A. Appendix

A.1. Experiment Details

We cover details of experiments in this section. We use PyTorch framework (Paszke et al., 2019) to train and evaluate our models. MAML meta training is implemented with Higher (Grefenstette et al., 2019) library. We explain the datasets and the baselines used for evaluations. We implement FTML baseline since there is no official implementation from Finn et al. (2019).

Algorithm 1 stores a state as $\nabla f^{t-1} \circ U^{t-1}(\boldsymbol{\theta}^t)$. Updating this state requires one full pass on the training dataset. To avoid this, we utilize the first order condition stated in Eq. 10. Using the first order condition, we linearly update $\nabla f^{t-1} \circ U^{t-1}(\boldsymbol{\theta}^t)$ states as $\nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1}) = \nabla f^{t-1} \circ U^{t-1}(\boldsymbol{\theta}^t) - \alpha \left(\boldsymbol{\theta}^{t+1} - \boldsymbol{w}^t\right)$ in our implementation.

Performance metrics. We give details of CTM, LTM and Task Learning Efficiency Metrics.

Current Task Metric (CTM). We consider the performance with respect to the current revealed task. At each round, the meta model is adapted using the revealed limited supervised data of the current task and the performance is recorded on the test set. Since we are using the meta model trained on previous losses, this metric shows how models perform in a new task. CTM corresponds to $\frac{1}{T}\sum_{t=1}^{T} f^t \circ U^t(\boldsymbol{\theta}^t)$ performance where $\boldsymbol{\theta}^t$ is the meta model that is not trained on loss t.

Long-Term Task Metric (LTM). We consider the performance with respect to previous tasks. At each round, the meta model is adapted using the revealed limited supervised data of the each previous and the performance is recorded on the test set. Then we compute the mean performance of all previous tasks as the LTM. Since we are using the current meta model as a initialization of previous tasks, this metric measures the ability of catastrophic forgetting. LTM corresponds to $\frac{1}{T}\sum_{t=1}^{T} f^t \circ U^t\left(\boldsymbol{\theta}^{T+1}\right)$ performance.

Task Learning Efficiency Metric. At each round before updating the meta model, we record the number of data points required to achieve a sufficient performance on the current revealed task instance. For S-MNIST dataset, each task contains fixed number training data samples. We randomly sample a subset of training data of current task and adapt the meta model on this subset of data and record performance on the test data of current task. We use $[0, 10, 20, 30, 40, 50, 60]$ as the size of subset where 0 means we directly test current task on meta model. For CIFAR-100, we use $[0, 50, 100, 150, 200, 250]$ as the size of subset.

Task Learning Efficiency is recorded on top of the meta model trained on previous losses. In this metric, we allow meta models to be updated using the current task. It is a metric to measure the ability to adapt to a new task. Figure 5 shows comparison of smoothed task efficiency results of MOML, FTML and FS on both S-MNIST and 5-way CIFAR-100 where shadow corresponds to the standard deviation of multiple runs. In general, both meta learning methods need less number of data samples to reach a threshold accuracy with more rounds. This means after revealing more tasks, both methods improve Task Learning Efficiency. We note that FS is not a meta learning methods as such its task efficiency does not improve with rounds as expected.

MOML needs to see less data as new tasks arrive. For example, in S-MNIST, MOML requires maximum 45 datapoints to achieve 80% accuracy after 300 tasks whereas FTML achieves the same point with 600 tasks. Similarly in CIFAR-100 dataset, MOML only needs 180 datapoints to achieve 55% accuracy after 110 rounds which is better than FTML.

Hyperparameters. We use SGD optimizer with learning rate of 0.1 and with weight decay of 10^{-3}. The batch size is set to 10, 20, and 50 for S-MNIST, CIFAR-100, and miniImageNet datasets respectively.

We notice that MOGD baseline improves if we do more than one gradient descent update for meta backbone. Hence, we use $K > 1$ updates in the experiments. Similar to other meta learning methods, we consider the number of gradient descent updates as an hyper parameter for MOGD method.

We search K values in range $\{10, 20\}$, $\{100, 200\}$, and $\{100, 200\}$ for S-MNIST, CIFAR-100, and miniImageNet. In MAML adaptation, we consider the number of gradient steps in $\{1, 5\}$ and the adaptation learning rates in $\{0.1, 0.01\}$. Different from these parameters, MOML has one extra parameter as α. We search α values in range $\{1, 5, 10\}$, $\{1, 0.1.0.01\}$, and $\{0.1, 0.01, 0.001\}$ for S-MNIST, CIFAR-100, and miniImageNet respectively.

Parameter tuning. Tuning hyper parameters is a problem in online learning as such we do not observe the tasks beforehand. To address this issue, we test the methods using Hedge algorithm (Freund & Schapire, 1997). For each method, we train experts using different configurations by changing the hyperparameters. Then, we consider these experts as black boxes and

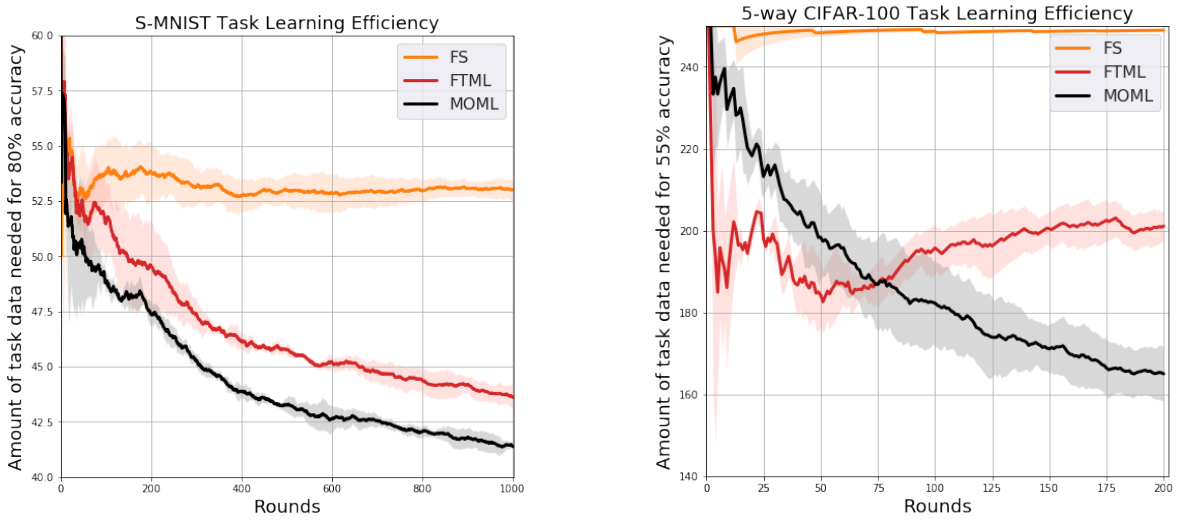

Figure 5. Experiment results of Task Learning Efficiency. **Left**: S-MNIST with efficiency threshold as 80%, **Right**: 5-way CIFAR-100 with proficiency threshold as 55%.

Algorithm 2 Hedge

Input: $\gamma \in [0,1], T, \{e_i\}_{i=1}^n$experts, $\boldsymbol{p}^1 = [0,1]^n$ probability vector initialized to uniform probability,
for $t = 1, 2, \dots T$ **do**
Observe new loss and adaptation function f^t, U^t,
Suffer the average loss of the experts using $\boldsymbol{p}^t$ as $\sum_{i=1}^n \boldsymbol{p}_i^t f^t \circ U^t(\boldsymbol{\theta}_i^t)$ where $\boldsymbol{\theta}_i^t$ is the recent model in expert e_i,
Let experts update their model based on their configurations,
Update $\boldsymbol{p}$ vector as $\boldsymbol{p}_i^{t+1} = \frac{1}{Z}\boldsymbol{p}_i^t \gamma^{\ell_i^t}$ where $Z = \sum_{i=1}^n \boldsymbol{p}_i^t \gamma^{\ell_i^t}$ and ℓ_i^t is the loss of expert i in the current round.
end for

run Hedge method, Algorithm 2, on top of them. Hedge has one parameter γ that trade-offs the confidence on the experts. We fix it to be $\gamma = 0.1$ in our setting.

Hedge algorithm allows us to avoid parameter tuning as such we do not set a configuration in advance. However, it requires to train all experts in parallel. Our tables are based on Hedge Algorithm as such we do not pick a hyperparameter configuration instead, we report the configuration Hedge algorithm chooses in each round.

A.2. Further Ablative Analysis of MOML

Different Model Architectures. We examine the effect of the model architecture. We consider 5-way CIFAR-100 setting and test the methods using with a model that has 5 CNN layers different from the model with 2 CNN layers. Table 5 shows the results for this architecture. As seen from Table 5, our results indicate that MOML is superior to the competitors in this setting as well.

Non-overlapping Classes Experiment. We consider a setting where tasks have completely non-overlapping classes. We constructed tasks with 5 classes out of CIFAR-100 dataset. Since the tasks are non-overlapping, we have in total 20 tasks. Table 6 shows the results for this non-overlapping class setting. We note that this setting is not ideal. As such meta learning needs large number of tasks whereas there are only 20 tasks. Nevertheless, MOML outperforms the competitors.

Ablation study on ∇ state. MOML introduces a regularizer $\mathcal{R}^t(\boldsymbol{\theta})$ (Eq. 6) to modify task objectives. $\mathcal{R}^t(\boldsymbol{\theta})$ consists of two states which are $\nabla f^{t-1} \circ U^{t-1}(\boldsymbol{\theta}^t)$ and $\boldsymbol{w}^t$. According to Proposition 1, $\boldsymbol{w}$ and $\boldsymbol{\theta}$ converge to the same model. The linear term with $\nabla f^{t-1} \circ U^{t-1}(\boldsymbol{\theta}^t)$ state can be interpreted as an adjustment to the optimization problem. To see the effect of this term, we test a variant of MOML where this linear term is assumed to be $\mathbf{0}$ in Algorithm 3 . Table 7 shows comparison between original MOML and the variant with eliminating the linear term (MOML $_{[\nabla=\mathbf{0}]}$) on S-MNIST and 5-way CIFAR-100 settings.

Algorithm 3 MOML $_{\nabla=\mathbf{0}}$

Input: $T, \boldsymbol{w}^1 = \boldsymbol{\theta}^1 = \mathbf{0}, \alpha, K, \beta$
for $t = 1, 2, \dots T$ **do**
  Output $\boldsymbol{\theta}^t$, reveal f^t and U^t, suffer $f^t \circ U^t(\boldsymbol{\theta}^t)$,
  $\mathcal{R}^t(\boldsymbol{\theta}) = \frac{\alpha}{2} \|\boldsymbol{\theta} - \boldsymbol{w}^t\|^2$,
  $\boldsymbol{\theta}_1^{t+1} = \boldsymbol{\theta}^{t-1}$,
  for $k = 1, 2, \dots K$ **do**
    $\boldsymbol{\theta}_{k+1}^{t+1} = \boldsymbol{\theta}_k^{t+1} - \beta \left(\nabla f^t \circ U^t(\boldsymbol{\theta}_k^{t+1}) + \nabla \mathcal{R}^t(\boldsymbol{\theta}_k^{t+1}) \right)$
  end for
  $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}_{K+1}^{t+1}$,
  $\boldsymbol{w}^{t+1} = \frac{1}{2} \left(\boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} \right)$,
end for

Table 5. MOML and FTML performances for 2 CNN layers and 5 CNN layers backbones for 5 way-CIFAR-100 settings.

Test Accuracy		**2 CNN**	**5 CNN**
FTML	CTM	50.68	54.90
	LTM	54.50	62.92
MOML	CTM	55.83	56.17
	LTM	60.78	64.72

Table 6. Non-overlapping classes performance for 5 way-CIFAR-100 settings.

Test Accuracy	**MOGD**	**FTML**	**MOML**
CTM	36.89	38.38	39.32
LTM	45.82	50.90	50.54

Table 7. Ablative study on the regularizer term of MOML.

	Test Accuracy	
S-MNIST	**CTM**	**LTM**
MOML	85.82	87.49
MOML $_{\nabla=\mathbf{0}}$	84.09	86.53
5 way-CIFAR-100		
MOML	55.83	60.78
MOML $_{\nabla=\mathbf{0}}$	51.33	57.87

MOML is better than MOML $_{\nabla=\mathbf{0}}$ variant. The linear term improves the performance of MOML. For instance, in 5 way-CIFAR-100, there is an increase of 4.5% in CTM accuracy as seen in Table 7. This improvement is consistent with our theory. As such the linear term debiases the current loss and it is essential in deriving regret guarantees of MOML.

A.3. Proof

A.3.1. Convex Analysis

Assumption 1 *(Stationary point) We assume that MOML finds a stationary point of the risk it minimizes. Formally, at each round, MOML satisfies*

$$\nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1}) - \nabla f^{t-1} \circ U^{t-1}(\boldsymbol{\theta}^t) + \alpha \left(\boldsymbol{\theta}^{t+1} - \boldsymbol{w}^t \right) = \mathbf{0}.$$

This assumption can be achieved by tuning parameter K. Parameter K controls the correctness of the solution as such it would introduce $O\left(\frac{1}{K}\right)$ noise at each round. Choosing $K = O\left(\sqrt{T}\right)$ would be sufficient.

Assumption 2 *(Bounded gradients with G)* $\{f^t \circ U^t\}_{t=1}^T$ *functions have bounded gradients with* G *.i.e*

$$\|\nabla f^t \circ U^t(\boldsymbol{\theta})\| \leq G \quad \forall \boldsymbol{\theta}, t$$

Theorem 3 *For adversarial convex* $\{f^t \circ U^t\}_{t=1}^T$ *functions we have, Algorithm 1 satisfies,*

$$R_T = \sum_{t=1}^{T} f^t \circ U^t(\boldsymbol{\theta}^t) - \sum_{t=1}^{T} f^t \circ U^t(\boldsymbol{\theta}^*) \leq \alpha \|\boldsymbol{\theta}^*\|^2 + \frac{3}{\alpha} \sum_{t=1}^{T} \|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2,$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{t=1}^{T} f^t \circ U^t(\boldsymbol{\theta})$.

If we plug in $\alpha = O\left(\sqrt{T}\right)$ and bound gradients with G in Theorem 3, we recover Theorem 1.

Proof.

We divide LHS with two terms as,

$$R_T = \left(\sum_{t=1}^{T} f^t \circ U^t(\boldsymbol{\theta}^t) - f^t \circ U^t(\boldsymbol{\theta}^{t+1})\right) + \left(\sum_{t=1}^{T} f^t \circ U^t(\boldsymbol{\theta}^{t+1}) - f^t \circ U^t(\boldsymbol{\theta}^*)\right)$$

where first term corresponds to the cost we incur by not using $\boldsymbol{\theta}^{t+1}$ and the second term quantifies how good $\boldsymbol{\theta}^{t+1}$ is with respect to the competitor.

For the sake of simplicity, let us define the local states with the gradient as

$$\boldsymbol{\lambda}^t = \nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1}). \tag{12}$$

We bound individual terms with the following Lemmas,

Lemma 1 *Algorithm 1 satisfies,*

$$\begin{aligned} f^t \circ U^t(\boldsymbol{\theta}^t) - f^t \circ U^t(\boldsymbol{\theta}^{t+1}) \leq & \frac{3}{\alpha} \|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2 \\ & + \frac{\alpha}{4} \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t\right) \right\|^2 + \frac{\alpha}{4} \left\| \boldsymbol{\theta}^t - \left(\boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1}\right) \right\|^2 + \frac{\alpha}{16} \|\boldsymbol{\lambda}^t\|^2 \end{aligned}$$

Lemma 2 *Algorithm 1 satisfies,*

$$\begin{aligned} f^t \circ U^t(\boldsymbol{\theta}^{t+1}) - f^t \circ U^t(\boldsymbol{\theta}^*) \leq & \alpha \left(\left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1}\right) \right\|^2 - \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t\right) \right\|^2 \right) \\ & + \frac{1}{4\alpha} \left(\|\boldsymbol{\lambda}^{t-1}\|^2 - \|\boldsymbol{\lambda}^t\|^2 \right) - \frac{\alpha}{2} \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t\right) \right\|^2 - \frac{1}{4\alpha} \|\boldsymbol{\lambda}^t\|^2 \end{aligned}$$

If we plug in Lemma 1 and 2 in the regret statement we get,

$$\begin{aligned} R_T \leq & \alpha \left(\sum_{t=1}^{T} \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1}\right) \right\|^2 - \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t\right) \right\|^2 \right) \\ & + \frac{\alpha}{4} \left(\sum_{t=1}^{T} \left\| \boldsymbol{\theta}^t - \left(\boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1}\right) \right\|^2 - \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t\right) \right\|^2 \right) \\ & + \frac{1}{4\alpha} \left(\sum_{t=1}^{T} \|\boldsymbol{\lambda}^{t-1}\|^2 - \|\boldsymbol{\lambda}^t\|^2 \right) + \frac{3}{\alpha} \sum_{t=1}^{T} \|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2 - \frac{3}{16\alpha} \sum_{t=1}^{T} \|\boldsymbol{\lambda}^t\|^2 \end{aligned}$$

$$\leq \alpha \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^1 + \frac{1}{2\alpha}\boldsymbol{\lambda}^0 \right) \right\|^2 + \frac{\alpha}{4} \left\| \boldsymbol{\theta}^1 - \left(\boldsymbol{w}^1 + \frac{1}{2\alpha}\boldsymbol{\lambda}^0 \right) \right\|^2 + \frac{1}{4\alpha}\|\boldsymbol{\lambda}^0\|^2 + \frac{3}{\alpha}\sum_{t=1}^{T} \|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2$$

$$= \alpha \|\boldsymbol{\theta}^*\|^2 + \frac{3}{\alpha}\sum_{t=1}^{T} \|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2$$

where the second inequality is due to telescoping and ignoring non-positive terms and the last equality comes from the initial conditions $\boldsymbol{\lambda}^0 = \boldsymbol{w}^1 = \boldsymbol{\theta}^1 = \mathbf{0}$. This completes the proof of Theorem 3. □

We give proof of Lemmas here. We first state two relations that are useful in the proof.

$$\left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right) = \left(\frac{1}{2}\left(\boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} - \frac{1}{\alpha}\boldsymbol{\lambda}^t \right) + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right)$$

$$= \frac{1}{2}\left(-\boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} - \frac{1}{\alpha}\boldsymbol{\lambda}^{t-1} \right) = -\frac{1}{2\alpha}\boldsymbol{\lambda}^t, \tag{13}$$

where the equalities are due to $\boldsymbol{w}$ update (Eq. 8) and first order condition (Eq. 10) respectively.

Proof of Lemma 1

$$f^t \circ U^t(\boldsymbol{\theta}^t) - f^t \circ U^t(\boldsymbol{\theta}^{t+1}) \leq \langle \nabla f^t \circ U^t(\boldsymbol{\theta}^t), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{t+1} \rangle \leq \|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\| \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t+1}\|$$

$$\leq \frac{3}{\alpha}\|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2 + \frac{\alpha}{12}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t+1}\|^2$$

$$\leq \frac{3}{\alpha}\|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2 + \frac{\alpha}{4}\left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) \right\|^2 + \frac{\alpha}{4}\left\| \boldsymbol{\theta}^t - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right) \right\|^2$$

$$+ \frac{\alpha}{4}\left\| \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right) \right\|^2$$

$$= \frac{3}{\alpha}\|\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\|^2$$

$$+ \frac{\alpha}{4}\left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) \right\|^2 + \frac{\alpha}{4}\left\| \boldsymbol{\theta}^t - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right) \right\|^2 + \frac{\alpha}{16}\|\boldsymbol{\lambda}^t\|^2$$

where inequalities come from convexity, Cauchy–Schwarz Eq., $2\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \rangle \leq \frac{1}{c}\|\boldsymbol{\theta}_1\|^2 + c\|\boldsymbol{\theta}_2\|^2$ and triangular inequality respectively. Last equality is due to Eq. 13. □

We state a useful relation to be used in the proof of Lemma 2.

Proposition 2

$$-\alpha \left\| \boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right\|^2 + \alpha \left\| \boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right\|^2 + \langle \boldsymbol{\lambda}^t, \boldsymbol{\theta}^{t+1} \rangle = \frac{1}{4\alpha}\left(\|\boldsymbol{\lambda}^{t-1}\|^2 - \|\boldsymbol{\lambda}^t\|^2 \right) - \alpha \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) \right\|^2 - \frac{1}{4\alpha}\|\boldsymbol{\lambda}^t\|^2$$

Proof of Lemma 2

$$f^t \circ U^t(\boldsymbol{\theta}^{t+1}) - f^t \circ U^t(\boldsymbol{\theta}^*) \leq \langle \nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1}), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^* \rangle = \langle \boldsymbol{\lambda}^t, \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^* \rangle$$

$$= \alpha \left(\left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right) \right\|^2 - \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) \right\|^2 \right)$$

$$- \alpha \left\| \boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right\|^2 + \alpha \left\| \boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right\|^2 + \langle \boldsymbol{\lambda}^t, \boldsymbol{\theta}^{t+1} \rangle$$

$$= \alpha \left(\left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^t + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \right) \right\|^2 - \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) \right\|^2 \right)$$

$$+ \frac{1}{4\alpha}\left(\|\boldsymbol{\lambda}^{t-1}\|^2 - \|\boldsymbol{\lambda}^t\|^2 \right) - \alpha \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha}\boldsymbol{\lambda}^t \right) \right\|^2 - \frac{1}{4\alpha}\|\boldsymbol{\lambda}^t\|^2$$

$$\leq \alpha \left(\left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1} \right) \right\|^2 - \left\| \boldsymbol{\theta}^* - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t \right) \right\|^2 \right) + \frac{1}{4\alpha} \left(\|\boldsymbol{\lambda}^{t-1}\|^2 - \|\boldsymbol{\lambda}^t\|^2 \right) - \frac{\alpha}{2} \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t \right) \right\|^2 - \frac{1}{4\alpha} \|\boldsymbol{\lambda}^t\|^2$$

where we use convexity, Eq. 12, 13 and Proposition 2.

We give the proof of the proposition here.

Proof of Proposition 2

Let us expand LHS and state it as a polynomial of $\boldsymbol{\theta}^{t+1}$, .i.e $A\|\boldsymbol{\theta}^{t+1}\|^2 + \langle \boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}' \rangle + \boldsymbol{\theta}''$

$$\begin{aligned}
LHS =& -\alpha \left\| \boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1} \right\|^2 + \alpha \left\| \frac{1}{2} \left(\boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} - \frac{1}{\alpha} \boldsymbol{\lambda}^t \right) + \frac{1}{2\alpha} \boldsymbol{\lambda}^t \right\|^2 + \langle \boldsymbol{\lambda}^t, \boldsymbol{\theta}^{t+1} \rangle \\
=& -\alpha \left\| \boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1} \right\|^2 + \frac{\alpha}{4} \left\| \boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} \right\|^2 + \langle \boldsymbol{\lambda}^t, \boldsymbol{\theta}^{t+1} \rangle \\
=& -\alpha \left\| \boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1} \right\|^2 + \frac{\alpha}{4} \left\| \boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} \right\|^2 + \langle \boldsymbol{\lambda}^{t-1} + \alpha \boldsymbol{w}^t - \alpha \boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+1} \rangle \\
=& \|\boldsymbol{\theta}^{t+1}\|^2 \left(\frac{\alpha}{4} - \alpha \right) + \left\langle \boldsymbol{\theta}^{t+1}, \frac{\alpha}{2} \boldsymbol{w}^t + \alpha \boldsymbol{w}^t + \boldsymbol{\lambda}^{t-1} \right\rangle - \alpha \left\| \boldsymbol{w}^t + \frac{1}{2\alpha} \boldsymbol{\lambda}^{t-1} \right\|^2 + \frac{\alpha}{4} \left\| \boldsymbol{w}^t \right\|^2 \\
=& \|\boldsymbol{\theta}^{t+1}\|^2 \left(-\frac{3\alpha}{4} \right) + \left\langle \boldsymbol{\theta}^{t+1}, \frac{3\alpha}{2} \boldsymbol{w}^t + \boldsymbol{\lambda}^{t-1} \right\rangle - \frac{3\alpha}{4} \left\| \boldsymbol{w}^t \right\|^2 - \frac{1}{4\alpha} \left\| \boldsymbol{\lambda}^{t-1} \right\|^2 - \langle \boldsymbol{w}^t, \boldsymbol{\lambda}^{t-1} \rangle
\end{aligned} \tag{14}$$

where we use $\boldsymbol{w}$ and $\boldsymbol{\lambda}$ updates. Similarly if we expand RHS we get,

$$\begin{aligned}
RHS =& \frac{1}{4\alpha} \left(\|\boldsymbol{\lambda}^{t-1}\|^2 - \|\boldsymbol{\lambda}^t\|^2 \right) - \alpha \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t \right) \right\|^2 - \frac{1}{4\alpha} \|\boldsymbol{\lambda}^t\|^2 \\
=& \frac{1}{4\alpha} \|\boldsymbol{\lambda}^{t-1}\|^2 - \frac{1}{2\alpha} \|\boldsymbol{\lambda}^{t-1} + \alpha \boldsymbol{w}^t - \alpha \boldsymbol{\theta}^{t+1}\|^2 - \alpha \left\| \boldsymbol{\theta}^{t+1} - \left(\boldsymbol{w}^{t+1} + \frac{1}{2\alpha} \boldsymbol{\lambda}^t \right) \right\|^2 \\
=& \frac{1}{4\alpha} \|\boldsymbol{\lambda}^{t-1}\|^2 - \frac{1}{2\alpha} \|\boldsymbol{\lambda}^{t-1} + \alpha \boldsymbol{w}^t - \alpha \boldsymbol{\theta}^{t+1}\|^2 - \alpha \left\| \boldsymbol{\theta}^{t+1} - \left(\frac{1}{2} \left(\boldsymbol{w}^t + \boldsymbol{\theta}^{t+1} - \frac{1}{\alpha} \boldsymbol{\lambda}^t \right) + \frac{1}{2\alpha} \boldsymbol{\lambda}^t \right) \right\|^2 \\
=& \frac{1}{4\alpha} \|\boldsymbol{\lambda}^{t-1}\|^2 - \frac{1}{2\alpha} \|\boldsymbol{\lambda}^{t-1} + \alpha \boldsymbol{w}^t - \alpha \boldsymbol{\theta}^{t+1}\|^2 - \frac{\alpha}{4} \left\| \boldsymbol{\theta}^{t+1} - \boldsymbol{w}^t \right\|^2 \\
=& \|\boldsymbol{\theta}^{t+1}\|^2 \left(-\frac{\alpha}{2} - \frac{\alpha}{4} \right) + \left\langle \boldsymbol{\theta}^{t+1}, \alpha \boldsymbol{w}^t + \boldsymbol{\lambda}^{t-1} + \frac{\alpha}{2} \boldsymbol{w}^t \right\rangle + \frac{1}{4\alpha} \|\boldsymbol{\lambda}^{t-1}\|^2 - \frac{\alpha}{4} \|\boldsymbol{w}^t\|^2 - \frac{1}{2\alpha} \|\boldsymbol{\lambda}^{t-1} + \alpha \boldsymbol{w}^t\|^2 \\
=& \|\boldsymbol{\theta}^{t+1}\|^2 \left(-\frac{3\alpha}{4} \right) + \left\langle \boldsymbol{\theta}^{t+1}, \frac{3\alpha}{2} \boldsymbol{w}^t + \boldsymbol{\lambda}^{t-1} \right\rangle - \frac{3\alpha}{4} \left\| \boldsymbol{w}^t \right\|^2 - \frac{1}{4\alpha} \left\| \boldsymbol{\lambda}^{t-1} \right\|^2 - \langle \boldsymbol{w}^t, \boldsymbol{\lambda}^{t-1} \rangle
\end{aligned} \tag{15}$$

where we use $\boldsymbol{w}$ and $\boldsymbol{\lambda}$ updates.

We have Eq. 14 is equal to Eq. 15 so the statement holds. □

A.4. Nonconvex Analysis

We prove regret statements of B-MOML and Theorem 2 with the following subsections.

A.4.1. (HAZAN ET AL., 2017) REGRET WITH B-MOML

Assumption 3 *(Stationary point) Similar to Assumption 1, we assume that B-MOML finds a stationary point of the risk it minimizes. Formally, at each round, B-MOML satisfies*

$$\frac{1}{B} \sum_{i=0}^{B-1} \nabla f^{t-i} \circ U^{t-i}(\boldsymbol{\theta}^{t+1}) - \frac{1}{B} \sum_{i=0}^{B-1} \nabla f^{t-i-1} \circ U^{t-i-1}(\boldsymbol{\theta}^{t-i}) + \alpha \left(\boldsymbol{\theta}^{t+1} - \boldsymbol{w}^t \right) = \mathbf{0}.$$

We use Assumption 2 and 3 in this subsection.

Theorem 4 *For adversarial nonconvex functions, B-MOML satisfies,*

$$\sum_{t=1}^{T}\left\|\frac{1}{B}\sum_{i=0}^{B-1}\nabla f^{t-i}\circ U^{t-i}(\boldsymbol{\theta}^t)\right\|^2 \leq 8\frac{T}{B^2}G^2$$

Proof. Let us define $S_B^t(\boldsymbol{\theta}) = \frac{1}{B}\sum_{i=0}^{B-1}\nabla f^{t-i}\circ U^{t-i}(\boldsymbol{\theta})$. Then, we have,

$$\begin{aligned}\left\|\frac{1}{B}\sum_{i=0}^{B-1}\nabla f^{t-i}\circ U^{t-i}(\boldsymbol{\theta}^t)\right\|^2 &= \|\nabla S_B^t(\boldsymbol{\theta}^t)\|^2 = \|\left(\nabla S_B^t(\boldsymbol{\theta}^t) - \nabla S_B^{t-1}(\boldsymbol{\theta}^t)\right) + \nabla S_B^{t-1}(\boldsymbol{\theta}^t)\|^2\\ &\leq 2\|\nabla S_B^{t-1}(\boldsymbol{\theta}^t)\|^2 + 2\|\nabla S_B^t(\boldsymbol{\theta}^t) - \nabla S_B^{t-1}(\boldsymbol{\theta}^t)\|^2\\ &= 2\|\nabla S_B^{t-1}(\boldsymbol{\theta}^t)\|^2 + \frac{2}{B^2}\left\|\nabla f^t\circ U^t(\boldsymbol{\theta}^t) - \nabla f^{t-B}\circ U^{t-B}(\boldsymbol{\theta}^t)\right\|^2\\ &\leq 2\|\nabla S_B^{t-1}(\boldsymbol{\theta}^t)\|^2 + \frac{8}{B^2}G^2 \end{aligned} \tag{16}$$

where inequalities come from triangular Inq. and Assumption 2.

We assume $\alpha = 0$. If $\alpha = 0$ we see that $\frac{1}{B}\sum_{i=0}^{B-1}\nabla f^{t-i}\circ U^{t-i}(\boldsymbol{\theta}^{t+1}) = \nabla S_B^t(\boldsymbol{\theta}^{t+1}) = \mathbf{0}$.

Plugging these relations in Eq. 16, we get,

$$\left\|\frac{1}{B}\sum_{i=0}^{B-1}\nabla f^{t-i}\circ U^{t-i}(\boldsymbol{\theta}^t)\right\|^2 \leq 2\|\nabla S_B^{t-1}(\boldsymbol{\theta}^t)\|^2 + \frac{8}{B^2}G^2 = 2\|\boldsymbol{\lambda}^{t-1}\|^2 + \frac{8}{B^2}G^2 = \frac{8}{B^2}G^2 \tag{17}$$

Summing Inq. 17 over time gives the statement in Theorem 4. □

A.4.2. $\mathcal{T}$ COLLECTION OF TASKS TYPE REGRET

Assumption 4 *(Smoothness) We assume $\{f^t\circ U^t\}_{t=1}^T$ functions to be L smooth .i.e*

$$\|\nabla f^t\circ U^t(\boldsymbol{\theta}_1) - \nabla f^t\circ U^t(\boldsymbol{\theta}_2)\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \;\; \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, t$$

Smoothness imply the following inequality,

$$f^t\circ U^t(\boldsymbol{\theta}_2) - f^t\circ U^t(\boldsymbol{\theta}_1) \leq \left\langle \nabla f^t\circ U^t(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\right\rangle + \frac{L}{2}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1|^2 \;\; \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, t \tag{18}$$

We use Assumption 1, 2 and 4 as well definition in Eq. 12 for this subsection.

Theorem 5 *Suppose $\mathcal{T}$ is a collection of tasks, and for each $\tau\in\mathcal{T}$, $f^\tau\circ U^\tau$ is L smooth. We choose tasks i_t from some task distribution, $P_{\mathcal{T}}$ in an IID fashion. Then it follows that,*

$$\sum_{t=1}^{T} E\left\|E_\tau[\nabla f^\tau\circ U^\tau(\boldsymbol{\theta}^t)]\right\|^2 \leq 4\alpha\Delta + T\frac{G^2L^2}{\alpha^2} + T\frac{1}{\alpha}\frac{11}{2}G^2L.$$

where $\Delta = E_\tau[f^\tau\circ U^\tau(\boldsymbol{\theta}^1)] - \min_{\boldsymbol{\theta}} E_\tau[f^\tau\circ U^\tau(\boldsymbol{\theta})]$

If we have $\alpha = O\left(\sqrt{T}\right)$, we get the bound in Theorem 2. Theorem 5 is can be derived from the following Lemma,

Lemma 3 *Algorithm 1 satisfies,*

$$\begin{aligned}E\left\|E_\tau\nabla f^\tau\circ U^\tau(\boldsymbol{\theta}^t)\right\|^2 \leq& 4\alpha\left(E\left[E_\tau f^\tau\circ U^\tau\left(\frac{1}{2}\left(\boldsymbol{w}^{t-1} + \boldsymbol{\theta}^t\right)\right)\right] - E\left[E_\tau f^\tau\circ U^\tau\left(\frac{1}{2}\left(\boldsymbol{w}^{t} + \boldsymbol{\theta}^{t+1}\right)\right)\right]\right)\\ &+ \frac{G^2L^2}{\alpha^2} + \frac{1}{\alpha}\frac{11}{2}G^2L\end{aligned}$$

If we sum Lemma 3 over time we get

$$\begin{aligned}\sum_{t=1}^{T} E\left\|E_\tau \nabla f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\right\|^2 \leq & 4\alpha\left(E\left[E_\tau f^\tau \circ U^\tau\left(\frac{1}{2}\left(\boldsymbol{w}^0+\boldsymbol{\theta}^1\right)\right)\right]-E\left[E_\tau f^\tau \circ U^\tau\left(\frac{1}{2}\left(\boldsymbol{w}^T+\boldsymbol{\theta}^{T+1}\right)\right)\right]\right) \\ & +T\frac{G^2L^2}{\alpha^2}+T\frac{1}{\alpha}\frac{11}{2}G^2L \\ \leq & 4\alpha\Delta+T\frac{G^2L^2}{\alpha^2}+T\frac{1}{\alpha}\frac{11}{2}G^2L\end{aligned}$$

which is the statement in Theorem 5.

We use smoothness bound (Eq. 18) to prove Lemma 3. First, we present a set of Lemmas that are useful to handle various terms and invoke these to prove this statement.

Lemma 4 *Define* $\boldsymbol{\rho}^t=\frac{1}{2}\left(\boldsymbol{w}^t+\boldsymbol{\theta}^{t+1}\right)$. *Then, in Algorithm 1, we have,*

$$\boldsymbol{\rho}^t-\boldsymbol{\rho}^{t-1}=-\frac{1}{2\alpha}\boldsymbol{\lambda}^t$$

Proof.

$$\frac{1}{2}\left(\boldsymbol{w}^t+\boldsymbol{\theta}^{t+1}\right)-\frac{1}{2}\left(\boldsymbol{w}^{t-1}+\boldsymbol{\theta}^t\right)=\frac{1}{2}\left(\boldsymbol{w}^t+\boldsymbol{\theta}^{t+1}\right)-\boldsymbol{w}^t-\frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1}=\frac{1}{2}\left(-\boldsymbol{w}^t+\boldsymbol{\theta}^{t+1}\right)-\frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1}=-\frac{1}{2\alpha}\boldsymbol{\lambda}^t$$

where first equality uses definition of $\boldsymbol{w}^t$ (Eq. 8) and the last one comes from the first order condition (Eq. 10). □

Lemma 5 *Algorithm 1 satisfies,*

$$E\left[\left\langle\nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right),-\boldsymbol{\lambda}^t\right\rangle\right] \leq \frac{1}{\alpha}\frac{5}{2}G^2L+\frac{1}{2\alpha^2}G^2L^2-\frac{1}{2}E\|E_\tau\nabla f^\tau\circ U^\tau(\boldsymbol{\theta}^t)\|^2$$

We use smoothness Eq. 18 on $E_\tau\left[f^\tau \circ U^\tau\right]$as,

$$\begin{aligned}E\left[E_\tau\left[f^\tau \circ U^\tau\left(\boldsymbol{\rho}^t\right)\right]\right]-E\left[E_\tau\left[f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right)\right]\right] \leq & E\left[\left\langle\nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right),\boldsymbol{\rho}^t-\boldsymbol{\rho}^{t-1}\right\rangle\right]+\frac{L}{2}E\|\boldsymbol{\rho}^t-\boldsymbol{\rho}^{t-1}\|^2 \\ = & \frac{1}{2\alpha}E\left[\left\langle\nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right),-\boldsymbol{\lambda}^t\right\rangle\right]+\frac{L}{8\alpha^2}E\|\boldsymbol{\lambda}^t\|^2 \\ \leq & \frac{1}{\alpha^2}\frac{5}{4}G^2L+\frac{1}{4\alpha^3}G^2L^2-\frac{1}{4\alpha}E\|E_\tau\nabla f^\tau\circ U^\tau(\boldsymbol{\theta}^t)\|^2+\frac{LG^2}{8\alpha^2}\end{aligned}$$

where we use Lemma 4, 5 and bound $\|\boldsymbol{\lambda}\|$ with Assumption 2. Rearranging terms give the statement in Lemma 3. □

Corollary 1 is a direct consequence of Theorem 5 in case of Polyak- Lojasiewicz (PL) functions. It gives a statement with respect to the best fixed competitor.

Proof of Corollary 1 Let's apply PL condition on LHS of Theorem 5. We get,

$$\sum_{t=1}^{T} E\left[E_\tau[f^\tau \circ U^\tau(\boldsymbol{\theta}^t)]-\min_{\boldsymbol{\theta}} E_\tau[f^\tau \circ U^\tau(\boldsymbol{\theta})]\right] \leq \frac{1}{2\mu}\sum_{t=1}^{T} E\left\|E_\tau[\nabla f^\tau \circ U^\tau(\boldsymbol{\theta}^t)]\right\|^2=O\left(\sqrt{T}\frac{1}{\mu}\left(\Delta+G^2L\right)\right)$$

This is the Corollary statement. □

We give proof of Lemma 5 here. We state two more Lemmas that are used in the proof.

Lemma 6 *Difference of consecutive meta models can be bounded as,*

$$\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^t\| \leq \frac{5}{2}\frac{1}{\alpha}G$$

Proof.

If we subtract the first order condition (Eq. 10) for consecutive times, we get,

$$\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-1} = \boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^{t-2} - \alpha(\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t) + \alpha(\boldsymbol{w}^t - \boldsymbol{w}^{t-1}) \tag{19}$$

Rearranging Eq. 19 gives,

$$\begin{aligned}\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t &= \frac{1}{\alpha}(2\boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-2}) + (\boldsymbol{w}^t - \boldsymbol{w}^{t-1}) = \frac{1}{\alpha}(2\boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-2}) + \frac{1}{2}(\boldsymbol{\theta}^t - \boldsymbol{w}^{t-1}) - \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} \\ &= \frac{1}{\alpha}(2\boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-2}) + \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-2} - \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} - \frac{1}{2\alpha}\boldsymbol{\lambda}^{t-1} = \frac{1}{\alpha}\left(\boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^t - \frac{1}{2}\boldsymbol{\lambda}^{t-2}\right)\end{aligned}$$

where we use update rule of $\boldsymbol{w}^t$ and $\boldsymbol{\lambda}^t$ in the second respectively. Since $\boldsymbol{\lambda}$ states store gradient information as in Eq. 12 and gradients are bounded (Assumption 2), we can write,

$$\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\| = \frac{1}{\alpha}\left\|\boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^t - \frac{1}{2}\boldsymbol{\lambda}^{t-2}\right\| \leq \frac{1}{\alpha}\left\|\boldsymbol{\lambda}^{t-1}\right\| + \frac{1}{\alpha}\left\|\boldsymbol{\lambda}^t\right\| + \frac{1}{2}\frac{1}{\alpha}\left\|\boldsymbol{\lambda}^{t-2}\right\| \leq \frac{5}{2}\frac{1}{\alpha}G.$$

where we relax norm as $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$. □

Lemma 7

$$E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle\right] \leq \frac{1}{2\alpha^2}G^2L^2 - \frac{1}{2}E\|E_\tau \nabla f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\|^2$$

Proof.

$$\begin{aligned}E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle\right] =& E\left[E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle | \mathcal{H}_t\right]\right] \\ =& E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla E_\tau f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\right\rangle\right] \\ \leq& -\frac{1}{2}E\|E_\tau \nabla f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\|^2 \\ &+ \frac{1}{2}E\left\|\nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right) - \nabla E_\tau f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\right\|^2\end{aligned}$$

where first equality comes from tower property noting that both $\boldsymbol{\rho}^{t-1}$ and $\boldsymbol{\theta}^t$ are independent of loss observed at time t, $f^t \circ U^t$. The inequality comes from $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{b} + \boldsymbol{a}\|^2 - \frac{1}{2}\|\boldsymbol{a}\|^2$. We bound the second term as

$$\begin{aligned}E\left\|\nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right) - \nabla E_\tau f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\right\|^2 \leq& L^2 E\left\|\boldsymbol{\rho}^{t-1} - \boldsymbol{\theta}^t\right\|^2 \leq \frac{1}{4}L^2 E\left\|\boldsymbol{w}^{t-1} - \boldsymbol{\theta}^t\right\|^2 \\ \leq& \frac{1}{4\alpha^2}L^2 E\left\|\boldsymbol{\lambda}^{t-1} - \boldsymbol{\lambda}^{t-2}\right\|^2 \leq \frac{L^2G^2}{\alpha^2}\end{aligned}$$

where we use smoothness, definition of $\boldsymbol{\rho}^{t-1}$, the first order condition (Eq. 10) and bound on gradients. Combining both inequalities gives the statement in the lemma. □

Proof of Lemma 5

$$\begin{aligned}E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\boldsymbol{\lambda}^t\right\rangle\right] =& E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1})\right\rangle\right] \\ =& E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle\right] \\ &+ E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), \nabla f^t \circ U^t(\boldsymbol{\theta}^t) - \nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1})\right\rangle\right] \\ \leq& E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle\right] \\ &+ E\left[\left\|\nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right)\right\| \left\|\nabla f^t \circ U^t(\boldsymbol{\theta}^t) - \nabla f^t \circ U^t(\boldsymbol{\theta}^{t+1})\right\|\right] \\ \leq& E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle\right] + E\left[GL\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t+1}\|\right] \\ \leq& E\left[\left\langle \nabla E_\tau f^\tau \circ U^\tau\left(\boldsymbol{\rho}^{t-1}\right), -\nabla f^t \circ U^t(\boldsymbol{\theta}^t)\right\rangle\right] + \frac{5}{2}\frac{1}{\alpha}G^2L \\ \leq& \frac{5}{2}\frac{1}{\alpha}G^2L + \frac{1}{2\alpha^2}G^2L^2 - \frac{1}{2}E\|E_\tau \nabla f^\tau \circ U^\tau(\boldsymbol{\theta}^t)\|^2\end{aligned}$$

where we use Cauchy–Schwarz Eq., smoothness, gradient bound, Lemma 6 and 7 respectively. □