Interaction Effects: Helpful or Harmful?

Ben Lengerich

CMU AI Seminar

Feb 18, 2020

Carnegie Mellon University Computer Science Department

Microsoft Research

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Today

1. What is an Interaction Effect?

2. Interaction Effects in Neural Networks

Based on:

- Purifying Interaction Effects with the Functional ANOVA. AISTATS 2020
  - Lengerich, Tan, Chang, Hooker, Caruana

- On Dropout, Overfitting, and Interaction Effects in Deep Neural Networks. Under Review 2020.

  - Lengerich, Xing, Caruana

Why do we care about interaction effects?

- Interpreting models
  - Identifiability
- Understanding how big machine learning models work

What is an Interaction Effect?

Intuitively:

“Effect of one variable changes based on the value of another variable”

But this definition is incomplete: 3 stories

Is “AND” an Interaction Effect?

Suppose we data:

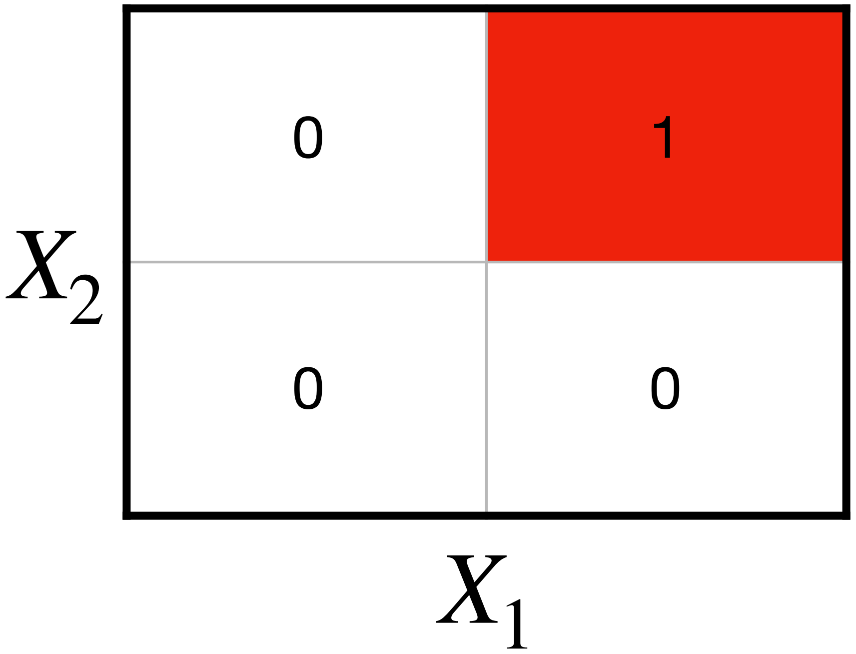
$Y = \text{AND}(X_1, X_2)$

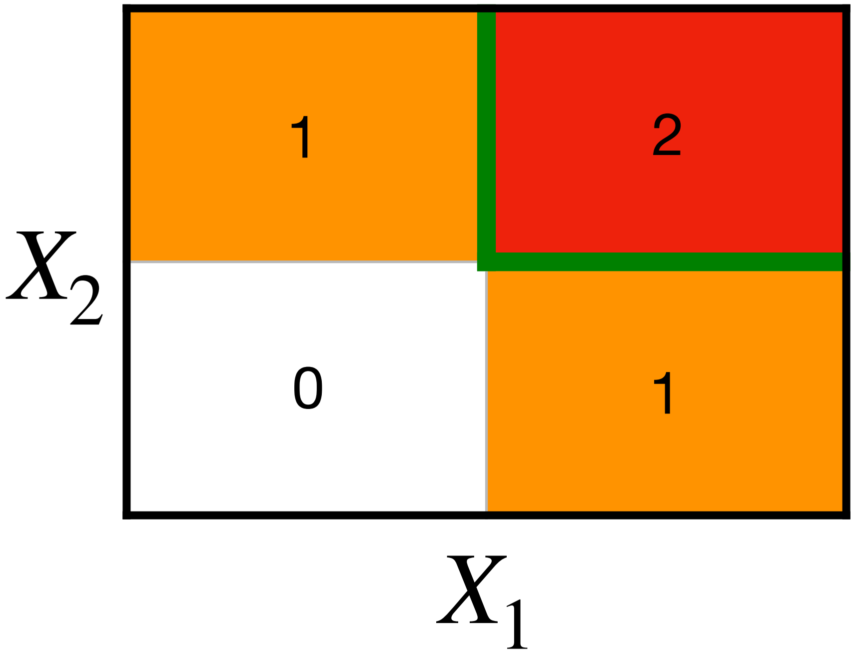
with Boolean X_1, X_2. Let’s fit an additive model (no interactions):

$Y = f_0 + f_1(X_1) + f_2(X_2)$

How well can we fit the data?

Perfectly*!





Is Multiplication an Interaction?

Common model:

$$Y = a + bX_1 + cX_2 + dX_1X_2$$

But this is equivalent to:

$$Y = (a - d\alpha\beta) + (b + d\beta)X_1 + (c + d\alpha)X_2 + d(X_1 - \alpha)(X_2 - \beta)$$

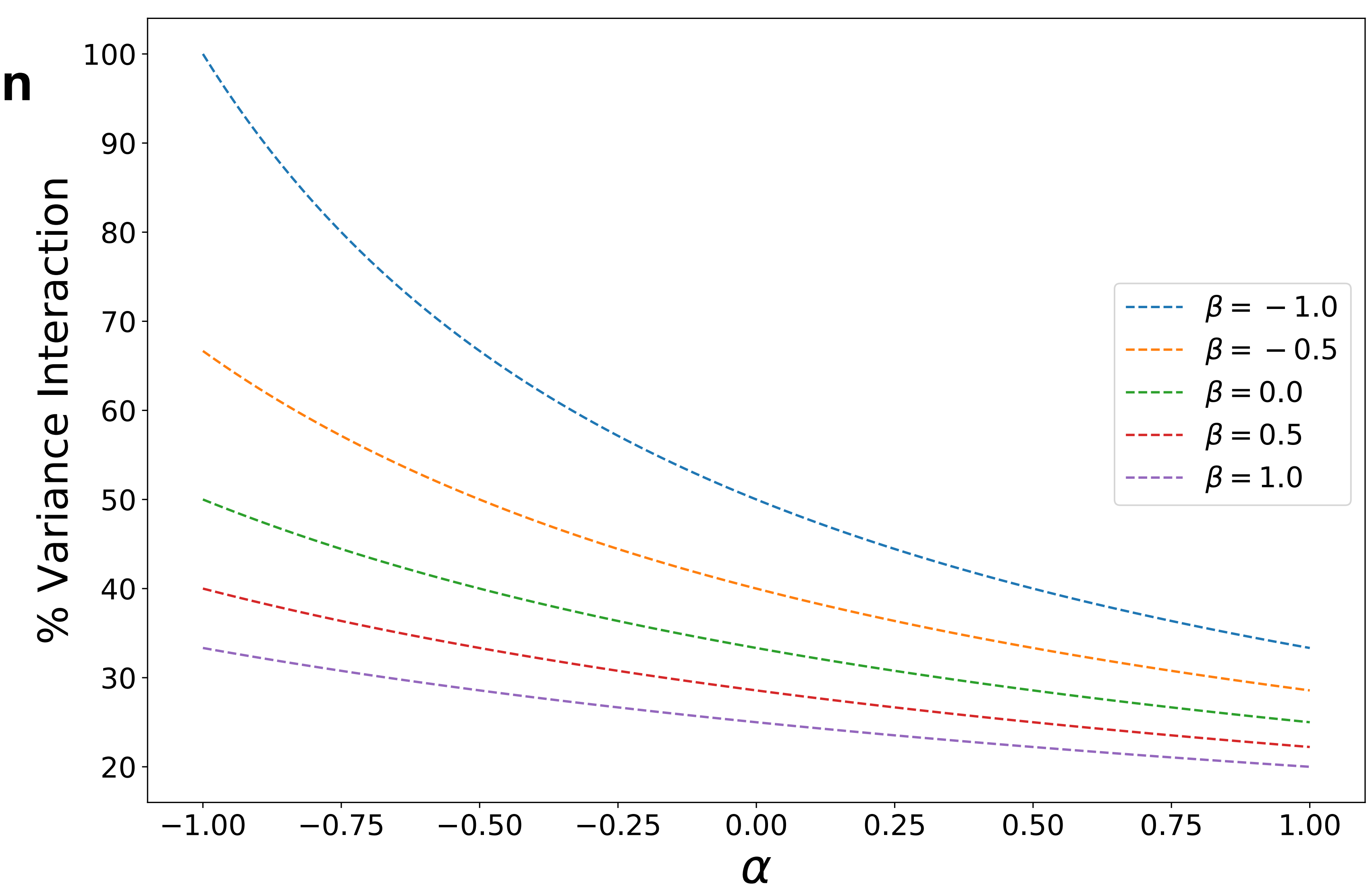
We can pick any offsets α, β without changing the function output.
Picking different values of α, β drastically changes the interpretation.

Is Multiplication an Interaction?

$$Y = (a - d\alpha\beta) + (b + d\beta)X_1 + (c + d\alpha)X_2 + d(X_1 - a)(X_2 - b)$$

Picking different values of α, β drastically changes the interpretation:

100% interaction effect



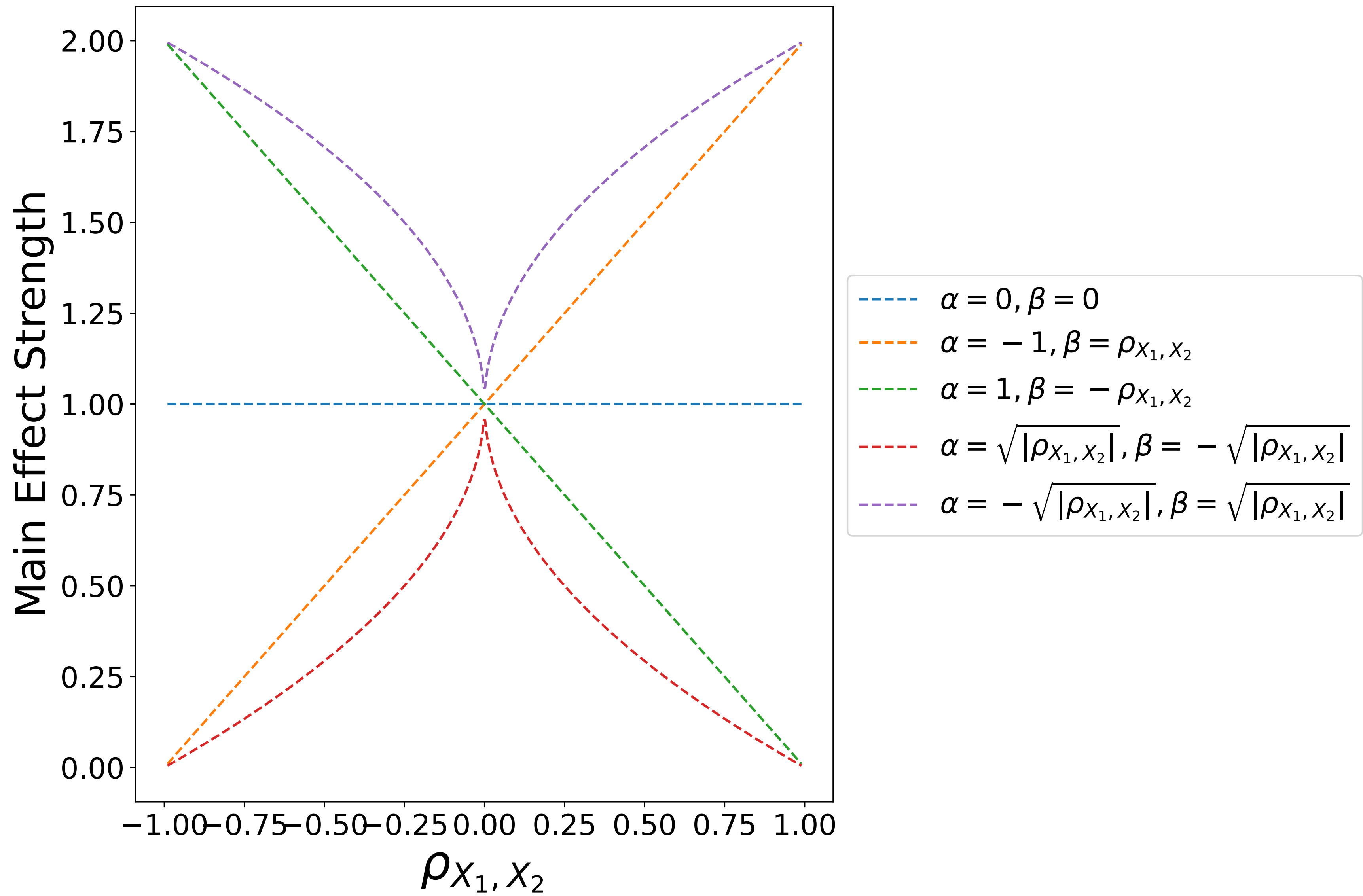
20% interaction effect

Is Multiplication an Interaction? Mean-Center?

- Does mean-centering solve this problem?

- No — If the correlation $\rho(X_1, X_2)$ is not zero, then we can't simultaneously center X_1, X_2, X_1X_2.

- Choosing which term to center changes the interpretation!



Is Multiplication an Interaction? One more wrinkle

If we say that

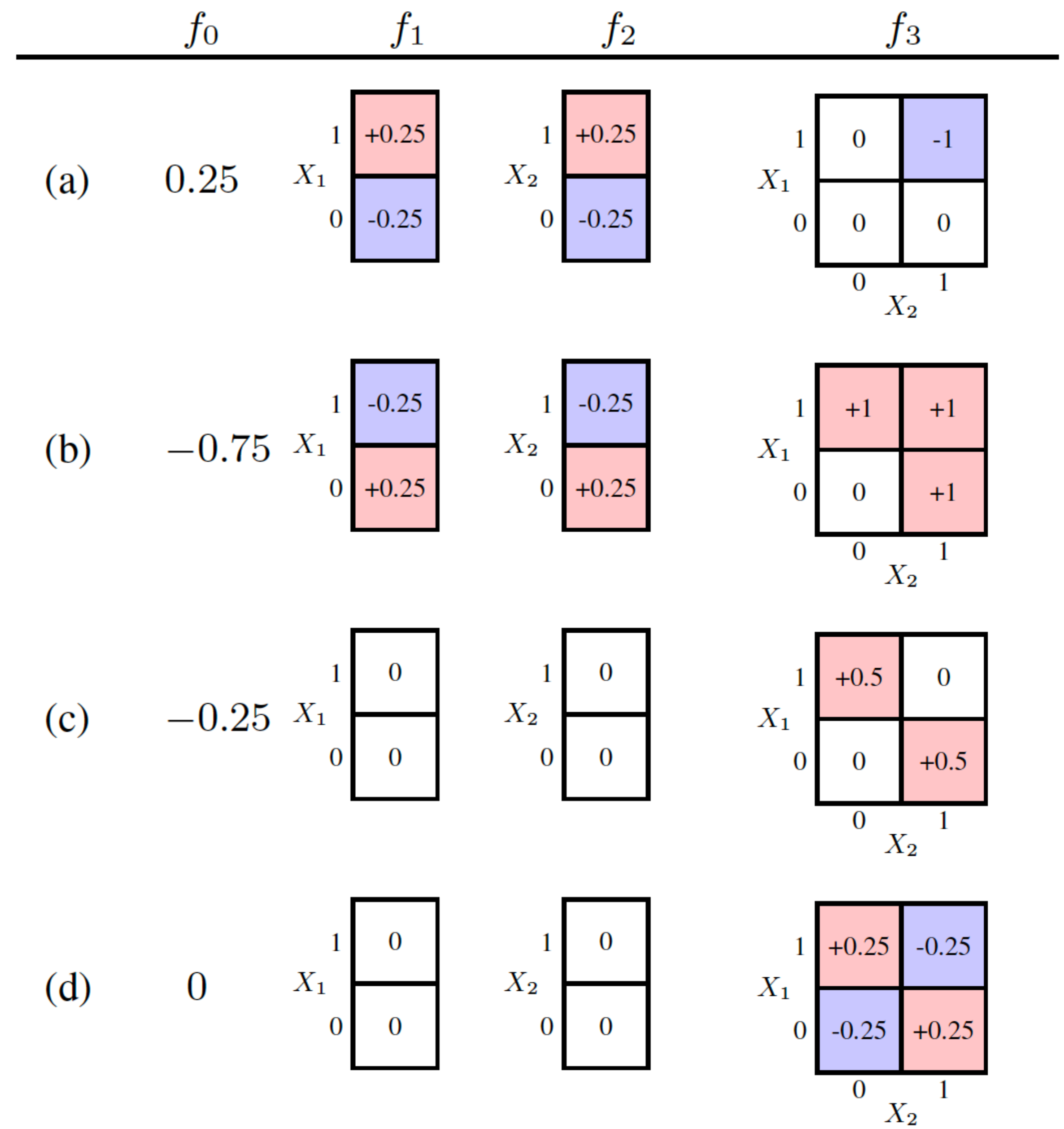
$$Y = X_1 X_2$$

is an interaction effect, then is

$$\log(Y) = \log(X_1 X_2) = \log(X_1) + \log(X_2)$$

an interaction effect?

Are “AND”, “OR”, “XOR” the same or different?



Suppose we have:

$$Y = f_0 + f_1(X_1) + f_2(X_2) + f_3(X_1, X_2)$$

Equivalent realizations can look like “AND”, “OR”, or “XOR”

Pure Interaction Effects

To make things identifiable, let's define a *Pure Interaction Effect of k Variables* as variance in the outcome which cannot be explained any function of fewer than k variables.

This gives us an optimization criterion: **maximize the variance of lower-order terms**.

Functional ANOVA

Statistical framework designed to decompose a function into orthogonal functions on sets of input variables.

Deep roots: [Hoeffding 1948, Huang 1998, Cuevas 2004, Hooker 2004, Hooker 2007]

Functional ANOVA

Given $F(X)$ where $X = (X_1, \ldots, X_d)$, the weighted fANOVA decomposition [Hooker 2004,2007] of $F(X)$ is:

$$\{f_u(X_u) \,|\, u \subseteq [d]\} = \text{argmin}_{\{g_u \in \mathcal{F}^u\}_{u \in [d]}} \int \Big(\sum_{u \subseteq [d]} g_u(X_u) - F(X) \Big)^2 w(X) dX,$$

where $[d]$ indicates the power set of d features, such that

$$\forall \; v \subseteq \; u, \quad \int f_u(X_u) g_v(X_v) w(X) dX = 0 \quad \forall g_v$$

Functional ANOVA

Key property 1 (Orthogonality): [Hooker 2004]

$$\forall\ v \subseteq\ u, \quad \int f_u(X_u) g_v(X_v) w(X) dX = 0$$

Every function f_u is **orthogonal** to any function f_v which operates on any subset of variables in u.

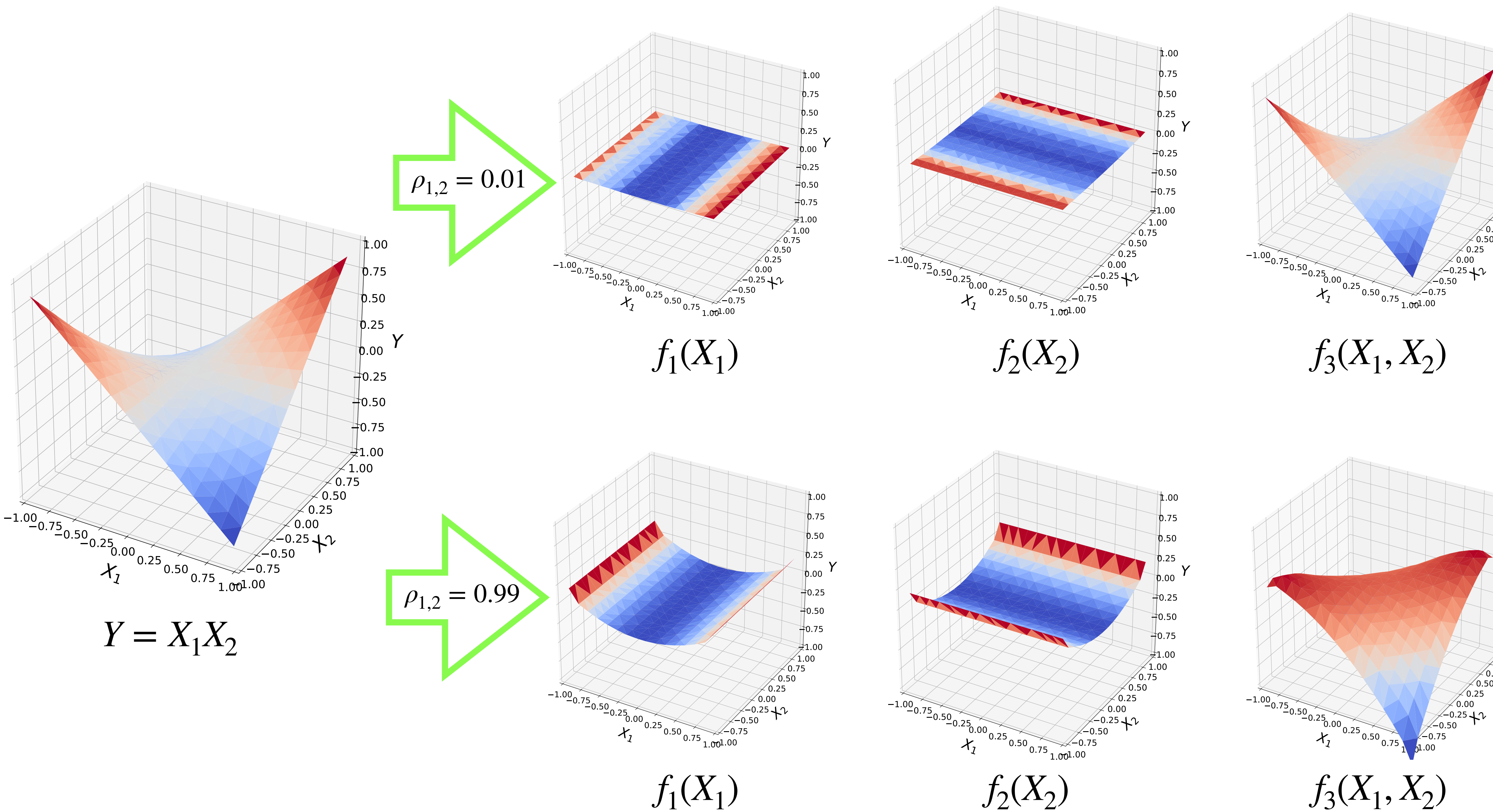
When $w(X) = P(X)$, this means that the functions in the decomposition are all mean-centered and uncorrelated with functions on fewer variables.

Functional ANOVA

Key property 2 (Existence and Uniqueness): [Hooker 2004]

Under reasonable assumptions on the joint distribution $P(X, Y)$, (e.g. no duplicated variables), the functional ANOVA decomposition **exists** and is **unique**.

Functional ANOVA Example



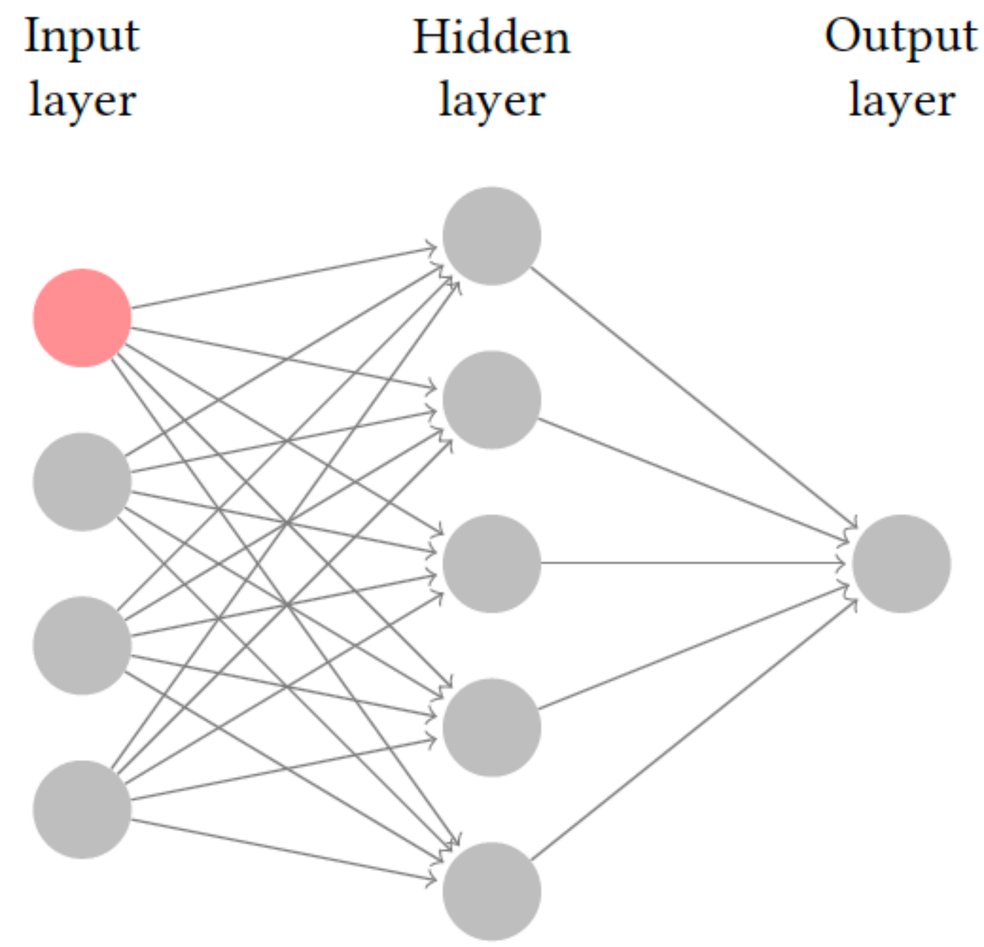
Interaction Effects in Neural Networks

The Challenge of Finding Interaction Effects

- Define: a k-order interaction effect f_u has $|u| = k$
- Give d input variables, there are a potential:
  - $O(d)$ interaction effects of order 1
  - $O(d^2)$ interaction effects of order 2
  - $O(d^3)$ interaction effects of order 3
  - ...
- How do deep nets learn? How do they generalize to test sets?

Dropout

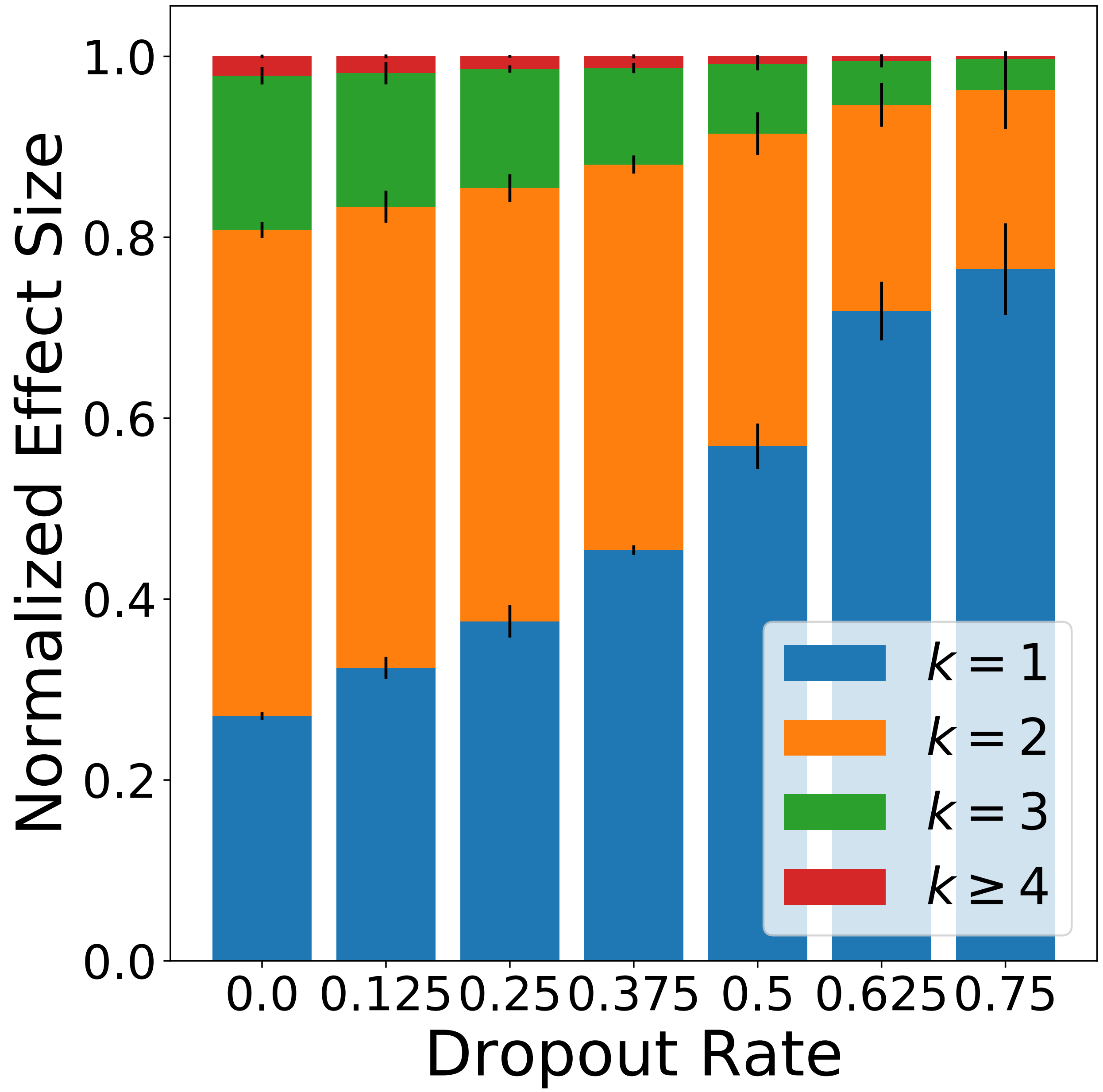
- “Input Dropout” if we drop input features.
- “Activation Dropout” if we drop hidden activations.
- Dropout rate will refer to the probability that the variable is set to 0.



Dropout Regularizes Interaction Effects

- With fANOVA, we can decompose the function estimated by each network into orthogonal functions of k variables.

- As we increase the Dropout rate, the estimated function is increasingly made up of low-order effects.



Dropout Preferentially Targets High-Order Effects

Intuition:

Let's consider Input Dropout. For a pure interaction effect of k variables, all k variables must be retained for the interaction effect to survive.

The probability that k variables all survive Input Dropout decays exponentially with k.

This balances out the exponential growth in k of the size of the hypothesis space.

Dropout Preferentially Targets High-Order Effects

Let $\mathbb{E}[Y \mid X] = F(X) + \epsilon$ with $F(X) = \sum_{u \in [d]} f_u(X_u)$ the fANOVA decomposition, with $\mathbb{E}[Y] = 0$. Let $\tilde{X}$ be X perturbed by Input Dropout, and define $v = \{j : \tilde{X}_j = 0\}$. Then

$$\boxed{\mathbb{E}_{X_u}[f_u(X_u) \mid \tilde{X}_u] = \begin{cases} f_u(\tilde{X}_u) & |v| = 0 \\ 0 & \text{otherwise} \end{cases}}$$

If a single variable in u has been dropped, then we have no information about $f_u(X_u)$

Dropout Preferentially Targets High-Order Effects

$$\mathbb{E}_{X_u}[f_u(X_u) \mid \tilde{X}_u] = \begin{cases} f_u(\tilde{X}_u) & |v| = 0 \\ 0 & \text{otherwise} \end{cases}$$

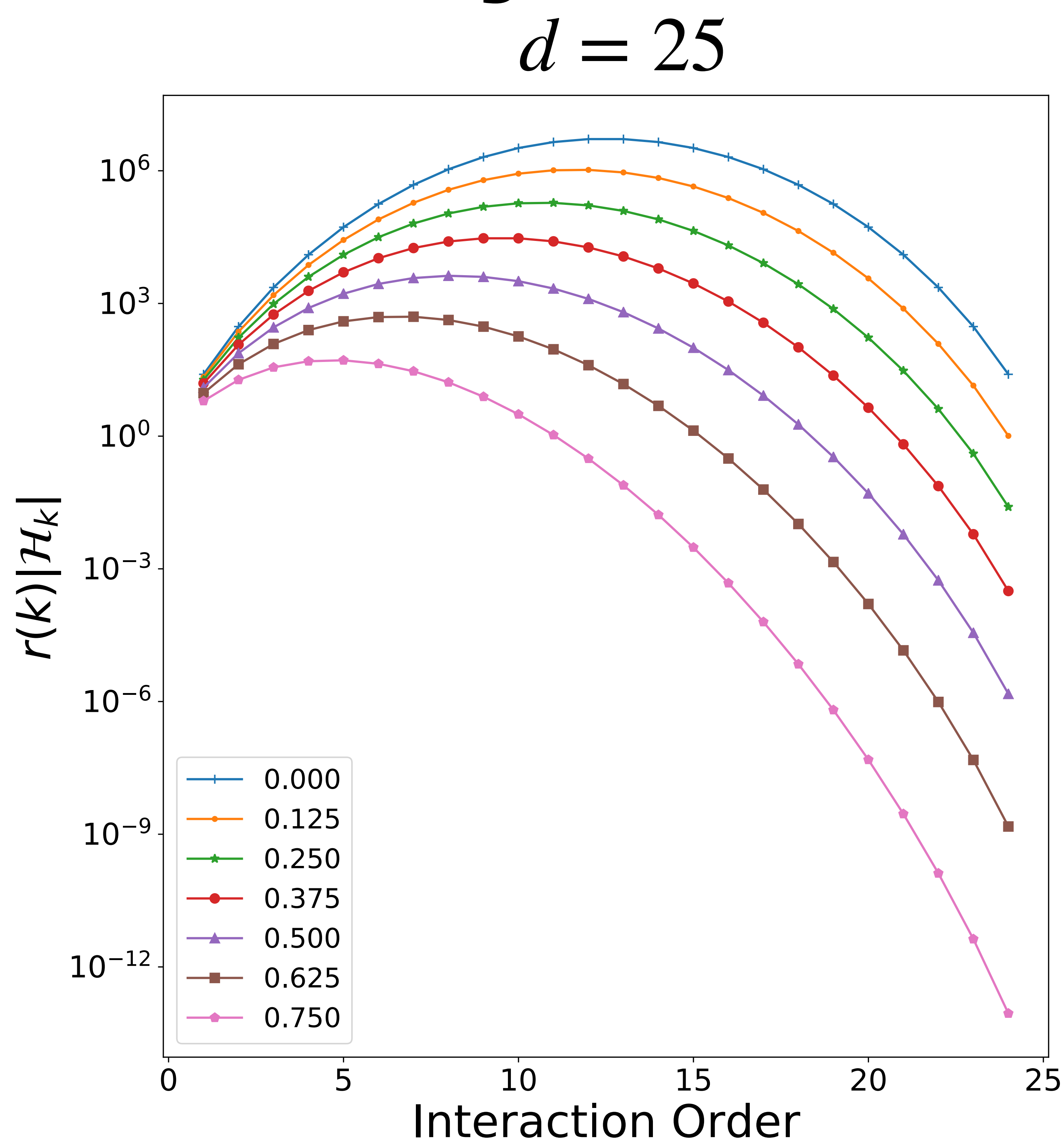
- What is the probability that $|v| = 0$?
  - $(1 - p)^{|u|}$
- Define: $r_p(k) = (1 - p)^k$ the ***effective learning rate*** of a k-order effect.

A Symmetry

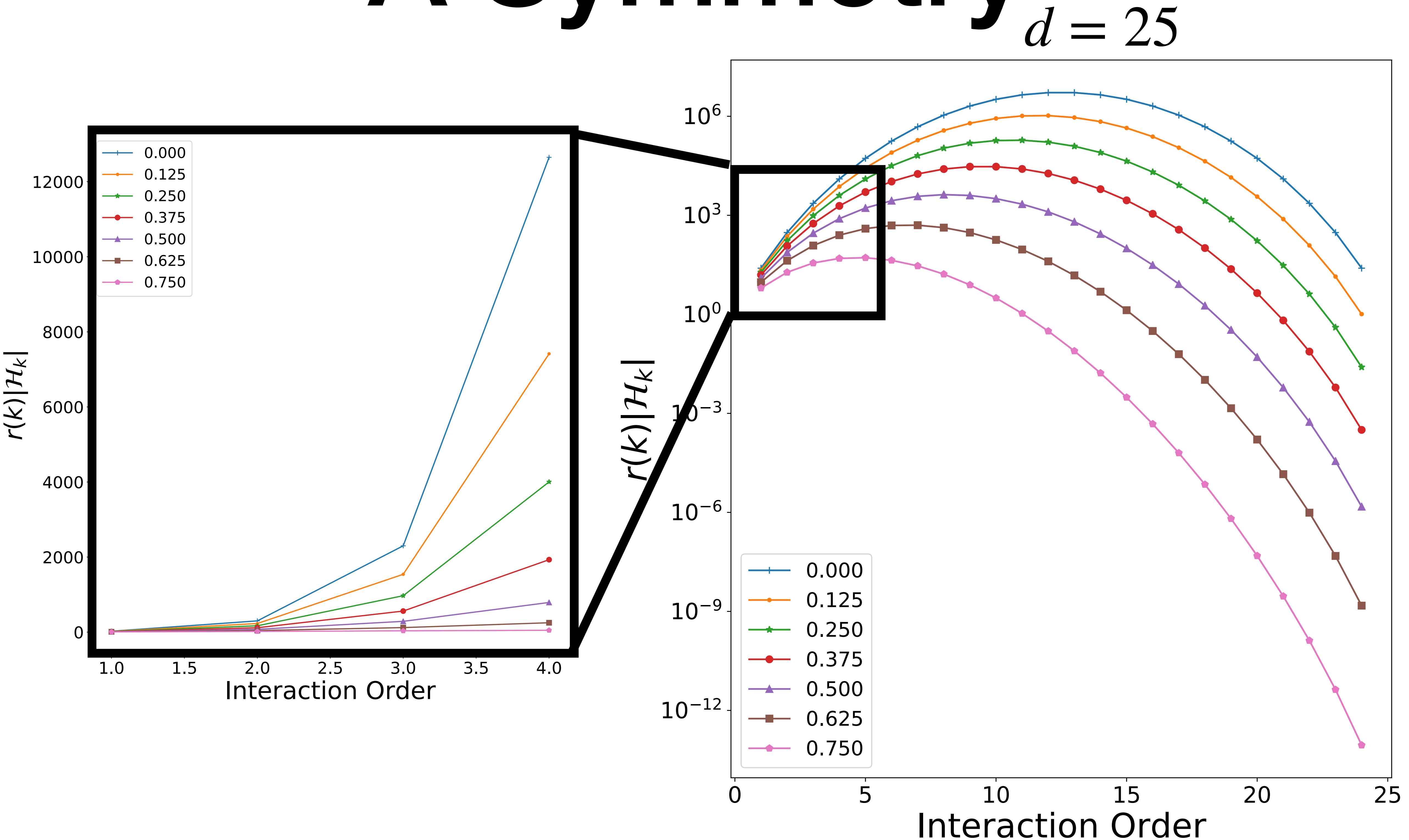
- Define: $r_p(k) = (1-p)^k$ the ***effective learning rate*** of a k-order effect.

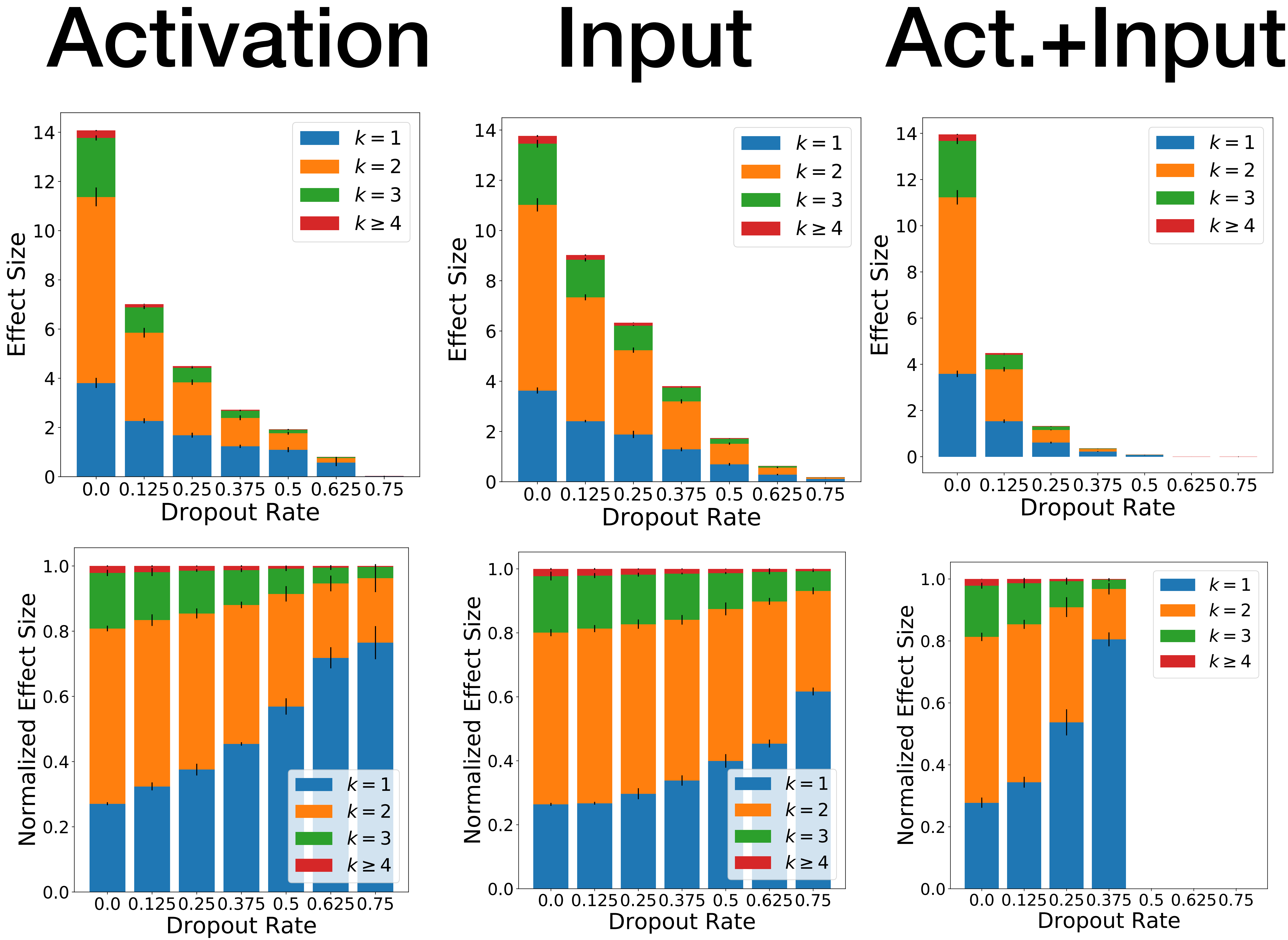
- $|\mathscr{H}_k| = \binom{d}{k}$ hypothesis space size

- Effective learning rate decay and hypothesis space growth in k balance each other out!

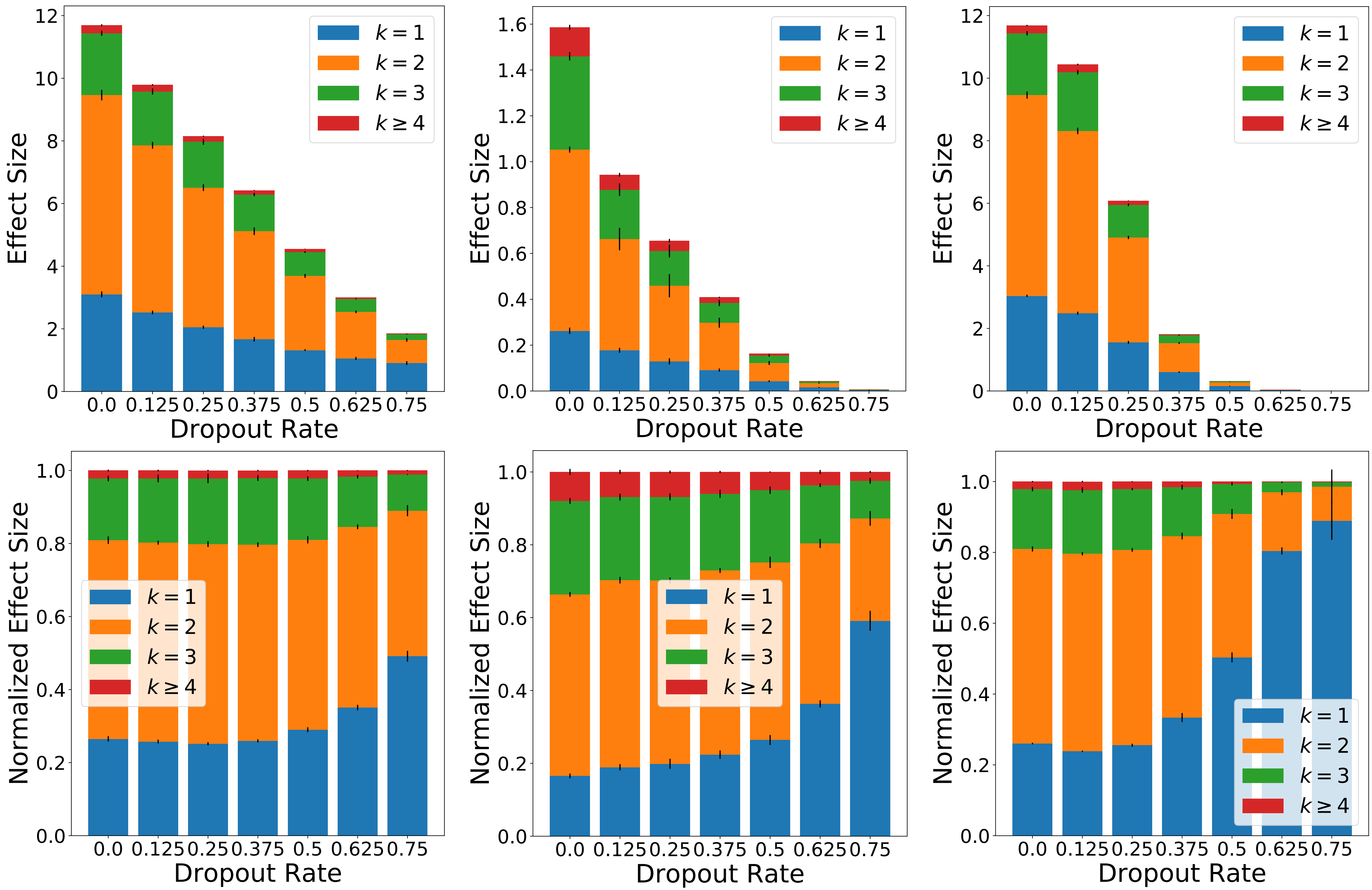


A Symmetry


$d = 25$
0.000
0.125
0.250
0.375
0.500
0.625
0.750
12000
10000
8000
6000
4000
2000
0
$r(k)|\mathcal{H}_k|$
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Interaction Order
10^{6}
10^{3}
10^{0}
10^{-3}
10^{-6}
10^{-9}
10^{-12}
$r(k)|\mathcal{H}_k|$
0
5
10
15
20
25
Interaction Order


Activation
Input
Act.+Input
Effect Size
Normalized Effect Size
Dropout Rate
0.0 0.125 0.25 0.375 0.5 0.625 0.75
k = 1
k = 2
k = 3
k ≥ 4

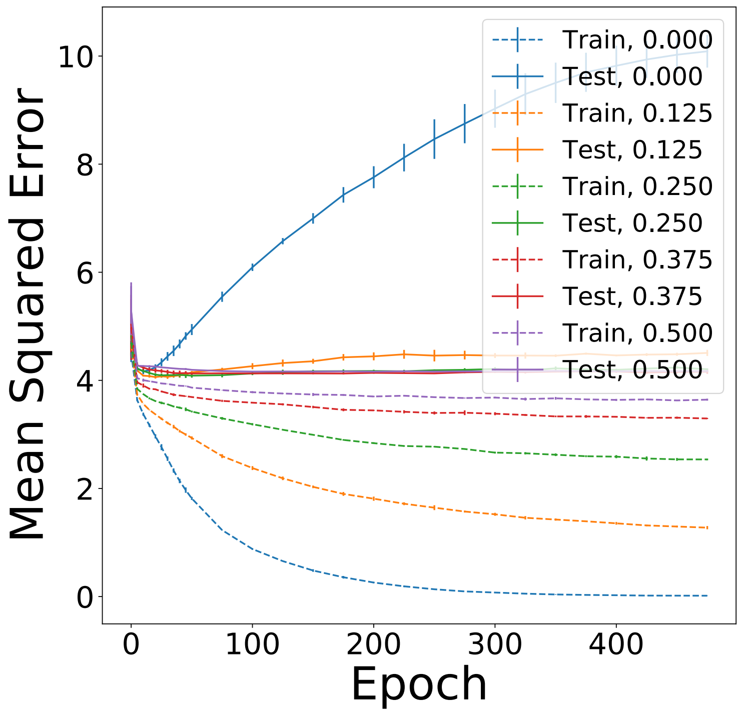
Activation Input Act.+Input

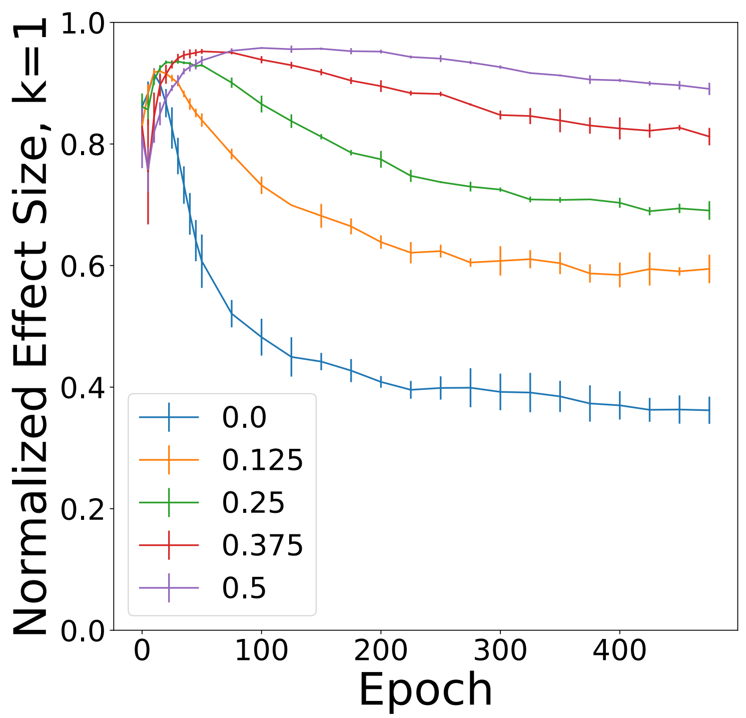


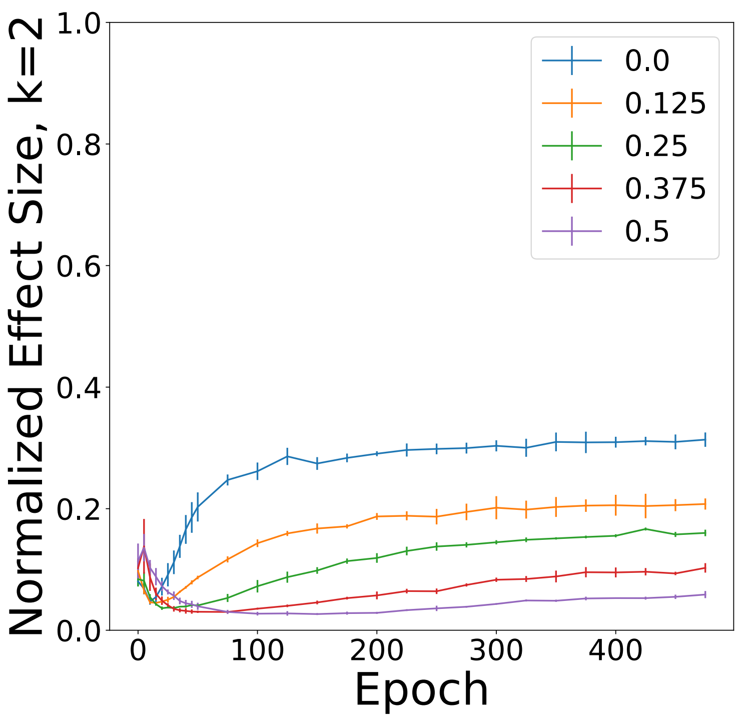
Early Stopping

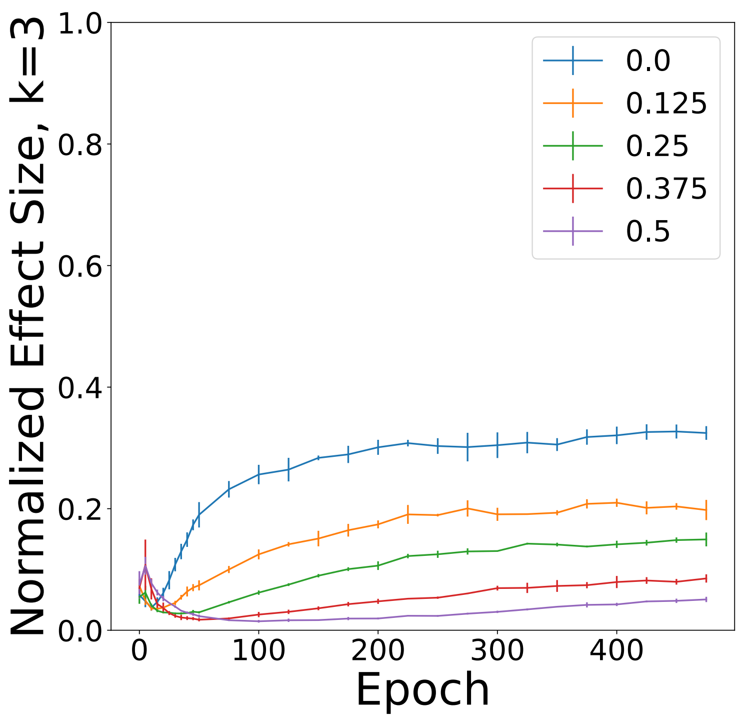
Neural networks tend to start near simple functions, and train toward complex functions [Weigand 1994, De Palma 2019, Nakkiran 2019].

Dropout slows down the training of high-order interactions, making early stopping even more effective.









Implications

- When should we use higher Dropout rates?
  - Higher in Later Layers
  - Lower in ConvNets
- Explicitly modeling interaction effects
- Dropout for explanations / saliency?

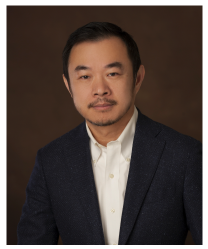
Conclusions

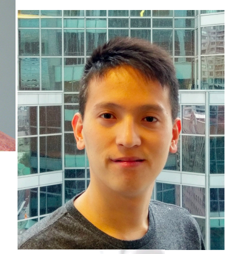
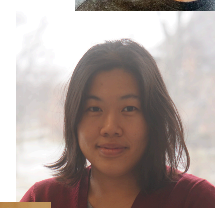
- Interaction effects are tricky — not everything that looks like an interaction is fully interaction.
- Defining pure interaction effects according to the Functional ANOVA gives us an identifiable form.
- The number of potential interaction effects explodes exponentially with order, so searching for high-order interaction effects from data is impossible in practice.
- Dropout is an effective regularizer against interaction effects. It penalizes higher-order effects more than lower-order effects.

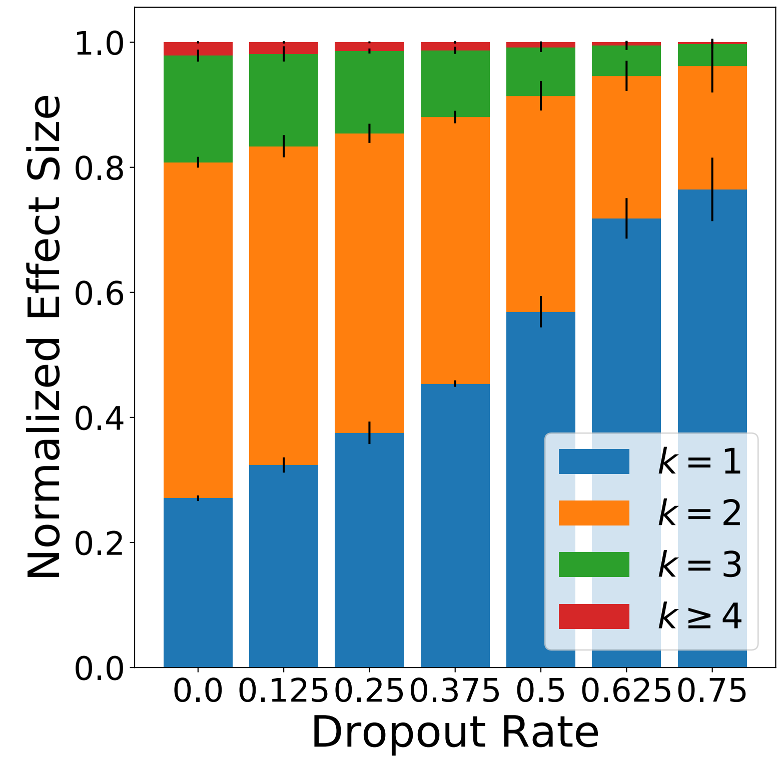
Thank You

Collaborators:

- **Eric Xing**
- **Rich Caruana (MSR)**
- **Chun-Hao Chang (Toronto)**
- **Sarah Tan (Facebook)**
- **Giles Hooker (Cornell)**



- Purifying Interaction Effects with the Functional ANOVA. AISTATS 2020
  - Lengerich, Tan, Chang, Hooker, Caruana

- On Dropout, Overfitting, and Interaction Effects in Deep Neural Networks. Under Review 2020.
  - Lengerich, Xing, Caruana

Carnegie Mellon University Computer Science Department

Microsoft Research

Pittsburgh Health Data Alliance Carnegie Mellon University University of Pittsburgh UPMC

Center for Machine Learning and Health
Carnegie Mellon University

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Dropout Preferentially Targets High-Order Effects

Let $\mathbb{E}[Y|X] = F(X) + \epsilon$ with $F(X) = \sum_{u \in [d]} f_u(X_u)$ the fANOVA decomposition, with $\mathbb{E}[Y] = 0$. Let $\tilde{X}$ be X perturbed by Input Dropout, and define $v = \{j : \tilde{X}_j = 0\}$. Then

$$\mathbb{E}_{X_u}[f_u(X_u)\,|\,\tilde{X}_u] \;= \int f_u(X_u)P(X_u\,|\,\tilde{X})dX_u$$

$$= \int f_u(X_u)I(X_{u \setminus v} = \tilde{X}_{u \setminus v})P(X_v\,|\,\tilde{X})dX_u$$

$$= \int f_h(X_v, \tilde{X}_{u \setminus v})P(X_v\,|\,\tilde{X})dX_v$$

$$= \begin{cases} f_u(\tilde{X}_u) & |v| = 0 \\ 0 & \text{otherwise} \end{cases}$$

Advantage of using fANVOA to define f_u — these are zero!