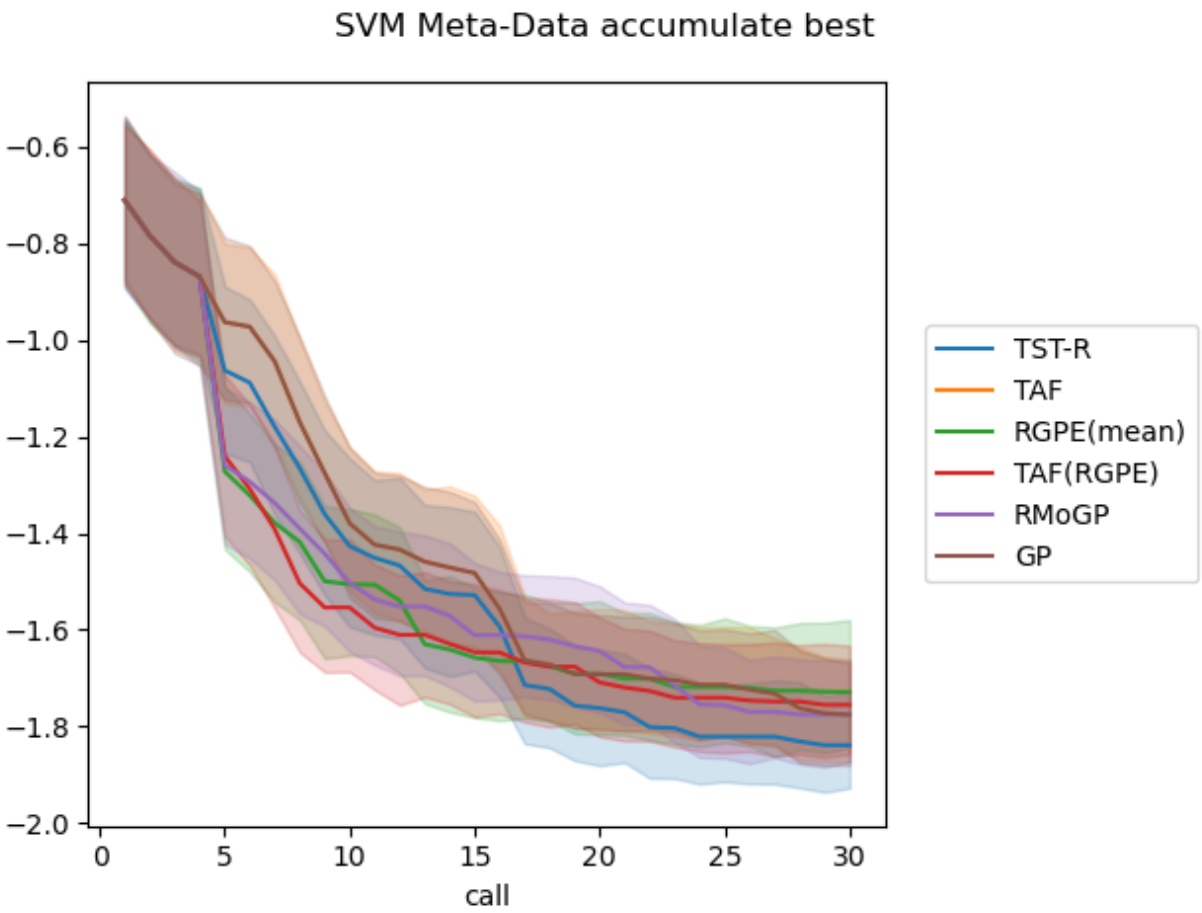
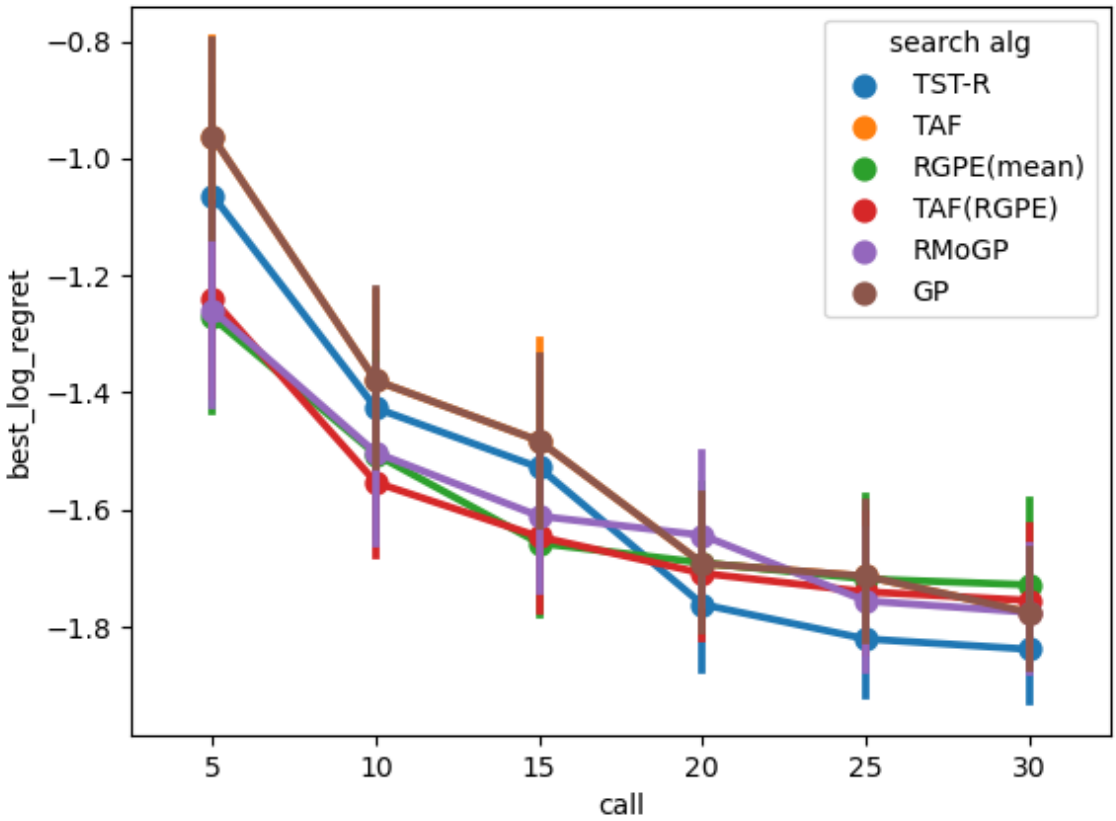
Transfer Bayesian Optimization

Transfer BO方法主要差异：组合的形式、组合的权重





代理模型集成方式

$$\mu = \frac{\sum_{i=1}^{M+1} w_i \mu_i \left(x_*\right)}{\sum_{i=1}^{M+1} w_i}$$

$$\sigma = \sqrt{\left(\sum_{i=1}^{M+1} v_i \sigma_i^{-2} \left(x_\star\right)\right)^{-1}}$$

TST-R 【组合代理模型形式】	权重计算方式：Kernel regression · 缺陷 ◦ 即时对不同task的观测做归一化，仍然有scale不一致问题，因为新任务的观测难以做准确的归一化 ◦ bandwidth 需要人为设定（且sensitive） ◦ weight是base model与target model之间确定，忽略了base model之内的相关性 ◦ 只对模型的均值进行组合，方差直接用的target model方差
TAF 【组合采集函数形式】	· 优势 ◦ TAF并不像transfer surrogate 那样直接使用surrogate的输出具体绝对的值，而是使用相对获得的提升（注：base model使用的不是期望提升），解决scale不一致。一个是组合surrogate输出（TST），一个是组合“surrogate提升”（TAF）

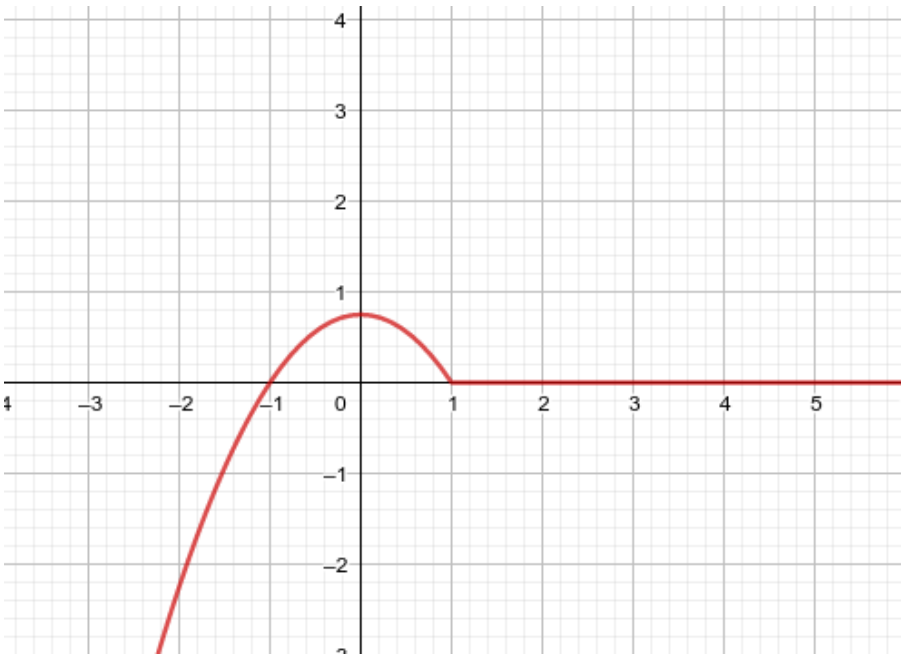
	◦ old data与new data的相似性来决定old data之间的相对权重大小，这个与TST-R一致。TAF能够解决的是随着new knowledge增加，自适应淡化 old knowledge(因为观测增多，base model的impovement会降低) · combine knowledge系数 ◦ POE（product of expert）：权重使用的是每个代理模型的精度Precision（inverse of covariance） ◦ Kernel regression ▪ meta-feature 自始至终权重固定，没有考虑新观测点 ▪ **rank（即TST-R的权重计算方式）** 动态调整权重，能够考虑进新观测的点 ◦ Ranking weighted ▪ 通过采样计算（100个样本足够获得好的性能）
RGPE（Ranking-Weighted Gaussian Process Ensemble） 【组合代理模型形式】	· 使用ranking loss 来计算权重 · 对所有model的均值和方差都进行了组合 · Preventing Weight Dilution
RMoGP 【组合采集函数形式】	· 使用ranking loss来计算权重 · 相比于TAF组合形式，对base model使用的期望提升 · Preventing Weight Dilution

权重计算

Meta data相似性（TST-M中使用）	构建task的feature，使用feature的距离
rank一致数目（TST-R、TAF中使用）	过去task上拟合的BO模型在新task上的预测与实际观测rank一致性，类似于kendall秩**相关系数**
Ranking loss（RGPE、RMoGP、TAF（RGPE）中使用）	权重表示该BO模型是所有BO模型（base model+target model）中最佳（最低泛化误差）的**概率**

计算细节

generalized Product of experts	$p\left(y_{\star} \mid x_{\star}, X, y\right)=\prod_{i=1}^{M+1} p_{i}^{\beta_{i}}\left(y_{\star} \mid x_{\star}, X^{(i)}, y^{(i)}\right)$ ，似然的乘积。

	$w_i = \beta_i \sigma_i^{-2}(x_\star)$ $v_i = \beta_i$ 均值的权重正比于精度（precision）
Kernel regression	$\gamma(x)$ ：  $k_\rho(\chi_i, \chi_j) = \gamma\left(\frac{\|\chi_i - \chi_j\|_2}{\rho}\right)$ base model Base model与target model之间进行计算相似性 · χ 的距离可以使用meta feature、肯德尔相关系数计算
RGPE（Ranking-Weighted Gaussian Process Ensemble）	1. Ranking loss（即validation上prediction和gt的rank不一致数）： $\mathcal{L}(f, \mathcal{D}_t) = \sum_{k=1}^{n_t}\sum_{l=1}^{n_t} 1\left(\left(f\left(\mathbf{x}_k^t\right) < f\left(\mathbf{x}_l^t\right)\right) \oplus \left(y_k^t < y_l^t\right)\right)$ 2. 计算权重（即最小泛化误差的概率）： $w_i = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\mathbb{I}\left(i \in \arg\min_i, l_{i',s}\right)}{\sum_{j=1}^{t}\mathbb{I}\left(j \in \arg\min_{i'} l_{i',s}\right)}\right)$ 相比于kernel regression的rank方式，考虑了base models之间的竞争关系，权重即表示model是best model（泛化的loss最小）的概率（mutinomial distribution） 3. 利用计算出的权重来组合均值与方差
RMoGP （Ranking-weighted Mixture of Gaussian Processes）	$EI_{\text{mix}}(\mathbf{x}) = \mathbb{E}_{f(\mathbf{x})\sim p_{\text{mix}}}[I(\mathbf{x})] = \mathbb{E}_i\mathbb{E}_{f(\mathbf{x})\sim p_i}[I(\mathbf{x})] = \sum_{i=1}^{t} w_i EI_i(\mathbf{x})$

Preventing Weight Dilution 阻止权重稀释

motivation：随着观测点增多，target model理应增加权重

1. 去掉明显比target model表现差的base model：如果base model的loss（采样方式得到）中位数比target model的95分位数还高，则丢弃
2. 按概率随机丢弃base model： $p_{\text{drop}}(i) = 1 - \left(\left(1 - \frac{n_t}{H}\right) \frac{\sum_{s=1}^{S} 1\left(l_{i,s} < l_{t,s}\right)}{S + \alpha S} \right)$
3. Kernel regression中的bandwith也可以理解为阻止权重稀释的一个调整策略。距离大于bandwidth的BO模型不参与集成（ensemble）

Grid benchmark

Surrogate benchmark

Metric 评价指标

- Average rank：在grid 搜索空间中，优化器给出的策略的score排名，在不同数据集上取平均
- Average distance to the global minimum（ADTM）：也是在数据集上取平均（regret加上归一化）

$$\text{ADTM} = \frac{1}{|\mathcal{D}|} \sum_{D \in \mathcal{D}} \min_{x \in \mathcal{X}_l} \frac{y_D(x) - y_D^{\min}}{y_D^{\max} - y_D^{\min}}$$

- Fraction of unsolved data sets：算法能找到grid 搜索空间中的最优配置在数据集中的占比