Minimum norm interpolation by perceptra: Explicit regularization and implicit bias

Jiyoung Park[1] Ian Pelakh[2] Stephan Wojtowytsch[3]

[1]Department of Statistics
Texas A&M University

[2]Department of Mathematics
Iowa State University

[3]Department of Mathematics
University of Pittsburgh

NeurIPS 2023

Table of Contents

Why we study a perceptron (Two-Layer neural network, shallow neural network, ...)?

- A tractable model for the theoretical analysis.
  - Implicit bias
    - How model settings (activation functions, optimization algorithms, initializations, ...) affect the solution we obtain?
    - Model settings 'implicitly' give us some 'bias' toward the certain solutions.
  - Depth separation
    - What kind of problems are solved with shallow networks, and better solved by deeper networks?

- A neural network is a function of the following form:

$$f(x) = g \circ h_K \circ h_{K-1} \circ \cdots h_1(x)$$

where $h_k : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ s.t. $h_k(z) = \sigma_k(W_k z + b_k)$.

  - W is called a 'weight'.
  - b is called a 'bias'.
  - σ is a pre-defined non-linear function called an 'activation function'.
    - Examples: $\tanh(x)$, $\mathrm{ReLU}(x) = \max\{0, x\}$.
    - From now on, we fix $\sigma = \mathrm{ReLU}$.
  - g is called a final layer, and convert the output into desired form (linear map if regression, softmax function if classification, ...).
  - Each h_k is called a layer, d_k is called a width of the kth layer. K is called a depth.

- Why neural network? Universal Approximation Theorems.
  - Any continuous function with compact support can be approximated by a neural network with suitable width, depth, and activation w.r.t. sup-norm topology (i.e. a set of neural network is dense in $C(K)$ w.r.t. sup-norm topology).
  - Limitations:
    - Compact support is necessary.
    - Lacking quantatitive bound or obtaining very loose bound.
    - Only for sup-norm topology.
- In shallow neural network cases, these limitations can be resolved.

- A Two-Layer neural network (perceptron, shallow neural network, ...) is the neural network with two layers including a final layer.

$$f_m(x) = \frac{1}{m}\sum_{j=1}^{m} a_j\sigma(w_j^T x + b_j).$$

- Which functions can be approximated well by a Two-Layer ReLU neural network?
    - A measure representation of m-width Two-Layer neural networks:

$$
\begin{aligned}
f_m(x) &= \frac{1}{m}\sum_{j=1}^{m} a_j\sigma(w_j^T x + b_j) \\
&= \int a\sigma(w^T x + b)\, d\left(\frac{1}{m}\sum_{j=1}^{m}\delta_{\theta_m}\right) \\
\Rightarrow f(x) &= \int_\theta a\sigma(w^T x + b) d\pi(\theta) \qquad (1)
\end{aligned}
$$

    where $\theta = (a, w, b)$ and $\pi(\theta)$ is a probability measure in Θ space.
    - Denote $\mathcal{B}$: a set of functions that can be expressed by the form (1).

- We can assign a natural norm in $\mathcal{B}$ (Barron norm).

$$\|f\|_{\mathcal{B}} := \inf_{\pi} \int |a|(\|w\| + |b|)d\pi.$$

- The normed space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ is indeed a Banach space.
  $\Rightarrow$ 'Barron Space'
- Barron space is a space can be approximated well by perceptra (i.e. perceptra is dense subset of Barron space w.r.t. Barron norm).

- Barron space has a relationship with other function spaces $\rightarrow$ convenient theoretical analysis ([EW21]).
  - E.g. $H^s(\mathbb{R}^d) \subseteq \mathcal{B}$ for $s > d/2 + 2$ if μ has a bounded support. $\mathcal{B} \subset C^{0,1}(\mathbb{R}^d)$.
  - Intuition: The form in (1) with $\sigma(\cdot) = \cos(\cdot)$ is a $\mathbb{R}$-valued Fourier inversion $\rightarrow \|\cdot\|_{\mathcal{B}}$ resembles the fractional Sobolev norm.

$\Rightarrow$ Is everything good now?

- The existence of the bias term in Barron norm causes discrepancies with practical settings. Why?
  - Barron Norm: Not invariant under translations in the data space.
    - $\Rightarrow\!\Leftarrow$ In practice we frequently center the data.
  - Bias term has no contribution to overfitting.
    - $\therefore$ Want to make a regularization without controlling the bias.
- Due to above facts, in practice 'weight decay' penalty is used instead of Barron norm penalty.

$$R_{WD}(\theta) = \frac{\|a\|_{\ell^2}^2 + \|W\|_F^2}{2m}.$$

$\Rightarrow$ Any concept corresponding to this Weight Decay regularizer?

Barron Space VI

- Construct continuum extension of Weight Decay regularizer:

$$[f]_{\mathcal{B}} = \inf_{\pi} R_{WD}(\pi) := \inf_{\pi} \frac{1}{2} \int |a|^2 + \|w\|^2 d\pi.$$

This $[\cdot]_{\mathcal{B}}$ is not a norm but is a semi-norm. $\Rightarrow$ 'Barron semi-norm'.

- Benefits of Barron semi-norm:
  - Any f with $f(0) < \infty$ and $[f]_{\mathcal{B}} < \infty$ is a Barron function $\Rightarrow$ Can import theoretical benefits of Barron norm.
  - If $f \in \mathcal{B}$, then $Lip(f) \leq [f]_{\mathcal{B}}$.
- Minimum norm interpolant: A function with $\min[f]_{\mathcal{B}}$ under data fitting constraint.

Main questions addressed in the work:

- Can we obtain the approximation error between a Two-Layer ReLU network and a target function in terms of number of parameters and data points, under more general conditions (unbounded & non-Lipschitz loss, non-compact & sub-Gaussian data)?
- How do Two-Layer ReLU networks interpolate where there is no data?
- Can we use theoretical solutions to compare different learning schemes (optimization algorithms, initialization)?

Notations

- $x \sim \mu$: Data distribution, $\|x\|$ is assumed to be σ^2-sub-Gaussian.
- n: Dataset size.
- m: Width of the neural network.
- λ: Strength of weight decay regularizer in the risk functional.
- f^*: Target function.
- Two-layer ReLU net $f_\theta(x) = \frac{1}{m}\sum_{j=1}^m a_j \sigma(w_j^T x + b_j)$
- (Regularized) empirical risk

$$\widehat{\mathcal{R}}_{n,m,\lambda}(\theta) = \frac{1}{2n}\sum_{i=1}^n \ell^2\left(f_\theta(x_i),\ f^*(x_i)\right) + \lambda R_{WD}(\theta)$$

- f_n = Empirical Risk Minimizer (ERM) for $\widehat{\mathcal{R}}_{n,m_n,\lambda_n}$.
- $[\cdot]_{\mathcal{B}}$: Barron semi-norm (infinite width weight decay norm).

General Convergence Result

Theorem (Convergence Theorem)

If m and λ scale with n as

$$\frac{\log n}{\sqrt{n}} \ll \lambda \ll 1, \qquad \frac{1}{m} \ll \lambda,$$

then almost surely over the choice of data points, f_n converges to f_∞ (1) in $L^p(\mu)$ for $p < \infty$ and (2) uniformly on compact subsets of $\mathbb{R}^d$, where $f_\infty \equiv f^*$ μ-almost everywhere and $[f_\infty]_\mathcal{B} \leq [f^*]_\mathcal{B}$. Also, with probability $1 - 1/n^2$:

$$\|f_{(a,W,b)_n} - f^*\|^2_{L^2(\mu)} \leq C\left(\frac{[f^*]^2_\mathcal{B}}{m}\,\mathbb{E}_\mu\left[\|x\|^2\right] + [f^*]^2_\mathcal{B}\left(\mathbb{E}_\mu\|x\| + \sigma^2\right)\frac{\log n}{\sqrt{n}} + \lambda\,[f^*]_\mathcal{B}\right).$$

- Comparison with the previous result ([EMW19]).
  1. We allow for general sub-Gaussian rather than compactly supported data distributions.
  2. We do not control the magnitude of the bias variables.
  3. Our results apply to ℓ^2-loss, which is neither globally Lipschitz-continuous nor bounded.
  4. In a limiting regime, we characterize how the empirical risk minimizers interpolate in the region where no data is given by proving uniform convergence to a minimum norm interpolant.
- Minimum norm interpolant is not unique.

Analysis of the bound

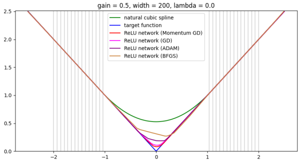
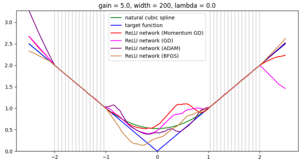
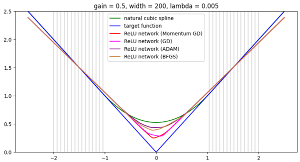
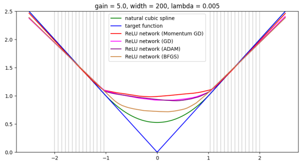
$$C\left(\frac{[f^*]_{\mathcal{B}}^2}{m}\,\mathbb{E}_\mu\left[\|x\|^2\right] + [f^*]_{\mathcal{B}}^2\left(\mathbb{E}_\mu\|x\| + \sigma^2\right)\frac{\log n}{\sqrt{n}} + \lambda\,[f^*]_{\mathcal{B}}\right).$$

- $1/m$ term comes from the risk competitor (a 'good' Two-Layer ReLU network approximator need not be an ERM).
- $\log n/\sqrt{n}$ term comes from sub-Gaussian condition and Rademacher Complexity of Two-Layer ReLU neural networks.
- λ term comes from the weight decay regularizer.
- Dependency on the dimension is implicit in $\mathbb{E}_\mu\|x\|$ and σ terms (and it is 'not' sharp).

- There is no guarantee that a training algorithm (often a ‘local’ algorithm) finds a ‘global’ ERM.
- When the set of minimum norm interpolants is known, we can compare numerical solutions to theoretical predictions to figure out how optimization algorithms work.
  - Examples of known minimum norm interpolants:
    - Convex data in one dimension.
    - Radially symmetric bump functions in odd dimensions.

In 1d, any convex function is a minimum norm interpolant of convex data.

- Consider $f^*(x) = |x|$ with data given for $1 < |x| < 2$.

	Small initialization	Large initialization
No regularization		
Regularization		

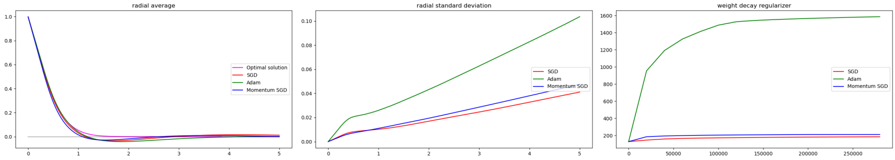
Small initialization and/or regularization lead to minimum norm bias for all optimization algorithms we studied.

Radially Symmetric Bump Function I

For radially symmetric ‘bump function’ data

$$f(0) = 1 \qquad \text{and} \qquad \|x\|_2 \geq 1 \quad \Rightarrow \quad f(x) = 0$$

there exists a *unique* radially symmetric minimum norm interpolant for odd dimensions (but there may be solutions without radial symmetry).



1. All algorithms (SGD, SGD+Momentum, Adam) find a solution with the correct radial average shape without regularization (left).
2. Adam exhibits the lowest degree of symmetry (middle).
3. Adam has by far the highest value for the norm of the weights (right).

$\because$ Adam has a coordinate-wise update, unlike SGD (+Momentum).

$$
\begin{aligned}
g_i &\leftarrow \partial f / \partial \theta_i \\
m_i &\leftarrow \beta_1 m_i + (1-\beta_1) g_i \\
v_i &\leftarrow \beta_2 v_i + (1-\beta_2) g_i^2 \\
\theta_i &\leftarrow \theta_i - \alpha \left(\frac{m_i}{1-\beta_1^t} \right) \Big/ \left(\sqrt{\frac{v_i}{1-\beta_2^t}} + \epsilon \right)
\end{aligned}
$$

Figure 1: Adam updates. Element-wise squaring and re-scaling steps of Adam depend on coordinates.

As in this example, our theoretical results can be used to empirically demonstrate implicit bias of optimization algorithms.

- Obtained convergence of Two-Layer ReLU neural network ERMs with a rate for unbounded data and unbounded, non-Lipschitz loss.
- Proved locally uniform convergence to a minimum norm interpolant (no rate), esepcially even away from the support of the data.
- Demonstrated that known minimum norm interpolants can be used to study implicit bias in optimization.
- In several settings, empirically demonstrated implicit bias towards minimum norm solutions *without* regularization – large initialization exhibits less such bias.

- Generalizing & Sharpening the bound.
  - Different activations.
  - Sharper rate for different norm (e.g. $L^\infty(\mu)$ norm)?

Acknowledgements

This work was partially supported by the NSF DMS-2210689.

Thank You!

Weinan E, Chao Ma, and Lei Wu, *A priori estimates of the population risk for two-layer neural networks*, Communications in Mathematical Sciences **17** (2019), no. 5, 1407–1425.

Weinan E and Stephan Wojtowytsch, *Representation formulas and pointwise properties for barron functions*, 2021.

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

Stephan Wojtowytsch, *Optimal bump functions for shallow relu networks: Weight decay, depth separation and the curse of dimensionality*, arXiv preprint arXiv:2209.01173 (2022).

Appendices

Theorem (Convergence Theorem)

If m and λ scale with n as

$$\frac{\log n}{\sqrt{n}} \ll \lambda \ll 1, \qquad \frac{1}{m} \ll \lambda,$$

then almost surely over the choice of data points, f_n converges to f_∞ (1) in $L^p(\mu)$ for $p < \infty$ and (2) uniformly on compact subsets of $\mathbb{R}^d$, where $f_\infty \equiv f^*$ μ-almost everywhere and $[f_\infty]_{\mathcal{B}} \leq [f^*]_{\mathcal{B}}$. Also, with probability $1 - 1/n^2$:

$$\|f_{(a,W,b)_n} - f^*\|^2_{L^2(\mu)} \leq C\left(\frac{[f^*]^2_{\mathcal{B}}}{m}\,\mathbb{E}_\mu\left[\|x\|^2\right] + [f^*]^2_{\mathcal{B}}\left(\mathbb{E}_\mu\|x\| + \sigma^2\right)\frac{\log n}{\sqrt{n}} + \lambda\,[f^*]_{\mathcal{B}}\right).$$

Proof Ingredients

1. $L^p(\mu)$-convergence:
    1. Rademacher complexity of Two-Layer ReLU networks with bounded weights (but not biases)
    2. Concentration inequalities to bound the magnitude of observed data (with high probability)
    3. 1, 2 $\Rightarrow$ generalization bound with high probability.
    4. Direct approximation theorem to construct a risk competitor.
    5. 2, 3, 4 $\Rightarrow$ $L^2(\mu)$ bound $\leq$ (ERM - risk competitor) $+$ (risk competitor - f^*). The first term is controlled by 3. The second term is controlled from 4.
    6. For $p \neq 2$: Interpolation using the a priori Lipschitz bound from regularization.
2. Minimum norm interpolation (via Γ-convergence):
    - lim inf-inequality: Compact embedding theorem, $L^2(\mu)$-bound, Generalization bound.
    - lim sup-inequality: Direct approximation theorem and concentration for risk competitor.

Definition (Rademacher Complexity)

Let $S_n = \{x_1, \dots, x_n\}$ be a set of points in $\mathbb{R}^d$ (a data sample) and $\mathcal{F}$ a real-valued function class. We define the empirical Rademacher complexity of $\mathcal{F}$ on the data sample as

$$\widehat{\mathrm{Rad}}(\mathcal{F}; S_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right]$$

where ϵ_i are iid random variables which take the values ± 1 with equal probability $\frac{1}{2}$. The population Rademacher complexity is defined as

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}_{S_n \sim \mu^n} \left[\widehat{\mathrm{Rad}}(\mathcal{F}; S) \right],$$

i.e. as the expected empirical Rademacher complexity over a set of n iid data points.

Consider the function classes $\mathcal{F}_Q$ and $\mathcal{F}_Q(R)$:

$$\mathcal{F}_Q = \overline{\mathrm{conv}\left\{a\left(\sigma(w \cdot x + b) - \sigma(b)\right) : a^2 + \|w\|^2 \leq 2Q\right\}}$$

$$\mathcal{F}_Q(R) = \overline{\mathrm{conv}\left\{a\left(\sigma(w \cdot x + b) - \sigma(b)\right) : a^2 + \|w\|^2 \leq 2Q, |b| \leq \sqrt{Q}\,R\right\}}.$$

Lemma

$$\widehat{\mathrm{Rad}}(\mathcal{F}_Q, S_n) \leq \frac{(1+3\sqrt{2})\,Q}{\sqrt{n}} \max_{1\leq i\leq n} \|x_i\|.$$

If in addition μ is a σ^2 sub-Gaussian distribution in $\mathbb{R}^d$. Then for all $n \geq 2$

$$\mathrm{Rad}(\mathcal{F}_Q) \leq (1+3\sqrt{2})\,Q\left(\frac{\mathbb{E}_{x\sim\mu}\left[\|x\|\right]}{\sqrt{n}} + \sigma\sqrt{2\frac{\log n}{n}}\right).$$

Some techniques for this calculation:

- The extreme points of $\sup_{\mathcal{F}_Q} \sum_i \epsilon_i f(x_i)$ are achieved in the boundary of the convex hull (i.e. single width neural network).
    - $\because$ Combining the facts that (1) The functional $f \mapsto \sum_{i=1}^n \epsilon_i f(x_i)$ is a continuous linear functional, and (2) $\mathcal{F}_Q$ is a compact set in $C^0(K)$ and $L^2(\mu)$.
- $\widehat{\mathrm{Rad}}(\mathcal{F}_Q; S_n) = \widehat{\mathrm{Rad}}\big(\mathcal{F}_Q(R); S_n\big)$.
    - $\because$ if $|b| \geq \|w\| R$, then $\sigma(w \cdot x + b) - \sigma(b) = \sigma\big(sgn(b)\big) w \cdot x \Rightarrow$ substitute the cases $|b| \geq \|w\| R$ to $|b| = \|w\| R$.

RC of Two-Layer network II

- Split the variation by the following:

$$\mathbb{E}_\epsilon \left[\sup_{a^2+\|w\|^2 \le Q,\, |b| \le \|w\| R} \sum_{i=1}^{n} \epsilon_i a \big(\sigma(w \cdot x_i + b) - \sigma(b)\big) \right]$$

$$\le \mathbb{E}_\epsilon \left[\sup_{|a|=\|w\| \le \sqrt{Q},\, |b| \le \sqrt{Q} R} \left(\left| \sum_i \epsilon_i a \sigma(w \cdot x_i + b) \right| + \left| \sum_i \epsilon_i a \sigma(b) \right| \right) \right]$$

- First term bound:
  - ReLU: 1-Lipschitz $\Rightarrow$ Contraction Lemma for Rademacher complexity $\Rightarrow$ Can bound by Rademacher complexity of the class of linear functions on Hilbert space (We get $\max_i \|x_i\|$ term here).
- Second term bound:
  - Use ReLU is 1-Lipschitz again and bound on $\mathbb{E}|\sum_i \epsilon_i|$.
- For population quantity, use the concentration of $\max_i \|x_i\|$ to $\mathbb{E}\|x\|$, due to the sub-Gaussian condition (We get $\log n/\sqrt{n}$ term here).

A slight modification of the previous Lemma leads to Rademacher complexity of general Barron functions with controlled bias.

Corollary (RC of Two-Layer ReLU)

Let

$$\mathcal{F}_{A,Q} := \{f \in \mathcal{B} : [f]_{\mathcal{B}} \leq Q, |f(0)| \leq A\}.$$

Under the same conditions as the above Lemma, we have

$$\mathrm{Rad}(\mathcal{F}_{A,Q}) \leq (1 + 3\sqrt{2})\, Q \left(\frac{\mathbb{E}_{x\sim\mu}\left[\|x\|\right]}{\sqrt{n}} + \sigma\sqrt{2\,\frac{\log n}{n}} \right) + \frac{A}{\sqrt{n}}$$

$\Rightarrow$ Rademacher complexity we obtained enable us to obtain a generalization bound.

TEXAS A&M UNIVERSITY

Corollary (Generalization bound)

Let

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \big|f(X_i) - f^*(X_i)\big|^2, \qquad \mathcal{R}(f) = \mathbb{E}_{x\sim\mu}\big[|f(x) - f^*(x)|^2\big].$$

If f^* satisfies $|f^*(x) - f^*(0)| \leq B_1 + B_2\|x\|$ μ-almost everywhere, then with probability at least $1 - 2\delta$,

$$\sup_{f - f^*(0) \in \mathcal{F}_{A,Q}} \big(\mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)\big) \leq C^*\Big(\big(Q + B_2\big)\big(\mathbb{E}_{x\sim\mu}\|x\| + \sigma^2 + 1\big) + A + B_1\Big)^2 \frac{\log(n/\delta)}{\sqrt{n}}$$

Generalization bound II

Proof techniques:

- Split $\mathbb{E}_\mu\left[(f(x) - f^*(x))^2\right]$ to $\mathbb{E}_\mu\left[(f(x) - f^*(x))^2 \mathbb{1}_{\|x\|\leq R}\right] + \mathbb{E}_\mu\left[(f(x) - f^*(x))^2 \mathbb{1}_{\|x\|>R}\right]$.
- First term: Regard it as using bounded Lipschitz loss $\Rightarrow$ Apply canonical method of obtaining generalization bound from Rademacher complexity (See [SSBD14] Thm 26.5.).
- Second term: Use the fact $|f(x) - f^*(x)| \leq \left|f(x) - f^*(0)\right| + |f^*(x) - f^*(0)| \leq (A + B_1) + (Q + B_2)\|x\|$ and sub-Gaussian properties of $\|x\|$.

Risk competitor

Theorem (Direct approximation, [Woj22] Prop. 2.6.)

Let $f \in \mathcal{B}$ and μ a measure on $\mathbb{R}^d$ with finite second moments. Then for any $m \in \mathbb{N}$ there exist $c \in \mathbb{R}$ and $(a_i, w_i, b_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ such that

$$\sum_{i=1}^{m} a_i^2 + \|w_i\|^2 \leq [f]_{\mathcal{B}},$$

$$\left\| f - c - \sum_{i=1}^{m} a_i \sigma(w_i^T x + b_i) \right\|_{L^2(\mu)} \leq \frac{2\,[f]_{\mathcal{B}}}{\sqrt{m}} \sup_{\|w\|=1} \sqrt{\int_{\mathbb{R}^d} |w^T x|^2 \,\mathrm{d}\mu_x}.$$

- Given f^*, we can obtain $f_{\tilde{\theta}}$ from the direct approximation theorem, which we call a risk competitor.

$L^2(\mu)$ bound I

Theorem (L^2-convergence)

Let $\hat\theta \in \operatorname{argmin}_\theta \widehat{\mathcal{R}}_{n,m,\lambda}(\theta)$. *If* $\delta \geq e^{-n}$, *and* $f^* \in \mathcal{F}_{Q^*}$, *then with probability at least* $1 - 4\delta$ *over the choice of random points* $x_1, \dots, x_n$ *we have*

$$\mathcal{R}(f_{\hat\theta}) \leq C \left(\frac{(Q^*)^2}{m} \left(\mathbb{E}\left[\|x\|^2\right]\right) + \lambda Q^* + Q^* \left(\mathbb{E}\|x\| + \sigma^2 + [f^*]_{\mathcal{B}}\right) \frac{\log(n/\delta)}{\sqrt{n}} \right)$$

up to higher order terms in the small quantities $(\lambda m)^{-1}, m^{-1}, n^{-1/2}\log n$.

Idea of proof:

1. Note the following:

$$\begin{aligned}\mathcal{R}(f_{\hat\theta}) &= \widehat{\mathcal{R}}_n(f_{\hat\theta}) + \mathcal{R}(f_{\hat\theta}) - \widehat{\mathcal{R}}_n(f_{\hat\theta})\\ &\le \widehat{\mathcal{R}}_{n,m,\lambda}(\hat\theta) + \mathcal{R}(f_{\hat\theta}) - \widehat{\mathcal{R}}(f_{\hat\theta})\\ &\le \widehat{\mathcal{R}}_{n,m,\lambda}(\tilde\theta) + \mathcal{R}(f_{\hat\theta}) - \widehat{\mathcal{R}}_n(f_{\hat\theta})\end{aligned}$$

2. First term in 1 is directly bounded using direct approximation theorem.
3. Second term: use generalization bound with suitable choice of Q and A.
   - Q: $[f_{\hat\theta}]_{\mathcal{B}} \le \frac{1}{\lambda}\widehat{\mathcal{R}}_{n,m,\lambda}(\hat\theta) \le \frac{1}{\lambda}\widehat{\mathcal{R}}_{n,m,\lambda}(\tilde\theta) = [f^*]_{\mathcal{B}} + O((\lambda m)^{-1})$, which implies we can use $Q = C[f^*]_{\mathcal{B}}$
   - A: Use the fact that Barron function is $[\cdot]_{\mathcal{B}}$-Lipschitz and apply the above.

General $L^p(\mu)$ convergence I

Corollary (L^p-convergence)

Let $p \in [1, \infty]$ and $\hat{\theta}$ is ERM. Then there exists a constant $\tilde{C} > 0$ depending on $\mathbb{E}\|x\|, \mathbb{E}[\|x\|^2], \sigma^2$ and p such that

$$\|f_{\hat{\theta}} - f^*\|_{L^p(\mu)} \leq \tilde{C} \left(\widehat{\mathcal{R}}_{n,m,\lambda}(\hat{\theta})^{1/2} + [f^*]_{\mathcal{B}}\right)^{1-1/p} \|f_{\hat{\theta}} - f^*\|_{L^2(\mu)}^{1/p}.$$

- $p < 2$: From the fact $L^2(\mu)$ embeds continuously into $L^p(\mu)$.
- $p > 2$: Apply the following fact with $g = f_{\hat{\theta}} - f^*$:
  - if g is a measurable function which satisfies $|g(x)| \leq C_g(1 + \|x\|)$ for some $C_g > 0$, then

$$\|g\|_{L^p(\mu)}^p = \mathbb{E}\left[g \cdot g^{p-1}\right] \leq \mathbb{E}\left[g^2\right]^{1/2} \mathbb{E}\left[g^{2(p-1)}\right]^{1/2} = \|g\|_{L^2} \|g\|_{L^{2(p-1)}}^{p-1}.$$

  - For $\|f_{\hat{\theta}} - f^*\|_{L^{2(p-1)}}$, we use the following from the Lipschitz condition:

$$\|f_{\hat{\theta}} - f^*\|_{L^{2(p-1)}(\mu)} \leq C\left(|f_{\hat{\theta}} - f^*|(0) + [f_{\hat{\theta}} - f^*]_{\mathcal{B}}\right).$$

    - Bound of the first term was already derived as $\widehat{\mathcal{R}}_{n,m,\lambda}(\hat{\theta})^{1/2} + C[f^*]_{\mathcal{B}}$ in $L^2(\mu)$ convergence analysis (when figuring out A in generalization bound).
    - $[f_{\hat{\theta}} - f^*]_{\mathcal{B}} \leq [f_{\hat{\theta}}]_{\mathcal{B}} + [f^*]_{\mathcal{B}} \leq \widehat{\mathcal{R}}_{n,m,\lambda}(\hat{\theta}) + [f^*]_{\mathcal{B}} \lesssim \widehat{\mathcal{R}}_{n,m,\lambda}(\hat{\theta})^{1/2} + [f^*]_{\mathcal{B}}$.

Overview so far...

1. $L^p(\mu)$-convergence: Done!
   1. Rademacher complexity of Two-Layer ReLU networks with bounded weights (but not biases)
   2. Concentration inequalities to bound the magnitude of observed data (with high probability)
   3. 1, 2 $\Rightarrow$ generalization bound with high probability.
   4. Direct approximation theorem to construct a risk competitor.
   5. 2, 3, 4 $\Rightarrow$ $L^2(\mu)$ bound $\leq$ (ERM - risk competitor) + (risk competitor - f^*). The first term is controlled by 3. The second term is controlled from 4.
   6. For $p > 2$: Interpolation using the a priori Lipschitz bound from regularization.
2. Minimum norm interpolation (via Γ-convergence): Next Step!
   - lim inf-inequality: Compact embedding theorem, $L^2(\mu)$-bound, Generalization bound.
   - lim sup-inequality: Direct approximation theorem and concentration for risk competitor.

A concept from the calculus of variation that is useful for the convergence of minimization problems.

Definition

Let (X, d) be a metric space and $F_n, F : X \to \mathbb{R} \cup \{-\infty, \infty\}$ be functions. We say that F_n converges to F in the sense of Γ-convergence if two conditions are met:

1. (lim inf-inequality) If x_n is a sequence in X and $x_n \to x$, then $\liminf_{n\to\infty} F_n(x_n) \geq F(x)$.
2. (lim sup-inequality) For every $x \in X$, there exists a sequence $x_n^* \in X$ such that $x_n^* \to x$ and $\limsup_{n\to\infty} F_n(x_n^*) \leq F(x)$.

- Remark: Γ-convergence depends on the convergence of the base space X.

Usefulness of Γ-convergence: Guarantees an empirical minimizer converging to a population minimizer (but without explicit rate).

Lemma

Assume that $F_n \to F$ in the sense of Γ-convergence, $\epsilon_n \to 0^+$ and $x_n \in X$ is a sequence such that

$$F_n(x_n) \leq \inf_{x\in X} F_n(x) + \epsilon_n.$$

Assume that $x_n \to x^*$. Then $F(x^*) = \inf_{x\in X} F(x)$. In particular, if x_n is a minimizer of F_n and the sequence x_n converges, then the limit point is a minimizer of F.

Γ-expansion of the problem

- Convergence in the base space:
  - Define $f_k \xrightarrow{good} f$ if $f_k \to f$ uniformly on compact sets and in $L^2(\mu)$.
  - Define $\theta_k \xrightarrow{good} f$ if $f_{\theta_k} \xrightarrow{good} f$.
- Γ-functional:
  - $F_n(\theta) := \widehat{\mathcal{R}}_{n,m_n,\lambda_n}(f_\theta)/\lambda_n$
  - $F(f) = [f]_{\mathcal{B}}$ if $f = f^* \ \mu - a.s.$ and $+\infty$ o.w.

Theorem (Γ-convergence of the risk functional)

Given above constructions, $\Gamma - \lim_{n\to\infty} F_n = F$ a.s. with respect to the notion of convergence $\theta_k \xrightarrow{good} f$.

$\Rightarrow$ Since F's minimizer is a minimum norm interpolant, the theorem gives ERM's convergence to a minimum norm interpolant.

lim inf-inequality:

- Strategy: Divide the case when $f = f^* \ \mu - a.s.$ and not.
  - $f = f^* \ \mu - a.s.$:

$$\liminf_{n\to\infty} F_n(\theta_n) \geq \liminf_{n\to\infty} R_{WD}(\theta_n) \geq \liminf_{n\to\infty} [f_{\theta_n}]_{\mathcal{B}} \geq [f]_{\mathcal{B}} = F(f)$$

    where third inequality comes from lower semi-continuity of Barron semi-norm.
  - $f \neq f^* \ \mu - a.s.$: Show $\liminf_n F_n(\theta_n) \geq F(f) = \infty$ for $\forall \theta_n \xrightarrow{good} f$:

$$F_n(\theta_n) \geq \frac{\widehat{\mathcal{R}}_n(f_{\theta_n}) - \mathcal{R}(f_{\theta_n})}{\lambda_n} + \frac{\|f - f^*\|^2_{L^2(\mu)} + \|f_{\theta_n} - f\|^2_{L^2(\mu)}}{\lambda_n} + [f_{\theta_n}]_{\mathcal{B}}$$

$$\geq O\left(\frac{\log n}{\lambda_n \sqrt{n}}\right) + \frac{\|f - f^*\|^2_{L^2(\mu)}}{\lambda_n} + [f_{\theta_n}]_{\mathcal{B}} \to \infty.$$

lim sup-inequality:

- Strategy: Only one sequence is sufficient
  $\Rightarrow$ for given f take $\tilde{\theta}_n$ from Direct approximation.
    - $f = f^* \ \mu - a.s.$:

$$F_n(\tilde{\theta}_n) \leq \frac{C}{\lambda_n m_n}\left(1 + \frac{\log n}{\sqrt{n}}\right) + [f]_{\mathcal{B}} \to [f]_{\mathcal{B}} = F(f).$$

    - $f \neq f^* \ \mu - a.s.$: Since $F(f) = \infty$, any sequence θ_n satisfy $F_n(\theta_n) \leq \infty = F(f)$.