1 Theoretical Analysis

The core of gcp-sampling is to adaptively sample tasks during meta-training. Hence, in this section, we theoretically analyze the advance of such a sampling method in terms of generalization bound. We first provide a generic generalization bound for task sampling. Then, we connect the generalization bound to the proposed task adaptive sampling (cp-sampling and gcp-sampling).

1.1 The Generalization Bound for Task Sampling Distribution

Given a meta-training dataset $D_{tr}$ with a category set $C_{tr}$ and each class including $L$ images, we assume a sequence of different meta-training tasks $T = \{(S_1, Q_1), \ldots, (S_{n_0}, Q_{n_0})\}$. Each task is generated by first sampling $K$ classes $L^K \sim C_{tr}$ and then sampling $M$ and $N$ images per class. Therefore, we have $n_0 = \binom{|C_{tr}|}{K} \left( \frac{L}{M+N} \right)^K \binom{M+N}{M}^{K}$ different tasks, where $\binom{j}{i}$ denotes the number of combinations of $j$ objects chosen from $i$ objects.

Let $\ell(\theta, S, Q)$ denote the task loss w.r.t model parameter $\theta$ and task $(S, Q)$. The ultimate goal of meta-learning algorithm is to have low expected task error, i.e. $er(\theta) = \mathbb{E}_{S, Q} \ell(\theta, S, Q)$. Since the underlying task distribution is unknown, we approximate it by the empirical task error over the meta-training tasks $T$, i.e. $er(\theta) = \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(\theta, S_i, Q_i)$. By bounding the difference of the two, we obtain an upper bound on $er(\theta)$.

In the meta-learning framework, we formulate the episodic training algorithm as $A(T, \sigma) \rightarrow \theta$, which produces the model parameter $\theta$ based on $T$ and some hyperparameters $\sigma$. Similar to [6], we could view the randomized episodic training algorithm as a deterministic learning algorithm whose hyperparameters are randomized. In particular, the episodic training performs a sequence of updates, for $t = 1, \ldots, T$, in the following way,

$$\theta_t \leftarrow U_t(\theta_{t-1}, S_{i_t}, Q_{i_t}),$$

where $U_t(\cdot)$ is an optimizer. It deals with a sequence of random task indices $\sigma = (i_1, \ldots, i_T)$, sampled according to a distribution $P$ on hyperparameter space.
\[ \Sigma = \{1, \ldots, n_0\}^T. \] This can be viewed as drawing \( \sigma \sim P \) based on \( T \) first, and then executing a sequence of updates by running a deterministic algorithm \( A(T, \sigma) \). Based on this, the expected task error and empirical task error are given by averaging over task distribution \( P \), namely \( er(P) = \mathbb{E}_{\theta \sim P} \mathbb{E}_{S, Q} \ell(\theta, S, Q) \) and \( \hat{er}(P) = \mathbb{E}_{\theta \sim P} \sum_{i=1}^{n_0} \ell(\theta_i, S_i, Q_i) \).

The distribution on the hyperparameter space \( \Sigma \) induces a distribution on hypothesis space. Then, we can find a direct connection between \( \mathbb{E}_{\theta \sim P} \ell(\theta, S, Q) \) and the Gibbs loss, which has been studied extensively using PAC-Bayes analysis \([3, 1, 7]\). According to the Catoni’s PAC-Bayes bound \([1]\), we could derive a generalization bound w.r.t. adaptive task sampling distribution \( Q \) on hyperparameter space \( \Sigma \).

**Theorem 1** Let \( P \) be some prior distribution over hyperparameter space \( \Sigma \). Then for any \( \delta \in (0, 1] \), and any real number \( c > 0 \), the following inequality holds uniformly for all posteriors distribution \( Q \) with probability at least \( 1 - \delta \),

\[
er(Q) \leq \frac{c}{1 - e^{-c}} \left[ \hat{er}(Q) + \frac{KL(Q\|P) + \log \frac{1}{n_0}}{n_0 c} \right]. \tag{2}
\]

Theorem 1 indicates that the expected task error \( er(Q) \) is upper bounded by the empirical task error plus a penalty \( KL(Q\|P) \). Since the bound holds uniformly for all \( Q \), it also holds for data-dependent \( Q \). By choosing \( Q \) that minimizes the bound, we obtain a data-dependent task distribution with generalization guarantees.

### 1.2 Connection to cp-sampling (gcp-sampling)

According to Theorem 1, to improve the generalization performance, the posterior sampling distribution \( Q \) should put its attention on the important task which is valuable for reducing empirical error. On the other hand, the posterior sampling distribution \( Q \) should be close to the prior \( P \) to control the divergence penalty. Moreover, the posterior is required to dynamically adapt to episodic training, which is a dynamic conditional distribution on the previous iteration \( Q(t)(i) = Q(t)(i_1, \ldots, i_{t-1}) \). Therefore, we choose the task sampling distribution at \( t + 1 \) by maximizing the expected utility over tasks while minimizing the KL penalty w.r.t. a reference distribution. It can be formulated as the following optimization problem:

\[
\max_{Q_t^{t+1} \in \Delta_{n_0}} \sum_{i=1}^{n_0} Q_i^{t+1}(i) f(\theta_t, S_i, Q_i) - \frac{1}{\alpha} KL(Q_t^{t+1}\|Q_t^t), \tag{3}
\]

where \( Q^0 \) is a uniform distribution, \( \alpha \) and \( \tau \) are hyperparameters that control the impact of current update and previous updates, \( f(\theta_t, S_i, Q_i) \) denotes the utility function of the chosen task and current model parameter. However, the two-level sampling for generating task makes \( n_0 \) quite large (\( n_0 = \ldots \)
\( \binom{|C_t|}{K} \left( \binom{L}{M+N} \binom{M+N}{M} \right)^K \). It is infeasible to maintain a distribution \( Q \) on \( \{1, \ldots, n_0\} \). Therefore, we propose to sample \( K \) classes \( L_K \) for each task and adopt uniform sampling to generate the support set and query set for each class, respectively. Then, we consider the following optimization problem w.r.t category set \( L_{t+1}^K \):

\[
\max_{p(\mathbb{L}_{t+1}^K) \in \Delta^{n_1}} \sum_{\mathbb{S}, \mathbb{Q}} p(\mathbb{L}_{t+1}^K) \mathbb{E}_{\mathbb{S}, \mathbb{Q}} f(\theta_t, \mathbb{S}, \mathbb{Q}) - \frac{1}{\alpha} KL(p(\mathbb{L}_{t+1}^K) || (p(\mathbb{L}_K))^\top),
\]

where \( n_1 = \binom{|C_t|}{K} \) and \((\mathbb{S}, \mathbb{Q})\) are the support set and the query set constructed by randomly sampling from category set \( L_K^t \). We can solve this problem by using the Lagrange multipliers, which yields:

\[
(p_\star(\mathbb{L}_{t+1}^K)) \propto (p(\mathbb{L}_K))^\top e^{-\frac{\alpha}{n_2} \mathbb{E}_{\mathbb{S}, \mathbb{Q}} f(\theta_t, \mathbb{S}, \mathbb{Q})}.
\]

It is impractical to compute the expectation of utility function over \( \mathbb{S} \) and \( \mathbb{Q} \) and all the possibilities of \( \mathbb{L}_K \), so we approximate the above solution by only computing the utility function on last sampled support set \( \mathbb{S}_t \) and query set \( \mathbb{Q}_t \) and updating the probability for the last sampled category set \( L_{t+1}^K \). Since \( p(\mathbb{L}_{t+1}^K) \) is proportional to the product of class-pair potentials \( \prod_{(i,j) \subset L_{t+1}^K} C_t(i,j) \). Substituting \( \bar{p}(i,j) | \mathbb{S}_t, \mathbb{Q}_t \) into the utility function, we obtain the updating rule for class-pair potentials:

\[
C_{t+1}(i,j) \leftarrow (C_t(i,j))^\top e^{\frac{1}{n_2} \bar{p}(i,j) | \mathbb{S}_t, \mathbb{Q}_t},
\]

where \( n_2 = \binom{2}{K} \). This derives the updating rule for the proposed adaptive task sampling methods (cp-sampling and gcp-sampling).

2 More Experimental Results

2.1 Evaluation on tieredImageNet Dataset

To further validate the effectiveness of gcp-sampling. We evaluate it on tieredImageNet. This dataset [8] is a larger subset of ILSVRC-12, which contains 608 classes and 779,165 images totally. As in [8], we split it into 351, 97, and 160 classes for training, validation, and test, respectively. The comparative results are shown in Table 1.

2.2 Evolution of Class-Pair Potentials

We demonstrate the evolution of class-pair potentials about 16 classes of CIFAR-FS dataset. We plot the evolving correlation matrix w.r.t. class-pair potentials in the first 600 iterations at the interval of every 40 iterations. By observing Figure 1, we can find that gcp-sampling is initialized with uniform sampling and gradually put its attention to the valuable class-pairs.
Table 1: Average 5-way, 1-shot and 5-shot classification accuracies (%) on the tieredImageNet dataset.

| Backbone    | 5way-1shot     | 5way-5shot     |
|-------------|---------------|---------------|
| Relation Network [10] | CONV-4 54.48 ± 0.93 71.32 ± 0.78 |
| PN [9]      | CONV-4 53.31 ± 0.89 72.69 ± 0.74 |
| MAML [2]    | CONV-4 51.57 ± 1.81 70.30 ± 1.75 |
| TPN [5]     | CONV-4 59.91 ± 0.94 73.30 ± 0.75 |
| TapNet [11] | ResNet-12 63.08 ± 0.15 80.26 ± 0.12 |
| PN [4]      | ResNet-12 61.74 ± 0.77 80.00 ± 0.55 |
| PN with gcp-sampling | ResNet-12 **62.80** ± 0.73 **80.52** ± 0.56 |
| MetaOptNet-RR [4] | ResNet-12 65.36 ± 0.71 81.34 ± 0.52 |
| MetaOptNet-RR with gcp-sampling | ResNet-12 **66.21** ± 0.73 **81.93** ± 0.48 |
| MetaOptNet-SVM [4] | ResNet-12 65.99 ± 0.72 81.56 ± 0.53 |
| MetaOptNet-SVM with gcp-sampling | ResNet-12 **66.92** ± 0.72 **82.10** ± 0.52 |

Fig. 1: Correlation matrix w.r.t. class-pair potentials for 16 classes of CIFAR-FS dataset. Each element indicates the class-pair potential. We plot the evolving correlation matrix of the first 600 iterations at the interval of every 40 iterations.

References

1. Catoni, O.: PAC-Bayesian supervised classification: The thermodynamics of statistical learning. institute of mathematical statistics lecture notes—monograph series 56. IMS, Beachwood, OH. MR2483528 (2007)
2. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017)
3. Guedj, B.: A primer on pac-bayesian learning. arXiv preprint arXiv:1901.05353 (2019)
4. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10657–10665 (2019)
5. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002 (2018)
6. London, B.: A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In: Advances in Neural Information Processing Systems. pp. 2931–2940 (2017)
7. McAllester, D.A.: Pac-bayesian model averaging. In: COLT. vol. 99, pp. 164–170. Citeseer (1999)
8. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. arXiv preprint arXiv:1803.00676 (2018)
9. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
10. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018)
11. Yoon, S.W., Seo, J., Moon, J.: Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. arXiv preprint arXiv:1905.06549 (2019)