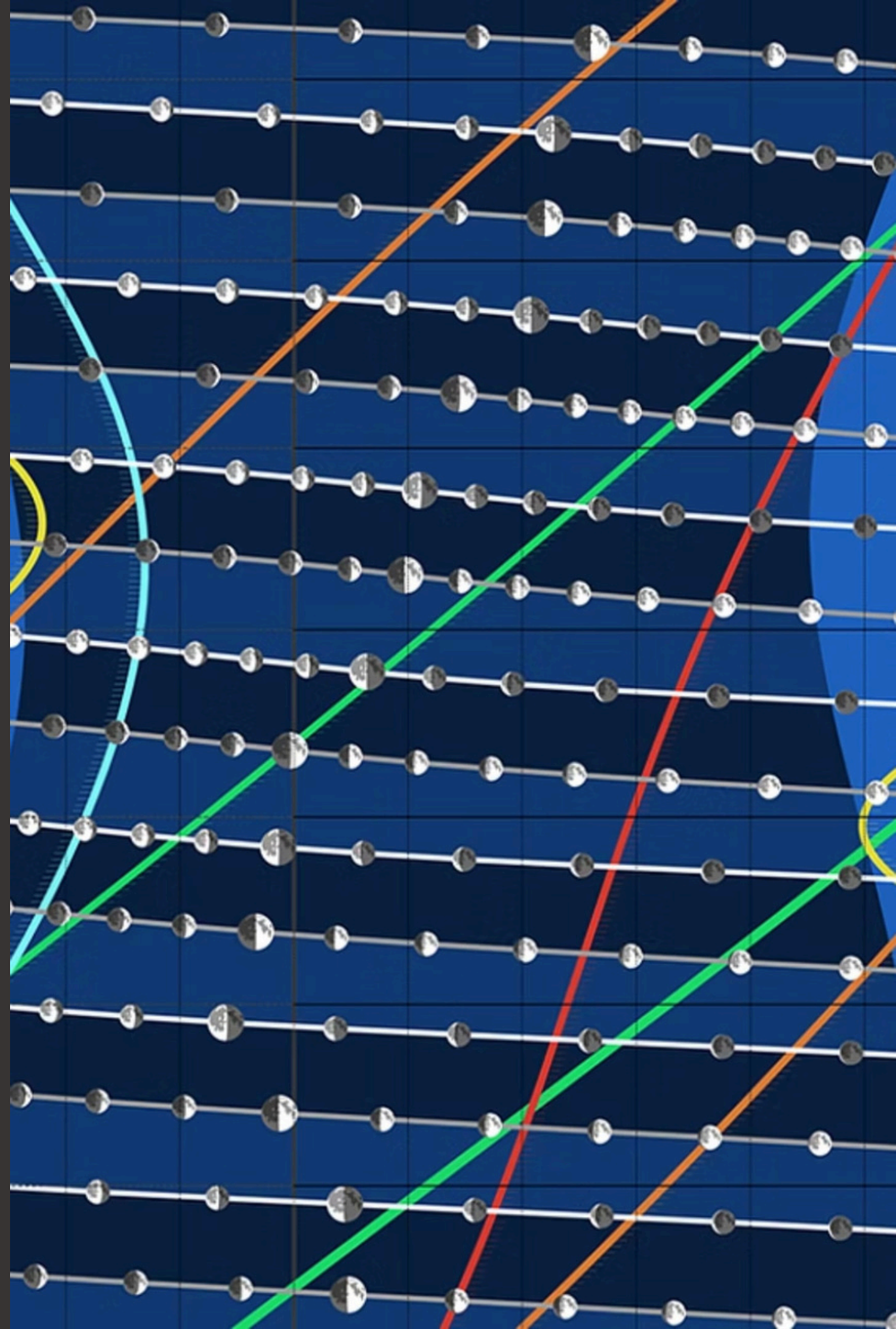
Bias-Variance และ Learning Curves

การวิเคราะห์ความเอนเอียงและความแปรปรวนของแบบจำลองเชิงเส้น

รายวิชา Machine Learning · ภาคการศึกษาที่ 2569

จัดทำโดย: นาย วีรภัทร สาลีผล รหัสนักศึกษา: 6610301006
นาย พชร เพิ่มพูล รหัสนักศึกษา: 6610301008



โจทย์และเป้าหมายของงาน

งานนี้มุ่งวิเคราะห์ **Bias** และ **Variance** ของแบบจำลอง 3 ประเภท ด้วย 2 วิธีคำนวณ บนฟังก์ชันเป้าหมาย 2 รูปแบบ โดยสุ่มข้อมูลแบบเอกรูป 2 จุดในช่วง *[-1, 1]* และใช้ **Normal Equation** เท่านั้น (ห้ามใช้ Gradient Descent)

3 แบบจำลองที่ศึกษา

- แบบจำลองค่าคงที่: $h(x) = b$
- แบบจำลองเชิงเส้น: $h(x) = ax + b$
- แบบจำลองเชิงเส้นผ่านจุดกำเนิด: $h(x) = ax$

2 ฟังก์ชันเป้าหมาย

- $\sin(\pi x)$— ฟังก์ชันตรีโกณมิติไม่เป็นเชิงเส้น
- x^2 — ฟังก์ชันพหุนามกำลังสอง

2 วิธีคำนวณ

- **Analytical Method:** คำนวณค่าคาดหวัง E[·] ด้วยการอินทิเกรต
- **Simulation Method:** Monte Carlo สุ่มชุดข้อมูล 100,000 รอบแล้วหาค่าเฉลี่ย

ทฤษฎีพื้นฐาน: Bias, Variance และ E_out

กรอบทฤษฎี Bias-Variance Decomposition อธิบายองค์ประกอบของข้อผิดพลาดนอกชุดฝึก *E_{out}* ซึ่งแบ่งออกเป็น 3 ส่วนหลักที่มีความสัมพันธ์กันโดยตรง

Bias (ความเอนเอียง)

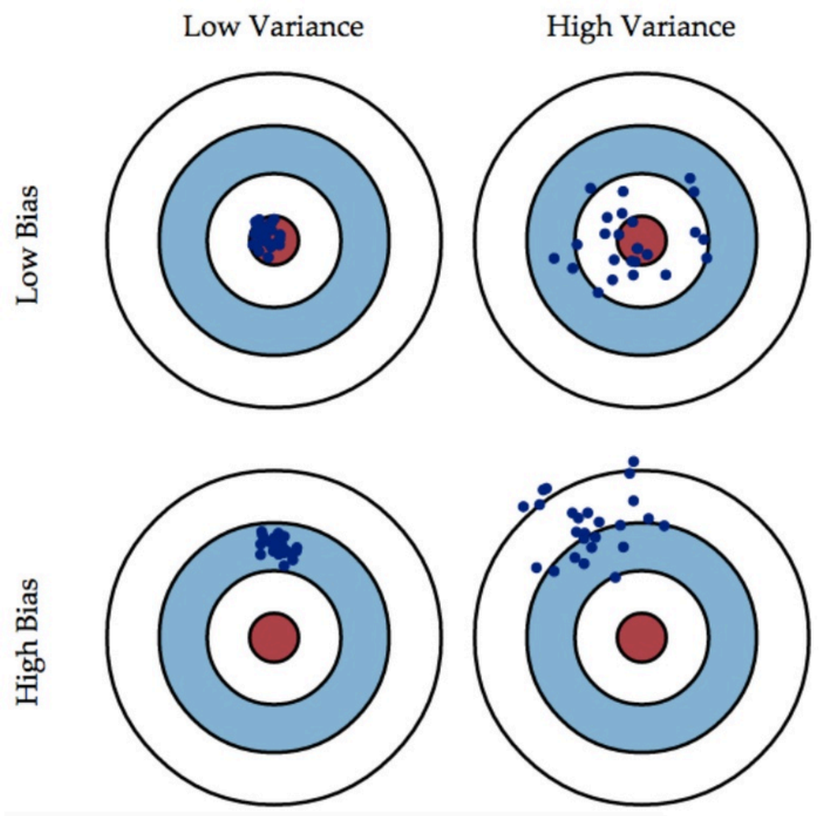
ค่าความคลาดเคลื่อนเฉลี่ยระหว่าง **ค่าทำนายเฉลี่ยของแบบจำลอง** กับ **ฟังก์ชันเป้าหมายจริง** $\text{Bias}(x) = \bar{g}(x) - f(x)$ ค่า Bias สูงหมายถึงแบบจำลองมีความยืดหยุ่นไม่เพียงพอที่จะประมาณฟังก์ชันจริงได้

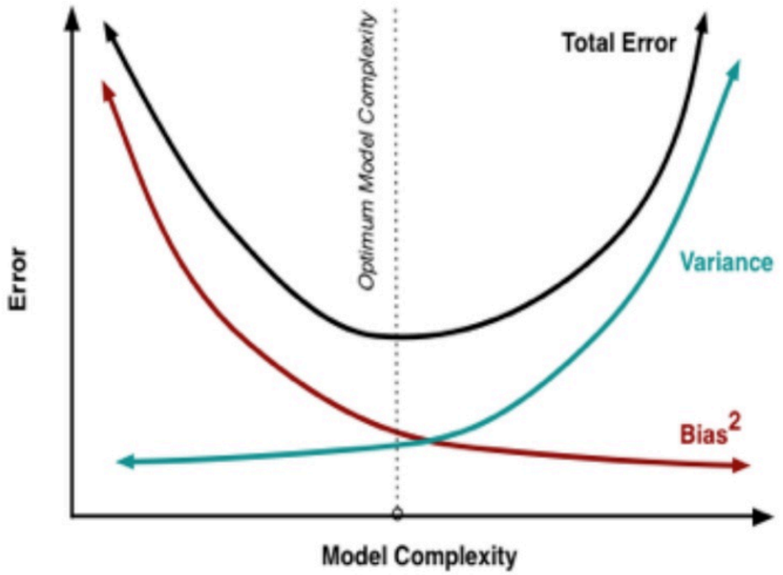
Variance (ความแปรปรวน)

ค่าความผันผวนของแบบจำลองเมื่อ **เปลี่ยนชุดข้อมูลฝึก** $\text{Var}(x) = E_D[(g_D(x) - \bar{g}(x))^2]$ ค่า Variance สูงหมายถึงแบบจำลองไวต่อการเปลี่ยนแปลงข้อมูลฝึกมากเกินไป (Overfitting)

การแยกองค์ประกอบของ E_out

ข้อผิดพลาดนอกชุดฝึกสามารถแยกได้เป็น : $E_{out} = \text{Bias}^2 + \text{Variance} + \sigma^2$ โดยที่ σ^2 คือค่าความแปรปรวนของ Noise ในข้อมูล ซึ่งไม่สามารถลดได้ด้วยแบบจำลองใดๆ





เปรียบเทียบ 3 แบบจำลอง

ตารางด้านล่างสรุปคุณสมบัติหลักของแบบจำลองทั้ง 3 ประเภทที่ใช้ในการทดลอง ครอบคลุมสมการ จำนวนพารามิเตอร์ และระดับความซับซ้อนของแบบจำลอง

ชื่อแบบจำลอง	สมการ	พารามิเตอร์	ความซับซ้อน	ดีกรี
ค่าคงที่ (Constant)	$h(x) = b$	1 ตัว (b)	ต่ำสุด	0
เชิงเส้น (Linear)	$h(x) = ax + b$	2 ตัว (a, b)	ปานกลาง	1
เชิงเส้นผ่านจุดกำเนิด	$h(x) = ax$	1 ตัว (a)	ต่ำ–ปานกลาง	1

Code Snippet — ฟังก์ชัน fit (Python / NumPy): ใช้ Normal Equation $w = (X^TX)^{-1}X^Ty$ (ใช้ np.linalg.lstsq ซึ่งแก้ normal equation $X^TXw = X^Ty$) สำหรับทุกแบบจำลอง โดยสร้างเมทริกซ์ design matrix X ให้สอดคล้องกับรูปแบบของแบบจำลองแต่ละประเภท

2 วิธีในการคำนวณ Bias และ Variance

Analytical Method

คำนวณค่าคาดหวัง E[·] โดยการ **อินทิเกรต**ทั่วทั้งช่วงข้อมูล x ∈ [-1, 1] และทุกความเป็นไปได้ของชุดฝึก

- ได้ผลลัพธ์ที่ **แม่นยำสมบูรณ์** (Exact)
- ไม่ต้องสุ่มข้อมูล จึงไม่มี Sampling Error
- คำนวณได้ยากสำหรับปัญหาที่ซับซ้อน
- เหมาะสำหรับการพิสูจน์ทางทฤษฎี

Simulation Method (Monte Carlo)

สุ่มชุดข้อมูลฝึก *N = 2* จุด จำนวน **100,000 รอบ** แล้วหาค่าเฉลี่ยของผลลัพธ์เพื่อประมาณค่าคาดหวัง

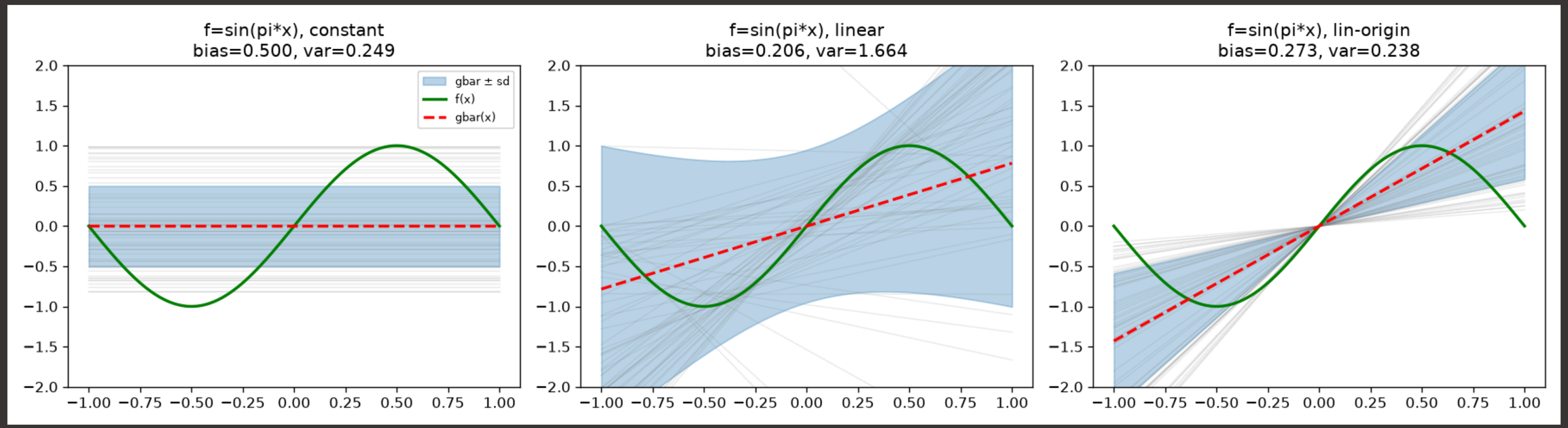
- ได้ผลลัพธ์ที่ **ประมาณค่า** (Approximate)
- มีความคลาดเคลื่อนจากการสุ่ม (Sampling Noise)
- นำไปใช้กับปัญหาทั่วไปได้ง่าย
- เหมาะสำหรับการทดลองเชิงปฏิบัติ

ทั้งสองวิธีควรให้ผลลัพธ์ที่ **ใกล้เคียงกันมาก** หากจำนวนรอบการจำลองมากพอ ซึ่งเป็นการตรวจสอบความถูกต้องของผลลัพธ์ซึ่งกันและกัน

ผลลัพธ์ข้อที่ 1: ฟังก์ชันเป้าหมาย sin(πx)

ตารางค่า Bias และ Variance ของแบบจำลองทั้ง 3 ประเภท เปรียบเทียบระหว่างวิธี Analytical และ Simulation (100,000 รอบ) บนฟังก์ชัน f(x) = sin(πx) ช่วง $x \in [-1, 1]$

แบบจำลอง	**Bias² (Analytical)**	**Bias² (Simulation)**	**Variance (Analytical)**	**Variance (Simulation)**
ค่าคงที่ $h(x)=b$	0.5000	0.4995	0.2500	0.2494
เชิงเส้น $h(x)=ax+b$	0.2067	0.2063	1.6763	1.8830
ผ่านจุดกำเนิด $h(x)=ax$	0.2706	0.2732	0.2380	0.5072

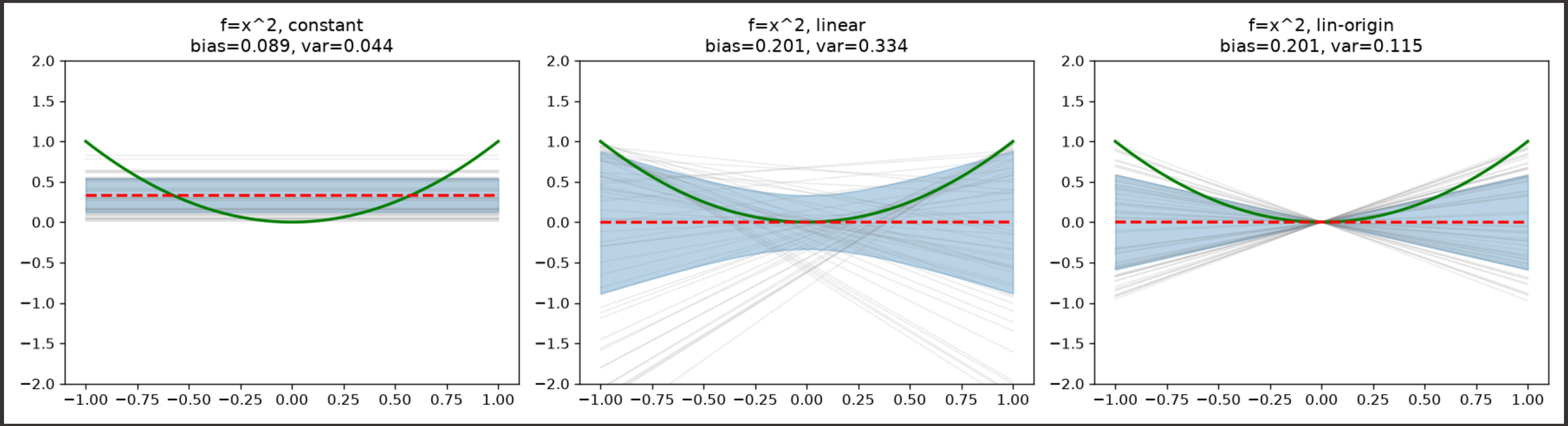


กราฟ Bias-Variance — **sin(πx):** [พื้นที่สำหรับแทรกรูปกราฟเปรียบเทียบ E_in และ E_out ของทั้ง 3 แบบจำลอง แยกกรณีไม่มี Noise และมี Noise]

ผลลัพธ์ข้อที่ 2: ฟังก์ชันเป้าหมาย x^2

ตารางค่า Bias และ Variance ของแบบจำลองทั้ง 3 ประเภท เปรียบเทียบระหว่างวิธี Analytical และ Simulation (100,000 รอบ) บนฟังก์ชัน $f(x) = x^2$ ช่วง $x \in [-1, 1]$

แบบจำลอง	Bias² (Analytical)	Bias² (Simulation)	Variance (Analytical)	Variance (Simulation)
ค่าคงที่ $h(x)=b$	0.0889	0.0892	0.0444	0.0444
เชิงเส้น $h(x)=ax+b$	0.2000	0.2011	0.3333	0.3340
ผ่านจุดกำเนิด $h(x)=ax$	0.2000	0.2008	0.1149	0.1153



กราฟ Bias-Variance — **x^2 :** [พื้นที่สำหรับแทรกรูปกราฟเปรียบเทียบ E_in และ E_out ของทั้ง 3 แบบจำลอง แยกกรณีไม่มี Noise และมี Noise]

สรุป Bias-Variance Tradeoff

ความสัมพันธ์ระหว่างความซับซ้อนของแบบจำลองกับค่า Bias และ Variance แสดงให้เห็นถึง **Tradeoff** ที่ไม่สามารถหลีกเลี่ยงได้ในการเลือกแบบจำลอง

แบบจำลองง่าย (ค่าคงที่)

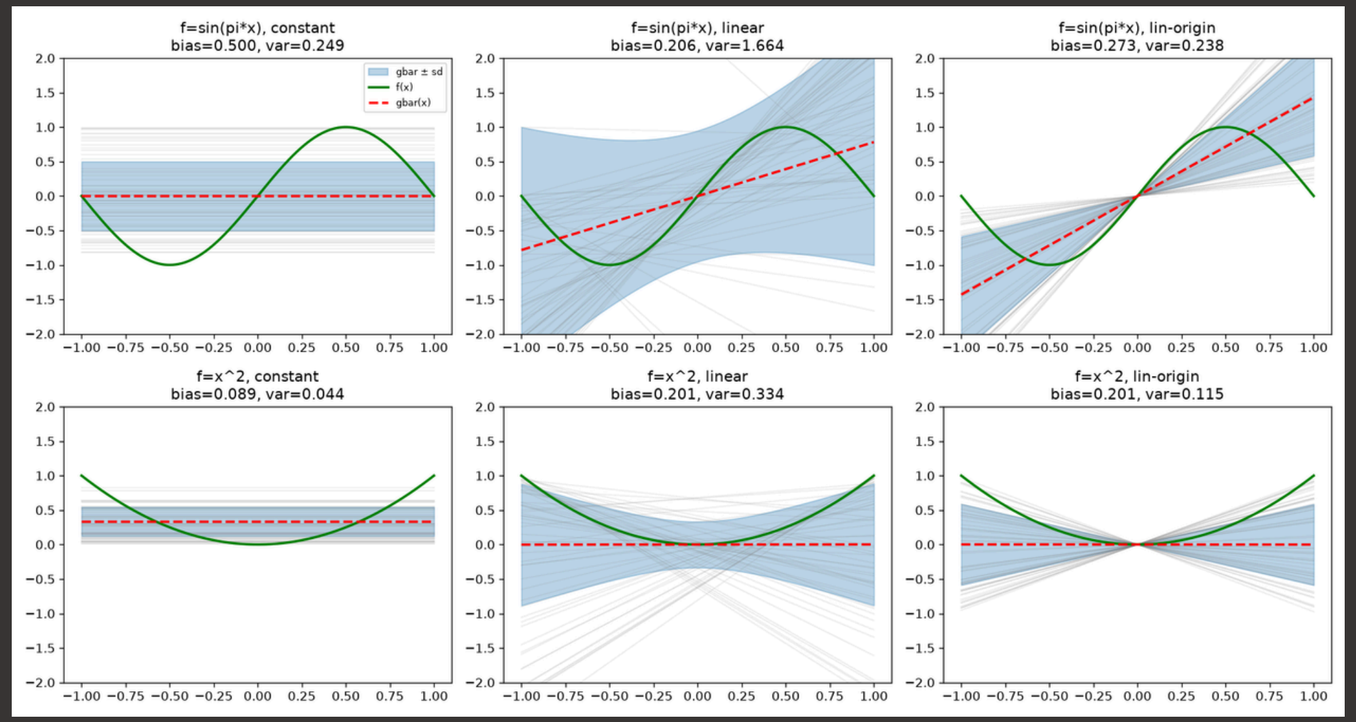
Bias: สูง — ไม่สามารถประมาณฟังก์ชันไม่เป็นเชิงเส้นได้ดี
Variance: ต่ำ — แบบจำลองไม่เปลี่ยนแปลงมากเมื่อเปลี่ยนชุดข้อมูล
เหมาะกับข้อมูลที่มี Noise สูงหรือข้อมูลน้อยมาก

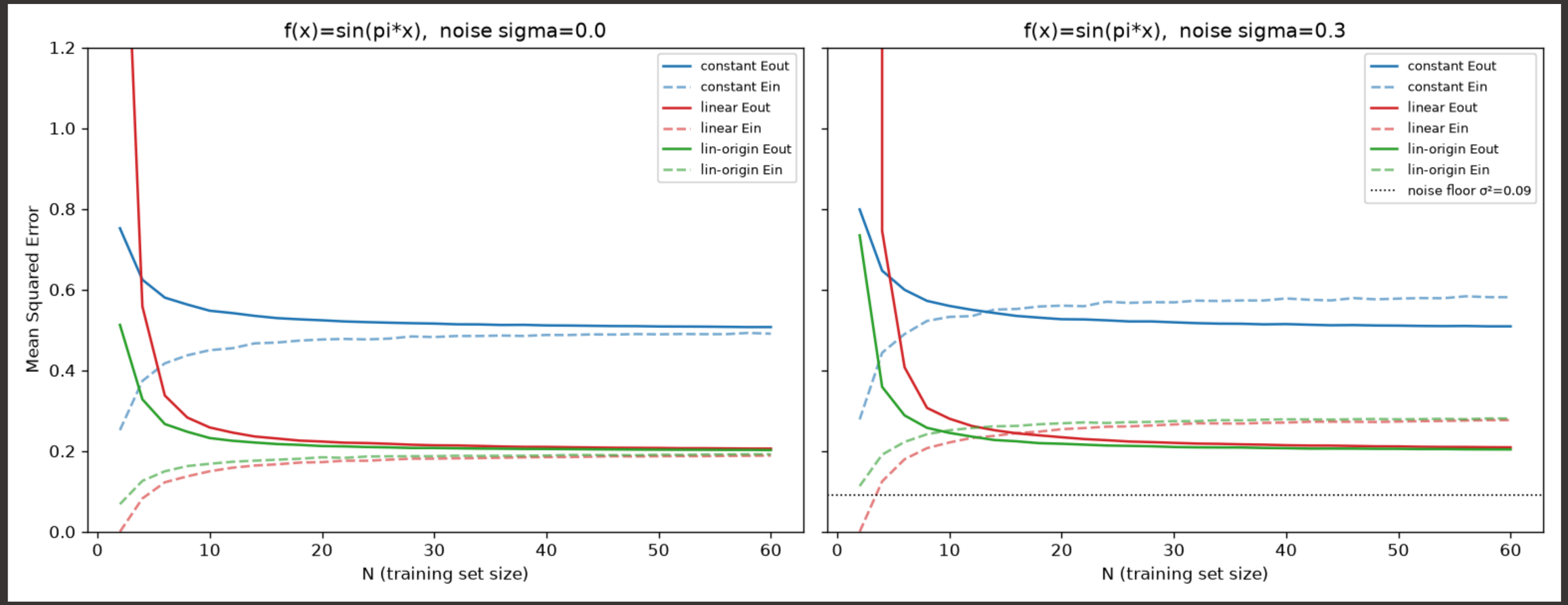
แบบจำลองปานกลาง (เชิงเส้นผ่านจุดกำเนิด)

Bias: ปานกลาง — ยืดหยุ่นกว่าค่าคงที่ แต่ยังมีข้อจำกัด
Variance: ปานกลาง — ไวต่อข้อมูลน้อยกว่าแบบจำลองเชิงเส้นเต็มรูปแบบ
เป็นทางเลือกกลางที่สมดุลในหลายสถานการณ์

แบบจำลองซับซ้อน (เชิงเส้นเต็ม)

Bias: ต่ำ — ประมาณฟังก์ชันจริงได้ดีกว่า
Variance: สูง — โดยเฉพาะเมื่อ N=2 แบบจำลองไวต่อการเปลี่ยนแปลงข้อมูลมาก
เสี่ยง Overfitting เมื่อข้อมูลฝึกมีจำนวนน้อย





กราฟ Learning Curves เปรียบเทียบ: [พื้นที่สำหรับแทรกรูปกราฟ Learning Curves ของทั้ง 3 แบบจำลอง แยกกรณีไม่มี Noise และมี Noise แสดง E_in และ E_out]

เส้นโค้งการเรียนรู้ (Learning Curves)

องค์ประกอบของกราฟ

- **แกน X:** ขนาดชุดฝึก N (จำนวนจุดข้อมูล)
- **แกน Y:** ค่า E_in และ E_out
- **เส้นทึบ:** E_out (ข้อผิดพลาดนอกชุดฝึก)
- **เส้นประ:** E_in (ข้อผิดพลาดในชุดฝึก)

2 เงื่อนไขการทดลอง

- **ไม่มี Noise:** *E_out* ลดลงจนเข้าใกล้ 0 เมื่อ N เพิ่มขึ้น
- **มี Noise:** *E_out* ลดลงแต่ไม่สามารถต่ำกว่า σ^2 (Noise Floor)

สิ่งที่คาดหวังจาก Learning Curves

- เมื่อ N เพิ่มขึ้น *E_in* จะ **เพิ่มขึ้น** (แบบจำลองเห็นข้อมูลมากขึ้น)
- เมื่อ N เพิ่มขึ้น *E_out* จะ **ลดลง** (แบบจำลอง generalize ได้ดีขึ้น)
- ช่องว่างระหว่าง *E_in* และ *E_out* บ่งชี้ระดับ **Variance**
- แบบจำลองซับซ้อนจะมีช่องว่างกว้างกว่าเมื่อ N น้อย

กราฟ Learning Curves เปรียบเทียบ: [พื้นที่สำหรับแทรกรูปกราฟ Learning Curves ของทั้ง 3 แบบจำลอง แยกกรณีไม่มี Noise และมี Noise แสดง E_in และ E_out]

ข้อสังเกตสำคัญจากการทดลอง

1

E_out ระเบิดสูงมากเมื่อมี Noise ที่ N=2

แบบจำลองเชิงเส้น $h(x) = ax + b$ เมื่อฝึกด้วยข้อมูลเพียง 2 จุดที่มี Noise ค่า *E_out* จะ **สูงผิดปกติ** เนื่องจาก Variance สูงมากเมื่อข้อมูลน้อย แบบจำลองพยายาม Fit ข้อมูลที่มี Noise จนเกิด Overfitting รุนแรง ทำให้การทำนายบนข้อมูลใหม่น่าเชื่อถือได้น้อยมาก

2

Noise Floor = σ^2 คือขีดจำกัดล่างของ Error

เมื่อข้อมูลมี Noise ค่า *E_out* ที่ดีที่สุดที่เป็นไปได้จะไม่ต่ำกว่า σ^2 (ความแปรปรวนของ Noise) ไม่ว่าแบบจำลองจะซับซ้อนเพียงใดก็ตาม นี่คือ **Irreducible Error** ที่ไม่สามารถกำจัดได้ด้วยการปรับปรุงแบบจำลอง เป็นข้อจำกัดพื้นฐานของปัญหาที่มีข้อมูลไม่สมบูรณ์

บทสรุป: การเข้าใจ Bias-Variance Tradeoff และ Learning Curves ช่วยให้นักพัฒนาเลือกความซับซ้อนของแบบจำลองได้เหมาะสมกับปริมาณและคุณภาพของข้อมูลที่มีอยู่