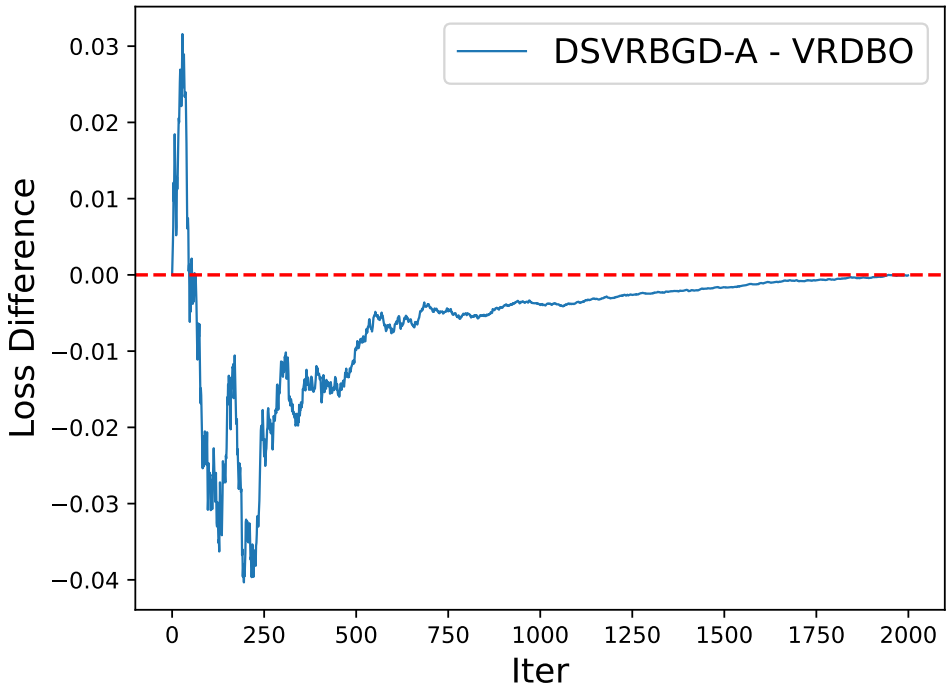
On the Communication Complexity of Decentralized Bilevel Optimization

Yihan Zhang My T. Thai Jie Wu Hongchang Gao*

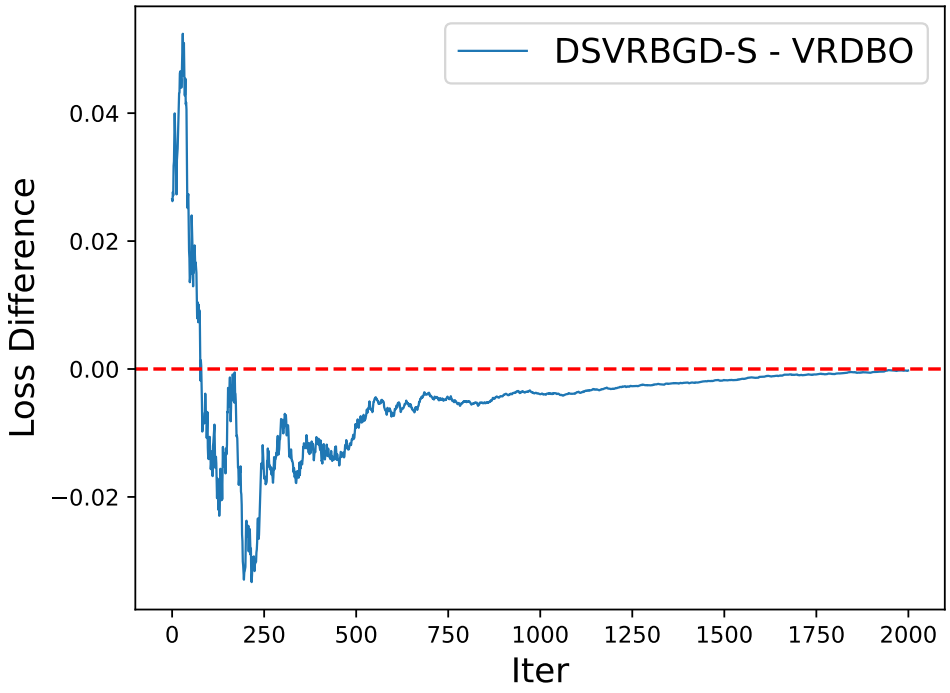
Abstract

Stochastic bilevel optimization finds widespread applications in machine learning, including meta-learning, hyperparameter optimization, and neural architecture search. To extend stochastic bilevel optimization to distributed data, several decentralized stochastic bilevel optimization algorithms have been developed. However, existing methods often suffer from slow convergence rates and high communication costs in heterogeneous settings, limiting their applicability to real-world tasks. To address these issues, we propose two novel decentralized stochastic bilevel gradient descent algorithms based on *simultaneous* and *alternating* update strategies. Our algorithms can achieve faster convergence rates and lower communication costs than existing methods. Importantly, our convergence analyses do not rely on strong assumptions regarding heterogeneity. More importantly, our theoretical analysis clearly discloses how the additional communication required for estimating hypergradient under the heterogeneous setting affects the convergence rate. To the best of our knowledge, this is the first time such favorable theoretical results have been achieved with mild assumptions in the heterogeneous setting. Furthermore, we demonstrate how to establish the convergence rate for the alternating update strategy when combined with the variance-reduced gradient. Finally, experimental results confirm the efficacy of our algorithms.

*Temple University, hongchang.gao@temple.edu



(a) $\text{LOSS}_{\text{DSVRBGD-A}} - \text{LOSS}_{\text{VRDBO}}$



(b) $\text{LOSS}_{\text{DSVRBGD-S}} - \text{LOSS}_{\text{VRDBO}}$

Figure 1: The loss difference between our two methods and VRDBO. **A negative difference indicates that our methods converge faster to the stationary point than VRDBO** (The potential reason for the positive difference in the first several steps is that the one-step gradient descent for estimating Hessian-inverse-vector product is not as good as the Neuman series expansion method in the initial stage.). There are 8 workers in this experiment. The training sample's feature x on the k-th worker is generated from a Guassian distribution $\mathcal{N}(\mu_k, \sigma_k)$, where μ_k is generated from a Uniform distribution $\mathcal{U}(-3, 3)$ and σ_k is generated from a Uniform distribution $\mathcal{U}(1, 25)$.