Convergence of Bayesian Bilevel Optimization

Shi Fu[1], Fengxiang He[2], Xinmei Tian[1] and Dacheng Tao[3]

[1]University of Science and Technology of China
[2]University of Edinburgh
[3]Nanyang Technological University

ICLR 2024

Motivation

- **Bayesian bilevel optimization** (BBO), integrating Bayesian optimization (BO) for outer-level hyperparameter tuning with inner-level model parameter optimization via SGD, shows significant promise in engineering applications.

- However, the working mechanisms and theoretical convergence guarantees of this approach remain unclear. Additionally, properly configuring the inner unit horizon presents challenges.

- In this paper, we demonstrate the sublinear regret bounds, showing simultaneous convergence of both inner-level model parameters and outer-level hyperparameters towards optimal configurations for **generalization capability**.

Problem Formulation

We consider bilevel optimization involving model parameters and hyperparameters:

$$\lambda^* = \arg\min_{\lambda \in \Lambda} L\left(\lambda, \theta_\lambda^*\right), \quad \text{where} \quad \theta_\lambda^* = \arg\min_{\theta \in \Theta} L\left(\lambda, \theta\right).$$

- The inner-level objective is to determine the optimal model parameters via SGD, denoted as θ_λ^*, for a given hyperparameter λ.
- At the outer level, the goal is to identify the optimal hyperparameter λ^* by BO, which determines its associated model parameters $\theta_{\lambda^*}^*$, collectively minimizing the expected error.

Definitions

Let constants $K, \gamma > 0$. Let $\ell : \Lambda \times \Theta \times \mathcal{Z} \to \mathbb{R}$. We have:

Lipschitzness: The loss $\ell(\lambda, \theta, z)$ is said to be K-Lipschitz continuous with respect to θ if $\|\ell(\lambda, \theta_1, z) - \ell(\lambda, \theta_2, z)\| \leq K\|\theta_1 - \theta_2\|$ for any $\theta_1, \theta_2, z, \lambda$.

Smoothness: The loss $\ell(\lambda, \theta, z)$ is said to be γ-Smooth with respect to θ if $\|\nabla_\theta \ell(\lambda, \theta_1, z) - \nabla_\theta \ell(\lambda, \theta_2, z)\| \leq \gamma\|\theta_1 - \theta_2\|$ for any $\theta_1, \theta_2, z, \lambda$.

Convexity: The loss $\ell(\lambda, \theta, z)$ is said to be convex with respect to θ if $\ell(\lambda, \theta_1, z) \geq \ell(\lambda, \theta_2, z) + \langle \nabla_\theta \ell(\lambda, \theta_2, z), \theta_1 - \theta_2 \rangle$ for any $\theta_1, \theta_2, z, \lambda$.

Regret

During t-th iteration at the outer level, the output hyperparameters are denoted as λ_t^+. Simultaneously, we optimize the model parameters through N steps of SGD. The regret is denoted as

$$R_T = \sum_{t=1}^{T} r_t, \quad \text{where} \quad r_t = L(\lambda^*, \theta^*_{\lambda^*}) - L(\lambda_t^+, \theta^N_{\lambda_t^+}).$$

Inner Level of BBO: Excess Risk Bound

Excess Risk Bound

Suppose that the function $\ell(\lambda, \theta, z)$ is K-Lipschitz continuous, γ-smooth and convex with respect to θ, uniformly bounded by M. We perform SGD with step sizes $\eta_j = \eta \asymp \frac{1}{\sqrt{N}} \leq 2/\gamma$ in the inner level. Choose $N \asymp n \asymp m$. Then, with a probability of at least $1 - \delta$, we have:

$$L^{val}(\lambda, \theta_\lambda^N, S^{val}) - L(\lambda, \theta_\lambda^*) = \mathcal{O}\left(N^{-\frac{1}{2}} \log^{3/2} N\right)$$

The difference between $L^{val}(\lambda, \theta_\lambda^N, S^{val})$ and $L(\lambda, \theta_\lambda^*)$ represents noise when evaluating $L(\lambda, \theta_\lambda^*)$ in outer optimization.

Outer Level of BBO: Regret Bound with EI Functions

Regret Bound with EI

Assume that $L(\lambda, \theta^*_\lambda)$ lies in the RKHS $\mathcal{H}_k(\Lambda)$. Consider the noise $\varepsilon_t = L^{val}\left(\lambda_t, \theta^N_{\lambda_t}, S^{val}\right) - L\left(\lambda_t, \theta^*_{\lambda_t}\right)$. Assume that $\|L(\cdot, \theta^*_\cdot)\|_k \leq B$ and define $\beta_t = \sqrt{B^2 + \sigma^{-2} t \varphi^2(N) N^{-1}}$. By using EI acquisition and the prior $GP_\Lambda(0, k(\cdot,\cdot))$, with probability at least $1-\delta$, the regret is bounded as:

$$R_T = \mathcal{O}\left(\frac{\beta_T^2 \sqrt{T\gamma_T}}{\tau(\beta_T) - \beta_T} + TN^{-\frac{1}{2}}\right),$$

If we select $N \asymp T$, we attain: $R_T = \mathcal{O}\left(\sqrt{T\gamma_T}\right)$.

Comparisions with Previous Works: Our regret bound saves $\sqrt{\gamma_T}$ compared to previous state-of-the-art results with noise assumptions more aligned with reality.

Practical Insights: The optimal number of inner-level SGD iterations is chosen as $N \asymp T$. Limited iterations such as $N \asymp \sqrt{T}$ lead to divergence, while excessive iterations waste resources. Moreover, we also demonstrate that for non-convex functions, the optimal is $N \asymp T^2$.

Outer Level of BBO: Regret Bound with UCB Functions

Regret Bound with UCB

Assume that $L(\lambda, \theta_\lambda^*)$ lies in the RKHS $\mathcal{H}_k(\Lambda)$. Furthermore, consider the noise $\varepsilon_t = L^{val}\left(\lambda_t, \theta_{\lambda_t}^N, S^{val}\right) - L\left(\lambda_t, \theta_{\lambda_t}^*\right)$. Then, assume that $\|L(\cdot, \theta_\cdot^*)\|_k \leq B$ and let $\beta_t = \sqrt{B^2 + \sigma^{-2} t \varphi^2(N) N^{-1}}$. By using UCB acquisition functions and the prior $GP_\Lambda(0, k(\cdot,\cdot))$, with probability at least $1-\delta$, the regret is bounded as:

$$R_T = \mathcal{O}\left(\sqrt{\left(B^2 + TN^{-1}\right) T\gamma_T} + TN^{-\frac{1}{2}}\right)$$

If we select $N \asymp T$, we can obtain that: $R_T = \mathcal{O}\left(B\sqrt{T\gamma_T}\right)$.

Comparisions with Previous Works: We obtain a regret bound $\mathcal{O}\left(B\sqrt{T\gamma_T}\right)$, which is tighter than the previous state-of-the-art bounds $\mathcal{O}\left(B\sqrt{T\gamma_T} + \gamma_T\sqrt{T}\right)$.

Practical Insights: The choice $N \asymp T$ is also reasonable for UCB function. However, UCB offers greater flexibility in choosing the number of inner-level SGD iterations compared to EI. For instance, if we choose $N \asymp \sqrt{T}$, the regret bound still exhibits sublinear growth in some cases.

Experiments

We conducted numerical experiments, using SGD to train a CNN on the MNIST, while employing BO with EI and UCB functions to adjust hyperparameters.

Expected Improvement

SGD Steps	**100**	**500**	**1000**	**2000**	**3000**
20 BO steps	3.15 ± 0.75	2.73 ± 0.03	2.70 ± 0.04	2.43 ± 0.12	2.46 ± 0.13
SGD Steps	**500**	**1000**	**2000**	**3000**	**4000**
40 BO steps	3.45 ± 0.85	2.42 ± 0.23	2.44 ± 0.55	2.39 ± 0.09	2.51 ± 0.09
SGD Steps	**1000**	**2000**	**3000**	**4000**	**6000**
60 BO steps	2.58 ± 0.58	2.40 ± 0.35	2.34 ± 0.28	2.10 ± 0.08	2.25 ± 0.17

The experiments are aligned with our theoretical analysis. Fixing the Bayesian optimization's iteration number, the loss function decreases as the SGD steps rise initially, while suboptimal hyperparameters cause high loss.

Experiments

Then we present the results when utilizing the UCB acquisition function.

Upper Confidence Bound

SGD Steps	**100**	**500**	**1000**	**2000**	**3000**
20 BO steps	2.93 ± 0.60	2.59 ± 0.47	2.49 ± 0.34	2.39 ± 0.07	2.31 ± 0.13
SGD Steps	**500**	**1000**	**2000**	**3000**	**4000**
40 BO steps	2.56 ± 0.53	2.34 ± 0.45	2.29 ± 0.29	2.27 ± 0.30	2.22 ± 0.16
SGD Steps	**1000**	**2000**	**3000**	**4000**	**6000**
60 BO steps	2.57 ± 0.27	2.26 ± 0.28	2.23 ± 0.17	2.19 ± 0.19	2.20 ± 0.10

We observe that initially, as the number of SGD steps increases, the loss decreases gradually. However, when the number of SGD steps becomes excessively large, the decrease in loss tends to plateau.

Thank you!