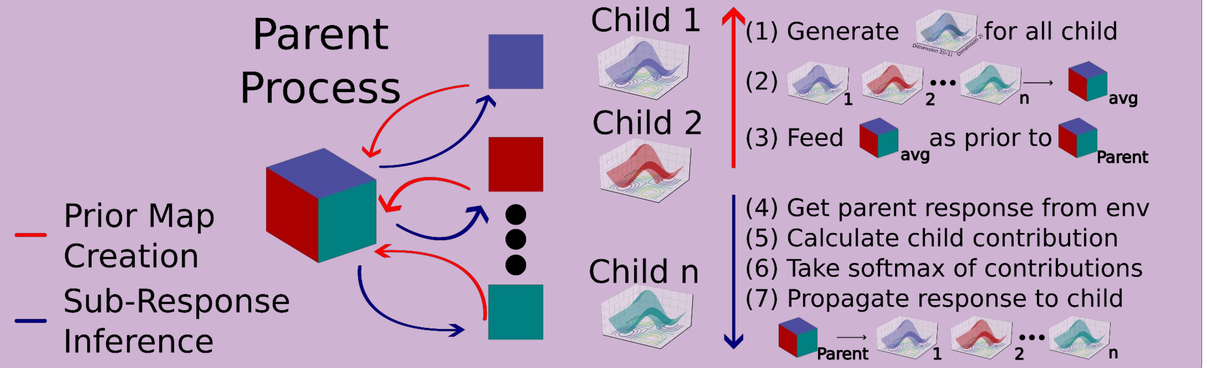
Bidirectional Information Flow

A Sample Efficient Hierarchical Gaussian Process for Bayesian Optimization

JD. Guerra[1,2,3] T. Garbay[1,2] N. Dancause[3] G. Lajoie[2,3] M. Bonizzato[1,2,3]
[1]Polytechnique Montréal, [2]Mila - Quebec AI institute, [3]Université de Montréal



TL;DR BIF improves optimization performance for hierarchical problems and establishes regret bounds for learned kernels according to kernel mismatch Δ

Introduction

Hierarchical Gaussian Process (H-GP) models divide problems into different sub-tasks

Integration of GPs into Bayesian Optimization (BO) allows for **efficient learning in online environments**

Previous works predominantly send information in one direction for hierarchy training

We establish a **novel bidirectional exchange of information**, boosting performance in problems with hierarchical structure

Methods

Upward Flow - children create prior maps for parent

$$\alpha_s(x;\theta_s) = \mu_s(x;\theta_s) + \gamma \cdot \frac{\sigma_s(x;\theta_s)}{\sqrt{n_s(x;\theta_s)}}$$

Downward Flow - parent creates child-level rewards by partitioning true parent response

$$c_s(x) = \frac{\alpha_s(x)}{\max_{x'} \alpha_s(x')} \qquad w_s(x) = \frac{\exp(c_s(x))}{\sum_{j\in S}\exp(c_j(x))}$$

$$y_s(x) = w_s(x) \cdot y_p(x)$$

Regret Bounds

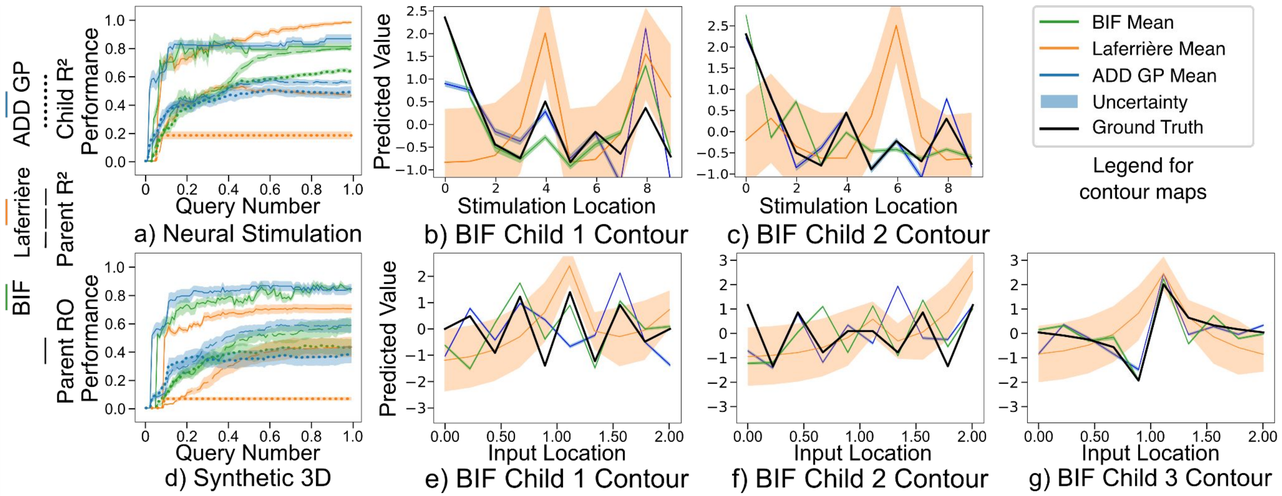
Theorem: GPU-UCB learned kernel regret bound according to kernel mismatch term Δ

$$R_T \leq 2\sqrt{T\beta_T\left(C\gamma_T(k^*) + \sum_{t=1}^{T}\delta_{\sigma,t}\right)} + \sum_{t=1}^{T}\left(2\delta_{\mu,t} + 2\sqrt{\beta_t\delta_{\sigma,t}}\right)$$

Corollary: Bound tightened to sum of child mismatch in hierarchical kernels

$$\hat{\delta}_{\mu,t} := \frac{\sqrt{t-1}(\Delta_{1,t}+\Delta_{2,t})}{\sigma^2}(B\kappa + \sigma\sqrt{t-1} + \sigma\sqrt{2\log(\tfrac{2T}{\delta})})(1+\tfrac{\kappa(t-1)}{\sigma^2})$$

$$\hat{\delta}_{\sigma,t} := \frac{\kappa(t-1)(\Delta_{1,t}+\Delta_{2,t})}{\sigma^2}\left(\frac{\kappa(t-1)}{\sigma^2}+2\right)$$



Datasets	Model	Parent RO ↑	Parent R^2 ↑	Child RO ↑	Child R^2 ↑	AUC ↑
HPO - GAN	Vanilla	96.04 ± 0.45	9.67 ± 1.00	-	-	2,035.21
	Deep GP	94.31 ± 0.85	1.53 ± 0.35	-	-	1,889.41
	ADD GP	94.71 ± 0.77	9.24 ± 1.34	73.18 ± 19.47	26.59 ± 2.81	2,430.96
	Laferrière	94.72 ± 0.55	6.76 ± 1.08	75.73 ± 16.10	23.08 ± 2.42	2,244.58
	BIF - All	**96.50 ± 0.45**	**12.19 ± 0.86**	**76.45 ± 17.05**	**30.04 ± 1.92**	**2,609.34**
HPO - Lasso	Vanilla	99.82 ± 0.11	43.72 ± 1.40	-	-	14,139.57
	Deep GP	99.86 ± 0.04	**89.27 ± 5.89**	-	-	14,655.16
	ADD GP	99.26 ± 0.17	57.35 ± 1.50	63.06 ± 19.66	51.62 ± 4.17	19,572.69
	Laferrière	99.81 ± 0.09	48.44 ± 3.35	10.82 ± 14.48	16.99 ± 3.79	15,302.20
	BIF - All	**99.98 ± 0.02**	79.96 ± 0.94	**78.49 ± 21.70**	**71.63 ± 3.40**	**20,249.31**

Conclusion

BIF achieves up to **4× higher parent R2 and ~100% AUC improvement** over vanilla GPBO

Achieves **top performance** or is runner-up on **38/40 tests**

GP regret bounded by the mismatch of the parent and child kernels to the true kernels

Continual training of children allows for ***plug and play*** style modular learning