Adversarial Projections to Tackle Support-Query Shifts in Few-Shot Meta-Learning

Aroof Aimen[1⋆](✉), Bharat Ladrecha[1⋆], and Narayanan C Krishnan[2]

[1] Indian Institute of Technology, Ropar
{2018csz0001,2018csb1080}@iitrpr.ac.in
[2] Indian Institute of Technology, Palakkad
ckn@iitpkd.ac.in

Abstract. Popular few-shot Meta-learning (ML) methods presume that a task's support and query data are drawn from a common distribution. Recently, Bennequin et al. [4] relaxed this assumption to propose a few-shot setting where the support and query distributions differ, with disjoint yet related meta-train and meta-test support-query shifts (SQS). We relax this assumption further to a more pragmatic SQS setting (SQS+) where the meta-test SQS is anonymous and need not be related to the meta-train SQS. The state-of-the-art solution to address SQS is transductive, requiring unlabelled meta-test query data to bridge the support and query distribution gap. In contrast, we propose a theoretically grounded inductive solution - Adversarial Query Projection (AQP) for addressing SQS+ and SQS that is applicable when unlabeled meta-test query instances are unavailable. AQP can be easily integrated into the popular ML frameworks. Exhaustive empirical investigations on benchmark datasets and their extensions, different ML approaches, and architectures establish AQP's efficacy in handling SQS+ and SQS.

Keywords: Meta-learning · Task · Support · Query · Projection · Shift.

1 Introduction

Learning Deep neural networks (DNN) from limited training data is of increasing relevance due to its ability to mitigate the challenges posed by the costly data annotation process for various real-world problems. A popular framework for learning with limited training data is few-shot learning, i.e., learning a model from few shots (examples) of data. Meta-learning (ML) approaches for few-shot learning have proven to be robust at handling data scarcity [25, 10, 28, 1]. A typical ML setup follows an episodic training regimen. An episode or a task is an N-way K-shot learning problem, where N is the number of classes in a task and K is the number of examples per class. Each task comprises of a task-train data (support set) and task-test data (query set), containing disjoint examples from the same classes. Models are adapted separately for the tasks using the support set. The adapted model's loss on the query set is used to update the meta-model.

⋆ Equal Contribution

The model meta-trained in this fashion extracts rich class discriminative features [14] that can quickly adapt to a new unseen test task.

The ML approach assumes that the meta-train and meta-test tasks are drawn from a common distribution. The shared distribution assumption prevents the use of meta-learned models in evolving test environments deviating from the training set. Recent ML works attempt at relaxing this assumption [30, 27]. However, these ML approaches assume a common distribution inside the tasks, i.e., the task-train and task-test data come from the same distribution. But a distribution shift may exist between the support and query set because of the evolving or deteriorating nature of real-world objects or environments, differences in the data acquisition techniques from support to query sets, extreme data deficiency from one distribution, etc. Addressing support query shift (SQS) inside a task has gained attention very recently [4]. However, this pioneering work assumes the prior knowledge of SQS in the meta-test set and induces a related although disjoint SQS in the meta-train set. The model trained on such a meta-train set is accustomed to handle the SQS and, to some extent, becomes robust to the related unseen meta-test SQS. In this paper, we consider, SQS+, a more generic SQS problem where the prior knowledge of the meta-test SQS is absent. We expect an unknown SQS in the meta-test set and therefore cannot induce any related SQS in the meta-train set. The earlier work on addressing SQS [4] is a limiting case of SQS+.

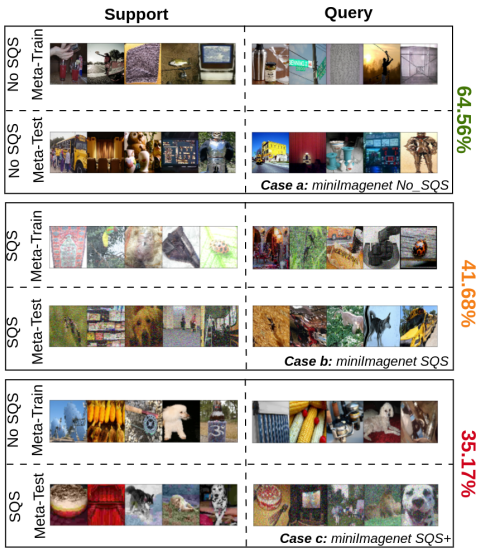


Fig. 1: Performance drop of a ProtoNet model due to SQS. **Case a)** No SQS in meta-train/test set **Case b)** Related but disjoint SQS in both meta-train/test sets **Case c)** Meta-train set lacks SQS, but meta-test set contains SQS.

We illustrate the significance of SQS+ in Figure 1 on a 5-way 5-shot problem: *Case a)* miniImagenet with No SQS [28] where both meta-train and meta-test sets do not contain SQS; *Case b)* miniImagenet with SQS [4] where meta-train and meta-test sets have related but disjoint SQS and *Case c)* miniImagenet with SQS+ (*ours*) where meta-train set lacks SQS, but meta-test set possesses SQS. The average performance of a meta-trained prototypical network (ProtoNet) [25] for the cases *(a)*, *(b)*, and *(c)* is 64.56%,

41.68%, and 35.17% respectively. The nearly 29% performance drop from *case a* to *case c* indicates that the naive ML model is vulnerable to SQS and cannot extrapolate its training experience to comparable scenarios. Bennequin et al., [4] initiated the research on SQS to address the problem specified in *case (b)*. We extend it to a more generic an challenging setting where there is no SQS during meta-training, but meta-test tasks may contain a distribution shift between the support and the query sets. The approximately 6% drop in the accuracy of the ProtoNets trained for settings *case b* and *case c* reinforces our challenging problem setting.

The solution proposed by Bennequin et al., [4] uses optimal transport (OT) to bridge the gap between support and query distributions, but assumes the availability of labeled and unlabelled query data during meta-training and testing respectively. While this solution can be adopted for our proposed problem, access to unlabelled query data during meta-test may be unrealistic in many real-world scenarios. Our solution to address the support query (SQ) shift problem - Adversarial Query Projection (AQP), does not require transduction during meta-testing. During meta-training, we induce a distribution shift between support and query sets by adversarially perturbing the query sets to create more "challenging" virtual query sets. New virtual tasks are constructed from the original support and virtual query sets. Due to the disparity between the initial and perturbed distributions, a distribution mismatch occurs between the support and query set of a virtual task. The adversarial perturbations are dynamic and adaptive, seeking to inhibit the model's learning. A model trained in such a setup performs well only if it learns to be resilient to the SQS in a task. As adversarial perturbations lack a static structure, the model is forced to learn various shift-invariant representations and thus becomes robust to various unknown distribution shifts. Overall, we make the following contributions:

- We propose, SQS+, a practical SQS setting for few-shot meta-learning. The shift between support and query sets during meta-testing is unknown while meta-training the model.
- We contribute to the FewShiftBed [4] realistic datasets for evaluating methods that address SQS and SQS+. In these datasets, meta-train data lacks SQS while meta-test data contains SQS.
- We design an inductive solution for tackling SQS+ using adversarial query projections (AQP). We theoretically justify the feasibility of meta-optimizing the model using adversarially projected query sets and verify the existence of an adversarial query projection for each query set. The AQP module is standalone and could be integrated with any few-shot ML episodic training regimen. We verify this capability by integrating AQP into ProtoNet and Matching Networks (MatchingNet).
- Exhaustive empirical investigation validates the effectiveness of the AQP on various settings and datasets, preventing a negative impact even in the absence of SQS.

2 Related Work

We segregate the discussion of the related work into approaches for cross-domain few-shot learning and tackling support-query shift in few-shot learning.

2.1 Cross-Domain Few-shot Learning (CDFSL)

Classical few-shot learning (FSL) [7, 15] does not expect distribution shifts between train and test sets. Domain generalization approaches that generate examples from a fictitious hard domain through adversarial training [29] or synthesize virtual train and test domains to simulate a shift during the training process using a critic network [17] aim to encourage generalization on unseen target domain. Typical domain generalization setting assumes abundant training data and shared labels between train and test domains, which need not hold in a cross domain FSL setup. The early approaches to bridge the domain shifts in FSL relied on adaptive batch normalization [9] and batch spectral normalization [19]. Recent work [4] suggests limitations of batch normalization as a strategy for handling SQS. A common hypothesis among cross domain FSL approaches is that a model's over-reliance on the meta-train domain inhibits its generalizability to unseen test domains. While some cross domain FSL approaches relied on model's generalizability by enhancing diversity in the feature representations [27, 26], others have tried ensembles [20], large margin enforcement [31], and adversarial perturbations [30]. Though these approaches handle domain discrepancy between meta-train and meta-test sets, they assume a common distribution over support and query sets. On the other hand, we focus on the scenarios where support and query distributions vary.

2.2 Support-Query Shift in FSL

Transductive meta-learning approaches that utilize unlabeled query data in the training process are effective baselines for handling SQS in FSL. Ren et al., [23] introduce a transductive prototypical network that refines the learned prototypes with cluster assignments of unlabelled query examples. Boudiaf et al. [6] induce transduction by maximizing the mutual information between query features and their predicted labels in conjunction with minimizing cross-entropy loss on the support set. Minimizing the entropy of the unlabeled query instance predictions during adaptation [8] also achieves a similar goal. Liu et al., [21] propose a graph based label propagation from the support to the unlabeled query set that exploits the data manifold properties to improve the efficiency of adaptation . Antoniou et al., [2] show that minimizing a parameterized label-free loss function that utilizes unlabelled query data during training can also bridge SQS. Inspired from learning invariant representations [12, 3, 11], Bennequin et al. [4] use Optimal Transport (OT) [22] during meta-training and meta-testing to address SQS. In contrast, we propose an inductive method to tackle SQS in few-shot meta-learning where access to the unlabelled meta-test query instances is not required.

Inductive approaches to tackle train-test domain shifts have relied on adversarial methods for data/task augmentations. Goldblum et al., [13] propose adversarial data augmentation for few-shot meta-learning and demonstrate the robustness of the model trained on augmented tasks to adversarial attacks at meta-test time. Wang et al. [30] bridge the shift between meta-train and meta-test domains by adversarial augmentation by constructing virtual tasks learned through adversarial perturbations. A model trained on such virtual tasks becomes resilient to meta-train and meta-test domain shifts. While adversarial perturbations are central to our approach, we use it to tackle a different problem, support query distribution shifts inside a task for few-shot meta-learning.

3 Methodology

3.1 Preliminaries

Notations A typical ML setup has three phases - meta-train M, meta-validation M_v and meta-test M_t. A model is trained on M and evaluated on M_t. M_v is used for hyperparameter tuning and model selection. The dataset $(C, \mathcal{D})$ comprising of classes and domains is partitioned into $(C_M, \mathcal{D}_M)$, $(C_{M_v}, \mathcal{D}_{M_v})$, and $(C_{M_t}, \mathcal{D}_{M_t})$ corresponding to the phases M, M_v and M_t, respectively. Each phase is a collection of tasks and every task T_0 is composed of a support set T_{S_0} and a query set T_{Q_0}. The support set $T_{S_0} = \{\{x_k^c, y_k^c\}_{k=1}^K\}_{c=1}^N$ and query set $T_{Q_0} = \{\{x_q^{*c}, y_q^{*c}\}_{q=1}^Q\}_{c=1}^N$ contain (example x, label y) pairs from N-classes with K and Q examples per class, with the label of meta-test query instances being used only for evaluation.

The classical few-shot learning setup does not consider diverse domains. The tasks are sampled from a common distribution $\mathcal{T}_0$. A model meta-trained on tasks sampled from $\mathcal{T}_0$ learns representations that extend to the disjoint meta-test tasks from the same distribution. Given a task T_0 (support-query pair $\{T_{S_0}, T_{Q_0}\}$), few-shot learning learns a classifier ϕ using T_{S_0}, which correctly categorizes instances of T_{Q_0}. A model parameterized by θ is meta-trained on a collection of tasks sampled from $\mathcal{T}_0$ using a bi-level optimization procedure. First, θ is adapted on the tasks' support set T_{S_0} to obtain ϕ. Then ϕ is evaluated on the query set T_{Q_0} to estimate query loss L^*, which is used to update θ. The model is meta-trained according to the objective $\min_{\theta \in \Theta} \mathbb{E}_{T_{Q_0}} [L^*(\phi, T_{Q_0})]$, where $\phi \leftarrow \theta - \alpha \nabla_\theta L(\theta; T_{S_0})$; L and L^* are the losses of the model on the support and query sets respectively. Note that ML approaches such as ProtoNet [16] and MatchingNet [28] do not require adaptation, and hence $\theta = \phi$.

Support-Query Distribution Shift In a classical few-shot learning setup, the domain is constant across M, M_v, M_t phases and within the tasks. So, in addition to a common distribution $\mathcal{T}_0$ over tasks, a shared distribution exists even at the task composition level, i.e., $\mathcal{T}_{S_0} = \mathcal{T}_{Q_0}$, where $\mathcal{T}_{S_0}$ and $\mathcal{T}_{Q_0}$ are the distributions on support and query sets respectively. A more pragmatic case is that of SQS, wherein a distribution mismatch occurs between the support and

query sets within a task. Let $\mathcal{D}_M$ and $\mathcal{D}_{M_t}$ be the set of domains for the M and M_t phases. We skip M_v for convenience, but it follows the same characteristics as M and M_t. We define our version of the support query shift problem termed SQS+ illustrated in the Figure 1 (*case c*) as follows.

Definition 1. *(SQS+) The support and query sets of every meta-train task come from the domain $\mathcal{D}_M$ and share a common distribution $\mathcal{T}_{S_0} = \mathcal{T}_{Q_0}$. Let $D_S^{M_t}, D_Q^{M_t} \in \mathcal{D}_{M_t}$ be the support and query domains for a meta-test task. The SQS+ setting is characterized by an unknown shift in the support and query domains of a meta-test task, $D_S^{M_t} \neq D_Q^{M_t}$ (introducing a shift in the support and query distributions $\mathcal{T}_{S_0} \neq \mathcal{T}_{Q_0}$), along with the standard SQS assumption of disjoint meta-train and meta-test domains - $\mathcal{D}_M \cap \mathcal{D}_{M_t} = \emptyset$.*

Bennequin et al. [4] identified the SQS problem, but assumed only a similar but disjoint SQS in the meta-train and meta-test datasets. A model learned on such a meta-train set is compelled to extract shift-invariant features during adaptation on the support set to reduce L^* on query sets. Although $\mathcal{D}_M$ and $\mathcal{D}_{M_t}$ are disjoint, they share a latent structure that facilitates learning of shift-invariant features on $\mathcal{D}_M$ that can be extended to $\mathcal{D}_{M_t}$. This makes the learned model impervious to SQS in the meta-test set. SQS+, on the other hand, is more general and challenging. We neither anticipate the occurrence of SQS in the meta-test set nor maintain a common structure between the meta-train and meta-test SQS's. Relaxing the shared structure constraint between $\mathcal{D}_M$ and $\mathcal{D}_{M_t}$ removes the need for prior access to the meta-test set (consequently its domains) to imbibe SQS in meta-train tasks. Hence, we tackle a more challenging problem of learning a resilient model for an unknown meta-test SQS.

A model trained using the classical ML objective has not encountered support and query set shifts during meta-training. Thus the learned representations are not shift-invariant, due to which the model does not generalize to the unknown meta-test SQS. Bennequin et al.,'s [4] transductive optimal transport (OT)-based solution to bridge the gap between the support and query shifts could also be adopted SQS+. However, the solution needs access to unlabeled query sets during meta-training and meta-testing, which is unavailable in our setting. We propose an inductive adversarial query projection (AQP) strategy to address SQS+ that can also work in the vanilla SQS setting.

3.2 Adversarial Query Projection (AQP)

Without leveraging unlabelled meta-test query instances, our solution induces the hardest distribution shift for the meta-model's current state. For a task T_0, we simulate the worst distribution shift by adversarially perturbing its query set T_{Q_0} such that the model's query loss L^* maximizes. Let H be the task composition space, i.e., H is the distribution of support and query distributions such that $\mathcal{T}_{Q_0} \sim H$ and $\mathcal{T}_Q \sim H$. Let T_{Q_0} and T_Q be the samples belonging to $\mathcal{T}_{Q_0}$ and $\mathcal{T}_Q$ respectively (we occasionally denote $T_Q \sim H$ because $T_Q \sim \mathcal{T}_Q \sim H$, to improve readability). Also, let Θ be the parameter space with $\theta, \phi \sim \Theta$, and

$d : H \times H \rightarrow R_+$ be the distance metric that satisfies $d(T_{Q_0}, T_{Q_0}) = 0$ and $d(T_Q, T_{Q_0}) \geq 0$. We consider a Wasserstein ball B centered at $\mathcal{T}_{Q_0}$ with radius ρ denoted by $B_\rho(\mathcal{T}_{Q_0})$ such that:

$$B_\rho(\mathcal{T}_{Q_0}) = \{\mathcal{T}_Q \in H : W_d(\mathcal{T}_Q, \mathcal{T}_{Q_0}) \leq \rho\}$$

where $W_d(\mathcal{T}_Q, \mathcal{T}_{Q_0}) = \inf_{M \in \pi(\mathcal{T}_Q, \mathcal{T}_{Q_0})} \mathbb{E}_M \left[d(T_Q, T_{Q_0})\right]$ is the Wasserstein distance that measures the minimum transportation cost required to transform $\mathcal{T}_{Q_0}$ to $\mathcal{T}_Q$, and $\pi(\mathcal{T}_Q, \mathcal{T}_{Q_0})$ denotes all joint distributions for $(\mathcal{T}_Q, \mathcal{T}_{Q_0})$ with marginals $\mathcal{T}_Q$ and $\mathcal{T}_{Q_0}$.

AQP aims to find the most challenging query distribution $\mathcal{T}_Q$ for an original query distribution $\mathcal{T}_{Q_0}$ that lies within or on the Wasserstein ball $B_\rho(\mathcal{T}_{Q_0})$. The hardest perturbation to the query distribution $\mathcal{T}_{Q_0}$ is the one that maximizes the model's query loss L^*. Updating the model using such difficult query distribution $\mathcal{T}_Q$ improves its generalizability. Further, the transformation of $\mathcal{T}_{Q_0}$ into $\mathcal{T}_Q$ induces a distributional disparity in a new virtual task comprising of the original support set from $\mathcal{T}_{S_0}$ and the projected query set from $\mathcal{T}_Q$. A model adapted to such virtual tasks is compelled to extract the shift-invariant representations from $T_{S_0} \sim \mathcal{T}_{S_0}$ transferable to $T_Q \sim \mathcal{T}_Q$ to reduce the query loss L^*. As adversarial perturbations are adaptive to the model's state, they do not have a monotonic structure throughout the meta-training phase. The evolving augmentations expose the model to diverse SQS. A model meta-trained on such virtual tasks with different SQ shifts learns to extract diverse shift-invariant representations increasing the model's endurance to unknown meta-test SQS. The simultaneous restrain of $\mathcal{T}_Q$ to a Wasserstein ball radius ρ ensures $\mathcal{T}_Q$ does not deviate extensively from $\mathcal{T}_{Q_0}$, and $\mathcal{T}_Q, \mathcal{T}_{Q_0}$ share the label space, and $\mathcal{T}_{Q_0}, \mathcal{T}_Q \in H$ is maintained. Thus the newly-framed meta-objective is:

$$\min_{\theta \in \Theta} \sup_{W_d(\mathcal{T}_Q, \mathcal{T}_{Q_0}) \leq \rho} \mathbb{E}_{(T_Q \sim\ \mathcal{T}_Q)} \left[L^*(\phi, T_Q)\right] \tag{1}$$

where $\phi \leftarrow \theta - \alpha \nabla_\theta L(T_{S_0}; \theta)$. As equation 1 is intractable for an arbitrary ρ, we aim to convert this constrained optimization problem to an unconstrained optimization problem for a fixed penalty parameter $\gamma \geq 0$ as given below:

$$\min_{\theta \in \Theta} \sup_{\mathcal{T}_Q} \left\{\mathbb{E}_{\mathcal{T}_Q}[L^*(\phi, T_Q)] - \gamma W_d(\mathcal{T}_Q, \mathcal{T}_{Q_0})\right\} \tag{2}$$

We first show that the unconstrained objective is strongly concave and then define a shift robust surrogate, $\psi_\gamma(\phi, T_{Q_0})$, that is easy to optimize.

Theorem 1. *For the loss function $L^*(\phi, T_Q)$ smooth in T_Q, a distance metric $d : H \times H \rightarrow R_+$ convex in T_Q and a large penalty γ (by duality small ρ), the function $L^*(\phi; T_Q) - \gamma d(T_Q, T_{Q_0})$ is $\gamma - \mathcal{L}$ strongly concave for $\gamma \geq \mathcal{L}$.*

Proof. Deferred to the supplementary material.

We next define a robust surrogate inspired from Sinha et al., [24] for this unconstrained objective that is the dual of the minimax problem in equation 1.

Theorem 2. *Let* $L^* : \Theta \times H \rightarrow R$ *and* $d : H \times H \rightarrow R_+$ *be continuous. Let* $\psi_\gamma(\phi; T_{Q_0}) = \sup_{T_Q \in H} \{L^*(\phi, T_Q) - \gamma d(T_Q, T_{Q_0})\}$ *be a shift robust surrogate. For any query set distribution* $\mathcal{T}_Q$ *and any* $\rho > 0$,

$$\sup_{\mathcal{T}_Q : W_d(\mathcal{T}_Q, \mathcal{T}_{Q_0}) \leq \rho} \mathbb{E}_{T_Q \sim \mathcal{T}_Q} [L^*(\phi, T_Q)] = \inf_{\gamma \geq 0} \left\{ \gamma\rho + \mathbb{E}_{\mathcal{T}_{Q_0}} [\psi_\gamma(\phi; T_{Q_0})] \right\}$$

and for any $\gamma \geq 0$,

$$\sup_{\mathcal{T}_Q} \left\{ \mathbb{E}_{\mathcal{T}_Q} [L^*(\phi, T_Q)] - \gamma W_d(\mathcal{T}_Q, \mathcal{T}_{Q_0}) \right\} = \mathbb{E}_{\mathcal{T}_{Q_0}} [\psi_\gamma(\phi; T_{Q_0})]$$

Using Theorem 2, we arrive at the following surrogate meta-objective:

$$\min_{\theta \in \Theta} \left\{ \mathbb{E}_{\mathcal{T}_{Q_0}} [\psi_\gamma(\phi; T_{Q_0})] \right\} \tag{3}$$

Thus, meta-optimizing the robust surrogate involves maximizing the loss L^* on adversarial query projections T_Q while simultaneously restraining T_Q to a ρ distance from T_{Q_0}. We now show the existence of the adversarial projection for an original query set using the results from [30, 5].

Theorem 3. *Let* $L^* : \Theta \times H \rightarrow R$ *be* $\mathcal{L}$*-Lipshitz smooth and* $d(., T_{Q_0})$ *be a* μ*-strongly convex for each* $T_{Q_0} \in H$*. If* $\gamma > \dfrac{\mathcal{L}}{\mu}$ *then there exists a unique* $\hat{T}_Q$ *satisfying*

$$\hat{T}_Q = arg\, \underset{T_Q \in H}{Sup} \{L^*(\phi, T_Q) - \gamma d(T_Q, T_{Q_0})\} \tag{4}$$

and

$$\nabla_\theta \psi_\gamma(\phi, T_{Q_0}) = \nabla_\theta L^*(\theta; \hat{T}_Q) \tag{5}$$

Proof. Deferred to the supplementary material.

Remark 1. $L^*(\phi, T_Q) - \gamma d(T_Q, T_{Q_0})$ is a $\gamma - \mathcal{L}/\mu$ strongly concave function for $\gamma \geq \mathcal{L}/\mu$ and so $L^*(\phi, T_Q) - \gamma d(T_Q, T_{Q_0})$ admits one and only one unique maximizer $\hat{T}_Q$ ($\mu = 1$ for Euclidean distance).

Estimation of AQP To find the adversarial query projection, we approximate equation 4 by employing gradient ascent with early stopping on the query set. We consider a task $T_0 = \{T_{S_0} \cup T_{Q_0}\}$ and let $\{X, Y\}$ and $\{X^*, Y^*\}$ be the set of all instance-label pairs in T_{S_0} and T_{Q_0}, respectively. We propose algorithm 1 to induce SQS in the meta-train tasks. The original query instances X^* initialize the worst-case query augmentations X_w^*. We perform an iterative gradient ascent on X^* using L^*, resulting in an augmented query set X_w^*. This augmented query set X_w^* has distributional disparity with original support set X. Early stopping by *Adv_iter* and initializing X_w^* with X^* regularizes ($-\gamma d(T_Q, T_{Q_0})$) and ensures X_w^* does not deviate extensively from X^*. The algorithm returns a virtual task with original support X and projected query X_w^*, which is used to update θ.

Algorithm 1: Adversarial Query Projection $AQP(T_{S_0}, T_{Q_0})$

Input: Task Support and Query Sets - ($T_{S_0} = \{X, Y\}$, $T_{Q_0} = \{X^*, Y^*\}$), model parameters ϕ
$X^*_w \leftarrow X^*$
for $i = 0$ *to* *Adv_iter* **do**
  $X^*_w \longleftarrow X^*_w + \eta \nabla_{X^*_w} L^*_i(\phi, X^*_w)$
end
$T_Q = \{X^*_w, Y^*\}$
return $(T_{S_0} \cup T_Q)$

4 Experiments and Results

We design experiments to investigate the challenging nature of our proposed SQS+ benchmark and empirically validate the efficacy of the proposed AQP over the state-of-the-art approach [4] to address SQS in inductive settings. We consider Cifar 100, miniImagenet, tieredImagenet, FEMNIST, and their state-of-the-art SQS variants for evaluation. We also demonstrate the AQP's efficiency on our proposed datasets (introduced in section 4.1 and elaborated in the supplementary material). We used Conv4 models [4] for Cifar 100, FEMNIST and their variants, and ResNet-18 [16] for miniImagenet, tieredImagenet, and their extensions. We use 32×32 images for Cifar 100, 28×28 for FEMNIST, and 84×84 for miniImagenet and tieredImagenet. We next present the implementation details, followed by the contributions to FewShiftBed and empirical investigations.

4.1 Implementation Details

Following [4], we fix the meta-learning rate as 0.001 for all approaches (Ind_OT, AQP), models (ProtoNet, MatchingNet), and datasets (Cifar 100, miniImagenet, tieredImagenet, FEMNIST, and their variants) and learn the models for 60,000 episodes. We perform a grid search using Ray over 35 configurations for 12000 episodes to find the optimal hyper-parameters. The search space is shared for all approaches, datasets, and models. The hyper-parameters of regularization in Ind_OT and AQP's adversarial learning rate are sampled from log uniform distribution in the ranges [15, 50] and [0.001, 1.0], respectively. Further, the number of iterations required to project data in AQP and Ind_OT iterations are randomly sampled from ranges [2,9] and [500, 1500] (increments of 100), respectively. However, for Ind_OT, we obtained better results on default hyper-parameters than tuned ones on miniImagenet and its SQS variants. So we fixed its parameters as mentioned in [4] for all the cases. We report the hyper-parameters (learning rate (η) and number of iterations (*Adv_iter*)) for AQP in table 1.

4.2 Contributions to FewShiftBed

We make significant contributions to the FewShiftBed [4]. Firstly, we have created challenging datasets wherein SQS is present only at meta-test time (SQS+).

Table 1: Hyperparameter details of AQP for different datasets and approaches.

Dataset	**ProtoNet**						**MatchingNet**					
	No SQS		SQS		SQS+		No SQS		SQS		SQS+	
	η	*Adv_iter*	η	*Adv_iter*	η	*Adv_iter*	η	*Adv_iter*	η	*Adv_iter*	η	*Adv_iter*
Cifar 100	22.0	4	31.0	3	22.0	4	22.0	4	31.0	3	32.0	2
miniImagenet	22.0	4	31.0	3	22.0	4	22.0	4	41.0	8	24.5	5
tieredImagenet	22.0	4	17.0	4	22.0	4	22.0	4	41.0	9	22.0	4
FEMNIST	22.0	4	16.5	2	24.0	2	22.0	4	25.8	5	30.0	8

The SQS+ versions of Cifar 100, miniImagenet, and tieredImagenet datasets are constructed from their SQS counterparts [4] by removing perturbations from the meta-train datasets. Similarly, the SQS+ variant of FEMNIST also follows its SQS counterpart, but the meta-train set contains alpha-numerals from users randomly. We add these SQS+ versions of benchmark datasets to the testbed. The perturbations applied to the tasks are entirely modular, i.e., a task may have augmentation in support, query, both, or none. More details about the datasets are available in the supplementary material. Secondly, we integrate our theoretically grounded inductive solution, Adversarial Query Projections (AQP), into the testbed. The AQP implementation is standalone and can be integrated with any episodic training regimen. We have successfully integrated AQP with ML approaches like Prototypical and Matching networks [25, 28]. Thirdly, we have also added a hyperparameter optimization module that uses RAY [18] for tuning parameters. We believe these additions improve the usability and coverage of FewShiftBed to study SQS. The modified FewShiftBed, which includes the proposed solution, datasets, and experimental setup, is publicly available. [3]

4.3 Evaluation of SQS+

We first validate that SQS+ is more challenging than the SQS problem [4]. We train Prototypical and Matching networks on Cifar 100, miniImagenet, tieredImagenet, and FEMNIST on all three settings - No SQS, SQS, and SQS+. We report the results in Table 2 and observe that for all the datasets, models trained with both the approaches (Prototypical and Matching network) perform best in the No SQS setting, followed by SQS and SQS+. In the classical few-shot setting, meta-train and meta-test phases share the domain, due to which the meta-knowledge is easily transferable across the phases. However, in SQS, each task's support and query set represent different domains, but share a latent structure, during the meta-train and meta-test phases. In SQS versions of Cifar 100, miniImagenet, and tieredImagenet, both meta-train and meta-test SQS are characterized by different types of data perturbations. However, in FEMNIST's SQS variant, meta-train and meta-test SQS is induced due to different writers. A meta-model trained in this setup becomes partially resilient to the related but disjoint SQS during meta-testing. A common SQS structure across meta-train

[3] https://github.com/Few-Shot-SQS/adversarial-query-projection

Table 2: Comparison of ML methods with their Ind_OT and AQP counterparts across Cifar 100, miniImagenet, tieredImagenet, FEMNIST datasets, and their SQS and SQS+ variants. The results are obtained on 5-way tasks with 5 support and 8 query instances per class except for FEMNIST and its variants, which contains only one support and one query instance per class. The ± represents the 95% confidence intervals over 2000 tasks. AQP outperforms classic, and Ind_OT-based ML approaches approximately on all datasets.

Method	**Test Accuracy**					
	No SQS	SQS	SQS+	No SQS	SQS	SQS+
	Cifar 100			miniImagenet		
ProtoNeT	48.07 ± 0.44	43.15 ± 0.48	40.59 ± 0.69	64.56 ± 0.42	41.68 ± 0.76	35.17 ± 0.78
Ind_OT+ ProtoNeT	48.62 ± 0.44	43.62 ± 0.49	41.74 ± 0.65	63.74 ± 0.42	39.84 ± 0.78	34.75 ± 0.80
AQP+ ProtoNeT	**48.70 ± 0.42**	**45.09 ± 0.46**	**45.06 ± 0.46**	**66.81 ± 0.42**	**42.65 ± 0.57**	**40.61 ±0.60**
MatchingNet	46.03 ± 0.42	39.89 ± 0.44	36.63 ± 0.45	59.68 ± 0.43	39.66± 0.54	35.40 ±0.52
Ind_OT+ MatchingNet	45.77 ± 0.42	40.82 ± 0.45	37.13 ± 0.47	59.64 ± 0.44	38.25± 0.54	33.22± 0.50
AQP+ MatchingNet	**46.53 ± 0.43**	**42.40 ± 0.46**	**41.26 ± 0.46**	**62.29 ± 0.42**	**42.32 ± 0.52**	**37.90 ± 0.53**
	tieredImagenet			FEMNIST		
ProtoNeT	**71.04 ± 0.45**	41.59 ± 0.57	38.57 ± 0.65	93.09 ± 0.51	84.36 ± 0.74	82.67 ± 0.77
Ind_OT+ ProtoNeT	69.56 ± 0.46	40.08 ± 0.56	35.81 ± 0.58	91.66 ± 0.55	79.64 ± 0.80	76.37 ± 0.84
AQP+ ProtoNeT	69.62 ± 0.45	**45.34 ± 0.60**	**40.94 ± 0.66**	**94.61 ± 0.45**	**85.92 ± 0.69**	**84.42 ± 0.74**
MatchingNet	67.85 ± 0.46	43.30 ± 0.56	37.57 ± 0.57	93.69 ± 0.49	85.88 ± 0.69	83.48 ± 0.74
Ind_OT+ MatchingNet	67.79 ± 0.46	44.27 ± 0.56	39.24 ± 0.59	**93.76 ± 0.48**	84.08 ± 0.71	83.09 ± 0.74
AQP+ MatchingNet	**68.40 ± 0.45**	**45.26 ± 0.56**	**39.39 ± 0.58**	93.69 +- 0.49	**87.24 ± 0.67**	**84.98 ± 0.72**

and meta-test sets may not exist. Thus, SQS+ datasets are more challenging, which is empirically validated by the baseline approach's poor performance.

4.4 Evaluation of AQP

We compare the efficiency of the proposed AQP and optimal transport (OT) based state-of-the-art solution for handling vanilla SQS and SQS+ on the benchmark datasets. A strong baseline for SQS+ is the inductive version of OT (Ind_OT), where we employ OT only in the meta-train phase to generate projected support sets using support and query instances of a task. We evaluate ProtoNet and Matching Networks versions of Ind_OT and AQP. Table 2 presents the results for this evaluation. We observe that the models learned on projected support data obtained by Ind_OT are less robust to both SQS and SQS+ than

the models learned on AQP for all approaches and datasets. Hence, AQP is better at addressing SQS+ (and SQS), when meta-test unlabeled query instances are unavailable.

To inspect whether the proposed AQP negatively impacts the models' generalization in the absence of meta-test SQS, we evaluate the ML approaches and their Ind_OT and AQP counterparts on classic datasets containing no support query shifts (No SQS). We observe from Table 2 that AQP does not lead to degradation in the performance in the absence of SQS, instead improves the generalizability of the model even when SQS is absent. We note that Ind_OT sometimes deteriorates the model's performance when SQS is missing. AQP outperforms both classic methods and their Ind_OT versions in almost all cases.

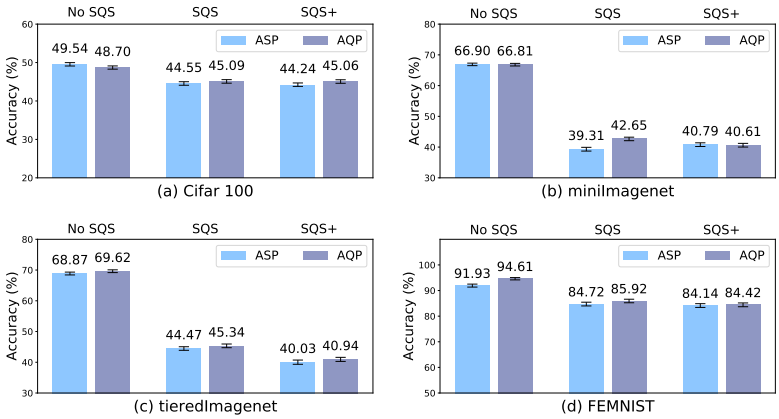


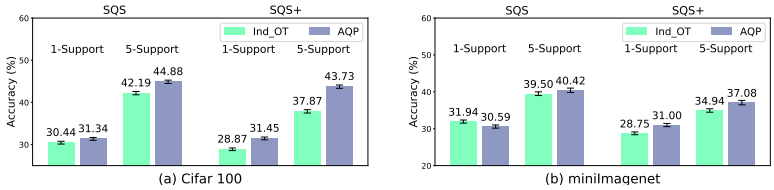
Fig. 2: Impact of adversarially projecting support and query data in a task on the model's performance across No SQS and SQS and SQS+ variants of Cifar 100, miniImagenet, tieredImagenet, and FEMNIST datasets.

Following [4], we used a Conv4 backbone for Cifar 100, FEMNIST and their transformations, and a ResNet-18 [16] backbone for miniImagenet, tieredImagenet, and their variants. Thus, Table 2 not only shows the robustness of a model trained via AQP on different SQ shifts but also its thoroughness across architectures. We randomly projecting 25% of the tasks with AQP to reduce the computational cost. Extending this idea to Ind_OT, resulted in a significant decline in the performance. We thus maintain the standard-setting [4] for Ind_OT, wherein support sets of all the tasks are projected.

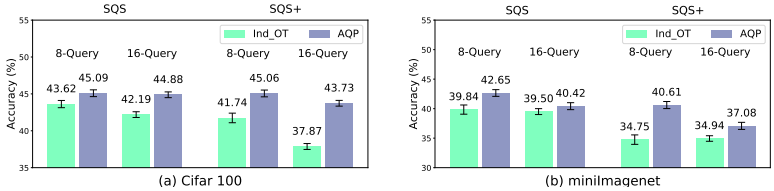
4.5 Ablations

We perform ablations to investigate the sensitivity of the proposed approach to task characteristics (varying number of support and query shots) and design choices (support vs query projections).

Ablation on Projections We study the impact of adversarially perturbing support *vs* query set in a task and evaluate the model's (ProtoNet) performance across all settings and datasets. From Figure 2 we observe perturbing query sets is empirically more meritorious in 9 out of 12 settings. We measure the model's generalizability from in-distribution support to out-of distribution (OOD) query set in a task by perturbing a query set. The magnitude of loss and hence gradients on the OOD query set is high, resulting in more meaningful meta-updates. As performance on the query set directly impacts the meta-update, the model's invariance to SQS is directly reflected in the meta-update. On the other hand, projecting support sets creates potent prototypes (robust adaptation) as adversarial projections distort the images. However, the meta-update may not be impactful due to the model's good performance on clean query images.



(a) Ablation on support instances.



(b) Ablation on query instances.

Fig. 3: Ablation on the number of support and query instances per class on SQS and SQS+ variants of Cifar 100 and miniImagenet datasets. In (a), we consider 5-way tasks with 1 and 5 support instances with 16 query instances. In (b), we vary query instances between 8 and 16 with 5 support instances per class.

Ablation on Support and Query Shots We ablate the number of shots per class in the support and query sets, limited to Cifar 100 and miniImagenet

datasets, to inspect the efficacy of our proposed AQP employing a ProtoNet. AQP outperforms Ind_OT when the number of query instances are fixed to 16 per class, and support shots per class vary from 1 to 5 (Figure 3a). We also vary the number of query instances per class from 8 to 16 and observe that AQP surpasses Ind_OT with varying query instances (Figure 3b).

4.6 Visual Analysis of AQP

We visualize the impact of AQP on the query instances across meta-training iterations. We train a Prototypical network in a 5-way 5-shot setting on the SQS+ version of miniImagenet for 150 epochs. Extended results on No SQS and SQS versions of miniImagenet are presented in the supplementary material. For better illustration, we fix one task and one instance per class and show the transformation in the query images over meta-train iterations (Figure 4). The images

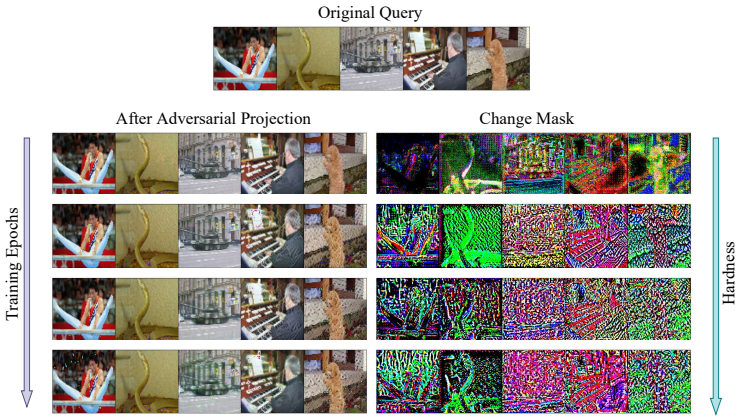


Fig. 4: Evolution of Adversarial Query Projections across training epochs for SQS+ version of miniImagenet.

in the top row are the original query set, the left column are the query images impacted by AQP with increasing iterations, and the right column represents the change mask (in the increasing order of iterations), which is the difference between the pixel intensities of the original image and its adversarially perturbed counterpart. We observe gradual increase in the distortions with increasing iterations. This in turn makes the model robust to query instances' degradation and thus to the distribution shifts between support and query. As AQP is adaptive and seeks to inhibit the model's learning, it increases the degradation in the query images to maximize the query loss with increasing iterations. This shows that following an easy to hard curriculum to distort the query contributes to AQP's success.

However, this experiment also reflects the potential limitations of the proposed AQP. We evaluated AQP in the cases where SQS is characterized by the perturbations in data (SQS variants of Cifar 100, miniImagenet, and tieredImagenet), and for a small-realistic dataset (FEMNIST and its variants) where different writers characterize SQS. The masks (Figure 4) reflect that AQP adds varying noise to distort the images, which may not resemble complex SQ shifts. Investigating AQP in more complex SQ shifts, e.g., real to sketch or caricature pictures, is part of our future work.

5 Conclusion and Future Directions

This paper proposes SQS+ - a more challenging distribution shift between the support and query sets of a task in a few-shot meta-learning setup. SQS+ includes an unknown SQ shift in the meta-test tasks, and empirical evidence suggests SQS+ is a complex problem than the prevalent SQS notion. We propose a theoretically grounded solution - Adversarial Query Projection (AQP) to address SQS+ without leveraging unlabelled meta-test query instances. Exhaustive experiments involving AQP on multiple benchmark datasets (Cifar 100, miniImagenet, tieredImagenet, and FEMNIST - their SQS and proposed SQS+ variants), different architectures, and ML approaches demonstrate its effectiveness. The future work lies in verifying the effectiveness of AQP in complex SQ shifts, e.g., shift from real to sketch images and creating datasets corresponding to these difficult SQ shifts, and integrating AQP with gradient and transductive ML approaches. We incorporate proposed AQP and SQS+ versions of Cifar 100, miniImagenet, tieredImagenet, and FEMNIST to FewShiftBed and make it publicly available to encourage research in this direction.

Acknowledgements The resources provided by 'PARAM Shivay Facility' under the National Supercomputing Mission, Government of India at the Indian Institute of Technology, Varanasi are gratefully acknowledged.

References

1. Aimen, A., Sidheekh, S., Ladrecha, B., Krishnan, N.C.: Task attended meta-learning for few-shot learning. In: Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems (2021)
2. Antoniou, A., Storkey, A.J.: Learning to learn by self-critique. Advances in Neural Information Processing Systems (2019)
3. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. Advances in Neural Information Processing Systems (2006)
4. Bennequin, E., Bouvier, V., Tami, M., Toubhans, A., Hudelot, C.: Bridging few-shot learning and adaptation: New challenges of support-query shift. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 554–569 (2021)
5. Bonnans, J.F., Shapiro, A.: Perturbation analysis of optimization problems. Springer Science & Business Media (2013)

6. Boudiaf, M., Ziko, I., Rony, J., Dolz, J., Piantanida, P., Ben Ayed, I.: Information maximization for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 2445–2457 (2020)
7. Chen, W., Liu, Y., Kira, Z., Wang, Y.F., Huang, J.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
8. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. In: International Conference on Learning Representations (2020)
9. Du, Y., Zhen, X., Shao, L., Snoek, C.G.: Metanorm: Learning to normalize few-shot batches across domains. In: International Conference on Learning Representations (2020)
10. Finn, C., Xu, K., Levine, S.: Probabilistic model-agnostic meta-learning. In: Advances in Neural Information Processing Systems (2018)
11. Flamary, R., Courty, N., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016)
12. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. pp. 1180–1189 (2015)
13. Goldblum, M., Fowl, L., Goldstein, T.: Adversarially robust few-shot learning: A meta-learning approach. In: Advances in Neural Information Processing Systems (2020)
14. Goldblum, M., Reich, S., Fowl, L., Ni, R., Cherepanova, V., Goldstein, T.: Unraveling meta-learning: Understanding feature representations for few-shot tasks. In: International Conference on Machine Learning. pp. 3607–3616 (2020)
15. Guo, Y., Codella, N., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: European Conference on Computer Vision. pp. 124–141 (2020)
16. Laenen, S., Bertinetto, L.: On episodes, prototypical networks, and few-shot learning. Advances in Neural Information Processing Systems (2021)
17. Li, Y., Yang, Y., Zhou, W., Hospedales, T.M.: Feature-critic networks for heterogeneous domain generalization. In: International Conference on Machine Learning. pp. 3915–3924 (2019)
18. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: A research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118 (2018)
19. Liu, B., Zhao, Z., Li, Z., Jiang, J., Guo, Y., Ye, J.: Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. arXiv preprint arXiv:2005.08463 (2020)
20. Liu, B., Zhao, Z., Li, Z., Jiang, J., Guo, Y., Ye, J.: Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. arXiv preprint arXiv:2005.08463 (2020)
21. 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. In: International Conference on Learning Representations (2019)
22. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning pp. 355–607 (2019)
23. 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. In: International Conference on Learning Representations (2018)

24. Sinha, A., Namkoong, H., Duchi, J.C.: Certifying some distributional robustness with principled adversarial training. In: International Conference on Learning Representations (2018)
25. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems (2017)
26. Sun, J., Lapuschkin, S., Samek, W., Zhao, Y., Cheung, N., Binder, A.: Explanation-guided training for cross-domain few-shot classification. In: International Conference on Pattern Recognition. pp. 7609–7616 (2020)
27. Tseng, H., Lee, H., Huang, J., Yang, M.: Cross-domain few-shot classification via learned feature-wise transformation. In: International Conference on Learning Representations (2020)
28. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems (2016)
29. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: Advances in Neural Information Processing Systems. pp. 5339–5349 (2018)
30. Wang, H., Deng, Z.: Cross-domain few-shot classification via adversarial task augmentation. In: International Joint Conference on Artificial Intelligence. pp. 1075–1081 (2021)
31. Yeh, J.F., Lee, H.Y., Tsai, B.C., Chen, Y.R., Huang, P.C., Hsu, W.H.: Large margin mechanism and pseudo query set on cross-domain few-shot learning. arXiv preprint arXiv:2005.09218 (2020)