Practical Transfer Learning for Bayesian Optimization

- 作者：Matthias Feurer, Benjamin Letham, Frank Hutter, et.al
- 机构：freiburg，Facebook
- 会议：Under review
- 地址：https://arxiv.org/abs/1802.02219
- 代码：无

摘要

论文提出了一种hyperparameter-free的贝叶斯优化集成，基于高斯过程的线性组合与Agnostic Bayesian Learning of Ensembles（ABE），并且给出了worst-case bound。在超参优化benchmark上实验表明论文的方法能显著减少优化时间、提升warm-start BO上的当前SOTA水平。

论文主要内容

创新点：

1. 防止权重稀释机制（概率drop base model ）
2. Ranking loss来计算权重

问题建立

> 高斯过程-BO
>
> $$\mu\left(\mathbf{x}_*\right)=\mathbf{k}_*^T\left(\mathbf{K}+\sigma_n^2 I\right)^{-1}\mathbf{y},\quad \sigma^2\left(\mathbf{x}_*\right)=k\left(\mathbf{x}_*,\mathbf{x}_*\right)-\mathbf{k}_*^T\left(\mathbf{K}+\sigma_n^2 I\right)^{-1}\mathbf{k}_*$$
>
> $$\alpha(\mathbf{x}\mid\mathcal{D})=\mathbb{E}_{f(\mathbf{x}\mid\mathcal{D})}\left[\max\left(0,f\left(\mathbf{x}_{\text{best}}\right)-f(\mathbf{x})\right)\right]=\sigma(\mathbf{x})z\Phi(z)+\sigma(\mathbf{x})\phi(z)$$

在不同的Task（数据集）上最优的超参一般是不同的，但相似的Task可能有相似的response surface。如何在新的Task上利用过去已经学习过的Task上的知识（Transfer）加速搜索？

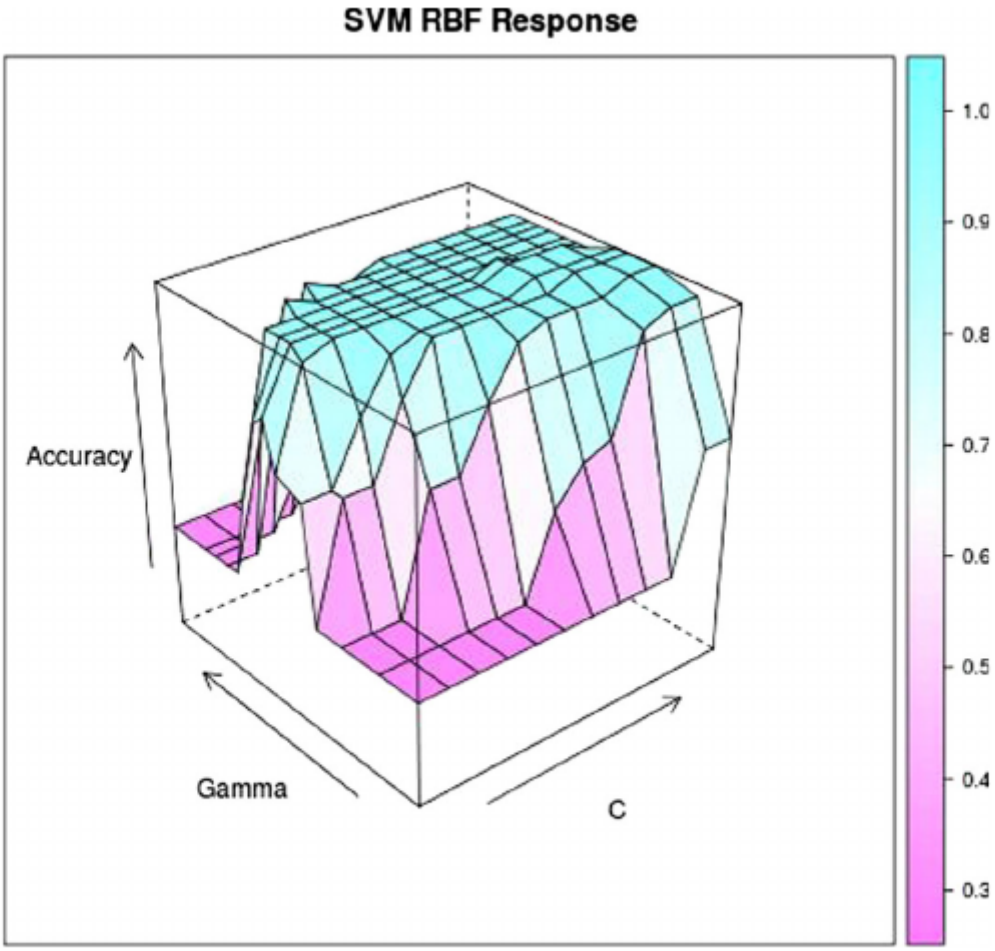


Fig. 1 Response surface of an RBF-SVM on the famous Iris data set. Hyperparameters are the cost of slack (C), and the kernel width γ

Base model：过去已学习任务上的BO model

Target model：目标任务上的BO model

已有工作：

TST-R：Two-stage transfer surrogate model for automatic hyperparameter optimization.

TAF：$\alpha(\mathbf{x}) = w_t EI(\mathbf{x}) + \sum_{i=1}^{t-1} w_i I_i(\mathbf{x}) = w_t EI(\mathbf{x}) + \sum_{i=1}^{t-1} \max\left(0, \arg\min_{\mathbf{x}_k \in \mathcal{D}_t} f^i(\mathbf{x}_k) - f^i(\mathbf{x})\right)$

1. 能够随着观测点增多，淡化base model里的知识
2. 权重使用TST-R中的计算方式，依赖一个敏感超参bandwidth

提出的方法

Preventing Weight Dilution

$$p_{drop}(i) = 1 - \left(\left(1 - \frac{n_t}{H}\right) \frac{\sum_{s=1}^{S} 1\left(l_{i,s} < l_{t,s}\right)}{S + \alpha S}\right)$$

1. Base model和target model竞争，删掉比target model更差的model，防止权重稀释且能降低开销
2. 加入先验知识：target model是'correct model'
3. $\left(1-\frac{n_t}{H}\right)$ 随着观测点的增加线性递减，意味着逐渐增加base model被删掉的概率
4. α 增加会增大base model被删掉的概率（超参）

Theorem 4.1. *Bayesian optimization using a linear combination of Gaussian processes with weights learned according to Section 4.2 is at most a factor of* $1/\left(\frac{1}{H}\sum_{h=1}^{H}\left(1-\left(\left(1-\frac{h}{H}\right)\frac{S}{S+\alpha S}\right)\right)^{t-1}\right)$ *slower than Bayesian optimization in the worst case.*

理论分析：概率drop base model策略不会比直接使用target model差太多（在最坏的情况下）

Ranking-weighted

Ranking loss：

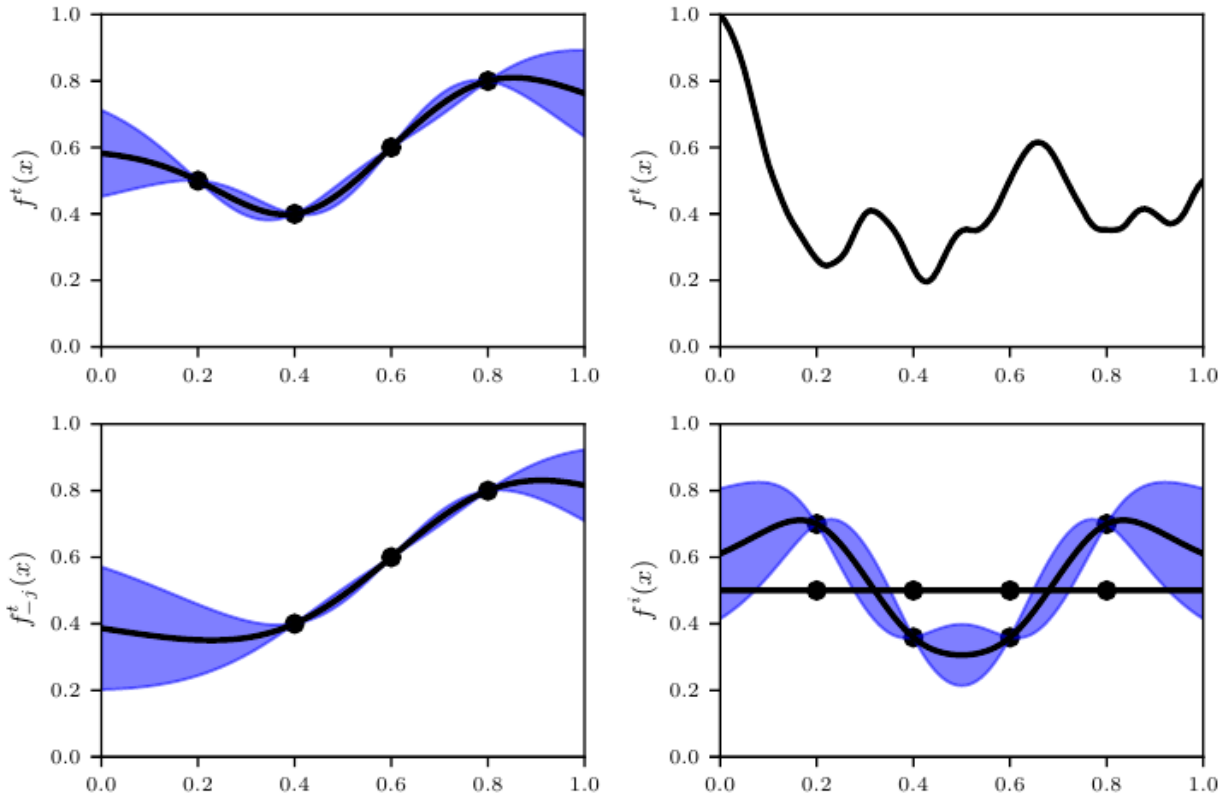
$$\mathcal{L}\left(f,\mathcal{D}_t\right)=\sum_{k=1}^{n_t}\sum_{l=1}^{n_t}1\left(\left(f\left(\mathbf{x}_k^t\right)<f\left(\mathbf{x}_l^t\right)\right)\oplus\left(y_k^t<y_l^t\right)\right)$$

agnostic Bayesian ensemble learning： $model_i$ 集成的权重由 $model_i$ 是最佳模型(validation上最佳)的概率确定。这里包括了确定target model的权重（TST-R的target model始终是个固定值）

$$w_i=\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\mathbb{I}\left(i\in\arg\min_{i'}l_{i',s}\right)}{\sum_{j=1}^{t}\mathbb{I}\left(j\in\arg\min_{i'}l_{i',s}\right)}\right)$$

B.1. Illustration of the Ranking Loss

Figure 2 provides an illustration of the ranking loss.



Gaussian Process Ensemble

RGPE

$$\bar{f}(\mathbf{x} \mid \mathcal{D}) = \sum_{i=1}^{t} w_i f^i \left(\mathbf{x} \mid \mathcal{D}_i\right)$$

$$\bar{f}(\mathbf{x} \mid \mathcal{D}) \sim \mathcal{N}\left(\sum_{i=1}^{t} w_i \mu_i(\mathbf{x}), \sum_{i=1}^{t} w_i^2 \sigma_i^2(\mathbf{x})\right)$$

TST-R方式

均值使用加权求和，方差只使用target model

Mixture of Gaussian Processes

RMoGP混合高斯过程: $p_{\text{mix}}\left(f(\mathbf{x})\right) = \sum_{i=1}^{t} p_i(f(\mathbf{x}))w_i$

$$EI_{\text{mix}}(\mathbf{x}) = \mathbb{E}_{f(\mathbf{x})\sim p_{\text{mix}}}[I(\mathbf{x})] = \mathbb{E}_i \mathbb{E}_{f(\mathbf{x})\sim p_i}[I(\mathbf{x})] = \sum_{i=1}^{t} w_i EI_i(\mathbf{x})$$

- 和TAF形式相似，区别在于对base model使用的是EI不是I

实验

Scaling study

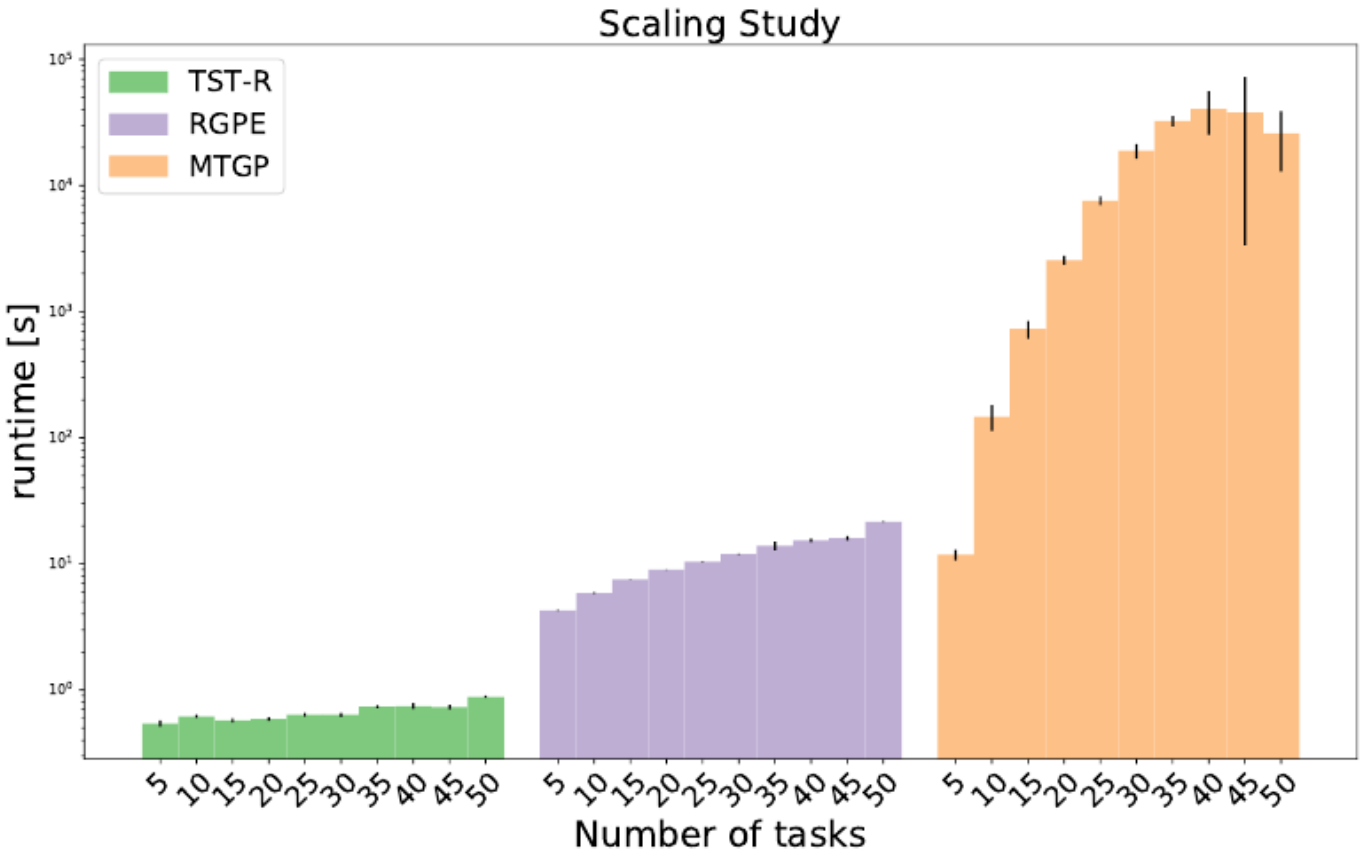


Figure 1: Caption

- Multi-task Gaussian processes (MTGP)
  - 时间复杂度： $O(t^3n^3)$
  - 只建立一个高斯过程模型，考虑所有tasks的观测点
- RGPE的额外开销仅在计算权重

benchmark

Table 1: Final results on four hyperparameter optimization benchmarks. The numbers reported are the average normalized regret [66]. Top groups: single task baselines; middle groups: multitask baselines; **bottom groups: our methods**.

	10	20	30	40	50	10	20	30	40	50
	GLMNET ($2d$)					SVM ($4d$)				
GP(MAP)	1.91	0.59	**0.31**	**0.21**	**0.15**	5.30	1.70	0.97	0.70	0.56
Random(1x)	4.79	1.39	0.88	0.70	0.52	5.56	2.98	2.12	1.77	1.59
Random(4x)	**0.70**	**0.36**	0.29	0.24	0.21	**1.77**	1.16	0.90	0.79	0.70
TST-R	1.63	0.86	0.65	0.40	0.21	3.91	1.85	1.17	0.92	0.75
TAF(TST-R)	0.99	**0.35**	**0.23**	**0.20**	**0.18**	3.24	1.73	0.99	0.70	0.57
TAF(RGPE)	0.76	**0.34**	**0.27**	**0.20**	**0.17**	3.18	**1.19**	**0.76**	**0.60**	**0.47**
RMoGP	**0.65**	**0.31**	**0.24**	**0.19**	**0.16**	3.16	**1.16**	**0.74**	**0.51**	**0.42**
RGPE(mean)	1.60	0.85	**0.67**	**0.62**	**0.14**	3.29	**1.03**	**0.72**	**0.53**	**0.43**
	NN ($7d$)					XGB ($10d$)				
GP(MAP)	16.40	11.51	9.69	8.62	7.84	5.29	2.46	1.66	1.20	0.98
Random(1x)	14.66	11.73	10.09	9.05	8.07	5.69	3.41	2.46	2.08	1.78
Random(4x)	9.05	6.27	5.00	4.11	3.49	2.08	1.50	1.28	1.20	1.14
TST-R	8.50	6.95	6.12	5.60	4.99	**1.37**	**1.00**	0.85	**0.77**	**0.73**
TAF(TST-R)	6.11	5.01	4.27	3.79	3.41	1.55	1.05	0.85	0.75	0.71
TAF(RGPE)	**4.52**	**3.26**	**2.72**	**2.36**	**2.14**	**1.53**	**0.99**	**0.80**	**0.72**	**0.67**
RMoGP	5.72	3.83	3.29	2.86	2.53	**1.49**	**0.95**	**0.80**	**0.72**	**0.67**
RGPE(mean)	6.50	4.67	3.90	3.37	3.02	**1.41**	**0.94**	**0.77**	**0.69**	**0.66**

- Wilcoxon signed-rank test
  - 带下划线是best method，不带下划线但加粗是 与best method没有显著差异
- TST-R的方法比最好方法显著表现差（敏感的bandwith超参）

> - We provided each meta-learning method with a single meta-learned configuration and each other method **with a Latin hypercube design of size 10**.

Grid data benchmark

Table 9: Results for Adaboost$_g$ and SVM$_g$ benchmarks. Top group: non-multitask baselines; middle group: multitask baselines; **bottom group: our methods**.

	$Adaboost_g$					SVM_g				
	10	20	30	40	50	10	20	30	40	50
GP(MAP)	5.42	**2.28**	**1.27**	**0.83**	**0.68**	9.66	3.64	2.06	1.45	1.13
Random(1x)	5.61	3.41	2.43	1.88	1.58	11.91	6.54	4.60	3.82	3.04
Random+box	5.91	3.59	2.50	1.94	1.47	9.79	5.51	4.32	3.65	3.10
GP(MAP,Box)	5.44	**2.32**	**1.37**	**0.92**	0.77	9.58	3.06	1.85	1.23	0.92
SMFO	4.43	2.93	2.15	1.59	1.29	5.45	3.69	2.50	2.02	1.71
WAC	9.10	6.52	5.10	3.87	3.10	11.62	9.33	8.09	7.35	6.59
ABLR	5.17	**2.38**	1.56	**0.99**	**0.67**	5.57	3.36	1.92	1.27	1.07
TST-R	4.87	**2.49**	**1.40**	**0.84**	**0.53**	4.56	2.20	**0.95**	**0.66**	**0.39**
TAF	**4.00**	2.60	1.63	1.15	0.73	4.59	2.52	1.53	1.03	**0.56**
TAF(RGPE)	**3.90**	**2.26**	**1.39**	**0.96**	**0.60**	3.81	**2.05**	**1.26**	**0.75**	**0.39**
RMoGP	4.52	2.60	**1.40**	**0.87**	**0.51**	**3.35**	**1.75**	**1.15**	**0.67**	**0.39**
RGPE(mean)	5.43	2.74	1.65	1.06	**0.65**	5.35	2.99	1.43	**0.79**	**0.51**
CFNEI	6.01	2.78	1.68	**1.02**	**0.51**	4.81	**2.01**	**0.95**	**0.61**	**0.38**

消融实验

Table 10: Ablation study on all benchmarks.

	$Adaboost_g$					SVM_g				
GP(MAP)	5.42	**2.28**	**1.27**	**0.83**	**0.68**	9.66	3.64	2.06	1.45	1.13
TST-R(HPO)	4.81	2.46	**1.59**	**1.00**	0.70	4.85	2.85	1.79	1.27	0.69
TST-R	4.87	2.49	**1.40**	**0.84**	**0.53**	4.56	2.20	**0.95**	**0.66**	**0.39**
TAF(TST-R,HPO)	**3.84**	**1.99**	**1.31**	**1.01**	0.74	4.43	2.32	**1.17**	**0.62**	**0.43**
TAF(TST-R)	**4.00**	2.60	1.63	1.15	0.73	4.59	2.52	1.53	1.03	0.56
RMoGP	4.52	2.60	**1.40**	**0.87**	**0.51**	**3.35**	**1.75**	**1.15**	0.67	**0.39**
RMoGP(None)	4.28	**2.37**	**1.54**	**0.95**	0.76	**3.51**	**1.80**	**1.10**	**0.78**	0.63
RMoGP(α=0.1)	4.56	2.68	**1.48**	**0.94**	**0.55**	**3.30**	**1.82**	**0.94**	**0.55**	**0.45**
RMoGP(α=0.2)	4.37	2.42	**1.48**	**1.01**	**0.67**	**3.38**	**1.83**	1.14	0.71	0.53
RMoGP(α=0.5)	4.33	2.62	**1.51**	**0.99**	**0.50**	3.51	**1.84**	**1.07**	**0.70**	**0.45**
RMoGP(α=1)	4.32	2.51	1.60	1.14	0.70	**3.38**	**1.91**	**1.09**	**0.57**	**0.33**
RMoGP(α=2)	4.34	2.46	1.55	1.15	0.73	3.59	**1.79**	**1.00**	**0.63**	**0.39**
RMoGP(α=5)	4.54	2.73	1.62	**1.11**	**0.77**	3.63	**1.78**	**0.98**	**0.66**	0.52
RMoGP(S=100)	4.36	2.40	**1.47**	**1.04**	**0.67**	**3.36**	**1.95**	**1.09**	**0.64**	**0.43**
RMoGP(S=1000)	4.63	2.58	1.68	**1.03**	**0.66**	**3.32**	**1.80**	**1.07**	**0.75**	0.56
RMoGP(S=100000)	4.41	2.55	**1.54**	1.15	0.73	**3.25**	**1.85**	**1.13**	**0.74**	0.56
	GLMNET					SVM				
GP(MAP)	1.91	0.59	0.31	**0.21**	**0.15**	5.30	1.70	0.97	0.70	0.56
TST-R(HPO)	1.92	0.78	0.29	0.23	0.19	**3.34**	1.92	1.33	1.10	1.00
TST-R	1.63	0.86	0.65	0.40	0.21	3.91	1.85	1.17	0.92	0.75
TAF(TST-R,HPO)	**0.76**	0.37	0.26	**0.18**	0.16	3.11	1.20	0.72	0.56	**0.47**
TAF(TST-R)	0.99	0.35	0.23	0.20	0.18	3.24	1.73	0.99	0.70	0.57
RMoGP	**0.65**	**0.31**	**0.24**	0.19	0.16	3.16	1.16	0.74	**0.51**	**0.42**
RMoGP(None)	**0.66**	**0.30**	**0.23**	0.19	0.17	**2.88**	**1.06**	0.70	0.55	**0.47**
RMoGP(α=0.1)	**0.70**	0.31	**0.21**	**0.17**	**0.15**	**2.91**	**1.16**	**0.68**	**0.50**	**0.39**

RMoGP(α=0.1)	**0.70**	0.31	**0.21**	**0.17**	**0.15**	**2.91**	**1.16**	**0.68**	**0.50**	**0.39**
RMoGP(α=0.2)	0.83	0.34	0.25	0.19	0.17	**2.75**	**1.02**	**0.64**	**0.48**	**0.40**
RMoGP(α=0.5)	**0.73**	**0.29**	**0.21**	**0.17**	**0.14**	**3.04**	**1.14**	0.77	**0.52**	**0.44**
RMoGP(α=1)	0.77	**0.28**	**0.22**	0.18	0.16	3.44	1.21	0.74	0.60	**0.53**
RMoGP(α=2)	0.80	0.37	0.25	0.20	0.18	3.12	1.15	0.69	0.55	0.49
RMoGP(α=5)	**0.71**	0.35	0.25	0.20	0.18	**3.09**	1.27	**0.71**	**0.50**	**0.42**
RMoGP(S=100)	**0.70**	0.32	**0.20**	**0.16**	**0.13**	**2.62**	**0.97**	**0.59**	**0.47**	**0.42**
RMoGP(S=1000)	**0.71**	**0.31**	**0.22**	**0.18**	**0.16**	**2.73**	**1.12**	0.76	**0.59**	**0.51**
RMoGP(S=100000)	**0.67**	**0.33**	**0.24**	**0.20**	0.17	**2.93**	**1.05**	**0.67**	**0.50**	**0.43**
	NN					XGB				
GP(MAP)	16.40	11.51	9.69	8.62	7.84	5.29	2.46	1.66	1.20	0.98
TST-R(HPO)	7.88	6.37	5.25	4.82	4.36	**1.52**	1.18	1.01	0.93	0.88
TST-R	7.91	6.36	5.49	4.87	4.40	**1.37**	1.00	0.85	0.77	0.73
TAF(TST-R,HPO)	**4.54**	**3.10**	**2.56**	**2.25**	**2.00**	**1.36**	**0.94**	**0.79**	**0.73**	**0.69**
TAF(TST-R)	6.11	5.01	4.27	3.79	3.41	1.55	1.05	0.85	**0.75**	**0.71**
RMoGP	5.72	3.83	3.29	2.86	2.53	1.49	**0.95**	**0.80**	**0.72**	**0.67**
RMoGP(None)	5.95	3.97	3.25	2.77	2.47	1.58	1.04	**0.82**	**0.74**	**0.68**
RMoGP(α=0.1)	5.71	4.05	3.38	2.93	2.56	1.56	**0.96**	**0.81**	**0.76**	0.72
RMoGP(α=0.2)	5.45	4.02	3.36	2.99	2.65	1.54	**0.98**	**0.80**	**0.75**	**0.69**
RMoGP(α=0.5)	5.77	3.84	3.11	2.76	2.42	**1.50**	**0.91**	**0.79**	**0.71**	**0.67**
RMoGP(α=1)	5.87	4.04	3.27	2.79	2.52	1.55	**0.97**	0.84	0.77	**0.72**
RMoGP(α=2)	5.47	4.04	3.48	3.00	2.73	1.53	**0.98**	**0.82**	**0.75**	0.72
RMoGP(α=5)	6.05	4.27	3.54	3.14	2.80	**1.41**	**0.95**	**0.80**	**0.73**	**0.69**
RMoGP(S=100)	5.59	3.84	3.18	2.71	2.46	1.58	**0.95**	**0.78**	**0.72**	**0.67**
RMoGP(S=1000)	5.95	4.21	3.31	2.83	2.43	**1.51**	**0.94**	**0.77**	**0.71**	**0.67**
RMoGP(S=100000)	5.82	3.92	3.10	2.65	2.46	**1.53**	**0.91**	**0.79**	**0.73**	**0.69**

- leave-one-task-out hyperparameter tuning (HPO)
  - 使用grid search在t-1个tasks上找到最好的超参，应用在target task上
  - 代价昂贵、不切实际
  - 实验结果说明TST-R中的bandwidth超参如果能设置好，效果就很好。
- None表示不使用减枝base model
  - 实验说明防止权重稀释的方法有用，并且在最坏的情况下比纯BO慢一个乘数比例
- α 不敏感
- S 样本数增加不敏感，实验中100个就足够