

8차 모임

2022.11.04.

한국에너지기술연구원 계산과학연구실

이제현

2022 에너지+AI 학습 조직 : 총 9회 + α

2차 모임 (5월)

김준우

Seeing is believing?

3차 모임 (6월)

한광우

강화학습 기반
에너지 시스템 운영 제어

4차 모임 (7월)

윤민혜 → 김경현

신소재 개발 관련
데이터 기법 및 동향

5차 모임 (8월)

손은국 → 박성수

강화학습 기술 &
풍력단지제어 적용 사례

6차 모임 (8월)

김보영 → 윤경호

AI 기반 메디컬 트윈

7차 모임 (9월)

주영환

GNN 위상설계 최적화

8차 모임 (10월)

이효진

(실험 설계)

9차 모임 (11월)

오명찬

(전문가 초청)

- **발표자** : 1. 개인 일정에 따라 당사자간 합의 하에 양도 가능,
2. 외부 연사 초빙 가능.
- **발표** : 자료 및 발표 영상 원내 공개 기본. 발표자 재량에 따라 비공개 및 유튜브 등 외부 공개 가능.

머신 러닝 강좌

1차 모임 (4월)

머신러닝 기본 개념

- base : Scikit-learn MOOC @inria

- 머신 러닝 위주. 딥 러닝은 skip (기회가 되면 한번쯤은 다룰 수도..?)
- 소스 코드 포함 강의 자료 : 원내 게시판 공개
- 강의 영상 : KIER-Tube & Youtube 공개

2차 모임 (5월)

Modeling pipeline

3차 모임 (6월)

Best Model?

4차 모임 (7월)

Tree Models

5차 모임 (8월)

Hyperparameter

6차 모임 (8월)

Nonlinear Models

7차 모임 (9월)

Ensemble Models

8차 모임 (10월)

Neural Network (?)

Gaussian Process

9차 모임 (11월)

마무리

+ Open API

NLP APIs

Browse the best premium and free APIs on the world's largest API Hub. Read about the latest API news, tutorials, SDK documentation, and API examples. RapidAPI offers free APIs all within one SDK. One API key. One dashboard.



NLP Translation

Neural Translate. 110+ Langs. HTML/Text/JSON. Protected words. Multiple target langs in single request.

🔗 9.9 ⌚ 1,632 ms ✓ 100%



Recipe Search and Diet

Since we understand the recipe – ingredients, diets, allergies, nutrition, taste, techniques & more. We can connect your users with the

🔗 9.6 ⌚ 714 ms ✓ 100%



Edamam Food and Grocery Database

This API provides you with tools to find nutrition and diet data for generic foods, packaged foods and

🔗 9.7 ⌚ 863 ms ✓ 100%



Medium

Medium API helps you quickly extract data from Medium's Website (https://medium.com). You

Verified ✓
🔗 9.6 ⌚ 309 ms ✓ 100%



Wiki Briefs

Briefs about anything you search. No need to read lengthy articles, we summarize.

🔗 9.5 ⌚ 849 ms ✓ 100%



Address Parser - Advanced NLP

Parse unstructured address strings into address components, using advanced NLP methods.

🔗 8.8 ⌚ 132 ms ✓ 100%



TextAPI

A text extraction, manipulation, and analysis api. Putting the power of natural language processing (nlp) in every developers

🔗 8.8 ⌚ 26,670 ms ✓ 100%



Twinword Text Analysis Bundle

One API for all your text analysis needs. Sentiment Analysis, Topic Tagging,

Verified ✓
🔗 9.2 ⌚ 565 ms ✓ 100%



TextAnalysis

TextAnalysis API provides customized Text Analysis,Text Mining and Text Processing Services like Text Summarization,

🔗 9.4 ⌚ 32,539 ms ✓ 76%



Sentence Case Converter - Truecaser

Corrects/restores the capitalization of text, taking proper nouns into account.

🔗 8.7 ⌚ 63 ms ✓ 100%



Rewriter Paraphraser

Advanced rewriter and paraphraser that gives you high manipulation capabilities.

🔗 9.2 ⌚ 21,206 ms ✓ 100%



Summarize Texts

Use One AI's NLP models to automatically summarize texts. Configure this API in [the Language Studio] (https://studio.oneai.com/?

🔗 8.1 ⌚ 1,498 ms ✓ 100%



Text-Analysis

A Natural Language Processing (NLP) API enabling users to perform a variety of NLP tasks on a great amount of news

🔗 7.6 ⌚ 498 ms ✓ 100%



Sentino

Sentino API is used to understand personality traits (Big5, NEO etc.) using NLP.

🔗 7.5 ⌚ 3,151 ms ✓ 100%



Article Parser and Summary Free 1000 Requests

This API can parse any article and provide summary be it news article, blog post,

🔗 8.4 ⌚ 3,897 ms ✓ 81%

https://rapidapi.com/hub

Gaussian Process

실습 코드 : <https://bit.ly/3NvEwIO>



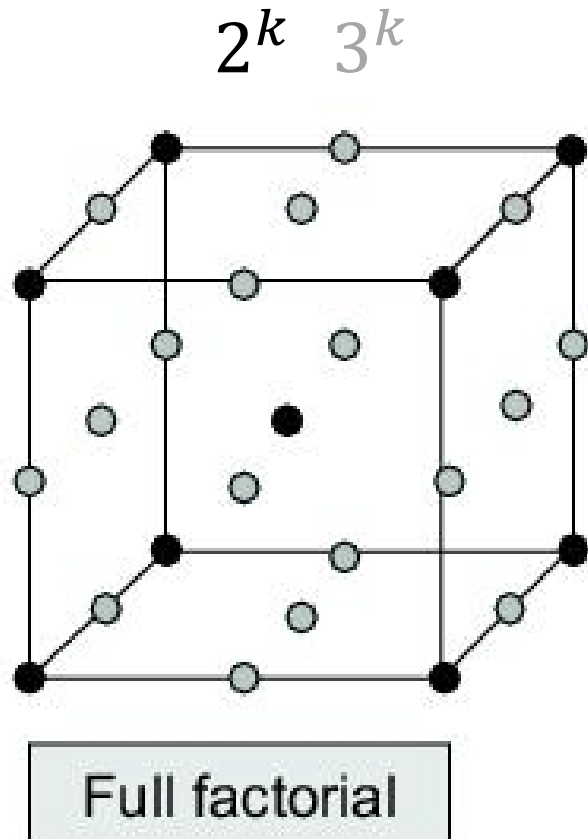
2022.11.04.

한국에너지기술연구원 계산과학연구실

이제현

Design of Experiment 실험 계획법

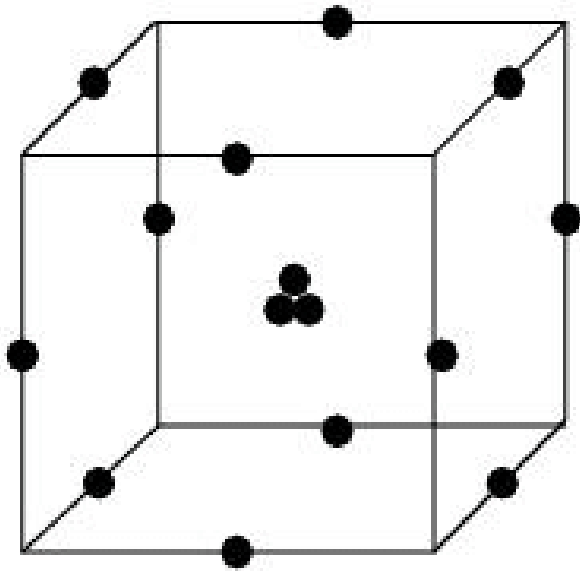
- k factors vs responses



Design of Experiment 실험 계획법

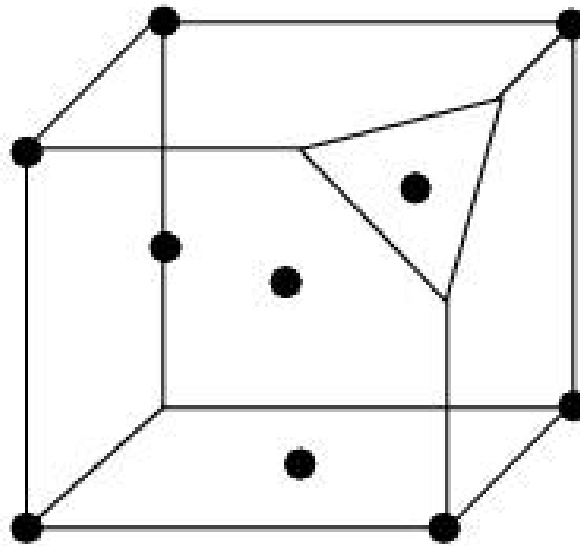
- k factors vs responses

Multiple Center Points
+ Rotatability



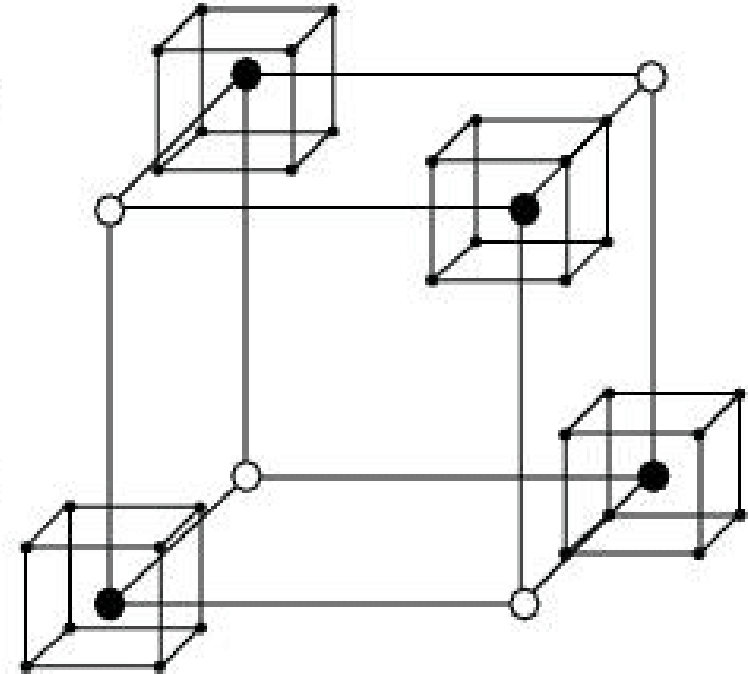
Box Behnken

Irregular Design Space



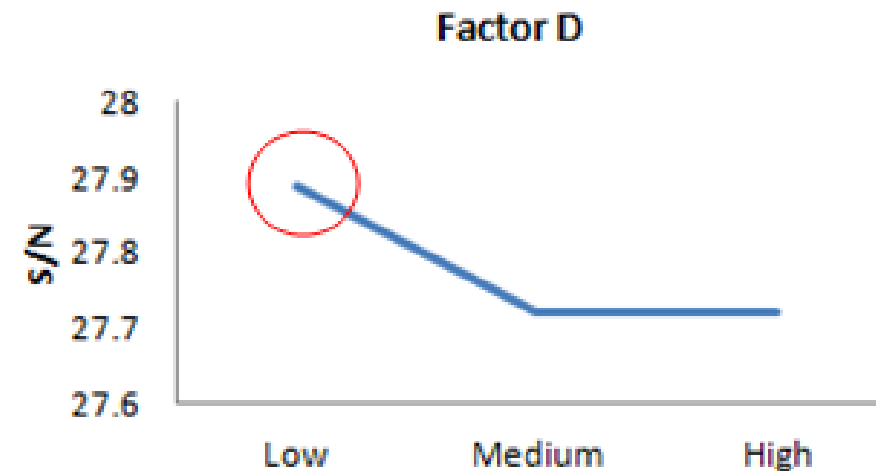
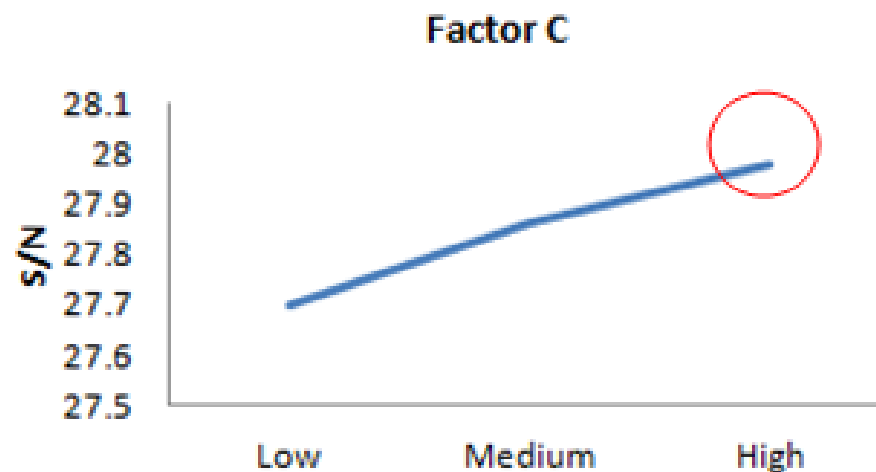
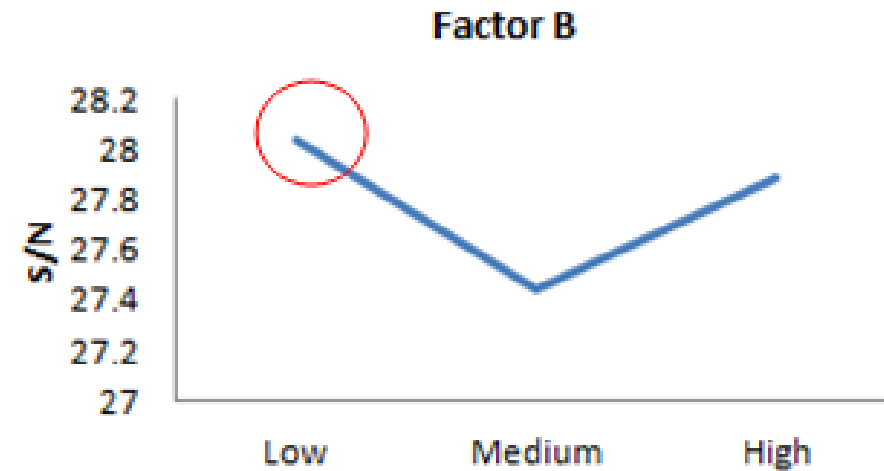
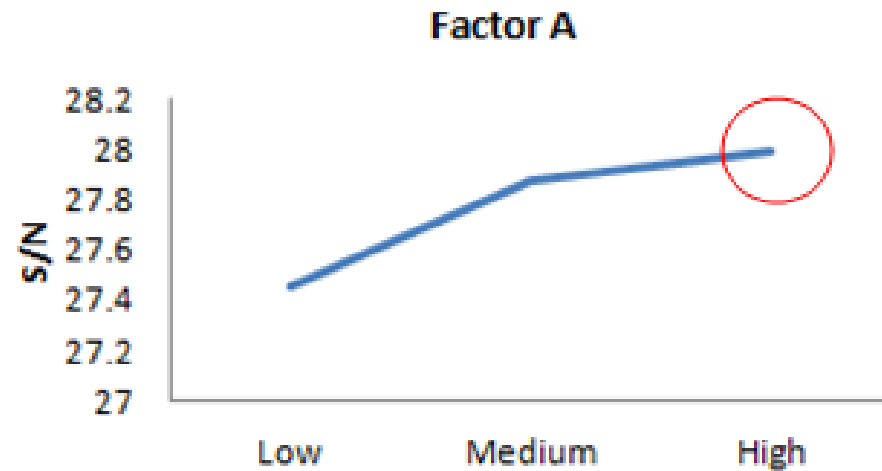
D-optimal design

“Robust Design”
Fractional + Deviation



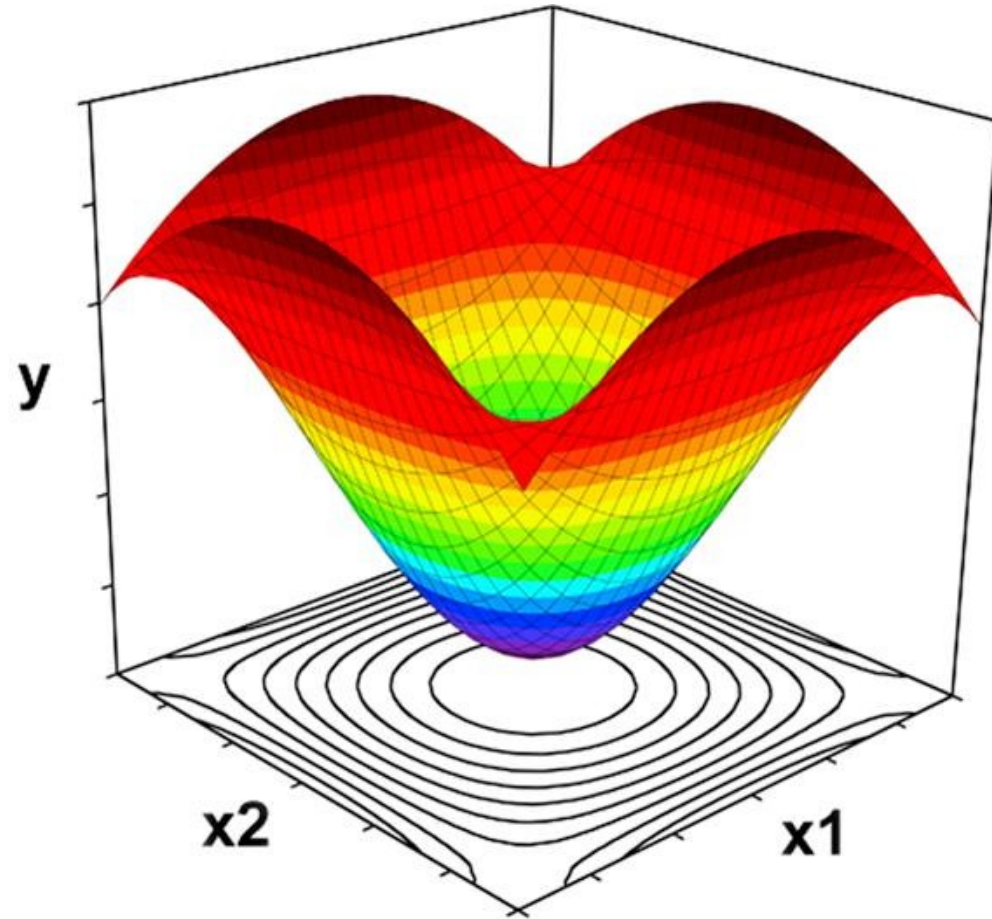
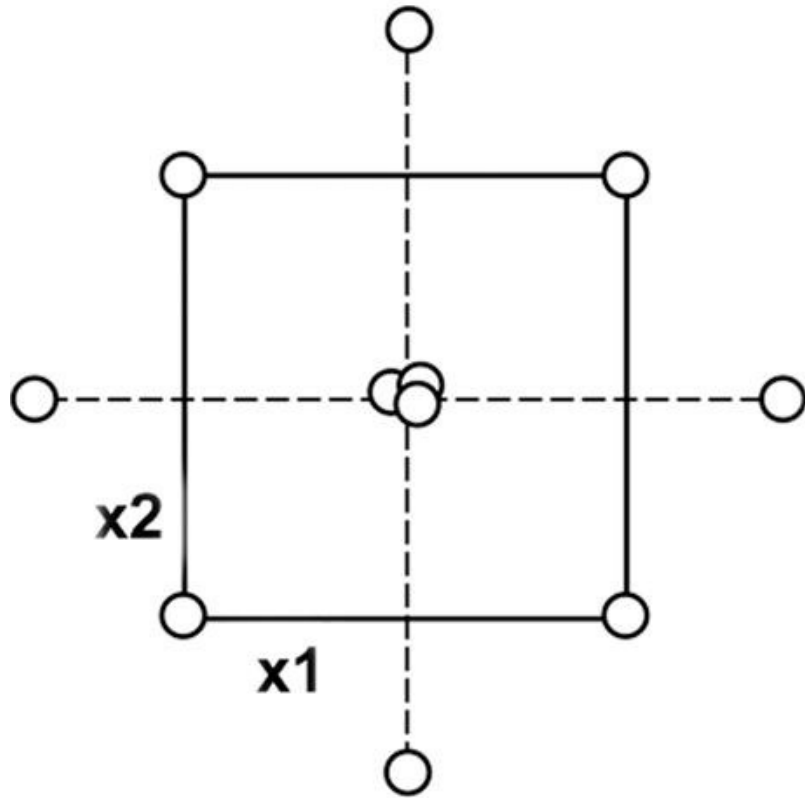
Taguchi design

Taguchi Method



RSM Response Surface Method

- DOE (실험 설계) → results → RSM (더 나은 조건 탐색)



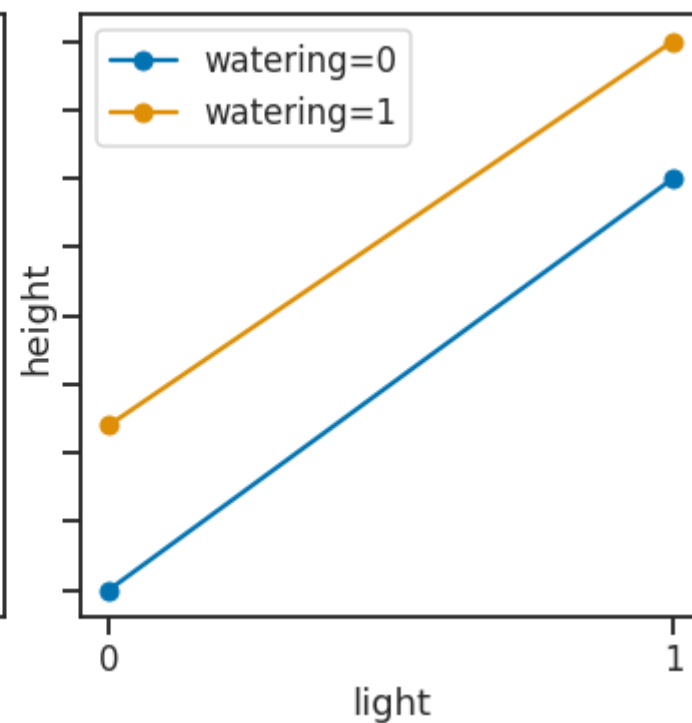
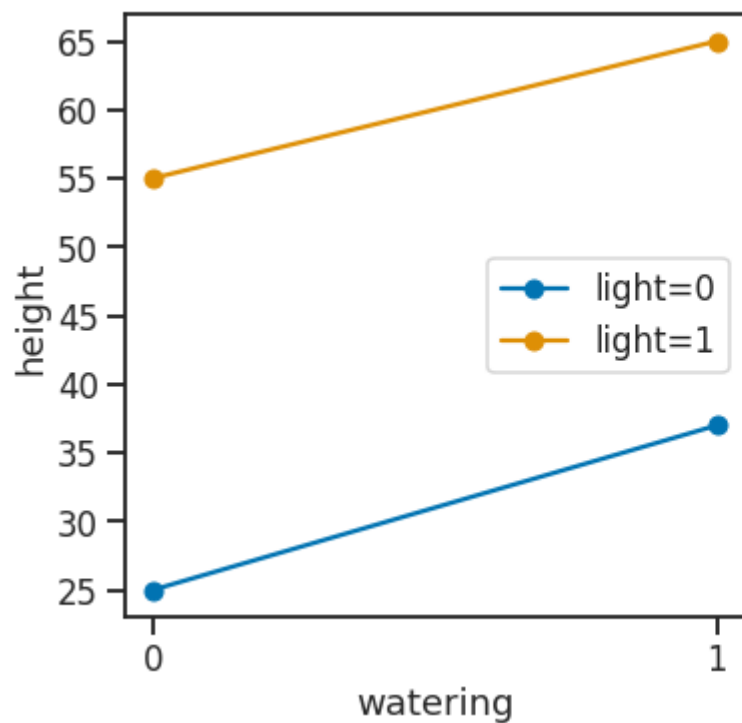
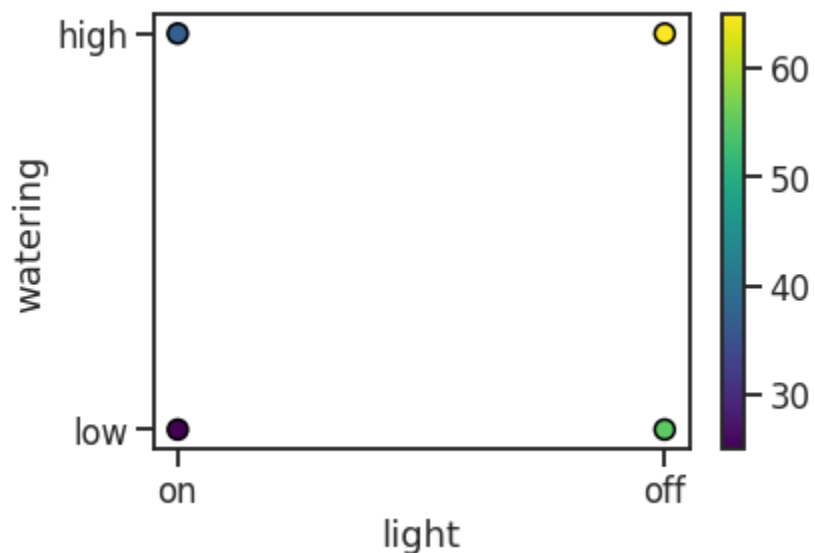
DOE + RSM

- 초간단 예제 : 식물 키우기

1차 선형 회귀

$$\text{height} = 45.5 + 5.5 \times \text{light} + 14.5 \times \text{watering}$$

	true	predict
0	55	54.5
1	65	65.5
2	25	25.5
3	37	36.5



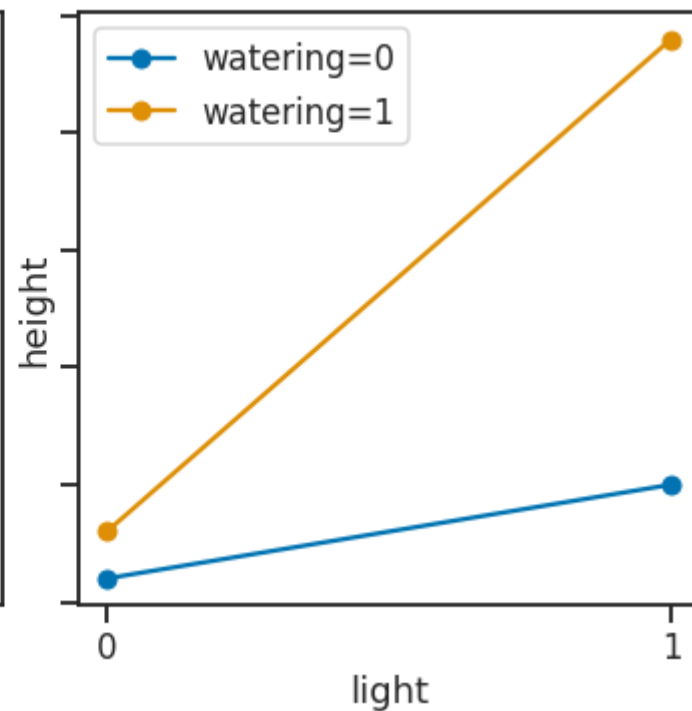
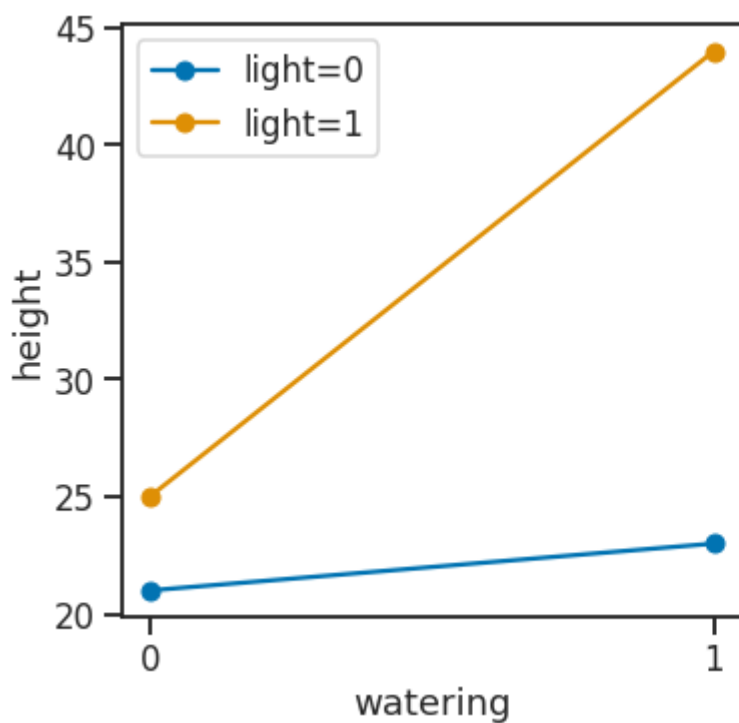
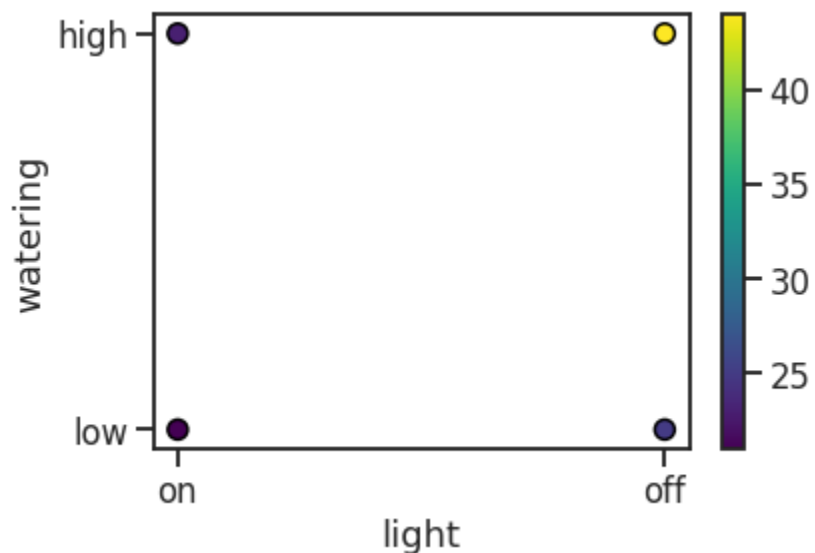
DOE + RSM

- 초간단 예제 : 식물 키우기
- Interaction 포함

1차 선형 회귀 + 교차항

$$\text{height} = 0 + 4 \times \text{light} + 2 \times \text{watering} + 17 \times \text{light} \cdot \text{watering}$$

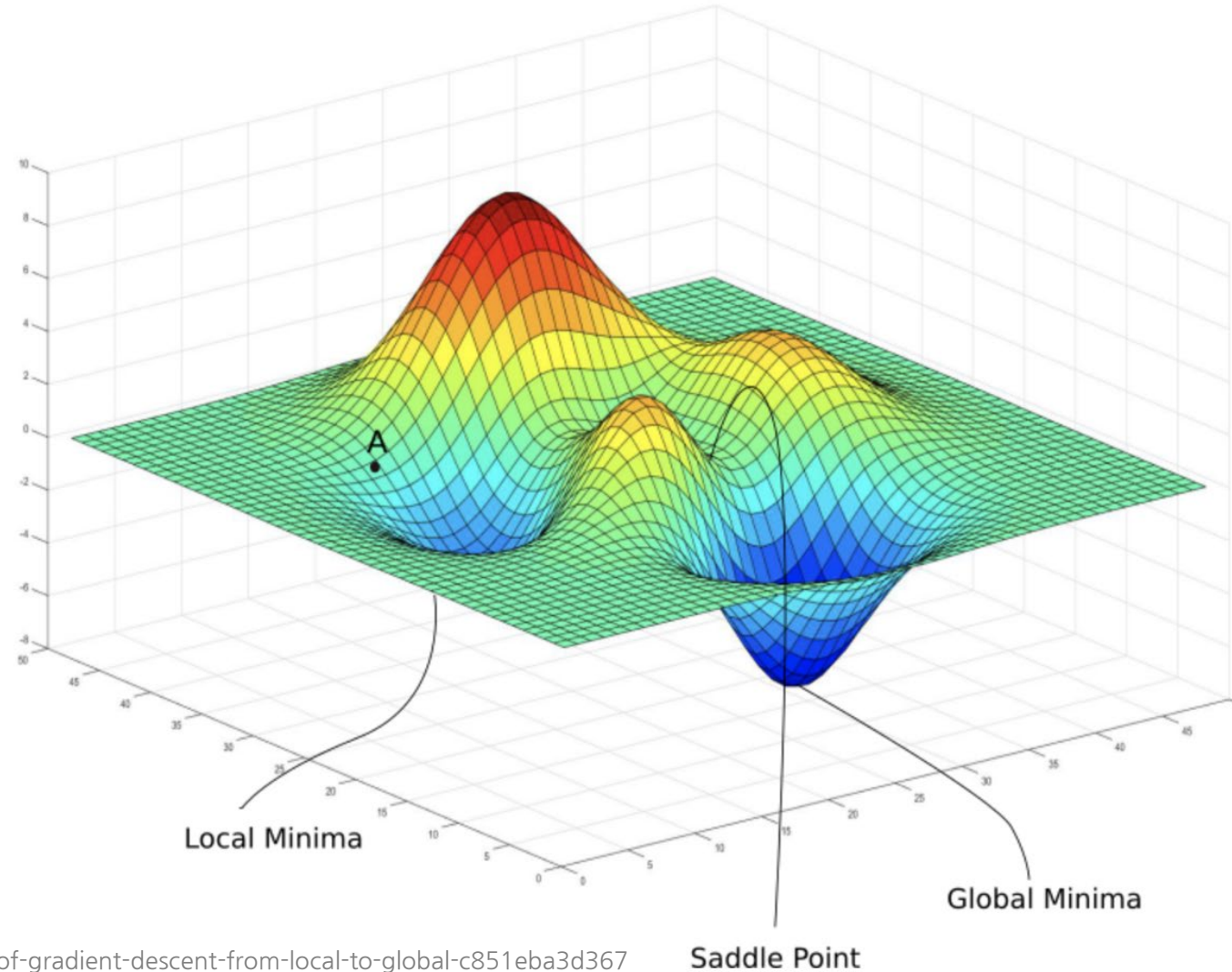
	true	predict
0	25	25.0
1	44	44.0
2	21	21.0
3	23	23.0



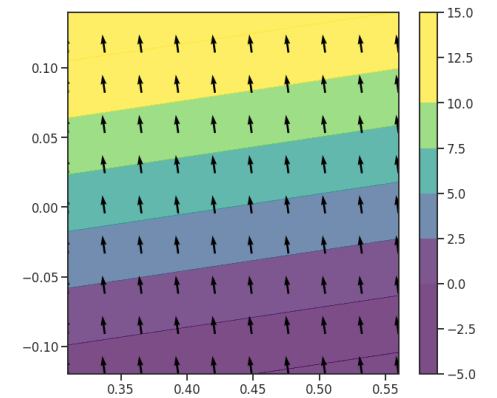
DOE + RSM

단점 : ① 시작점에 따라 결과가 다를 수 있음
② 인자 수가 많아지면 시간이 오래 걸림

- 2변수 최적화

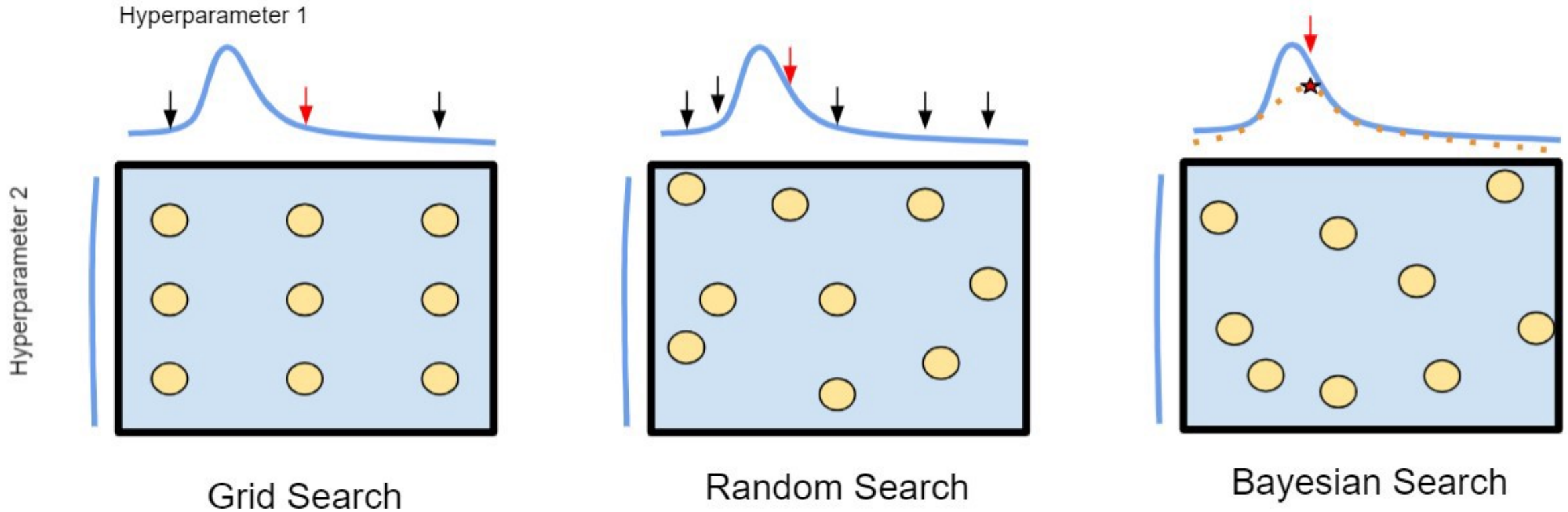


2차 다항식 fitting



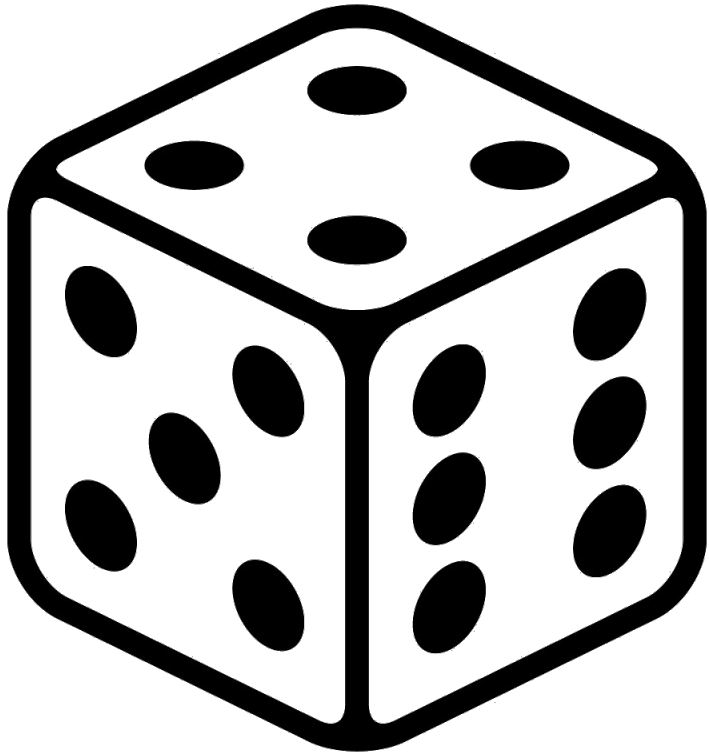
Hyperparameter Optimization

Comparison of Hyperparameter Search Methods (Optimal Selection in Red)



Frequentist vs Bayesianism

- 주사위를 한 번 던져 6이 나올 확률은?



“무한히 많이 던지면 알 수 있다” : Frequentist (빈도주의)

- 60번 던지면? 14번 = 23.3%
- 600번 던지면? 98번 = 16.3%
- 6000번 던지면? 989번 = 16.5%
- 60000번 던지면? 9965번 = 16.6%

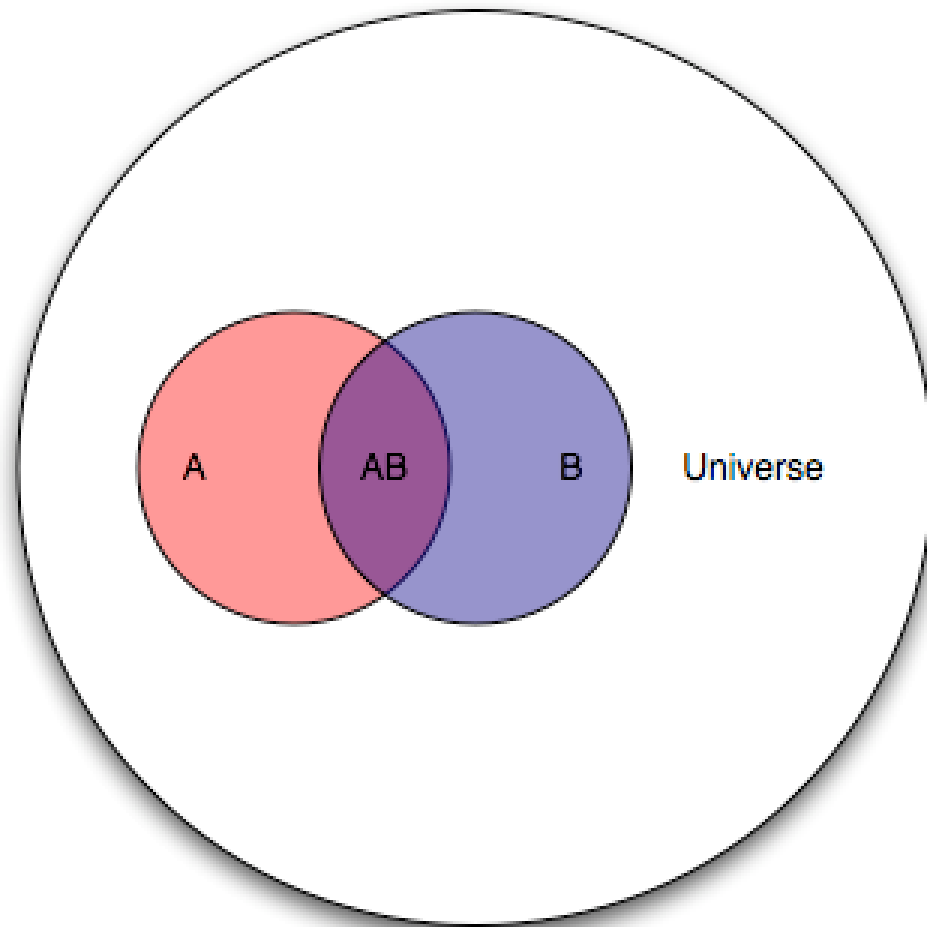
1/6에 수렴

“1/6 (16.7%)만큼 확신할 수 있다.” : Bayesianism (베이즈주의)

- “면이 여섯 개니까 공정한 주사위라면 1/6이 나올 거라 믿는다”



Bayes' Rule



$$P(A) = \frac{|A|}{|U|} \quad P(B) = \frac{|B|}{|U|} \quad P(AB) = \frac{|AB|}{|U|}$$

$$P(A|B) = \frac{|AB|}{|B|} = \frac{\frac{|AB|}{|U|}}{\frac{|B|}{|U|}} = \frac{P(AB)}{P(B)}$$

$$P(B|A) = \frac{P(AB)}{P(A)}$$

$$P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule

Posterior 사후확률

사건 B가 발생했을 때 사건 A가 발생할 확률

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Likelihood 가능도

사건 A가 발생했을 때
사건 B가 발생할 확률

Prior 사전확률

사건 B와 무관하게
사건 A가 발생할 확률

Evidence 증거

사건 A와 무관하게
사건 B가 발생할 확률

Bayes' Rule

- 어떤 병을 진단하는 키트의 양성 진단율이 90%라고 하자.
- 해당 증상이 있어 검사 결과 양성이 나왔다.
- 진짜 병에 걸려 있을 확률은 얼마일까?

병에 걸림

↓

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

↑

테스트가 양성

- 알고 있는 것 : 키트의 양성 진단율 $P(B|A) = 0.9$

- 알아야 할 것 : ① $P(A)$ = 병에 걸릴 확률

② $P(B)$ = 누구에게나 테스트가 양성이 나올 확률

$$\frac{0.9 \times 0.002}{0.9 \times 0.002 + 0.05 \times (1 - 0.002)} = 3.5\%$$

위양성률 = 5%

Bayes' Rule

- O. J. Simpson 살인 사건 (1994)



Bayes' Rule with Python

- 사기 동전 의심 : “나는 저 동전의 앞면 확률이 70%, 뒷면이 30%라고 믿는다.”
- 동전을 N 번 던져가면서 확인하자.
 - **Beta distribution** : 0~1을 가질 수 있는 베르누이 분포의 사후확률 분포

$$\text{Beta}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

✓
0초

“편견이 데이터로 교정됩니다”

Prior Beliefs

Posterior Beliefs

Evidence

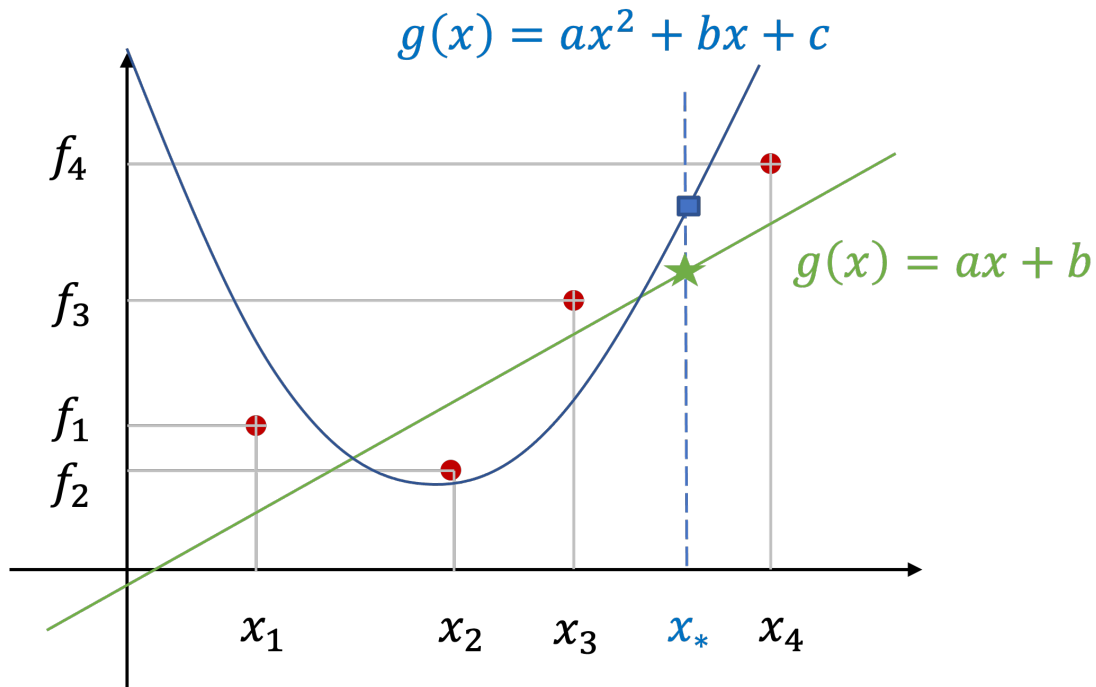


or

Gaussian Process Regression

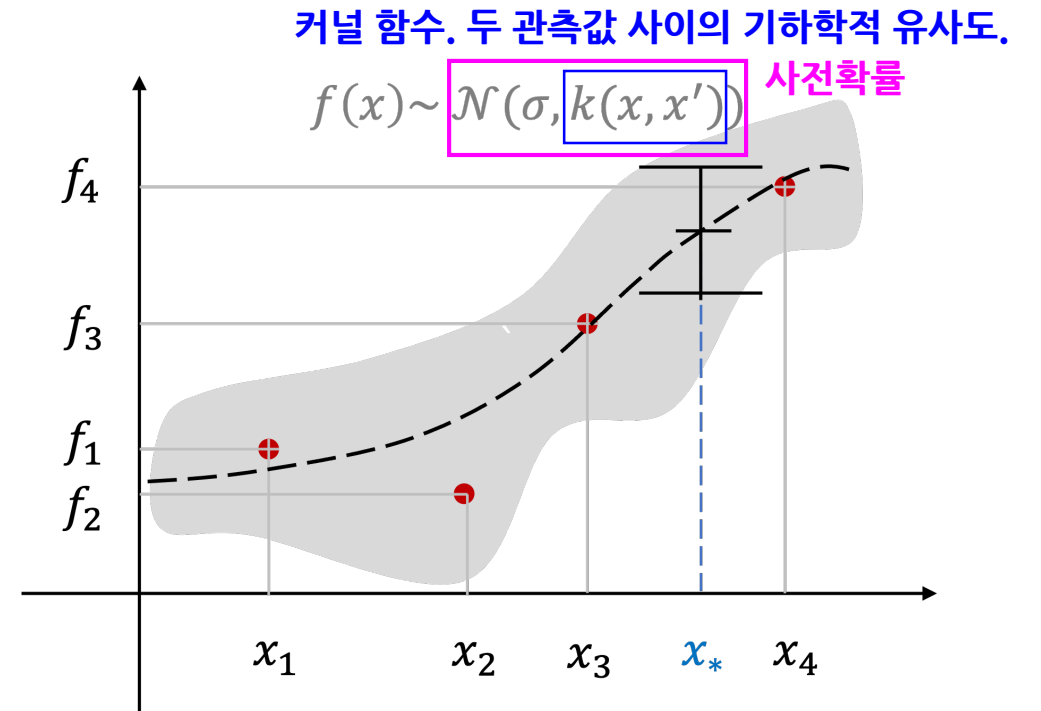
일반 Regression

목표 함수가 특정 형태를 이루고 있음.
최소제곱법 등으로 parameter 추정



Gaussian Process Regression

목표 함수가 특정 형태를 이루고 있지 있음.
확률 변수 : 값 대신 평균과 분산을 가짐

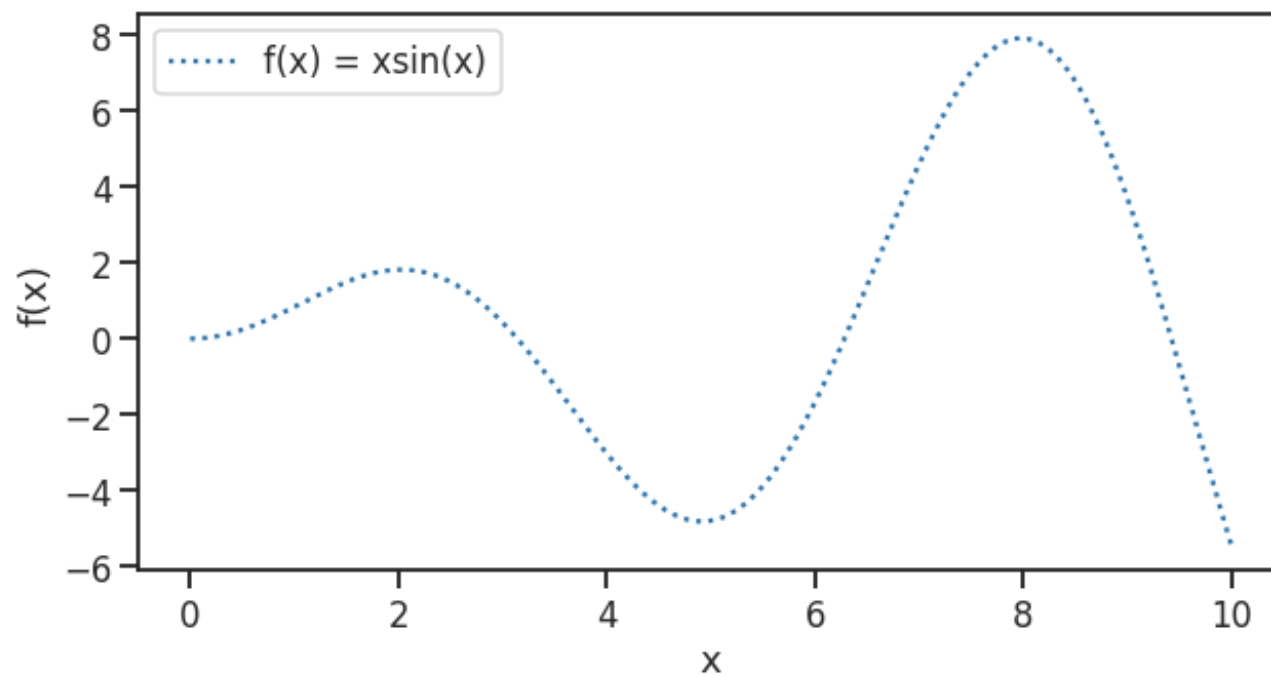


Gaussian Process

- Sample data (1,000)

```
1 X = np.linspace(start=0, stop=10, num=1_000).reshape(-1, 1)  
2 y = np.squeeze(X * np.sin(X))  
3  
4 fig, ax = plt.subplots(figsize=(10, 5))  
5 ax.plot(X, y, ":", label="f(x) = xsin(x)")  
6 ax.set(xlabel="x", ylabel="f(x)")  
7 ax.legend()
```

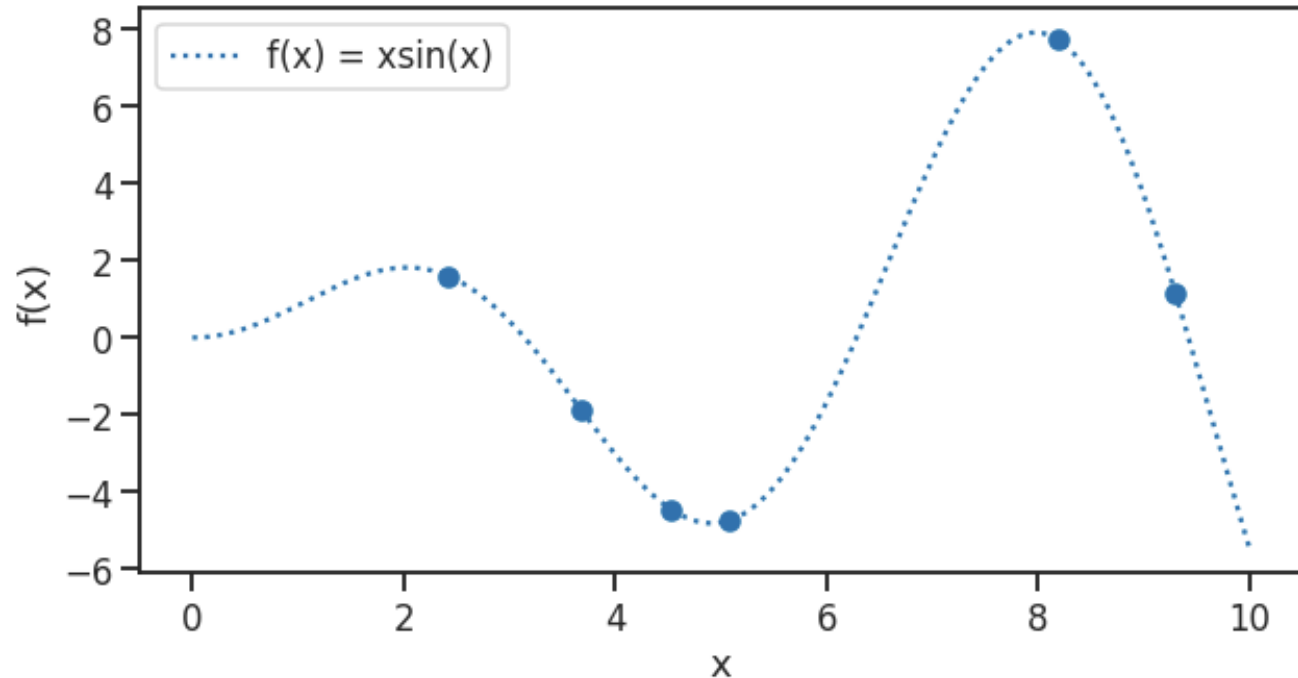
<matplotlib.legend.Legend at 0x7f9d87bfbad0>



Gaussian Process

- Sample data (1,000)
- Train set (6)

```
1 rng = np.random.RandomState(1)      # Random Number Generator
2 training_indices = rng.choice(np.arange(y.size), size=6, replace=False)
3 X_train, y_train = X[training_indices], y[training_indices]
4
5 ax.scatter(X_train, y_train)
6 display(fig)
```



Gaussian Process

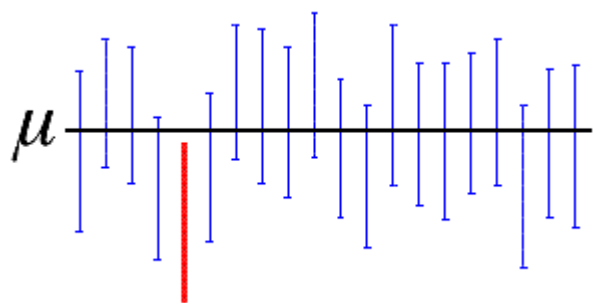
- Sample data (1,000)
- Train set (6)
- Set kernel & training

```
1 from sklearn.gaussian_process import GaussianProcessRegressor
2 from sklearn.gaussian_process.kernels import RBF
3
4 kernel = RBF(length_scale=1.0, length_scale_bounds=(1e-2, 2e2))
5 gaussian_process = GaussianProcessRegressor(kernel=kernel, n_restarts_optimizer=9)
6 gaussian_process.fit(X_train, y_train)
7 gaussian_process.kernel_

RBF(length_scale=0.687)
```

Gaussian Pro

- Sample data (1,000)
- Train set (6)
- Set kernel & training
- Prediction

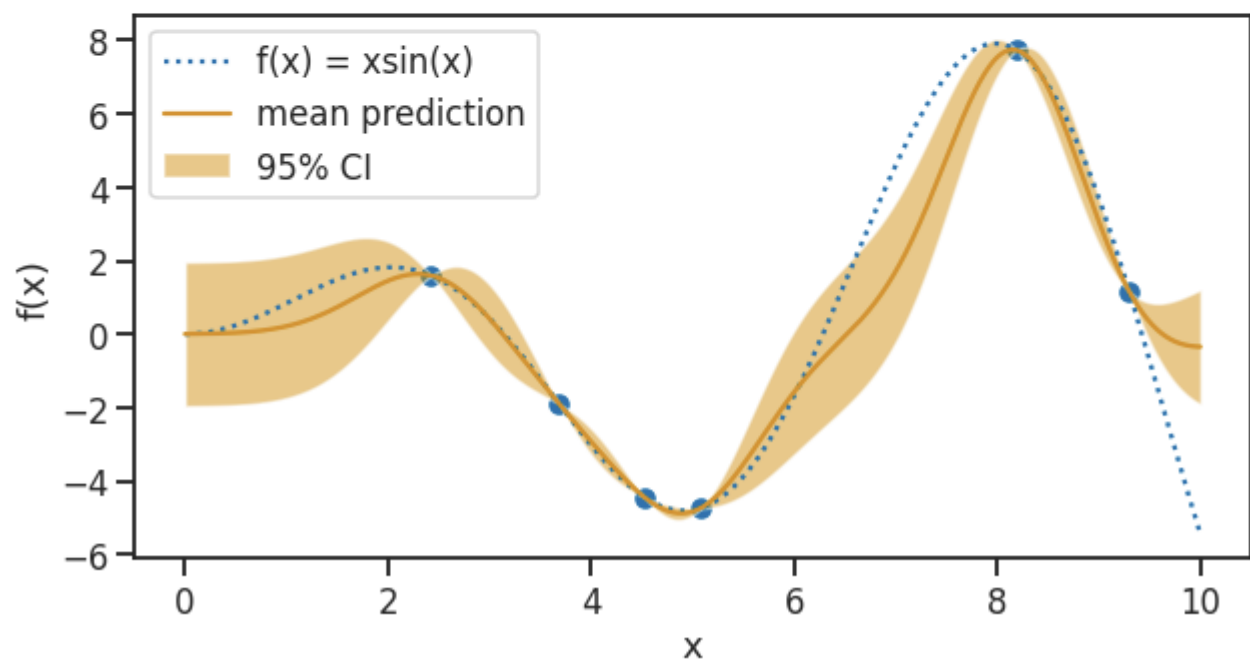


Confidence	Z
99%	2.576
98%	2.326
95%	1.960
90%	1.645

$$\text{Confidence Interval: } \bar{X} \pm Z \frac{s}{\sqrt{n}}$$

```
1 mean_prediction, std_prediction = gaussian_process.predict(X, return_std=True)
2
3 fig, ax = plt.subplots(figsize=(10, 5))
4 ax.plot(X, y, ":", label="f(x) = xsin(x)")
5 ax.scatter(X_train, y_train)
6 ax.plot(X, mean_prediction, label="mean prediction")
7 ax.fill_between(X.ravel(),
8                 mean_prediction-1.96*std_prediction,
9                 mean_prediction+1.96*std_prediction,
10                alpha=0.5,
11                label="95% CI")
12 ax.set(xlabel="x", ylabel="f(x)")
13 ax.legend()
```

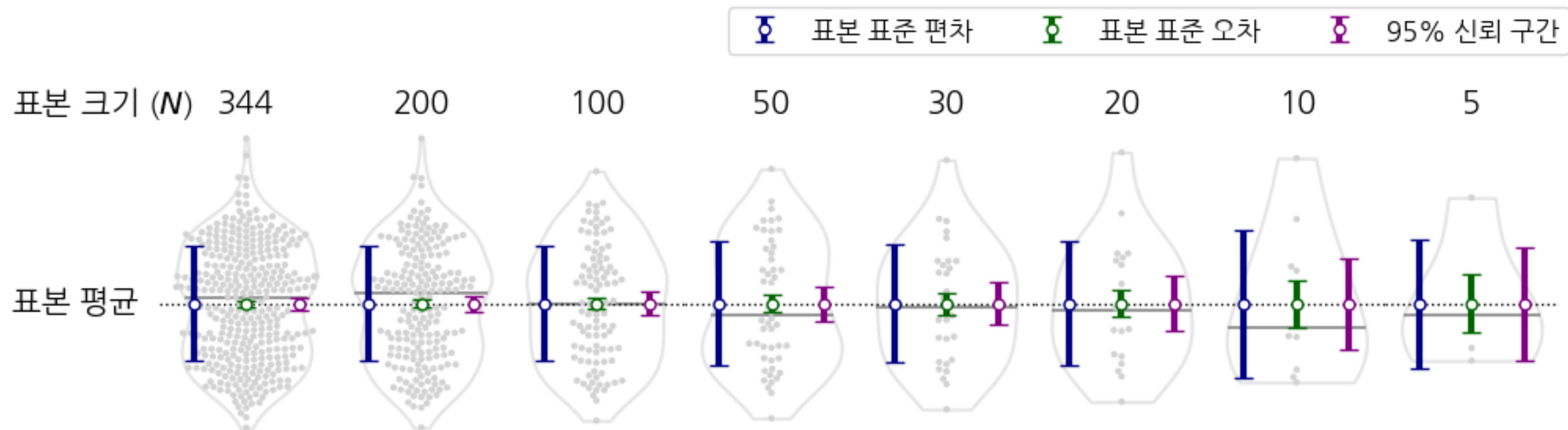
<matplotlib.legend.Legend at 0x7f9d874cbf90>



불확실성 지표

용어	기호, 약어	설명	수식
평균	μ	통계 집단을 모두 더한 후 크기(N, 데이터 수)로 나눈 산술 평균	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$
표준 편차	σ	통계 집단의 분산 정도. 각 데이터에서 평균을 뺀 값을 제곱하여 평균을 구한 후에 제곱근을 씌운 값.	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$
표준 오차	SE	표본 분포의 표준 편차.	$SE = \frac{\sigma}{\sqrt{N}}$
신뢰 구간	CI	모집단의 평균이 실제로 포함될 것으로 예측되는 범위.	$\mu - Z \frac{\sigma}{\sqrt{N}} \sim \mu + Z \frac{\sigma}{\sqrt{N}}$

불확실성 지표



Points of Significance

nature

nature > col

Collection

Statist

There is no
may be too

This collecti
work.

Nature Met
online plott

Image Cred

POINTS OF SIGNIFICANCE

Error bars

The meaning of error bars is often misinterpreted, as is the statistical significance of their overlap.

Last month in Points of Significance, we showed how samples are used to estimate population statistics. We emphasized that, because of chance, our estimates had an uncertainty. This month we focus on how uncertainty is represented in scientific publications and reveal several ways in which it is frequently misinterpreted.

The uncertainty in estimates is customarily represented using error bars. Although most researchers have seen and used error bars, misconceptions persist about how error bars relate to statistical significance. When asked to estimate the required separation between two points with error bars for a difference at significance $P = 0.05$, only 22% of respondents were within a factor of 2 (ref. 1). In light of the fact that error bars are meant to help us assess the significance of the difference between two values, this observation

THIS MONTH

Login

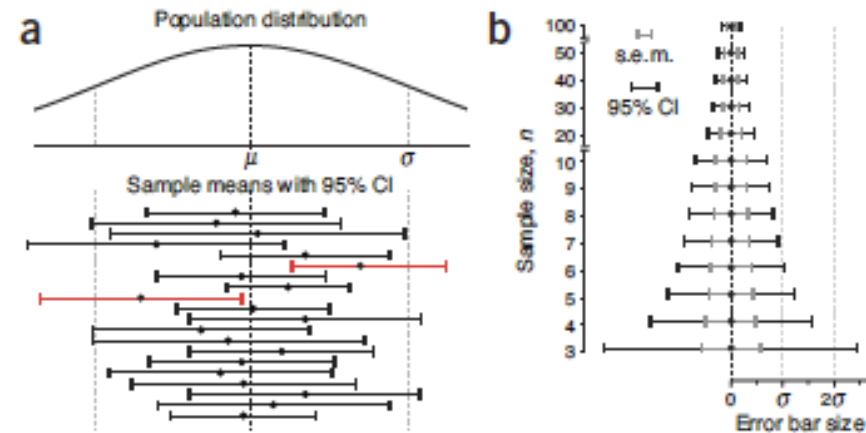


Figure 2 | The size and position of confidence intervals depend on the sample. On average, CI% of intervals are expected to span the mean—about 19 in 20 times for 95% CI. (a) Means and 95% CIs of 20 samples ($n = 10$) drawn from a normal population with mean μ and s.d. σ . By chance, two of the intervals (red) do not capture the mean. (b) Relationship between s.e.m. and 95% CI error bars with increasing n .

does not ensure significance, nor does overlap rule it out—it depends on the type of bar. Chances are you were surprised to learn this unintuitive result.

The first step in avoiding misinterpretation is to be clear about which measure of uncertainty is being represented by the error bar.

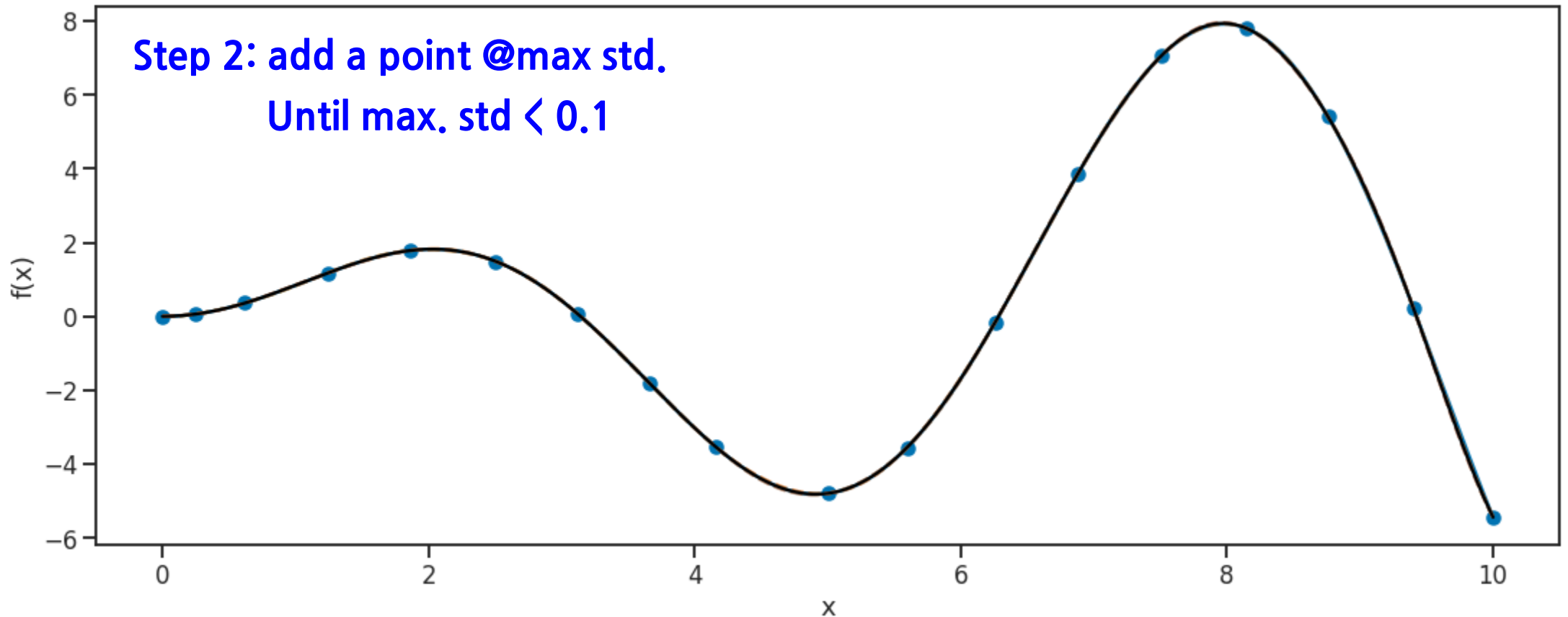
ed, when it

or of their

ude an

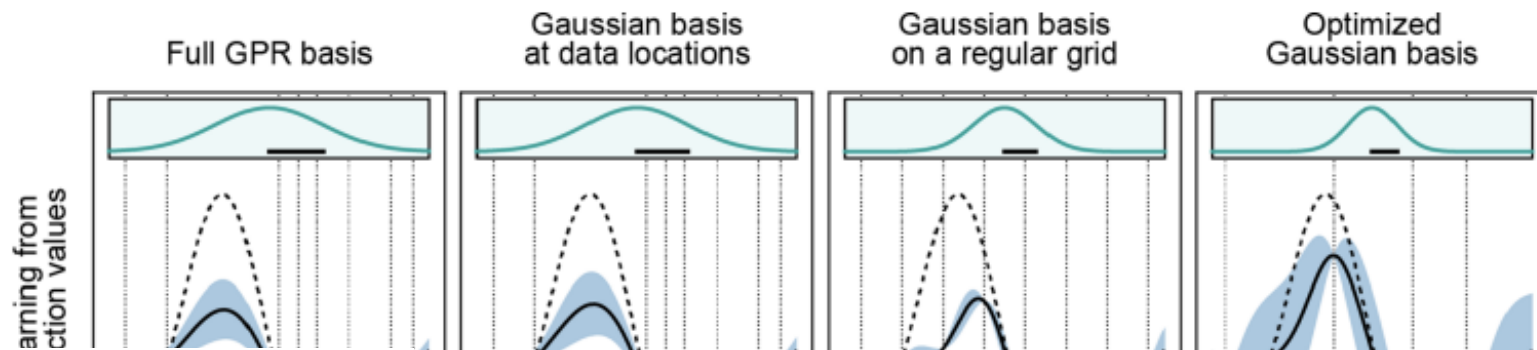
Gaussian Process

- Gaussian Process + DOE + iteration

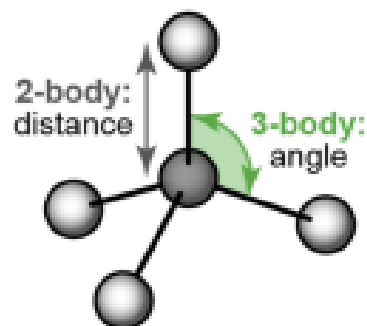


Reference

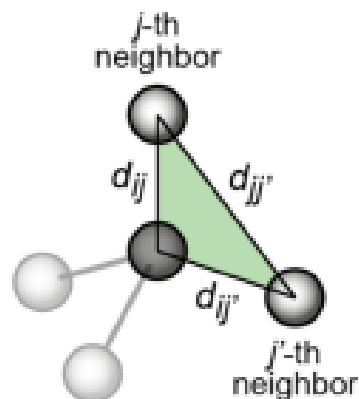
- 69 pages, 420 references



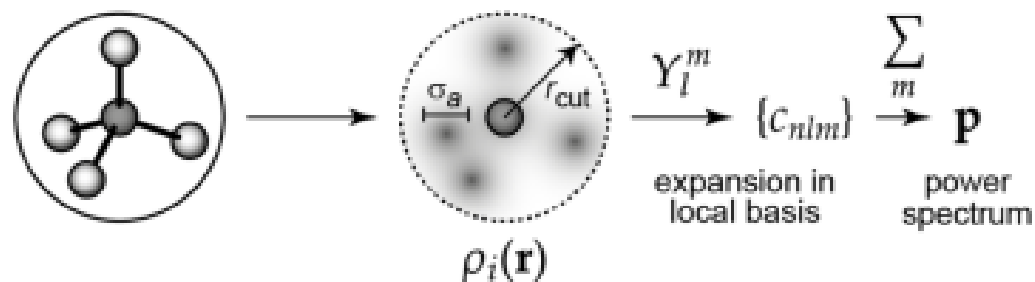
a Conventional N -body terms



b General 3-body descriptor



c Smooth Overlap of Atomic Positions (SOAP) descriptor



d

