# A direct proof of a unified law of robustness for Bregman divergence losses

Santanu Das* Jatin Batra† Piyush Srivastava‡

**Abstract**

In contemporary deep learning practice, models are often trained to near zero loss i.e. to nearly *interpolate* the training data. However, the number of parameters in the model is usually far more than the number of data points $n$, the theoretical minimum needed for interpolation: a phenomenon referred to as *overparameterization*. In an interesting piece of work that contributes to the considerable research that has been devoted to understand overparameterization, Bubeck and Sellke considered a natural notion of what it means for a model to interpolate: the model is said to interpolate when the model's training loss goes below the loss of the conditional expectation of the response given the covariate. For this notion of interpolation and for a broad class of covariate distributions (specifically those satisfying a natural notion of concentration of measure), they showed that overparameterization is necessary for *robust* interpolation i.e. if the interpolating function is required to be *Lipschitz*. Their main proof technique applies to regression with *square* loss against a scalar response, but they remark that via a connection to Rademacher complexity and using tools such as the Ledoux-Talagrand contraction inequality, their result can be extended to more general losses, at least in the case of scalar response variables. In this work, we recast the original proof technique of Bubeck and Sellke in terms of a bias-variance type decomposition, and show that this view directly unlocks a generalization to Bregman divergence losses (even for vector-valued responses), without the use of tools such as Rademacher complexity or the Ledoux-Talagrand contraction principle. Bregman divergences are a natural class of losses since for these, the best estimator is the conditional expectation of the response given the covariate, and in particular, include other practical losses such as the cross entropy loss. Our work thus gives a more general understanding of the main proof technique of Bubeck and Sellke and demonstrates its broad utility.

## 1 Introduction

The recent revolution in deep learning was driven by models that are highly *overparameterized* [15, 19, 38], i.e. models where the number of parameters far exceeds $n$, the number of training data points.[1] Since this is the naive theoretical condition needed to interpolate the training data, classical statistical theory suggests that this situation may make these models susceptible to the risk of *overfitting* to the idiosyncrasies of the training data, and thereby suffer in terms of generalizing to new inputs. On the other hand, experience with such models suggests that such overfitting does not tend to happen. Understanding the mystery of overparameterization by resolving this apparent conflict has thus attracted a lot of research, see e.g. [7, 10, 30, 40].

Another line of research focuses on *(adversarial) robustness*, i.e. whether models are susceptible to small (possibly adversarially chosen) perturbations in the input. Several existing models are known to be brittle to adversarial perturbations [4,8,31,33], which is a major issue for security [8] and reliability [31,33]. At the same time, understanding

*Sanatanu Das. Tata Institute of Fundamental Research, Mumbai. Email: dassantanu315@gmail.com.

†Jatin Batra. Tata Institute of Fundamental Research, Mumbai. Email: jatinbatra50@gmail.com.

‡Piyush Srivastava. Tata Institute of Fundamental Research, Mumbai. Email: piyush.srivastava@tifr.res.in.

We acknowledge support from DAE, India under project no. RTI4001. PS acknowledges support from Adobe Systems Incorporated via a gift to TIFR, from DST, India under project number MTR/2023/001547, and from the Infosys-Chandrasekharan Virtual Centre for Random Geometry at TIFR. The contents of this paper do not necessarily reflect the views of the funding agencies listed above.

[1] However, this may possibly not be as true for the current LLMs, see e.g. [16], where the trend is to train on web-scale data of heterogenous quality.

the nature of adversarial perturbations is an interesting tool [18, 20] to understand deep learning. In fact, in an interesting set of experiments, Madry et al. [28] gave strong evidence that overparameterization plays an important role in adversarial robustness: their emphasis was on training models robust to adversarial attacks and they observed that increasing the number of parameters in the model alone helps significantly.

In a recent line of work, Bubeck and Sellke [7] proved an exciting theorem that the requirement of robustness can be used to *explain* overparameterization in a certain sense, and they used their theorem to explain the experimental results of Madry et al. [28]. They considered the interpolation task on $n$ data points (i.e. fitting the data to "very small" loss), and any "smoothly" parameterized class of models, which includes neural networks under certain boundedness assumptions. (Note that the interpolation task deals with only the training data, with no concern for generalization: since modern deep learning models are often trained to near zero loss [3], this setup is nonetheless an interesting setup.) They showed that overparameterization is *necessary* for a certain notion of robustness for the interpolation task, namely, a low Lipschitz constant for the model. More precisely, for $d$-dimensional covariates (assuming certain concentration properties on the covariate distribution) and models with $p$ parameters, the Lipschitz constant of any model that interpolates the training data must be at least $\Omega(\sqrt{nd/p})$, with high probability over the choice of the training data (for precise notions of smooth parameterization, covariate distribution assumptions and interpolation, see Section 2).

The main proof technique of [7] is designed to understand interpolation with *square* loss, although [7] also sketches a proof for general Lipschitz losses (at least for the case of scalar responses) by exploiting a connection to Rademacher complexity and then applying the Ledoux-Talagrand contraction principle (see [23, Theorem 4.12], which improves upon [22, Theorem 5]). It is, however, also important to understand in a similar depth as the square loss other practically motivated losses. For example, the experiments of [28] (discussed above) use the cross-entropy loss for the classification problem on MNIST and CIFAR10. Some other often used problem-dependent losses are logistic loss [44], KL divergence loss [28], Mahalanobis loss [35], etc. In this work, we seek a more complete understanding of the main technique of [7] beyond the square loss, without the detour through tools such as Rademacher complexity and the Ledoux-Talagrand contraction principle.

## 1.1 Our Contribution

Looking for a generalization of the main proof technique of Bubeck and Sellke [7] from the square loss to more general losses presents two conflicting requirements: on the one hand, we would like to generalize to a sufficiently rich family of losses; on the other hand, we would need this family of losses to share with the square loss those properties which make the notion of interpolation of [7] make sense and their main technique work.

One of the many nice properties enjoyed by the square loss is that the optimal predictor of an observation $Y$ with respect to this loss, given a covariate $X$, has a crisp characterization: it is the conditional expectation $\mathbb{E}[Y|X]$ (see, e.g., [36, Sections 9.3-9.4]). As highlighted in more detail towards the end of this section, our main observation is that this simple fact and its consequences play a central role in the interpolation notion and the main technique of Bubeck and Sellke [7].

This observation leads us to the class of *Bregman divergence losses*. The Bregman divergence $D_\phi$ on $\mathbb{R}^K$, corresponding to a differentiable convex function $\phi : \mathbb{R}^K \to \mathbb{R}$ is given by

$$D_\phi(y_1, y_2) = \phi(y_1) - \phi(y_2) - \langle \nabla\phi(y_2), y_1 - y_2 \rangle .$$

This is a rich and often-used family of losses (which are not necessarily metrics; see Section 2 for examples), and includes, in particular, losses such as the square loss and the cross-entropy loss as special cases. Importantly for our purposes, this family shares with the square loss the same optimal predictor: it was shown by Banerjee, Guo and Wang [1] that this is essentially the class of losses for which the conditional expectation is the optimal predictor.

In this work, we show that this conceptual property is sufficient: we generalize the main technique of Bubeck and Sellke [7] to *Bregman divergence* losses. As described above, these include the square loss and several other commonly used losses (see Section 2 for definitions). We now proceed to an informal technical account of the result and the proof techniques; the formal statements can be found in Section 3.

Extending the notion of [7], in our setup, a function $f$ is said to $\epsilon$-overfit a set of samples with respect to a given loss if its empirical loss over the samples is at least $\epsilon$ lower than the minimum expected loss over the distribution of any function of the covariate. As in [7], our regularity condition on the distribution of the covariates is that it should be a mixture of distributions satisfying measure concentration analogous to a normalized high-dimensional Gaussian: a condition that is referred to in [7] as being a mixture of isoperimetric distributions. We defer the precise definitions of these technical terms to Section 2, and proceed to give an informal statement of our main result.

**Theorem 1.1 (Main theorem (informal, see Theorem 3.1)).** *Let $\Omega$ be a compact convex subset of $\mathbb{R}^K$ for some $K > 0$ and let $\phi : \Omega \to \mathbb{R}$ be a continuously differentiable strictly convex function. Let $D_\phi$ denote the corresponding Bregman divergence loss. For $\Delta \subseteq \mathbb{R}^d$, let $\mathcal{D}$ be a probability distribution on $\Delta \times \Omega$ such that its marginal $\mathcal{D}_X$ on $\Delta \subseteq \mathbb{R}^d$ is a mixture of $r$ isoperimetric distributions. Let $(X_i, Y_i)_{i=1}^n$ be $n$ i.i.d samples from $\mathcal{D}$. Let $\mathcal{F}$ be a family of functions that admits a bounded Lipschitz parameterization with $p$ parameters. If $n \geq \tilde{O}(K^2 r/\epsilon^2)$, then w.h.p. over the random choice of these samples, the Lipschitz constant $L$ of any function $f \in \mathcal{F}$ that $\epsilon$-overfits these samples with respect to $D_\phi$ satisfies*

$$L \geq \frac{O(1) \cdot \epsilon\sqrt{nd}}{K\sqrt{p \log(1 + O(\sqrt{K})/\epsilon)}}. \tag{1}$$

*Here, the hidden constant factors depend upon the properties of $\phi$ and the Lipschitz parameterization of $\mathcal{F}$.*

**Comparison to Bubeck and Sellke [7].**

*Approach of Bubeck and Sellke [7] in our setting.* Bubeck and Sellke [7] sketched a proof of how a Rademacher complexity view of their main technique yields generalization error bounds via Ledoux-Talagrand contraction [32,34], and remarked in passing that this view yields laws of robustness for general (Lipschitz) losses for scalar responses. Extending this view to the case of vector-valued response turns out to be trickier since the usual notion of Rademacher complexity does not enjoy a contraction principle [27, Section 6]. One must resort to coordinate-wise Rademacher complexity for which contraction does hold [27], or employ a lossy conversion from Rademacher to Gaussian complexity followed by contraction via Slepian's Lemma [2].

*Our proof technique.* Our approach side-steps use of contraction principles by recognizing that a bias-variance like decomposition lies at the heart of the main proof technique of Bubeck and Sellke [7]. Using the simple but important fact that among all $X$-measurable predictors for a random variable $Y$, the conditional expectation $\mathbb{E}[Y|X]$ minimizes not just the square loss, but any Bregman divergence loss (see, e.g., Theorem 2.3 from the work of Banerjee, Guo and Wang [1] below), we are able to replace this with a more general decomposition (see Lemmas 3.2 and 3.6 and Section 5 below). After this conceptual modification, we show that a structure similar to that of the main technique of Bubeck and Sellke [7] yields an elementary proof of the law of robustness for Bregman divergence losses, even for vector-valued responses. (The technical details of implementing the strategy are necessarily somewhat different because of the more general decomposition that we use.) Further, as we elaborate below, this more direct and elementary approach gives more flexibility for obtaining finer-grained laws of robustness for specific losses.

*Remark on loss classes.* While Bubeck and Sellke sketch an approach to go beyond square losses to Lipschitz losses, our proof is about Lipschitz Bregman divergence losses. While this may make our approach seem restrictive at first glance, as stated earlier, Bregman divergence losses are essentially the class of losses for which the notion of interpolation of Bubeck and Sellke [7] can be defined *independently of the function class* (since the conditional expectation $\mathbb{E}[Y|X]$, which is the best estimator for these losses, depends only on the data distribution and not on the function class), and hence are the natural class of losses for Bubeck and Sellke's notion of interpolation. In Section 4, we present corollaries of the main theorem for a couple of specific losses (including cross-entropy loss for vector valued responses and also the case of the square loss already considered in [7]). Our approach is flexible enough to be adapted to provide stronger bounds in specific cases, as we show for the cross-entropy loss in Corollary 4.2. For details, please refer to the remark after Corollary 4.2.

### 1.2 Related work

**Adversarial robustness experiments and overparameterization** Several works other than [28] have studied experimentally the relationship between model capacity and adversarial robustness. (Model capacity, in this context, is an informal notion that tries to capture how rich a class of functions the model captures; in studies involving neural networks, quantities such as the number of learnable parameters in the network are typically used to quantify model capacity.) Liu et al. [26] studied the loss landscape of adversarial training and observed easier adversarial training in large capacity models. Similar observations for model capacity and adversarial robustness were made by Kurakin, Goodfellow and Bengio [21] and Xie and Yuille [39].

**Other theoretical works** There is a rich body of recent theoretical work on different aspects of the link between robustness and neural network parameterizations along different lines; here we mention a few. Bubeck, Lee, and Nagaraj [6] introduced the idea of using the Lipschitz constant to measure robustness in the interpolation regime and formed the foundation for the work of Bubeck and Sellke [7]. For the square loss, Wu et al. [37] give a weaker law of robustness while relaxing the condition that the covariate distribution follows the isoperimetry condition of Bubeck and Sellke. Zhu et al. [45] gave theoretical results for robustness along more fine-grained lines of depth, width, and initialization. Gao et al. [11] and Zhang et al. [43] studied adversarial training and overparameterization from the perspective of convergence of gradient descent. Gao et al. [11] also showed interesting lower bounds for the VC dimension of a model class that can robustly interpolate arbitrary well-separated data. Another interesting line of work is studying the relationship between overparameterization and robust *generalization* (i.e. quantities such as $\mathbb{E}[\sup_{x \in \mathcal{N}(X)} \ell(f(X), Y)]$, where $\mathcal{N}(x)$ denotes a neighborhood of $x$, and $\ell$ the loss function), rather than merely robust interpolation. Recent works such as [9, 24] have studied the correlation of the Lipschitz constant of the model with the robust generalization properties of the model; see also the survey [46]. More broadly, Hassani and Javanmard [13] gave a precise analysis of robust generalization for random features regression.

## 2 Preliminaries

**Notation** We denote the standard Euclidean norm of a vector $v$ in $\mathbb{R}^d$ as $\|v\|$. The standard inner product of vectors $u, v \in \mathbb{R}^d$ is denoted $\langle u, v \rangle$.

**Bregman divergence losses** To unify different kinds of losses such as square loss and cross-entropy loss, we will use the notion of Bregman divergence losses [5] which we now describe.

**Definition 2.1 (Bregman divergence).** *Given a convex set $\Omega \subset \mathbb{R}^K$, let $\phi : \Omega \to \mathbb{R}$ be a strictly convex continuously differentiable function defined on $\Omega$. Then, the* Bregman divergence $D_\phi : \Omega \times \Omega \to \mathbb{R}$ *between two points $x, y \in \Omega$ is defined as*

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y)\ ,\ x - y \rangle, \tag{2}$$

*where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^K$.*

The Bregman divergence between two points may be viewed as a measure, depending upon the function $\phi$, of the distance between $x$ and $y$. It is however not a metric in general. It is well known that several commonly used losses may be expressed as a Bregman divergence for an appropriate choice of $\phi$: we recall some examples below.

**Example 2.2.** The following losses can be expressed as Bregman divergences:

1. **Square loss.** For $\phi : \mathbb{R}^d \to \mathbb{R}$ given as $\phi(x) = \|x\|^2$, $D_\phi(y, \hat{y}) = \|\hat{y} - y\|^2$, the square loss for regression.
2. **Mahalanobis loss.** More generally, for $\phi : \mathbb{R}^d \to \mathbb{R}$ given as $\phi(x) = x^T A x$, where $A$ is a positive definite matrix, $D_\phi(y, \hat{y}) = (y - \hat{y})^T A (y - \hat{y})$, the Mahalanobis loss for regression. Note that Mahalanobis loss is a symmetric Bregman divergence loss.

3. **KL-divergence and cross-entropy loss.** Let $\Delta_K$ denote the probability simplex in $K$ dimensions. Then, for $\phi : \Delta_K \to \mathbb{R}$ given as $\phi(x) = \sum_i^K x_i \log x_i$, $D_\phi(y, \hat{y}) = KL(y, \hat{y})$. For a 1-hot vector $y$ (i.e. $y_i = 1$ for some $i$ and $y_j = 0$ for $j \neq i$) and $\hat{y} \in (0, 1]^K \cap \Delta_K$, we slightly abuse notation and write $D_\phi(y, \hat{y})$ in the limit of the non-1 coordinates of $y$ tending to zero, and obtain $D_\phi(y, \hat{y}) = -\sum_{i=1}^K \mathbb{I}_{y_i=1} \log(\hat{y}_i)$, the cross-entropy loss for $K$-class classification. Note that both these losses are asymmetric Bregman divergence losses.

4. **Logistic loss.** This may be seen as a special case of the previous example. Consider the binary classification problem such that true label $y \in \{0, 1\}$ and predicted probability $\hat{y} \in (0, 1)$. For $\phi : [0, 1] \to \mathbb{R}$ given as $\phi(p) = p \log(p) + (1 - p) \log(1 - p)$, $D_\phi(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$, the logistic loss for classification.

Next, we state some standard and easily verified properties of the Bregman divergence that we use frequently.

1. **(Continuity).** $D_\phi$ is a continuous function on $\Omega \times \Omega$.
2. **(Nonnegativity).** $D_\phi(x, y) \geq 0$ for all $x, y$ and $D_\phi(x, y) = 0$ iff $x = y$: this is essentially equivalent to the strict convexity of $\phi$.
3. **(Triangle equality).** For any $x, y, z \in \Omega$ the following holds

$$D_\phi(x, y) = D_\phi(x, z) + D_\phi(z, y) - \langle x - z\ ,\ \nabla\phi(y) - \nabla\phi(z)\rangle. \tag{3}$$

We will also use the following Theorem of Banerjee, Guo and Wang [1]. Note that for the case of "well-behaved" random variables, this is a direct consequence of the properties above (especially of item 3).

**Theorem 2.3 ([1, Theorem 1]).** *Let $(\Omega^o, \mathcal{F}, P)$ be an arbitrary probability space, let $X$ be a random variable taking values in $\mathbb{R}^d$ and $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$ generated by $X$. Let $Y$ be any $\mathcal{F}$-measurable random variable taking values in $\mathbb{R}^K$ for which both $\mathbb{E}[Y]$ and $\mathbb{E}[\phi(Y)]$ are finite. Then, among all $\mathcal{G}$-measurable random variables of the form $f(X)$ such that $f : \mathbb{R}^d \to \mathbb{R}^K$, the conditional expectation is the unique minimizer (up to a.s. equivalence) of the expected Bregman divergence loss, i.e.,*

$$\arg\min_{f:\mathbb{R}^d \to \mathbb{R}^K} \mathbb{E}\left[D_\phi(Y, f(X))\right] = \mathbb{E}[Y|X]. \tag{4}$$

**Realistic function classes** Bubeck and Sellke [7] considered the following notion of function classes for interpolating given data. We use the same notion and name these classes as *realistic* function classes.

**Definition 2.4 (Realistic function class).** *Let $\Delta \subseteq \mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^K$ be compact sets. A class $\mathcal{F}$ of function from $\Delta$ to $\Omega$ is said to be a* $(p, J)$-realistic *function class if $\mathcal{F}$ admits a $J$-Lipschitz-parametrization by $p$ parameters. Formally, there exists a compact set $B_p \subseteq \mathbb{R}^p$ and a map $\tau : B^p \to \mathcal{F}$ such that for all $w_1, w_2 \in B_p$ and all $x \in \Delta$,*

$$\|\tau(w_1)(x) - \tau(w_2)(x)\|_2 \leq J \|w_1 - w_2\|_2 . \tag{5}$$

*The set $B_p$ is called the* parameter domain *of $\mathcal{F}$, the set $\Delta$ is called the* input domain *of $\mathcal{F}$, and the set $\Omega$ is called the* co-domain *of $\mathcal{F}$.*

We now observe that both regression and classification for bounded domains using neural networks with bounded parameters can be modeled using $(p, J)$-realistic function classes.

**Example 2.5** (Neural networks for regression.)**.** Let $B_p$ be a subset of $\mathbb{R}^p$ bounded in some unspecified norm and $\mathcal{F}_1$ be the class of $p$-parameter neural networks with parameters in $B_p$. Let $\mathcal{X}$ be a bounded domain for the covariates $x \in \mathbb{R}^d$. Then, as in [7], the natural map mapping parameter space to neural networks $\psi_1 : B_p \to \mathcal{F}_1$ satisfies (5) for all $w_1, w_2 \in \mathbb{R}^p$ and $x \in \mathcal{X}$, for some choice of $J$.

**Example 2.6** (Neural networks for classification.)**.** Let $B_p$ be a subset of $\mathbb{R}^p$ bounded in some unspecified norm and $\mathcal{X}$ be a bounded domain for covariates. Construct $\mathcal{F}_2$ by applying the Softmax operator on $\mathcal{F}_1$ as in Example 2.5 (hence $\mathcal{F}_2$ has range in $\Delta^K$). Then, since Softmax is Lipschitz (specifically with Lipschitz constant as 1 for $\ell_2$ norms on the input and output), Softmax $\circ\ \psi_1$ gives a $J$-Lipschitz parameterization for $\mathcal{F}_2$.

**Concentration of measure** We will need the notion of *sub-Gaussian* random variables. A random variable $X$ with mean $\mu$ is said to have *sub-Gaussian* parameter $\sigma > 0$ if for all real $\lambda$, it holds that $\mathbb{E}[\exp(\lambda(X-\mu))] \leq \lambda^2\sigma^2/2$. Thus, if $X_1, X_2, \ldots, X_n$ are independent sub-Gaussian random variables with parameters $\sigma_1, \sigma_2, \ldots, \sigma_n$, then their sum is also sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \sigma_i^2}$. If $X$ is sub-Gaussian with parameter $\sigma$ then the *Hoeffding inequality* states that for all $t > 0$, it holds that

$$\mathbb{P}\left[X \leq \mathbb{E}[X] - t\right] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{6}$$

It is also well known that if a random variable $X$ has support in the interval $[a, b]$, then it is sub-Gaussian with parameter $\frac{b-a}{2}$. Combined with the above discussion this leads to the usually stated form of Hoeffding's inequality: if $X_1, X_2, \ldots X_n$ are i.i.d. copies of such an $X$ then for all $t > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i < \mathbb{E}[X] - t\right] \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right). \tag{7}$$

It is well known that the assumption that $X_1, X_2, \ldots, X_n$ are i.i.d can be relaxed so that one gets the following result (usually called *Azuma's inequality*): if $Y_0 = 0, Y_1, Y_2, \ldots, Y_n$ form a martingale sequence (with respect to some filtration) such that $|Y_i - Y_{i-1}| \leq c$ holds with probability 1 for each $1 \leq i \leq n$, then $\mathbb{P}\left[Y_n \leq -nt\right] \leq \exp\left(-\frac{nt^2}{2c^2}\right)$. Via the Doob martingale construction, this inequality leads to the *bounded differences inequality*, one special case of which is the following [41]: let $V_1, V_2, \ldots V_n$ be independent mean zero random vectors in $\mathbb{R}^d$ such that $\|V_i\| \leq b$ holds with probability 1 for $1 \leq i \leq n$. Then for every $t > 0$, $\mathbb{P}\left[\left\|\sum_{i=1}^n V_i\right\| \geq \mathbb{E}\left[\left\|\sum_{i=1}^n V_i\right\|\right] + nt\right] \leq \exp\left(-nt^2/(8b^2)\right)$. Using the fact that the $V_i$ are independent and have mean zero, one has $\mathbb{E}\left[\left\|\sum_{i=1}^n V_i\right\|\right] \leq b\sqrt{n}$. Combining this with a bit of algebra, the above inequality can be simplified to the following:[2]

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n V_i\right\| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{16b^2}\right) \text{ for all } t \geq 0. \tag{8}$$

We also record the following well-known fact.

**Fact 2.7.** *There exists a positive constant $C$ such that the following is true. If $X$ is a mean-zero random variable with sub-Gaussian parameter $\sigma$, and $Z$ is any random variable (not necessarily independent of $X$) such that $\mathbb{E}[ZX] = 0$ and $|Z| \leq M$ a.s., then $ZX$ has sub-Gaussian parameter at most $CM\sigma$.*

All of the above facts are standard, and those for which no reference has been provided above can be found, e.g., in [34, Chapter 2]. In particular, a proof of the previous fact is implicit in the calculations in [34, pp. 46–47].

The following important notion (isolated in this form by Bubeck and Sellke [7]) will play an important role.

**Definition 2.8 ($c$-isoperimetry).** *Given $c > 0$, a distribution $\mathcal{D}$ on $\mathbb{R}^d$ is said to satisfy $c$-*isoperimetry *if for very $L$-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, the random variable $f(X)$, when $X$ is sampled according to $\mathcal{D}$, is sub-Gaussian with parameter $L\sqrt{c/d}$.*

The important factor to note here is the dimension-dependent factor of $1/\sqrt{d}$ in the sub-Gaussian parameter. Roughly speaking, this says that the distribution $\mathcal{D}$ exhibits a concentration of measure phenomenon similar to the *normalized* $d$-dimensional standard Gaussian distribution (i.e., with mean 0 and co-variance matrix $\frac{1}{d}\mathbf{I}$). For further discussion on this assumption, see [7]. Our results extend to mixtures of isoperimetric distributions as in [7].

## 2.1 Overfitting

We are now ready to define our notion of overfitting to the training set. Our starting point is the optimality of the conditional expectation as an estimator (Theorem 2.3). In particular, from eq. (4) in Theorem 2.3, we can conclude that $\mathbb{E}[D_\phi(Y, f(X))]$ is at least $\sigma_\phi^2 := E[D_\phi(Y, \mathbb{E}[Y|X])]$

[2] See [42] and [23, Section 6.3] for a history of related, more sophisticated inequalities and [14] and [12, Theorem 12] for related statements.

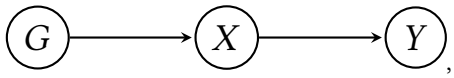
We thus say that $f$ $\epsilon$-overfits training data $\{x_i, y_i\}_{i=1}^n$ if

$$\frac{1}{n}\sum_{i=1}^{n} D_\phi(y_i, f(x_i)) < \sigma_\phi^2 - \epsilon, \tag{9}$$

i.e., if the empirical Bregman divergence loss is lower (by at least $\epsilon$) than the minimum, over all functions $f : \mathbb{R}^d \to \mathbb{R}^K$, of the expected Bregman divergence loss where $x_i \in \mathbb{R}^d$ and $y_i \in \Omega$. This notion of overfitting for Bregman divergence loss generalizes that for the square loss used in [7]; in particular, when $D_\phi$ is the square loss, it reduces to their notion.

# 3 Proof of the main theorem

We are now ready to state the main result of our paper. We assume that the model for generating the covariates $(X, Y)$ is described by the following graphical model



, (10)

where $G$ denotes the label of the mixture component. In particular, we assume that the label $Y$ is independent of the index of the mixture component, conditioned on the covariate $X$. This is the same model as the one used by Bubeck and Sellke [7, Theorem 3, points 2 and 3]. (See also Section 5.)

**Theorem 3.1.** *Let $\Omega$ be a compact convex subset of $\mathbb{R}^K$ for some $K > 0$, with $\ell_\infty$-diameter at most $d_\Omega$. Let $\phi : \Omega \to \mathbb{R}$ be a continuously differentiable strictly convex function. Let $D_\phi$ denote the corresponding Bregman divergence loss. For $\Delta \subseteq \mathbb{R}^d$, let $\mathcal{D}$ be a probability distribution on $\Delta \times \Omega$ such that $\mathcal{D}$ obeys the graphical model in eq. (10) and such that the marginal $\mathcal{D}_X$ of $\mathcal{D}$ is a mixture of $r$ distributions $(\mathcal{D}_i)_{i=1}^r$ each of which is $c$-isoperimetric for some $c > 0$. Assume that $\phi$ satisfies the following regularity condition: there exists a subset $A \subseteq \Omega$, such that a version of $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[Y|X]$ takes values only in $A$, and such that $A$ satisfies the following:*[3]

$$a_0 := \sup_{a\in A} \|a\| \le m_0 := \max_{\omega\in\Omega} \|\omega\| < \infty, \text{ and} \tag{11}$$

$$m_2 := \sup_{a\in A} |\phi(a)| \le m_1 := \max_{\omega\in\Omega} |\phi(\omega)| < \infty, \text{ and} \tag{12}$$

$$m_3 := \sup_{a\in A} \|\nabla\phi(a)\| < \infty. \tag{13}$$

*Let $\mathcal{F}$ be a $(p, J)$-realistic function class with parameter domain $B_p$, input domain $\Delta \subseteq \mathbb{R}^d$, and co-domain $\Omega$. Assume that*

1. *$\phi$ is $L_\phi$-Lipschitz on the range $R := \{f(x) | x \in \Delta \text{ and } f \in \mathcal{F}\}$ of $\mathcal{F}$, and*
2. *For each $1 \le \ell \le K$, the derivative $(\nabla\phi)_\ell$ of the $\ell^{th}$ coordinate of $\phi$ is $L_g$-Lipschitz on the range $R$.*

*Further define*

$$W := \operatorname{diam}(B_p) \text{ and } \gamma := \sup_{\substack{f\in\mathcal{F}\\ x\in\Delta}} \|\nabla\phi(f(x))\|_2 . \tag{14}$$

*Given $\epsilon, \delta \in (0, 1)$, fix a positive integer $n$ satisfying*

$$n \ge \max\left\{ \frac{300\log(\frac{10K}{\delta})}{\epsilon^2}\left(m_1 + m_2 + 2\max\{3\gamma, m_3\}(m_0 + a_0)\right)^2, \frac{2048K^2\gamma^2 r d_\Omega^2 \log(\frac{10Kr}{\delta})}{\epsilon^2} \right\}.$$

[3] Since $\Omega$ is a compact convex subset of $\mathbb{R}^K$, it can be assumed without loss of generality that a version of $\mathbb{E}[Y|X]$ taking values only in $\Omega$ exists.

*Let $(X_i, Y_i)_{i=1}^n$ be n i.i.d samples from $\mathcal{D}$. Then, with probability at least $1-\delta$ over the random choice of these samples, the Lipschitz constant L of any function $f \in \mathcal{F}$ that $\epsilon$-overfits (see eq. (9)) these samples satisfies*

$$L \geq \frac{\epsilon}{32CKd_\Omega L_g\sqrt{2c}}\sqrt{\frac{nd}{p\log(1+8JW(d_\Omega L_g K + L_\phi + \gamma)/\epsilon) + \log(5K/\delta)}}. \tag{15}$$

Note that when $\phi$ is a $C^2$ function then we can take $L_g = \max_{x\in\Omega}\left\|\nabla^2\phi(x)\right\|_{2\to 2}$. Almost all losses used in machine learning are $C^2$ losses, for example, square loss, cross-entropy defined on a bounded set away from the $\vec{0}$ vector and coordinate axes.

**Remark** Theorem 3.1 tells us that when a Bregman divergence loss (such as cross-entropy loss, logistic loss, etc.) is used for training, overparameterization becomes essential for robust interpolation (i.e., achieving a low Lipschitz constant). Note that all the regularity assumptions of Theorem 3.1 are satisfied by the cross-entropy loss, provided that the probability of each label given the covariate is bounded away from zero. For $K$-class classification, the assumption $\mathbb{P}[Y_i = 1 \mid X = x] \geq \alpha > 0$ for all $i$ and all $x$ implies the existence of $A \subseteq \tilde{\Delta}_K \cap [\alpha, 1-\alpha)^K$, where $\Omega = \tilde{\Delta}_K$ denotes the $K$-dimensional probability simplex. For further details, please refer to Corollary 4.2.

We now proceed to describe the steps of the proof of Theorem 3.1, and begin by setting up some notational conventions. We denote the sample $(X_i, Y_i)$ by $S_i$ for $1 \leq i \leq n$. Given a function $f \in \mathcal{F}$, we let $Z_i$ denote the random variable $D_\phi(Y_i, f(X_i))$ (the function $f$ would be understood from the context). When discussing a single sample, we will often drop the index and denote these as $S = (X, Y)$, and $Z = D_\phi(Y, f(X))$. Similarly, we will use the corresponding small case letters $s_i, s, x_i, x$, etc. to denote specific realizations of the corresponding random variables $S_i, S, X_i, X$, etc. The starting point of the proof is the following simple but important decomposition.

**Lemma 3.2** (Decomposition)**.** *Fix $f \in \mathcal{F}$, and with the above notation, define (in accordance with eq. (9))*

$$\sigma_\phi^2 := \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[D_\phi(Y, \mathbb{E}[Y|X])]. \tag{16}$$

*We then have*

$$Z - \sigma_\phi^2 = \Phi_1 + \Phi_2 + \sum_{i=1}^{3}\Gamma_i, \tag{17}$$

*where $\Phi_1, \Phi_2$, and $(\Gamma_i)_{i=1}^3$ are random variables defined as*

$$\Phi_1 := D_\phi(\mathbb{E}[Y|X], f(X)),\ \Phi_2 := D_\phi(Y, \mathbb{E}[Y|X]) - \sigma_\phi^2, \tag{18}$$

$$\Gamma_1 := \langle Y - \mathbb{E}[Y|X], \nabla\phi(\mathbb{E}(Y|X))\rangle,\ \Gamma_2 = \Gamma_2(f) := -\langle Y - \mathbb{E}[Y|X], \mathbb{E}\left[\nabla\phi(f(X))\right]\rangle,\ \textit{and} \tag{19}$$

$$\Gamma_3 = \Gamma_3(f) := -\langle Y - \mathbb{E}[Y|X], \nabla\phi(f(X)) - \mathbb{E}\left[\nabla\phi(f(X))\right]\rangle. \tag{20}$$

*Further, $\Phi_2$ and the $\Gamma_i$ $(1 \leq i \leq 3)$ have mean 0, and $\Phi_1$ is non-negative.*

**Remark** As discussed earlier, our proof technique relies on a bias-variance type decomposition of Bregman divergence losses. In the decomposition Lemma 3.2, the terms $\Phi_1$ and $\Phi_2$ correspond to the bias and variance components, respectively, while $\Gamma_1$, $\Gamma_2$, and $\Gamma_3$ are mean-zero terms. Specifically, the term $\Gamma_3$ involves the function $f$ evaluated at random points.

*Proof.* The decomposition follows by applying the triangle decomposition for the Bregman divergence (eq. (3)) to $Z = D_\phi(Y, f(X))$ with the "third point" chosen as $\mathbb{E}[Y|X]$. The non-negativity of $\Phi_1$ follows from the non-negativity of Bregman divergences. $\mathbb{E}[\Phi_2]$ is zero by the definition of $\sigma_\phi^2$ (eqs. (9) and (16)). We also have, for example,

$$\mathbb{E}[\Gamma_3] = \mathbb{E}[\langle Y - E[Y|X], \nabla\phi(f(X)) - \mathbb{E}[\nabla\phi(f(X))]\rangle] \tag{21}$$

$$= \mathbb{E}[\mathbb{E}[\langle Y - \mathbb{E}[Y|X], \nabla\phi(f(X)) - \mathbb{E}[\nabla\phi(f(X))]\rangle\,|X] \tag{22}$$

$$\stackrel{(\star)}{=} \mathbb{E}[\langle \mathbb{E}[Y - \mathbb{E}[Y|X]|X], \nabla\phi(f(X)) - \mathbb{E}[\nabla\phi(f(X))]\rangle] = 0, \tag{23}$$

where the equality marked ($\star$) follows because $\nabla\phi(f(X)) - \mathbb{E}[\nabla\phi(f(X))]$ is measurable with respect to the $\sigma$-field generated by $X$. The computations for $\Gamma_1$ and $\Gamma_2$ are similar. □

We now prove appropriate concentration results for sample estimates of $\Phi_2$ and the $\Gamma_i$, beginning with $\Phi_2$.

**Observation 3.3.** *Let $\Phi_2^{(1)}, \Phi_2^{(2)}, \ldots, \Phi_2^{(n)}$ be $n$ i.i.d. samples of $\Phi_2$ generated using $n$ i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from $\mathcal{D}$. Let $M_0 := \sup_{(x,y)\in\Delta\times\Omega} D_\phi(y, \mathbb{E}_{(X,Y)\sim\mathcal{D}}[Y|X = x])$. Then for any $\epsilon > 0$.*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \Phi_2^{(i)} \le -\epsilon\right) \le e^{-2n\epsilon^2/M_0^2}. \tag{24}$$

*Further, in the notation of eqs. (11) to (13) in Theorem 3.1, we can take $M_0 = m_1 + m_2 + m_3(m_0 + a_0)$.*

*Proof.* The concentration claim follows directly from the Hoeffding inequality applied to the $n$ i.i.d. random variables $D_\phi(Y_i, \mathbb{E}[Y_i|X_i])$, $1 \le i \le n$, which are all constrained to lie in the interval $[0, M_0]$. The bound on $M_0$ follows as given below. For $p \in \Omega$ and $q \in A \subseteq \Omega$ (with $\Omega$ and $A$ as defined in Theorem 3.1),

$$\begin{aligned} D_\phi(p,q) &= |\phi(p) - \phi(q) - \langle \nabla\phi(q)\,,\ p - q\rangle| && (25)\\ &\le |\phi(p)| + |\phi(q)| + \|\nabla\phi(q)\| \cdot \big(\|p\| + \|q\|\big) && (26)\\ &\le m_1 + m_2 + m_3 \cdot (m_0 + a_0)\,. \end{aligned}$$

□

An analogous application of the Hoeffding bound gives the following.

**Observation 3.4.** *Let $\Gamma_1^{(1)}, \Gamma_1^{(2)}, \ldots, \Gamma_1^{(n)}$ be $n$ i.i.d. samples of $\Gamma_1$ generated using $n$ i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from $\mathcal{D}$. Then for any $\epsilon > 0$.*

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n \Gamma_1^{(j)} \le -\epsilon\right) \le e^{-2n\epsilon^2/M_1^2}, \tag{27}$$

*where, in the notation of eqs. (11) to (13) in Theorem 3.1, we can take $M_1 = 2m_3 \cdot (m_0 + a_0)$.*

The proof for $\Gamma_2$ is also similar, since it does not depend upon evaluations of $f$ on a random point.

**Observation 3.5.** *For $f \in \mathcal{F}$, let $\Gamma_2^{(1)}(f), \Gamma_2^{(2)}(f), \ldots, \Gamma_2^{(n)}(f)$ be $n$ i.i.d. samples of $\Gamma_2$ generated using $n$ i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from $\mathcal{D}$. Then for any $\epsilon > 0$.*

$$\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } \frac{1}{n}\sum_{j=1}^n \Gamma_2^{(j)}(f) \le -\epsilon\right) \le 2e^{-2n\epsilon^2/M_2^2}, \tag{28}$$

*where, in the notation of eqs. (11) to (14) in Theorem 3.1, we can take $M_2 = 6\gamma \cdot (m_0 + a_0)$.*

*Proof.* Consider the mean-zero i.i.d random vectors $V_i := Y_i - \mathbb{E}[Y_i|X_i]$ where $1 \le i \le n$. From eq. (11), we also have that $\|V_i\| \le m_0 + a_0$ for each $i$. Let $\gamma$ be as defined in eq. (14). Applying the version of the bounded differences inequality from eq. (8) then gives

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i|X_i])\right\| \ge \frac{\epsilon}{\gamma}\right] \le 2\exp\left(-\frac{2n\epsilon^2}{M_2^2}\right). \tag{29}$$

Denote by $V$ the random vector $\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i|X_i])$. Note that

$$\frac{1}{n}\sum_{i=1}^n \Gamma_2^{(i)}(f) = -\langle V, \mathbb{E}[\nabla\phi(f(X))]\rangle \ge -\|V\| \cdot \|\mathbb{E}[\nabla\phi(f(X))]\|\,. \tag{30}$$

Further, for each $f \in \mathcal{F}$, we have, from eq. (14), that $\|\mathbb{E}[\nabla\phi(f(X))]\| \le \gamma$. Combining this bound with eqs. (29) and (30) gives the claim. □

Finally, we study $\Gamma_3$. Note that of the terms in the decomposition, this is the only term which depends upon the evaluation of a function $f$ from $\mathcal{F}$ at a random input. It is here that the notion of $c$-isoperimetry enters the picture. For simplicity of presentation, we assume in the statement of Lemma 3.6 that the number $r$ of mixture components is *one* (so that the marginal $\mathcal{D}_X$ is itself $c$-isoperimetric). The full proof for the case $r > 1$ is presented in Section 5.

**Lemma 3.6.** *Fix $f \in \mathcal{F}$ as in Theorem 3.1 and assume that $f$ is $L$-Lipschitz. Let $\Gamma_3^{(1)}(f), \Gamma_3^{(2)}(f), \ldots, \Gamma_3^{(n)}(f)$ be $n$ i.i.d. samples of $\Gamma_3(f)$ generated using $n$ i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from $\mathcal{D}$. Then for any $\epsilon > 0$.*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma_3^{(i)}(f) \le -\epsilon\right] \le K\exp\left(-\frac{nd\epsilon^2}{2cC^2K^2d_\Omega^2L^2L_g^2}\right). \tag{31}$$

*where $C$ is the universal constant from Fact 2.7, $d_\Omega$ represents the $\ell_\infty$-diameter of $\Omega$, and the remaining notation comes from item 2 and eqs. (11) to (14) in Theorem 3.1.*

*Proof.* Since $f$ remains fixed through out this proof, we drop the dependence of $\Gamma_3$ on $f$ from the notation. For $1 \le i \le n$, and $1 \le \ell \le K$, define

$$U_{i,\ell} := -(Y_{i,\ell} - \mathbb{E}[Y_{i,\ell}|X_i]) \cdot (\nabla\phi(f(X_i))_\ell - \mathbb{E}[\nabla\phi(f(X_i))_\ell]). \tag{32}$$

Note that $\sum_{i=1}^n \Gamma_3^{(i)} = \sum_{\ell=1}^K \sum_{i=1}^n U_{i,\ell}$, so that by an union bound,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma_3^{(i)} \le -\epsilon\right] \le \sum_{\ell=1}^{K}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}U_{i,\ell} \le -\epsilon/K\right]. \tag{33}$$

Further, note that for any fixed $\ell \in [K]$, the random variables $U_{1,\ell}, U_{2,\ell}, \ldots, U_{n,\ell}$ are i.i.d, and are also mean-zero (by the same argument as in eq. (23)). We now proceed to estimate their sub-Gaussian parameter. From item 2 in Theorem 3.1, we see that the function $(\nabla\phi)_\ell \circ f : \mathbb{R}^d \to \mathbb{R}$ is $L_g \cdot L$-Lipschitz. From the $c$-isoperimetric assumption on the distribution of the $X_i$, it therefore follows that for each $i \in [n], \ell \in [K]$, the mean-zero random variable $V_{i,\ell} := \nabla\phi(f(X_i))_\ell - \mathbb{E}[\nabla\phi(f(X_i))_\ell]$ is sub-Gaussian with parameter $L_gL\sqrt{c/d}$. Further, the random variables $T_{i,\ell} := -(Y_{i,\ell} - \mathbb{E}[Y_{i,\ell}|X_i])$ satisfy $\left|T_{i,\ell}\right| \le \min\{d_\Omega, m_0 + a_0\}$ (by eq. (11) and the assumption on the diameter of $\Omega$). From Fact 2.7, it thus follows that the $U_{i,\ell}$ are all sub-Gaussian with parameter $Cd_\Omega L_gL\sqrt{c/d}$. Applying the Hoeffding equality to each of the $K$ terms on the right hand side of eq. (33) then gives

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma_3^{(i)} \le -\epsilon\right] \le K\exp\left(-\frac{nd\epsilon^2}{2cC^2K^2d_\Omega^2L^2L_g^2}\right). \qquad \square$$

We also need the following simple lemma.

**Lemma 3.7.** *With the setup of Theorem 3.1, let $f$ and $g$ be functions in $\mathcal{F}$ such that $\sup_{x\in\Delta}\|f(x) - g(x)\| \le \nu$. Then for any $(x, y)$ in the support of $\mathcal{D}$, we have*

$$\left|D_\phi(y, f(x)) - D_\phi(y, g(x))\right| \le \nu \cdot \left(d_\Omega L_g K + L_\phi + \gamma\right) \tag{34}$$

*Proof.* It is enough to prove the upper bound for $D_\phi(y, f(x)) - D_\phi(y, g(x))$ (the bound on the absolute value follows by interchanging the role of $f$ and $g$). For this we again employ the triangle decomposition eq. (3).

$$D_\phi(y, f(x)) - D_\phi(y, g(x)) = D_\phi(g(x), f(x)) - \langle y - g(x), \nabla\phi(f(x)) - \nabla\phi(g(x))\rangle\,. \tag{35}$$

Here, the first term is at most $\nu \cdot (L_\phi + \gamma)$, while the second term is at most $d_\Omega \nu L_g K$, where the parameters are as defined in the statement of Theorem 3.1. The claim thus follows. $\square$

We are now ready to prove Theorem 3.1.

*Proof of Theorem 3.1.* Given the above lemmas, the structure of the proof is similar to that in the work of Bubeck and Sellke [7]. The goal is to show that if $L$ is not too large, then with high-probability, no $L$-Lipschitz function in $\mathcal{F}$ can $\epsilon$-overfit the observed data. Lemma 3.6 essentially establishes this for any particular $f \in \mathcal{F}$, and to perform a "union bound" over the uncountable set $\mathcal{F}$, one needs to pass to an appropriate finite net. We now proceed to the details.

For a given $f \in \mathcal{F}$, let $\left(\Phi_1^{(i)}\right)_{i=1}^n$, $\left(\Phi_2^{(i)}\right)_{i=1}^n$, $\left(\Gamma_1^{(i)}\right)_{i=1}^n$, $\left(\Gamma_2^{(i)}(f)\right)_{i=1}^n$ and $\left(\Gamma_3^{(i)}(f)\right)_{i=1}^n$ be i.i.d. sequences obtained by decomposing each $Z_i - \sigma_\phi^2$ according to eq. (17).

Fix $L$ to be equal to the lower bound claimed in Theorem 3.1. Let $\mathcal{F}_L \subseteq \mathcal{F}$ denote the set of all $L$-Lipschitz functions in $\mathcal{F}$. Now, since the events considered in Observations 3.3 to 3.5 do not depend upon the choice of $f$, we have

$$\mathbb{P}\left(\exists f \in \mathcal{F}_L \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \Phi_2^{(i)} \le -\epsilon/8\right) \le e^{-2n(\epsilon/8)^2/M_0^2}, \text{ and} \tag{36}$$

$$\mathbb{P}\left(\exists f \in \mathcal{F}_L \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \Gamma_1^{(i)} \le -\epsilon/8\right) \le e^{-2n(\epsilon/8)^2/M_1^2}, \text{ and} \tag{37}$$

$$\mathbb{P}\left(\exists f \in \mathcal{F}_L \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \Gamma_2^{(i)}(f) \le -\epsilon/8\right) \le 2Ke^{-2n(\epsilon/8)^2/M_2^2}. \tag{38}$$

Let the events on the LHS above be denoted $E_0, E_1$ and $E_2$ for future reference. We now proceed to analyze $\Gamma_3$. Recall that by a standard argument a $\nu/2$-net for $\mathcal{F}$ can be modified to form a $\nu$-net of $\mathcal{F}_L$ of the same size, all of whose elements are also elements of $\mathcal{F}_L$. Further an $\epsilon'$-net of $B_p$ maps (under the map $\tau$ in the definition of a $(p, J)$-realistic class) to a $J\epsilon'$-net for $\mathcal{F}$ (under the sup norm). Note also that by standard arguments $B_p$ has an $\epsilon'$-net of size at most $(1 + 2W/\epsilon')^p$. Set $\nu := \frac{\epsilon}{2(d_\Omega L_g K + L_\phi + \gamma)}$. Then, from the above arguments, we see that $\mathcal{F}_L$ has a $\nu$-net $\mathcal{F}_{L,\nu} \subseteq \mathcal{F}_L$ of size at most $(1 + 4WJ/\nu)^p \le \exp(4pWJ/\nu)$. Applying Lemma 3.6 and taking a union bound over $\mathcal{F}_{L,\nu}$ then yields

$$\mathbb{P}\left[\exists f \in \mathcal{F}_{L,\nu} \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \Gamma_3^{(i)}(f) \le -\epsilon/8\right] \le K\left|\mathcal{F}_{L,\nu}\right| \cdot \exp\left(-\frac{nd(\epsilon/8)^2}{2cC^2K^2d_\Omega^2L^2L_g^2}\right). \tag{39}$$

Let the event above be denoted $E_3$. Now, from Lemma 3.2,

$$\frac{1}{n}\sum_{i=1}^n Z_i(f) - \sigma_\phi^2 = \frac{1}{n}\sum_{i=1}^n \Phi_1^{(i)} + \frac{1}{n}\sum_{i=1}^n \Phi_2^{(i)} + \frac{1}{n}\sum_{i=1}^n \Gamma_1^{(i)} + \frac{1}{n}\sum_{i=1}^n \Gamma_2^{(i)}(f) + \frac{1}{n}\sum_{i=1}^n \Gamma_3^{(i)}(f) \tag{40}$$

Since the $\Phi_1^{(i)}$ are all non-negative, a union bound gives

$$\mathbb{P}\left[\exists f \in \mathcal{F}_{L,\nu} \text{ s.t. } \frac{1}{n}\sum_{i=1}^n Z_i(f) - \sigma_\phi^2 \le -\epsilon/2\right] \le \sum_{j=0}^3 \mathbb{P}\left[E_i\right], \tag{41}$$

where the $E_i$ are the events considered in eqs. (36) to (39) above. Finally, since $\mathcal{F}_{L,\nu}$ is a $\nu$-net for $\mathcal{F}_L$ under the sup norm, it follows from Lemma 3.7 and the choice of $\nu$ that (recall also that $Z_i(f) = D_\phi(Y_i, f(X_i))$)

$$\mathbb{P}\left[\exists f \in \mathcal{F}_L \text{ s.t. } \frac{1}{n}\sum_{i=1}^n Z_i(f) - \sigma_\phi^2 \le -\epsilon\right] \le \mathbb{P}\left[\exists f \in \mathcal{F}_{L,\nu} \text{ s.t. } \frac{1}{n}\sum_{i=1}^n Z_i(f) - \sigma_\phi^2 \le -\epsilon/2\right]. \tag{42}$$

Combining eq. (42) with eq. (41) and using eqs. (36) to (39), we thus get

$$\mathbb{P}\left[\exists f \in \mathcal{F}_L \text{ s.t. } \frac{1}{n}\sum_{i=1}^n Z_i(f) - \sigma_\phi^2 \le -\epsilon\right] \le K\left|\mathcal{F}_{L,\nu}\right| \cdot e^{-\frac{nd(\epsilon/8)^2}{2cC^2K^2d_\Omega^2L^2L_g^2}} + 2K\sum_{j=0}^2 e^{-2n(\epsilon/8)^2/(M_j)^2}. \tag{43}$$

Using the size bound derived above for $\mathcal{F}_{L,\nu}$, and the choice of $L$ and $n$, we conclude that the right-hand side is at most $\delta$, since each term is at most $\delta/4$. □

# 4 Specializing the result to specific losses

In this section we show how Theorem 3.1 leads to laws of robustness for specific losses. We first verify that we can obtain the result of Bubeck and Sellke [7] for the square loss as a special case. The calculations underlying the following verifications can be found in Appendix A.

Consider the regression setting using the mean squared error (MSE) loss. Assume that $n$ input covariates and labels $((x_i, y_i))_{i=1}^n$ in $\mathbb{R}^d \times [-M, M]^K$ are drawn from the distribution $\mathcal{D}$. Let the hypothesis class for the regression problem be a $(p, J)$-realistic function class $\mathcal{F} = \{f : \mathbb{R}^d \to [-M, M]^K\}$ with parameter domain $B_p$ having diameter $W$. We specialize eq. (9) to this setting (using item 1 of Example 2.2), and say that a function $f \in \mathcal{F}$ $\epsilon$-overfits the data if

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \leq \sigma^2 - \epsilon, \tag{44}$$

where $\sigma_\phi^2 = \sigma^2 = \mathbb{E}[\mathrm{Var}[Y|X]]$.

**Corollary 4.1 (Law of robustness for regression).** *Consider the above regression setting, and assume that for some $c > 0$, the marginal $\mathcal{D}_X$ of $\mathcal{D}$ on the covariates is a mixture of $r$ $c$-isoperimetric distributions.*

*Given $\epsilon, \delta \in (0, 1)$, assume that the number $n$ of samples satisfies $n \geq \left(C_1 M^4 K^3 r \log\left(\frac{10Kr}{\delta}\right)\right)/\epsilon^2$, where $C_1$ is an absolute constant. Then, with probability at least $1 - \delta$ over the samples, the Lipschitz constant $L$ of any function $f \in \mathcal{F}$ that $\epsilon$-overfits the samples satisfies*

$$L \geq \frac{\epsilon}{128CKM\sqrt{2c}}\sqrt{\frac{nd}{p\log(1 + 64JWKM/\epsilon) + \log(5K/\delta)}}. \tag{45}$$

As another example, we consider the classification problem with a suitable loss. Consider the $K$-class classification setting with the cross-entropy loss. Assume that $n$ input covariates and labels $((x_i, y_i))_{i=1}^n$ in $\mathbb{R}^d \times \{0, 1\}^K$ are drawn from the distribution $\mathcal{D}$: here we assume that the labels are one-hot encoded as $K$-dimensional binary vectors. Let the hypothesis class for the classification problem be a $(p, J)$-realistic function class $\mathcal{F} = \{\mathrm{Softmax}(g) | g : \mathbb{R}^d \to [-M, M]^K\}$ such that there exists a compact set $B_p \subseteq \mathbb{R}^p$ and a $J-$Lipschitz map $\tau : B^p \to \mathcal{F}$, where $W = \mathrm{diam}(B_p)$. Again, we specialize eq. (9) to this setting (using item 3 of Example 2.2) and say that a function $f \in \mathcal{F}$ $\epsilon$-overfits the data if

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{K} -\mathbb{I}_{\{y_i=\ell\}}\log(f(x_i)_\ell) \leq \sigma_\phi^2 - \epsilon, \tag{46}$$

where $\sigma_\phi^2 = H(Y|X)$ is the conditional entropy of $Y$ given $X$. In addition to the positivity of $\sigma_\phi^2$, we also need another regularity condition on $\mathcal{D}$, which is that the probability of each label is bounded away from 0, even conditioned on the covariate: the number of samples needed for our result then has a mild poly-logarithmic dependence on this lower bound.

**Corollary 4.2 (Law of robustness for classification).** *Consider the above $K-$class classification setting, and assume that for some $c > 0$, the marginal $\mathcal{D}_X$ of $\mathcal{D}$ on the covariates is a mixture of $r$ $c$-isoperimetric distributions. We further assume the regularity condition that there exists an $\alpha > 0$ such that $\mathbb{P}[Y_i = 1|X = x] \geq \alpha$ for all $i \in [K]$ and $x \in \mathbb{R}^d$ (recall that the label $Y$ is a one-hot encoded vector).*

*Define $a_0 := 1 + 2M + \log K$. Given $\epsilon, \delta \in (0, 1)$, assume that the number $n$ of samples satisfies*

$$n \geq \frac{C_1 K^3 r \log\left(\frac{10Kr}{\delta}\right)}{\epsilon^2} \cdot \max\left\{a_0, 1 + |\log\alpha|\right\}^2,$$

*where $C_1$ is an absolute constant. Then, with probability at least $1 - \delta$ over the random choice of these samples, the Lipschitz constant $L$ of any function $f$ for which $\mathrm{Softmax}(f) \in \mathcal{F}$ and which $\epsilon$-overfits these samples (according to*

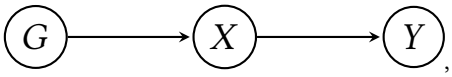
*eq.* (46)*) satisfies*

$$L \geq \frac{\epsilon}{64CK\sqrt{2c}} \sqrt{\frac{nd}{p \log(1 + 8JW(e^{2M}K^2 + \sqrt{K}(1 + 2M + \log K))/\epsilon) + \log(5K/\delta)}}. \tag{47}$$

**Remark** A direct application of the Theorem 3.1 gives only a lower bound on the Lipschitz constant of interpolating estimators of the form Softmax($f$). However, in this case, a more appropriate quantity might be the Lipschitz constant of the function $f$, before the Softmax layer is applied. Our direct proof adapts to this and gives a better bound (by a factor of $Ke^{2M}$) for the Lipschitz constant of $f$ whenever Softmax($f$) is interpolating. For details, please look at the proof of Corollary 4.2 in the Appendix A.

# 5 Extension to mixtures

*Mixture model*: We assume that the model for generating the covariates $(X, Y)$ is described by the following graphical model




, (48)

where $G$ denotes the label of the mixture component. In particular, we assume that the label $Y$ is independent of the index of the mixture component, conditioned on the covariate $X$. This is the same model as the one used by Bubeck and Sellke [7, Theorem 3, points 2 and 3].[4]

The extension of the proof of Theorem 3.1 given in Section 3 to the case of $r > 1$ mixture components has a structure similar to the similar extension by Bubeck and Sellke [7]; however, we provide the details for completeness. The main technical step is replacing Lemma 3.6 by a more general analog. Towards this end, we proceed to set up some notation. In the following, we also import all the notation from the statements of Theorem 3.1 and Lemma 3.2.

Let $f \in F$ be given. For $1 \leq i \leq n$ and $1 \leq \ell \leq K$, define the following random variables.

$$T_{i,\ell} := -(Y_{i,\ell} - \mathbb{E}[Y_{i,\ell}|X_i]), \tag{49}$$

$$V(f)_{i,\ell} := \nabla\phi(f(X_i))_\ell - \mathbb{E}[\nabla\phi(f(X_i))_\ell], \tag{50}$$

$$\hat{V}(f)_{i,\ell} := \nabla\phi(f(X_i))_\ell - \mathbb{E}[\nabla\phi(f(X_i))_\ell|G_i], \text{ and} \tag{51}$$

$$\tilde{V}(f)_{i,\ell} := \mathbb{E}[\nabla\phi(f(X_i))_\ell|G_i] - \mathbb{E}[\nabla\phi(f(X_i))_\ell], \tag{52}$$

and set

$$U(f)_{i,\ell} := T_{i,\ell}V(f)_{i,\ell} = T_{i,\ell}\hat{V}(f)_{i,\ell} + T_{i,\ell}\tilde{V}(f)_{i,\ell}. \tag{53}$$

Note that

$$\sum_{i=1}^{n} \Gamma_3^{(i)}(f) = \sum_{\ell=1}^{K} \sum_{i=1}^{n} U(f)_{i,\ell}. \tag{54}$$

Note that for any fixed $\ell \in [K]$ and $f \in \mathcal{F}$, the random variables $U(f)_{1,\ell}, U(f)_{2,\ell}, \ldots, U(f)_{n,\ell}$ are i.i.d, and are also mean-zero (by the same argument as in eq. (23)).

To further study the distribution of the $U_{i,\ell}$, we will use the random variables $G = (G_i)_{i=1}^n$, taking values in $[r]$, which denote the mixture component distribution $\mathcal{D}_{G_i}$ from which the $i$th covariate $X_i$ is sampled. Note that the $G_i$ are i.i.d.

[4] See also the computation leading to and following eq. (2.5) in [7], where this assumption has been used to (implicitly) deduce that the random variable in the left hand side of their eq. (2.5) has mean 0: this can fail if $Y$ is not independent of the mixture component when conditioned on $X$.

Note that we have the uniform bound $\left|T_{i,\ell}\right| \leq \min\{d_\Omega, m_0 + a_0\}$ (by eq. (11) and the assumption on the diameter of $\Omega$). We also note that

$$\mathbb{E}[Y_{i,\ell}|G] = \mathbb{E}[Y_{i,\ell}|G_i] = \mathbb{E}[\mathbb{E}[Y_{i,\ell}|G_i, X_i]|G_i] = \mathbb{E}[\mathbb{E}[Y_{i,\ell}|X_i]|G_i], \tag{55}$$

where the first equality follows from the independence of the samples, the second from the tower property of conditional expectation, and the last from the conditional independence of $Y_i$ from $G_i$ given $X_i$. Combined again with the independence of the samples, eq. (55) implies that

$$\mathbb{E}[T_{i,\ell}|G] = \mathbb{E}[T_{i,\ell}|G_i] = 0. \tag{56}$$

We now have the following two lemmas.

**Lemma 5.1.** *With the notation above, we have, for every $L$-Lipschitz $f \in \mathcal{F}$ every $1 \leq \ell \leq K$, and every $\epsilon > 0$*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\hat{V}(f)_{i,\ell} \leq -\epsilon\right] \leq \exp\left(-\frac{nd\epsilon^2}{2cC^2 d_\Omega^2 L^2 L_g^2}\right). \tag{57}$$

*Proof.* Since $f$ remains fixed through the proof of the lemma, we drop the dependence of $\hat{V}$ on $f$ from the notation. We first note that $\mathbb{E}[\hat{V}_{i,\ell}|G_i] = 0$, and also that $\hat{V}_{i,\ell}$ is measurable with respect to the $\sigma$-field generated by the random variables $X_i$ and $G_i$. Note also that

$$\mathbb{E}[T_{i,\ell}|G_i, X_i] = \mathbb{E}[Y_{i,\ell}|G_i, X_i] - \mathbb{E}[\mathbb{E}[Y_{i,\ell}|X_i]|G_i, X_i] = \mathbb{E}[Y_{i,\ell}|X_i] - \mathbb{E}[Y_{i,\ell}|X_i] = 0, \tag{58}$$

where the first term has been simplified using the conditional independence of $Y_i$ from $G_i$ given $X_i$, and the second term using the tower property of conditional expectation. We thus get

$$\mathbb{E}[T_{i,\ell}\hat{V}_{i,\ell}|G_i] = \mathbb{E}[\mathbb{E}[T_{i,\ell}\hat{V}_{i,\ell}|X_i, G_i]|G_i] = \mathbb{E}[\hat{V}_{i,\ell}\mathbb{E}[T_{i,\ell}|X_i, G_i]|G_i] \overset{\text{eq. (58)}}{=} 0, \tag{59}$$

where the first equality is the tower property, and the second uses the observation from above that $\hat{V}_{i,\ell}$ is measurable with respect to the $\sigma$-field generated by the random variables $G_i$ and $X_i$. From the independence of the samples, we also have that the above equalities hold when conditioning on $G$:

$$\mathbb{E}[T_{i,\ell}\hat{V}_{i,\ell}|G] = \mathbb{E}[T_{i,\ell}\hat{V}_{i,\ell}] = 0. \tag{60}$$

Now, from item 2 in Theorem 3.1, we see that the function $(\nabla\phi)_\ell \circ f : \mathbb{R}^d \to \mathbb{R}$ is $L_g \cdot L$-lipschitz. From the $c$-isoperimetric assumption on each mixture component of the co-variate distribution, it therefore follows that *conditioned on $G$*, for each $i \in [n]$, $\ell \in [K]$, the (conditionally) mean-zero random variable $\hat{V}_{i,\ell}$ is sub-gaussian with parameter $L_g L\sqrt{c/d}$. Combining eq. (60) with the absolute bound of $d_\Omega$ on $T_{i,\ell}$ given above and with Fact 2.7, it thus follows that *conditioned on $G$*, the random variables $T_{i,\ell}\hat{V}_{i,\ell}$ are mean-zero and sub-gaussian with parameter $Cd_\Omega L_g L\sqrt{c/d}$. Further, by the independence of samples, it also follows that for any fixed $\ell$, (even when conditioned on $G$) they are independent. Thus, the Hoeffding inequality gives that for any $1 \leq \ell \leq k$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\hat{V}_{i,\ell} \leq -\epsilon \middle| G\right] \leq \exp\left(-\frac{nd\epsilon^2}{2cC^2 d_\Omega^2 L^2 L_g^2}\right). \tag{61}$$

We take expectations on both sides to get the claimed "un-conditioned" bound. □

**Lemma 5.2.** *With the notation above, we have for every $1 \leq \ell \leq K$ and every $\epsilon > 0$*

$$\mathbb{P}\left[\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\tilde{V}(f)_{i,\ell} \leq -\epsilon\right] \leq 2r\exp\left(\frac{-n\epsilon^2}{8\gamma^2 r d_\Omega^2}\right). \tag{62}$$

*Proof.* Note that for every $f \in \mathcal{F}$ the random variables $\tilde{V}(f)_{i,\ell}$ are $G_i$-measurable and are bounded as $\left|\tilde{V}(f)_{i,\ell}\right| \leq 2\gamma$ (by eq. (14)). We thus see that for any fixed $\ell$, *conditioned on the mixture labels $G$*,

1. the random variables $\tilde{V}(f)_{i,\ell}$ become deterministic with absolute value at most $2\gamma$. In fact, $\tilde{V}(f)_{i,\ell} = \tilde{V}(f)_{j,\ell}$ whenever $G_i = G_j$.
2. the $T_{i,\ell}$ are *independent* (though not identically distributed), have absolute value at most $d_\Omega$, and have (conditional) mean 0 (as argued above in eq. (56), and the paragraph preceding it).

Following Bubeck and Sellke [7], we thus define the (random) sets $S_k := \{i \in [n] | G_i = k\}$, for $1 \le k \le r$ (note that the random sets $S_k$ are deterministic conditioned on $G$), and then use the Cauchy-Schwarz inequality to obtain

$$\sum_{k=1}^{r} \sqrt{|S_k|} \le \sqrt{r \cdot \sum_{k=1}^{r} |S_k|} \le \sqrt{nr}. \tag{63}$$

From item 1 above, $\tilde{V}(f)_{i,\ell}$ is the same for each $i \in S_k$, and is at most $2\gamma$ in absolute value. Thus, for each $1 \le k \le r$ we have

$$\mathbb{P}\left[\inf_{f\in\mathcal{F}} \sum_{i\in S_k} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le -\epsilon\sqrt{\frac{n\,|S_k|}{r}}\middle| G\right] \le \mathbb{P}\left[\left|\sum_{i\in S_k} T_{i,\ell}\right| \ge \frac{\epsilon}{2\gamma}\sqrt{\frac{n\,|S_k|}{r}}\middle| G\right]. \tag{64}$$

From item 2 above, the Hoeffding inequality can be applied to the right hand side above, so that we get

$$\mathbb{P}\left[\inf_{f\in\mathcal{F}} \sum_{i\in S_k} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le -\epsilon\sqrt{\frac{n\,|S_k|}{r}}\middle| G\right] \le 2\exp\left(\frac{-n\epsilon^2}{8\gamma^2 r d_\Omega^2}\right). \tag{65}$$

Finally, by a union bound we have

$$\begin{aligned}
\mathbb{P}\left[\inf_{f\in\mathcal{F}} \sum_{i=1}^{n} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le -n\epsilon\middle| G\right] &\le \sum_{k=1}^{r} \mathbb{P}\left[\inf_{f\in\mathcal{F}} \sum_{i\in S_k} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le \frac{-n\epsilon\sqrt{|S_k|}}{\sum_{t=1}^{r}\sqrt{|S_t|}}\middle| G\right] \\
&\overset{\text{eq. (63)}}{\le} \sum_{k=1}^{r} \mathbb{P}\left[\inf_{f\in\mathcal{F}} \sum_{i\in S_k} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le -\epsilon\sqrt{\frac{n\,|S_k|}{r}}\middle| G\right] \\
&\overset{\text{eq. (65)}}{\le} 2r\exp\left(\frac{-n\epsilon^2}{8\gamma^2 r d_\Omega^2}\right).
\end{aligned}$$

The claim now follows by taking expectations of both sides in the above. □

We can now describe the modifications needed to complete the proof of Theorem 3.1. These modifications are described in terms of the notation set up above and in the proof for the case $r = 1$ given in Section 3. In particular, the $\nu$-net $\mathcal{F}_{L,\nu}$ of the subset $\mathcal{F}_L$ of $L$-lipschitz functions in $\mathcal{F}$ is as defined in that proof.

*Proof of Theorem 3.1: Case $r > 1$.* Most of the proof of the theorem remains the same, except for use of Lemma 3.6 in the derivation of eq. (39), which now has to be replaced by applications of Lemmas 5.1 and 5.2. First, using the decompositions in eqs. (53) and (54) above and a union bound, we have

$$\begin{aligned}
&\mathbb{P}\left[\exists f \in \mathcal{F}_{L,\nu} \text{ s.t. } \frac{1}{n}\sum_{i=1}^{n} \Gamma_3^{(i)}(f) \le -\epsilon/8\right] \\
&\qquad \le \sum_{\ell=1}^{K}\left(\mathbb{P}\left[\inf_{f\in\mathcal{F}_{L,\nu}} \frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le \frac{-\epsilon}{16K}\right] + \mathbb{P}\left[\inf_{f\in\mathcal{F}_{L,\nu}} \frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\hat{V}(f)_{i,\ell} \le \frac{-\epsilon}{16K}\right]\right).
\end{aligned} \tag{66}$$

By construction, $\mathcal{F}_{L,\nu} \subseteq \mathcal{F}$, so we bound the first term using Lemma 5.2.

$$\begin{aligned}
\sum_{\ell=1}^{K} \mathbb{P}\left[\inf_{f\in\mathcal{F}_{L,\nu}} \frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le \frac{-\epsilon}{16K}\right] &\le \sum_{\ell=1}^{K} \mathbb{P}\left[\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\tilde{V}(f)_{i,\ell} \le \frac{-\epsilon}{16K}\right] \\
&\le 2Kr\exp\left(\frac{-n(\epsilon/16)^2}{8K^2\gamma^2 r d_\Omega^2}\right).
\end{aligned} \tag{67}$$

For the second term, we use Lemma 5.1 and a union bound over the $\nu$-net $\mathcal{F}_{L,\nu}$ of the set of $L$-lipschitz function in $\mathcal{F}$, exactly as in the argument leading to eq. (39). This gives,

$$\sum_{i=1}^{K} \mathbb{P}\left[\inf_{f\in\mathcal{F}_{L,\nu}} \frac{1}{n}\sum_{i=1}^{n} T_{i,\ell}\hat{V}(f)_{i,\ell} \le \frac{-\epsilon}{16K}\right] \le K\left|\mathcal{F}_{L,\nu}\right| \cdot \exp\left(-\frac{nd(\epsilon/16)^2}{2cC^2K^2d_\Omega^2L^2L_g^2}\right). \tag{68}$$

Together, these two computations, when substituted into eq. (66), give the following more general version of eq. (39).

$$\mathbb{P}\left[\exists f \in \mathcal{F}_{L,\nu} \text{ s.t. } \frac{1}{n}\sum_{i=1}^{n}\Gamma_3^{(i)}(f) \le -\epsilon/8\right] \quad \le \quad 2Kr\exp\left(\frac{-n(\epsilon/16)^2}{8K^2\gamma^2 r d_\Omega^2}\right) \quad + \quad K\left|\mathcal{F}_{L,\nu}\right| \quad \cdot \quad \exp\left(-\frac{nd(\epsilon/16)^2}{2cC^2K^2d_\Omega^2L^2L_g^2}\right). \tag{69}$$

We then proceed with the argument exactly as before, replacing all usages of eq. (39) by eq. (69). The final bound in eq. (43) then gets modified to the following.

$$\mathbb{P}\left[\exists f \in \mathcal{F}_L \text{ s.t. } \frac{1}{n}\sum_{i=1}^{n} Z_i(f) - \sigma_\phi^2 \le -\epsilon\right] \le K\left|\mathcal{F}_{L,\nu}\right| \cdot e^{-\frac{nd(\epsilon/16)^2}{2cC^2K^2d_\Omega^2L^2L_g^2}} + 2Kr\exp\left(\frac{-n(\epsilon/16)^2}{8K^2\gamma^2 r d_\Omega^2}\right) + 2K\sum_{j=0}^{2} e^{-2n(\epsilon/8)^2/(M_j)^2}. \tag{70}$$

As in the proof for the $r = 1$ case, each term above is at most $\delta/5$ by the choice of the parameters. □

# 6 Discussion

In this paper, we gave a more comprehensive understanding of the law of robustness of Bubeck and Sellke [7] for interpolation by considering Bregman divergence losses. In applications, the objective of interest is usually robust generalization rather than robust interpolation. We leave the extension of the line of work of this paper to robust generalization as open. We also suggest a few specific directions of inquiry:

1. To understand robust generalization in practice, *local* notions of the Lipschitz constant are often tighter than the global notion [17, 29]. Bubeck and Sellke [7] remark that the expected squared norm of the gradient instead of the (global) Lipschitz constant does not lead to a similar law of robustness. How can we get a better understanding of robustness for local notions of Lipschitz constants?

2. The current line of work focuses on models that overfit training data, as such models heralded the recent deep learning revolution. However, modern model training procedures often do not fall into this paradigm. In fact they may practice 'early stopping', for which a theory of overparameterization has recently been proposed [25]. Connecting the line of work of robust interpolation to practical 'underfitting' setups such as early stopping is an important research problem.

# References


[1] Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

[2] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[3] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

[4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013.

[5] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, January 1967.

[6] Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, pages 804–820. PMLR, 2021.

[7] Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

[8] Alesia Chernikova, Alina Oprea, Cristina Nita-Rotaru, and BaekGyu Kim. Are self-driving cars secure? Evasion attacks against deep neural networks for steering angle prediction. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 132–137. IEEE, 2019.

[9] Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, and Simon Kornblith. Generalised Lipschitz regularisation equals distributional robustness. In *International Conference on Machine Learning*, pages 2178–2188. PMLR, 2021.

[10] Ouns El Harzli, Bernardo Cuenca Grau, Guillermo Valle-Pérez, and Ard A. Louis. Double-descent curves in neural networks: A new perspective using Gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11856–11864, March 2024.

[11] Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] David Gross. Recovering Low-Rank Matrices From Few Coefficients in Any Basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

[13] Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, 52(2):441–465, 2024.

[14] Thomas P Hayes. A large-deviation inequality for vector-valued martingales. Available from https://www.cs.unm.edu/~hayes/papers/VectorAzuma/VectorAzuma20030207.pdf, 2003.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[17] Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local Lipschitz bounds. *Advances in Neural Information Processing Systems*, 34:22745–22757, 2021.

[18] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[20] Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Theoretical understanding of learning from adversarial perturbations. In *International Conference on Learning Representations*, 2024.

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

[22] M. Ledoux and M. Talagrand. Comparison Theorems, Random Geometry and Some Limit Theorems for Empirical Processes. *The Annals of Probability*, 17(2):596–631, April 1989.

[23] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Springer, Berlin, Heidelberg, 1991.

[24] Binghui Li, Jikai Jin, Han Zhong, John Hopcroft, and Liwei Wang. Why robust generalization in deep learning is difficult: Perspective of expressive power. *Advances in Neural Information Processing Systems*, 35:4370–4384, 2022.

[25] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR, 2020.

[26] Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems*, 33:21476–21487, 2020.

[27] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

[28] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[29] Ramchandran Muthukumar and Jeremias Sulam. Adversarial robustness of sparse local Lipschitz predictors. *SIAM Journal on Mathematics of Data Science*, 5(4):920–948, 2023.

[30] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings*, 2015.

[31] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[32] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, 2014.

[34] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[35] Siyang Wen, Wei Guo, Yi Liu, and Ruijie Wu. Rotated object detection via scale-invariant mahalanobis distance in aerial images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[36] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, 1991.

[37] Yihan Wu, Heng Huang, and Hongyang Zhang. A law of robustness beyond isoperimetry. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37439–37455. PMLR, July 2023.

[38] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5393–5402. PMLR, 10–15 Jul 2018.

[39] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020.

[40] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

[41] V. V. Yurinskii. Exponential bounds for large deviations. *Theory of Probability & Its Applications*, 19(1):154–155, December 1974.

[42] V. V Yurinskiĭ. Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis*, 6(4):473–499, December 1976.

[43] Yi Zhang, Orestis Plevrakis, Simon S Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. *Advances in Neural Information Processing Systems*, 33:679–688, 2020.

[44] Zihan Zhang, Lei Shi, and Ding-Xuan Zhou. Classification with deep neural networks and logistic loss. *Journal of Machine Learning Research*, 25(125):1–117, 2024.

[45] Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). *Advances in neural information processing systems*, 35:36094–36107, 2022.

[46] Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial robustness of neural networks from the perspective of Lipschitz calculus: A survey. *ACM Computing Surveys*, 57(6):1–41, 2025.

# A Proofs of corollaries of the main theorem

*Proof of Corollary 4.1.* The square loss is a symmetric Bregman divergence loss $D_\phi$ with $\phi(x) = \|x\|^2$. We consider the domain $\Omega = [-M, M]^K$ for $\phi$, which has $l_\infty$ diameter at most $d_\Omega = 2M$. Further, on this domain, $\phi$ is a $2\sqrt{K}M$-Lipschitz function, so we can take $L_\phi = 2\sqrt{K}M$. Since $\nabla\phi(x) = 2x$, we can take $L_g = 2$.

The set $A = \{\mathbb{E}[Y|x] : x \in \mathbb{R}^d\}$ is (by definition) contained in $[-M, M]^K$. We thus get

$$a_0 = \max_{a\in A} \|a\| \le m_0 = \max_{a\in[-M,M]^K} \|a\| = \sqrt{K}M,$$

$$m_2 = \max_{a\in A} |\phi(a)| \le m_1 = \max_{a\in[-M,M]^K} |\phi(a)| = KM^2,$$

$$m_3 = \max_{a\in A} ||\nabla\phi(a)|| \le \max_{a\in[-M,M]^K} ||\nabla\phi(a)|| = 2\sqrt{K}M,\ \gamma \le 2\sqrt{K}M.$$

We have thus verified all assumptions of Theorem 3.1, and using that Theorem, we now get that for any $(p, J)$-realistic function class $\mathcal{F}$, w.h.p. over the sampling of $n$ independent samples (where $n$ is as in the statement of the corollary), if there exists an $L$-Lipschitz function $f \in \mathcal{F}$ that $\epsilon$-overfits the training data, then the following lower bound on $L$ holds:

$$L \ge \frac{\epsilon}{128CKM\sqrt{2c}} \sqrt{\frac{nd}{p\log(1 + 64JWKM)/\epsilon) + \log(5K/\delta)}}.$$ □

*Proof of Corollary 4.2.* Let $\tilde{\Delta}_K$ denote the $K$ dimensional probability simplex. The cross-entropy loss is an asymmetric Bregman divergence loss $D_\phi$ with $\phi : \tilde{\Delta}_K \to \mathbb{R}$ defined as $\phi(x) := \sum_{i=1}^K x_i \log(x_i)$. Note also that the $\ell_\infty$-diameter of $\Omega = \tilde{\Delta}_K$ is at most $d_\Omega = 1$. We also recall that the labels $Y$ are one-hot encoded vectors, and the assumption that $\mathbb{P}[Y_i = 1|x] \ge \alpha$ for all $i \in [K]$ and $x$. This implies that $A = \{\mathbb{E}[Y|x] = (\mathbb{P}[Y_1 = 1|x], \mathbb{P}[Y_2 = 1|x], \ldots, \mathbb{P}[Y_K = 1|x]) : x \in \mathbb{R}^d\} \subseteq \tilde{\Delta}_K \cap [\alpha, 1 - \alpha)^K$.

Note also that the range of any function in the $(p, J)$-realistic function class $\mathcal{F}$ is a subset of $B := \{\text{Softmax}(x) : x \in [-M, M]^K\} \subset B_1 := \{x \in \tilde{\Delta}_K : x_i \ge \frac{e^{-2M}}{K}\ \forall 1 \le i \le K\}$, and $\nabla\phi(x) = (\log(x_1) + 1, \log(x_2) + 1, \ldots, \log(x_K) + 1)$ is well

defined for all $x \in B_1$. From this it follows that $\phi$ is an $L_\phi$-Lipschitz function on $B_1$ with $L_\phi = \sqrt{K}(1 + 2M + \log K)$. We also have

$$a_0 = \max_{a \in A} \|a\| \leq m_0 = \max_{b \in \tilde{\Delta}_K} \|b\| = 1,$$

$$m_2 = \max_{a \in A} |\phi(a)| \leq m_1 = \max_{b \in \tilde{\Delta}_K} |\phi(b)| \leq \log K,$$

$$m_3 = \max_{a \in A} \|\nabla \phi(a)\| \leq \sqrt{K}(1 + |\log(\alpha)|), \gamma \leq \max_{b \in B_1} \|\nabla \phi(b)\| = \sqrt{K}(1 + 2M + \log K).$$

Further, we also see that each coordinate $(\nabla \phi)_\ell$ of the gradient of $\phi$ is an $L_g$-Lipschitz function on the set $B_1$ with $L_g = Ke^{2M}$. Thus, we have verified all assumptions of Theorem 3.1, and using that theorem, we get that for any $(p, J)$-realistic function class $\mathcal{F}$, w.h.p. over the sampling of $n$ independent samples (where $n$ is as in the statement of the corollary), if there exists an $L$-Lipschitz function $f \in \mathcal{F}$ that $\epsilon$-overfits then the following lower bound on $L$ holds:

$$L \geq \frac{\epsilon}{32CK^2 e^{2M} \sqrt{2c}} \sqrt{\frac{nd}{p \log(1 + 8JW(e^{2M}K^2 + 2\sqrt{K}(1 + 2M + \log K))/\epsilon) + \log(5K/\delta)}}. \tag{71}$$

This bound can however be improved if we take a careful look at the internals of the proof of the Theorem 3.1. The crucial observation is that in Lemmas 3.6 and 5.1, we only need the Lipschitz constant of the composition $(\nabla \phi)_\ell \circ f$, which we estimate in general by multiplying the worst-case Lipschitz constants of $\nabla \phi_\ell$ and $f \in \mathcal{F}$. However, in this case, for $\text{Softmax}(f) \in \mathcal{F}$, we can get a better direct estimate of this Lipschitz constant of $(\nabla \phi)_\ell \circ \text{Softmax}(f)$ since $(\nabla \phi)_\ell \circ \text{Softmax}(f) = \log(e^{f(x)_\ell} / \sum_{t=1}^K e^{f(x)_t}) + 1$ is a $2L$-Lipschitz function from $\mathbb{R}^d$ to $\mathbb{R}^K$ whenever $f$ is an $L$-Lipschitz function. Thus, we can replace the factor $LL_g$ appearing in the proof of Theorem 3.1 by $2L$. This propagates through the proof of Theorem 3.1 and we thus obtain the following w.h.p. lower bound on the Lipschitz constant $L$ of $g$ when $\text{Softmax}(g) \in \mathcal{F}$ over-fits the data:

$$L \geq \frac{\epsilon}{64CK\sqrt{2c}} \sqrt{\frac{nd}{p \log(1 + 8JW(e^{2M}K^2 + \sqrt{K}(1 + 2M + \log K))/\epsilon) + \log(5K/\delta)}}. \tag{72}$$

□