The Price of Robustness

Stable Classifiers Need Overparameterization

Jonas von Berg, Adalbert Fono, Massimiliano Datres, Sohir Maskey, Gitta Kutyniok

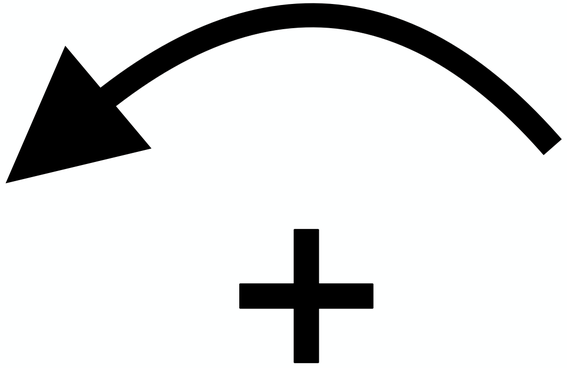
The Price of Robustness

Stable Classifiers Need Overparameterization

Jonas von Berg, Adalbert Fono, Massimiliano Datres, Sohir Maskey, Gitta Kutyniok

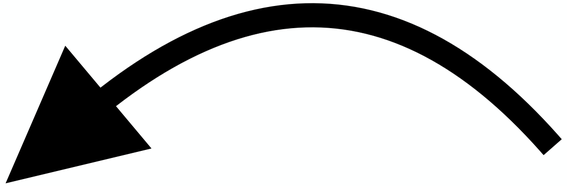
Context

Stable Classifiers Need Overparameterization



Stable Classifiers Generalise Better

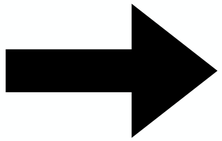
Context



Stable Classifiers Need Overparameterization

+

Stable Classifiers Generalise Better



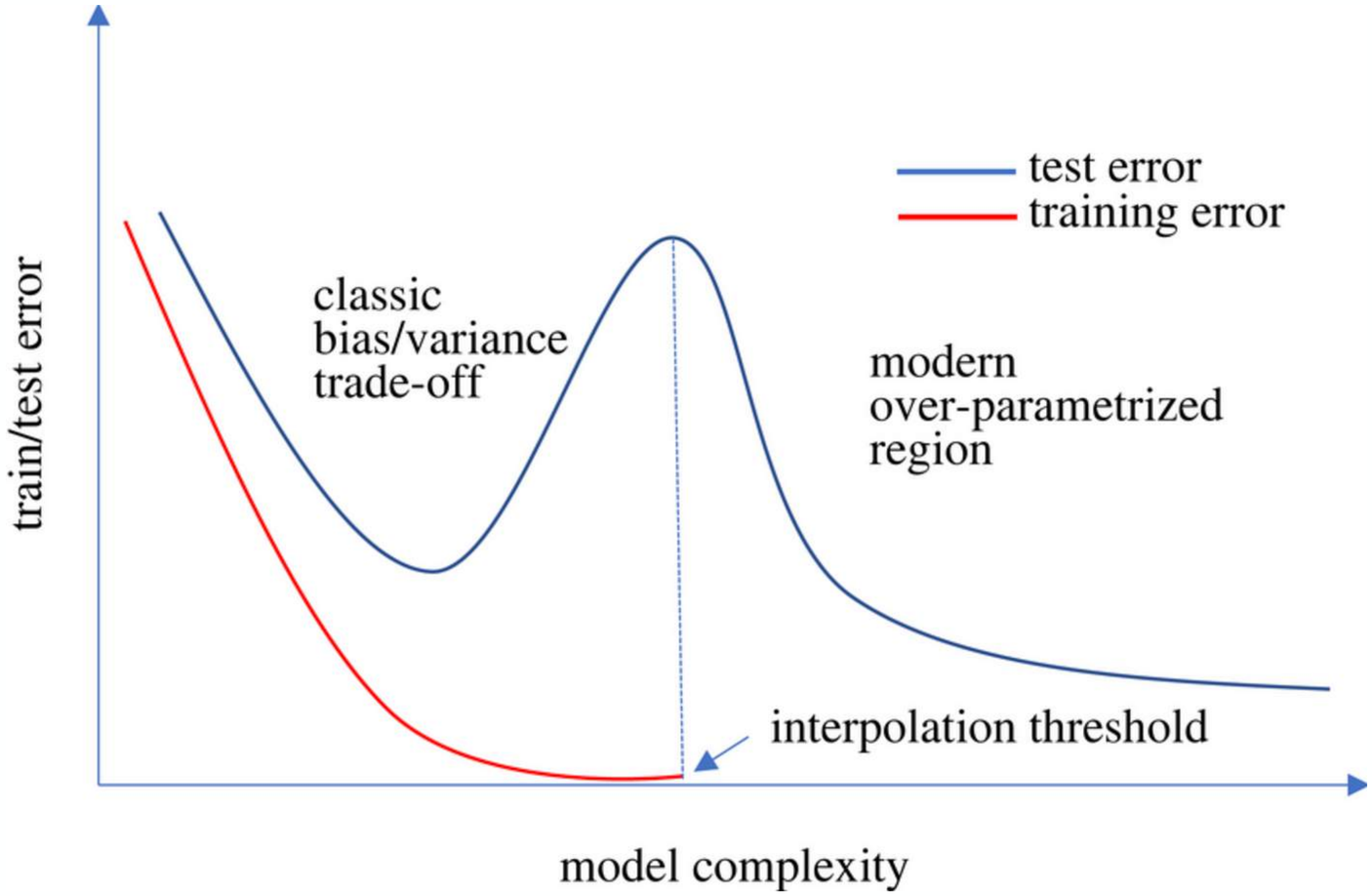
Overparameterization can lead to better generalisation

Context

Stable Classifiers Need Overparameterization

+

Stable Classifiers Generalise Better

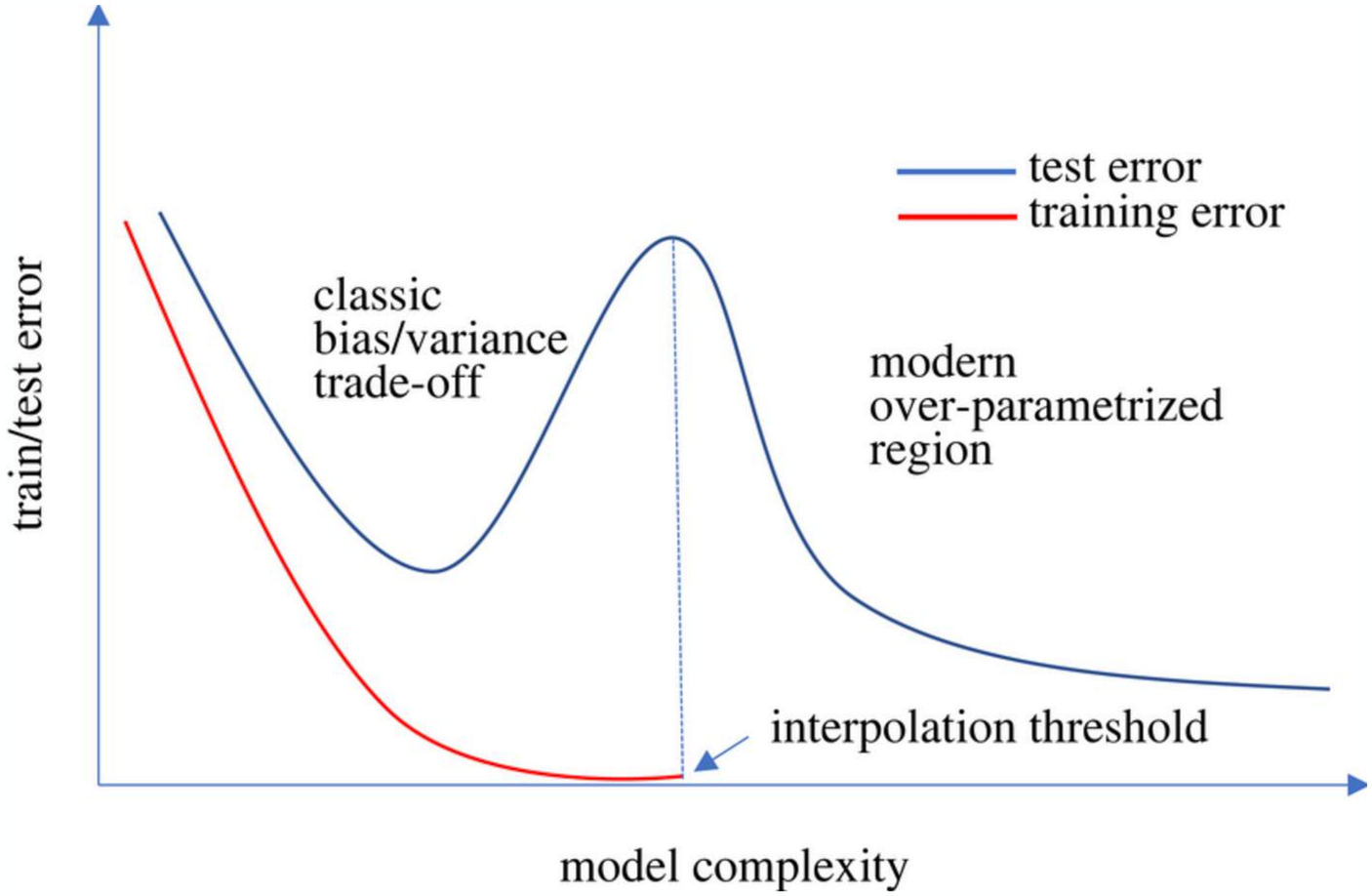


Context

Stable Classifiers Need Overparameterization

+

Stable Classifiers Generalise Better



Outline

1. Stability
    1.1. Stability as a measure of Robustness
    1.2. Lipschitz Constant and Robustness
    1.3. Class Stability
2. Isoperimetry
3. A Law of Robustness for Classification
    3.1. Generalization Perspective
    3.2. Universal Law of Robustness for Discontinuous Functions
4. Experiments
5. Outlook

1. Stability

1.1. Stability as a measure of Robustness

Robustness denotes the capacity of a model to sustain stable predictive perfomance against variations in the input data.

Crab-eating macaque

Crab-eating macaque

1.2. Lipschitz Constant and Robustness

$$L^{\varepsilon}(f,x) := \sup_{0< \ y-x \ \leq \varepsilon} \frac{f(y)-f(x)}{y-x}$$

Rich Literature on Lipschitz constant as a measure of Robustness

- Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks [https://arxiv.org/abs/1802.04034, Yusuke Tsuzuku, Issei Sato, Masashi Sugiyama]

- Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100 [https://arxiv.org/abs/2108.04062, Sahil Singla, Surbhi Singla, Soheil Feizi]

1.2. Lipschitz Constant and Robustness

Stable Regression Needs Overparameterization

Theorem 1.2 (Sébastien Bubeck, Mark Sellke 2022). *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i\in[n]}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$. Assume that:*

1. *$\mathcal{F}$ admits a Lipschitz parametrization by p real parameters, each of size at most poly(n, d).*
2. *The distribution μ of the covariates x_i satisfies isoperimetry (or is a mixture theoreof).*
3. *The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted $\sigma^2 = \mathbb{E}_\mu[Var(y|x))] > 0$*

Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i\in[n]}(f(x_i) - y_i)^2 \leq \sigma^2 - \varepsilon \implies L(f) \gtrsim \frac{\varepsilon}{\sigma}\sqrt{\frac{nd}{p}}$$

„Universal Law of Robustness“

1.2. Lipschitz Constant and Robustness

Stable Regression Needs Overparameterization

Theorem 1.2 (Sébastien Bubeck, Mark Sellke 2022). *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i\in[n]}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$. Assume that:*

1. *$\mathcal{F}$ admits a Lipschitz parametrization by p real parameters, each of size at most poly(n, d).*
2. *The distribution μ of the covariates x_i satisfies isoperimetry (or is a mixture theoreof).*
3. *The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted $\sigma^2 = \mathbb{E}_\mu[Var(y|x))] > 0$*

Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i\in[n]}(f(x_i)-y_i)^2 \leq \sigma^2 - \varepsilon \implies L(f) \gtrsim \frac{\varepsilon}{\sigma}\sqrt{\frac{nd}{p}}$$

„Universal Law of Robustness“

1.2. Lipschitz Functions and Robustness

Stable **Regression** Needs Overparameterization

- Classification functions are necessarily discontinuous : no Lipschitz constant
- In NN-setting the classifier f is usually of the form $f := argmax \circ g$, for a continuous score function g. However, g can be arbitrarily rescaled without altering the output of f

→ $L(g)$ alone does not reflect the geometry of the decision boundary

1.2. Lipschitz Functions and Robustness

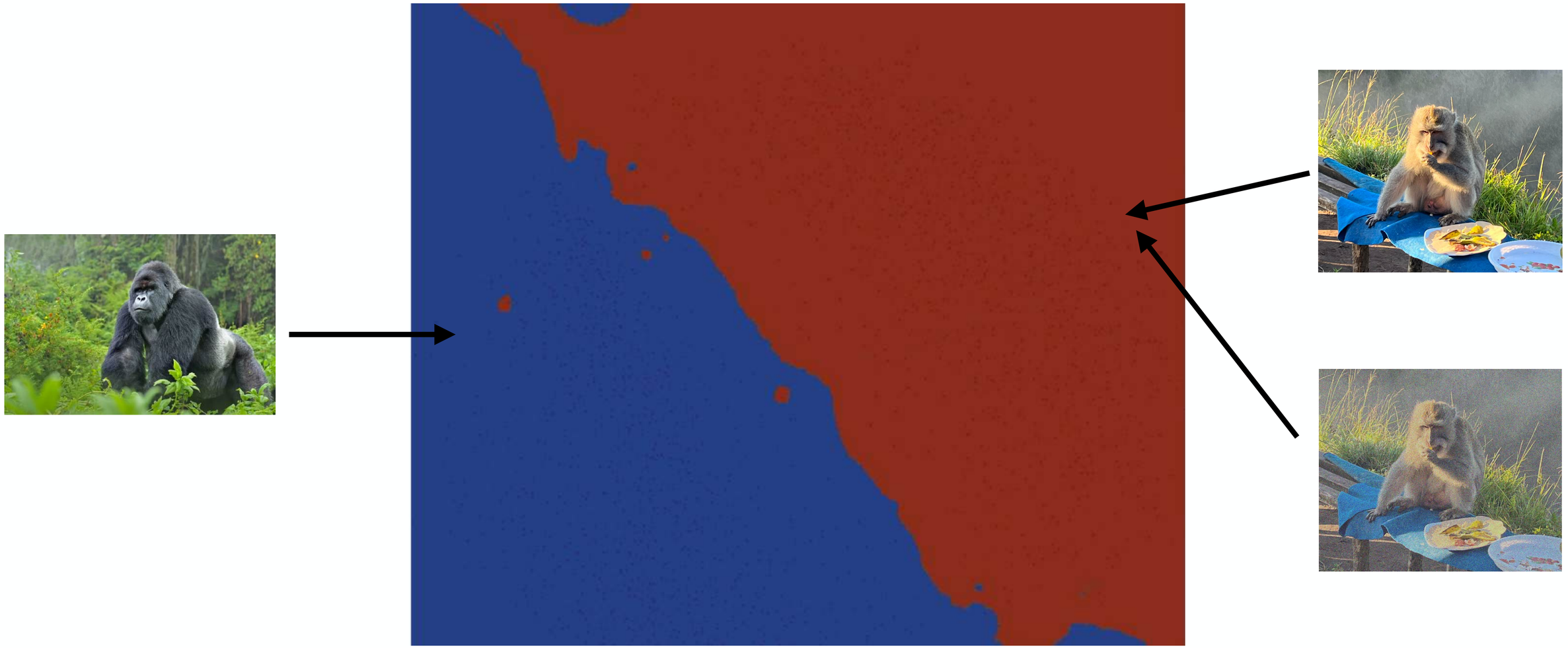
Stable **Regression** Needs Overparameterization

- Classification functions are necessarily discontinuous : no Lipschitz constant
- In NN-setting the classifier f is usually of the form $f := argmax \circ g$, for a continuous score function g. However, g can be arbitrarily rescaled without altering the output of f

→ $L(g)$ alone does not reflect the geometry of the decision boundary

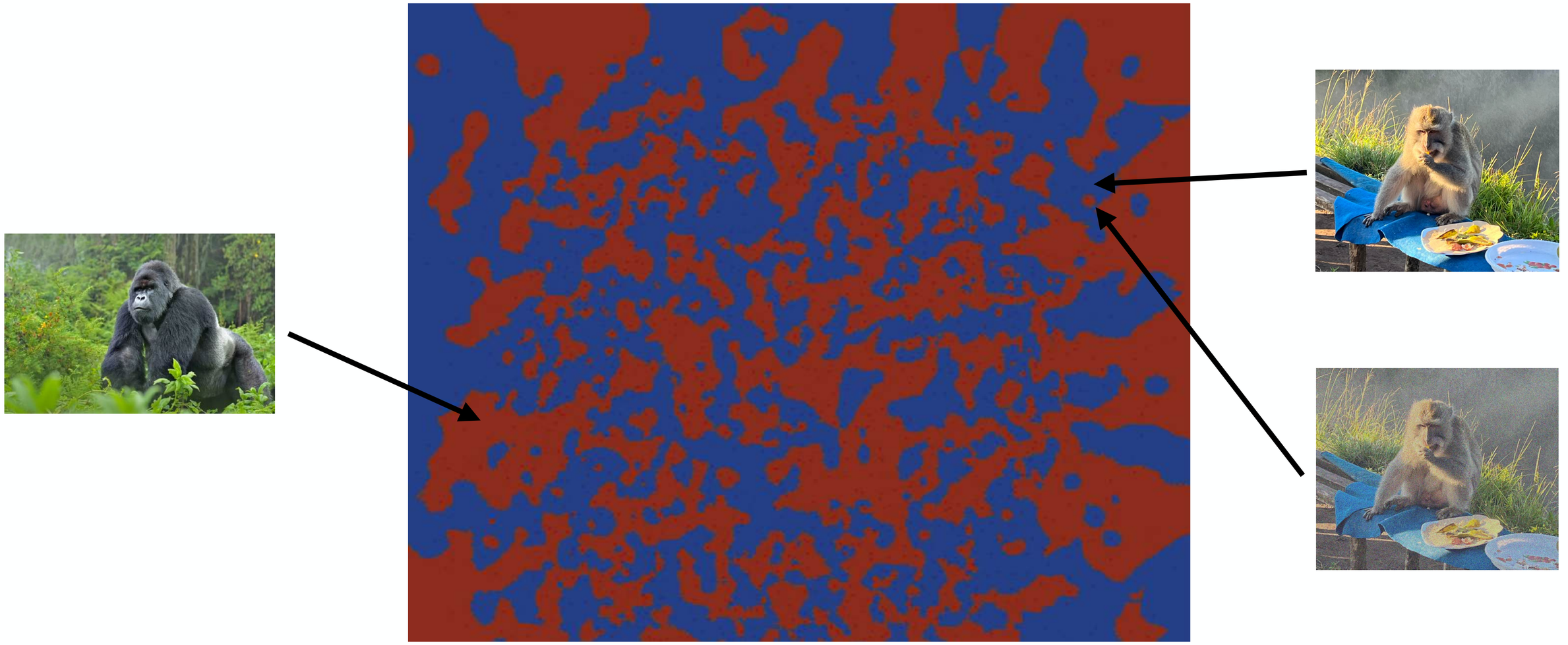
→ Need a different measure of stability for classification

1.3. Class Stability



Constructing and Visualizing High-Quality Classifier Decision Boundary Maps, Figure [2] | https://www.mdpi.com/2078-2489/10/9/280

1.3. Class Stability

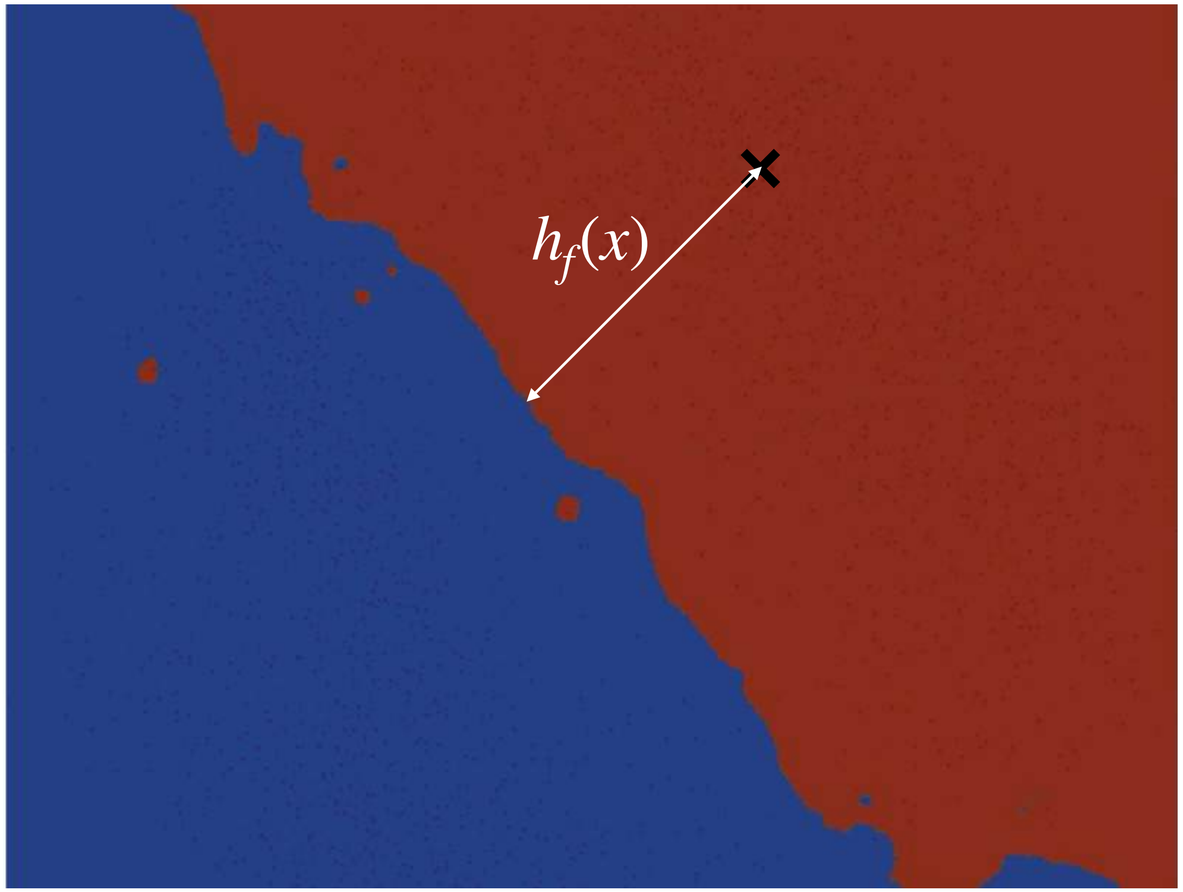


Constructing and Visualizing High-Quality Classifier Decision Boundary Maps, Figure [2] | https://www.mdpi.com/2078-2489/10/9/280

1.3. Class Stability

Definition 1 (Local Stability) *Let* $f : \mathcal{X} \to \{-1, 1\}$. *Its* extension $\bar{f} : \mathbb{R}^d \to \{-1, 0, 1\}$ *is defined by* $\bar{f}(x) = f(x)$ *for* $x \in \mathcal{X}$, *and* $\bar{f}(x) = 0$ *otherwise. We define the* local stability *of* f *at* $x \in X$ *as:*

$$h_f(x) := \inf\{\|x - z\|_2 : \bar{f}(x) \neq \bar{f}(z),\ z \in \mathbb{R}^d\}.$$



1.3. Class Stability

Definition 2 (Class Stability) *The* class stability *of* $f : \mathcal{X} \to \{-1, 1\}$ *on* $\mathcal{X}$ *is given by the average local stability with respect to the marginal distribution of* μ*, that is*

$$S(f) := \int_{\mathcal{X}} h_f^p(x)\, d\mu = \mathbb{E}[h_f^p]\,.$$

2. Isoperimetry

Stable Classifiers in **High Dimensions** Need Overparameterization

Goal: Bound the probability $\mathbb{P}(\exists f \in \mathscr{F} : \hat{R}(f) \approx 0 \wedge L(f) \leq L_*(p, n, d)) \leq \delta$

→ Need to control the ability of a function f to correlate with the data in terms of its Lipschitz constant!

Definition 3 (Isoperimetry) *A probability measure μ on $\mathcal{X} \subset \mathbb{R}^d$ satisfies* c-isoperimetry *if for any bounded L-Lipschitz $f : \mathcal{X} \to \mathbb{R}$, and any $t \geq 0$,*

$$\mathbb{P}[|f(x) - \mathbb{E}[f]| \geq t] \leq 2e^{-\frac{dt^2}{2cL^2}}.$$

Examples: Gaussian measures, Manifolds with positive Ricci curvature, log-concave measures

2. Isoperimetry

Stable Classifiers in **High Dimensions** Need Overparameterization

Goal: Bound the probability $\mathbb{P}(\exists f \in \mathscr{F} : \hat{R}(f) \approx 0 \wedge L(f) \leq L_*(p, n, d)) \leq \delta$

⟶ Need to control the ability of a function f to correlate with the data in terms of its Lipschitz constant!

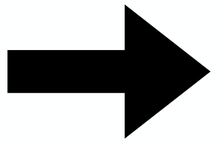
Definition 3 (Isoperimetry) *A probability measure μ on $\mathcal{X} \subset \mathbb{R}^d$ satisfies* c-isoperimetry *if for any bounded L-Lipschitz $f : \mathcal{X} \to \mathbb{R}$, and any $t \geq 0$,*

$$\mathbb{P}[|f(x) - \mathbb{E}[f]| \geq t] \leq 2e^{-\frac{dt^2}{2cL^2}}.$$

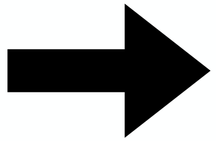
Examples: Gaussian measures, Manifolds with positive Ricci curvature, log-concave measures

2. Isoperimetry

Hoeffding's inequality: $\mathbb{P}(\ \ f(x) - \mathbb{E}(f)\ \ \geq t) \leq 2\exp(-\frac{t^2}{8})$



We need $L(f) < 2\sqrt{\frac{d}{c}}$ to get a non-vacuous bound.



In high dimensions, this constraint is not particularly restrictive.

3. A Law of Robustness for Classification

From now on we will make the following two assumptions:

(H1) $(\mathcal{X} \times \{-1.1\}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure $\mu_{\mathcal{X}}$;

(H2) the considered hypothesis class $\mathcal{F}$ of classifiers $f : \mathbb{R}^d \to \{-1, 1\}$ is finite, that is $|\mathcal{F}| < \infty$, and there exists $S > 0$ such that $\min_{f \in \mathcal{F}} S(f) > S$.

3. A Law of Robustness for Classification

From now on we will make the following two assumptions:

$$\mathbb{P}[|f(x) - \mathbb{E}[f]| \geq t] \leq 2e^{-\frac{dt^2}{2cL^2}}.$$

(H1) $(\mathcal{X} \times \{-1.1\}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure $\mu_{\mathcal{X}}$;

(H2) the considered hypothesis class $\mathcal{F}$ of classifiers $f : \mathbb{R}^d \to \{-1, 1\}$ is finite, that is $|\mathcal{F}| < \infty$, and there exists $S > 0$ such that $\min_{f \in \mathcal{F}} S(f) > S$.

$$S(f) := \int_{\mathcal{X}} h_f^p(x)\, d\mu = \mathbb{E}[h_f^p].$$

3.1. Generalization Perspective

Theorem 4 (Generalization Bound) *Under Assumptions (H1), (H2) and* $\log(|\mathcal{F}|) \geq n$, *we have:*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \lesssim \max\left\{\sqrt{\frac{1}{n}}, \frac{1}{S}\frac{c\ \log(|\mathcal{F}|)}{n\sqrt{d}}\right\}, \tag{3}$$

where $\lesssim$ *means that the inequality holds up to an absolute constant independent of* $\log(|\mathcal{F}|), n, d, S, c$.

Simple but crucial observation:

If $h_f(x) > \alpha$ then for all $y \in \mathcal{X}$: $\left\| f(x) - f(y) \right\| \leq \frac{2}{\alpha} \left\| x - y \right\|$.

→ f is $\frac{2}{S-t}$ - Lipschitz continuous on $A_t(f) := \{x \in \mathcal{X} : h_f(x) > S - t\}$

3.1. Generalization Perspective

Idea: Apply the regression result of Bubeck and Selke on $A_t(f)$ and control the measure $\mathbb{P}(A_t(f)^c)$ using isoperimetry.

Indeed, $h_f(x)$ is Lipschitz continuous and bounded:

Isoperimetry

$$\mathbb{P}(A_t(f)^c) = \mathbb{P}(S(f) - h_f^p(x) \geq t) \leq \exp\left(-\frac{dt^2}{2^3 c}\right) = \exp\left(-\frac{dS^2}{2^5 c}\right).$$

Choose $t = S/2$ and write $A_S(f)$

3.1. Generalization Perspective

Proof (Sketch):

$$\mathscr{R}_{n,\mu}(\mathscr{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

3.1. Generalization Perspective

Proof (Sketch):

$$\mathscr{R}_{n,\mu}(\mathscr{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

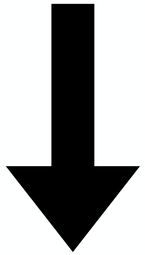
$$\leq \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i F_f(x_i)\right|\right] + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i (f-F_f)(x_i)\right|\right]$$

3.1. Generalization Perspective

Proof (Sketch):

$$\mathscr{R}_{n,\mu}(\mathscr{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

$$\leq \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i F_f(x_i)\right|\right] + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i (f - F_f)(x_i)\right|\right]$$



Bubeck and Selke!

3.1. Generalization Perspective

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right|\right]$$

$\longrightarrow$ On $A_S(f):\ f = F_f$

$\longrightarrow$ $\mathbb{P}(A_S(f)^c) \to 0$ for large S

Define the set:

$$A_S^I(f) := \{x \in \mathcal{X}^n : i \in I \iff h_f(x_i) \geq S/2\} \quad \Big| \quad \mathcal{X}^n = \bigcup_{I\in\mathcal{P}([n])} A_S^I(f)$$

$$\Rightarrow \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \wedge x \in A_S^I(f)\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \ \ x \in A_S^I(f)\right)\mathbb{P}(A_S^I(f))$$

3.1. Generalization Perspective

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right|\right]$$

→ On $A_S(f): f = F_f$

→ $\mathbb{P}(A_S(f)^c) \to 0$ for large S

Define the set:

$$A_S^I(f) := \{x \in \mathcal{X}^n : i \in I \iff h_f(x_i) \geq S/2\} \quad \Big| \quad \mathcal{X}^n = \bigcup_{I\in\mathcal{P}([n])} A_S^I(f)$$

$$\Rightarrow \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \wedge x \in A_S^I(f)\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \;\; x \in A_S^I(f)\right)\mathbb{P}(A_S^I(f))$$

small for $I \to n$ small for $I \to 0$

3.1. Generalization Perspective

sub-exponential!

$$\Rightarrow \quad \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \right) \leq 2 \exp\left(- \frac{rS\sqrt{d}C}{log(\ |\mathcal{F}|\ 2^n)} \right)$$

3.1. Generalization Perspective

sub-exponential!

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right|\right] \lesssim \frac{1}{S}\left(\frac{log(\ \ \mathscr{F}\ \ ) + nlog(2)}{n\sqrt{d}}\right)$$

3.1. Generalization Perspective

sub-exponential!

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F)(x_i)\right|\right] \lesssim \frac{1}{S}\left(\frac{log(\ \mathscr{F}\ ) + nlog(2)}{n\sqrt{d}}\right)$$

Theorem 4 (Generalization Bound) *Under Assumptions (H1), (H2) and* $log(|\mathcal{F}|) \geq n$, *we have:*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \lesssim \max\left\{\sqrt{\frac{1}{n}}, \frac{1}{S}\frac{c\ \log(|\mathcal{F}|)}{n\sqrt{d}}\right\}, \tag{3}$$

where $\lesssim$ *means that the inequality holds up to an absolute constant independent of* $\log(|\mathcal{F}|), n, d, S, c$.

3.2. Universal Law of Robustness for Discontinuous functions

Corollary 5 (Law of Robustness for Discontinuous Functions) *Under Assumptions (H1) and (H2), let us fix* $\varepsilon, \delta \in (0,1)$ *and the loss* ℓ_{0-1} *to be the 0-1 loss. Then there exists an absolute constant* $K > 0$ *such that under the additional conditions*

1. *The minimal risk, defined as* $\sigma^2 := \min_{f \in \mathcal{F}} R_{0-1}(f)$*, satisfies* $\sigma^2 \geq \varepsilon$*,*
2. *The number of samples* n *is large enough such that:* $\frac{K}{\sqrt{n}} < \frac{\varepsilon}{3}$ *and* $\sqrt{\frac{2\log(2/\delta)}{n}} < \varepsilon/2$*; furthermore, we assume that* $p := \log(|\mathcal{F}|) \geq n$*,*

we have: With probability at least $1 - \delta$ *with respect to the sample, the following holds uniformly for all* $f \in \mathcal{F}$*:*

$$\hat{R}_{0-1}(f) \leq \sigma^2 - \varepsilon \implies S(f) < \frac{Kc}{\varepsilon} \frac{p}{n\sqrt{d}}$$

„Universal Law of Robustness for Classification“

3.2. Universal Law of Robustness for Discontinuous functions

Idea: Suppose there exists an $f \in \mathscr{F}$ that overfits, that is $\hat{R}_{0-1}(f) \leq \sigma^2 - \varepsilon$

$$\Rightarrow \quad R_{0-1}(f) - \hat{R}_{0-1}(f) \geq \varepsilon$$

But we can bound the generalization error with the Rademacher complexity!

$$R_{0-1}(f) - \hat{R}_{0-1}(f) \leq \mathscr{R}_{n,\mu}(\mathscr{F}) + \sqrt{\frac{2\log(2/\delta)}{n}} < \varepsilon, \qquad \text{w/ probability } \geq 1 - \delta$$

$$\text{for } S \geq S_* = \frac{Kc}{\varepsilon}\frac{p}{n\sqrt{d}}$$

$$\Rightarrow \quad \mathbb{P}(\exists f \in \mathscr{F} : \hat{R}_{0-1}(f) \leq \sigma^2 - \varepsilon \wedge S \geq S_*) \leq \delta$$

Contraposition gives the universal law of robustness for discontinuous functions!

Some concerns…

- Is (H2, finiteness of $\mathscr{F}$) too restrictive?
- Can we improve $p \geq n\sqrt{d}$ to $p \geq nd$ as in Bubeck? [1]
- Empirical relevance: Do margins truly correlate with generalization in realistic settings?

4. A Law of Robustness for infinite functions

(H1) $(\mathcal{X} \times \{-1.1\}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure $\mu_{\mathcal{X}}$;

(H2) the considered hypothesis class $\mathcal{F}$ of classifiers $f : \mathbb{R}^d \to \{-1, 1\}$ is finite, that is $|\mathcal{F}| < \infty$, and there exists $S > 0$ such that $\min_{f \in \mathcal{F}} S(f) > S$.

4. A Law of Robustness for infinite functions

(H1) $(\mathcal{X} \times \{-1.1\}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure $\mu_{\mathcal{X}}$;

(H2) the considered hypothesis class $\mathcal{F}$ of classifiers $f : \mathbb{R}^d \to \{-1, 1\}$ is finite, that is $|\mathcal{F}| < \infty$, and there exists $S > 0$ such that $\min_{f \in \mathcal{F}} S(f) > S$.

4. A Law of Robustness for infinite functions

(H1) $(\mathcal{X} \times \{-1.1\}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure $\mu_{\mathcal{X}}$;

(H2)' The function class $\mathcal{F} = sgn \circ \mathcal{G}$, where $\mathcal{G} = \{g_w : w \in \mathcal{W}\}$ is a parameterized class of Lipschitz continuous functions. The parameter space $\mathcal{W} \subset \mathbb{R}^p$ is bounded with $diam(\mathcal{W}) \leq W$ and the parameterization is Lipschitz continuous

$$\|g_{w_1} - g_{w_2}\|_\infty \leq J \|w_1 - w_2\| \tag{16}$$

4. A Law of Robustness for infinite functions

(H1) $(\mathcal{X} \times \{-1.1\}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure $\mu_{\mathcal{X}}$;

(H2)' The function class $\mathcal{F} = sgn \circ \mathcal{G}$, where $\mathcal{G} = \{g_w : w \in \mathcal{W}\}$ is a parameterized class of Lipschitz continuous functions. The parameter space $\mathcal{W} \subset \mathbb{R}^p$ is bounded with $diam(\mathcal{W}) \leq W$ and the parameterization is Lipschitz continuous

$$\|g_{w_1} - g_{w_2}\|_\infty \leq J \|w_1 - w_2\| \tag{16}$$

Problem: Prediction $g_w(x)$ may concentrate around the discontinuity at 0

4.1 Class Margin

Definition 7 (Local and Class Margin) *For* $f = \operatorname{sgn} \circ g : \mathcal{X} \to \{-1, 1\}$, *we define the* local prediction margin *of* f *at* $x \in \mathcal{X}$ *as*

$$\tilde{h}_f(x) := |g(x)|.$$

The class margin *of* f *on* $\mathcal{X}$ *is defined as the expected local prediction margin with respect to the marginal distribution* μ*, that is,*

$$\tilde{S}(f) := \int_{\mathcal{X}} \tilde{h}_f(x)\, d\mu = \mathbb{E}[\tilde{h}_f].$$

4.1 Class Margin

Definition 7 (Local and Class Margin) *For* $f = \operatorname{sgn} \circ g : \mathcal{X} \to \{-1, 1\}$, *we define the* local prediction margin *of* f *at* $x \in \mathcal{X}$ *as*

$$\tilde{h}_f(x) := |g(x)|.$$

The class margin *of* f *on* $\mathcal{X}$ *is defined as the expected local prediction margin with respect to the marginal distribution* μ, *that is,*

$$\tilde{S}(f) := \int_{\mathcal{X}} \tilde{h}_f(x)\, d\mu = \mathbb{E}[\tilde{h}_f].$$

$$S(f) \geq \frac{\tilde{S}(f)}{L(g)}$$

$$\mathscr{R}_{n,\mu}(\mathscr{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

$$\mathscr{R}_{n,\mu}(\mathscr{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

$$\leq \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i F_f(x_i)\right|\right] + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i (f-F_f)(x_i)\right|\right]$$

$$\mathscr{R}_{n,\mu}(\mathscr{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

$$\leq \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i F_f(x_i)\right|\right] + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i (f-F_f)(x_i)\right|\right]$$

Bubeck and Selke!

$f - F_f$ is $\frac{L(g)}{\tilde{S}}$ – *Lipschitz*!

Theorem 8 *Under assumptions (H1) and (H2)', suppose that* $\tilde{S}(f) > \tilde{S} > 0$ *and* $L(g) \leq L$ *for all* $f \in \mathcal{F}$ *and* $g \in \mathcal{G}$. *Furthermore, assume that* $\log|\mathcal{F}| \geq n$. *Then,*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \leq K \max\left\{\sqrt{\tfrac{1}{n}},\ \tfrac{L}{\tilde{S}}\sqrt{\tfrac{p}{nd}}\sqrt{c\log(1+60WJ\varepsilon^{-1})},\ 2\exp\left(-\tfrac{d\tilde{S}^2}{8cL^2}\right),\ \tfrac{J}{\tilde{S}}\varepsilon\right\},$$

where $K > 0$ *is an absolute constant independent of* $\log|\mathcal{F}|, n, d, \tilde{S}, c$.

Theorem 8 *Under assumptions (H1) and (H2)', suppose that* $\tilde{S}(f) > \tilde{S} > 0$ *and* $L(g) \leq L$ *for all* $f \in \mathcal{F}$ *and* $g \in \mathcal{G}$. *Furthermore, assume that* $\log|\mathcal{F}| \geq n$. *Then,*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \leq K \max\left\{\sqrt{\tfrac{1}{n}},\ \tfrac{L}{\tilde{S}}\sqrt{\tfrac{p}{nd}}\sqrt{c\log(1+60WJ\varepsilon^{-1})},\ 2\exp\left(-\tfrac{d\tilde{S}^2}{8cL^2}\right),\ \tfrac{J}{\tilde{S}}\varepsilon\right\},$$

where $K > 0$ *is an absolute constant independent of* $\log|\mathcal{F}|, n, d, \tilde{S}, c$.

Corollary 11 (Law of Robustness for Infinite Function Classes) *Under Assumptions (H1), (H2)', let us fix* $\varepsilon, \delta \in (0,1)$ *and consider the 0-1 loss* $\ell_{0\text{-}1}$. *Then there exists an absolute constant* $K > 0$ *such that under the additional conditions*

1. *the minimal risk, defined as* $\sigma^2 := \min_{f\in\mathcal{F}} R_{0\text{-}1}(f)$, *satisfies* $\sigma^2 \geq \varepsilon$,
2. *the number of samples* n *is large enough such that* (i) $\frac{K}{\sqrt{n}} < \frac{\varepsilon}{3}$ *and* (ii) $\sqrt{\frac{2\log(2/\delta)}{n}} < \frac{\varepsilon}{2}$,

we have: With probability at least $1-\delta$ *with respect to the sample, the following holds uniformly for all* $f \in \mathcal{F}$:

$$\hat{R}_{0\text{-}1}(f) \leq \sigma^2 - \varepsilon \implies \frac{\tilde{S}(f)}{L(g)} < \frac{3K\sqrt{c}}{\varepsilon}\sqrt{\frac{p}{nd}}\sqrt{c\log(1+60WJ\varepsilon^{-1})}. \tag{19}$$

4. Experiments

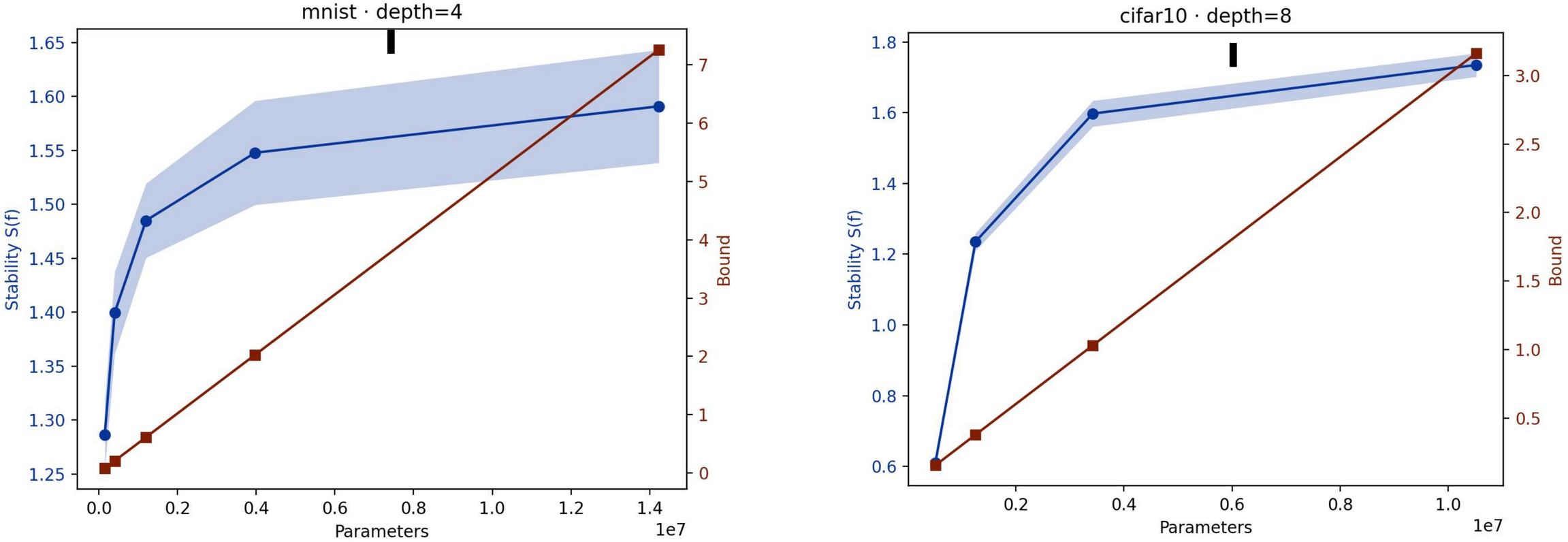


Figure 1: Empirical class stability $S(f)$ (blue, left axis) and theoretical bound from Corollary 5 (red, right axis) as a function of the numbers of parameters p.

4. Experiments

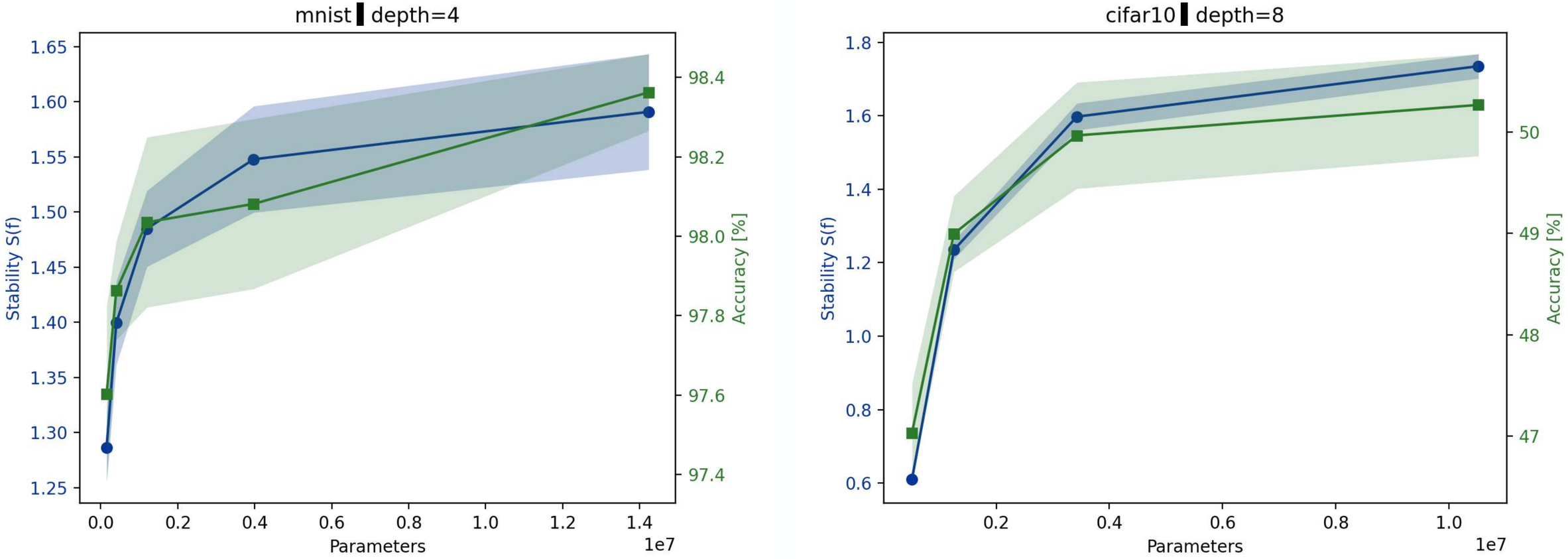


Figure 2: Empirical class stability $S(f)$ (blue, left axis) and test accuracy (green, right axis) as a function of the numbers of parameters p.

5. Outlook

- Experiments for $\frac{\tilde{S}}{L}$
-

H1
- → Test isoperimetric concentration properties of real-world datasets.
- → Alternative concentration inequalities beyond isoperimetry.

H2'
- → Can we prove an infinite function case result without changing $S \to \tilde{S}$
- → Problem: Prediction $g_w(x)$ may concentrate around the discontinuity at 0

- Relate class stability to robust generalisation error [5] and related concepts of robustness.

- How to compute $S(f)$ effectively?

6. References

(1) Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id= z71OSKqTFh7.

(2) Z. N. D. Liu and A. C. Hansen. Do stable neural networks exist for classification problems? – a new view on stability in ai, 2024. URL https://arxiv.org/abs/2401.07874.

(3) Peter L Bartlett, Philip M Long, Gabor Lugosi, and Alexander Tsigler. Benign overfitting in ´ linear regression. Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.

(4) Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machinelearning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019

(5) Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https: //arxiv.org/abs/ 1706.06083.

3.1. Generalization Perspective

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right|\right]$$

$\longrightarrow$ On $A_S(f):\ f = F_f$

$\longrightarrow$ $\mathbb{P}(A_S(f)^c) \to 0$ for large S

Define the set:

$$A_S^I(f) := \{x \in \mathcal{X}^n : i \in I \iff h_f(x_i) \geq S/2\} \quad \Big| \quad \mathcal{X}^n = \bigcup_{I\in\mathcal{P}([n])} A_S^I(f)$$

$$\Rightarrow \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \wedge x \in A_S^I(f)\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \ \ x \in A_S^I(f)\right)\mathbb{P}(A_S^I(f))$$

3.1. Generalization Perspective

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right|\right]$$

→ On $A_S(f): f = F_f$

→ $\mathbb{P}(A_S(f)^c) \to 0$ for large S

Define the set:

$$A_S^I(f) := \{x \in \mathcal{X}^n : i \in I \iff h_f(x_i) \geq S/2\} \quad \Big| \quad \mathcal{X}^n = \bigcup_{I\in\mathcal{P}([n])} A_S^I(f)$$

$$\Rightarrow \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \wedge x \in A_S^I(f)\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right| > r \;\; x \in A_S^I(f)\right)\mathbb{P}(A_S^I(f))$$

small for $I \to n$ small for $I \to 0$

3.1. Generalization Perspective

$$\mathcal{X}^n = \bigcup_{I \in \mathcal{P}([n])} A_S^I(f)$$

$$\Rightarrow \mathbb{P}\left(\left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \right) = \sum_{I \subset [n]} \mathbb{P}\left(\left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \;\; x \in A_S^I(f) \right) \mathbb{P}(A_S^I(f))$$

$$\leq \sum_{I \in \mathcal{P}([n])} 2 \exp\left(- \frac{r^2 c_1}{n - m} \right) \exp\left(- (n - m) S^2 d c_2 \right)$$

$$\leq \sum_{I \subset [n]} exp(-k(\;\; I \;\;)) \leq 2^{n+1} exp(-k(m_*))$$

3.1. Generalization Perspective

$$\Rightarrow \quad \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \right) \leq 2 \;\; \mathcal{F} \;\; 2^n \exp\left(- rS\sqrt{d}C \right)$$

3.1. Generalization Perspective

sub-exponential!

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \right) \leq 2 \ \mathcal{F} \ 2^n \exp\left(- rS\sqrt{d}C \right)$$

3.1. Generalization Perspective

sub-exponential!

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \right) \leq 2 \ \ |\mathcal{F}| \ \ 2^n \exp\left(- rS\sqrt{d}C \right)$$

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right| > r \right) \leq 2 \exp\left(- \frac{rS\sqrt{d}C}{log(\ |\mathcal{F}| \ 2^n)} \right)$$

3.1. Generalization Perspective

sub-exponential!

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \sigma_i (f - F)(x_i) \right|\right] \lesssim \frac{1}{S}\left(\frac{log(\ \ \mathscr{F}\ \ ) + nlog(2)}{n\sqrt{d}}\right)$$

3.1. Generalization Perspective

sub-exponential!

$$\frac{1}{n}\mathbb{E}^{\sigma_i, x_i}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F)(x_i)\right|\right] \lesssim \frac{1}{S}\left(\frac{log(\ \mathscr{F}\ ) + nlog(2)}{n\sqrt{d}}\right)$$

Theorem 4 (Generalization Bound) *Under Assumptions (H1), (H2) and* $log(|\mathcal{F}|) \geq n$, *we have:*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \lesssim \max\left\{\sqrt{\frac{1}{n}}, \frac{1}{S}\frac{c\ \log(|\mathcal{F}|)}{n\sqrt{d}}\right\}, \tag{3}$$

where $\lesssim$ *means that the inequality holds up to an absolute constant independent of* $\log(|\mathcal{F}|), n, d, S, c$.