

On Bayesian Optimization

Part 2 of 3

July 2, 2025

Donghun Lee

Department of Mathematics
Korea University

BO Tutorial Part 2: Current

1. 획득함수 + Information Theory

2. High-dimensional BO

3. Multi-objective BO

4. Closing Remarks

획득함수란?

- 베이지안 최적화는 평가 횟수가 매우 제한된 상황에서 사용됨
- **획득함수(acquisition function)**는 다음 평가할 점을 결정하는 역할
- 대표적 획득함수: UCB, EI, PI, TS 등

정보이론 기반 접근이란?

- 정보이론 기반 BO는 **최적값 또는 함수에 대한 불확실성 감소**를 목표로 설계
- 핵심 개념: **엔트로피(entropy)**를 줄이는 방향으로 탐색
- 핵심 빈자리: 무엇이 엔트로피를 줄일까?
- 특징: 탐색-활용 균형을 이론적으로 정당화 가능
- 특징2: 앞서 본 직관적인 획득함수들인 PI, EI 계열 대비 효과가 좋다는 연구결과 [1]

엔트로피?

Shannon Entropy (정보이론)

X 가 $p(x)$ 를 확률질량함수로 갖는 이산적 확률변수일때, 엔트로피 $H(X)$:

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \frac{1}{\log p(x)} = - \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

X 가 $p(x)$ 를 확률밀도함수로 갖는 연속적 확률변수일때, 엔트로피 $H(X)$:

$$H(X) = \int_{x \in \mathcal{X}} p(x) \frac{1}{\log p(x)} dx = - \int_{\mathcal{X}} p(x) \log p(x) dx$$

- 단위: bits ($\log_2$ 사용시) / nats ($\ln$ 사용시)
- $H(X)$ 는 변수 X 의 불확정성을 표현하는 (최적화에 매우 좋은) 함수

Entropy Search (ES) [2]

Journal of Machine Learning Research 13 (2012) 1809-1837

Submitted 12/11; Published 6/12

Entropy Search for Information-Efficient Global Optimization

Philipp Hennig

Christian J. Schuler

Department of Empirical Inference

Max Planck Institute for Intelligent Systems

Spemannstraße

72076 Tübingen, Germany

PHILIPP.HENNIG@TUEBINGEN.MPG.DE

CHRISTIAN.SCHULER@TUEBINGEN.MPG.DE

- 엔트로피를 사용해서 보다 똑똑하게 다음 샘플링을 결정할 수 있을까?

Entropy Search (ES) [2]

- 개념: 다음 샘플링은, 샘플링한 정보를 취합하는 것으로 최적점 x^* 의 엔트로피가 가장 많이 감소시키는 점 (x, y) 로 한다
- 획득함수가 가질 의미: (x, y) 를 관측해 x^* 의 엔트로피 감소를 최대화하는 점을 선택
- 획득함수: (x, y) 과 x^* 의 “mutual information”을 최대화

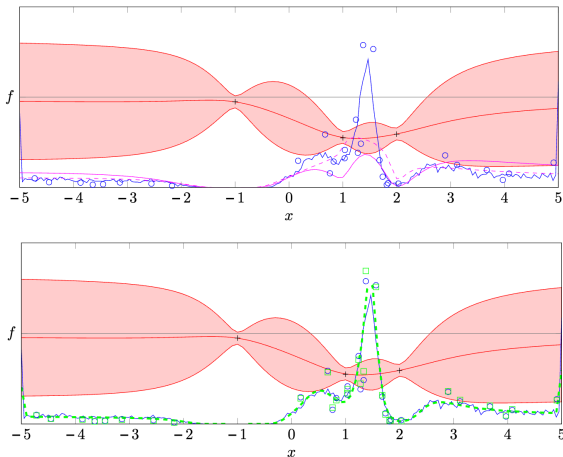
$$\arg \max_{x \in \mathcal{X}} I((x, y); x^* \mid \mathcal{D}_n)$$

- 엔트로피와 획득함수 간의 연계점

$$I((x, y); x^* \mid \mathcal{D}_n) = H[x^* \mid \mathcal{D}_n] - \mathbb{E}_{y \mid \mathcal{D}_n, x} [H[x^* \mid \mathcal{D}_n \cup (x, y)]]$$

Entropy Search (ES) [2]

- 자주색 실선: EI
- 자주색 점선: PI
- 빨간색 중간실선: Unknown $f(x)$
- 파란 실선: $\min f(x)$
확률의 Monte-Carlo
예상치
- 파란 원: PI기반으로
샘플링한 예시
- 녹색 점선: ES



ES의 매우 큰 문제

- 획득함수의 계산이 매우 난감함:

$$\arg \max_{x \in \mathcal{X}} \left\{ H[x^* \mid \mathcal{D}_n] - \underbrace{\mathbb{E}_{y \mid \mathcal{D}_n, x}}_{!} [H[x^* \mid \mathcal{D}_n \cup (x, y)]] \right\}$$

- $\mathbb{P}[y \mid \mathcal{D}_n, x]$ 를 사용해서 (x, y) 를 샘플링하여 !를 계산해야 함
 - 고차원 (x 가 매우 많아짐)에서 계산량/샘플요구량 폭발
 - noisy setting (!의 추정값이 흔들림)에서 계산량/샘플요구량 폭발

ES를 개선해봅시다

- ES 획득함수에서 ! 이 문제였으니

$$I((x, y); x^* | \mathcal{D}_n) = H[x^* | \mathcal{D}_n] - \overbrace{\mathbb{E}_{y|\mathcal{D}_n, x}}^! [H[x^* | \mathcal{D}_n \cup (x, y)]]$$

- 정보이론 한토막: $I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
- = 를 사용해서 뒤집는다

$$\underbrace{H[x, y | \mathcal{D}_n]}_{(A)} - \underbrace{\mathbb{E}_{x^*|\mathcal{D}_n}}_{!!} \left[\underbrace{H[x, y | \mathcal{D}_n, x^*]}_{(B)} \right]$$

- !!는 더이상 (x, y) 샘플링 안해도 되네? (대신 $\mathbb{P}[x^*|\mathcal{D}_n]$ 을 써야됨)

ES를 개선해봅시다 (계속)

- 다듬기 작업

$$\underbrace{H[x, y | \mathcal{D}_n]}_{(A)} - \mathbb{E}_{x^* | \mathcal{D}_n} \left[\underbrace{H[x, y | \mathcal{D}_n, x^*]}_{(B)} \right]$$

- (A,B): 사실 x 는 고르는거니 변수가 아님. 하지만 y 에 영향을 준다. 그러니 조건부로 이동

$$\underbrace{H[y | \mathcal{D}_n, x]}_{(A2)} - \underbrace{\mathbb{E}_{x^* | \mathcal{D}_n} [H[y | \mathcal{D}_n, x^*, x]]}_{(B2)}$$

- (A2): GP에서 바로 계산이 가능함 (posterior marginal i.e. $\mathbb{P}[y | \mathcal{D}_n, x]$)
- (B2): $H[y | \mathcal{D}_n, x^*, x]$ 를 계산하기. 먼저 x^* 를 GP에서 $\mathbb{P}[x^* | \mathcal{D}_n]$ 사용해 샘플링하고 $\mathbb{E}_{x^* | \mathcal{D}_n}$ 는 “expectation propagation” [3] 으로 근사

Predictive Entropy Search (PES) [1]

Predictive Entropy Search for Efficient Global Optimization of Black-box Functions

José Miguel Hernández-Lobato

jmh233@cam.ac.uk

University of Cambridge

Matthew W. Hoffman

mwh30@cam.ac.uk

University of Cambridge

Zoubin Ghahramani

zoubin@eng.cam.ac.uk

University of Cambridge

- ES의 획득함수 계산을 더 쉽게 할 방법이 있다 (GP를 사용하니까)

Predictive Entropy Search (PES) [1]

- ES를 개선: 획득함수 계산을 더 쉽게. $\mathbb{P}[y|\mathcal{D}_n, x]$ 대신에 $\mathbb{P}[x^*|\mathcal{D}_n]$ 사용해 근사
- **주의: 근사값임** ($\mathbb{P}[x^*|\mathcal{D}_n]$ 는 $\mathcal{D}_n$ 가 주어졌을 때의 최적값 x^* 의 **예측**분포)
- 획득함수: $I((x, y); x^* | \mathcal{D}_n)$ 를 $\mathbb{P}[x^*|\mathcal{D}_n]$ 사용하여 근사한 값을 최대화

$$\arg \max_{x \in \mathcal{X}} \left\{ \underbrace{H[y | \mathcal{D}_n, x]}_{(A2)} - \underbrace{\mathbb{E}_{x^*|\mathcal{D}_n} [H[y | \mathcal{D}_n, x^*, x]]}_{(B2)} \right\}$$

- 결과적으로 PES는, predictive 분포를 써서 (B2)를 쉽게 계산하는 ES

Max-value Entropy Search (MES) [4]

Max-value Entropy Search for Efficient Bayesian Optimization

Zi Wang¹ Stefanie Jegelka¹

- 최적 선택 x^* 의 엔트로피 말고, **최대값** $y^* = f(x^*)$ 의 엔트로피를 줄이는건 어때?

Max-value Entropy Search (MES) [4]

- 최적 선택 x^* 의 엔트로피 말고, **최대값** $y^* = f(x^*)$ 의 엔트로피를 줄이는건 어때?
- 획득함수: (x, y) 과 y^* 의 “mutual information”을 최대화

$$\arg \max_{x \in \mathcal{X}} I((x, y); y^* \mid \mathcal{D}_n)$$

- PES에서 사용한 뒤집기와 같이 접근하면:

$$\arg \max_{x \in \mathcal{X}} \left\{ \underbrace{H[y \mid \mathcal{D}_n, x]}_{(A2)} - \underbrace{\mathbb{E}_{y^* \mid \mathcal{D}_n} [H[y \mid \mathcal{D}_n, y^*, x]]}_{(B2')} \right\}$$

- (B2')을 근사할때 Gumbel분포 (cf. Fisher-Tippett-Gnedenko theorem) 사용

Max-value Entropy Search (MES) [4]

이론적 특징: 함수값 사용 획득함수들과 동치 조건 존재

- MES, (B2')을 단일 샘플로 근사할 때
- GP-UCB, E-E parameter $\beta = \min_{x \in \mathcal{X}} \frac{y^* - \mu_n(x)}{\sigma_n(x)}$ 사용할 때
- PI, $\theta = y^*$ 사용할 때

실험적 특징: 다양한 방면의 실용성 높음

- 간단하고 병렬화가 쉬우며 (다차원 x 대신 1차원 y 사용), 실성능도 우수
- Gumbel 샘플링 기반으로 근사 (다차원 x 문제(조정 불가)를 버리고, 샘플링 수 문제(조정 가능)를 가져옴)
- PS. 다차원 x 에 대한 대응이 가능할까 (Hint: Add-MES [4])?

엔트로피 사용한 획득함수 3종 비교/요약

방법	엔트로피의 대상	계산량	적용성
ES	x^* 위치	높음	이론적 연구
PES	x^* 위치 (근사)	중간	실용적 개선
MES	$f(x^*)$ 값	낮음	가장 실용적

- 정보이론적 BO는 엔트로피를 사용해 불확실성을 계측한다는 단단한 이론적 기반 제공
- 현실은 근사를 하기때문에 계산 비용 및 surrogate model 품질에 민감
- 실전에서는 현실과 타협을 본 MES와 PES가 널리 사용됨

BO Tutorial Part 2: Current

1. 획득함수 + Information Theory

2. High-dimensional BO

3. Multi-objective BO

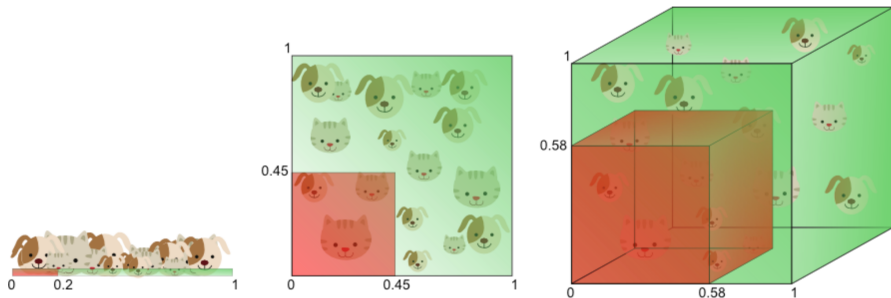
4. Closing Remarks

고차원에서 BO?

- 베이지안 최적화는 일반적으로 x 가 10-30차원 이하의 문제에 적합
 - 일단, 샘플이 N 개일때 predictive분포 계산을 위해 $K \in \mathbb{R}^{N \times N}$ 의 역행렬 필요
 - x 의 차원이 커지면 샘플이 정말 많이 필요함
- 고차원(high-dimensional) 문제에서는 BO예측성능 저하 발생
- 굳이 해야 된다면, 고차원 BO의 핵심: 고차원 공간에서의 **효율적인 탐색 전략** 설계

고차원이 되면 문제가 많다 1

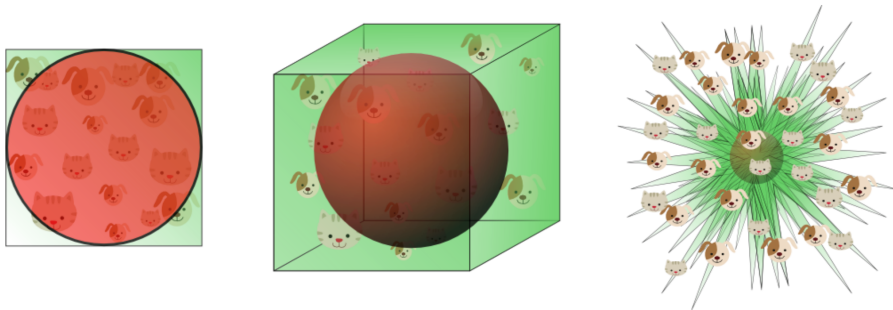
- 차원의 저주 (Curse of Dimensionality): 탐색공간 $\mathcal{X}$ 가 희소(sparse)해짐



- 동일한 정확도를 갖도록 샘플링하려면, 계산 비용 증가 (특히 획득함수 최적화때)

고차원이 되면 문제가 많다 2

- 차원의 저주 (Curse of Dimensionality): 탐색공간 $\mathcal{X}$ 가 희소(sparse)해짐



- Gaussian Process의 학습수렴 및 불확실성 추정값 정확도 하락

접근법 1: 가산 모델 (Additive Models)

- 최적점을 찾아낼 고차원 함수 f 를
저차원 함수 f_i 들의 합으로
표현된다고 가정

$$f(x) = \sum_{i=1}^k f_i(x_{S_i})$$

- 중요: 각 저차원 함수 f_i 의 변수는
원래 f 의 변수보다 차원이 낮음
- 예시 알고리즘: Add-MES [4],
Add-GP-UCB [5], EBO [6]

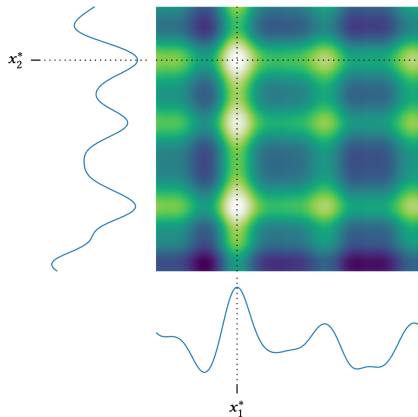


Figure: $k = 2$ 개의 GP의 합으로 표현된다면?

Add-GP-UCB [5]

High Dimensional Bayesian Optimisation and Bandits via Additive Models

Jeff Schneider

Carnegie Mellon University, Pittsburgh, PA, USA

SCHNEIDE@CS.CMU.EDU

- GP-UCB를 가산 모델로 만들 수 있다?

Add-GP-UCB [5]

- 핵심: 전체 함수를 d 차원이 아닌 k 개의 저차원 블록으로 구성하고, GP-UCB 틀에 넣음
- 장점1: 각 블록에서 개별적으로 BO를 수행하므로 병렬 처리 가능
- 장점2: GP-UCB와 마찬가지로, no-regret 특성 가짐
- 문제: 구성은 어떻게 정하지? *a priori*로 주거나, 자체학습(!)

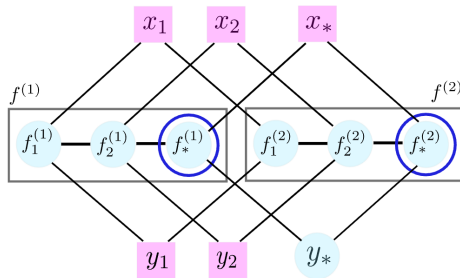


Figure: $k = 2$ 개의 블록으로 쪼갠 예시

접근법 2: 능동적 부분공간 탐색

- 함수 f 가 실제로는 특정 **저차원 선형 부분공간**에서만 변화할 수 있음
- 핵심 아이디어**: 입력 차원을 축소할 수 있나? e.g. $x = Az$, where $A \in \mathbb{R}^{D \times d}$, $d \ll D$
- 보다 구체적으로: 목적함수 f 가 고차원 도메인 $X \subset \mathbb{R}^D$ 에서 정의되었다면, 아래와 같은 $A \in \mathbb{R}^{D \times d}$ 가 있을/찾을/만들 수 있을까? (그리고 상응하는 g 도?)

$$f(x) = g(Az)$$

- 예시 알고리즘: Active Learning of Linear Embedding [7], REMBO [8], SAASBO [9]

Active Learning of Linear Embedding [7]

Active Learning of Linear Embeddings for Gaussian Processes

Roman Garnett
University of Bonn
Römerstraße 164
53117 Bonn, Germany
rgarnett@uni-bonn.de

Michael A. Osborne
University of Oxford
Parks Road
Oxford OX1 3PJ, UK
mosb@robots.ox.ac.uk

Philipp Hennig
MPI for Intelligent Systems
Spemannstraße
72076 Tübingen, Germany
phennig@tue.mpg.de

Active Learning of Linear Embedding [7]

- 보다 구체적으로: 목적함수 f 가 고차원 도메인 $\mathcal{X} \subset \mathbb{R}^D$ 에서 정의되었다면, 아래와 같은 $A \in \mathbb{R}^{D \times d}$ 가 있을/찾을/만들 수 있을까? (그리고 상응하는 g 도?)

$$f(x) = g(Az)$$

- 있긴 한거같고, 되기도 하네요?

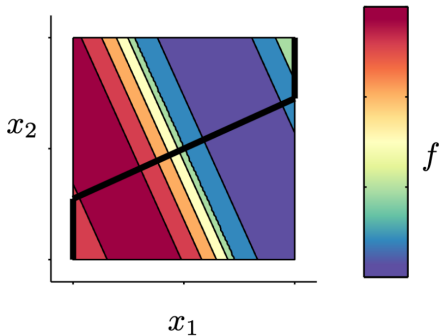


Figure: $\mathbb{R}^2$ 상의 함수를 잘 표현하는 1차원 찾기

Bayesian Optimization in a Billion Dimensions via Random Embeddings

Ziyu Wang

Department of Computer Science, University of Oxford

ZIYU.WANG@CS.OX.AC.UK

Frank Hutter

Department of Computer Science, University of Freiburg

FH@CS.UNI-FREIBURG.DE

Masrour Zoghi

Department of Computer Science, University of Amsterdam

M.ZOGHI@UVA.NL

David Matheson

Department of Computer Science, University of British Columbia

DAVIDM@CS.UBC.CA

Nando de Freitas

Department of Computer Science, University of Oxford

Canadian Institute for Advanced Research

NANDO@CS.OX.AC.UK

REMBO [8]: Random Embedding

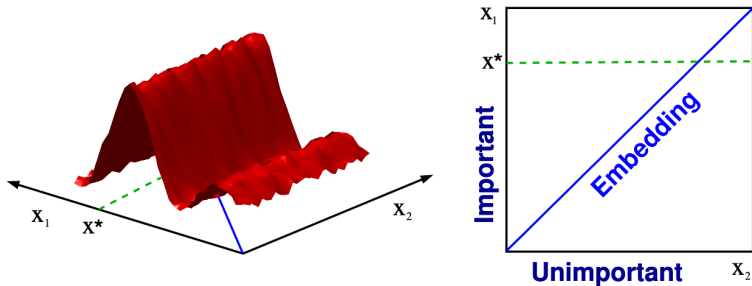
기본 아이디어:

- 주어진 고차원 공간 $\mathcal{X}$ 을 **랜덤** 선형맵으로 저차원으로 투영 (할 수 있으니까!)
- 투영된 공간에서 BO를 수행 (저차원이니 더 쉬우니까!)
- 찾은 최적점 결과를 원래 공간에 매핑 (어..?! 이게 된다고?)
- Food for thought: 이게 안 되는 상황도 분명 있을텐데요?

요약정리:

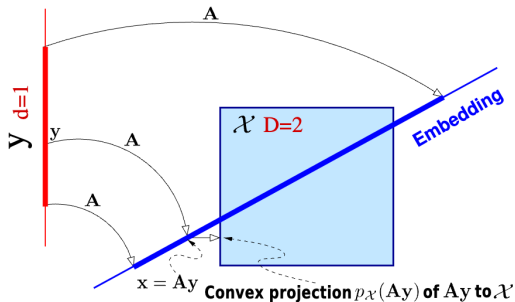
- 되긴 됩니다. 은근히 잘 되던데요?
- 이게 되는 근거에 대한 이론적인 이야기도 할 수 있습니다!
- 제약사항: 고차원 공간의 유효영역을 설정하는데, 그 바깥에 있는 점들은 아쉽습니다

REMBO [8]: 기본 원리



- 주어진 고차원 공간 $\mathcal{X}$ 을 **랜덤** 선형맵으로 저차원으로 투영하고나서 BO를 한다?
- 그래도 x^* 를 찾을수 있다..? 이게 된다고?! (특정 조건 하에서)

REMBO [8]: 아이디어의 구체화



- f 가 고차원 $\mathcal{X} \subset \mathbb{R}^D$ 에서 정의되었다면, $f(x) = g(Ay)$ 를 만족하는 g 를 찾게 해줄 좋은 $A \in \mathbb{R}^{D \times d}$ 가 있을텐데 ...
- 그냥 A 를 랜덤으로 만들고, Ay 중에서 바깥에 빠져나간부분은 $\mathcal{X}$ 안으로 우겨넣자!

High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces

David Eriksson¹

Martin Jankowiak²

¹Facebook, Menlo Park, California, USA

²Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

SAASBO [9]

가능성 타진

- 아마도 현실 문제에서는 **일부 중요한 차원만**이 목적함수에 큰 영향을 줄것같은데..
- 최적화를 하면서, 중요한 차원을 자동으로 찾아낼수 있을까?

문제 셋팅 리뷰

- 목적: 함수 $f : \mathcal{X} \rightarrow \mathbb{R}$ 의 최대값 추정
- $\mathcal{X} = [0, 1]^D$ 는 고차원 공간
- $y = f(x) + \epsilon$ 형태로, 노이즈 포함한 관측만 가능
- 도함수 ∇f 정보 없음, $f(x)$ 는 볼록함수가 아닐수도 있음

SAASBO 핵심 원리

희소성 가정 (Sparsity Assumption)

x 를 구성하는 대부분의 차원은 $f(x)$ 에 거의 영향을 주지 않을거다

- 이 희소성을 알고리즘 단에서 **명시적으로** 모델링
- **S**parse **A**xis-**A**ligned **S**ubspaces (**SAAS**) prior 을 GP에 적용
 - Axis-aligned 부분이 매우 중요한 특징임: 차원들을 뒤섞지 말고, 보존해서 선택하겠다!

SAAS Prior를 적용한 GP 모델

- Automatic Relevance Determination (ARD) 커널을 살짝 뜯어고쳐서:

$$k(x, x') = \sigma_k^2 \exp \left(- \sum_{d=1}^D \rho_d^2 (x_d - x'_d)^2 \right)$$

- SAAS prior: $\rho_d \sim \text{HalfCauchy}(\tau)$, $\tau \sim \text{HalfCauchy}(\alpha)$
 - “Shrinkage” 특성: 대부분의 ρ_d 는 작지만 ($\rho_d \lesssim \tau$), 일부는 크게 유지 (Half Cauchy 분포)
 - 위 특성은 하이퍼파라미터 τ 로 조정
- 문제: α 는 하이퍼-하이퍼파라미터. 기본값은 $\alpha = 0.1$
- 문제2: 그 외 여러 구성 가정들 (e.g. $\sigma_k^2 \sim \text{Log-Normal}(0, 10^2)$)

SAASBO 특징 정리

- 유관 차원들을 선택하고, 중요도를 학습할 수 있도록 한 Hierarchical Bayesian 구성
- 특징/장점: 차원을 뭉개버리지 않고, 주어진 차원 중에서 일부만을 선별할 수 있음
- Bayesian의 장점과 단점을 더욱 강렬하게 가짐
- BO 획득함수로는 TI를 사용 (계산은 Hamiltonian Monte Carlo와 Sobol sequence sampling, L-BFGS-B를 사용해 근사)
- 하이퍼파라미터 튜닝을 통해 수백차원 규모에도 잘 작동

접근법 3: 잠재 임베딩 기반 BO

지금까지

- 가산 모델: f 를 구성하는 BO들을 저차원 기반으로 쪼개버린 뒤 개별적으로 BO 수행
- 능동적 부분공간 탐색: $\mathcal{X}$ 보다 작은 차원으로 보내놓고 BO 수행. 선형변환이나 차원축 선택을 생각함

자연스러운 생각의 흐름: 비선형이 더 파워풀할듯한데?

- 오토인코더 등의 비선형 고차원 데이터를 잠재 공간(latent space)으로 변환
- 잠재 공간에서 BO 수행하고 원래 입력 복원하는 비선형함수도 학습하면 되겠지?
- 사례: VAE-DGP [10], VAE-BO [10], Deep Kernel Learning [11]

VAE-DGP [12]

VARIATIONAL AUTO-ENCODED DEEP GAUSSIAN PROCESSES

Zhenwen Dai, Andreas Damianou, Javier González & Neil Lawrence

Department of Computer Science,

University of Sheffield, UK

{z.dai, andreas.damianou, j.h.gonzalez, n.lawrence}@sheffield.ac.uk

- 시조새 같은 위치에 있는, 덜 정제된 아이디어 하이브리드 (Variational + auto-encoding + GP)

VAE-DGP [12] 핵심 포인트

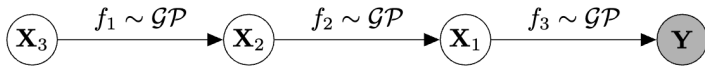
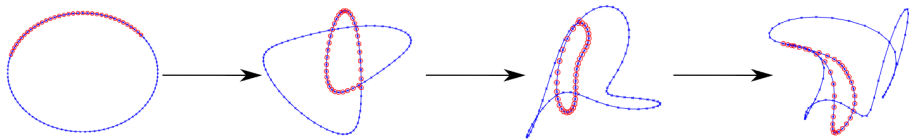
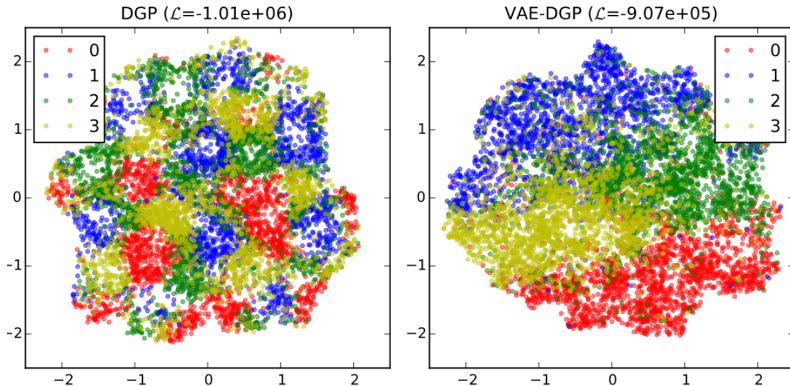


Figure 1: A deep Gaussian process with two hidden layers.



- GP를 깊게 쌓은 deep-GP는 단순한 입력차원의 구조도 더욱 복잡하게 표현할 수 있다
- latent의 방향성이 좀 달랐던 것 (반대로 돌린거니?)

VAE-DGP [12] 핵심 결과



- latent의 방향성이 달랐지만, 결과적으로는 깔끔하게 학습된 latent를 보여줌

VAE-BO [12]

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,^{†,‡,§} Jennifer N. Wei,^{‡,§,¶} David Duvenaud,^{¶,¶} José Miguel Hernández-Lobato,^{§,¶} Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,^{‡,¶} Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,^{∇,||} and Alán Aspuru-Guzik^{*,‡,⊥,¶}

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States

[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

[¶]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada

[§]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

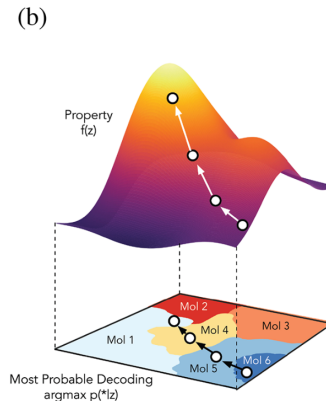
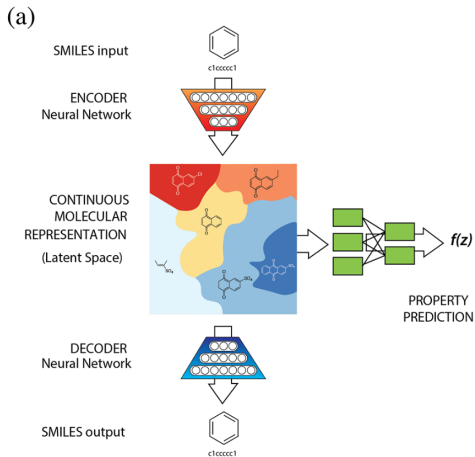
[∇]Google Brain, Mountain View, California, United States

^{||}Princeton University, Princeton, New Jersey, United States

[⊥]Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

- 의도한 방향으로 적용한 것
- 놀랍게도, 화학 저널!

VAE-BO 핵심 아이디어 [12]



- VAE를 통해 $\mathcal{X}$ 를 저차원 latent로 내려서 BO를 수행!

Deep Kernel Learning in Chemistry [11]

communications chemistry

Article



<https://doi.org/10.1038/s42004-024-01219-x>

Deep Kernel learning for reaction outcome prediction and optimization

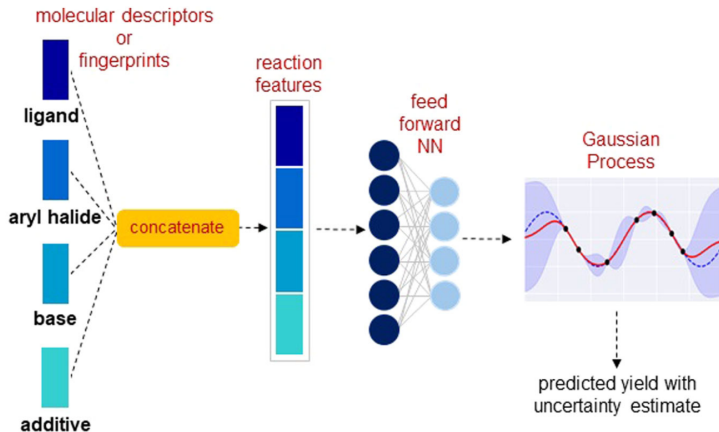


Check for updates

Sukriti Singh & José Miguel Hernández-Lobato

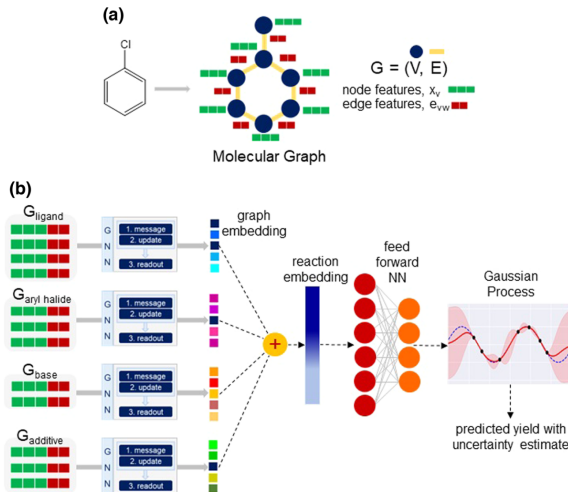
- 다시한번 화학 저널

Deep Kernel Learning in Chemistry [11]



- 아까 본 VAE-BO와 매우 유사한 기본구조

Deep Kernel Learning in Chemistry [11]



접근법 비교 요약 및 발전방향

방법	구조 가정	특징	제약점
Additive Model들 부분공간 기법들 Latent Space BO 계열	가산 구조 선형변환 비선형 임베딩	병렬화 용이 계산 효율적 유연성 높음	구조를 손으로 정해줘야됨 투영 손실, 언저리 효과 등.. 학습 필요한 모듈 존재

- 원래 고차원에 약한 BO를 돕기 위한 노력이었으나,
- 현재는 Latent space를 도메인에 맞게 만들어내는 기법들로 더욱 발전시키는 방향
- Generative model들을 차용하여 GP와 합치는 것도 최근 등장 (BO는 어디로?)

BO Tutorial Part 2: Current

1. 획득함수 + Information Theory

2. High-dimensional BO

3. Multi-objective BO

4. Closing Remarks

다목적 베이지안 최적화 (MOBO)

- 하나의 f 가 아니라, 여러 개의 (상충할수도 있는) f 들을 동시에 최적화해야 된다면?
- 심플한 예시들: 정확도 vs. 연산속도, 비용 vs. 품질, bias vs. variance
- **핵심 해결방안 아이디어들:**
 1. 가장 직관적이고 단순하게, scalarization
 2. Pareto최적화 and/or Hypervolume 개념
 3. Custom design

다목적 최적화의 근본적 차이

- **다변수** 목적함수: $f(x) = (f_1(x), f_2(x), \dots, f_k(x))$
- 목적이 다변수가 되면 $\arg \max f(x)$ 의 개념이 모호해짐
 - e.g. 0, 1은 대소가 명확함. 하지만 (0, 1)과 (1, 0)은 관점에 따라 대소가 달라짐

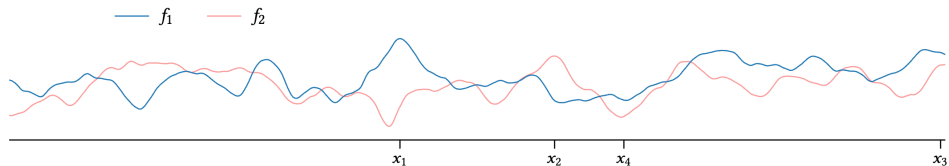
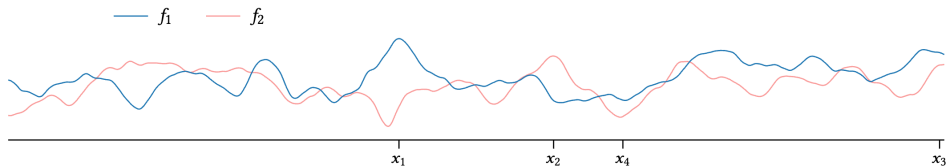


Figure: 2변수 목적함수를 강제로 한 축에 그려보았습니다

다목적 최적화 해결방안 1

해결 방안 1: 가장 직관적이고 단순하게, scalarization

- 목적함수가 벡터라 어려워? 그럼 스칼라로 바꿔.



다목적 최적화 해결방안 1

해결 방안 1: 가장 직관적이고 단순하게, scalarization

- 목적함수가 벡터라 어려워? 그럼 스칼라로 바꿔.

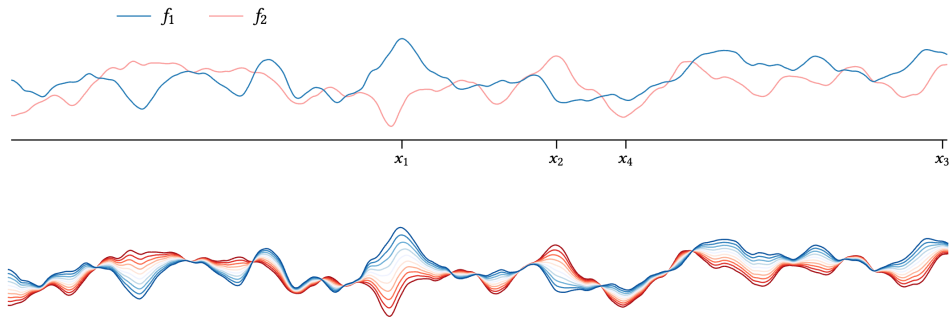


Figure: Scalarization을 어떻게 하느냐에 따라 결과물이 달라진다는 함정이 숨어있습니다

ParEGO [13]

IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 10, NO. 1, FEBRUARY 2005

ParEGO: A Hybrid Algorithm With On-Line Landscape Approximation for Expensive Multiobjective Optimization Problems

Joshua Knowles

- trivia: 이번에는 저자가 화학과 소속!

ParEGO [13]

해결 방안 1: 가장 직관적이고 단순하게, scalarization

- 목적함수가 k 개다? 그럼 하나로 합쳐라

$$g(x) := \sum_i^k \lambda_i f_i(x)$$

- 그리고 나서 저 g 를 갖고 BO를 하면 되지?

ParEGO [13]

해결 방안 1: 가장 직관적이고 단순하게, scalarization

- 목적함수가 k 개다? 그럼 하나로 합쳐라

$$g(x) := \sum_i^k \lambda_i f_i(x)$$

- 그리고 나서 저 g 를 갖고 BO를 하면 되지?

진지한 문제점

- 그럼 저 λ_i 들은 어떻게 정하지?

ParEGO [13]

해결 방안 1: 가장 직관적이고 단순하게, scalarization

- 목적함수가 k 개다? 그럼 하나로 합쳐라

$$g(x) := \sum_i^k \lambda_i f_i(x)$$

- 그리고 나서 저 g 를 갖고 BO를 하면 되지?

진지한 문제점

- 그럼 저 λ_i 들은 어떻게 정하지? 알잘딱깔센.

A Flexible Framework for Multi-Objective Bayesian Optimization using Random Scalarizations

Biswajit Paria

MLD, Carnegie Mellon University

bparia@cs.cmu.edu

Kirthevasan Kandasamy*

EECS, UC Berkeley

kandasamy@eecs.berkeley.edu

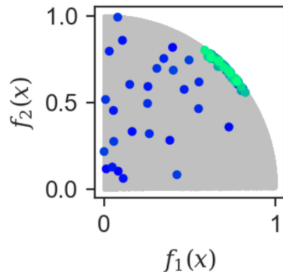
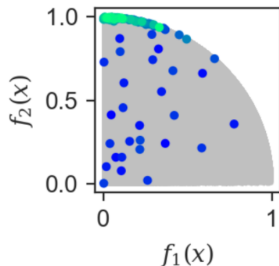
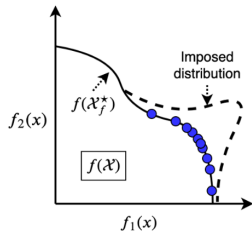
Barnabás Póczos

MLD, Carnegie Mellon University

bapoczos@cs.cmu.edu

- 플렉시블하게 알잘딱깔센 scalarization을 해주면 되겠죠?

MOBO-RS [14]

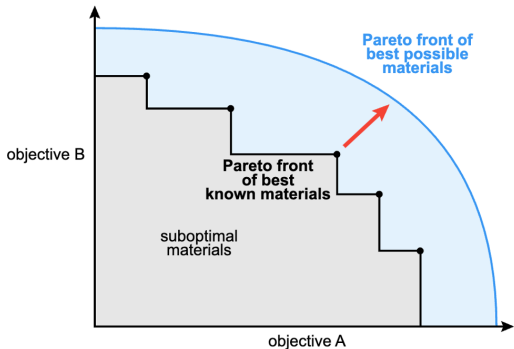


- 왼쪽 (기본 아이디어): λ_i 들의 분포를 하나 prior로 정해주면, 비슷한 영역을 샘플링하게 된다는 점
- 중간 & 오른쪽 (핵심 아이디어): 학습을 통해서 λ 들의 분포를 잘 바꿔서, 후반부에는 매우 좋아보이는 가장자리(Pareto frontier)만 샘플링하도록 하자

다목적 최적화 해결방안 2

해결 방안 2: Pareto최적화 그리고 Hypervolume 개념

- 목적함수가 벡터라 어려워? 다 잘하면 되겠네?
- “꼴리는건 없고 더 잘하는건 있다”는 개념을 Pareto dominance라 하겠습니다
- Front가 넓어지면, 그 밑의 부피가 커집니다



다목적 최적화 해결방안 2

해결 방안 2: Pareto최적화 그리고 Hypervolume 개념

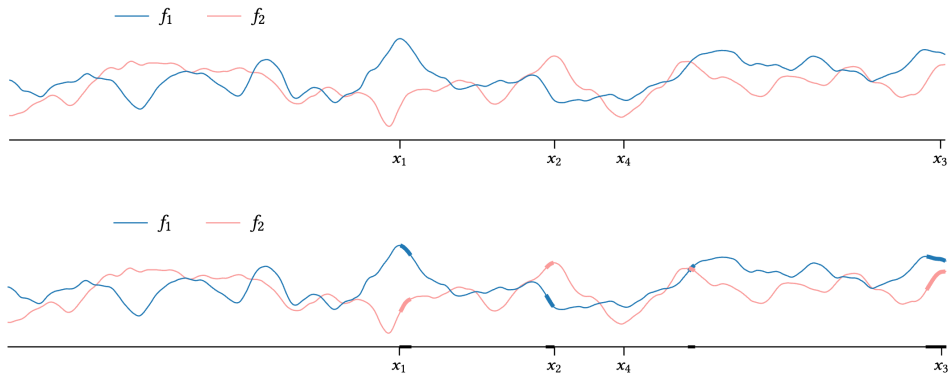


Figure: Pareto front들이 진하게 표시되어 있습니다. 생각했던 거랑 좀 다르죠?

다목적 최적화 해결방안 2

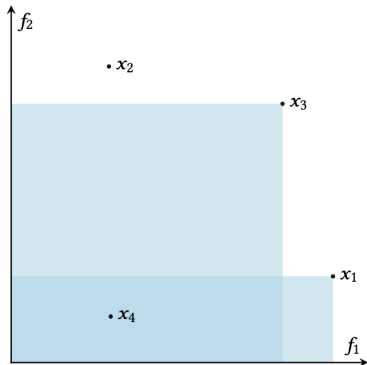


Figure: Pareto-dominated regions

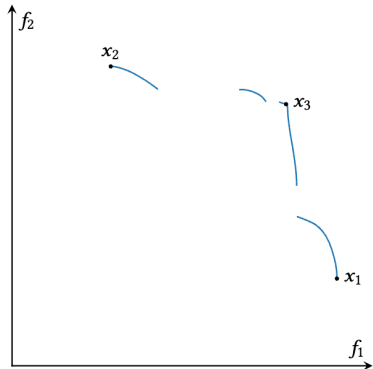


Figure: Pareto front components

- Pareto front라는게 이제 좀 더 납득이 갑니다

아이디어: Hypervolume Improvement

- 문제: Pareto최적화를 하는데, Pareto front를 찾기는 쉽지 않음
- 아이디어1: Pareto front를 근사해보자. Lower bound를 만드는거지
- 아이디어2: 단계적으로 Lower bound를 개량하자. Hypervolume이 커지게

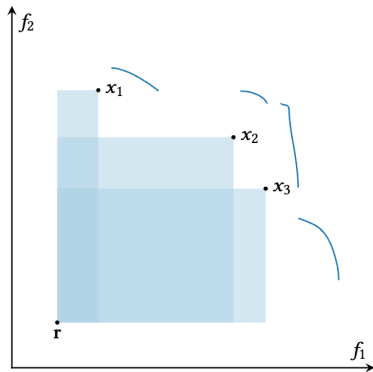


Figure: LB of Pareto front by x_1, x_2, x_3

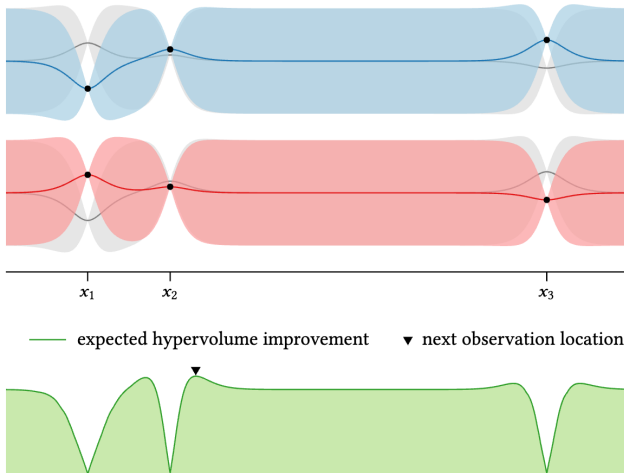
Expected Hypervolume Improvement (EHVI) [?] [?]

Single- and Multi-objective Evolutionary Optimization Assisted by Gaussian Random Field Metamodels

Michael Emmerich, Kyriakos Giannakoglou, Boris Naujoks

EHVI 예시

- posterior mean, f_1
- posterior mean, f_2
- observations
- posterior 95% credible interval, f_1
- posterior 95% credible interval, f_2



EHVI는 계산 품이 많이 든다

그래서 병렬화하고 빨리 계산이 되도록 만들었습니다: qEHVI [?]

Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization

Samuel Daulton

Facebook

sdaulton@fb.com

Maximilian Balandat

Facebook

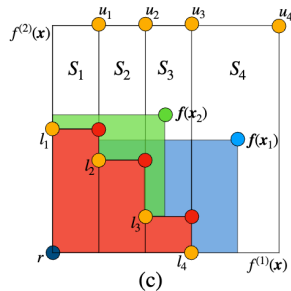
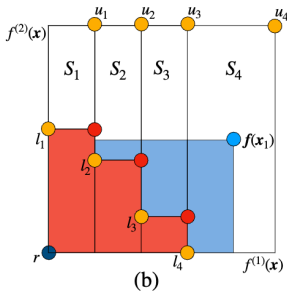
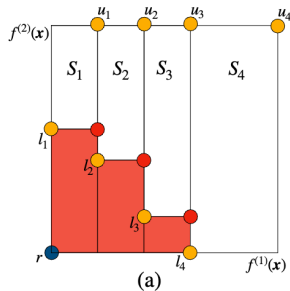
balandat@fb.com

Eytan Bakshy

Facebook

ebakshy@fb.com

qEHVI [?] 개념도



- 색깔 상자들이 차지한 공간의 union을 계산하는데, 이를 빠르게 진행하는 계산기법들을 동원함
- BoTorch, Ax등에서 multi-objective BO에서 기본 예제로 사용되는 기법

BO Tutorial Part 2: Current

1. 획득함수 + Information Theory

2. High-dimensional BO

3. Multi-objective BO

4. Closing Remarks

못다한 이야기들 (현재-미래 부분인 제 3파트!)

- Cost-aware BO 및 관련된 Multi-fidelity BO 시나리오들
- Robust, No-regret, Anytime BO 알고리즘들과 Bandit과의 연결점들
- Exotic surrogates: TPE, RF, BNN 등
- 도함수 정보를 사용할 수 있다면 BO가 어떻게 달라지는지
- 구현 도구들에 대한 이야기: GPU사용한 가속화, MCMC와 다중코어 사용한 가속화
- 실제 활용사례들에 대한 구체적인 case study

P.S. 원래 3시간으로 기획했지만.. 첫날 첫타임은 2시간입니다 (점심시간은 중요하니까요)

- BO는 정말 바로 쓸수있는 도구입니다. 이미 패키지화도 많이 되어있어요
- 도구가 최신화되어도 근본 개념은 바뀌지 않습니다. 오늘은 그 근본개념을 봤구요
- 최신 도구가 항상 더 잘 되는건 아닙니다. 도구-문제의 fit이 더 중요해요
- 풀어야 되는 문제와 잘 맞는 BO 기법이 무엇인지를 고민해보세요 (BO가 안 맞는 도구일수도 있습니다!)

Thank You!

- Questions? » holy@korea.ac.kr
- Slides? » QR
- AIML@K? » Barcode



References I

-  J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, “Predictive Entropy Search for Efficient Global Optimization of Black-box Functions,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.
-  P. Hennig and C. J. Schuler, “Entropy Search for Information-Efficient Global Optimization,” *Journal of machine learning research: JMLR*, vol. 13, no. 57, pp. 1809–1837, 2012.
-  T. P. Minka and R. Picard, *A family of algorithms for approximate bayesian inference*. PhD thesis, USA, 2001.
-  Z. Wang and S. Jegelka, “Max-value Entropy Search for Efficient Bayesian Optimization,” in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3627–3635, PMLR, 2017.

References II

-  K. Kandasamy, J. Schneider, and B. Póczos, “High dimensional Bayesian optimisation and bandits via additive models,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, (Lille, France), p. 295–304, JMLR.org, 2015.
-  Z. Wang, C. Gehring, P. Kohli, and S. Jegelka, “Batched Large-scale Bayesian Optimization in High-dimensional Spaces,” in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (A. Storkey and F. Perez-Cruz, eds.), vol. 84 of *Proceedings of Machine Learning Research*, pp. 745–754, PMLR, 2018.
-  R. Garnett, M. A. Osborne, and P. Hennig, “Active learning of linear embeddings for Gaussian processes,” in *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI’14, (Arlington, Virginia, USA), p. 230–239, AUAI Press, 2014.
-  Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. De Freitas, “Bayesian optimization in a billion dimensions via random embeddings,” *J. Artif. Int. Res.*, vol. 55, p. 361–387, Jan. 2016.

References III

-  D. Eriksson and M. Jankowiak, “High-dimensional Bayesian optimization with sparse axis-aligned subspaces,” in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* (C. de Campos and M. H. Maathuis, eds.), vol. 161 of *Proceedings of Machine Learning Research*, pp. 493–503, PMLR, 2021.
-  Z. Dai, A. Damianou, J. Gonzalez, and N. D. Lawrence, “Variationally Auto-Encoded Deep Gaussian Processes,” in *Proceedings of the International Conference on Learning Representations* (H. Larochelle, B. Kingsbury, and S. Bengio, eds.), vol. 3, (Caribe Hotel, San Juan, PR), 2016.
-  S. Singh and J. M. Hernández-Lobato, “Deep Kernel learning for reaction outcome prediction and optimization,” *Communications Chemistry*, vol. 7, p. 136, June 2024.
-  R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules,” *ACS Central Science*, vol. 4, pp. 268–276, Feb. 2018.

References IV

-  J. Knowles, “ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems,” *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 1, pp. 50–66, 2006.
-  B. Paria, K. Kandasamy, and B. Póczos, “A Flexible Framework for Multi-Objective Bayesian Optimization using Random Scalarizations,” 2019.