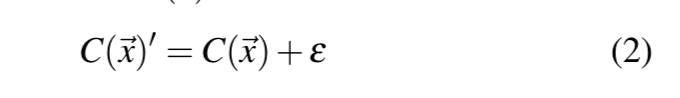
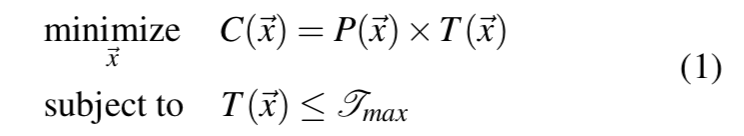
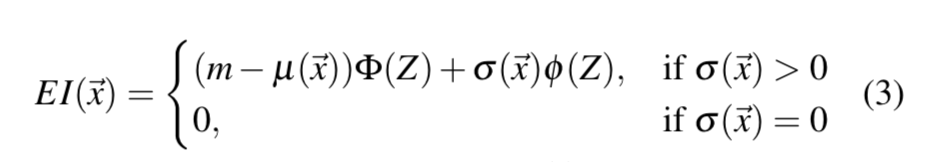
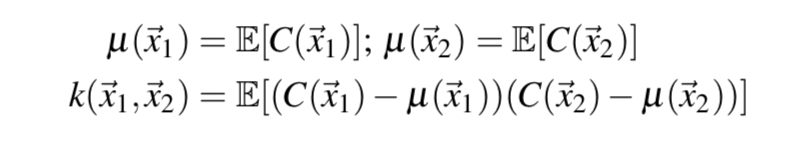
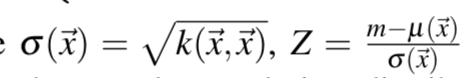
Cherrypick

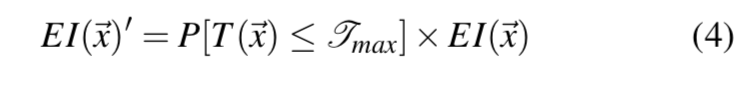
基于朴素贝叶斯优化的配置策略。



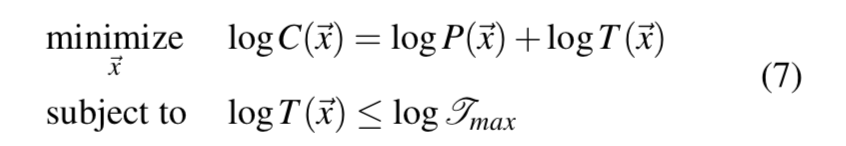
ε ∼ N (0,σε2)

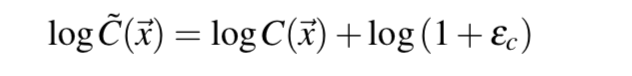






使用协方差度量相似度，使用预期提升作为BO的挑选函数，替代模型使用高斯过程。对运行时间给定上界，超出会受到惩罚。为了衡量网络波动，添加了噪声，并取log对数。





测试

TPC-DS bench- mark on Spark SQL with a scale factor of 20

TPC-H on Hadoop with a scale factor of 100

trace runs 20 queries

TeraSort on Hadoop with 300 GB of data

regression workload in SparkML with 250k examples, 10k features, and 5 iteration

250k observations with 10k fea- tures

配置

Amazon EC2: M4 (general purpose), C4 (compute op- timized), R3 (memory optimized), I2 (disk optimized) instances.

初始化

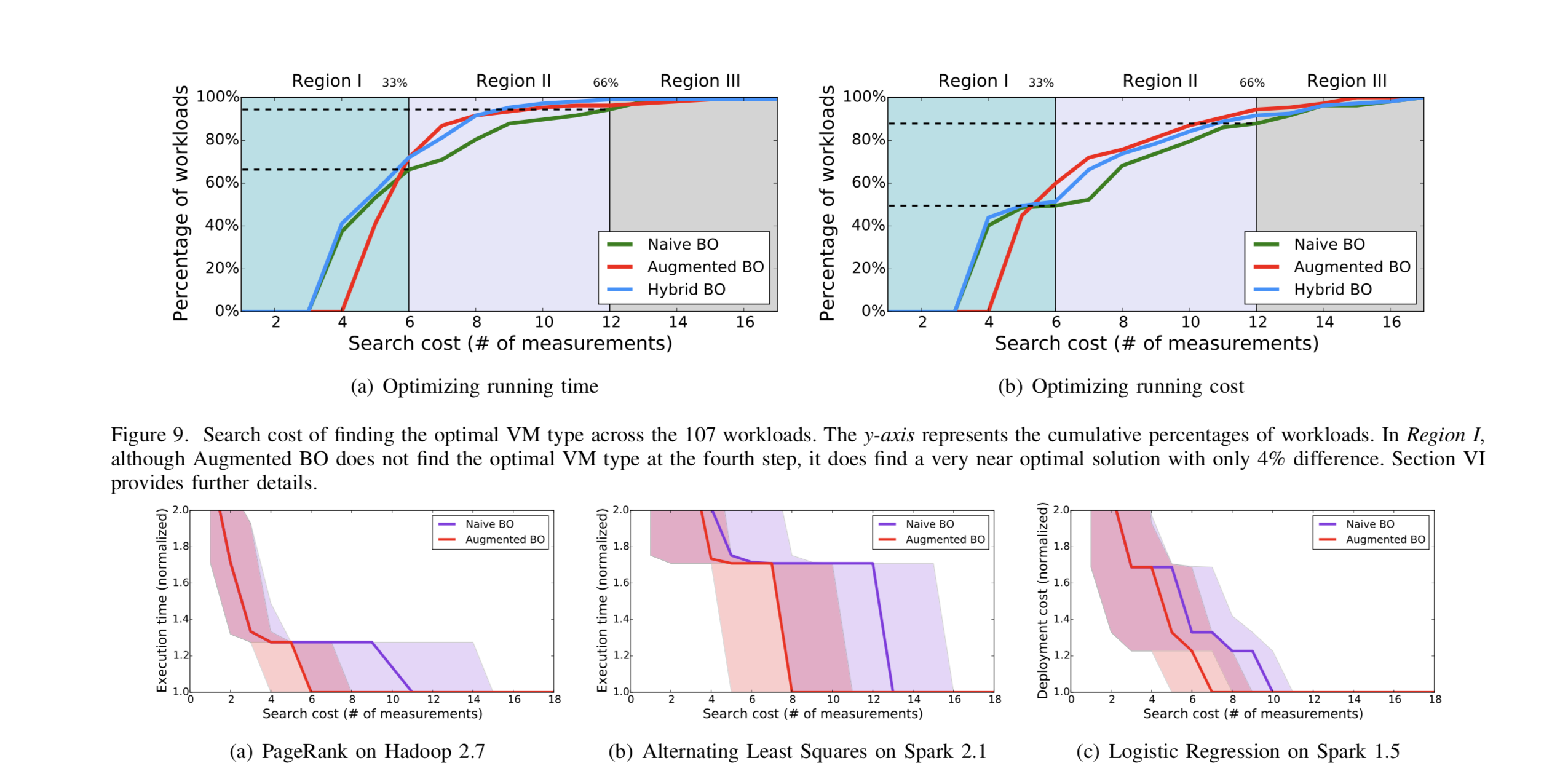
EI= 10%, *N* = 6, and 3 initial samples

与CherryPick相比，诸如坐标下降和随机搜索等替代解决方案的运行时间最多会提高75％，搜索时间最多会提高45％。

Arrow

指出了cherrypick使用的朴素贝叶斯优化的不足。例如工作负载大小会影响成本，相似的工作流要求的最优配置不一定相同。复杂模型的缺点：要求无噪声，训练成本高，准确率不够。

在BO中引入低级指标，工作负载压力（cpu），内存压力，io压力。启动更慢但更多情况的搜索成本更低。



Paris

使用随机森林。让工作流在给定vm上测试得到指纹。依据指纹寻找合适的设备。

EMSE较大，50%。

Ernest

使用非负最小二乘法。交叉验证。添加额外实例，排除最差者。

搜索成本高。