Penalty Method for Inversion-Free Deep Bilevel Optimization

Akshay Mehra and Jihun Hamm
Department of Computer Science, Tulane University.

Motivation

- Many machine learning applications have a bilevel problem at their core. For e.g.
  - Learning in presence of noisy labels:

$$\min_u \mathcal{L}_{val}(u,w^*)\ s.t.\ w^* = \arg\min_w \mathcal{L}_{w\_train}(u,w).$$

  - Few-shot and meta learning:

$$\min_u \sum_i \mathcal{L}_{val}(u,w^*)\ s.t.\ w_i^* = \arg\min_{w_i} \mathcal{L}_{train}(u,w_i)\ i=1,\ldots N.$$

  - Data poisoning attack:

$$\min_u \mathcal{L}_{val}(u,w^*)\ s.t.\ w^* = \arg\min_w \mathcal{L}_{poison}(u,w).$$

General bilevel formulation: $\min_{u\in\mathcal{U}} f(u,v^*(u))\ s.t.\ v^*(u) = \arg\min_{v\in\mathcal{V}(u)} g(u,v).$

- The steepest descent direction, $\frac{df}{du} = \nabla_u f - \nabla^2_{uv} g\, (\nabla^2_{vv} g)^{-1} \nabla_v f$, requires computing an inverse Hessian gradient product which can be computationally expensive for large-scale bilevel problems.

- Methods based on automatic differentiation or approximate inversion exist to compute the hypergradient, but their computational complexity can be quite high.

Contributions

- We propose a novel algorithm to solve bilevel problems which relies on the classical penalty function approach.

- We present a convergence analysis of our method and show that it converges to the KKT solution of the single level reformulation of a bilevel problem.

- Our algorithm avoids computing the hypergradient and uses a sequence of alternating minimizations, which finds the exact hypergradient asymptotically.

- Our algorithm is computationally efficient and achieves better or comparable performance to previous methods on several machine learning applications.

Penalty method for bilevel optimization

We consider bilevel problems with additional upper-level constraints:

$$\min_u f(u,v^*(u))$$
$$s.t.\ h(u,v^*(u)) = 0\ and\ v^*(u) = \arg\min_{v\in\mathcal{V}(u)} g(u,v).$$

Assuming the solution to the lower-level problem is unique, we can replace the lower-level problem with its necessary condition for minimality:

$$\min_u f(u,v)\ s.t.\ h(u,v) = 0\ and\ \nabla_v g = 0.$$

We apply the penalty function approach to this problem and show that the sequence of approximate (ϵ_k-optimal) solutions to the penalized unconstrained problem (below) converges to the KKT solution of the single level problem (above).

$$\min_{u,v} \tilde{f}(u,v) = \min_{u,v} f(u,v) + \frac{\gamma_k}{2}(\|h(u,v)\|^2 + \|\nabla_v g\|^2).$$

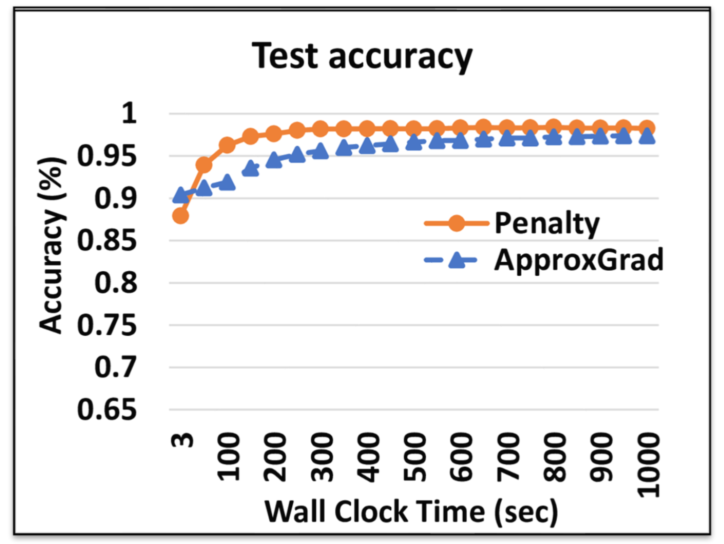
Main algorithm (Penalty)

for $k = 0,\ldots,K-1$ **do**
  while $\|\nabla_u \tilde{f}\|^2 + \|\nabla_v \tilde{f}\|^2 > \epsilon_k^2$ **do**
    for $t = 0,\ldots,T-1$ **do**
      $v_{t+1} \leftarrow v_t - \rho_{k,t}\nabla_v \tilde{f}$
    $u_{t+1} \leftarrow u_t - \sigma_k \nabla_u \tilde{f}$
  $\gamma_{k+1} \leftarrow c_\gamma \gamma_k,\ \epsilon_{k+1} \leftarrow c_\epsilon \epsilon_k$

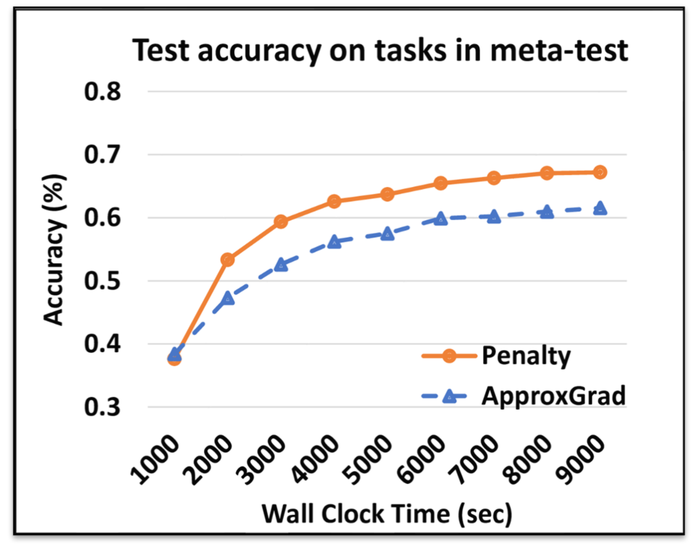
Complexity comparison

Method	Time	Space
FMD	O(cUT)	O(UV)
RMD	**O(cT)**	O(U+VT)
ApproxGrad	**O(cT)**	**O(U+V)**
Penalty	**O(cT)**	**O(U+V)**

Convergence speed comparison



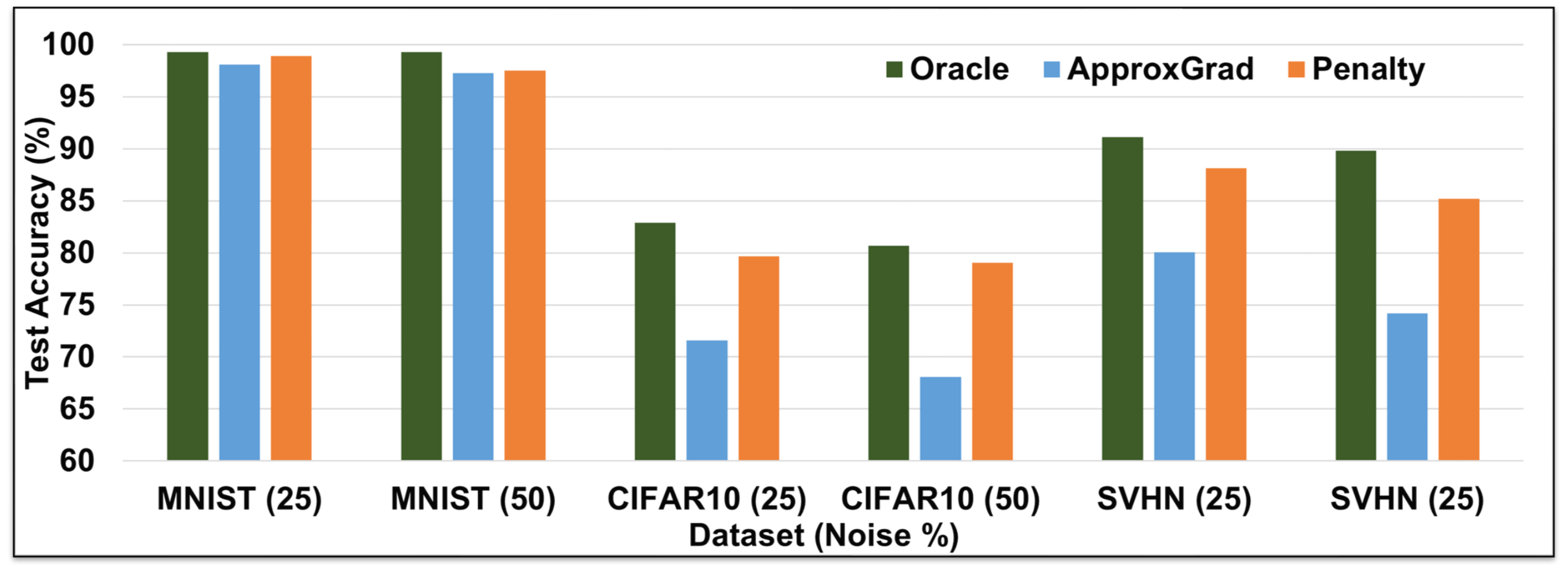
Data denoising



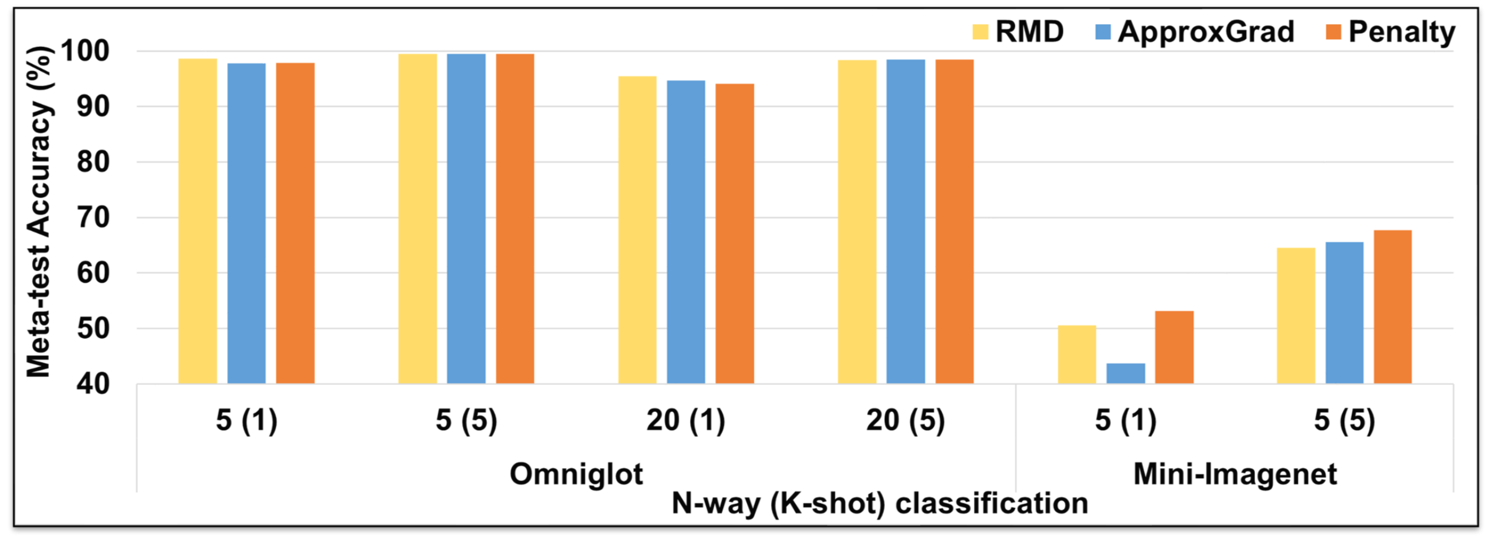
Few-shot learning

Performance comparison on machine learning applications

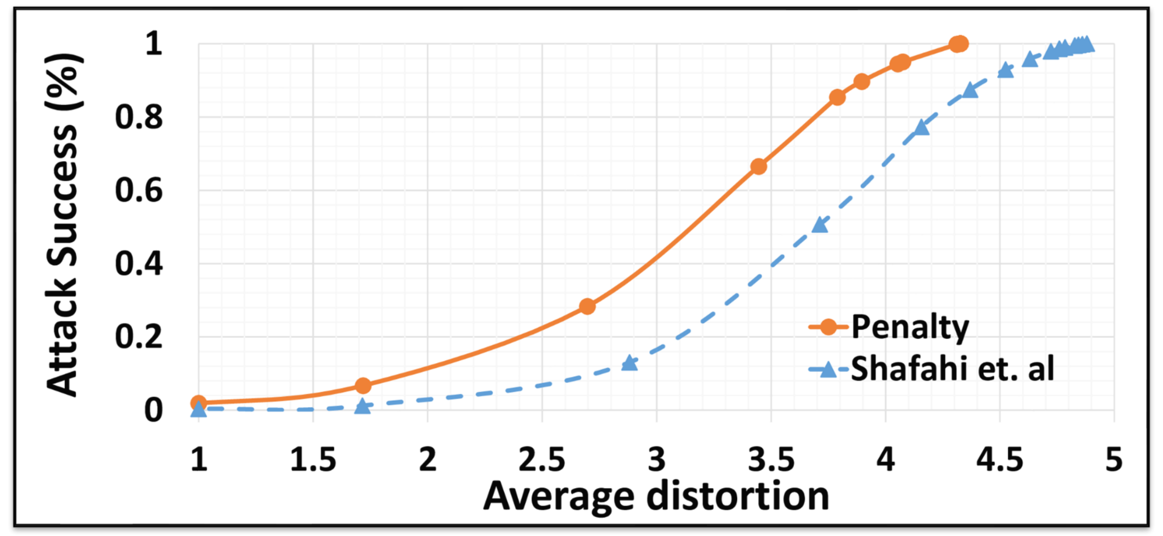
- Data denoising by importance re-weighting:



- Learning a common representation for few-shot learning:



- Clean label data poisoning attack:



References

- Bard. Practical bilevel optimization: algorithms and applications, volume 30. Springer Science & Business Media, 2013.
- Yo Ishizuka and Eitaro Aiyoshi. Double penalty method for bilevel optimization problems. Annals of Operations Research.
- Luca Franceschi et al. Forward and reverse gradient-based hyperparameter optimization. In International Conference on Machine Learning.
- Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In International conference on machine learning.
- Luca Franceschi et al. Bilevel programming for hyperparameter optimization and meta-learning.

This work was supported by the NSF EPSCoR-Louisiana Materials Design Alliance (LAMDA) program #OIA1946231