Feature Geometry and Applications in Deep Learning

Lizhong Zheng

IAS Workshop, Hong Kong, July, 2023

Shashank Jere, Lingjia Liu (VT)

Our Interest/Belief on Deep Learning

- Applications of DNN in engineering problems are different from NLP/Image Processing
  - Limited training;
  - Domain knowledge and structures, do not re-learn what is known;
  - Guarantees;
  - Parameterized optimal solutions;
  - Targetted performance enhancement (performance comparison table is often not the right way.)

The Role of Information Theory

The Role of Information Theory

- Information theoretic quantities, entropy, mutual information, K-L divergence, etc., are pleasant concepts, and therefore used in many learning problems as a part of the loss function.

The Role of Information Theory

- Information theoretic quantities, entropy, mutual information, K-L divergence, etc., are pleasant concepts, and therefore used in many learning problems as a part of the loss function.
- The operational meanings for information-theoretic quantities: the coding theorems, "max rate with $P_e \to 0$ as $n \to \infty$".

The Role of Information Theory

- Information theoretic quantities, entropy, mutual information, K-L divergence, etc., are pleasant concepts, and therefore used in many learning problems as a part of the loss function.
- The operational meanings for information-theoretic quantities: the coding theorems, "max rate with $P_e \to 0$ as $n \to \infty$".
- The current operational meaning of IT quantities in ML: when used in the loss function, the performance is sometimes better.

Need Some Re-Thinking

Need Some Re-Thinking

- The carrier of information: bits $\longrightarrow$ real-valued features $f(x), \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Need Some Re-Thinking

- The carrier of information: bits $\longrightarrow$ real-valued features $f(x), \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.
- Without coding, every step of the processing strictly loses information.

Need Some Re-Thinking

- The carrier of information: bits $\longrightarrow$ real-valued features $f(x)$, $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$.
- Without coding, every step of the processing strictly loses information.
- Have to talk about semantics, which Shannon didn't.

Need Some Re-Thinking

- The carrier of information: bits $\longrightarrow$ real-valued features $f(x)$, $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$.
- Without coding, every step of the processing strictly loses information.
- Have to talk about semantics, which Shannon didn't.
- Quantify the meaning of a feature: what binary question does it answer?

Need Some Re-Thinking

- The carrier of information: bits $\longrightarrow$ real-valued features $f(x), \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.
- Without coding, every step of the processing strictly loses information.
- Have to talk about semantics, which Shannon didn't.
- Quantify the meaning of a feature: what binary question does it answer?
- Naturally a geometric concept.

The Plan for This Short Talk

- Our attempt for new information-metric for real-valued features $f : \mathcal{X} \to \mathbb{R}$,

The Plan for This Short Talk

- Our attempt for new information-metric for real-valued features $f : \mathcal{X} \to \mathbb{R}$,
- A Hilbert space $\langle f_1, f_2 \rangle = \mathbb{E}[f_1(\mathsf{X}) \cdot f_2(\mathsf{X})]$ and the norm, subspace, angle, and projection based on these.

The Plan for This Short Talk

- Our attempt for new information-metric for real-valued features $f : \mathcal{X} \to \mathbb{R}$,
- A Hilbert space $\langle f_1, f_2 \rangle = \mathbb{E}[f_1(X) \cdot f_2(X)]$ and the norm, subspace, angle, and projection based on these.
- The operational meanings: all learning algorithms try to learn optimal things with induced metrics.

The Plan for This Short Talk

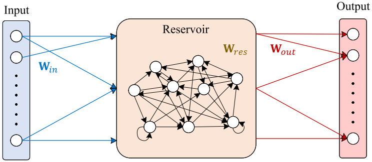
- Our attempt for new information-metric for real-valued features $f : \mathcal{X} \to \mathbb{R}$,
- A Hilbert space $\langle f_1, f_2 \rangle = \mathbb{E}[f_1(\mathrm{X}) \cdot f_2(\mathrm{X})]$ and the norm, subspace, angle, and projection based on these.
- The operational meanings: all learning algorithms try to learn optimal things with induced metrics.
- Examples of what we can do with this geometric language.

The Plan for This Short Talk

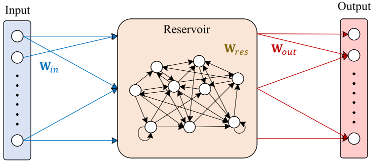
- Our attempt for new information-metric for real-valued features $f : \mathcal{X} \to \mathbb{R}$,
- A Hilbert space $\langle f_1, f_2 \rangle = \mathbb{E}[f_1(\mathsf{X}) \cdot f_2(\mathsf{X})]$ and the norm, subspace, angle, and projection based on these.
- The operational meanings: all learning algorithms try to learn optimal things with induced metrics.
- Examples of what we can do with this geometric language.

A few steps we need to change our thinking

Example: The Story of Reservoir Computing

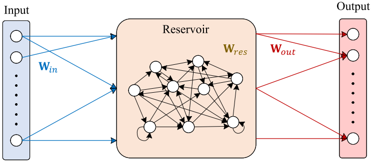


Example: The Story of Reservoir Computing



- In general, we need to generate a rich set of time sequences driven by the inputs.

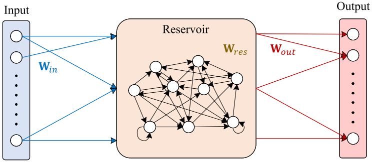
Example: The Story of Reservoir Computing



- In general, we need to generate a rich set of time sequences driven by the inputs.
- Typical implementation: like a state space model

$$q[n+1] = \sigma(w_{\text{res}} \cdot q[n] + W_{\text{in}} \cdot x[n])$$
$$y[n] = W_{\text{out}} \cdot q[n]$$

Example: The Story of Reservoir Computing



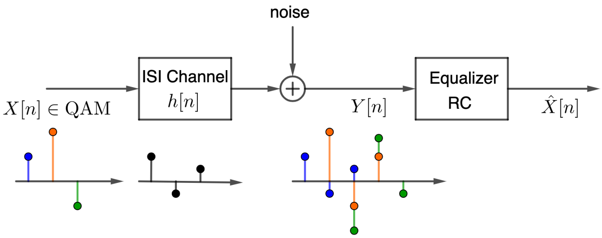
- In general, we need to generate a rich set of time sequences driven by the inputs.
- Typical implementation: like a state space model

$$q[n+1] = \sigma(w_{\mathrm{res}} \cdot q[n] + W_{\mathrm{in}} \cdot x[n])$$
$$y[n] = W_{\mathrm{out}} \cdot q[n]$$

- Only train the input/output weights.

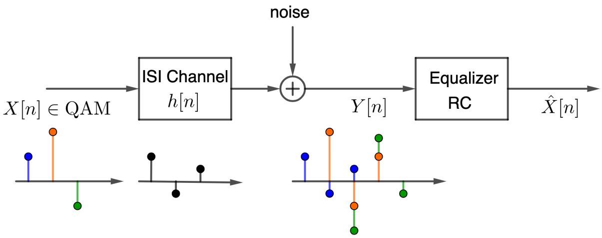
The Equalization Problem

- The symbol detection in ISI channel.



The Equalization Problem

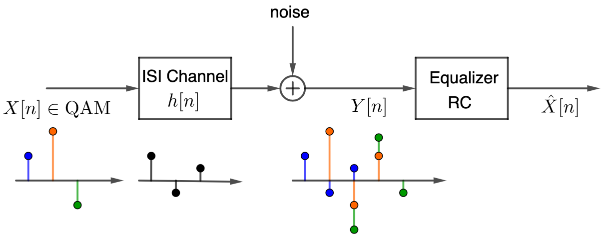
- The symbol detection in ISI channel.



- Why is this difficult for conventional solutions?

The Equalization Problem

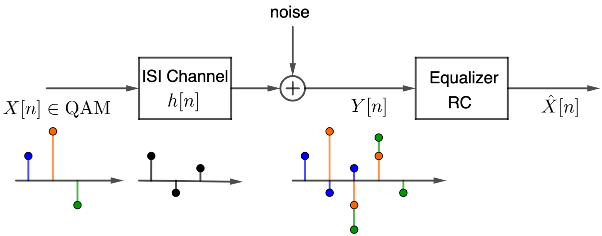
- The symbol detection in ISI channel.



- Why is this difficult for conventional solutions?
  - Without additive noise, reduces to deconvolution

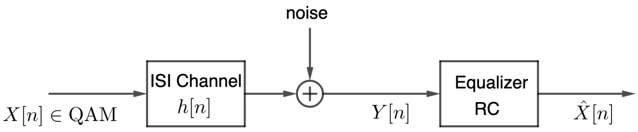
The Equalization Problem

- The symbol detection in ISI channel.



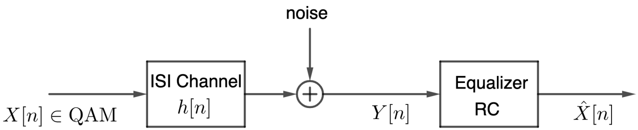
- Why is this difficult for conventional solutions?
  - Without additive noise, reduces to deconvolution
  - If the interference were Gaussian, L2 estimation is optimal

(Good) Blackboxes Work, Sort of.



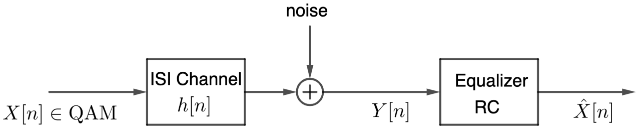
- Train a network, with $Y[n]$ as input and try to predict $X[n]$ (sorry for the convention).

(Good) Blackboxes Work, Sort of.



- Train a network, with $Y[n]$ as input and try to predict $X[n]$ (sorry for the convention).
- Reservoir computing works quite well.

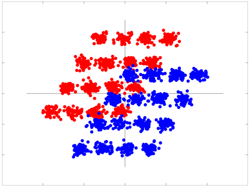
(Good) Blackboxes Work, Sort of.



- Train a network, with $Y[n]$ as input and try to predict $X[n]$ (sorry for the convention).
- Reservoir computing works quite well.
- There is an issue of error floor, performance gap to the optimal at high SNR: the deconvolution didn't work too well.

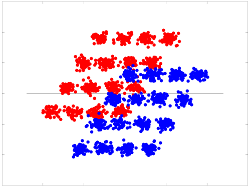
Moving towards Understanding

- Performance metrics might be misleading, both learning performance metrics and communication metrics.
  - Weak interference can be handled with classical approaches.
  - Strong interference occurs rarely.
  - Switching is not hard for engineers.



Moving towards Understanding

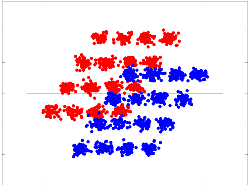
- Performance metrics might be misleading, both learning performance metrics and communication metrics.
  - Weak interference can be handled with classical approaches.
  - Strong interference occurs rarely.
  - Switching is not hard for engineers.



- Using the learning-based method: can we resolve the interference?

Moving towards Understanding

- Performance metrics might be misleading, both learning performance metrics and communication metrics.
  - Weak interference can be handled with classical approaches.
  - Strong interference occurs rarely.
  - Switching is not hard for engineers.



- Using the learning-based method: can we resolve the interference?
- Wish list: training costs, use of structure, prior statistical knowledge, change with parameters, optimality, ...

Hansel and Gretel's Bread Crumbs

Hansel and Gretel's Bread Crumbs

- Try deconvolution (switch to conventional methods when needed.) Minmax vs. Average

$$\min \ \mathbb{E}_h\left[\left\|\delta[\cdot] - h * \widehat{h_{\mathrm{res}}^{-1}}\right\|^2\right] \iff \min \mathbb{E}_h\left[\left\|h * (h^{-1} - \widehat{h_{\mathrm{res}}^{-1}})\right\|^2\right]$$

Hansel and Gretel's Bread Crumbs

- Try deconvolution (switch to conventional methods when needed.) Minmax vs. Average

$$\min\ \mathbb{E}_h\left[\left\|\delta[\cdot]-h*\widehat{h^{-1}_{\mathrm{res}}}\right\|^2\right]\iff\min\mathbb{E}_h\left[\left\|h*(h^{-1}-\widehat{h^{-1}_{\mathrm{res}}})\right\|^2\right]$$

- L2 distance

$$\min\mathbb{E}_h\left[\left\|h^{-1}-\widehat{h^{-1}_{\mathrm{res}}}\right\|^2\right]$$

Hansel and Gretel's Bread Crumbs

- Try deconvolution (switch to conventional methods when needed.) Minmax vs. Average

$$\min\ \mathbb{E}_h \left[\left\| \delta[\cdot] - h * \widehat{h_{\text{res}}^{-1}} \right\|^2 \right] \iff \min \mathbb{E}_h \left[\left\| h * (h^{-1} - \widehat{h_{\text{res}}^{-1}}) \right\|^2 \right]$$

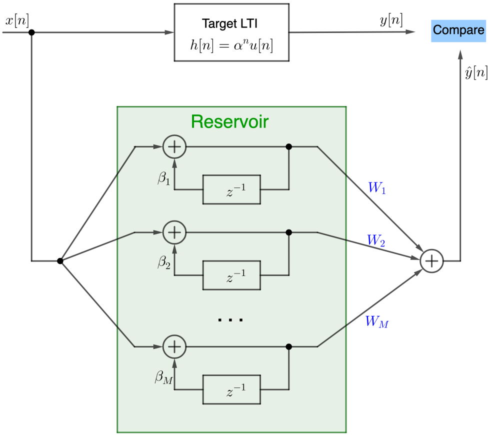
- L2 distance

$$\min \mathbb{E}_h \left[\left\| h^{-1} - \widehat{h_{\text{res}}^{-1}} \right\|^2 \right]$$

- Inverse z-transform by partial fraction expansion

$$\min \mathbb{E}_\alpha \left[\left\| \frac{1}{1 - \alpha z^{-1}} - \widehat{h_{\text{res}}^{-1}} \right\|^2 \right]$$

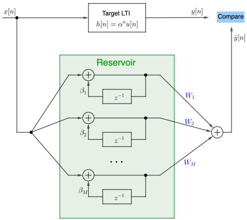
A Problem We Can Do



Simplest reservoir: no connection, no non-linear.

Specific Functional Approximation

Need to choose $\beta_1, \ldots, \beta_M$, a random choice of the target α with a given prior p_α,



$$\min_{\beta_1^M} \mathbb{E}_{p_\alpha}\left[\left\|\left(\frac{1}{1-\alpha z^{-1}}\right) - \sum_{i=1}^{M} w_i^* \cdot \left(\frac{1}{1-\beta_i z^{-1}}\right)\right\|^2\right]$$

What We Can See From the Solution?

What We Can See From the Solution?

- Why is reservoir a good idea? If $\beta_i = \alpha$, we have no error.

What We Can See From the Solution?

- Why is reservoir a good idea? If $\beta_i = \alpha$, we have no error.
- The scaling law of performance (training loss), how does $\|\mathrm{error}\|^2$ decrease with M: $O(M^{-4})$.

What We Can See From the Solution?

- Why is reservoir a good idea? If $\beta_i = \alpha$, we have no error.
- The scaling law of performance (training loss), how does $\|\mathrm{error}\|^2$ decrease with M: $O(M^{-4})$.
- Optimal reservoir size with limited training: standard generalization error analysis.

What We Can See From the Solution?

- Why is reservoir a good idea? If $\beta_i = \alpha$, we have no error.
- The scaling law of performance (training loss), how does $\|\mathrm{error}\|^2$ decrease with M: $O(M^{-4})$.
- Optimal reservoir size with limited training: standard generalization error analysis.
- How to place β's? Randomly with density $\propto \frac{1}{(1-\beta^2)^{\frac{5}{4}}}$ if p_α is uniform.

What We Can See From the Solution?

- Why is reservoir a good idea? If $\beta_i = \alpha$, we have no error.
- The scaling law of performance (training loss), how does $\|\mathrm{error}\|^2$ decrease with M: $O(M^{-4})$.
- Optimal reservoir size with limited training: standard generalization error analysis.
- How to place β's? Randomly with density $\propto \frac{1}{(1-\beta^2)^{\frac{5}{4}}}$ if p_α is uniform.
- When we know p_α(3GPP/LTE), easily fold in the prior knowledge.

With all these, the error floor is pushed down.

What is Hidden in the Black Box Solution?

What is Hidden in the Black Box Solution?

- The value of connecting the neurons?

$$q[n+1] = \sigma(W_{\mathrm{res}} \cdot q[n] + W_{\mathrm{in}} \cdot x[n])$$
$$y[n] = W_{\mathrm{out}} \cdot q[n]$$

What is Hidden in the Black Box Solution?

- The value of connecting the neurons?

$$q[n+1] = \sigma(W_{\text{res}} \cdot q[n] + W_{\text{in}} \cdot x[n])$$
$$y[n] = W_{\text{out}} \cdot q[n]$$

- The issue of L2 loss: channel inversion to all-pass filter.

What is Hidden in the Black Box Solution?

- The value of connecting the neurons?

$$q[n+1] = \sigma(W_{\mathrm{res}} \cdot q[n] + W_{\mathrm{in}} \cdot x[n])$$
$$y[n] = W_{\mathrm{out}} \cdot q[n]$$

- The issue of L2 loss: channel inversion to all-pass filter.
- The value of having an activation function?

What is Hidden in the Black Box Solution?

- The value of connecting the neurons?

$$q[n+1] = \sigma(W_{\mathrm{res}} \cdot q[n] + W_{\mathrm{in}} \cdot x[n])$$
$$y[n] = W_{\mathrm{out}} \cdot q[n]$$

- The issue of L2 loss: channel inversion to all-pass filter.
- The value of having an activation function?
- A parameterized optimal solution: the topic of a different talk.

Concluding Remarks

- Apply ML to engineering problems, maybe I have a narrow view here.
- Side information, structure of the problem, constraints: separate what we want to learn and what we don't.
- Do Not always want a more complex design.
- Either performance metric does not tell the full story.
- Using non-linear units.