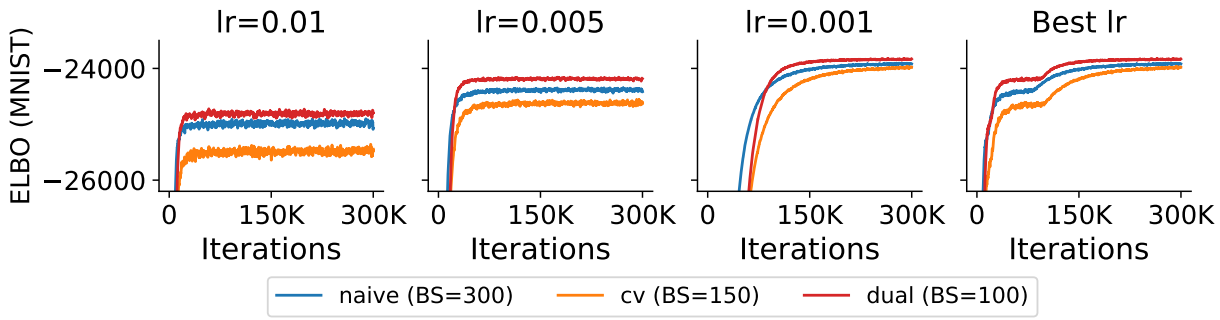

Figure 1. **Under the same number of budget per iteration, dual still converges faster than baseline estimators on MNIST.** Dual requires three and two times more gradient evaluations than naive and cv respectively. Therefore we increase the batch size of baseline estimators to ensure the same budget at each iteration. In particular, the batch size (BS) is 300, 150, and 100 for naive, cv and dual respectively. Dual performs better the baseline estimators with a smaller number of mini-batch samples.

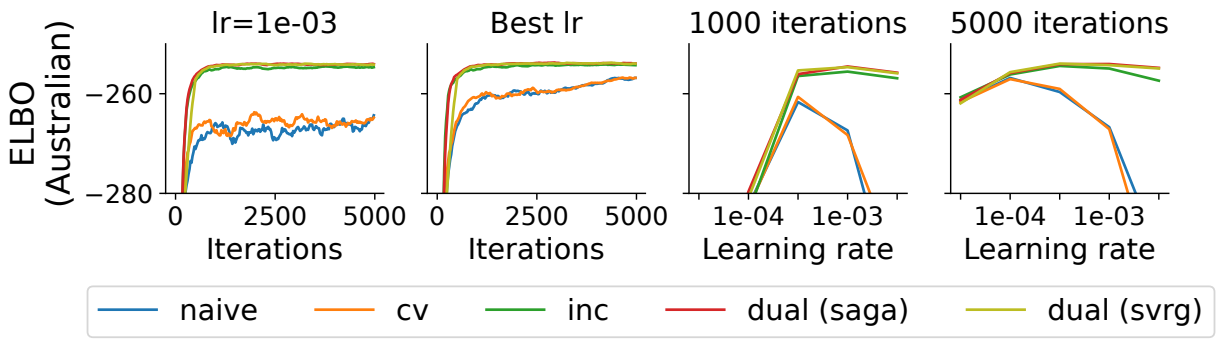

Figure 2. **The SVRG version of dual shows performance similar to the SAGA version on Australian.** The origin version of SAGA-based dual control variate requires $O(ND)$ memory cost. It is possible to alleviate the additional memory cost by using the SVRG version of dual, which cost no extra memory but would require extra gradient evaluation at each step. In the experiments above, we update the SVRG cache every 5 epochs, equivalent to 0.2 extra gradient evaluations per iteration. Overall, we observe dual (svrg) showing results similar to the saga version of dual.