Rebuttal Response

Anonymous Authors[1]

1. Comparison Table of Stochastic Bilevel Optimization Algorithms

Table 1. Comparison of stochastic bilevel optimization algorithms in the nonconvex-strongly-convex setting under different smoothness assumptions on f and g. The oracle complexity stands for the number of oracle calls to stochastic gradients and stochastic Hessian/Jacobian-vector products to find an ϵ-stationary point. $\mathcal{C}_L^{a,k}$ denotes a-times differentiability with Lipschitz k-th order derivatives. "SC" means "strongly-convex". $\widetilde{O}(\cdot)$ compresses logarithmic factors of $1/\epsilon$ and $1/\delta$, where $\delta \in (0,1)$ denotes the failure probability.

Method [1]	Loop Style	Stochastic Setting	Oracle Complexity	Upper-Level f	Lower-Level g	Batch Size
BSA (Ghadimi & Wang, 2018)	Double	General expectation	$\widetilde{O}(\epsilon^{-6})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\widetilde{O}(1)$
StocBio (Ji et al., 2021)	Double	General expectation	$\widetilde{O}(\epsilon^{-4})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\widetilde{O}(\epsilon^{-2})$
AmIGO (Arbel & Mairal, 2021)	Double	General expectation	$\widetilde{O}(\epsilon^{-4})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$O(\epsilon^{-2})$
ALSET (Chen et al., 2021)	Double / Single [2]	General expectation	$O(\epsilon^{-4})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$O(1)$
TTSA (Hong et al., 2023)	Single	General expectation	$\widetilde{O}(\epsilon^{-5})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\widetilde{O}(1)$
F^{2}SA (Kwon et al., 2023)	Single	General expectation	$O(\epsilon^{-7})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$O(1)$
SOBA (Dagréou et al., 2022)	Single	Finite sum	$O(\epsilon^{-4})$	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	$O(1)$
SABA (Dagréou et al., 2022)	Single	Finite sum	$O(N^{4/3}\epsilon^{-2})$ [3]	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	$O(1)$
MA-SOBA (Chen et al., 2023)	Single	General expectation	$O(\epsilon^{-4})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$O(1)$
BO-REP (Hao et al., 2024)	Double	General expectation	$\widetilde{O}(\epsilon^{-4})$	$(L_{x,0}, L_{x,1}, L_{y,0}, L_{y,1})$-smooth	SC and $\mathcal{C}_L^{2,2}$	$O(1)$
SLIP (This work)	Single	General expectation	$\widetilde{O}(\epsilon^{-4})$	$(L_{x,0}, L_{x,1}, L_{y,0}, L_{y,1})$-smooth	SC and $\mathcal{C}_L^{2,2}$	$O(1)$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]We omit the comparison with variance reduction-based methods (except for SABA (Dagréou et al., 2022)) that may achieve $\widetilde{O}(\epsilon^{-3})$ complexity under additional mean-squared smoothness assumptions on both upper-level and lower-level problems, e.g., VRBO, MRBO (Yang et al., 2021); SUSTAIN (Khanduri et al., 2021); SVRB (Guo et al., 2021); or under the finite sum setting, e.g., SRBA (Dagréou et al., 2023).

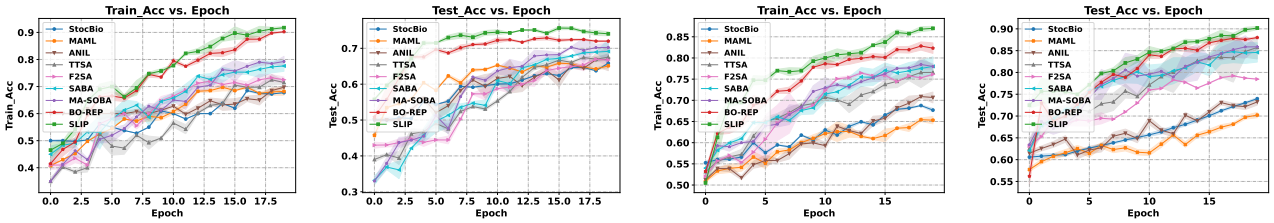
[2]ALSET can converge without the need for double loops, but at the cost of a worse dependence on $\kappa := l_{g,1}/\mu$ in oracle complexity.

[3]SABA (Dagréou et al., 2022) studies finite-sum problem and adapts variance reduction technique SAGA (Defazio et al., 2014), here $N = m + n$ denotes the total number of samples.

2. Experiments with Multiple Runs

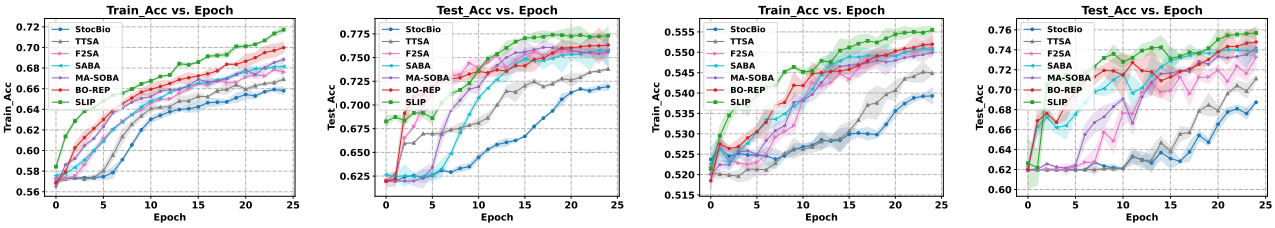
In this section, we rerun all experiments for multiple times and present the average accuracy and running time results. The results show that SLIP consistently exhibits low standard deviation and outperforms other baselines significantly.

2.1. Hyper-representation Learning and Data Hyper-cleaning



(a) Training accuracy on SNLI (b) Test accuracy on SNLI (c) Training accuracy on ARD (d) Test accuracy on ARD

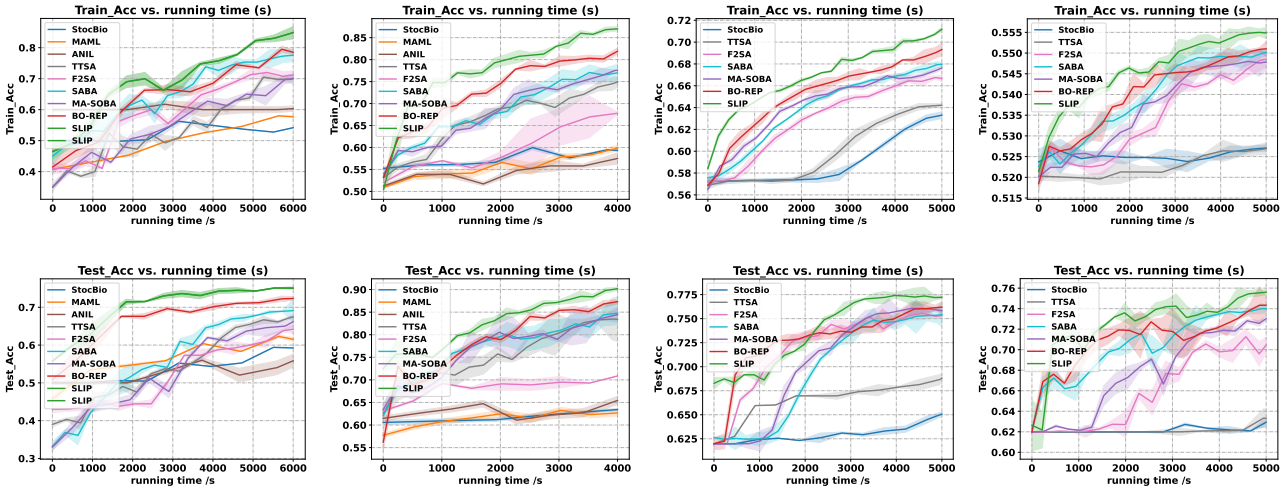
Figure 1. Comparison with bilevel optimization baselines on Hyper-representation. Figure (a) and (b) are the results in the SNLI dataset. Figures (c) and (d) are the results of the Amazon Review Dataset (ARD).



(a) Training acc with $p = 0.2$ (b) Test acc with $p = 0.2$ (c) Training acc with $p = 0.4$ (d) Test acc with $p = 0.4$

Figure 2. Comparison with bilevel optimization baselinses on data hyper-cleaning. Figure (a), (b) are the results with the corruption rate $p = 0.2$. Figure (c), (d) are the results with the corruption rate $p = 0.4$.

2.2. Running Time Comparison



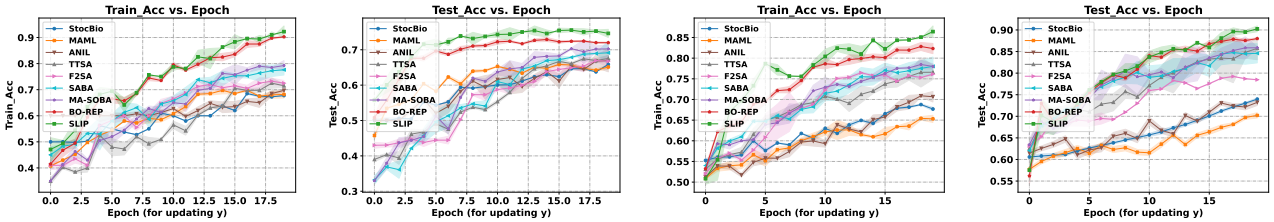
(a) Hyper-representation (SNLI) (b) Hyper-representation (ARD) (c) Data Hyper-cleaning (p=0.2) (d) Data Hyper-cleaning (p=0.4)

Figure 3. Comparision on running time. (a) Results of Hyper-representation on SNLI dataset. (b) Results of Hyper-representation on Amazon Review Dataset (ARD). (c), (d) Results of data Hyper-cleaning on Sentiment140 with corruption rate $p = 0.2$ and $p = 0.4$.

3. Experiments with Revised Epoch Definition

In this section, we clarify the notion of "epoch" in our previous experiments, where an epoch means a full pass over the validation set (for upper-level variable x update). For a more comprehensive comparison, we re-conduct experiments where an epoch means a full pass over the training set (for lower-level variable y update). In this case there are fewer x updates than the previous run due to the warm-start phase. The results show that SLIP is still empirically better than other baselines.

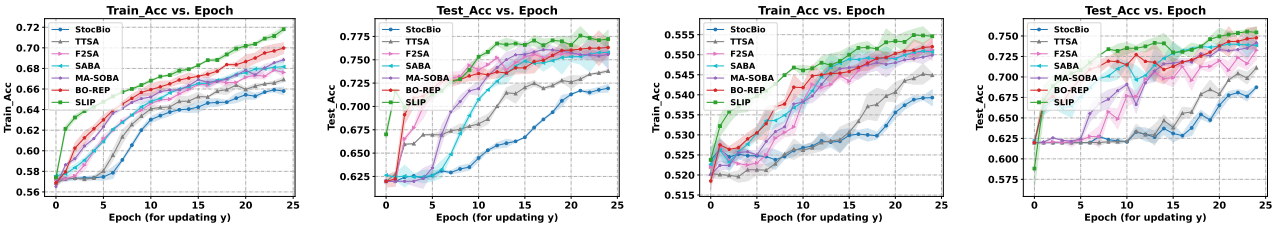
3.1. Hyper-representation Learning



(a) Training accuracy on SNLI (b) Test accuracy on SNLI (c) Training accuracy on ARD (d) Test accuracy on ARD

Figure 4. Comparison with bilevel optimization baselines on Hyper-representation. Figure (a) and (b) are the results in the SNLI dataset. Figures (c) and (d) are the results of the Amazon Review Dataset (ARD).

3.2. Data Hyper-cleaning



(a) Training acc with $p = 0.2$ (b) Test acc with $p = 0.2$ (c) Training acc with $p = 0.4$ (d) Test acc with $p = 0.4$

Figure 5. Comparison with bilevel optimization baselinses on data hyper-cleaning. Figure (a), (b) are the results with the corruption rate $p = 0.2$. Figure (c), (d) are the results with the corruption rate $p = 0.4$.

References

Arbel, M. and Mairal, J. Amortized implicit differentiation for stochastic bilevel optimization. *arXiv preprint arXiv:2111.14580*, 2021.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

Chen, X., Xiao, T., and Balasubramanian, K. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *arXiv preprint arXiv:2306.12067*, 2023.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

Dagréou, M., Moreau, T., Vaiter, S., and Ablin, P. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. *arXiv preprint arXiv:2302.08766*, 2023.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Ghadimi, S. and Wang, M. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Guo, Z., Hu, Q., Zhang, L., and Yang, T. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

Hao, J., Gong, X., and Liu, M. Bilevel optimization under unbounded smoothness: A new algorithm and convergence analysis. In *The Twelfth International Conference on Learning Representations*, 2024.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. D. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.