Appendix for "Online Constrained Meta-Learning: Provable Guarantees for Generalization"

A Notation checklist

Table 3: Notation checklist.

Notation	Definition
t	Round index
i	Constraint index ($1 \leq i \leq m$)
$\mathcal{T}_t$	Task at round t
$\mathcal{D}_{0,t}$	Data distribution of loss function for task $\mathcal{T}_t$
$\mathcal{D}_{i,t}$ $(1 \leq i \leq m)$	Data distribution of constraint functions for task $\mathcal{T}_t$
$\mathcal{D}_t = \{\mathcal{D}_{0,t}, \mathcal{D}_{1,t}, \ldots, \mathcal{D}_{m,t}\}$	Data distribution for task $\mathcal{T}_t$
$\mathcal{D}_{0,t}^{tr}$	Training dataset of loss function for task $\mathcal{T}_t$
$\mathcal{D}_{i,t}^{tr}$ $(1 \leq i \leq m)$	Training dataset of constraint functions for task $\mathcal{T}_t$
$\mathcal{D}_t = \{\mathcal{D}_{0,t}^{tr}, \mathcal{D}_{1,t}^{tr}, \ldots, \mathcal{D}_{m,t}^{tr}\}$	Training dataset for task $\mathcal{T}_t$
$\mathcal{D}_{0,t}^{val}$	Validation dataset of loss function for task $\mathcal{T}_t$
$\mathcal{D}_{i,t}^{val}$ $(1 \leq i \leq m)$	Validation dataset of constraint functions for task $\mathcal{T}_t$
$\mathcal{D}_t^{val} = \{\mathcal{D}_{0,t}^{val}, \cdots, \mathcal{D}_{m,t}^{val}\}$	Validation dataset for task $\mathcal{T}_t$
$\lvert\mathcal{D}_0^{tr}\rvert$	Number of data points in $\mathcal{D}_{0,t}^{tr}$ (same for task $\mathcal{T}_t$)
$\lvert\mathcal{D}_+^{tr}\rvert$	Number of data points in $\mathcal{D}_{i,t}^{tr}$ (same for task $\mathcal{T}_t$ and any i)
$\lvert\mathcal{D}_0^{val}\rvert$	Number of data points in $\mathcal{D}_{0,t}^{val}$ (same for task $\mathcal{T}_t$)
$\mathcal{L}^{val}(\theta, \mathcal{D}_{0,t}^{val})$	Loss of θ on $\mathcal{D}_{0,t}^{val}$
$p(\mathcal{T})$	Task distribution
θ_t^*	Optimal solution parameter for task $\mathcal{T}_t$
θ_t	Task-specific parameter for task $\mathcal{T}_t$
$\mathcal{D}ist(\phi, \mathcal{T}_{1:T})$	Distance between ϕ and optimal parameters of $\mathcal{T}_{1:T}$
$\mathcal{S}^*(p(\mathcal{T}))$	Task dissimilarity of $p(\mathcal{T})$
λ	regularization weight
$\mathcal{A}lg$	Within-task algorithm
ϕ / ϕ_t	Meta-parameter
$R_{0,t}$	Optimality gap
$R_{i,t}$	The i-th constraint violation
$\bar{R}_{0,[1:T]}$	Ttask-averaged optimality gap (TAOG)
$\bar{R}_{i,[1:T]}$	Task-averaged constraint violatio (TACV)

B Practical Algorithm and Implementation Details

In Sections 3 and 4, we develop a theoretically guaranteed algorithm with Assumptions 1 and 2. In many practical machine learning problems, the assumptions may not be satisfied. For example, some problems have non-convex loss functions ℓ_i. However, methods designed for convex loss functions, such as [47, 24], often perform well in non-convex settings. Inspired by these successes, we develop a practical instantiation of the online constrained meta-learning algorithm and evaluate its performance in Section 5.

Algorithm 2 states the practical algorithm of online constrained meta-learning and tests the performance of the deployed models for the sequentially revealed tasks. Compared with Algorithm 1, Algorithm 2 considers that (a) the constrained bilevel optimization problem in (4) has no closed-form solution; (b) the training data $\mathcal{D}_{0,t}^{tr}$ are limited and fixed for each task, which limits the generalization of the learned model; (c) since $\mathcal{S}^*(\mathcal{T}_{1:T})$ and $\mathcal{S}^*(p(\mathcal{T}))$ cannot be obtained before the task sequence is revealed, we cannot compute λ as shown in Theorem 1 and Corollary 1. To overcome these limitations, we propose Algorithm 2 that uses the stochastic optimization steps to optimize the meta-objective function.

Algorithm 2 Practical Algorithm of Online Constrained Meta-Learning

Require: Initial regularization weight $\lambda_1 > 0$; Initial meta parameter ϕ_1.

1: Initialize a empty task buffer $\mathcal{B}_{\mathcal{T}} \leftarrow \emptyset$
2: **for** $t = 1, \cdots, T$ **do**
3: Sample and restore the training datasets $\mathcal{D}_t^{tr}$ from the distributions $\mathcal{D}_t$ for task $\mathcal{T}_t$
4: Adapt task-specific parameter as $\theta_t = \mathcal{A}lg(\lambda_t, \phi_t, \mathcal{D}_t^{tr})$ and deploy θ_t for task $\mathcal{T}_t$
5: Test and record the performance of θ_t for task $\mathcal{T}_t$ if running a test for the algorithm
6: Sample and restore the evaluation dataset $\mathcal{D}_{0,t}^{val}$ from the distributions $\mathcal{D}_{0,t}$
7: Update the task buffer $\mathcal{B}_{\mathcal{T}} \leftarrow \mathcal{B}_{\mathcal{T}} \cup \{\mathcal{T}_t\}$
8: Initialize $\phi_{t+1}^{(0)} = \phi_t$, $\lambda_{t+1}^{(0)} = \lambda_t$
9: **for** $n = 1, \cdots, N_m$ **do**
10: Sample a task $\mathcal{T}_k$ from the buffer $\mathcal{B}_{\mathcal{T}}$, and pick the datasets $\mathcal{D}_k^{tr}$ and $\mathcal{D}_k^{val}$
11: Randomly allocate $\mathcal{D}_k^{tr} \cup \mathcal{D}_k^{val}$ to $\mathcal{D}_k^{tr+}$ and $\mathcal{D}_k^{val+}$ without overlapping
12: Solve $\theta^* = \mathcal{A}lg(\lambda_{t+1}^{(n-1)}, \phi_{t+1}^{(n-1)}, \mathcal{D}_k^{tr})$ defined in (3) to obtain θ^* and the multipliers μ^* by the primal-dual approach [7].
13: Compute $g^{(n)} = \nabla_\phi \mathcal{L}^{val}(\mathcal{A}lg(\lambda_{t+1}^{(n-1)}, \phi, \mathcal{D}_k^{tr+}), \mathcal{D}_{0,k}^{val+})|_{\phi=\phi_{t+1}^{(n-1)}}$ by (6)
14: Compute $q^{(n)} = \nabla_\lambda \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi_{t+1}^{(n-1)}, \mathcal{D}_k^{tr+}), \mathcal{D}_{0,k}^{val+})|_{\lambda=\lambda_{t+1}^{(n-1)}}$
15: $\phi_{t+1}^{(n)} = \phi_{t+1}^{(n-1)} - \eta_1 g^{(n)}$ and $\lambda_{t+1}^{(n)} = \lambda_{t+1}^{(n-1)} - \eta_2 q^{(n)}$ (or Adam [36])
16: **end for**
17: $\phi_{t+1} = \phi_{t+1}^{(N_m)}$ and $\lambda_{t+1} = \lambda_{t+1}^{(N_m)}$
18: **end for**

In the beginning of Algorithm 2, when a new task $\mathcal{T}_t$ is revealed at round t, similar to Algorithm 1, the agent samples the dataset $\mathcal{D}_t^{tr}$ and run $\theta_t = \mathcal{A}lg(\lambda_t, \phi_t, \mathcal{D}_t^{tr})$ to quickly adapt θ_t from the current meta-parameter ϕ_t and deploy it to $\mathcal{T}_t$. To test Algorithm 2 in Section 5, line 5 tests and records the performance of the deployed model on $\mathcal{T}_t$. Next, the evaluation dataset $\mathcal{D}_{0,t}^{val}$ used in the meta-objective function (4) is sampled. To optimize the meta-objective function, as shown in lines 7 to 15, we use multiple stochastic gradient descent steps. At step n, the agent randomly samples a task $\mathcal{T}_k$ from all previously revealed tasks and picks the training dataset $\mathcal{D}_k^{tr}$ and evaluation dataset $\mathcal{D}_k^{val}$, which is similar to the training data sampling at each step of the SGD. Then, the data included in $\mathcal{D}_k^{tr} \cup \mathcal{D}_k^{val}$ are randomly allocated to a new training dataset $\mathcal{D}_k^{tr+}$ and a new evaluation dataset $\mathcal{D}_k^{val+}$, and keep $|\mathcal{D}_k^{tr}| = |\mathcal{D}_k^{tr+}|$, $|\mathcal{D}_k^{val}| = |\mathcal{D}_k^{val+}|$, and no overlapping between $\mathcal{D}_k^{tr+}$ and $\mathcal{D}_k^{val+}$. As we expect the task-specific adaptation $\mathcal{A}lg(\lambda_t, \phi_t, \mathcal{D}_t^{tr})$ can perform well for any sampled dataset with the data number $|\mathcal{D}_t^{tr}|$, the dataset reallocation enables the agent to see a different training dataset $\mathcal{D}_k^{tr+}$, when the task $\mathcal{T}_k$ is repeatedly sampled at different step n (line 10), and thus improves the generalization of the model. When the task $\mathcal{T}_k$ is sampled at step n, the current stochastic gradient descent step focuses on the descent on $\mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+}), \mathcal{D}_{0,k}^{val+})$. According to the constrained bilevel optimization method shown in [55], we have

$$\nabla_\phi \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+}), \mathcal{D}_{0,k}^{val+}) = \nabla_\phi^\top \mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+}) \nabla_\theta \mathcal{L}^{val}(\theta, \mathcal{D}_{0,k}^{val+})|_{\theta=\theta^*} \tag{6}$$

where $\nabla_\phi \mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+}) = -M(\phi, \theta^*, \mu^*)^{-1} N(\phi, \theta^*, \mu^*)$,

$$M \triangleq \begin{bmatrix} \nabla_\theta^2 L & (\nabla_\theta p_1)^\top & \cdots & (\nabla_\theta p_m)^\top \\ \mu_1^* \nabla_\theta p_1 & p_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mu_m^* \nabla_\theta p_m & 0 & \cdots & p_m \end{bmatrix} \text{ and } N \triangleq \left[\nabla_{\theta\phi}^2 L^\top, 0, \cdots, 0\right]^\top.$$

Here, θ^* is the optimal solution of $\mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+})$, $p_i(\theta) \triangleq \frac{1}{|\mathcal{D}_{i,t}^{tr}|} \sum_{z \in \mathcal{D}_{i,t}^{tr}} \ell_i(\theta, z) - c_{i,t}$, $L(\phi, \theta, \mu^*) \triangleq \frac{1}{|\mathcal{D}_{0,t}^{tr}|} \sum_{z \in \mathcal{D}_{0,t}^{tr}} \ell_0(\theta, z) + \frac{\lambda}{2}\|\theta - \phi\|^2 + \sum_{i=1}^m \mu_i^* (\frac{1}{|\mathcal{D}_{i,t}^{tr}|} \sum_{z \in \mathcal{D}_{i,t}^{tr}} \ell_i(\theta, z) - c_{i,t})$ is the Lagrangian of the problem $\mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+})$, and μ^* is its Lagrangian multiplier. In the gradient computation (6), the optimal solution θ^* and the multiplier μ^* can be solved by the convex optimization algorithm [10, 6]. Then, in line 14, the gradient in (6) is applied to the gradient descent step with a learning rate η_1. After N_m steps of the stochastic gradient descent (or Adam [36]), the meta-parameter

ϕ_{t+1} for the next revealed task is updated. Similarly, the regularization weight λ_t can be optimized by the gradient descent (or Adam [36]), where the gradient $\nabla_\lambda \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_k^{tr+}), \mathcal{D}_{0,k}^{val+})$ is computed by replacing each ∇_ϕ by ∇_λ and replacing $\nabla_{\theta\phi}$ by $\nabla_{\theta\lambda}$ in (6).

C Experimental Supplement

All experiments are executed on a computer with a 4.10 GHz Intel Core i5 CPU and an RTX 3080 GPU.

C.1 Meta imitation learning

Problem formulation. Imitation learning [8] has been widely studied as a way to transfer human skills to robots. In [30, 29], imitation learning is formulated by kernelized movement primitives, and takes into account nonlinear hard constraints and obstacle avoidance. In particular, the states of the robot are modeled as the linear combination of basis functions; the demonstrations from humans are modeled by a Gaussian mixture model (GMM) where the parameters follow a Gaussian distribution. The approach minimizes the divergence between the distributions of the robot state model and the demonstrations, while the hard constraints are satisfied.

In this experiment, instead of the linear combination of basis functions, we model the states of the robot by a neural network. In particular, the robot's state $\xi(w,t)$, including the joint position $q(w,t) \in \mathbb{R}^{\mathcal{O}}$ and velocity $\dot{q}(w,t)$, is parameterized by w and is modeled as

$$\xi(w,t) = \begin{bmatrix} q(w,t) \\ \dot{q}(w,t) \end{bmatrix},$$

where $q(w,t)$ is a neural network and takes t as the input and the location $q(w,t)$ as the output. The demonstrations include H trajectories, where N time-state pairs are contained in each trajectory, and are denoted as $\{\{t_{n,h}, \hat{\xi}_{n,h}\}_{n=1}^N\}_{h=1}^H$ and are modeled by a GMM. Then, each demonstration state $\hat{\xi}_n$ associated with t_n is described by a conditional probability distribution with mean $\hat{\mu}_n$ and covariance $\hat{\Sigma}_n$, i.e., $\hat{\xi}_n \mid t_n \sim \mathcal{N}(\hat{\mu}_n, \hat{\Sigma}_n)$, where $\hat{\mu}_n$ and $\hat{\Sigma}_n$ can be computed by the GMM. Following the problem formulation in [30, 29], an imitation learning task is to solve the following constrained optimization problem:

$$\begin{aligned} &\min_w \mathop{\mathbb{E}}_{t\in[t_0,t_N]} \left[\frac{1}{2}(\xi(w,t) - \hat{\mu}_t)^\top \hat{\Sigma}_t^{-1}(\xi(w,t) - \hat{\mu}_t)\right] \\ &\text{s.t. } g_i(\xi(w,t)) \le c_i,\ \forall t \in [t_0, t_N],\ i = 1, \ldots, m, \end{aligned} \tag{7}$$

where g_i is the i-th state constraint and the total constraint number is m; $\hat{\mu}_t$ and $\hat{\Sigma}_t$ are the mean and variance of the demonstration state $\hat{\xi}(t)$, i.e., $\hat{\xi} \mid t \sim \mathcal{N}(\hat{\mu}_t, \hat{\Sigma}_t)$. Here, the training dataset $\{t_n, \hat{\mu}_n, \hat{\Sigma}_n\}_{n=1}^N$ is provided by the GMM. Note that the given demonstrations are collected under a no-collision environment and thus may not be able to avoid the collision.

Consider that a set of imitation learning tasks are revealed sequentially. In each round t, a new task of Problem (7) is revealed, and its data $\mathcal{D}_t^{tr} = \{t_n, \hat{\mu}_n, \hat{\Sigma}_n\}_{n=1}^N$ and the collision area denoted by $\{g_i\}_{i=1}^m$ and $\{c_i\}_{i=1}^m$ are given. In round T, a model w_T is required to be updated and deployed by the robot. We apply the proposed constrained meta-learning approach to solving this problem. In particular, we solve the following constrained bilevel optimization problem at round T:

$$\phi_T = \mathop{\mathrm{argmin}}_{\phi} \sum_{t=1}^{T-1} \mathcal{L}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}), \mathcal{D}_t^{val}),$$

with

$$\begin{aligned} \mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}) = &\mathop{\mathrm{argmin}}_{w}\ \mathcal{L}(w, \mathcal{D}_t^{tr}) + \frac{\lambda}{2}\|w - \phi\|^2 \\ \text{s.t. } &\frac{1}{N}\sum_{n=[N]} g_i(\xi(w, t_n)) \le c_i,\ i = 1, \ldots, m, \end{aligned}$$

where $\mathcal{L}(w, \mathcal{D}^{tr})$ is the loss function of the model parameter w on a dataset $\mathcal{D}^{tr} = \{t_n, \hat{\mu}_n, \hat{\Sigma}_n\}_{n=1}^N$ and $\mathcal{L}(w, \mathcal{D}^{tr}) = \frac{1}{N} \sum_{n=[N]} \frac{1}{2}(\xi(w, t_n) - \hat{\mu}_n)^\top \hat{\Sigma}_n^{-1}(\xi(w, t_n) - \hat{\mu}_n)$. After ϕ_T is updated, for a new task $\mathcal{T}_T$, the task-specific model can be computed by $w_T = \mathcal{A}lg(\lambda, \phi_T, \mathcal{D}_T^{tr})$.

Few-shot imitation learning. We use the demonstration data given by [30]. In different rounds, the robot needs to imitate the demonstration, and move and write the capital letter "A" with different sizes, angles, and locations, i.e., the different demonstration dataset $\mathcal{D}_t^{tr} = \{t_n, \hat{\mu}_n, \hat{\Sigma}_n\}_{n=1}^N$. The sizes, angles, and locations for the sequential tasks are sampled from a Gaussian distribution. The robot also needs to avoid a circle collision area, which defines m, g_1, and c_1 in (7) as $m = 1$, $c_1 = 0.5$ and $g_1(x) = \sigma(\sqrt{(x_1 - d_1)^2 + (x_2 - d_2)^2} - r)$, where $\sigma(x)$ is a barrier function and $= \frac{1}{\beta}\log(1 + \exp(\beta x))$, r is the radius of the collision area and $r = 6$, and (d_1, d_2) is the center of the collision area and is sampled from the Gaussian distribution $\mathcal{N}(0, 1)$ for each task. The full demonstrations for each imitation learning task contain 400 data points, but the robot only can obtain 20 data points in each round.

We model the position of the robot by a four-layer neural network with 128 which consists of an input layer of size 8, followed by 3 hidden layers of size 128 with the ReLU nonlinearities and an output layer of size 2. The neural network takes $\{t, t^2, t^3, t^4, \sin(t), \cos(2t), \sin(2t), \cos(2t)\}$ as the inputs and $q(w, t)$ as the outputs.

Improve full-shot imitation learning by meta-learning. In this experiment, the collision area is defined by $g_1(x) = \sigma(-\sqrt{(x_1 - d_1)^2 + (x_2 - d_2)^2} + r)$, where $r = 2$ and (d_1, d_2) is is sampled from the Gaussian distribution $\mathcal{N}(0, 1)$ for each task. For each task, the robot can access the full shot of demonstrations including 400 data points. Other settings are exactly the same as the few-shot imitation learning.

C.2 Few-shot image classification with robustness

Problem formulation. Similar to the problem formulation in [12], we formulate the problem of robust learning for a single image classification task $\mathcal{T}_t$ as:

$$\begin{aligned} \theta_t^* &= \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{z \sim \mathcal{D}_t} [\ell_0(\theta, z)] \\ \text{s.t. } & \mathbb{E}_{z \sim \mathcal{D}_{t,[P]}} [\ell(\theta, z)] - (1 + \alpha)\mathbb{E}_{z \sim \mathcal{D}_t} [\ell(\theta, z)] \leq 0. \end{aligned}$$

where $0 < \alpha < 1$ is the robustness tolerance parameter, and ℓ is the loss function. Here, $\mathcal{D}_t$ is the distribution of the original data, and $\mathcal{D}_{t,[P]}$ is the distribution of the perturbed data, which is generated by the Projected Gradient Descent (PGD) method on $\mathcal{D}_t$. This formulation minimizes the loss on the original data while maintaining the robustness, i.e., the loss on the perturbed data is constrained in an acceptable range.

In the problem of few-shot learning with robustness, only N-shot of training data is given, we expect the model to have both high accuracy on original test data and high accuracy on perturbed test data. We apply the proposed constrained meta-learning approach to solving this problem. In particular, we solve the following constrained bilevel optimization problem at round T:

$$\phi_T = \operatorname*{argmin}_{\phi} \sum_{t=1}^{T-1} \mathcal{L}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}), \mathcal{D}_t^{val}),$$

with

$$\begin{aligned} \mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}) = & \operatorname*{argmin}_{\theta} \mathcal{L}(\theta, \mathcal{D}_t^{tr}) + \frac{\lambda}{2}\|\theta - \phi\|^2 \\ \text{s.t. } & \mathcal{L}(\theta, \mathcal{D}_{t,[P]}^{tr}) - (1 + \alpha)\mathcal{L}(\theta, \mathcal{D}_t^{tr}) \leq 0. \end{aligned}$$

Here, $\mathcal{L}(\theta, \mathcal{D}^{tr})$ is the loss function of the model parameter θ on a dataset $\mathcal{D}^{tr}$. The dataset $\mathcal{D}_t^{tr}$ is the N-shot dataset for the meta-training (the support dataset), the dataset $\mathcal{D}_{t,[P]}^{tr}$ is also N-shot and generated by the PGD on $\mathcal{D}_t^{tr}$, and $\mathcal{D}_t^{val}$ is the dataset for the meta-validation (the query dataset). After ϕ_T is updated, for a new task $\mathcal{T}_T$, the task-specific model can be computed by $\mathcal{A}lg(\lambda, \phi_T, \mathcal{D}_T^{tr})$. In the experiment, α is selected as 0.3, λ is selected as $\lambda = 1.0$ for 5-shot learning, $\lambda = 8.0$ for 1-shot learning.

Experiments setting. Experiments are conducted on two datasets, CUB-200-2011 (referred to as CUB) [53] and mini-ImageNet [52]. We used the same class splits used in [27, 48, 57]. In particular, the mini-ImageNet dataset holds 64 classes for training data, 16 classes for validation data, and the remaining 20 classes as a test set; the CUB dataset holds 100 classes for training data, 50 classes for validation data, and the remaining 50 classes as a test set. The input images are resized to 84×84 for both two datasets and applied data augmentation following [57]. A four-layer convolutional neural network (Conv-4) is used as the backbone, which consists of four blocks. Each of which consists of a convolution layer with 64 kernels of size 3×3, stride 1, and zero padding, a batch normalization layer, a ReLU activation function, and a max-pooling layer with the pooling size 2×2. After the convolutional layers, a fully-connected linear layer with 5 neurons is used as a classifier to output the prediction for the input image.

During the online meta-training, in each round, we sample a task only from the training data classes and regard it as the revealed task, i.e., sample a 5-way k-shot learning task (5 classes and k images for each class). There are 200 rounds of online learning, and thus we sample 200 tasks from the training data. In the meta-test for Tables 2 and 4, we use the test dataset. From the test classes, we sample 600 tasks, i.e., 600 times of 5-way k-shot data sampling from the test classes, which means that the image classes in the 600 meta-test tasks are unseen in training tasks.

The optimizer Adam [36] with a learning rate of 0.001 is used for the optimization. The cross-entropy is selected as the loss function. The adversarial attack on the query set is performed by the PGD attack with a perturbation size $\epsilon = 2/255$ and it takes 7 iterative steps with the step size of 2.5ϵ. To guarantee a fair comparison, we keep all the above setting for all methods, including (i) MAML [23] with constraint penalty (CP); (ii) ProtoNet [49] with CP; (iii) BOIL [44] with CP; (iv) MAML with MOML [57]; (v) ProtoNet with MOML; (vi) BOIL with MOML; (vii) our constrained meta-learning approach.

Supplementary results. Table 4 and Fig. 5 show the test result on the dataset CUB. From the results, we can get similar conclusions to the experimental results on dataset mini-ImageNet. It is shown in Table 4 that our method significantly improves the PGD accuracy and the B-score than the benchmarks and keeps the clean accuracy comparable. Fig. 5 shows that our method outperforms the benchmarks in terms of both learning speed and test accuracy.

Table 4: Clean accuracy (abbreviated as "Clean Acc.") and PGD accuracy (abbreviated as "PGD Acc.") on the CUB dataset for 5-way 5-shot and 5-way 1-shot learning.

Setting	Method	Clean Acc.	PGD Acc.	B-score
1-shot	MAML + CP	49.60 ± 0.81	36.42 ± 0.49	41.89 ± 0.84
	MAML + MOML	48.66 ± 0.87	38.37 ± 0.90	42.75 ± 0.89
	ProtoNet + CP	48.04 ± 0.91	28.53 ± 0.85	35.42 ± 0.90
	ProtoNet + MOML	42.26 ± 0.89	32.19 ± 0.82	36.24 ± 0.85
	BOIL + CP	54.29 ± 0.83	33.65 ± 0.67	41.34 ± 0.71
	BOIL + MOML	52.15 ± 0.93	40.44 ± 0.94	45.55 ± 0.94
	CML (**ours**)	50.45 ± 0.73	$\mathbf{41.91} \pm 0.83$	$\mathbf{45.84} \pm 0.76$
5-shot	MAML + CP	68.50 ± 0.69	52.96 ± 0.87	59.63 ± 0.77
	MAML + MOML	67.57 ± 0.78	55.26 ± 0.87	60.68 ± 0.83
	ProtoNet + CP	72.51 ± 0.68	52.61 ± 0.77	60.81 ± 0.72
	ProtoNet + MOML	71.10 ± 0.74	56.11 ± 0.87	62.73 ± 0.76
	BOIL + CP	76.25 ± 0.60	44.86 ± 0.81	56.28 ± 0.73
	BOIL + MOML	71.03 ± 0.74	56.05 ± 0.84	62.65 ± 0.81
	CML (**ours**)	72.05 ± 0.73	$\mathbf{60.16} \pm 0.83$	$\mathbf{66.01} \pm 0.76$

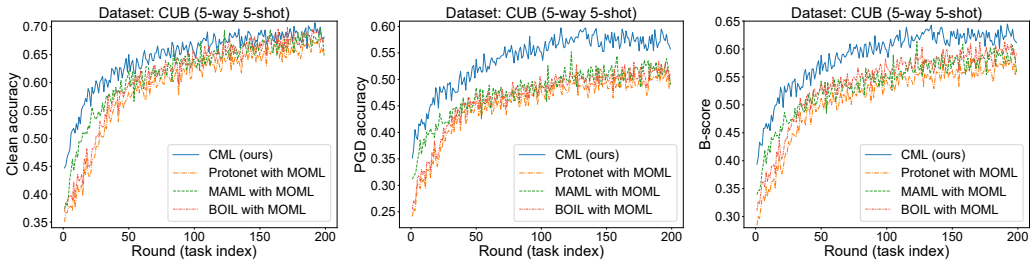

Figure 5: Test accuracy v.s. training task index on dataset CUB (5-way, 5-shot). **Left**: Clean accuracy; **Middle**: PGD accuracy; **Right**: B-score.

Analysis and Proof

D Intermediate Results

In this section, we list several results that will be helpful in proofs of our main results.

D.1 Sensitivity analysis for optimization

Consider that an optimization problem $P(\epsilon)$ which is parameterized by ϵ:

$$x^*(\epsilon) = \arg\min_x g(x,\epsilon) \ \text{ s.t. } \ h_i(x,\epsilon) \le 0, \ i \in I \triangleq \{1,\dots,m\}.$$

Suppose that g is twice continuously differentiable and h_i is twice continuously differentiable for all $i \in I$. We define the Lagrangian as

$$\mathcal{L}(x,\mu,\epsilon) \triangleq g(x,\epsilon) + \sum\nolimits_{i=1}^{m} \mu_i h_i(x,\epsilon),$$

where μ_i are Lagrange multipliers and $\mu \ge 0$. Suppose that the optimal solution $x^*(\epsilon)$ is unique for problem $P(\epsilon)$, we define the set $I(\epsilon) \triangleq \{i \in I \mid h_i(x^*(\epsilon),\epsilon) = 0\}$ and $I^C(\epsilon) \triangleq \{i \in I \mid h_i(x^*(\epsilon),\epsilon) < 0\}$.

Lemma 1 (Modified from [55],[22]). *Suppose that for all ϵ, the function $g(x,\epsilon)$ is strongly-convex w.r.t. x; $h_i(x,\epsilon)$ is convex w.r.t. x for each $i \in I$; the LICQ holds for $P(\epsilon)$. Then, the following properties hold for any ϵ.*

(i) The global minimum $x^*(\epsilon)$ of $P(\epsilon)$ exists and is unique. The Karush-Kuhn-Tucker (KKT) conditions hold at $x^*(\epsilon)$ with unique Lagrangian multipliers $\mu(\epsilon)$.

(ii) The vector function $z(\epsilon) \triangleq [x^*(\epsilon)^\top, \mu(\epsilon)^\top]^\top$ is continuous and locally Lipschitz. The directional derivative of $z(\epsilon)$ on any direction exists, and the directional derivative $\nabla_d x^*(\epsilon)$ is computed as

$$\begin{bmatrix} \nabla_d x^*(\epsilon) \\ \nabla_d \mu_{I(\epsilon,d)}(\epsilon) \end{bmatrix} = -M_D^{-1}(\epsilon,d) N_D(\epsilon,d) d$$

where $\mu_{I(\epsilon,d)}$ is a vector function that contain all μ_i with $i \in I(\epsilon,d) \subseteq I$, and

$$M_D(\epsilon,d) \triangleq \begin{bmatrix} \nabla_x^2 \mathcal{L} & \nabla_x h_{I(\epsilon,d)}^\top \\ \nabla_x h_{I(\epsilon,d)} & 0 \end{bmatrix} (x^*(\epsilon), \mu(\epsilon), \epsilon)$$

is nonsingular and

$$N_D(\epsilon,d) \triangleq [\nabla_{\epsilon x}^2 \mathcal{L}^\top, \nabla_\epsilon h_{I(\epsilon,d)}^\top]^\top (x^*(\epsilon), \mu(\epsilon), \epsilon).$$

Here, $h_{I(\epsilon,d)}$ is a vector function that contain all h_i with $i \in I(\epsilon,d) \subseteq I$.

(iii) If x^* is differentiable at ϵ, then the gradient is computed as

$$\begin{bmatrix} \nabla_\epsilon x^*(\epsilon) \\ \nabla_\epsilon \mu_{I(\epsilon)}(\epsilon) \end{bmatrix} = -M_+^{-1}(\epsilon) N_+(\epsilon)$$

and

$$\nabla_\epsilon \mu_{I^C(\epsilon)}(\epsilon) = 0,$$

where $\mu_{I(\epsilon)}$ is a vector function that contain all μ_i with $i \in I(\epsilon)$; $\mu_{I^C(\epsilon)}$ is a vector function that contain all μ_i with $i \in I^C(\epsilon)$;

$$M_+(\epsilon) \triangleq \begin{bmatrix} \nabla_x^2 \mathcal{L} & \nabla_x h_{I(\epsilon)}^\top \\ \nabla_x h_{I(\epsilon)} & 0 \end{bmatrix} (x^*(\epsilon), \mu(\epsilon), \epsilon)$$

is nonsingular and

$$N_+(\epsilon) \triangleq [\nabla_{\epsilon x}^2 \mathcal{L}^\top, \nabla_\epsilon h_{I(\epsilon)}^\top]^\top (x^*(\epsilon), \mu(\epsilon), \epsilon).$$

Here, $h_{I(\epsilon)}$ is a vector function that contain all h_i with $i \in I(\epsilon) \subseteq I$.

D.2 Sample average approximation for stochastic optimization

Consider the stochastic optimization problem with compound functions:

$$\min_{x \in X} H_0(x) = h_0 \left(x, \mathbb{E} f_1(x, \xi), \ldots, \mathbb{E} f_l(x, \xi)\right), \tag{8}$$

where $\xi \in \Xi \subseteq \mathbb{R}^c$ is a random variable defined on some probability space; the function $f_i(x, z) : X \times \Xi \to \mathbb{R}, i = 1, \ldots, l$, and $h_0(x, y) : X \times \mathbb{R}^l \to \mathbb{R}$.

The sample average approximation for the stochastic optimization problem (8) is:

$$\min_{x \in X} H_0^n \left(x, \xi^n\right) = h_0 \left(x, (1/n) \sum_{k=1}^n f_1 \left(x, \xi_k\right), \ldots, (1/n) \sum_{k=1}^n f_l \left(x, \xi_k\right)\right), \tag{9}$$

where each ξ_k is i.i.d sampled from its probability distribution. Let H^*, H_n^* be optimal values of problems (8) and (9), respectively; X^*, X_n^* be the sets of their optimal solutions; and

$$X_\epsilon^* = \{x \in X : H_0(x) \leq H^* + \epsilon\}$$
$$X_{n\epsilon}^* = \{x \in X : H_0^n \left(x, \xi^n\right) \leq H_n^* + \epsilon\}$$

be the sets of ϵ-approximate solutions in problems (8) and (9), respectively, and $\epsilon > 0$. Let

$$\Delta \left(X_n^*, X^*\right) = \sup_{x' \in X_n^*} \inf_{x \in X^*} \|x' - x\|, \quad \Delta \left(X_{\epsilon n}^*, X_\epsilon^*\right) = \sup_{x' \in X_{\epsilon n}^*} \inf_{x \in X_\epsilon^*} \|x' - x\|.$$

The following lemma states the rate of convergence of the sample average approximation method for compound stochastic optimization.

Lemma 2 (Simplified from Theorem 4.1 of [19])**.** *Suppose that $X \subset \mathbb{R}^d$ is a compact set with diameter D_X and the following assumptions are satisfied.*

(i) The function family $\{f_j(\cdot, z)\}_{z \in \Xi}$ is uniformly bounded by M for all j, i.e., $\sup_{x \in X, z \in \Xi} |f_j(x, z)| \leq M$; the function $f_j(\cdot, z)$ is Lipschitz continuous with constant L_c for all j and all $z \in \Xi$.

(ii) The function $h_0(x, y)$ satisfies that, $\sup_{x \in X, z \in \Xi} |h_0(x, 0)| < +\infty$, and $|h\left(x, y', z\right) - h\left(x, y'', z\right)| \leq L_h \|y' - y''\|$ for all $y', y'' \in \mathbb{R}^l$ and all $x \in X$.

Then, the following estimates hold ture:

$$\mathbb{E} |H_n^* - H^*| \leq \frac{C}{n^\alpha}$$

and

$$\mathbb{E} \Delta \left(X_{\epsilon n}^*, X_\epsilon^*\right) \leq \frac{2 C D_X}{\epsilon n^\alpha}.$$

where $C = 2 l N_f L_h$, $N_f = \sqrt{d}(L_c D_X + M/\sqrt{(1 - 2\alpha)e})$, and α can be arbitrarily determined in $\alpha \in (0, 1/2)$.

Remark 2. *In Lemma 2, we pick that $\alpha = \frac{1}{2} - \frac{1}{2 \ln n}$, we have that $\alpha \in (0, 1/2)$ with $n > 1$, then we have that*

$$\mathbb{E} |H_n^* - H^*| \leq \frac{2 l L_h \sqrt{d}(L_c D_X + M\sqrt{\ln n})}{\sqrt{n}} \tag{10}$$

and

$$\mathbb{E} \Delta \left(X_{\epsilon n}^*, X_\epsilon^*\right) \leq \frac{4 D_X l L_h \sqrt{d}(L_c D_X + M\sqrt{\ln n})}{\epsilon \sqrt{n}}. \tag{11}$$

D.3 Rademacher average

For a set of points $(z_1, \ldots, z_n) = z^n$ in Ξ and a sequence of functions $\{f(\cdot, z_i) : X \to \mathbb{R}\}_{i=1}^n$ define Rademacher average $R_n(f, z^n)$ as

$$R_n(f, z^n) \triangleq \mathbb{E}_\sigma \sup_{x \in X} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x, z_i) \right|$$

where σ_i are i.i.d random numbers such that $\sigma_i \in \{\pm 1\}$ with probabilities $1/2$; $\mathbb{E}_\sigma$ denotes mathematical expectation over $\sigma = (\sigma_1, \ldots, \sigma_n)$. Rademacher average of a family of functions $\{f(\cdot, z) : X \to \mathbb{R}\}_{z \in \Xi}$ is defined as

$$R_n(f, \Xi) \triangleq \sup_{z^n \sim \Xi} R_n(f, z^n).$$

The following results and their proofs are shown in Theorem 3.1 of [9] and Theorem 3.2 of [19].

Lemma 3. *For a random function $f(\cdot, \xi) : X \to \mathbb{R}$ and i.i.d random variables $(\xi_1, \ldots, \xi_n) = \xi^n$ sampled from Ξ. Then, we have*

$$\mathbb{E}_{\xi^n \sim \Xi} \left[\sup_{x \in X} \left| \frac{1}{n} \sum_{i=1}^n f(x, \xi_i) - \mathbb{E}_{\xi \in \Xi} f(x, \xi) \right| \right] \leq 2R_n(f, \Xi).$$

With probability at least $1 - \delta$,

$$\sup_{x \in X} \left| \frac{1}{n} \sum_{i=1}^n f(x, \xi) - \mathbb{E}_{z \in \Xi} f(x, z) \right| \leq 2R_n(f, \Xi) + \sqrt{\frac{2 \ln \frac{1}{\delta}}{n}}.$$

The following lemma states the Rademacher average of Lipschitz functions $f(\cdot, z_i)$.

Lemma 4 (Simplified from Theorem B.2 of [19])**.** *Let compact set $X \subset \mathbb{R}^d$ be contained in a cube with edge of length D_l, and functions $\{f(\cdot, z_i) : X \to \mathbb{R}\}_{i=1}^n$ be uniformly bounded by constant $M(z_i)$ and Lipschitz continuous on X with constants $L > 0$, i.e., $|f(x, z_i)| \leq M$ and $|f(x, z_i) - f(y, z_i)| \leq L\|x - y\|$ for any $x, y \in X$ and z_i. Then, for any z^n,*

$$R_n(f, z^n) \leq \sqrt{d}(LD + M\sqrt{\ln n})/\sqrt{n}. \tag{12}$$

Remark 3. *From the proofs of Lemmas 3 and 4 shown in [9, 19], it is easy to extend Lemmas 3 and 4 to the vector function $f(\cdot, z_i) : X \to \mathbb{R}^k$, i.e.,*

$$R_n(f, z^n) \triangleq \mathbb{E}_\sigma \sup_{x \in X} \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x, z_i) \right\|;$$

if $\|f(x, z_i)\| \leq M$ and $\|f(x, z_i) - f(y, z_i)\| \leq L\|x - y\|$ for any $x, y \in X$ and z_i.

$$\mathbb{E}_{\xi^n \sim \Xi} \left[\sup_{x \in X} \left\| \frac{1}{n} \sum_{i=1}^n f(x, \xi_i) - \mathbb{E}_{\xi \in \Xi} f(x, \xi) \right\| \right] \leq 2R_n(f, \Xi), \tag{13}$$

where $R_n(f, \Xi) = \sqrt{d}(LD + M\sqrt{\ln n})/\sqrt{n}$. With probability at least $1 - \delta$,

$$\sup_{x \in X} \left\| \frac{1}{n} \sum_{i=1}^n f(x, \xi) - \mathbb{E}_{z \in \Xi} f(x, z) \right\| \leq 2R_n(f, \Xi) + \sqrt{\frac{2 \ln \frac{1}{\delta}}{n}} \leq 2\sqrt{d}(LD + M\sqrt{\ln n})/\sqrt{n} + \sqrt{\frac{2 \ln \frac{1}{\delta}}{n}}. \tag{14}$$

Here, we show and prove a further result.

Lemma 5. *For the function f in Remark 3, we have*

$$\mathbb{E}_{\xi^n \sim \Xi} \left[\sup_{x \in X} \left\| \frac{1}{n} \sum_{i=1}^n f(x, \xi_i) - \mathbb{E}_{\xi \in \Xi} f(x, \xi) \right\|^2 \right] \leq 4R_n(f, \Xi)^2 + \frac{6R_n(f, \Xi)}{\sqrt{n}} + \frac{2}{n}.$$

Proof. When the random variable $Y > 0$, we compute the expectation as

$$E(Y) = \int_0^\infty (1 - F_Y(y))\,\mathrm{d}y,$$

where $F_Y(y)$ is the cumulative distribution function. Let $Y = \sup_{x\in X} \left\| \frac{1}{n}\sum_{i=1}^n f\left(x,\xi_i\right) - \mathbb{E}_{\xi\in\Xi} f(x,\xi)\right\|^2$. From (14), we have that,

$$F_Y\left((2R_n(f,\Xi) + \sqrt{\frac{2\ln\frac{1}{\delta}}{n}})^2\right) = P\left(Y \le (2R_n(f,\Xi) + \sqrt{\frac{2\ln\frac{1}{\delta}}{n}})^2\right) \ge 1-\delta,$$

i.e.,

$$1 - F_Y(y) \le \exp(-\frac{n(\sqrt{y} - 2R_n(f,\Xi))^2}{2}).$$

Moreover, when $\sqrt{y} \le 2R_n(f,\Xi)$, we have

$$1 - F_Y(y) \le 1.$$

Then,

$$\begin{aligned}
E(Y) \le& \int_0^{4R_n(f,\Xi)^2} 1\mathrm{d}y + \int_{4R_n(f,\Xi)^2}^\infty \exp(-\frac{n(\sqrt{y} - 2R_n(f,\Xi))^2}{2})\mathrm{d}y \\
=& 4R_n(f,\Xi)^2 + \int_0^\infty \exp(-\frac{nt^2}{2})\mathrm{d}(t + 2R_n(f,\Xi))^2 \\
=& 4R_n(f,\Xi)^2 + \int_0^\infty 4R_n(f,\Xi)\exp(-\frac{nt^2}{2})\mathrm{d}t + \int_0^\infty 2t\exp(-\frac{nt^2}{2})\mathrm{d}t \\
=& 4R_n(f,\Xi)^2 + 4R_n(f,\Xi)\sqrt{\frac{\pi}{2n}} + \frac{2}{n} \\
\le& 4R_n(f,\Xi)^2 + 6R_n(f,\Xi)\sqrt{\frac{1}{n}} + \frac{2}{n}.
\end{aligned}$$

□

Then, from (12), we have

$$\mathbb{E}_{\xi^n\sim\Xi}\left[\sup_{x\in X}\left\|\frac{1}{n}\sum_{i=1}^n f\left(x,\xi_i\right) - \mathbb{E}_{\xi\in\Xi} f(x,\xi)\right\|^2\right] \le \frac{\sqrt{d}(LD_l + M\sqrt{\ln n})(4LD_l\sqrt{d} + 4M\sqrt{d\ln n} + 6)}{n},$$

where $2/n$ is omitted.

D.4 Non-convex online learning

Here, we provide a review of the online algorithms we use. Paper [50] studies the problem of online learning with non-convex losses and proposes the Follow-the-Perturbed-Leader (FTPL) algorithm.

At each round t, the FTPL algorithm minimizes a perturbed summation of the loss functions revealed from round 1 to $t-1$. In particular, at each round t, the parameter x_t is obtained by following optimization problem:

$$x_{t+1} = \operatorname*{argmin}_x \sum_{t'=1}^t f_{t'}(x) - \sigma_t^\top \phi,$$

where the random perturbed vector $\sigma_t \in \mathbb{R}^d$ is i.i.d sampled by $\{\sigma_{t,j}\}_{j=1}^d \sim \mathrm{Exp}(\eta)$ with a contant $\eta > 0$ at each t.

Lemma 6 (Simplified from Theorem 1 of [50])**.** *Let D_l be the ℓ_∞ diameter of $\mathcal{X}$. Suppose the losses encountered by the learner are L-Lipschitz w.r.t ℓ_1 norm. For any fixed η, the predictions of FTPL satisfy the following regret bound:*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T f_t\left(x_t\right) - \frac{1}{T}\inf_{x\in\mathcal{X}}\sum_{t=1}^T f_t(x)\right] \le \mathcal{O}\left(\eta d^2 D_l L^2 + \frac{dD_l}{\eta T}\right).$$

E Proofs of Propositions 1 and 2

We first define some notations used in proofs of our results. For data distributions $\mathcal{D} = \{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_m\}$ and the training datasets $\{\mathcal{D}_1^{tr}, \mathcal{D}_0^{tr}, \ldots, \mathcal{D}_m^{tr}\}$, we denote $\mathcal{D}_+ \triangleq \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ and $\mathcal{D}_+^{tr} \triangleq \{\mathcal{D}_1^{tr}, \ldots, \mathcal{D}_m^{tr}\}$, which are data distributions and training datasets for constraint satisfaction. Similarly, for data distributions $\mathcal{D}_t = \{\mathcal{D}_{0,t}, \mathcal{D}_{1,t}, \ldots, \mathcal{D}_{m,t}\}$ and the training datasets $\{\mathcal{D}_{1,t}^{tr}, \mathcal{D}_{0,t}^{tr}, \ldots, \mathcal{D}_{m,t}^{tr}\}$, denote $\mathcal{D}_{+,t} \triangleq \{\mathcal{D}_{1,t}, \ldots, \mathcal{D}_{m,t}\}$ and $\mathcal{D}_{+,t}^{tr} \triangleq \{\mathcal{D}_{1,t}^{tr}, \ldots, \mathcal{D}_{m,t}^{tr}\}$. Recall the notations defined in Section 1, we use $\mathcal{D}_i^{tr} \sim \mathcal{D}_i$ to represent that all elements of datasets $\mathcal{D}_i^{tr}$ are i.i.d sampled from the distribution $\mathcal{D}_i$, $\mathcal{D}^{tr} \sim \mathcal{D}$ to represent $\mathcal{D}_i^{tr} \sim \mathcal{D}_i$ for all $0 \le i \le m$, and $\mathcal{D}_+^{tr} \sim \mathcal{D}_+$ to represent $\mathcal{D}_i^{tr} \sim \mathcal{D}_i$ for all $1 \le i \le m$. Similar for $\mathcal{D}_{i,t}^{tr} \sim \mathcal{D}_{i,t}$ and $\mathcal{D}_{+,t}^{tr} \sim \mathcal{D}_{+,t}$

Denote θ^* as the optimal solution of problem

$$\begin{aligned} &\min_{\theta\in\Theta} \ \mathbb{E}_{z\sim\mathcal{D}_0}\left[\ell_0(\theta, z)\right] \\ &\text{s.t.} \ \ \mathbb{E}_{z\sim\mathcal{D}_i}\left[\ell_i(\theta, z)\right] \le c_i, \ i = 1, \ldots, m. \end{aligned}$$

Denote $\theta_h(\mathcal{D}_0, \mathcal{D}_+)$ as the optimal solution of the optimization problem

$$\begin{aligned} &\min_{\theta\in\Theta} \ \mathbb{E}_{z\sim\mathcal{D}_0}\left[\ell_0(\theta, z)\right] + \frac{\lambda}{2}\|\theta - h\|^2 \\ &\text{s.t.} \ \ \mathbb{E}_{z\sim\mathcal{D}_i}\left[\ell_i(\theta, z)\right] \le c_i, \ i = 1, \ldots, m. \end{aligned} \tag{15}$$

Denote $\theta_h(\mathcal{D}_0, \mathcal{D}_+^{tr})$ as the optimal solution of the optimization problem

$$\begin{aligned} &\min_{\theta\in\Theta} \ \mathbb{E}_{z\sim\mathcal{D}_0}\left[\ell_0(\theta, z)\right] + \frac{\lambda}{2}\|\theta - h\|^2 \\ &\text{s.t.} \ \ \frac{1}{|\mathcal{D}_i^{tr}|}\sum_{z\in\mathcal{D}_i^{tr}} \ell_i(\theta, z) \le c_i, \ i = 1, \ldots, m. \end{aligned} \tag{16}$$

Denote $\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)$ as the optimal solution of the optimization problem

$$\begin{aligned} &\min_{\theta\in\Theta} \ \frac{1}{|\mathcal{D}_0^{tr}|}\sum_{z\in\mathcal{D}_0^{tr}} \ell_0(\theta, z) + \frac{\lambda}{2}\|\theta - h\|^2 \\ &\text{s.t.} \ \ \mathbb{E}_{z\sim\mathcal{D}_i}\left[\ell_i(\theta, z)\right] \le c_i, \ i = 1, \ldots, m. \end{aligned} \tag{17}$$

Denote $\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+^{tr}) = \mathcal{A}lg(\lambda, h, \mathcal{D}^{tr})$ as the optimal solution of the optimization problem

$$\begin{aligned} &\min_{\theta\in\Theta} \ \frac{1}{|\mathcal{D}_0^{tr}|}\sum_{z\in\mathcal{D}_0^{tr}} \ell_0(\theta, z) + \frac{\lambda}{2}\|\theta - h\|^2 \\ &\text{s.t.} \ \ \frac{1}{|\mathcal{D}_i^{tr}|}\sum_{z\in\mathcal{D}_i^{tr}} \ell_i(\theta, z) \le c_i, \ i = 1, \ldots, m. \end{aligned} \tag{18}$$

Define the feasible set in Problems (15) and (17) as $\mathcal{K}^* \triangleq \{\theta \mid \mathbb{E}_{z\sim\mathcal{D}_i}\left[\ell_i(\theta, z)\right] \le c_i, \ i = 1, \ldots, m\}$, and define the feasible set in Problems (16) and (18) as $\mathcal{K}^{tr} \triangleq \{\theta \mid \frac{1}{|\mathcal{D}_i^{tr}|}\sum_{z\in\mathcal{D}_i^{tr}} \ell_i(\theta, z) \le c_i, \ i = 1, \ldots, m\}$.

Define some functions for the following lemmas and the proof of Propositions 1 and 2.

$$\mathcal{L}_{\mathcal{D}_0}(\theta) \triangleq \mathbb{E}_{z\sim\mathcal{D}_0}\left[\ell_0(\theta, z)\right],$$

$$\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) \triangleq \frac{1}{|\mathcal{D}_0^{tr}|}\sum_{z\in\mathcal{D}_0^{tr}} \ell_0(\theta, z),$$

$$\mathcal{L}_{\mathcal{D}_0,h}(\theta) \triangleq \mathcal{L}_{\mathcal{D}_0}(\theta) + \frac{\lambda}{2}\|\theta - h\|^2,$$

$$\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) \triangleq \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) + \frac{\lambda}{2}\|\theta - h\|^2,$$

$$\mathcal{C}_{\mathcal{D}_i}(\theta) \triangleq \mathbb{E}_{z\sim\mathcal{D}_i}\left[\ell_i(\theta, z)\right],$$
$$\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta) \triangleq \frac{1}{|\mathcal{D}_i^{tr}|}\sum_{z\in\mathcal{D}_i^{tr}} \ell_i(\theta, z).$$

To simplify the notation, we restate Assumptions 1 and 2 for a single constrained learning task as Assumptions 3 and 4.

Assumption 3 (Constraint qualifications)**.** *(i) For the data distributions $\mathcal{D}$ and training datasets $\mathcal{D}^{tr}$ are sampled from $\mathcal{D}$, with probability* 1*, the feasible sets $\mathcal{K}^*$ in* (15) *and $\mathcal{K}^{tr}$ in* (18) *are bounded in a compact set with diameter $\mathcal{B}$.*

(ii) For the data distributions $\mathcal{D}$, the Slater's condition (SC) holds for Problem (15) *with the margin $\mathcal{C} > 0$. For datasets $\mathcal{D}^{tr}$ sampled from $\mathcal{D}$, with probability* 1*, the SC holds for Problem* (18) *with $\mathcal{C}$.*

(iii) For the data distributions $\mathcal{D}$, the linear independence constraint qualification (LICQ) holds for Problem (15)*.*

Assumption 4 (Function properties)**.** *(i) For any $z \in \mathcal{Z}$, the loss function $\ell_0(\cdot, z)$ and the constraint functions $\ell_1(\cdot, z), \cdots, \ell_m(\cdot, z)$ are twice continuously differentiable.*

(ii) For any $z \in \mathcal{Z}$, $\ell_0(\cdot, z)$ is L_0-Lipschitz, i.e., $\|\ell_0(w, z) - \ell_0(u, z)\| \leq L_0\|w - u\|$ for any $w, u \in \mathbb{R}^d$.

(iii) For any $z \in \mathcal{Z}$, $\ell_0(\cdot, z)$ is ρ-smooth, i.e., $\|\nabla\ell_0(w, z) - \nabla\ell_0(u, z)\| \leq \rho\|w - u\|$ for any $w, u \in \mathbb{R}^d$.

(iv) For any $z \in \mathcal{Z}$ and $1 \leq i \leq m$, $\ell_i(\cdot, z)$ is L_c-Lipschitz, i.e., $\|\ell_i(w, z) - \ell_i(u, z)\| \leq L_c\|w - u\|$ for any $w, u \in \mathbb{R}^d$.

(v) For any $z \in \mathcal{Z}$ and $0 \leq i \leq m$, $\ell_i(\cdot, z)$ is convex.

(vi) For any $z \in \mathcal{Z}$, and $1 \leq i \leq m$, $\ell_i(w, z) - c_i$ are bounded by M.

We introduce some lemmas required in the proof of Propositions 1 and 2.

Lemma 7. *Suppose that the vector function $\theta : [a, b] \to \mathbb{R}^d$ is Lipschitz on $[a, b]$ and the left derivative $\nabla_-\theta$ and the right derivative $\nabla_+\theta$ of θ always exists. Then, for any $\epsilon', \epsilon'' \in [a, b]$.*

$$\|\theta(\epsilon') - \theta(\epsilon'')\| \leq sup_{\epsilon\in[0,1]}\{\max\left(\|\nabla_-\theta(\epsilon)\|, \|\nabla_+\theta(\epsilon)\|\right)\}\|\epsilon' - \epsilon''\|.$$

Proof. Assume that there exists $\epsilon', \epsilon'' \in [a, b]$,

$$\|\theta(\epsilon') - \theta(\epsilon'')\| = sup_{\epsilon\in[a,b]}\{\max\left(\|\nabla_-\theta(\epsilon)\|, \|\nabla_+\theta(\epsilon)\|\right)\}\|\epsilon' - \epsilon''\| + c$$

where $c > 0$, then

$$\sum_{k=0}^{n}\|\theta(\epsilon_{k+1}) - \theta(\epsilon_k)\| \geq \|\theta(\epsilon'') - \theta(\epsilon')\| = sup_{\epsilon\in[a,b]}\{\max\left(\|\nabla_-\theta(\epsilon)\|, \|\nabla_+\theta(\epsilon)\|\right)\}\|\epsilon' - \epsilon''\| + c,$$

where $\epsilon_0 = \epsilon' < \cdots < \epsilon_k < \epsilon_{k+1} < \cdots < \epsilon_n = \epsilon''$. Then, for the largest Lipschitz L_k constant on $[\epsilon_k, \epsilon_{k+1}]$, there must have $L_k \geq sup_{\epsilon\in[0,1]}\{\max\left(\|\nabla_-\theta(\epsilon)\|, \|\nabla_+\theta(\epsilon)\|\right)\} + \frac{c}{\|\epsilon' - \epsilon''\|}$. For a sufficient large n, from the definition of the left and right derivative, we have $L_k \leq \max\left(\|\nabla_-\theta_k(\epsilon)\|, \|\nabla_+\theta(\epsilon_k)\|\right) + \delta$ for arbitrarily small $\delta > 0$, which contradicts that $L_k \geq sup_{\epsilon\in[0,1]}\{\max\left(\|\nabla_-\theta(\epsilon)\|, \|\nabla_+\theta(\epsilon)\|\right)\} + \frac{c}{\|\epsilon' - \epsilon''\|}$. Therefore,

$$\|\theta(\epsilon') - \theta(\epsilon'')\| \leq sup_{\epsilon\in[0,1]}\{\max\left(\|\nabla_-\theta(\epsilon)\|, \|\nabla_+\theta(\epsilon)\|\right)\}\|\epsilon' - \epsilon''\|.$$

□

Lemma 8. *Consider an optimization problem*

$$\min g(x) \;\; \textit{s.t. } h_i(x) \leq 0,\; i \in \{1, \ldots, m\}. \tag{19}$$

Suppose that the functions g and $h_1, \cdots, h_m$ are convex; the Slater's condition holds for Problem (19) *with the margin $\mathcal{C} > 0$; the constraint set $\mathcal{K} = \{x \mid h_i(x) \leq 0,\; i \in \{1, \ldots, m\}$ is bounded with*

the diameter $\mathcal{B}$; the function g is Lipschitz continuous with Lipschitz constant L_0. Then, Problem (19) *is equivalent to the following problem:*

$$\min g(x) + L^* \sum_{i=1}^{m} \max\{0, h_i(x)\}, \tag{20}$$

where $L^* = L_0\mathcal{B}/\mathcal{C}$.

Proof. Since the optimization problem (19) is convex and the Slater's condition holds, then the KKT condition holds at the optimal solution x^* with the Lagrangian multiplier μ^*, and the strong duality holds.

(a) We first show that Problem (19) is equivalent to Problem (20), if L^* is selected as $L^* \geq \max_i \{u_i^*\}$.

Let x^* be an optimal solution solution for Problem (19), for any x we have

$$g(x) + L^* \sum_{i=1}^{m} \max\{0, h_i(x)\} \geq g(x) + \sum_{i=1}^{m} \max\{0, u_i^* h_i(x)\} \geq g(x) + \sum_{i=1}^{m} u_i^* h_i(x).$$

From the Lagrangian saddle-point theorem in [6], since the the strong duality holds and (x^*, u^*) is a pair of primal and dual optimal solutions,

$$g(x) + \sum_{i=1}^{m} u_i^* h_i(x) \geq g(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) \tag{21}$$

for any x. From the KKT condition at (x^*, u^*), we have $\sum_{i=1}^{m} u_i^* h_i(x^*) = 0$. Then,

$$g(x) + L^* \sum_{i=1}^{m} \max\{0, h_i(x)\} \geq g(x^*).$$

for any x. Thus, x^* is an optimal solution of Problem (20).

Let $\bar{x}$ be an optimal solution solution for Problem (20), then

$$g(\bar{x}) + L^* \sum_{i=1}^{m} \max\{0, h_i(\bar{x})\} \leq g(x^*) + L^* \sum_{i=1}^{m} \max\{0, h_i(x^*)\} = g(x^*).$$

If $\bar{x} \in \mathcal{K}$, then $g(\bar{x}) + L^* \sum_{i=1}^{m} \max\{0, h_i(\bar{x})\} = g(\bar{x}) \leq g(x^*) \leq g(x)$ for any $x \in \mathcal{K}$, then $\bar{x}$ is an optimal solution of Problem (19).

If $\bar{x} \notin \mathcal{K}$, then

$$\begin{aligned} g(\bar{x}) + L^* \sum_{i=1}^{m} \max\{0, h_i(\bar{x})\} &\geq g(\bar{x}) + \sum_{i=1}^{m} \max\{0, u_i^* h_i(\bar{x})\} \\ &> g(\bar{x}) + \sum_{i=1}^{m} u_i^* h_i(\bar{x}) \geq g(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) = g(x^*). \end{aligned}$$

Then, $g(\bar{x}) + L^* \sum_{i=1}^{m} \max\{0, h_i(\bar{x})\} > g(x^*)$, which contradicts that $g(\bar{x}) + L^* \sum_{i=1}^{m} \max\{0, h_i(\bar{x})\} \leq g(x^*)$. Thus, $\bar{x} \in \mathcal{K}$ and $\bar{x}$ is an optimal solution of Problem (19).

Then, Problem (19) is equivalent to Problem (20), if L^* is selected as $L^* \geq \max_i \{u_i^*\}$.

(b) We next show that $\max_i \{u_i^*\} \leq L_0\mathcal{B}/\mathcal{C}$.

Since the Slater's condition holds for Problem (19) with the margin $\mathcal{C} > 0$, there exists $\tilde{x}$ such that $h_i(\tilde{x}) \leq -\mathcal{C}$ for all $i \in \{1, \ldots, m\}$. From (21), we have

$$g(\tilde{x}) + \sum_{i=1}^{m} u_i^* h_i(\tilde{x}) \geq g(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) = g(x^*).$$

Then,

$$-\mathcal{C}\sum_{i=1}^{m} u_i^* \geq \sum_{i=1}^{m} u_i^* h_i(\tilde{x}) \geq g(x^*) - g(\tilde{x}),$$

and then

$$\mathcal{C}\sum_{i=1}^{m} u_i^* \leq g(\tilde{x}) - g(x^*).$$

Then,

$$\max_i \{u_i^*\} \leq \sum_{i=1}^{m} u_i^* \leq \frac{g(\tilde{x}) - g(x^*)}{\mathcal{C}} \leq \frac{L_0\mathcal{B}}{\mathcal{C}}.$$

From (a)(b), the proof is finished. □

Lemma 9. *Suppose that Assumptions 3 and 4 are satisfied. Then,*

$$\|\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \theta_h(\mathcal{D}_0, \mathcal{D}_+)\| \leq \frac{1}{\lambda} \sup_{\theta \in \mathcal{K}^*} \|\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\theta)\|.$$

where λ is selected as $\lambda > 0$ for Problems (16) *and* (18)*.*

Proof. Consider the following optimization problem, which is parameterized by ϵ:

$$\begin{aligned} \bar{\theta}(\epsilon) = &\min_{\theta} (1-\epsilon)\mathcal{L}_{\mathcal{D}_0}(\theta) + \epsilon\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) + \frac{\lambda}{2}\|\theta - h\|^2 \\ &\text{s.t. } \mathbb{E}_{z\sim\mathcal{D}_i}[\ell_i(\theta, z)] \leq c_i,\ i = 1, \ldots, m. \end{aligned} \tag{22}$$

We denote the optimal solution of Problem (22) as $\bar{\theta}(\epsilon)$ and consider that $\epsilon \in [0, 1]$. We can see that, Problem (22) is reduced to Problem (16) when $\epsilon = 0$, and Problem (22) is reduced to Problem (18) when $\epsilon = 1$. Then, $\theta_h(\mathcal{D}_0, \mathcal{D}_+)) = \bar{\theta}(0)$ and $\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) = \bar{\theta}(1)$.

From part (iii) of Assumption 3, The LICQ holds for the optimization problem Problem (22), since Problem (15) and Problem (22) hold the same constraints.

From part (v) of Assumption 4, $\ell_0(\theta, z)$ is convex. Then, $\mathcal{L}_{\mathcal{D}_0}(\theta)$ is convex and $\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)$ is convex. Then,

$$(1-\epsilon)\mathcal{L}_{\mathcal{D}_0}(\theta) + \epsilon\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)$$

is convex. Since $\frac{\lambda}{2}\|\theta - h\|^2$ is λ-strongly convex, then

$$(1-\epsilon)\mathcal{L}_{\mathcal{D}_0}(\theta) + \epsilon\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) + \frac{\lambda}{2}\|\theta - h\|^2$$

is λ-strongly convex.

From part (v) of Assumption 4, $\ell_i(\cdot, z)$ is convex, then $\frac{1}{|\mathcal{D}_i^{tr}|}\sum_{z\in\mathcal{D}_i^{tr}} \ell_i(\theta, z) - c_i$ is convex for $i \leq i \leq m$.

From Lemma 1 and above conditions, the optimal solution $\bar{\theta}(\epsilon)$ of Problem (22) exists and is unique. The KKT conditions hold at $\bar{\theta}(\epsilon)$ with unique Lagrangian multipliers $\mu(\epsilon)$. The vector function $z(\epsilon) \triangleq [\bar{\theta}(\epsilon)^\top, \mu(\epsilon)^\top]^\top$ is continuous and locally Lipschitz at any $\epsilon \in [0, 1]$. Since the function $z(\epsilon) \triangleq [\bar{\theta}(\epsilon)^\top, \mu(\epsilon)^\top]^\top$ is locally Lipschitz on a compact set $[0, 1]$, then $z(\epsilon) \triangleq [\bar{\theta}(\epsilon)^\top, \mu(\epsilon)^\top]^\top$ is Lipschitz on $[0, 1]$.

From Lemma 1, The directional derivative of $\bar{\theta}(\epsilon)$ on any direction exists. For $\epsilon \in [0, 1] \subset \mathbb{R}$, the left derivative and the right derivative of $\bar{\theta}(\epsilon)$ always exists. We denote the left derivative as $\nabla_-\bar{\theta}(\epsilon)$, and the right derivative $\nabla_+\bar{\theta}(\epsilon)$. We have

$$\begin{bmatrix} \nabla_+\bar{\theta}(\epsilon) \\ \nabla_+\mu_{I(\epsilon,+)}(\epsilon) \end{bmatrix} = -\begin{bmatrix} \nabla_\theta^2\mathcal{L}(\bar{\theta}(\epsilon), \mu(\epsilon), \epsilon) & \nabla_\theta q_{I(\epsilon,d)}(\bar{\theta}(\epsilon), \epsilon)^\top \\ \nabla_\theta q_{I(\epsilon,+)}(\bar{\theta}(\epsilon), \epsilon) & 0 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_{\theta\epsilon}^2\mathcal{L}(\bar{\theta}(\epsilon), \mu(\epsilon), \epsilon) \\ \nabla_\epsilon q_{I(\epsilon,+)}(\bar{\theta}(\epsilon), \epsilon) \end{bmatrix} \tag{23}$$

and

$$\begin{bmatrix} \nabla_- \bar{\theta}(\epsilon) \\ \nabla_- \mu_{I(\epsilon,-)}(\epsilon) \end{bmatrix} = - \begin{bmatrix} \nabla^2_\theta \mathcal{L}(\bar{\theta}(\epsilon), \mu(\epsilon), \epsilon) & \nabla_\theta q_{I(\epsilon,d)}(\bar{\theta}(\epsilon), \epsilon)^\top \\ \nabla_\theta q_{I(\epsilon,-)}(\bar{\theta}(\epsilon), \epsilon) & 0 \end{bmatrix}^{-1} \begin{bmatrix} \nabla^2_{\theta\epsilon} \mathcal{L}(\bar{\theta}(\epsilon), \mu(\epsilon), \epsilon) \\ \nabla_\epsilon q_{I(\epsilon,-)}(\bar{\theta}(\epsilon), \epsilon) \end{bmatrix}. \tag{24}$$

Here,

$$q_i(\theta, \epsilon) = \frac{1}{|\mathcal{D}_i^{tr}|} \sum_{z \in \mathcal{D}_i^{tr}} \ell_i(\theta, z) - c_i,$$

and $q_{I(\epsilon,+)}$ is a vector function that contain all q_i with $i \in I(\epsilon,+) \subseteq I$; $\mu_{I(\epsilon,+)}$ $\mu_{I(\epsilon,+)}$ is a vector function that contain all μ_i with $i \in I(\epsilon,+) \subseteq I$; $q_{I(\epsilon,-)}$ is a vector function that contain all q_i with $i \in I(\epsilon,-) \subseteq I$; $\mu_{I(\epsilon,-)}$ is a vector function that contain all μ_i with $i \in I(\epsilon,+) \subseteq I$. Denote

$$g(\theta, \epsilon) = (1-\epsilon)\mathcal{L}_{\mathcal{D}_0}(\theta) + \epsilon \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) + \frac{\lambda}{2}\|\theta - h\|^2.$$

Then,

$$\mathcal{L}(\theta, \mu, \epsilon) = g(\theta, \epsilon) + \sum\nolimits_{i=1}^{m} \mu_i q_i(\theta, \epsilon).$$

We can compute

$$\nabla_\epsilon q_{I(\epsilon,+)} = 0$$

and

$$\nabla^2_{\theta\epsilon} \mathcal{L}(\theta, \mu, \epsilon) = \nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\theta)$$

By the computation of (23), we have

$$\begin{aligned} \nabla_+ \bar{\theta}(\epsilon) = &\Big(\nabla^2_\theta \mathcal{L}^{-1} - \nabla^2_\theta \mathcal{L}^{-1} \nabla q_{I(\epsilon,+)} (\nabla q_{I(\epsilon,+)}^T \nabla^2_\theta \mathcal{L}^{-1} \nabla q_{I(\epsilon,+)})^{-1} \nabla q_{I(\epsilon,+)}^T \nabla^2_\theta \mathcal{L}^{-1}\Big) \\ &\Big(\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\bar{\theta}(\epsilon)) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\bar{\theta}(\epsilon))\Big). \end{aligned} \tag{25}$$

Since $\nabla^2_\theta \mathcal{L}^{-1}$ is symmetric, then it can be represented as $\nabla^2_\theta \mathcal{L}^{-1} = M^T M$, then

$$\begin{aligned} &\|\nabla^2_\theta \mathcal{L}^{-1} - \nabla^2_\theta \mathcal{L}^{-1} \nabla q_{I(\epsilon,+)} (\nabla q_{I(\epsilon,+)}^T \nabla^2_\theta \mathcal{L}^{-1} \nabla q_{I(\epsilon,+)})^{-1} \nabla q_{I(\epsilon,+)}^T \nabla^2_\theta \mathcal{L}^{-1}\| \\ =&\|M^T (I - M \nabla q_{I(\epsilon,+)} (\nabla q_{I(\epsilon,+)}^T M^T M \nabla q_{I(\epsilon,+)})^{-1} \nabla q_{I(\epsilon,+)}^T M^T\| \\ =&\|M^T (I - N(N^T N)^{-1} N^T) M\| \\ \leq&\|M^T\| \|I - N(N^T N)^{-1} N^T\| \|M\|, \end{aligned} \tag{26}$$

where $N = M \nabla q_{I(\epsilon,+)}$.

Now, we have $\|M\| = \|M^T\| = \sqrt{\lambda_{max}(M^T M)} = \sqrt{\lambda_{max}(\nabla^2_\theta \mathcal{L}^{-1})} = \sqrt{\|\nabla^2_\theta \mathcal{L}^{-1}\|}$, where λ_{max} denotes the largest eigenvalue of the matrix. Then $\|M\|\|M^T\| = \|\nabla^2_\theta \mathcal{L}^{-1}\|$. We have $\mathcal{L}(\theta, \mu, \epsilon) = g(\theta, \epsilon) + \sum_{i=1}^m \mu_i q_i(\theta, \epsilon)$ where $q_i(\cdot, \epsilon)$ is convex for any ϵ and $g(\cdot, \epsilon)$ is λ-strongly convex, and $\mu_i \geq 0$, then $\mathcal{L}$ is λ-strongly convex. Therefore, $\|\nabla^2_\theta \mathcal{L}\| \geq \lambda$ and $\lambda_{min}(\nabla^2_\theta \mathcal{L}) \geq \lambda$, where λ_{min} denotes the smallest eigenvalue of the matrix. Then,

$$\|M\|\|M^T\| = \|\nabla^2_\theta \mathcal{L}^{-1}\| = \lambda_{max}(\nabla^2_\theta \mathcal{L}^{-1}) = \frac{1}{\lambda_{min}(\nabla^2_\theta \mathcal{L})} \leq \frac{1}{\lambda}. \tag{27}$$

Also, since all eigenvalues of $N(N^T N)^{-1} N^T$ are 0 and 1, then

$$\|I - N(N^T N)^{-1} N^T\| = \lambda_{max}(I - N(N^T N)^{-1} N^T) = 1. \tag{28}$$

From (25)(26)(27)(28),

$$\begin{aligned} \|\nabla_+ \bar{\theta}(\epsilon)\| \leq &\|\nabla^2_\theta \mathcal{L}^{-1} - \nabla^2_\theta \mathcal{L}^{-1} \nabla q_{I(\epsilon,+)} (\nabla q_{I(\epsilon,+)}^T \nabla^2_\theta \mathcal{L}^{-1} \nabla q_{I(\epsilon,+)})^{-1} \nabla q_{I(\epsilon,+)}^T \nabla^2_\theta \mathcal{L}^{-1}\| \\ &\|\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\bar{\theta}(\epsilon)) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\bar{\theta}(\epsilon))\| \\ \leq &\frac{1}{\lambda} \|\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\bar{\theta}(\epsilon)) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\bar{\theta}(\epsilon))\|. \end{aligned}$$

Similarly, we also have

$$\|\nabla_- \bar{\theta}(\epsilon)\| \leq \frac{1}{\lambda}\|\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\bar{\theta}(\epsilon)) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\bar{\theta}(\epsilon))\|.$$

Now, we know $\bar{\theta}(\epsilon)$ is Lipschitz on $[0,1]$, and

$$\|\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+^{tr})) - \theta_h(\mathcal{D}_0, \mathcal{D}_+^{tr}))\| = \|\bar{\theta}(1) - \bar{\theta}(0)\| \leq \sup_{\epsilon\in[0,1]} \|\bar{\theta}(\epsilon) - \bar{\theta}(0)\|.$$

We denote that $\epsilon^* = \arg\max_{\epsilon\in[0,1]} \|\bar{\theta}(\epsilon) - \bar{\theta}(0)\|$. From Lemma 7,

$$\begin{aligned}\|\bar{\theta}(\epsilon^*) - \bar{\theta}(0)\| &\leq \sup_{\epsilon\in[0,\epsilon^*]} \{\max\left(\|\nabla_-\bar{\theta}(\epsilon)\|, \|\nabla_+\bar{\theta}(\epsilon)\|\right)\}\epsilon^* \\ &\leq \sup_{\epsilon\in[0,\epsilon^*]} \{\max\left(\|\nabla_-\bar{\theta}(\epsilon)\|, \|\nabla_+\bar{\theta}(\epsilon)\|\right)\}.\end{aligned}$$

We donote that $\hat{\epsilon} = \arg\max_{\epsilon\in[0,\epsilon^*]}\{\max\left(\|\nabla_-\bar{\theta}(\epsilon)\|, \|\nabla_+\bar{\theta}(\epsilon)\|\right)\}$. Then, we have

$$\begin{aligned}&\|\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+^{tr})) - \theta_h(\mathcal{D}_0, \mathcal{D}_+^{tr}))\| = \|\bar{\theta}(1) - \bar{\theta}(0)\| \leq \ \|\bar{\theta}(\epsilon^*) - \bar{\theta}(0)\| \\ &\leq \frac{1}{\lambda}\|\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\bar{\theta}(\hat{\epsilon})) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\bar{\theta}(\hat{\epsilon}))\| \\ &\leq \frac{1}{\lambda}\sup_{\theta\in\mathcal{K}^*} \|\nabla_\theta \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta \mathcal{L}_{\mathcal{D}_0}(\theta)\|.\end{aligned} \tag{29}$$

□

Lemma 10. *Suppose that Assumptions 3 and 4 are satisfied. Then,*

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\left|\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+^{tr})) - \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+))\right|\right] \leq$$
$$2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0 + \lambda(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}},$$

where the constants L_c, $\mathcal{B}$, M, L_0, $\mathcal{C}$ are referred to Assumptions 3 and 4.

Proof. According to the part (i) of Assumption 3, the feasible sets $\mathcal{K}$ and $\mathcal{K}^{tr}$ are bounded in a compact set Θ with diameter $\mathcal{B}$. Then for any $\theta \in \Theta$, $\|\theta - \theta^*\| \leq \mathcal{B}$, since $\theta^* \in \Theta$. The Lipschitz constant of $\mathcal{L}_{\mathcal{D}_0^{tr},h}$ on Θ is

$$\begin{aligned}\sup_{\theta\in\Theta}\|\nabla\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta)\| &= \sup_{\theta\in\Theta}\|\frac{1}{|\mathcal{D}_0^{tr}|}\sum_{z\in\mathcal{D}_0^{tr}}\nabla_1\ell_0(\theta, z) + \lambda(\theta - h)\| \\ &\leq \sup_{\theta\in\Theta}\|\frac{1}{|\mathcal{D}_0^{tr}|}\sum_{z\in\mathcal{D}_0^{tr}}\nabla_1\ell_0(\theta, z)\| + \lambda\sup_{\theta\in\Theta}\|\theta - h\|.\end{aligned}$$

From part (ii) of Assumption 4, $\ell_0(\cdot, z)$ is L_0-Lipschitz continuous, then

$$\sup_{\theta\in\Theta}\|\frac{1}{|\mathcal{D}_0^{tr}|}\sum_{z\in\mathcal{D}_0^{tr}}\nabla_1\ell_0(\theta, z)\| \leq L_0.$$

Also, we have

$$\sup_{\theta\in\Theta}\|\theta - h\| \leq \|\theta^* - h\| + \sup_{\theta\in\Theta}\|\theta - \theta^*\| \leq \|\theta^* - h\| + \mathcal{B}.$$

Thus, the Lipschitz constant of $\mathcal{L}_{\mathcal{D}_0^{tr},h}$ on Θ is $L_0 + \lambda(\|\theta^* - h\| + \mathcal{B})$.

From the parts (i)(ii) of Assumption 3 and Lemma 8, the optimization problem in (18) is

$$\min_\theta \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) \quad \text{s.t.} \quad \mathcal{C}_{\mathcal{D}_i^{tr}}(\theta) - c_i \leq 0,$$

and equivalent to:

$$\min_\theta \mathcal{L}_1(\theta) = \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) + \frac{L_0\mathcal{B} + \lambda\mathcal{B}(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}}\sum_{i=1}^m \max\{\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta) - c_i, 0\}.$$

The optimization problem in (17) is

$$\min_{\theta} \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) \quad \text{s.t.} \quad \mathcal{C}_{\mathcal{D}_i}(\theta) - c_i \leq 0,$$

and equivalent to:

$$\min_{\theta} \mathcal{L}_2(\theta) = \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) + \frac{L_0\mathcal{B} + \lambda\mathcal{B}(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}} \sum_{i=1}^{m} \max\{\mathcal{C}_{\mathcal{D}_i}(\theta) - c_i, 0\}.$$

According to Lemma 2 and Remark 2,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[|\mathcal{L}_2(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)) - \mathcal{L}_1(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))|\right] \\
&\leq 2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0 + \lambda(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}.
\end{aligned}$$

Since $\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)$ is the solution of (17), then

$$\frac{L_0\mathcal{B} + \lambda\mathcal{B}(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}} \sum_{i=1}^{m} \max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)) - c_i, 0\} = 0,$$

then

$$\mathcal{L}_2(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)) = \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)).$$

Similarly,

$$\mathcal{L}_1(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})) = \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\left|\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))\right|\right] \\
&\leq 2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0 + \lambda(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}.
\end{aligned}$$

□

Lemma 11. *Suppose that Assumptions 3 and 4 are satisfied. Then,*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}) - \theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|] \\
&\leq 4\mathcal{B}\sqrt{m\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_0 + \lambda(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}},
\end{aligned}$$

where the constants L_c, $\mathcal{B}$, M, L_0, $\mathcal{C}$ are referred to Assumptions 3 and 4.

Proof. From the proof of Lemma 10, Problem (17) is equivalent to $\min_\theta \mathcal{L}_2(\theta)$ and Problem (18) is equivalent to $\min_\theta \mathcal{L}_1(\theta)$. According to Lemma 2 and Remark 2,

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\Delta\left(\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}), \Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)\right)\right] \leq 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}},$$

for any $\epsilon > 0$, and L_{max} is selected as

$$L_{max} = \frac{\mathcal{B}L_0 + \lambda\mathcal{B}(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}},$$

where the sets

$$\begin{aligned}
\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+) &= \left\{\theta \in \Theta : \mathcal{L}_1(\theta) \leq \mathcal{L}_1(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})) + \epsilon\right\}, \\
\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}) &= \left\{\theta \in \Theta : \mathcal{L}_2(\theta) \leq \mathcal{L}_2(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)) + \epsilon\right\}.
\end{aligned}$$

and

$$\mathcal{L}_1(\theta) = \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) + L_{max}\sum_{i=1}^{m} \max\{\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta) - c_i, 0\},$$

$$\mathcal{L}_2(\theta) = \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta) + L_{max}\sum_{i=1}^{m}\max\{\mathcal{C}_{\mathcal{D}_i}(\theta) - c_i, 0\},$$

and Δ is defined in Section D.2.

Then we have

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\Delta\left(\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}),\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)\right)\right] \le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}},$$

i.e.,

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\sup_{\theta_1\in\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})}\inf_{\theta\in\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)}\|\theta_1-\theta\|\right] \le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Then,

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\inf_{\theta\in\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)}\|\theta_1-\theta\|\right] \le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}}$$

for any selected $\theta_1 \in \Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})$. Then,

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\inf_{\theta\in\Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)}\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta\|\right] \le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}}$$

Then, there exists $\theta \in \Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)$, such that

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta\|] \le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}}$$

Then, for any sampled $\mathcal{D}_+^{tr}$, there exists $\theta \in \Theta_h^\epsilon(\mathcal{D}_0^{tr},\mathcal{D}_+)$, such that

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta\|] \le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Then,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|] &\le \mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta\| + \epsilon] \\ &\le 4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\epsilon\sqrt{|\mathcal{D}_+^{tr}|}} + \epsilon.\end{aligned}$$

Select that $\epsilon = \sqrt{4m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\sqrt{|\mathcal{D}_+^{tr}|}}}$, we have

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|] \le 4\sqrt{m\mathcal{B}\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_{max}}{\sqrt{|\mathcal{D}_+^{tr}|}}}.$$

Substitute $L_{max} = \frac{\mathcal{B}L_0+\lambda\mathcal{B}(\|\theta^*-h\|+\mathcal{B})}{\mathcal{C}}$ into the inequality, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|] \\ &\le 4\mathcal{B}\sqrt{m\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})\frac{L_0+\lambda(\|\theta^*-h\|+\mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}}.\end{aligned}$$

□

Propositions 3 and 4 state the generalization with respect to the objective risk and the generalization with respect to the constraint violations, respectively.

Proposition 3. *Suppose that Assumptions 3 and 4 are satisfied. Pick that $\lambda > 0$ for Problems* (15)-(18)*. Then,*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta^*),0\right\}\right] &\le \frac{\lambda}{2}\|\theta^*-h\|^2\\ &+\frac{\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})(4\rho\mathcal{B}\sqrt{d}+4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|}+6)}{\lambda|\mathcal{D}_0^{tr}|}\\ &+2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0+\lambda(\|\theta^*-h\|+\mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}\\ &+\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}}\cdot\frac{(\ln|\mathcal{D}_+^{tr}|)^{\frac{1}{4}}}{|\mathcal{D}_+^{tr}|^{\frac{1}{4}}}\right),\end{aligned}$$

where the constants ρ, L_c, $\mathcal{B}$, M, L_0, $\mathcal{C}$ are referred to Assumptions 3 and 4.

Proof. For $\mathcal{D}^{tr}\sim\mathcal{D}$, consider following decomposition:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta^*),0\right\}\right]\\ \le\ &\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}\right]\\ &+\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*),0\right\}\right]\end{aligned}$$

Since for any $\mathcal{D}_0^{tr}$, we have $\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*)\ge 0$, then

$$\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*),0\right\}\right]=\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*)\right].$$

Then,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta^*),0\right\}\right]\\ \le\ &\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}\right]\\ &+\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right]\\ &+\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*)\right]\\ \le\ &\underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}-\frac{\lambda}{2}\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)-\theta^*\|^2\right]}_{B}\\ &+\underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right]}_{A}+\underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*)\right]}_{C}.\end{aligned}$$

(i) Consider C, we have

$$\begin{aligned}C&=\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*)\right]\\ &=\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta^*)\right]\right]\\ &=\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0,h}(\theta^*)+\frac{\lambda}{2}\|\theta^*-h\|^2\right]\right]\\ &=\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0,h}(\theta^*)\right]\right]+\frac{\lambda}{2}\|\theta^*-h\|^2\\ &=\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right]-\mathcal{L}_{\mathcal{D}_0,h}(\theta^*)\right]+\frac{\lambda}{2}\|\theta^*-h\|^2\\ &=\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta^*)\right]\right]+\frac{\lambda}{2}\|\theta^*-h\|^2.\end{aligned}$$

The last equality comes from

$$\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta^*)\right]=\mathcal{L}_{\mathcal{D}_0,h}(\theta^*).$$

By the definition of $\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)$ in (17), we have

$$\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta^*) \leq 0.$$

Then,

$$C \leq \frac{\lambda}{2}\|\theta^* - h\|^2.$$

(ii) Consider A, we have

$$\begin{aligned}
A &= \mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+))\right] \\
&\leq \underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0, \mathcal{D}_+)) + \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+))\right]}_{A_1} \\
&\quad + \underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right]}_{A_2}.
\end{aligned}$$

For A_2, we have

$$\begin{aligned}
A_2 &= \mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right] \\
&= \mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0, \mathcal{D}_+)) - \mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right]\right] \\
&= \mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0, \mathcal{D}_+)) - \mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right]\right] \\
&= \mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}[0] = 0.
\end{aligned}$$

For A_1, we have

$$A_1 = \mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[(\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - (\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right]\right].$$

From Lemma 9,

$$\|\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \theta_h(\mathcal{D}_0, \mathcal{D}_+)\| \leq \frac{1}{\lambda}\sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|. \tag{30}$$

From the mean value theorem,

$$\begin{aligned}
&\left|(\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - (\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right| \\
&\leq \|\nabla\mathcal{L}_{\mathcal{D}_0}(\theta) - \nabla\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)\|\|\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \theta_h(\mathcal{D}_0, \mathcal{D}_+^{tr})\|,
\end{aligned}$$

where θ is in the middle of $\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)$ and $\theta_h(\mathcal{D}_0, \mathcal{D}_+)$, i.e., $\theta = (1-\alpha)\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+) + \alpha\theta_h(\mathcal{D}_0, \mathcal{D}_+)$, with $\alpha \in [0, 1]$. Then, we have

$$\|\nabla\mathcal{L}_{\mathcal{D}_0}(\theta) - \nabla\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)\| \leq \sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|.$$

Then,

$$\begin{aligned}
&\left|(\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - (\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right| \\
&\leq \sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|\|\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - \theta_h(\mathcal{D}_0, \mathcal{D}_+^{tr})\|,
\end{aligned}$$

From (30), we can get

$$\begin{aligned}
&\left|(\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+)) - (\mathcal{L}_{\mathcal{D}_0} - \mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0, \mathcal{D}_+))\right| \\
&\leq \sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|\frac{1}{\lambda}\sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\| \\
&\leq \frac{1}{\lambda}\sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta) - \nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|^2.
\end{aligned}$$

From Lemma 5, since the Lipschitz constant of $\nabla \ell_0(\cdot, z)$ is ρ and the upper bounded of $\nabla \ell_0(\cdot, z)$ is L_0, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\left|(\mathcal{L}_{\mathcal{D}_0}-\mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-(\mathcal{L}_{\mathcal{D}_0}-\mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0,\mathcal{D}_+))\right|\right]\\
\le&\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\frac{1}{\lambda}\sup_{\theta\in\mathcal{K}^*}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)-\nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|^2\right]\\
\le&\frac{\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})(4\rho\mathcal{B}\sqrt{d}+4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|}+6)}{\lambda|\mathcal{D}_0^{tr}|}.
\end{aligned}$$

Then,

$$\begin{aligned}
A\le A_1=&\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\left|(\mathcal{L}_{\mathcal{D}_0}-\mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-(\mathcal{L}_{\mathcal{D}_0}-\mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0,\mathcal{D}_+))\right|\right]\right]\\
\le&\frac{\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})(4\rho\mathcal{B}\sqrt{d}+4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|}+6)}{\lambda|\mathcal{D}_0^{tr}|}.
\end{aligned}$$

(iii) Consider B, we have

$$\begin{aligned}
B=&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}-\frac{\lambda}{2}\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)-\theta^*\|^2\right]\\
\le&\underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}-\frac{\lambda}{2}\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)-\theta^*\|^2\right]}_{B_1}\\
+&\underbrace{\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\left|\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))+\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right|\right]}_{B_2}.
\end{aligned}$$

Then,

$$\begin{aligned}
B_1=&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}-\frac{\lambda}{2}\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)-\theta^*\|^2\right]\\
\le&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)),0\right\}\right]\\
\le&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\left|\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right|\right]\\
=&\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\left|\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr},h}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right|\right]\right].
\end{aligned}$$

From Lemma 10, we have

$$B_1\le 2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0+\lambda(\|\theta^*-h\|+\mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}.$$

For B_2, we have

$$\begin{aligned}
B_2=&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\left|\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))+\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))-\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right|\right]\\
=&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\left|(\mathcal{L}_{\mathcal{D}_0}-\mathcal{L}_{\mathcal{D}_0^{tr}})(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+))\right|\right]\\
\le&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\sup_{\theta\in\mathcal{K}^*\cup\mathcal{K}^{tr}}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)-\nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|\right].
\end{aligned}$$

Then,

$$B_2^2\le\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\sup_{\theta\in\mathcal{K}^*\cup\mathcal{K}^{tr}}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)-\nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|^2\right]\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|^2\right].$$

From the computation of A_1, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\sup_{\theta\in\mathcal{K}^*\cup\mathcal{K}^{tr}}\|\nabla_\theta\mathcal{L}_{\mathcal{D}_0^{tr}}(\theta)-\nabla_\theta\mathcal{L}_{\mathcal{D}_0}(\theta)\|^2\right]\\
\le&\frac{\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})(4\rho\mathcal{B}\sqrt{d}+4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|}+6)}{|\mathcal{D}_0^{tr}|}
\end{aligned}$$

From Lemma 11,

$$\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\|\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})-\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+)\|^2\right]\leq\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right).$$

Then,

$$B_2\leq\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}}\cdot\frac{(\ln|\mathcal{D}_+^{tr}|)^{\frac{1}{4}}}{|\mathcal{D}_+^{tr}|^{\frac{1}{4}}}\right)\ll B_1.$$

From (i)(ii)(iii), we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{L}_{\mathcal{D}_0}(\theta^*),0\right\}\right]&\leq\frac{\lambda}{2}\|\theta^*-h\|^2\\&+\frac{\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})(4\rho\mathcal{B}\sqrt{d}+4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|}+6)}{\lambda|\mathcal{D}_0^{tr}|}\\&+2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0+\lambda(\|\theta^*-h\|+\mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}\\&+\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}}\cdot\frac{(\ln|\mathcal{D}_+^{tr}|)^{\frac{1}{4}}}{|\mathcal{D}_+^{tr}|^{\frac{1}{4}}}\right).\end{aligned}$$

□

Proposition 4. *Suppose that Assumptions 3 and 4 are satisfied.*

$$\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-c_i,0\}\right]\leq\frac{2\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}$$

for each $1\leq i\leq m$, where the constants L_c, $\mathcal{B}$, M, L_0, $\mathcal{C}$ are referred to Assumptions 3 and 4.

Proof. From Lemmas 3 and 4,

$$\mathbb{E}_{\mathcal{D}_i^{tr}\sim\mathcal{D}_i}\left[\sup_{\theta\in\mathcal{K}^{tr}\cup\mathcal{K}^*}|\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta)-\mathcal{C}_{\mathcal{D}_i}(\theta)|\right]\leq\frac{2\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Then,

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[|\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))|\right]\leq\frac{2\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Since

$$\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-c_i\leq 0,$$

and

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr})),0\}\right]\\&\leq\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[|\mathcal{C}_{\mathcal{D}_i^{tr}}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))|\right]\end{aligned}$$

we have

$$\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-c_i,0\}\right]\leq\frac{2\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Thus,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-c_i,0\}\right]\\&=\mathbb{E}_{\mathcal{D}_0^{tr}\sim\mathcal{D}_0}\left[\mathbb{E}_{\mathcal{D}_+^{tr}\sim\mathcal{D}_+}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr},\mathcal{D}_+^{tr}))-c_i,0\}\right]\right]\leq\frac{2\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}.\end{aligned}$$

□

Proposition 1 can be proven by Propositions 3 and 4. Proposition 2 can be proven by the following proposition.

Proposition 5. *Suppose that Assumptions 1 and 2 are satisfied. For a sequence $\phi_{1:T} = \{\phi_1, \cdots, \phi_T\}$, pick the regularization parameter*

$$\lambda = \frac{2\sqrt{d}(\rho\mathcal{B} + L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})}{\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T})\sqrt{|\mathcal{D}_0^{tr}|}},$$

where

$$\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T})^2 \triangleq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|\theta_t^* - \phi_t\|^2.$$

Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\left\{\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda, \phi_t, \mathcal{D}_t^{tr})) - \mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*), 0\right\}\right] \leq \mathcal{O}\left(\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T})\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}} + \sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right).$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\{\mathcal{C}_{\mathcal{D}_{i,t}}(\mathcal{A}lg(\lambda, \phi_t, \mathcal{D}_t^{tr})) - c_i, 0\}\right] \leq \mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right), \forall\, i = 1, \ldots, m.$$

Proof. From Propositions 3 and 4, for problems (15)-(18), we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\left\{\mathcal{L}_{\mathcal{D}_0}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+^{tr})) - \mathcal{L}_{\mathcal{D}_0}(\theta^*), 0\right\}\right] &\leq \frac{\lambda}{2}\|\theta^* - h\|^2 \\ &+ \frac{\sqrt{d}(\rho\mathcal{B} + L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})(4\rho\mathcal{B}\sqrt{d} + 4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|} + 6)}{\lambda|\mathcal{D}_0^{tr}|} \\ &+ 2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0 + \lambda(\|\theta^* - h\| + \mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}.\end{aligned}$$

and

$$\mathbb{E}_{\mathcal{D}^{tr}\sim\mathcal{D}}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\theta_h(\mathcal{D}_0^{tr}, \mathcal{D}_+^{tr})) - c_i, 0\}\right] \leq \frac{2\sqrt{d}(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Consider the task sequence $\{\mathcal{T}_1, \cdots, \mathcal{T}_T\}$ and

$$\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T})^2 \triangleq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|\theta_t^* - \phi_t\|^2,$$

for the optimization problems in (1) and (3), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\left\{\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda, \phi_t, \mathcal{D}_t^{tr})) - \mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*), 0\right\}\right] &\leq \lambda\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T})^2 \\ &+ \frac{\sqrt{d}(\rho\mathcal{B} + L_0\sqrt{\ln|\mathcal{D}_{0,t}^{tr}|})(4\rho\mathcal{B}\sqrt{d} + 4L_0\sqrt{d\ln|\mathcal{D}_0^{tr}|} + 6)}{\lambda|\mathcal{D}_0^{tr}|} \\ &+ 2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B} + M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0 + \lambda(\sqrt{2}\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T}) + \mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}.\end{aligned}$$

Consider $\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T}) > 0$, let

$$\lambda = \frac{2\sqrt{d}(\rho\mathcal{B} + L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})}{\mathcal{D}ist(\phi_{1:T}, \mathcal{T}_{1:T})\sqrt{|\mathcal{D}_0^{tr}|}},$$

Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*)\right]$$
$$\leq \frac{4\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})\mathcal{D}ist(\phi_{1:T},\mathcal{T}_{1:T})+3\mathcal{D}ist(\phi_{1:T},\mathcal{T}_{1:T})}{\sqrt{|\mathcal{D}_0^{tr}|}}$$
$$+2m\mathcal{B}\sqrt{d}\left(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|}\right)\frac{L_0+\lambda(\sqrt{2}\mathcal{D}ist(\phi_{1:T},\mathcal{T}_{1:T})+\mathcal{B})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}}.$$

We omit some constants with small quantities,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*)\right]\leq\frac{4\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})\mathcal{D}ist(\phi_{1:T},\mathcal{T}_{1:T})}{\sqrt{|\mathcal{D}_0^{tr}|}}$$
$$+\frac{2L_0\mathcal{B}m\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\mathcal{C}\sqrt{|\mathcal{D}_+^{tr}|}},$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\{\mathcal{C}_{\mathcal{D}_i}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-c_i,0\}\right]\leq\frac{2\sqrt{d}(L_c\mathcal{B}+M\sqrt{\ln|\mathcal{D}_+^{tr}|})}{\sqrt{|\mathcal{D}_+^{tr}|}}.$$

Then, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*)\right]$$
$$\leq\mathcal{O}\left(\mathcal{D}ist(\phi_{1:T},\mathcal{T}_{1:T})\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}}+\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right).$$
$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\{\mathcal{C}_{\mathcal{D}_{i,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-c_i,0\}\right]\leq\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right),\forall\, i=1,\ldots,m.$$

□

By selecting $\phi_t=\phi$ for each t in Proposition 5, Proposition 2 is proven.

F Proof of Theorem 1

As we apply the FTPL algorithm to the meta-objective function in the online constrained meta-learning problem and formulate the problem (4), we require some properties of the meta-objective function. Note that $\mathcal{L}^{val}(\mathcal{A}lg(\lambda,\phi,\mathcal{D}_t^{tr}),\mathcal{D}_{0,t}^{val})$ in (4) is a function of the solution $\mathcal{A}lg(\lambda,\phi,\mathcal{D}_t^{tr})$ of the optimization problem (3) and its property is shown in Proposition 6.

Proposition 6. *Suppose that Assumptions 1 and 2 are satisfied. Then, $\mathcal{L}^{val}(\mathcal{A}lg(\lambda,\cdot,\mathcal{D}_t^{tr}),\mathcal{D}_{0,t}^{val})$ is L_0-Lipschitz.*

Proof. Since $\ell_0(\cdot,z)$ is L_0-Lipschitz, i.e., $\|\ell_0(w,z)-\ell_0(u,z)\|\leq L_0\|w-u\|$, we have
$\|\mathcal{L}^{val}(\mathcal{A}lg(\lambda,w,\mathcal{D}_t^{tr}),\mathcal{D}_{0,t}^{val})-\mathcal{L}^{val}(\mathcal{A}lg(\lambda,u,\mathcal{D}_t^{tr}),\mathcal{D}_{0,t}^{val})\|\leq L_0\|\mathcal{A}lg(\lambda,w,\mathcal{D}_t^{tr})-\mathcal{A}lg(\lambda,u,\mathcal{D}_t^{tr})\|$.
Then, to show Proposition 6, we need to show

$$\|\mathcal{A}lg(\lambda,w,\mathcal{D}_t^{tr})-\mathcal{A}lg(\lambda,u,\mathcal{D}_t^{tr})\|\leq\|w-u\|.$$

Consider the problem

$$\mathcal{A}lg(\lambda,w,\mathcal{D}_t^{tr})=\operatorname*{argmin}_{\theta\in\Theta}\frac{1}{|\mathcal{D}_{0,t}^{tr}|}\sum_{z\in\mathcal{D}_{0,t}^{tr}}\ell_0(\theta,z)+\frac{\lambda}{2}\|\theta-w\|^2$$
$$\text{s.t. }\frac{1}{|\mathcal{D}_{i,t}^{tr}|}\sum_{z\in\mathcal{D}_{i,t}^{tr}}\ell_i(\theta,z)\leq c_{i,t},\ i=1,\ldots,m,$$

then, for $0 \leq \alpha \leq 1$, we have

$$\mathcal{A}lg(\lambda, \alpha w + (1-\alpha)u, \mathcal{D}_t^{tr}) = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_{0,t}^{tr}|} \sum_{z \in \mathcal{D}_{0,t}^{tr}} \ell_0(\theta, z) + \frac{\lambda}{2}\|\alpha w + (1-\alpha)u - \theta\|^2$$

$$\text{s.t. } \frac{1}{|\mathcal{D}_{i,t}^{tr}|} \sum_{z \in \mathcal{D}_{i,t}^{tr}} \ell_i(\theta, z) \leq c_{i,t},\ i = 1, \ldots, m,$$

From part (iii) of Assumption 1, The LICQ holds for the optimization problem. From part (v) of Assumption 2, $\ell_0(\cdot, z)$ is convex, then $\frac{1}{|\mathcal{D}_{0,t}^{tr}|} \sum_{z \in \mathcal{D}_{0,t}^{tr}} \ell_0(\theta, z) + \frac{\lambda}{2}\|\alpha w + (1-\alpha)u - \theta\|^2$ is λ-strongly convex w.r.t. θ.

By Lemma 1 and above conditions, similar to the proof of Lemma 9, we have

$$\begin{aligned}
&\|\mathcal{A}lg(\lambda, w, \mathcal{D}_t^{tr}) - \mathcal{A}lg(\lambda, u, \mathcal{D}_t^{tr})\| \\
&\leq \frac{1}{\lambda} \left\| \nabla_{\alpha\theta} \frac{\lambda}{2} \|\alpha w + (1-\alpha)u - \theta\|^2 \right\| = \frac{1}{\lambda}\|\lambda w - \lambda u\| \leq \|w - u\|.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\mathcal{L}^{val}(\mathcal{A}lg(\lambda, w, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) - \mathcal{L}^{val}(\mathcal{A}lg(\lambda, u, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val})\| \\
&\quad \leq L_0 \|\mathcal{A}lg(\lambda, w, \mathcal{D}_t^{tr}) - \mathcal{A}lg(\lambda, u, \mathcal{D}_t^{tr})\| \leq L_0 \|w - u\|.
\end{aligned}$$

The proof is done. □

Proof of Theorem 1. From Proposition 6, $\mathcal{L}^{val}(\mathcal{A}lg(\lambda, \cdot, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val})$ is L_0-Lipschitz.

By Lemma 6, we have

$$\begin{aligned}
&\mathbb{E}_\sigma \left[\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi_t, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) - \frac{1}{T} \inf_{\phi} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) \right] \\
&\qquad \leq \mathcal{O}\left(\eta d^2 D_l L_0^2 + \frac{dD}{\eta T} \right).
\end{aligned}$$

Select that

$$\eta = \frac{1}{L_0 \sqrt{dT}},$$

we have

$$\begin{aligned}
&\mathbb{E}_\sigma \left[\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi_t, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) - \frac{1}{T} \inf_{\phi} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) \right] \\
&\qquad \leq \mathcal{O}\left(\frac{D_l d^{\frac{3}{2}} L_0}{\sqrt{T}} \right).
\end{aligned}$$

Moreover, we have

$$\frac{1}{T} \inf_{\phi} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) \leq \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi^*, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}),$$

where $\mathcal{S}^*(\mathcal{T}_{1:T}) \triangleq \min_\phi \mathcal{D}ist(\phi, \mathcal{T}_{1:T})$ is the task dissimilarity and $\phi^* \triangleq \arg\min_\phi \mathcal{D}ist(\phi, \mathcal{T}_{1:T})$. Then,

$$\begin{aligned}
&\mathbb{E}_\sigma \left[\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi_t, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) \right] \\
&\leq \mathbb{E}_\sigma \left[\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}^{val}(\mathcal{A}lg(\lambda, \phi^*, \mathcal{D}_t^{tr}), \mathcal{D}_{0,t}^{val}) \right] + \mathcal{O}\left(\frac{D_l d^{\frac{3}{2}} L_0}{\sqrt{T}} \right).
\end{aligned} \tag{31}$$

From Lemmas 3 and 4, we have

$$\mathbb{E}_{\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\sup_{\theta}|\mathcal{L}^{val}(\theta,\mathcal{D}_{0,t}^{val})-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta)|\right]\leq\frac{2\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}}.$$

Here, we represent $\mathcal{L}_{\mathcal{D}_{0,t}}(\theta)=\mathbb{E}_{z\sim\mathcal{D}_{0,t}}\left[\ell_0(\theta,z)\right]$. Then, we have

$$\mathbb{E}_{\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[|\mathcal{L}^{val}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}),\mathcal{D}_{0,t}^{val})-\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))|\right]\leq\frac{2\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}},$$

and

$$\mathbb{E}_{\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[|\mathcal{L}^{val}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}),\mathcal{D}_{0,t}^{val})-\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))|\right]\leq\frac{2\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}}.$$

Combine with (31), we have

$$\begin{aligned}\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))\right]&\leq\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))\right]\\&+\frac{4\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}}+\mathcal{O}\left(\frac{D_l d^{\frac{3}{2}}L_0}{\sqrt{T}}\right)\end{aligned}$$

Then,

$$\begin{aligned}&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*))\right]\\\leq&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*))\right]\\&+\frac{4\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}}+\mathcal{O}\left(\frac{D_l d^{\frac{3}{2}}L_0}{\sqrt{T}}\right),\end{aligned}$$

which implies that

$$\begin{aligned}&\mathbb{E}[\bar{R}_{0,[1:T]}]\\=&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_t,\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\frac{1}{T}\sum_{t=1}^{T}R_{0,t}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))\right]\\=&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[R_{0,t}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))\right]\right]\\=&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\left\{\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi_t,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*),0\right\}\right]\right]\\\leq&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[\max\left\{\mathcal{L}_{\mathcal{D}_{0,t}}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))-\mathcal{L}_{\mathcal{D}_{0,t}}(\theta_t^*),0\right\}\right]\right]\\&+\frac{4\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}}+\mathcal{O}\left(\frac{D_l d^{\frac{3}{2}}L_0}{\sqrt{T}}\right)\\=&\mathbb{E}_{\sigma,\mathcal{D}_{0,t}^{val}\sim\mathcal{D}_{0,t}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[R_{0,t}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))\right]\right]\\&+\frac{4\sqrt{d}L_0\mathcal{B}(1+\sqrt{\ln|\mathcal{D}_0^{val}|})}{\sqrt{|\mathcal{D}_0^{val}|}}+\mathcal{O}\left(\frac{D_l d^{\frac{3}{2}}L_0}{\sqrt{T}}\right).\end{aligned}$$

By Proposition (2), pick the regularization parameter

$$\lambda=\frac{2\sqrt{d}(\rho\mathcal{B}+L_0\sqrt{\ln|\mathcal{D}_0^{tr}|})}{\mathcal{S}^*(\mathcal{T}_{1:T})\sqrt{|\mathcal{D}_0^{tr}|}},$$

we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[R_{0,t}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))\right]\leq\mathcal{O}\left(\mathcal{S}^*(\mathcal{T}_{1:T})\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}}+\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right).$$

Thus, we have

$$\mathbb{E}[\bar{R}_{0,[1:T]}]\leq\mathcal{O}\left(\mathcal{S}^*(\mathcal{T}_{1:T})\sqrt{\frac{\ln|\mathcal{D}_0^{tr}|}{|\mathcal{D}_0^{tr}|}}+\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}+\sqrt{\frac{\ln|\mathcal{D}_0^{val}|}{|\mathcal{D}_0^{val}|}}+\frac{1}{\sqrt{T}}\right).$$

By Proposition (2), we also have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_t^{tr}\sim\mathcal{D}_t}\left[R_{i,t}(\mathcal{A}lg(\lambda,\phi^*,\mathcal{D}_t^{tr}))\right]\leq\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right),\ \forall i=1,\ldots,m.$$

Then,

$$\mathbb{E}[\bar{R}_{i,[1:T]}]\leq\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{D}_+^{tr}|}{|\mathcal{D}_+^{tr}|}}\right),\ \forall i=1,\ldots,m.$$

□

Proof of Corollary 1. We have

$$\mathcal{S}^*(\mathcal{T}_{1:T})\triangleq\min_{\phi}\ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|\theta_t^*-\phi\|^2,$$

and

$$\mathcal{S}^*(p(\mathcal{T}))^2\triangleq\min_{\phi}\mathbb{E}_{\mathcal{T}_t\sim p(\mathcal{T})}\left[\frac{1}{2}\|\theta_t^*-\phi\|^2\right].$$

Then,

$$\mathcal{S}^*(p(\mathcal{T}))^2\triangleq\min_{\phi}\mathbb{E}_{\mathcal{T}_1,\cdots,\mathcal{T}_T\sim p(\mathcal{T})}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|\theta_t^*-\phi\|^2\right].$$

It is easy to see that

$$\mathbb{E}_{\mathcal{T}_t\sim p(\mathcal{T})}[\mathcal{S}^*(\mathcal{T}_{1:T})^2]\leq\mathcal{S}^*(p(\mathcal{T}))^2.$$

Similar to the proofs of Theorem 1 and Proposition 5, Corollary 1 is proven. □