^t$, we immediately see that

$$E_\infty(\alpha\eta)(x) \leq A(\alpha\eta, x) = \alpha E_\infty(\eta)(x) + \log\left(1 + \sum_{k \neq k^*(x)} e^{\left[\eta_k^T x - E_\infty(\eta)(x) \right]} \right) \leq E_\infty(\alpha\eta)(x) + K e^{-\Delta\alpha}.$$

It follows that

$$\mathbb{E}_{p(x)}\left[E_\infty(\alpha\eta)(X) \right] \leq \mathbb{E}_{p(x)}\left[A(\alpha\eta, X) \right] \leq \mathbb{E}_{p(x)}\left[E_\infty(\alpha\eta)(X) \right] + K e^{-\Delta\alpha}.$$

We thus have

$$-K e^{-\Delta\alpha} \leq \tilde{A}(\alpha\eta, x) - \tilde{E}_\infty(\alpha\eta)(x) \leq K e^{-\Delta\alpha}, \quad x \in \mathcal{X}.$$

The claim follows. $\qquad\square$

If we let

$$V_{\mathrm{E}}^*(\eta) = \inf_{\eta' \sim \eta} \mathrm{Var}_{p(x)}\left[\tilde{E}_\infty(\eta', X) \right].$$

Corollary 11. *For $\alpha > \frac{\log 2K}{\Delta}$, we have*

$$V^*(\alpha\eta) \geq V_{\mathrm{E}}^*(\eta)\alpha^2 - \left(1 + V_{\mathrm{E}}^*(\eta) \right)\alpha.$$

Proof. For this, observe first that if $\eta' \sim \eta$, then

$$\tilde{A}(\eta', x)^2 \geq \tilde{E}_\infty(\alpha\eta')(x)^2 - 2\left| \tilde{E}_\infty(\alpha\eta')(x) \right| \left| \tilde{A}(\eta', x) - \tilde{E}_\infty(\eta')(x) \right|.$$

By linearity of $E_\infty(\eta')$ in its η argument, and by Lemma 10, we therefore deduce

$$\tilde{A}(\eta', x)^2 \geq \tilde{E}_\infty(\eta')(x)^2\alpha^2 - 2K e^{-\Delta\alpha} \left| \tilde{E}_\infty(\eta')(x) \right| \alpha.$$

Then using the inequality $\mathbb{E}_{p(x)}\left[|f(X)| \right] \leq 1 + \mathrm{Var}_{p(x)}\left[f(X) \right]$, valid for any $f \in L^2\left(Q, \mathbb{R}^D \right)$ with $\mathbb{E}_{p(x)}\left[f \right] = 0$, we thus deduce

$$\mathrm{Var}_{p(x)}\left[A(\alpha\eta', X) \right] \geq \mathrm{Var}_{p(x)}\left[E_\infty(\eta')(X) \right]\alpha^2 - 2K e^{-\Delta\alpha}\left(1 + \mathrm{Var}_{p(x)}\left[E_\infty(\eta')(X) \right] \right)\alpha.$$

Taking the infimum over both sides, we get

$$V^*(\eta) \geq V_{\mathrm{E}}^*(\eta) - 2K e^{-\Delta\alpha}\left(1 + V_{\mathrm{E}}^*(\eta) \right)\alpha.$$

$\qquad\square$

We are now prepared to give the explicit example. It is defined by $\eta_k = 0$ if $k > 2$ and

$$\eta_{1j} = \begin{cases} -a & \text{if } d = 1, \\ \frac{a}{d-1} & \text{o.w.} \end{cases} \tag{15}$$

and for all j,

$$\eta_{2j} = \frac{a}{d(d-1)}, \tag{16}$$

where

$$a = \sqrt{1 - \frac{1}{d}}.$$

For convenience, also define

$$b(x) = \sum_d x_d$$

14

and observe that

$$E_\infty(\eta)(x) = \begin{cases} \frac{ab(x)}{d(d-1)} & \text{if } x_j = 1, \\ \frac{ab(x)}{d-1} & \text{o.w.} \end{cases},$$

Our goal will be to prove that

$$1 \geq V_{\mathrm{E}}^*(\eta) \geq \frac{1}{32d(d-1)}.$$

The claim will then follow by the above corollary.

To see that $V_{\mathrm{E}}^*(\eta) \leq 1$, we simply note that

$$\max_k \left| \eta_k^T x \right| \leq a < 1,$$

whence $\mathrm{Var}_{p(x)}\left[\eta^T x \right] \leq 1$ as well and we are done.

The other direction requires more work. To prove it, we first prove the following lemma

Lemma 12. *With η defined as in* (15)-(16)*, we have*

$$\inf_{\eta' \sim \eta} \mathbb{E}_{p(x)}\left[E_\infty(\eta')(X)^2 \right] \geq \frac{1}{16d(d-1)}.$$

Proof. Suppose $\eta_k - \eta_k' = \beta \in \mathbb{R}^d$. We can then write

$$\inf_{\eta' \sim \eta} \mathbb{E}_{p(x)}\left[E_\infty(\eta')(X)^2 \right] = \inf_{\beta \in \mathbb{R}^d} \frac{1}{2^d} \sum_{x \in \mathcal{H}} \sum_{x \in \mathcal{H}} \left[E_\infty(\eta)(x) - \beta^T x \right]^2$$

and we therefore define

$$\mathcal{L}(\beta) = \sum_{x \in \mathcal{H}} \sum_{x \in \mathcal{H}} \left[E_\infty(\eta)(x) - \beta^T x \right]^2$$

$$= \sum_{x \,:\, x_1 = 0} \left[\left(\beta_1 + \beta^T x - \frac{a}{d(d-1)} \right)^2 + \left(\frac{ab(x)}{d-1} - \beta^T x \right)^2 \right],$$

noting that

$$\inf \mathcal{L} = 2^d \cdot \inf_{\eta' \sim \eta} \mathbb{E}_{p(x)}\left[E_\infty(\eta')(X)^2 \right].$$

We therefore need to prove

$$\mathcal{L} \geq \frac{2^{d-4}}{d(d-1)}.$$

Holding $\beta_{2:d}$ fixed, we note that the optimal setting of β_1 is given by

$$\beta_1 = -\frac{1}{2} \sum_{j \geq 2} \beta_j + \frac{a}{d(d-1)}.$$

We can therefore work with the objective

$$\mathcal{L}(\beta) = \sum_{x \,:\, x_1 = 0} \left[\frac{(\beta^T x - \beta^T x^\neg)^2}{4} + \left(\frac{ab(x)}{d-1} - \beta^T x \right)^2 \right],$$

where we have defined

$$x_j^\neg = \begin{cases} 0 & \text{if } j = 1, \\ 1 - x_j & \text{o.w.} \end{cases}$$

Grouping into $\{x, x^\neg\}$ pairs, we end up with

$$\mathcal{L}(\beta_{2:d}) = \sum_{x \,:\, x_1 = x_2 = 0} \left[\frac{(\beta^T x - \beta^T x^\neg)^2}{2} + \left(\frac{ab(x)}{d-1} - \beta^T x \right)^2 + \left(\frac{ab(x^\neg)}{d-1} - \beta^T x^\neg \right)^2 \right]$$

15

Now, supposing $b(x) \leq \frac{d-1}{2} - \frac{3}{2}$ or $b(x) \geq \frac{D-1}{2} + \frac{3}{2}$, we have
$$|b(x^\neg) - b(x)| = |d - 1 - 2b(x)| \geq 3.$$

We will bound the terms that satisfy this property. Indeed, supposing we fix such an x, at least one of the following must be true: either
$$\max\left(\left(\frac{ab(x)}{d-1} - \beta^T x\right)^2, \left(\frac{ab(x^\neg)}{d-1} - \beta^T x^\neg\right)^2\right) \geq \frac{a^2}{(d-1)^2},$$

or
$$\left(\beta^T x - \beta^T x^\neg\right)^2 \geq \frac{a^2}{(d-1)^2}.$$

Indeed, suppose the first condition does not hold. Then necessarily
$$\left|\frac{ab(x)}{d-1} - \beta^T x\right| < \frac{a}{d-1}$$

and
$$\left|\frac{ab(x^\neg)}{d-1} - \beta^T x^\neg\right| < \frac{a}{d-1},$$

so that
$$\frac{a(b(x)-1)}{d-1} \leq \beta^T x \leq \frac{a(b(x)+1)}{d-1}$$

and
$$\frac{a(b(x^\neg)-1)}{d-1} \leq \beta^T x \leq \frac{a(b(x^\neg)+1)}{d-1}.$$

Now, if $b(x) \geq b(x^\neg) + 3$, this immediately implies
$$\beta^T x - \beta^T x^\neg \geq \frac{a}{d-1}$$

and, symmetrically, if $b(x^\neg) \geq b(x) + 3$, we get
$$\beta^T x^\neg - \beta^T x \geq \frac{a}{d-1}.$$

Either way, the second inequality holds, whence the claim. Since there are at least $2^{d-1} - \frac{3 \cdot 2^d}{\sqrt{\frac{3d}{2}+1}} \geq 2^{d-2}$ choices of x satisfying the requirements of our line of reasoning, we get 2^{d-3} pairs, whence
$$\mathcal{L}(\beta_{2:d}) \geq \frac{2^{d-4}a^2}{(d-1)^2} = \frac{2^{d-4}}{d(d-1)},$$

as claimed. $\qquad\square$

We can apply this lemma to derive a variance bound, viz.

Lemma 13. *With η as in (15)-(16), we have*
$$V_E^*(\eta) \geq \frac{1}{32d(d-1)}.$$

Proof. For this, observe that, with η' being the value corresponding to $\eta_k' - \eta_k = \beta$, we have
$$V_E^*(\eta) = \inf_\beta \frac{1}{2^d} \sum_{x \in \mathcal{H}} \tilde{E}_\infty(\eta')(x)^2 \geq \inf_\beta \frac{1}{2^d} \sum_{x \in \mathcal{H}: x_1=1} \tilde{E}_\infty(\eta')(x)^2.$$

Applying the previous result to the $(D-1)$-dimensional hypercube on which $x_1 = 1$, we deduce
$$V_E^*(\eta) \geq \frac{1}{2} \cdot \frac{1}{16(d-1)(d-2)} = \frac{1}{32(d-1)(d-2)} \geq \frac{1}{32d(d-1)}.$$

$\qquad\square$

Proof of Theorem 4 from Lemma 13. Putting everything together, we see first that

$$V^*(\alpha\eta) \geq V_{\mathrm{E}}^*(\eta)\alpha^2 - 4e^{-\Delta\alpha}\alpha,$$

where $\Delta = \frac{\sqrt{1-\frac{1}{d}}}{2(d-1)}$. But then this implies

$$V^*(\alpha\eta) \geq \frac{\alpha^2}{32d(d-1)} - 4e^{-\Delta\alpha}\alpha.$$

On the other hand, $||\eta||_2^2 \leq 2$, so $\alpha^2 = \frac{||\alpha\eta||_2^2}{||\eta||_2^2} \geq \frac{||\alpha\eta||_2^2}{2}$, whence

$$V^*(\alpha\eta) \geq \frac{||\alpha\eta||_2^2}{64d(d-1)} - 4e^{-\frac{\sqrt{1-\frac{1}{d}}||\alpha\eta||_2}{2(d-1)}}||\alpha\eta||_2 \,,$$

which is the desired result. $\qquad\qquad\square$