Complexity of Linear Regions in Deep Nets

Boris Hanin

Facebook AI Research and Texas A&M

March 5, 2019

Joint with **David Rolnick**

Theoretical vs. Practical Expressivity

Theoretical vs. Practical Expressivity

- **Brain:** Why deep nets, Pinky?

Theoretical vs. Practical Expressivity

- **Brain:** Why deep nets, Pinky?

- **Pinky:** Expressivity, Brain!

Theoretical vs. Practical Expressivity

- **Brain:** Why deep nets, Pinky?
- **Pinky:** Expressivity, Brain!
- **Brain:** What about learnability?

Numerical Instability for Large Numbers of Regions

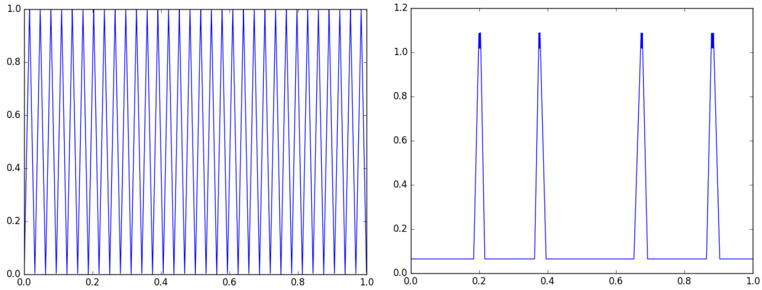
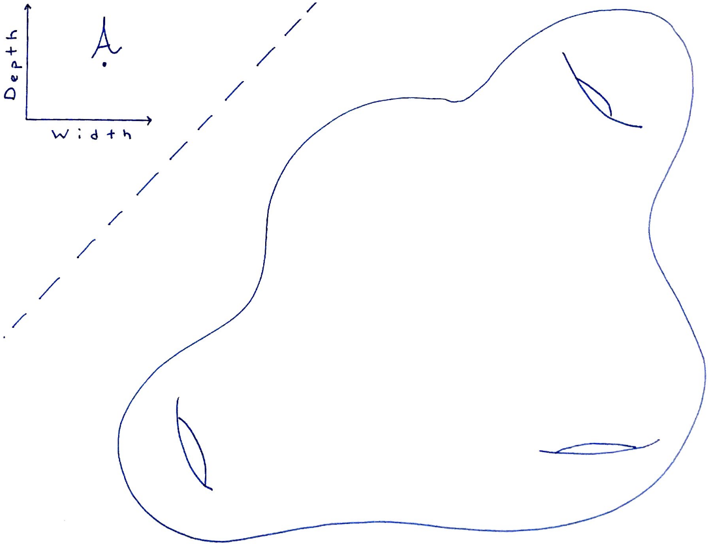


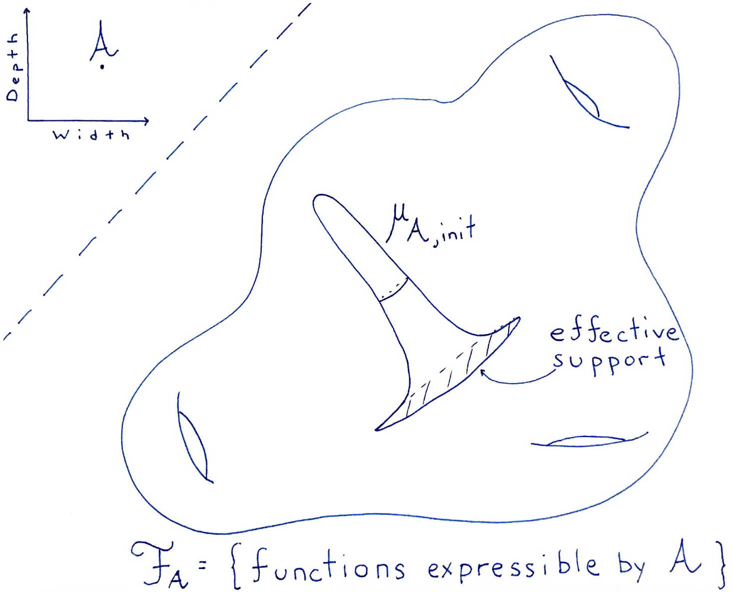
Figure: Random perturbation of example w/maximal number of regions.

Theoretical Expressivity

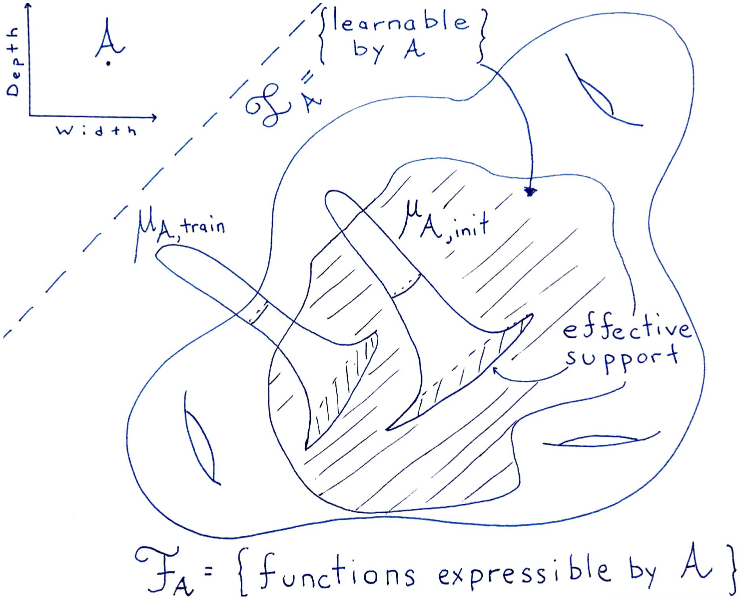


$\mathcal{F}_A$ = { functions expressible by A }

Practical Expressivity at Init



Practical Expressivity



How To Do Theory?

How To Do Theory?

- **Goal.** Characterize typical complexity of functions drawn from $\mu_{\mathcal{A},\text{init}}$, $\mu_{\mathcal{A},\text{train}}$.

How To Do Theory?

- **Goal.** Characterize typical complexity of functions drawn from $\mu_{\mathcal{A},\text{init}}$, $\mu_{\mathcal{A},\text{train}}$.

- **Intution.** Probabilty measures in high dimensions are often concentrated around low dimensional sets.

How To Do Theory?

- **Goal.** Characterize typical complexity of functions drawn from $\mu_{\mathcal{A},\text{init}}$, $\mu_{\mathcal{A},\text{train}}$.

- **Intution.** Probabilty measures in high dimensions are often concentrated around low dimensional sets.

- **Idea.** For networks with piecewise linear activations, complexity of $\mu_{\mathcal{A},\text{init}}$ and $\mu_{\mathcal{A},\text{train}}$ encoded in corresponding partition of input space.

Overview

Overview

- $\mathcal{N}$ – depth d ReLU net with $n_{out} = 1$

Overview

- $\mathcal{N}$ – depth d ReLU net with $n_{out} = 1$

- $x \mapsto \mathcal{N}(x)$ is continuous and piecewise linear function

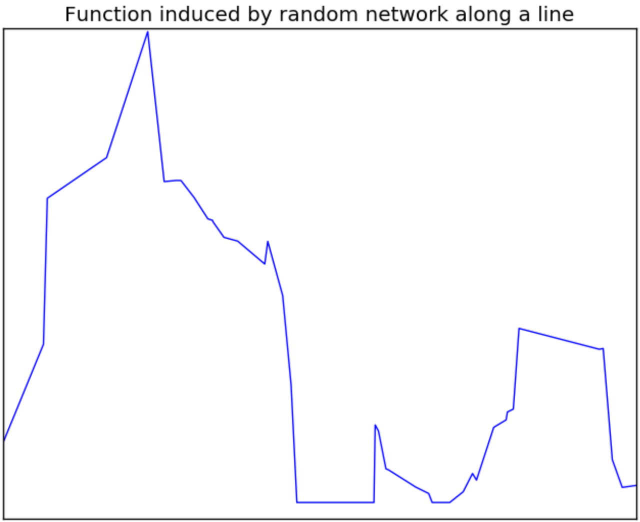
Overview

- $\mathcal{N}$ – depth d ReLU net with $n_{out} = 1$
- $x \mapsto \mathcal{N}(x)$ is continuous and piecewise linear function
- Fixed weights/biases partition $\mathbb{R}^{n_{in}}$ into convex pieces on which $\mathcal{N}$ is linear

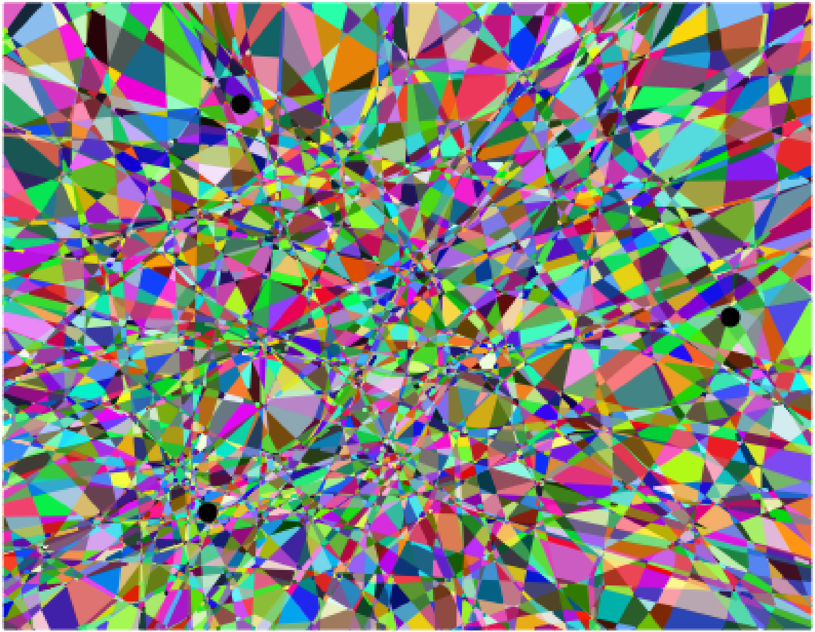
Overview

- $\mathcal{N}$ – depth d ReLU net with $n_{out} = 1$
- $x \mapsto \mathcal{N}(x)$ is continuous and piecewise linear function
- Fixed weights/biases partition $\mathbb{R}^{n_{in}}$ into convex pieces on which $\mathcal{N}$ is linear
- **Goal.** Understand average complexity of this partition

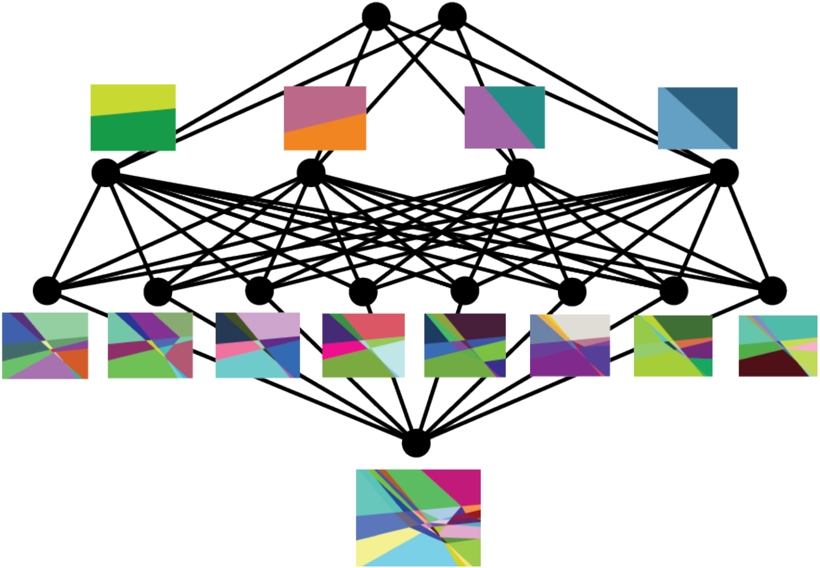
ReLU Net with $n_{in} = n_{out} = 1$ at Initialization



Input Space Partition with $n_{in} = 2$ at Initialization



Evolution of Input Partition Through Network



Complexity v1.0: Number of Regions

Complexity v1.0: Number of Regions

- **Deterministic Bounds**: $1 \leq \#\text{regions} \leq 2^{\#\text{neurons}}$

Complexity v1.0: Number of Regions

- **Deterministic Bounds**: $1 \leq \#\text{regions} \leq 2^{\#\text{neurons}}$

- **Moral of Prior Work.** There exist very special weight/bias settings for deep skinny nets that saturate upper bound.

Complexity v1.0: Number of Regions

- **Deterministic Bounds**: $1 \le \#\text{regions} \le 2^{\#\text{neurons}}$

- **Moral of Prior Work.** There exist very special weight/bias settings for deep skinny nets that saturate upper bound.

- **Q1.** What is the average number of regions at init?

Complexity v1.0: Number of Regions

- **Deterministic Bounds**: $1 \leq \#\text{regions} \leq 2^{\#\text{neurons}}$

- **Moral of Prior Work.** There exist very special weight/bias settings for deep skinny nets that saturate upper bound.

- **Q1.** What is the average number of regions at init?

- **Q2.** What happens to regions during training (practical vs. theoretical expressivity)?

Number of Regions when $n_{in} = n_{out} = 1$

Number of Regions when $n_{in} = n_{out} = 1$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\mathrm{Var}[\text{weights}] \;=\; 2/\text{fan-in}, \qquad \mathrm{Var}[\text{bias}] \;=\; \sigma_b^2 > 0.$$

Number of Regions when $n_{in} = n_{out} = 1$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\text{Var}[\text{weights}] = 2/\text{fan-in}, \qquad \text{Var}[\text{bias}] = \sigma_b^2 > 0.$$

For any compact $S \subset \mathbb{R}$ there are $c = c(\sigma_b)$, $C = C(\sigma_b)$ so that

$$c\,\#\left\{\text{neurons}\right\} \leq \frac{1}{|S|}\mathbb{E}\left[\#\left\{\text{regions in } S\right\}\right] \leq C\,\#\left\{\text{neurons}\right\}$$

Number of Regions when $n_{in} = n_{out} = 1$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\mathrm{Var}[\text{weights}] \;=\; 2/\text{fan-in}, \qquad \mathrm{Var}[\text{bias}] \;=\; \sigma_b^2 > 0.$$

For any compact $S \subset \mathbb{R}$ *there are* $c = c(\sigma_b)$, $C = C(\sigma_b)$ *so that*

$$c\,\#\,\{\text{neurons}\} \;\leq\; \frac{1}{|S|}\mathbb{E}\left[\#\,\{\text{regions in } S\}\right] \;\leq\; C\,\#\,\{\text{neurons}\}$$

Remark

1. *Comes from formula that holds throughout training*

Number of Regions when $n_{in} = n_{out} = 1$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\mathrm{Var}[\text{weights}] = 2/\text{fan-in}, \qquad \mathrm{Var}[\text{bias}] = \sigma_b^2 > 0.$$

For any compact $S \subset \mathbb{R}$ there are $c = c(\sigma_b)$, $C = C(\sigma_b)$ so that

$$c \, \# \{\text{neurons}\} \leq \frac{1}{|S|} \mathbb{E}\left[\# \{\text{regions in } S\}\right] \leq C \, \# \{\text{neurons}\}$$

Remark

1. *Comes from formula that holds throughout training*
2. *Holds for any network connectivity*

Number of Regions when $n_{in} = n_{out} = 1$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\text{Var}[\text{weights}] = 2/\text{fan-in}, \qquad \text{Var}[\text{bias}] = \sigma_b^2 > 0.$$

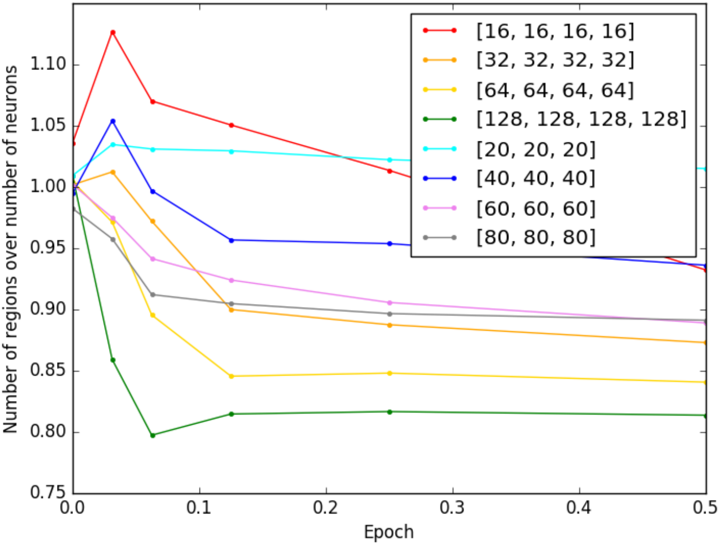
For any compact $S \subset \mathbb{R}$ there are $c = c(\sigma_b)$, $C = C(\sigma_b)$ so that

$$c\,\#\{\text{neurons}\} \le \frac{1}{|S|}\mathbb{E}\left[\#\{\text{regions in } S\}\right] \le C\,\#\{\text{neurons}\}$$

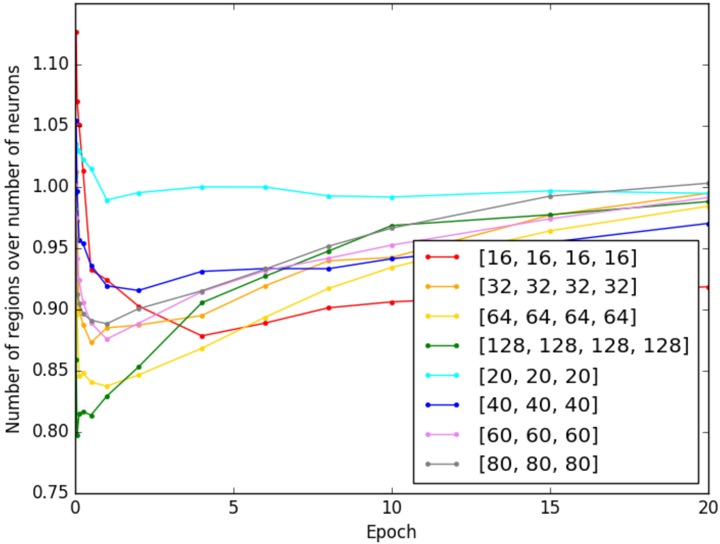
Remark

1. *Comes from formula that holds throughout training*
2. *Holds for any network connectivity*
3. *Holds for any 1D curve inside high dimensional input space*

Number of Regions on 1D Line Through Training



Number of Regions on 1D Line Through Training



Maximal # Regions on 2D Plane

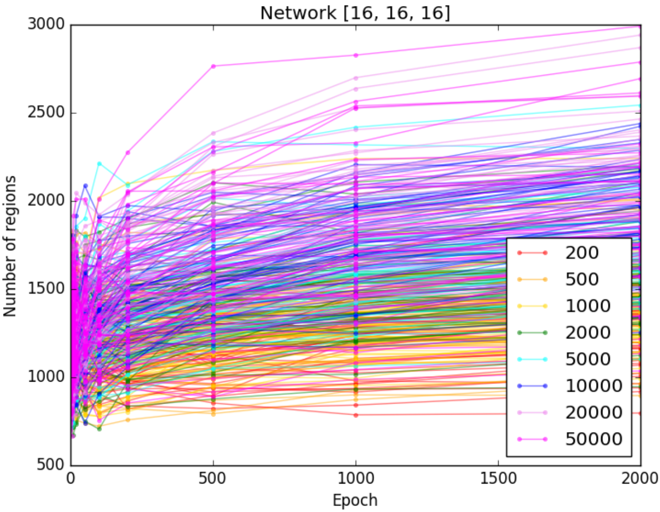


Figure: Heuristic: $\#\{\text{regions on k dim slice}\} \sim (\#\text{neurons})^k$. When $k = 2$, should have $\approx (16 * 3)^2 = 2304$ regions.

Maximal # Regions on 2D Plane

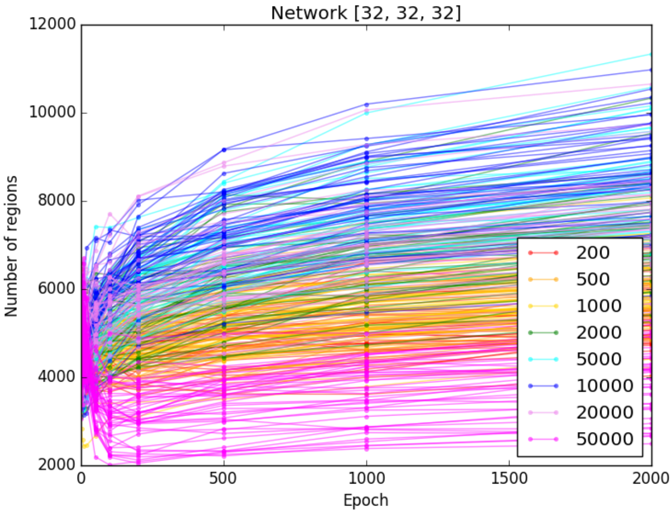


Figure: Heuristic: $\#\{\text{regions on k dim slice}\} \sim (\#\text{neurons})^k$. When $k = 2$, should have $\approx (32 * 3)^2 = 9216$ regions.

Maximal # Regions on 2D Plane

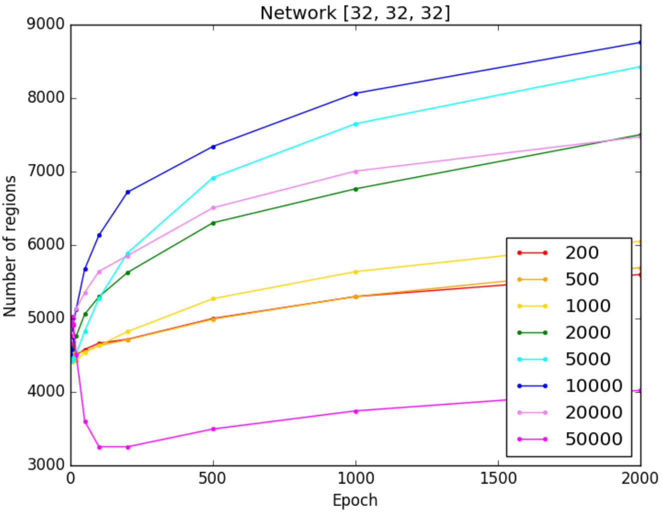
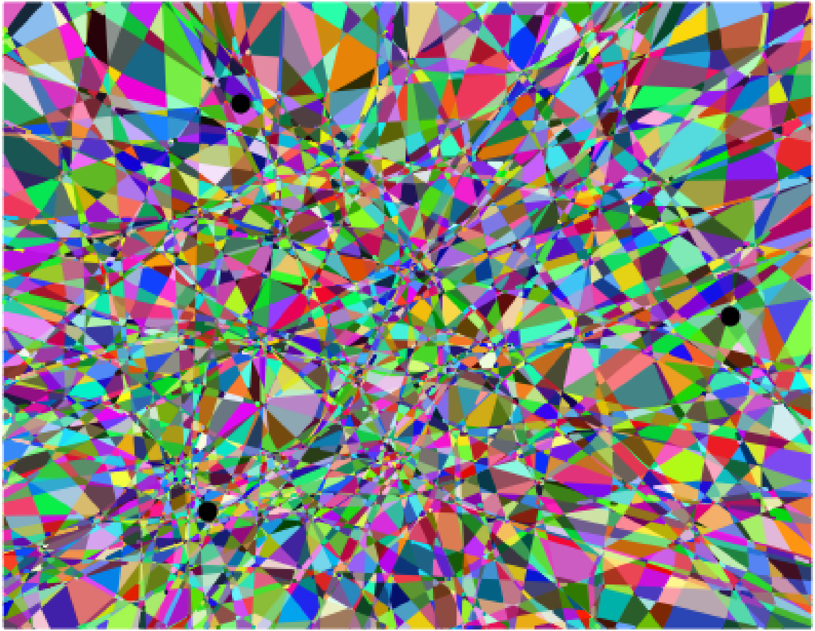


Figure: Heuristic: $\#\{\text{regions on k dim slice}\} \sim (\#\text{neurons})^k$. When $k = 2$, should have $\approx (32 * 3)^2 = 9216$ regions.

Complexity v2.0: Volume of Linear Region Boundaries



Complexity v2.0: Volume of Linear Region Boundaries

Complexity v2.0: Volume of Linear Region Boundaries

- **Basic Object of Study:**

$$\mathcal{B}_{\mathcal{N}} := \{\text{Linear region boundaries of } \mathcal{N}\}.$$

Complexity v2.0: Volume of Linear Region Boundaries

- **Basic Object of Study:**

$$\mathcal{B}_{\mathcal{N}} := \{\text{Linear region boundaries of } \mathcal{N}\}.$$

- $\underline{n_{in} = 1}$: $\quad \text{vol}(\mathcal{B}_{\mathcal{N}}) + 1 = \#\text{regions}$

Complexity v2.0: Volume of Linear Region Boundaries

- **Basic Object of Study:**

$$\mathcal{B}_{\mathcal{N}} := \{\text{Linear region boundaries of } \mathcal{N}\}.$$

- $\underline{n_{in} = 1}$: $\quad \text{vol}(\mathcal{B}_{\mathcal{N}}) + 1 = \#\text{regions}$

- $\underline{n_{in} > 1}$: $\quad \#\{\text{regions inside } S\} \neq \text{vol}(\mathcal{B}_{\mathcal{N}} \cap S)$

Complexity v2.0: Volume of Linear Region Boundaries

- **Basic Object of Study:**

$$\mathcal{B}_{\mathcal{N}} := \{\text{Linear region boundaries of } \mathcal{N}\}.$$

- $\underline{n_{in} = 1}$: $\quad \text{vol}(\mathcal{B}_{\mathcal{N}}) + 1 = \#\text{regions}$

- $\underline{n_{in} > 1}$: $\quad \#\{\text{regions inside } S\} \neq \text{vol}(\mathcal{B}_{\mathcal{N}} \cap S)$

- **Motivation 1.** $\text{vol}(\mathcal{B}_{\mathcal{N}})$ controls avg dist to boundary:

$$\mathbb{P}\left(\text{dist}(x, \mathcal{B}_{\mathcal{N}}) \leq \epsilon\right) \simeq \epsilon \, \text{vol}(\mathcal{B}_{\mathcal{N}} \cap S), \qquad x \sim \text{Unif}(S).$$

Complexity v2.0: Volume of Linear Region Boundaries

- **Basic Object of Study:**

$$\mathcal{B}_{\mathcal{N}} := \{\text{Linear region boundaries of } \mathcal{N}\}.$$

- $\underline{n_{in} = 1}$: $\quad \mathrm{vol}(\mathcal{B}_{\mathcal{N}}) + 1 = \#\mathrm{regions}$

- $\underline{n_{in} > 1}$: $\quad \#\{\text{regions inside } S\} \neq \mathrm{vol}(\mathcal{B}_{\mathcal{N}} \cap S)$

- **Motivation 1.** $\mathrm{vol}(\mathcal{B}_{\mathcal{N}})$ controls avg dist to boundary:

$$\mathbb{P}\left(\mathrm{dist}(x, \mathcal{B}_{\mathcal{N}}) \leq \epsilon\right) \simeq \epsilon\, \mathrm{vol}(\mathcal{B}_{\mathcal{N}} \cap S), \qquad x \sim \mathrm{Unif}(S).$$

- **Motivation 2.**: $\mathrm{vol}(\mathcal{B}_{\mathcal{N}})$ controls correlation length:

$$\text{corr. length of } \mathcal{N} \overset{?}{\approx} \mathrm{dist}(x, \mathcal{B}_{\mathcal{N}})$$

Volume of $\mathcal{B}_{\mathcal{N}}$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\mathsf{Var}[\text{weights}] \;=\; 2/\text{fan-in}, \qquad \mathsf{Var}[\text{bias}] \;=\; \sigma_b^2 > 0.$$

Volume of $\mathcal{B}_{\mathcal{N}}$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\mathrm{Var}[\text{weights}] \;=\; 2/\text{fan-in}, \qquad \mathrm{Var}[\text{bias}] \;=\; \sigma_b^2 > 0.$$

For compact $S \subset \mathbb{R}^{n_{in}}$ *there are* $c = c(\sigma_b)$, $C = C(\sigma_b)$ *so that*

$$c\,\#\{\text{neurons}\} \;\leq\; \frac{1}{\mathrm{vol}\,(S)}\,\mathbb{E}\left[\mathrm{vol}(\,\mathcal{B}_{\mathcal{N}}\;\cap\;S)\right] \;\leq\; C\,\#\{\text{neurons}\}$$

Volume of $\mathcal{B}_{\mathcal{N}}$

Theorem (H-Rolnick)

Suppose weights and biases are independent with

$$\text{Var}[\text{weights}] = 2/\text{fan-in}, \qquad \text{Var}[\text{bias}] = \sigma_b^2 > 0.$$

For compact $S \subset \mathbb{R}^{n_{in}}$ there are $c = c(\sigma_b),\ C = C(\sigma_b)$ so that

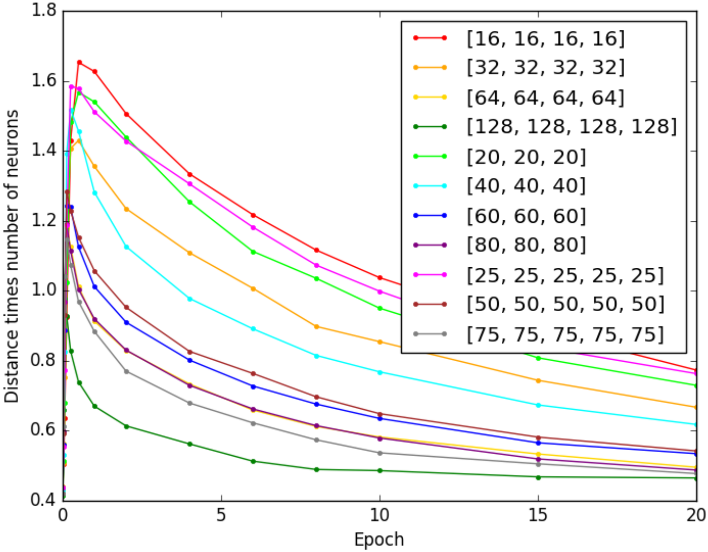
$$c\,\#\{\text{neurons}\} \leq \frac{1}{\text{vol}\,(S)}\,\mathbb{E}\Big[\text{vol}(\,\mathcal{B}_{\mathcal{N}} \cap S)\Big] \leq C\,\#\{\text{neurons}\}$$

Corollary

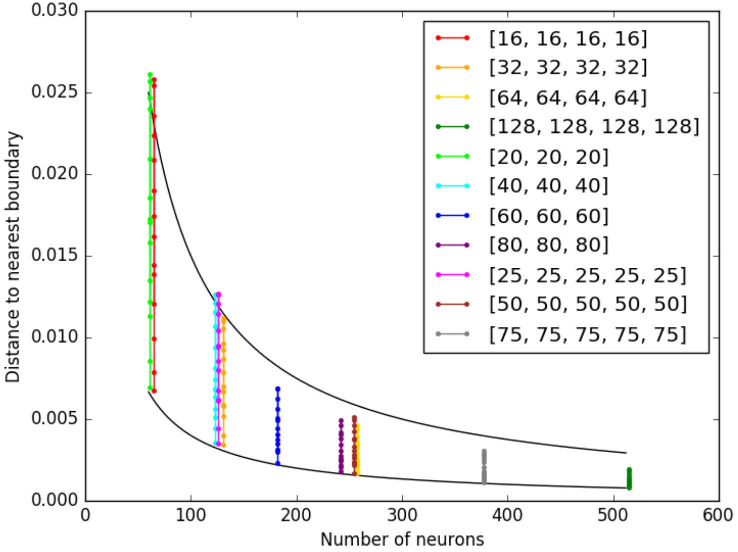
Let $x \in S = [0,1]^{n_{in}}$ be uniform. There exists $c = c(\sigma_b)$ so that

$$\mathbb{E}\left[\text{dist}(x, \mathcal{B}_{\mathcal{N}})\right] \geq \frac{c}{\#\{\text{neurons}\}}$$

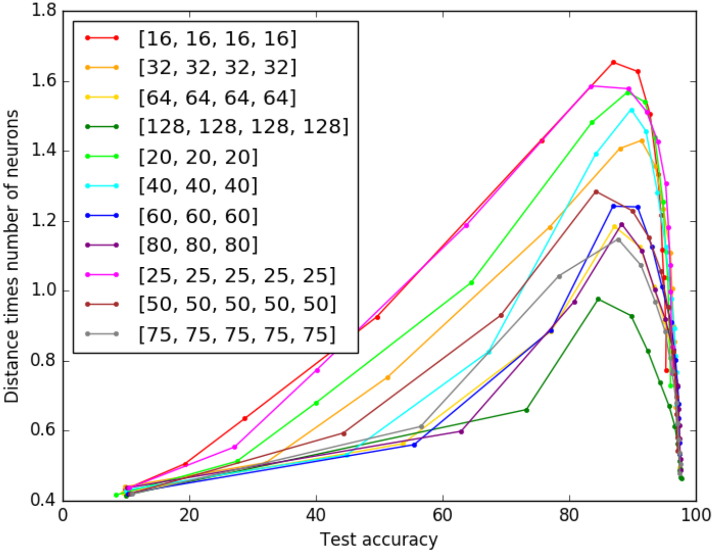
Distance to $\mathcal{B}_{\mathcal{N}}$ vs. Number of Neurons



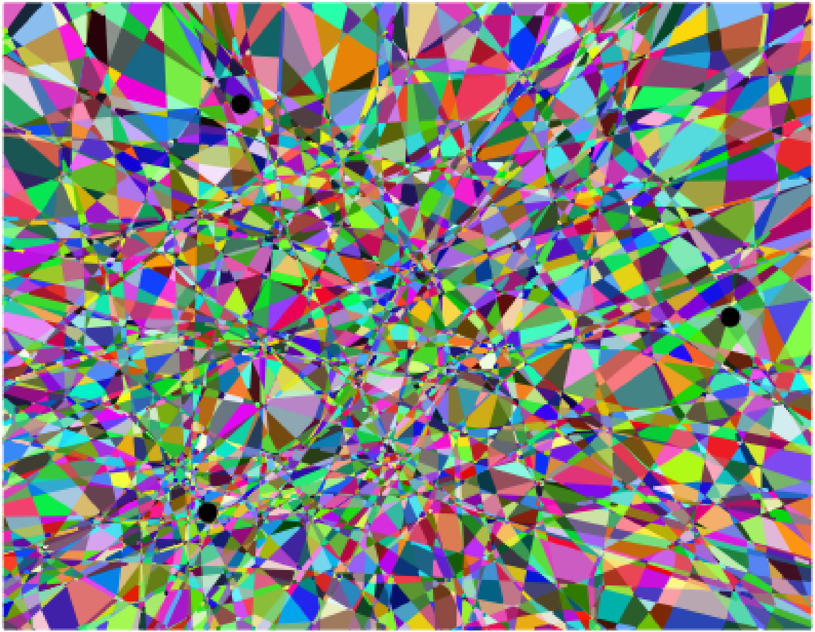
Distance to $\mathcal{B}_{\mathcal{N}}$ vs. Number of Neurons



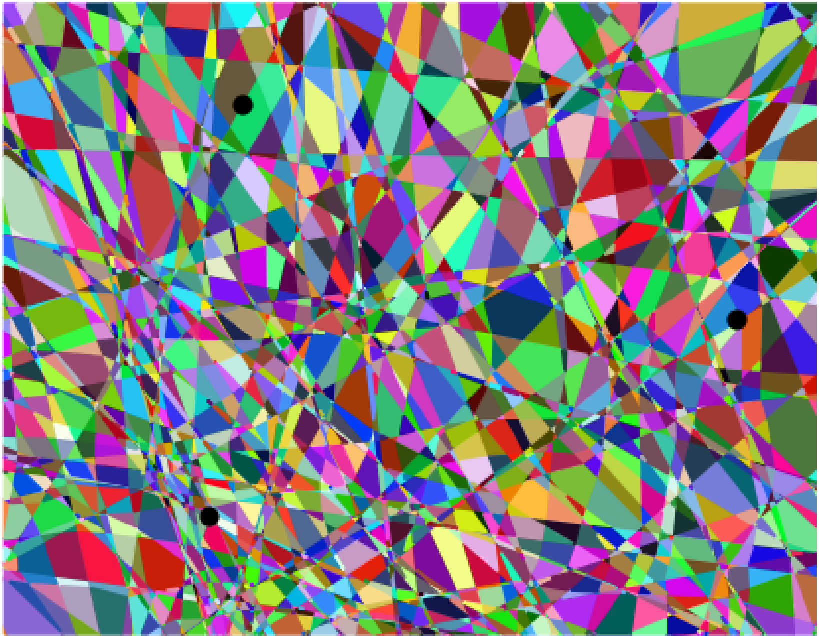
Distance to $\mathcal{B}_{\mathcal{N}}$ vs. Test Accuracy



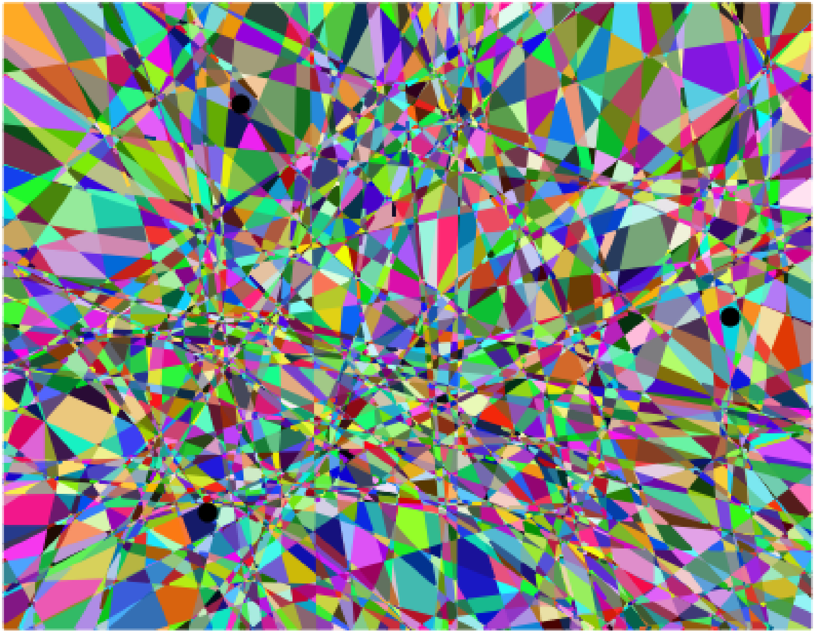
Input Space Partition with $n_{in} = 2$ at Initialization



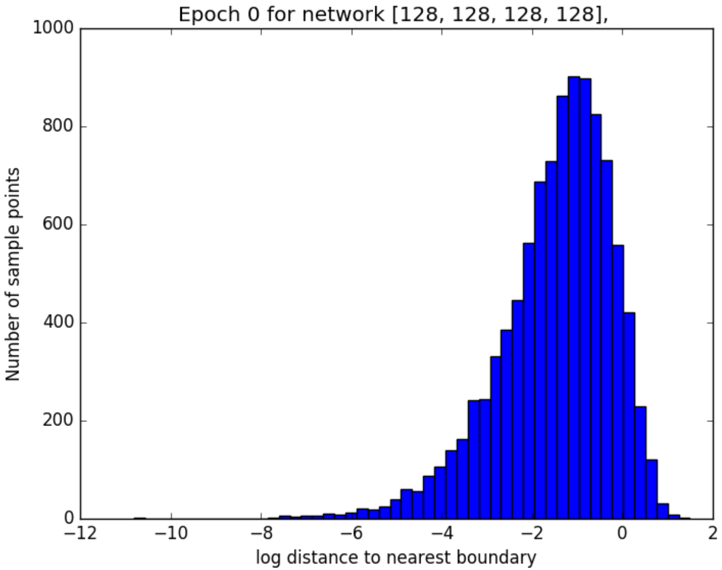
Input Space Partition with $n_{in} = 2$ after 1 Epoch



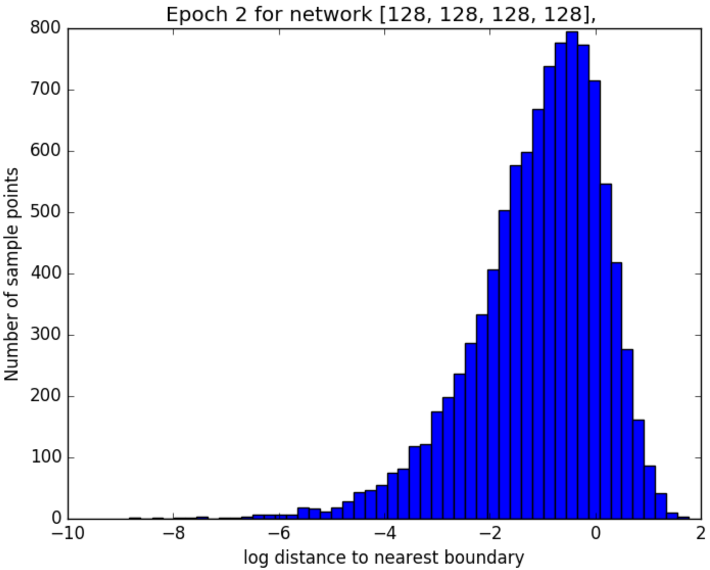
Input Space Partition with $n_{in} = 2$ after Training



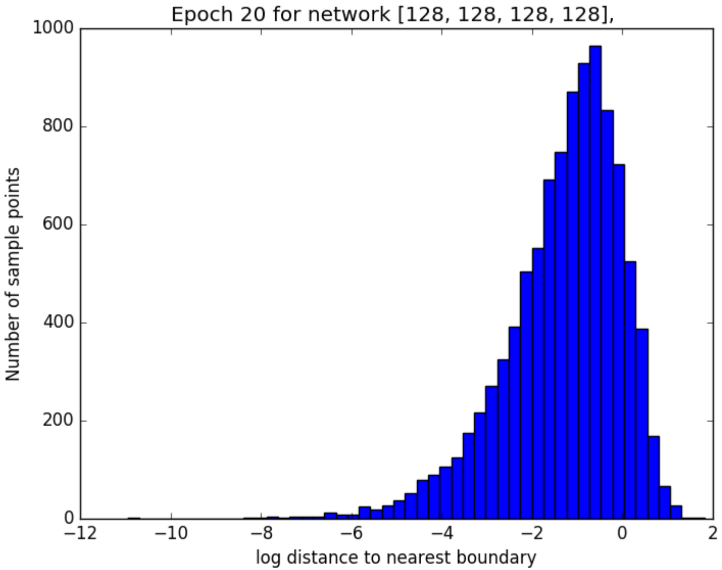
Distribution of Distance to Linear Region Boundary



Distribution of Distance to Linear Region Boundary



Distribution of Distance to Linear Region Boundary



Main Technical Theorem (for ReLU Nets)

Main Technical Theorem (for ReLU Nets)

Theorem (H-Rolnick)

Let $\mathcal{N}$ be a ReLU net with $n_{out} = 1$ and random weights/biases, so that bias b_z at neuron z has density ρ_{b_z}.

Main Technical Theorem (for ReLU Nets)

Theorem (H-Rolnick)

Let $\mathcal{N}$ be a ReLU net with $n_{out} = 1$ and random weights/biases, so that bias b_z at neuron z has density ρ_{b_z}. Then, for $S \subset \mathbb{R}^{n_{in}}$,

$$\mathbb{E}\left[\mathrm{vol}\left(\mathcal{B}_{\mathcal{N}} \cap S\right)\right]$$

$$= \sum_{\text{neurons } z} \int_S \mathbb{E}\left[\|\nabla z(x)\| \;\; \rho_{b_z}(z(x)) \;\; \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x) \neq 0\right\}} \right] dx,$$

Main Technical Theorem (for ReLU Nets)

Theorem (H-Rolnick)

Let $\mathcal{N}$ be a ReLU net with $n_{out} = 1$ and random weights/biases, so that bias b_z at neuron z has density ρ_{b_z}. Then, for $S \subset \mathbb{R}^{n_{in}}$,

$$\mathbb{E}\left[\mathrm{vol}\left(\mathcal{B}_{\mathcal{N}} \cap S\right)\right] = \sum_{\text{neurons } z} \int_S \mathbb{E}\left[\|\nabla z(x)\| \;\; \rho_{b_z}(z(x)) \;\; \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial Z}(x) \neq 0\right\}} \right] dx,$$

where $z(x)$ is the pre-activation for neuron z and

$$Z(x) = \max\left\{b_z, z(x)\right\} = \text{post-activation}.$$

Main Technical Theorem (for ReLU Nets)

Theorem (H-Rolnick)

Let $\mathcal{N}$ be a ReLU net with $n_{out} = 1$ and random weights/biases, so that bias b_z at neuron z has density ρ_{b_z}. Then, for $S \subset \mathbb{R}^{n_{in}}$,

$$\mathbb{E}\left[\mathrm{vol}\left(\mathcal{B}_{\mathcal{N}} \cap S\right)\right] = \sum_{\text{neurons } z} \int_S \mathbb{E}\left[\|\nabla z(x)\| \ \rho_{b_z}(z(x)) \ \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial Z}(x) \neq 0\right\}}\right] dx,$$

where $z(x)$ is the pre-activation for neuron z and

$$Z(x) = \max\left\{b_z, z(x)\right\} = \text{post-activation}.$$

Remark

1. *Analogous to Kac-Rice formula but easier because b_z random*

Main Technical Theorem (for ReLU Nets)

Theorem (H-Rolnick)

Let $\mathcal{N}$ be a ReLU net with $n_{out} = 1$ and random weights/biases, so that bias b_z at neuron z has density ρ_{b_z}. Then, for $S \subset \mathbb{R}^{n_{in}}$,

$$\mathbb{E}\left[\mathrm{vol}\left(\mathcal{B}_{\mathcal{N}} \cap S\right)\right] = \sum_{\text{neurons } z} \int_S \mathbb{E}\left[\|\nabla z(x)\| \;\; \rho_{b_z}(z(x)) \;\; \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial Z}(x) \neq 0\right\}}\right] dx,$$

where $z(x)$ is the pre-activation for neuron z and

$$Z(x) = \max\left\{b_z, z(x)\right\} = \text{post-activation}.$$

Remark

1. *Analogous to Kac-Rice formula but easier because b_z random*
2. *Holds throughout training as weights/biases can be correlated*

Main Technical Theorem (for ReLU Nets)

Theorem (H-Rolnick)

Let $\mathcal{N}$ be a ReLU net with $n_{out} = 1$ and random weights/biases, so that bias b_z at neuron z has density ρ_{b_z}. Then, for $S \subset \mathbb{R}^{n_{in}}$,

$$\mathbb{E}\left[\mathrm{vol}\left(\mathcal{B}_{\mathcal{N}} \cap S\right)\right] = \sum_{\text{neurons } z} \int_S \mathbb{E}\left[\|\nabla z(x)\| \ \rho_{b_z}(z(x)) \ \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial Z}(x) \neq 0\right\}} \right] dx,$$

where $z(x)$ is the pre-activation for neuron z and

$$Z(x) = \max\left\{b_z, z(x)\right\} = \text{post-activation}.$$

Remark

1. *Analogous to Kac-Rice formula but easier because b_z random*
2. *Holds throughout training as weights/biases can be correlated*
3. *Holds for any connectivity*

Interpretation and Intuition

Interpretation and Intuition

- For fixed $x \in S$, each term in

$$\mathbb{E}\left[\ \|\nabla z(x)\|\ \ \rho_{b_z}(z(x))\ \ \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x)\neq 0\right\}}\right]dx$$

has interpretation

- For fixed $x \in S$, each term in

$$\mathbb{E}\left[\;\|\nabla z(x)\|\;\;\rho_{b_z}(z(x))\;\;\mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x)\neq 0\right\}}\right]dx$$

has interpretation:

  - $\|\nabla z(x)\|\; dx$ $\quad - \quad$ size of dx under $x \mapsto z(x)$

Interpretation and Intuition

- For fixed $x \in S$, each term in

$$\mathbb{E}\left[\ \|\nabla z(x)\| \ \ \rho_{b_z}(z(x)) \ \ \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x) \neq 0\right\}}\right] dx$$

  has interpretation:

  - $\|\nabla z(x)\| \ dx$ — size of dx under $x \mapsto z(x)$
  - $\rho_{b_z}(z(x)) \ \|\nabla z(x)\| \ dx$ — $\mathbb{P}(b_z$ creates kink at $[x \pm dx])$

Interpretation and Intuition

- For fixed $x \in S$, each term in

$$\mathbb{E}\left[\|\nabla z(x)\| \;\; \rho_{b_z}(z(x)) \;\; \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x) \neq 0\right\}} \right] dx$$

has interpretation:

  - $\|\nabla z(x)\| \; dx$ — size of dx under $x \mapsto z(x)$
  - $\rho_{b_z}(z(x)) \; \|\nabla z(x)\| \; dx$ — $\mathbb{P}(b_z$ creates kink at $[x \pm dx])$
  - $\mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x) \neq 0\right\}}$ — event that kink at x survives to output

Interpretation and Intuition

- For fixed $x \in S$, each term in

$$\mathbb{E}\left[\|\nabla z(x)\| \ \ \rho_{b_z}(z(x)) \ \ \mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x) \neq 0\right\}} \right] dx$$

  has interpretation:

  - $\|\nabla z(x)\| \ dx$ — size of dx under $x \mapsto z(x)$
  - $\rho_{b_z}(z(x)) \ \|\nabla z(x)\| \ dx$ — $\mathbb{P}(b_z$ creates kink at $[x \pm dx])$
  - $\mathbf{1}_{\left\{\frac{\partial \mathcal{N}}{\partial z}(x) \neq 0\right\}}$ — event that kink at x survives to output

- **Intuition.** If $\|\nabla z(x)\| = O(1)$ and b_z is not too concentrated, then $z(x) = b_z$ can only be solved in $O(1)$ regions.