Mathematics of Deep Learning, Summer Term 2020

Week 7

Sparse Data Representation

Philipp Harms   Lars Niemann

University of Freiburg

Overview of Week 7

1. Rate-Distortion Theory
2. Hypercube Embeddings and Ball Coverings
3. Dictionaries as Encoders
4. Frames as Dictionaries
5. Networks as Encoders
6. Dictionaries as Networks
7. Wrapup

Acknowledgement of Sources

Sources for this lecture:

- Bölcskei, Grohs, Kutyniok, Petersen (2017): Optimal approximation with sparsely connected deep neural networks. In: SIAM Journal on Mathematics of Data Science 1.1, pp. 8–45
- Dahlke, De Mari, Grohs, Labatte (2015): Harmonic and Applied Analysis. Birkhäuser.

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 1

Rate-Distortion Theory

Philipp Harms Lars Niemann

University of Freiburg

Encoding, Decoding, and Distortion

Definition

Let $\mathcal{H}$ be a normed space, let $\mathcal{C} \subseteq \mathcal{H}$ be a signal class, and let $l \in \mathbb{N}$.

- The set of binary encoders of $\mathcal{C}$ with runlength l is defined as
$$\mathcal{E}^l := \left\{E : \mathcal{C} \to \{0,1\}^l\right\} .$$
- The set of binary decoders with runlength l is defined as
$$\mathcal{D}^l := \left\{D : \{0,1\}^l \to \mathcal{H}\right\} .$$
- The distortion of an encoder-decoder pair $(E, D) \in \mathcal{E}^l \times \mathcal{D}^l$ is defined as
$$\delta(E, D) := \sup_{f \in \mathcal{C}} \|f - D(E(f))\|_{\mathcal{H}} .$$

Remark: Alternatively, in probabilistic settings, one can consider the expected distortion $\mathbb{E}[\|f - D(E(f))\|_{\mathcal{H}}]$.

Encoding Rate

Definition

The optimal encoding rate of a signal class $\mathcal{C}$ in a normed space $\mathcal{H}$ is defined as

$$s^*_{\text{enc}}(\mathcal{C}) := \sup \left\{ s > 0 \middle| \inf_{(E,D)\in\mathcal{E}^l\times\mathcal{D}^l} \delta(E,D) = \mathcal{O}(l^{-s}) \right\}.$$

Remark:

- The optimal encoding rate quantifies the complexity of a signal class.
- The interpretation is information-theoretic: for any $s < s^*_{\text{enc}}(\mathcal{C})$, one can compress signals $f \in \mathcal{C}$ using l-bit encodings with distortion l^{-s}.
- Rate-distortion theory is the mathematical branch of information theory which studies data compression problems by analyzing the trade-off between compression rates and distortion.

Examples: Signal Classes

- Continuously differentiable functions:
  $\mathcal{C}_K^k(C) := \{f \in L^2(\mathbb{R}^d) \mid f \in C^k,\ \|f\|_{C^k} \le K,\ \operatorname{supp} f \subseteq C\}$, where $C \subseteq \mathbb{R}^d$ is a smooth bounded domain.
- Piecewise continuously differentiable functions:
  $\mathcal{C}_K^{k,pw}(I) := \{f_1 \mathbb{1}_{[0,c)} + f_2 \mathbb{1}_{[c,1)} \mid c \in I,\ f_1, f_2 \in \mathcal{C}_K^k(I)\}$, where $I = (a,b)$ is an open interval.
- Star-shaped images:
  $\mathrm{STAR}_K^2 := \{\mathbb{1}_B \mid B$ is interior of Jordan curve $\rho \in C^2,\ \|\rho\|_{C^2} \le K\}$.
- Cartoon images:
  $\mathrm{CART}_K^2 := \{f_1 \mathbb{1}_B + f_2 \mid \mathbb{1}_B \in \mathrm{STAR}_K^2,\ f_1, f_2 \in \mathcal{C}_K^2([0,1]^2)\}$.
- Textures: $\mathrm{TEXT}_{K,M}^k := \{\sin(Mf)g \mid f, g \in \mathcal{C}_K^k([0,1]^2)\}$.
- Mutilated functions: $\mathrm{MUTIL}_K^k := \{g(u\ \cdot)h \mid g \in \mathcal{C}_K^{k,pw}(\mathbb{R}),\ h \in \mathcal{C}_K^k([0,1]^d),\ u \in \mathbb{R}^d,\ \|u\| = 1\}$.

Remark: All introduced signal classes are relatively compact in $L^2(\mathbb{R}^d)$.

Examples: Optimal Encoding Rates

Remark: The main goal of this week's lecture is to establish the following optimal encoding rates and to show that they are achieved by deep neural networks.

Theorem

- $s^*_{\text{enc}}(\mathcal{C}^k_K(C)) = k/d.$
- $s^*_{\text{enc}}(\mathcal{C}^{k,pw}_K(I)) = k.$
- $s^*_{\text{enc}}(\text{STAR}^2_K) = 1.$
- $s^*_{\text{enc}}(\text{CART}^2_K) = 1.$
- $s^*_{\text{enc}}(\text{TEXT}^k_{K,M}) = k/2.$
- $s^*_{\text{enc}}(\text{MUTIL}^k_K) = k/d.$

Sketch of Proof:

- Upper bounds on encoding rates: Hypercubes are difficult to encode. If $\mathcal{C}$ contains hypercubes, then $\mathcal{C}$ is difficult to encode. See Video 2.
- Lower bounds on encoding rates: If signals in $\mathcal{C}$ have Banach frame coefficients with fast decay, then picking the n largest among the first n^k frame coefficients defines a good encoder. See Video 4. $\square$

Paradigm: Analysis by Synthesis

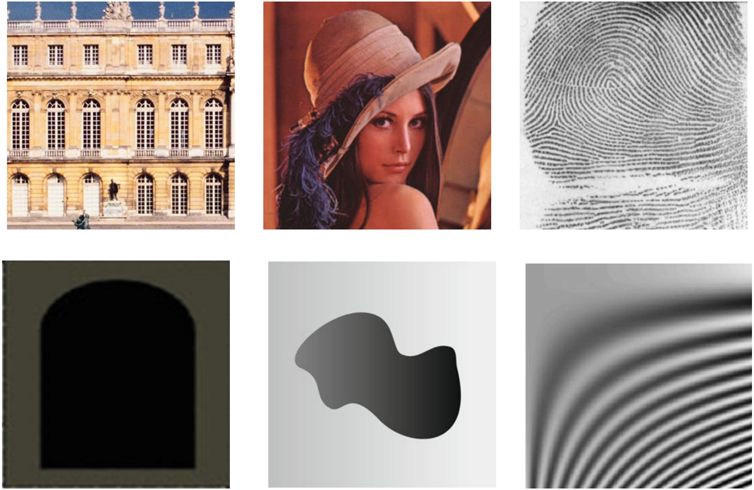

Figure: Real-world images (top) can be analyzed by synthesizing them from simpler image elements (bottom) such as star-shaped domains, cartoons, or textures. Additional benefits are compression and denoising. [Dahlke, Fig. 5.1–3]

Questions to Answer for Yourself / Discuss with Friends

- Repetition: What is an endoding-decoding pair, and how are optimal encoding rates defined?
- Check: How many bits are needed to encode a natural number in $\{1, \ldots, n\}$?
- Background: The definition of star-shaped images involves Jordan curves—can you recall their definition and main properties?
- Context: Read some introductory articles (e.g. on Wikipedia) on data compression and rate-distortion theory.

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 2

Hypercube Embeddings and Ball Coverings

Philipp Harms   Lars Niemann

University of Freiburg

Hypercube Embeddings

Definition (Donoho 2001)

Let $\mathcal{C}$ be a signal class in $\mathcal{H}$, and let $p > 0$.

- A hypercube of dimension $m \in \mathbb{N}$ and side-length $\delta > 0$ is a set of the form
$$\left\{ f + \sum_{i=1}^{m} \epsilon_i \psi_i \middle| \epsilon_i \in \{0,1\} \right\},$$
where $f \in \mathcal{C}$, and ψ_i are orthogonal functions in $\mathcal{H}$ with $\|\psi_i\|_{\mathcal{H}} \geq \delta$.
- The signal class $\mathcal{C}$ is said to contain a copy of ℓ_0^p if it contains for each $k \in \mathbb{N}$ a hypercube with dimension m_k and side-length δ_k such that
$$\delta_k \to 0 \quad \text{and} \quad m_k^{-1/p} = \mathcal{O}(\delta_k) \quad \text{as } k \to \infty.$$

Remark: A ball of radius r in ℓ^p contains hypercubes of dimension $m \in \mathbb{N}$ with side-length $rm^{-1/p}$.

Hypercube Embeddings and Encoding Rates

Remark: For many signal classes, hypercube embeddings are easy to construct and provide (sharp) upper bounds on the encoding rate.

Theorem

If a signal class $\mathcal{C}$ in $\mathcal{H}$ contains a copy of ℓ_0^p for some $p \in (0,2]$, then

$$s^*_{\mathrm{enc}}(\mathcal{C}) \le \frac{1}{p} - \frac{1}{2} \,.$$

Proof: Hypercube Embeddings and Encoding Rates

Idea of proof: (See [Dahlke e.a., Theorem 5.12] for a full proof.)

- Hypercubes of dimension m can be identified with bit streams in $\{0,1\}^m$.
- Recall that the Hamming distance (aka. ℓ^1 or Manhattan distance) between two bit streams is the number of unequal bits.
- Chernoff's bounds imply that for any compression rate $\alpha \in (0,1)$, there exists $C > 0$ such that for any $m \in \mathbb{N}$ and encoder-decoder

$$E\colon \{0,1\}^m \to \{0,1\}^{\lfloor \alpha m \rfloor}, \qquad D\colon \{0,1\}^{\lfloor \alpha m \rfloor} \to \{0,1\}^m,$$

  the distortion in the Hamming distance is lower-bounded by Cm.
- This translates into a lower bound on the encoding rate of a hypercube as well as its containing signal class. □

Examples: Upper Bounds on Optimal Encoding Rates

Remark: The following are special cases of the above theorem.

Corollary

The following upper bounds on encoding rates are achieved via hypercube embeddings:

- $s^*_{\mathrm{enc}}(\mathcal{C}^k_K(C)) \le k/d$ *via embedding of* $\ell_0^{1/(\frac{k}{d}+\frac{1}{2})}$
- $s^*_{\mathrm{enc}}(\mathcal{C}^{k,pw}_K(I)) \le k$ *via embedding of* $\ell_0^{1/(k+\frac{1}{2})}$
- $s^*_{\mathrm{enc}}(\mathrm{STAR}^2_K) \le 1$ *via embedding of* $\ell_0^{2/3}$
- $s^*_{\mathrm{enc}}(\mathrm{CART}^2_K) \le 1$ *via embedding of* $\ell_0^{2/3}$
- $s^*_{\mathrm{enc}}(\mathrm{TEXT}^k_{K,M}) \le k/2$ *via embedding of* $\ell_0^{2/(k+1)}$
- $s^*_{\mathrm{enc}}(\mathrm{MUTIL}^k_K) \le k/d$ *via embedding of* $\ell_0^{1/(\frac{k}{d}+\frac{1}{2})}$

Examples: Upper Bounds on Optimal Encoding Rates

Idea of proof: For a fixed bump function ψ, one uses hypercubes of the following forms:

- $\sum_{i=0}^{n-1} \epsilon_i \psi(nx - i)$ for piece-wise continuously differentiable functions,
- $\mathbb{1}_{\{\|x\| \leq 1\}} + \sum_{i=0}^{n-1} \epsilon_i \big(\mathbb{1}_{\{\|x\| \leq i/n\}} - \mathbb{1}_{\{\|x\| \leq 1\}} \big)$ for star-shaped images, or
- $\sum_{i,j=1}^{n-1} \epsilon_{i,j} \sin \big(n^{-k} \psi(nx - i) \psi(ny - j) \big)$ for textures, etc.

See [Dahlke e.a., Theorem 5.17] for a full proof. □

Digression: Kolmogorov Entropy

Remark:

- Encoding rates are closely related to covering numbers and Kolmogorov entropy.
- We have already encountered the Kolmogorov entropy in the context of statistical learning theory.
- Unfortunately, covering numbers are often difficult to compute and therefore of rather theoretical interest.

Definition

Let $\mathcal{H}$ be a metric space, and let $\mathcal{C} \subseteq \mathcal{H}$ be a relatively compact subset.

- The covering number of $\mathcal{C}$ is defined for any $\epsilon > 0$ as the smallest number $N_\epsilon(\mathcal{C})$ of ϵ-balls required to cover $\mathcal{C}$.
- The Kolmogorov entropy of $\mathcal{C}$ is defined as $H_\epsilon(\mathcal{C}) := \log_2(N_\epsilon(\mathcal{C}))$.

Digression: Kolmogorov Entropy and Encoding Rates

Lemma

Let $\mathcal{C} \subseteq \mathcal{H}$ be a relatively compact signal class in a normed space $\mathcal{H}$. Then the optimal encoding rate $s^*_{\mathrm{enc}}(\mathcal{C})$ is related to the Kolmogorov entropy $H_\epsilon(\mathcal{C})$ by

$$s^*_{\mathrm{enc}}(\mathcal{C}) = \sup\left\{ s > 0 : H_\epsilon(\mathcal{C}) = \mathcal{O}(\epsilon^{-\frac{1}{s}}) \right\}.$$

Proof:

- Given a pair (E, D) of length l that achieves distortion ϵ, the ϵ-balls centered at $D(\xi)$, $\xi \in \{0,1\}^l$, cover $\mathcal{C}$.
- Conversely, given $\epsilon > 0$, we can find $N_\epsilon := 2^{H_\epsilon(\mathcal{C})}$ centers whose ϵ-neighborhoods cover $\mathcal{C}$. Encode $\mathcal{C}$ using the binary representation of the nearest center, and decode by reversing this process. □

Questions to Answer for Yourself / Discuss with Friends

- Repetition: How are upper bounds on the encoding rate obtained from hypercube embeddings?
- Check: Show that relatively compact signal classes have finite covering numbers.
- Background: Skim through the construction of hypercube embeddings for specific signal classes in [Dahlke e.a., Theorem 5.17].
- Transfer: The upper bounds on the optimal encoding rates decay inversely proportional to the dimension—an instance of the curse of dimensionality.

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 3

Dictionaries as Encoders

Philipp Harms   Lars Niemann

University of Freiburg

Repetition: Approximation Rates of Dictionaries

Definition

A dictionary $(\phi_\lambda)_{\lambda \in \Lambda}$ in $\mathcal{H}$ achieves an approximation rate of $(h_n)_{n \in \mathbb{N}}$ if

$$\sigma(\Sigma_n(\phi), \mathcal{C}) := \sup_{f \in \mathcal{C}} \inf_{g \in \Sigma_n(\phi)} \|f - g\|_{\mathcal{H}} = \mathcal{O}(h_n) \quad \text{as } n \to \infty,$$

where $\Sigma_n(\phi)$ denotes the set of n-term linear combinations in ϕ.

Remark:

- A dense dictionary ϕ in $\mathcal{H}$ achieves any approximation rate for any signal class. Nevertheless, it is ill-suited for efficient encoding of functions.
- This motivates the requirement of polynomial-depth search, which is described next.
- We restrict ourselves to polynomial rates $h_n = n^{-s}$, $s > 0$, as these are most relevant.

Dictionary Approximation with Polynomial-Depth Search

Definition (Donoho 2001)

Let $\phi = (\phi_i)_{i\in\mathbb{N}}$ be a dictionary, π a univariate polynomial, $\mathcal{C}$ a signal class in $\mathcal{H}$, and $n \in \mathbb{N}$.

- The set of n-term linear combinations in ϕ with polynomial-depth search is defined as

$$\Sigma_n^\pi(\phi) = \left\{ \sum_{i=1}^{\pi(n)} c_i\phi_i \middle| c_i \in \mathbb{R} \text{ with } \|c\|_0 \leq n \right\}.$$

- The approximation rate of ϕ with polynomial-depth search is defined as

$$s^*_{\text{dict}}(\mathcal{C}, \phi) := \sup \left\{ s > 0 \middle| \exists \pi : \sup_{f\in\mathcal{C}} \inf_{g\in\Sigma_n^\pi(\phi)} \|g - f\|_{\mathcal{H}} = \mathcal{O}(n^{-s}) \right\}$$

Remark: Here, the dictionary needs to be ordered, i.e., indexed over $\mathbb{N}$.

Encoding via Dictionaries

Remark: Polynomial-depth search leads to the desired link between dictionary approximation rates and encoding rates:

Theorem

For any dictionary ϕ and bounded signal class $\mathcal{C}$ in $\mathcal{H}$,

$$s^*_{\mathrm{enc}}(\mathcal{C}) \geq s^*_{\mathrm{dict}}(\mathcal{C}, \phi) \ .$$

Remark:

- A dictionary ϕ is called rate-optimal if equality holds above.
- Explicit dictionary approximation rates can be obtained for Hilbert or Banach frames, as shown in the next video.

Proof: Encoding via Dictionaries

Proof:

- We start by constructing an encoder. For any $s < s^*_{\mathrm{dict}}(\mathcal{C}, \phi)$, there exists a polynomial π and a constant $C > 0$ such that for all $n \in \mathbb{N}$ and $f \in \mathcal{C}$, there exist coefficients $c_i \in \mathbb{R}$ with $\|c\|_0 \leq n$ such that

$$\left\| f - \sum_{i=1}^{\pi(n)} c_i \phi_i \right\|_{\mathcal{H}} \leq C n^{-s} .$$

- The set $\Lambda_n := \{i \in \mathbb{N} : c_i \neq 0\}$ can be encoded using $\mathcal{O}(n \log n)$ bits thanks to the assumption of polynomial-depth search.
- Applying the Gram-Schmidt orthonormalization to $\phi_{\Lambda_n} := (\phi_\lambda)_{\lambda \in \Lambda_n}$ yields an orthonormal set $\tilde{\phi}_{\Lambda_n} := (\tilde{\phi}_\lambda)_{\lambda \in \Lambda_n}$. Some $\tilde{\phi}_\lambda$ may be zero.

Proof: Encoding via Dictionaries (cont.)

- Determine coefficients $\tilde{c}_\lambda$ uniquely by

$$\sum_{\lambda \in \Lambda_n} \tilde{c}_\lambda \tilde{\phi}_\lambda = \sum_{\lambda \in \Lambda_n} c_\lambda \phi_\lambda, \qquad \tilde{c}_\lambda = 0 \text{ if } \tilde{\phi}_\lambda = 0.$$

- Note that

$$\Big\| f - \sum_{\lambda \in \Lambda_n} \tilde{c}_\lambda \tilde{\phi}_\lambda \Big\|_{\mathcal{H}} \leq C n^{-s}$$

  and that the sequence $\tilde{c}$ is ℓ^2-bounded uniformly in n and f. (Here enters the boundedness of $\mathcal{C}$.)

- Rounding the coefficients $\tilde{c}_\lambda$ up to multiples of $n^{-(s+\frac{1}{2})}$ encodes them with a bit string of length $\mathcal{O}(n \log n)$.
- Altogether, this gives an encoding procedure $E_l : \mathcal{C} \to \{0,1\}^l$ with length $l = \mathcal{O}(n \log n)$.

Proof: Decoding via Dictionaries

- Decoding is done by reversing this process: starting from a bit string ξ, reconstruct the set Λ_n and the rounded approximations $\hat{c}_\lambda$ of $\tilde{c}_\lambda$, and define the decoder

$$D_n \colon \{0,1\}^l \to \mathcal{H}, \qquad D_l(\xi) := \sum_{\lambda \in \Lambda_n} \hat{c}_\lambda \tilde{\phi}_\lambda \,.$$

- It remains to control the distortion:

$$\begin{aligned}
\|f - D_l(E_l(f))\|_{\mathcal{H}} &= \Big\| f - \sum_{\lambda \in \Lambda_n} \hat{c}_\lambda \tilde{\phi}_\lambda \Big\|_{\mathcal{H}} \\
&\leq \Big\| f - \sum_{\lambda \in \Lambda_n} \tilde{c}_\lambda \tilde{\phi}_\lambda \Big\|_{\mathcal{H}} + \Big\| \sum_{\lambda \in \Lambda_n} \hat{c}_\lambda \tilde{\phi}_\lambda - \sum_{\lambda \in \Lambda_n} \tilde{c}_\lambda \tilde{\phi}_\lambda \Big\|_{\mathcal{H}} \\
&\leq C n^{-s} + \max_{\lambda \in \Lambda_n} |\tilde{c}_\lambda - \hat{c}_\lambda| \, n^{\frac{1}{2}} \leq C n^{-s}.
\end{aligned}$$

□

Questions to Answer for Yourself / Discuss with Friends

- Repetition: How are lower bounds on encoding rates obtained from dictionary approximation rates?
- Check: The approximation rate of a dense dictionary is arbitrarily high—what about the approximation rate with polynomial-depth search?
- Check: Verify that the coefficients $\tilde{c}$ after Gram–Schmidt orthogonalization are ℓ^2-bounded uniformly in $n \in \mathbb{N}$ and $f \in \mathcal{C}$. Hint: $\|\tilde{c}\|_{\ell^2} = \|\sum_\lambda \tilde{c}_\lambda \tilde{\phi}_\lambda\|_{\mathcal{H}}$.
- Transfer: Nonlinear approximation spaces $\mathcal{C}$ are *defined* by the requirement that $s^*(\mathcal{C}, \phi) = s$ for given $s \in \mathbb{R}$.

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 4

Frames as Dictionaries

Philipp Harms   Lars Niemann

University of Freiburg

Repetition: Hilbert Frames

Remark: Recall that Hilbert frames are Banach frames in Hilbert spaces with respect to the sequence space ℓ^2; this boils down to the following:

Definition

- A Hilbert frame in a Hilbert space $\mathcal{H}$ is a dictionary $\phi = (\phi_\lambda)_{\lambda\in\Lambda}$ s.t.

$$\forall f \in \mathcal{H}: \qquad \|f\|_H^2 \lesssim \sum_{\lambda\in\Lambda} |\langle f, \phi_\lambda\rangle_{\mathcal{H}}|^2 \lesssim \|f\|_{\mathcal{H}}^2.$$

- A dual frame for ϕ is a complementary dictionary $\tilde{\phi} = (\tilde{\phi}_\lambda)_{\lambda\in\Lambda}$ s.t.

$$\forall f \in \mathcal{H}: \qquad f = \sum_{\lambda\in\Lambda} \langle f, \tilde{\phi}_\lambda\rangle_{\mathcal{H}} \phi_\lambda = \sum_{\lambda\in\Lambda} \langle f, \phi_\lambda\rangle_{\mathcal{H}} \tilde{\phi}_\lambda.$$

Remark: Every Hilbert frame has a dual frame, for instance the canonical one, which is determined by $\phi_\mu = \sum_\lambda \langle \tilde{\phi}_\mu, \phi_\lambda\rangle_{\mathcal{H}} \phi_\lambda$, or the one from the definition of Banach frames.

Weak ℓ^p Spaces

Remark: Recall that a quasi-norm is a norm without a triangle inequality.

Definition

The weak ℓ^p-quasinorm of a sequence $c := (c_k)_{k\in\mathbb{N}}$ is defined for any $p > 0$ as

$$\|c\|_{w\ell^p}^p := \sup_{t>0} t^p \; \#\{k \in \mathbb{N} : |c_k| > t\},$$

and the space $w\ell^p$ consists of all sequences with finite weak ℓ^p-quasinorm.

Remark:

- For any $p \geq 1$, the space ℓ^p embeds continuously in $w\ell^p$ because

$$\|c\|_{\ell^p}^p \geq \sum_k t^p \mathbb{1}_{\{k:|c_k|>t\}} + \sum_k |c_k|^p \mathbb{1}_{\{k:|c_k|\leq t\}} \geq t^p \#\{k : |c_k| > t\}\,.$$

- The space $w\ell^p$ coincides with the Lorentz space $\ell^{p,\infty}$, is complete, and is normable for $p > 1$. Weak L^p spaces are defined similarly.

Approximation via Frames

Remark: We next show that weak ℓ^p bounds on Hilbert frame coefficients translate into dictionary approximation rates.

Theorem

Let $(\phi_n)_{n\in\mathbb{N}}$ be a Hilbert frame with dual frame $(\tilde{\phi}_n)_{n\in\mathbb{N}}$ in a Hilbert space $\mathcal{H}$, and let $\mathcal{C}$ be a signal class in $\mathcal{H}$ which satisfies the weak ℓ^p bound

$$\sup_{f\in\mathcal{C}} \left\| (\langle f, \tilde{\phi}_n \rangle_{\mathcal{H}})_{n\in\mathbb{N}} \right\|_{w\ell^p} < \infty$$

and, for some $\alpha > 0$, the ℓ^2 tail bound

$$\sup_{f\in\mathcal{C}} \sum_{i\geq n} |\langle f, \tilde{\phi}_i \rangle|^2 = \mathcal{O}(n^{-\alpha}).$$

Then $s^*_{\mathrm{dict}}(\mathcal{C}, \phi) \geq \frac{1}{p} - \frac{1}{2}$.

Proof: Approximation via Frames

Proof: Claim 1: The $w\ell^p$ bound implies that $\sigma(\Sigma_n(\phi), \mathcal{C}) = \mathcal{O}(n^{-s})$.

- For any signal $f \in \mathcal{C}$, picking the n largest frame coefficients defines an n-term approximation

$$f_n := \sum_{i \leq n} c_{k_i} \phi_{k_i} \,,$$

  were c_{k_i} is a non-increasing rearrangement of $c_k := \langle f, \tilde{\phi}_k \rangle_{\mathcal{H}}$.

- The definition of the $w\ell^p$ norm implies $|c_{k_i}| \lesssim i^{-1/p}$ because

$$|c_{k_i}|^p \, i \leq |c_{k_i}|^p \, \#\{k \in \mathbb{N} : |c_k| \geq |c_{k_i}|\} \leq \|c\|_{w\ell^p}^p \,.$$

- Together with the frame property of ϕ this yields

$$\|f - f_n\|^2 \lesssim \sum_{i>n} |c_{k_i}|^2 \lesssim \sum_{i>n} i^{-2/p} \leq n^{-2s}, \quad \text{where } s := \tfrac{1}{p} - \tfrac{1}{2} \,,$$

  where the last inequality follows from an elementary calculation. This proves Claim 1.

Proof: Approximation via Frames

Claim 2: The ℓ^2 tail bound implies $\sigma(\Sigma_n^\pi(\phi), \mathcal{C}) = \mathcal{O}(n^{-s})$ for suitable π.

- Define $\pi(n) := n^{\lceil 2s/\alpha \rceil}$.
- For any signal $f \in \mathcal{C}$, picking the first $\pi(n)$ frame coefficients defines an approximation $\tilde{f}_n$ with

$$\|f - \tilde{f}_n\|_{\mathcal{H}}^2 \lesssim \sum_{i > \pi(n)} \left|\langle f, \tilde{\phi}_i \rangle_{\mathcal{H}}\right|^2 \leq \big(\pi(n)\big)^{-\alpha} \leq n^{-2s}.$$

- By the previous claim, picking the n largest frame coefficients of $\tilde{f}_n$ defines an approximation f_n with

$$\|\tilde{f}_n - f_n\|_{\mathcal{H}}^2 \lesssim n^{-2s}.$$

- Taken together, this implies

$$\|f - f_n\|_{\mathcal{H}} \lesssim n^{-s},$$

which proves Claim 2 and establishes the theorem. □

Examples: Lower Bounds on Optimal Encoding Rates

Remark: The following lower bounds are sharp and are obtained as special cases of the previous theorem:

Corollary

The following lower bounds on encoding rates are achieved via frames:

- $s^*_{\mathrm{enc}}(\mathcal{C}^k_K(C)) \geq k/d$ *via wavelets, shearlets, and many more*
- $s^*_{\mathrm{enc}}(\mathcal{C}^{k,pw}_K(I)) \geq k$ *via wavelets*
- $s^*_{\mathrm{enc}}(\mathrm{STAR}^2_K) \geq 1$ *via curvelets and shearlets*
- $s^*_{\mathrm{enc}}(\mathrm{CART}^2_K) \geq 1$ *via curvelets and shearlets*
- $s^*_{\mathrm{enc}}(\mathrm{TEXT}^k_{K,M}) \geq k/2$ *via wave atoms*
- $s^*_{\mathrm{enc}}(\mathrm{MUTIL}^k_K) \geq k/d$ *via ridgelets*

Proof: Verify the conditions of the previous theorem for the specified frames; see [Dahlke e.a., Theorem 5.51]. □

Questions to Answer for Yourself / Discuss with Friends

- Repetition: How are dictionary approximation rates obtained from weak ℓ^p bounds on Hilbert frame coefficients?
- Background: Find the definition of wave atoms and have a look at some pictures of wave atoms. Hint: [Demanet and Ying (2007): Wave atoms and sparsity of oscillatory patterns]
- Discussion: Are the encoders/decoders obtained via frame approximations constructive and numerically implementable?
- Discussion: How could the theory be generalized to Banach frames, and what kind of results would you expect from this?

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 5

Networks as Encoders

Philipp Harms   Lars Niemann

University of Freiburg

Neural Network Approximation Rates

Remark: Neural networks with constrained memory can be seen as encoders.

Definition

Let $\mathcal{C}$ be a signal class in a normed function space $\mathcal{H}$ on $\mathbb{R}^d$, let $M \in \mathbb{N}$, let π be a univariate polynomial, and let A be a subset of $\mathbb{R}$.

- The set $\mathcal{NN}_M^A$ of neural networks with quantized weights is defined as the set of neural networks Φ with input dimension d, output dimension 1, and at most M non-zero weights belonging to A.
- The effective network approximation rate of $\mathcal{C}$ is defined as

$$s^*_{\mathcal{NN}}(\mathcal{C}) := \sup \Big\{ s > 0 \Big| \exists \pi, \exists (A_M)_{M \in \mathbb{N}} : \# A_M = \mathcal{O}(\pi(M)), \sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{NN}_M^{A_M M}} \|\mathrm{R}(\Phi) - f\|_{\mathcal{H}} = \mathcal{O}(M^{-s}) \Big\},$$

where R is defined using some fixed activation function $\rho \in C(\mathbb{R})$.

Encoding via Neural Networks

Remark: The memory constraint imposed via weight quantization yields the desired link between network approximation rates and encoding rates:

Theorem

For any signal class $\mathcal{C}$,

$$s^*_{\mathrm{enc}}(\mathcal{C}) \geq s^*_{\mathcal{N}\mathcal{N}}(\mathcal{C}).$$

Remark:

- Neural networks are called rate-optimal for $\mathcal{C}$ if equality holds above.
- The theorem implies a lower bound on the network connectivity, namely, an approximation error of ϵ requires approximately $\epsilon^{1/s^*_{\mathrm{enc}}(\mathcal{C})}$ non-zero network weights.

Proof: Encoding via Neural Networks

Proof:

- Let $s < s^*_{\mathcal{NN}}(\mathcal{C})$, and choose π, $(A_M)_{M\in\mathbb{N}}$, and C such that

$$\forall M \in \mathbb{N}: \quad \sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}_M^{A_M}} \|\mathrm{R}(\Phi) - f\|_{\mathcal{H}} < CM^{-s}, \quad \#A_M \le \pi(M)\,.$$

- Thus, for any given $f \in \mathcal{C}$ and $M \in \mathbb{N}$, there exists a network $\Phi \in \mathcal{NN}_M^{A_M}$ with $\|\mathrm{R}(\Phi) - f\|_{\mathcal{H}} < CM^{-s}$.
- We write $E \le M$ for the number of edges, $L \le M$ for the number of layers, $N_0 := d$ for the input dimension, $N_1, \dots, N_L$ for the numbers of neurons per layer, and $N := \sum_{\ell=0}^{L} N_\ell \le 2E$.
- We will show that Φ can be encoded in a bit string of length $\mathcal{O}(M \log M)$. This yields an encoder-decoder pair with distortion

$$\|D(E(F)) - f\| = \|\mathrm{R}(\Phi) - f\| = \mathcal{O}(M^{-s})$$

thereby establishing the theorem.

Proof: Encoding via Neural Networks (cont.)

- We encode the architecture of Φ in a bit string:
  - The number E of edges is encoded by a string of E 1's, followed by a single 0.
  - The number L of layers is encoded by a string of $\lceil \log_2 E \rceil$ bits, namely, by the binary representation of $L-1$ with left-padded zeros.
  - Then $(N_0, \ldots, N_L)$ is encoded in a string of $(L+1)\lceil \log_2 E + 1 \rceil$ bits.
- We encode the topology of Φ in a bit string:
  - To each neuron, we associate a unique index $i \in \{1, \ldots, N\}$, noting that this index can be encoded in a string b_i of $\lceil \log_2 E \rceil + 1$ bits.
  - For each neuron i, we output the concatenation of the bit strings b_j of all children j, followed by a zero string of length $2\lceil \log_2 E \rceil + 2$ to signal the transition to neuron $i+1$.
- We encode the weights of Φ in a bit string:
  - Each weight requires $\lceil \log_2 \pi(M) \rceil$ bits.
  - The nodal weights are encoded in $(N_1 + \cdots + N_l)\lceil \log_2 \pi(M) \rceil$ bits.
  - The edge weights are encoded in $E\lceil \log_2 \pi(M) \rceil$ bits.
- Overall, this requires $\mathcal{O}(M \log_2 M)$ bits, as claimed. □

Questions to Answer for Yourself / Discuss with Friends

- Repetition: What is the effective network approximation rate, and why is it upper-bounded by the encoding rate?
- Check: Why can the logarithmic factors in the rate computations be ignored?
- Check: In the last proof we constructed an encoder—what does the corresponding decoder look like?
- Discussion: What does the result say about deep learning? What are limitations of the result?

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 6

Dictionaries as Networks

Philipp Harms Lars Niemann

University of Freiburg

Representation of Dictionaries by Neural Networks

Setting: $\mathcal{H} = L^2(\Omega)$ for some $\Omega \subseteq \mathbb{R}^d$, and $\rho \colon \mathbb{R} \to \mathbb{R}$ is globally Lipschitz continuous or differentiable with polynomially bounded first derivative.

Definition

A dictionary $\phi = (\phi_i)_{i \in \mathbb{N}}$ in $\mathcal{H}$ is said to be effectively representable by neural networks if there exists $L, M \in \mathbb{N}$ and a bi-variate polynomial π such that for every $\epsilon \in (0, 1/2)$ and $i \in \mathbb{N}$ there exists a neural network Φ with $\mathrm{M}(\Phi) \leq M$, $\mathrm{L}(\Phi) \leq L$, and weights bounded by $\pi(i, \epsilon^{-1})$, such that

$$\|\phi_i - \mathrm{R}(\Phi)\|_{\mathcal{H}} \leq \epsilon.$$

Remark:

- The crucial point, also compared to our former setting for dictionary learning, is the requirement of polynomially bounded weights.
- For affine systems, i.e., dictionaries of affine transformations of a mother function ψ, it suffices to check effective representability of ψ.

Quantization of Neural Networks

Remark: We will need a seemingly stronger property, namely effective representation by quantized networks:

Lemma

In the definition of effective representability, it can be assumed without loss of generality that the weights of Φ are quantized in the sense that they belong to the set

$$\pi(i,\epsilon)\mathbb{Z} \cap [-\pi(i,\epsilon^{-1}), \pi(i,\epsilon^{-1})].$$

Proof: Quantization of Neural Networks

Sketch of proof for Lipschitz activation functions ρ:

- For single-layer networks $x \mapsto A_1 x + b_1$, which by definition are just affine maps, the quantization error of the network is proportional to the quantization error of the weights.
- For double-layer networks $x \mapsto A_2 \rho(A_1 x + b_1) + b_2$, the quantization error of the single-layer sub-network is amplified polynomially via the multiplication by A_2.
- By induction, the same holds for multi-layer networks.
- Thus, the quantization error of the network is $\mathcal{O}(\epsilon)$ if the quantization error of the weights is $\mathcal{O}(\epsilon^k)$ for sufficiently high k, with additional polynomial dependence on i.

For activation functions with polynomially bounded first derivative we refer to [Bölcskei e.a., Lemma 3.3]. □

Transfer of Approximation

Remark: Approximation rates for dictionaries transfer to approximation rates for neural networks if the dictionary is effectively represented by neural networks.

Theorem

If ϕ is effectively representable by neural networks and $\mathcal{C}$ is bounded, then

$$s^*_{\mathcal{NN}}(\mathcal{C}) \geq s^*_{\mathrm{dict}}(\mathcal{C}, \phi).$$

Proof: Transfer of Approximation

Proof: Dictionary learning.

- For any $s < s^*_{\text{dict}}(\mathcal{C}, \phi)$, there are approximations f_n of $f \in \mathcal{C}$ s.t.

$$f_n := D_n(E_n(f)) := \sum_{i=1}^{\pi(n)} c_i \phi_i, \qquad \|f_n - f\|_{\mathcal{H}} = \mathcal{O}(n^{-s}).$$

- In the theorem on encoding via dictionaries in Video 3 we have shown that the coefficients c_i can be chosen in a set of cardinality polynomially bounded in n.
- The dictionary functions ϕ_i, $i \in \{1, \ldots, \pi(n)\}$, can be effectively represented by neural networks Φ_i, up to an approximation error of order $\mathcal{O}(n^{-s})$, with weights polynomially bounded in n.
- By the quantization lemma, it can be assumed without loss of generality that the weights of the networks Φ_i belong to a set of cardinality polynomially bounded in n.
- Taking linear combinations produces a network approximation of f_n with weights in a set of cardinality polynomially bounded in n and approximation error $\mathcal{O}(n^{-s})$. □

Rate-Optimal Approximation by Neural Networks

Corollary

If ϕ is a rate-optimal dictionary for $\mathcal{C}$, and ϕ is effectively represented by neural networks, then neural networks are rate-optimal for $\mathcal{C}$.

Proof: The following rates are equal,

$$s^*_{\text{dict}}(\mathcal{C}, \phi) \overset{①}{=} s^*_{\text{enc}}(\mathcal{C}) \overset{②}{\geq} s^*_{\mathcal{NN}}(\mathcal{C}) \overset{③}{\geq} s^*_{\text{dict}}(\mathcal{C}, \phi),$$

because

1. the dictionary ϕ is rate-optimal,
2. quantized neural networks are encoders, as shown in Video 5, and
3. quantized dictionary approximations are quantized neural networks, as shown in the last theorem. □

Remark: This corollary applies to all examples of signal classes and dictionaries discussed so far.

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Why and under what conditions is the effective network approximation rate lower-bounded by the dictionary approximation rate?
- Check: How wide and deep are the approximating networks?
- Check: How does the present transfer-of-approximation result differ from the one of Week 3?
- Discussion: What does the result say about deep learning? What are limitations of the result?

Mathematics of Deep Learning, Summer Term 2020

Week 7, Video 7

Wrapup

Philipp Harms Lars Niemann

University of Freiburg

Outlook on this week's discussion and reading session

- Reading:
  - Bölcskei, Grohs, Kutyniok, Petersen (2017): Optimal approximation with sparsely connected deep neural networks
  - Donoho (2001): Sparse Components of Images and Optimal Atomic Decompositions. In: Constructive Approximation 17, pp. 353–382
  - Shannon (1959): Coding Theorems for a Discrete Source with a Fidelity Criterion. In: International Convention Record 7, pp. 325–350

Summary by learning goals

Having heard this lecture, you can now . . .

- Derive lower bounds on effective network approximation rates from harmonic analysis.
- Derive upper bounds on effective network approximation rates from rate-distortion theory.
- Explain why neural networks are optimal descriptors of a wide variety of signal classes.