MetaBayes: A Meta-Learning Framework from a Bayesian Perspective

Tamara AlShammari, Anis Elgabli, and Mehdi Bennis
Centre for Wireless Communication, University of Oulu, Finland
Emails: {Tamara.Alshammari, Anis.Elgabli, mehdi.bennis}@oulu.fi

2021 55th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5828-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53345.2021.9723290

***Abstract*—Meta-learning is a powerful learning paradigm in which solving a new task can benefit from similar tasks for faster adaption (few shot learning). Stochastic gradient descent (SGD) based meta learning has emerged as an attractive solution in the few-shot learning. However, this approach suffers from significant computational complexity due to the double loop and matrix inversion operations which incurs a significant amount of uncertainty and poor generalization. To achieve lower complexity and better generalization, in this paper, we propose *MetaBayes*, a novel framework that views the original meta learning problem from a Bayesian perspective where the meta-model is cast as the prior distribution and the task-specific models are viewed as task-specific posterior distributions. The objective amounts to jointly optimizing the prior and the posterior distributions. With this, we obtain a closed-form expression to update the distributions at every iteration, to avoid the high computation cost issue of SGD based meta learning, and produce a more robust and generalized meta-model. Our simulations show that tasks with few training samples achieves higher accuracy when MetaBayes prior distribution is used as an initializer compared to the commonly-used Gaussian prior distribution.**

I. Introduction.

Humans are capable of inferring new information from very limited samples [1] owing to the innate ability of extracting related knowledge from previous tasks for faster learning on novel tasks. This methodology is known as *learning to learn* or *meta-learning* [2]. Meta-learning has recently received great attention as a powerful solution to few-shot learning problem [3]. This is mainly due to the fact that learning from limited data with zero prior knowledge results in poor performance, unlike the case when a learning algorithm is able to reuse previously acquired related knowledge. Specifically, model-agnostic meta-learning (MAML) algorithm [4] has made great strides in this regard. MAML is a few-shot learning algorithm which formulates the meta-learning problem as a bilevel optimization problem where both meta and task-specific models are optimized. The objective is to find a meta-model that minimizes the average validation loss over N tasks while maintaining K SGD steps to each task-specific model which is in the direction of the minimum training loss. Both inner and outer level problems are solved using SGD which introduces high computation cost in the step of updating the meta model where matrix manipulation is required. Although, MAML can achieve fast adaptation to a new task with a few samples, it incurs significant computations, and ignores uncertainty quantification, which makes it brittle.

Uncertainty quantification presents a critical component for enabling mission-critical applications such as healthcare and autonomous vehicles. In this regard, existing meta-learning algorithms overlook model uncertainty since they train a meta-model in a deterministic fashion which may not generalize well to new coming tasks. Moreover, starting from a deterministic model and running a few shot learning with very limited samples per task may further increase uncertainty. For example, the authors in [5] showed that existing few-shot learning algorithms tend to overfit. Therefore, a robust meta-learning algorithm that is able to intrinsically deal with such uncertainty is needed. To this end, Bayesian learning is a promising technique to obviate this issue [6]–[8]. Not only it offers robustness towards overfitting and uncertainty estimation, but it also enables efficient learning in the small data regime. [6], [9]. Hence, a Bayesian view of few-shot meta-learning approach presents a logical step towards achieving robustness.

Motivated by the above, in this paper we propose, *MetaBayes*, a Bayesian meta-learning framework in which each agent learns a task-specific posterior distribution for its own task, and where all tasks collaborate to jointly optimize a global prior distribution to produce a meta-prior distribution that can be used by new tasks for faster learning. In fact, *MetaBayes* serves not only as a generalization and a robust approach for meta-learning but it also serves as a promising solution for learning informative Bayesian priors. Generally, Bayesian learning encounters challenges related to the intractability of estimating the posterior distributions, and choosing an informative prior. Variational inference is intended to address the first challenge [9]–[11], nevertheless, limited work has addressed the problem of inferring a suitable prior distribution [12]. Hence, most works in Bayesian learning assume a naive zero-centered Gaussian distribution which results in bad generalization and inaccurate uncertainty etimates, when training over scarce data [13]. In contrast to this prior art, *MetaBayes* framework extracts the shared knowledge from the set of existing tasks to form an informative prior to be used by new tasks, offering a principled approach to learn informative Bayesian priors.

Specifically, we propose a joint objective function with respect to tasks' posterior distributions and the meta-prior distribution. The intution behind the objective is to find the optimal posterior and prior distributions that minimize the Kullback-Leibler (KL) divergence between the true generating

Authorized licensed use limited to: Oulu University. Downloaded on March 30,2023 at 13:22:10 UTC from IEEE Xplore. Restrictions apply.

likelihood functions and the tasks' likelihood function. The problem is solved analytically using alternating minimization, in which tasks optimize their own task-specific posterior independently over their own training datasets. Subsequently, tasks collaborate to learn a shared informative prior distribution. In our work, we derive a novel closed-form formula for updating the global prior distribution.

The rest of the paper is organized as follows. In section II, we introduce the system model and problem formulation. In section III, we describe our alternating minimization based algorithm to solve the proposed optimization problem. In section IV, we introduce and discuss our simulation results. Lastly, we conclude the paper in section V.

Notation: We use boldface lowercase symbol for vectors $\boldsymbol{s}$, and boldface uppercase symbol for matrices $\boldsymbol{S}$. In addition, we refer to the KL divergence between two probability distributions as $D_{\text{KL}}(P_r||P_r')$ such that $(P_r, P_r') \in \Delta R$ where ΔR denotes a set of prabability distributions. Moreover, for simplicity, and without loss of generality, we discretize the parameter space Θ with K representative points. Lastly, if a probability distribution is denoted as $\boldsymbol{p}$, then $[p]_k$ and $p(\theta_k)$ denote interchangeably the probability at θ_k.

II. System Model and Problem Formulation

Consider a set $\mathcal{M} = \{1, 2, ..., M\}$ of M tasks where each task $i \in \mathcal{M}$ holds a local dataset $\mathcal{D}_i = \{(\boldsymbol{x}, y) | \boldsymbol{x} \in \mathcal{X}_i, y \in \mathcal{Y}_i\}$ of cardinality D_i where $\mathcal{X}_i$ is the local instance space at task i, and $\mathcal{Y}_i$ is the local set of all possible labels. Each task i generates input-label samples according to a probabilistic model with distribution $P_i(\boldsymbol{x}) f_i(y|\boldsymbol{x})$. Task i's local samples, $\mathcal{X}_i = \{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, ..., \boldsymbol{x}_i^{(D_i)}\}$, are assumed to be independent and identically distributed (i.i.d).

Each task $i \in \mathcal{M}$ aims to learn the true parameter for its own dataset; i.e. to learn $\theta_i^* \in \Theta$ where $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$ denotes a finite set of possible states. It is also assumed that each task i holds a set of local likelihood functions of the labels $\{l_i(y|\boldsymbol{x}, \theta_k) | y \in \mathcal{Y}_i, \boldsymbol{x} \in \mathcal{X}_i, \theta_k \in \Theta\}$. Furthermore, we denote the posterior distribution of task i over parameter θ at time $t \geq 0$ by $\boldsymbol{\mu}_i^t \in \Delta\Theta$ where $\Delta\Theta$ is a probability distribution over the set Θ.

Assumption 1. *All tasks $i \in \mathcal{M}$ start with a prior distribution $\boldsymbol{\mu}_0$ such that at time $t = 0$, $[\mu_0^0]_k > 0, \ \forall \theta_k \in \Theta$.*

Assumption 1 is necessary to rule out the degenerate case where zero Bayesian prior prevents learning.

Definition 1. *Let $[\mu_i^t]_*$ denotes task's i posterior distribution at time t at task's i true parameter θ_i^*. Under Assumption 1, task $i \in \mathcal{M}$ asymptotically learns the true parameter for its dataset θ_i^* on a path $(\boldsymbol{X}_i^{(t)}, \boldsymbol{y}_i^{(t)})_{t=1}^{\infty}$ if along that path $\lim_{t\to\infty} [\mu_i^t]_* = 1$.*

In other words, tasks assign probability one to their true parameters θ_i^* [14]. Definition 1 implies that the true labeling function $f_i(y|\boldsymbol{x})$ is equivalent to $l_i(y|\boldsymbol{x}, \theta_i^*)$ since the optimal posterior distribution takes value one at θ_i^* and zero elsewhere. The goal of each task i is to learn a parameter θ_k that makes its likelihood distribution $l_i(y|\boldsymbol{x}, \theta_k)$ as close as possible to its true likelihood function $l_i(y|\boldsymbol{x}, \theta_i^*)$. To measure the divergence between both distributions, we use $D_{\text{KL}}\left(l_i(y|\boldsymbol{x}, \theta_i^*) || l_i(y|\boldsymbol{x}, \theta_k)\right)$ which represents the tasks' local relative entropy. To this end, we cast the following optimization problem,

$$\underset{\{\theta_k\}\in\Theta}{\text{minimize}} \sum_{i=1}^{M} \mathbb{E}_{\boldsymbol{x}\sim P_i(\boldsymbol{x})} D_{\text{KL}}\Big(l_i(y|\boldsymbol{x}, \theta_i^*) || l_i(y|\boldsymbol{x}, \theta_k)\Big) \quad (1)$$

The above objective is convex with respect to θ_k.

Lemma 1. *The problem of minimizing the KL divergence D_{KL} between two distributions w.r.t θ is equivalent to the problem of maximizing the expectation of the logarithm of one distribution with respect to the other one. i.e.,*

$$\min_{\theta_k\in\Theta} D_{KL}(l_i^* || l_i) = \min_{\theta\in\Theta} \{\mathbb{E}_{y\sim l_i^*}(\log l_i^*) - \mathbb{E}_{y\sim l_i^*}(\log l_i)\} \quad (2)$$

$$= \max_{\theta\in\Theta} \mathbb{E}_{y\sim l_i^*}(\log l_i), \quad (3)$$

where $l_i = l_i(y|\boldsymbol{x}, \theta_k)$ and $l_i^* = l_i(y|\boldsymbol{x}, \theta_i^*)$. The term $\mathbb{E}_{y\sim l_i^*}(\log l_i^*)$ in (2) is ignored since it is not a function of the estimated parameter θ.

Using Lemma 1, equation (1) can be recast in terms of the model parameter as follows:

$$\underset{\{\theta_k\}\in\Theta}{\text{maximize}} \quad \sum_{i=1}^{M} \mathbb{E}_{x\sim P_i(x)} \mathbb{E}_{y\sim l_i^*}\Big(\log(l_i(y|\boldsymbol{x}, \theta_k)\Big). \quad (4)$$

The Maximum Likelihood Estimation (MLE) problem presented above can be casted as an optimization problem over the prior and posterior distributions by reformulating it as an inner product of the posterior vector $\boldsymbol{\mu}_i$[1] and expectation of log likelihood [14] as follows:

$$\underset{\{\boldsymbol{\mu}_i\}\in\Delta\Theta, \boldsymbol{\mu_0}\in\Delta\Theta}{\text{maximize}} \sum_{i=1}^{M} \boldsymbol{\mu_i}^T(\boldsymbol{\mu_0}) \, \mathbb{E}_{x\sim P_i(x)} \mathbb{E}_{y\sim l_i^*}\Big(\log(\boldsymbol{l}_i(y|\boldsymbol{x})\Big), \quad (5)$$

where $\boldsymbol{l}_i(y|\boldsymbol{x}) = [l_i(y|\boldsymbol{x}, \theta_1), \cdots, l_i(y|\boldsymbol{x}, \theta_K)]^T$. Problem (5) jointly optimizes the prior distribution and each task's posterior distribution. The equivalence of (4) and (5) follows immediately from Definition 1. In detail, the goal of problem (4) is to find θ_k for each task $i \in \mathcal{M}$ that maximizes the scalar log-likelihood $\log\big(l_i(y|\boldsymbol{x}, \theta_k)\big)$; i.e. θ_i^* for all $i \in \mathcal{M}$. Whereas the goal of problem (5) is to find the optimal posterior distribution vector $\boldsymbol{\mu_i}$ for all $i \in \mathcal{M}$ that maximizes the inner product with the log-likeihood vector; i.e. the posterior distribution $\boldsymbol{\mu_i}$ that gives value one at θ_i^* and zero elsewhere for all $i \in \mathcal{M}$. Next, we introduce our alternating minimization algorithm to solve the proposed optimization problem defined in (5).

[1] The posterior distribution $\boldsymbol{\mu}_i$ is in terms of the prior distribution $\boldsymbol{\mu}_0$.

Authorized licensed use limited to: Oulu University. Downloaded on March 30,2023 at 13:22:10 UTC from IEEE Xplore. Restrictions apply.

III. Meta Learning via Alternating Minimization

In this section we describe our alternating minimization based approach to solve the proposed problem defined in (5). We alternate between updating the tasks' posterior distributions given the current prior distribution and updating the prior distribution given the current posterior distributions.

A. Bayesian Posterior Distribution Estimation

Here, we optimize the objective in (5) with respect to tasks' posterior distributions $\boldsymbol{\mu}_i$ given the prior distribution $\boldsymbol{\mu}_0$. However, the major challenge in optimizing this objective lies in the fact that $\mathbb{E}_{y\sim l_i^*}(\cdot)$ is unknown which means that the true gradient of the objective cannot be computed. A common approach to tackle the objective in (5) is to consider the empirical average as the cost function, and solve the online stochastic learning problem [14] as follows.

At iteration $t-1$, each task i draws a mini-batch of observations $(\boldsymbol{X}_i^{t-1}, \boldsymbol{y}_i^{t-1})$ of cardinality B from its local dataset $\mathcal{D}_i$. Hence, $\boldsymbol{g}_i^{t-1}$, the task i's stochastic gradient of the objective presented in (5) with respect to the posterior vector $\boldsymbol{\mu}_i$ at time $t-1$ is computed as follows:

$$\boldsymbol{g}_i^{t-1} = \log(\boldsymbol{\gamma}_i(\boldsymbol{y}_i^{t-1}|\boldsymbol{X}_i^{t-1})), \tag{6}$$

where

$$\boldsymbol{\gamma}_i(\boldsymbol{y}_i^{t-1}|\boldsymbol{X}_i^{t-1}) = \frac{1}{B}\sum_{e=1}^{B} \boldsymbol{l}_i(y_e^{t-1}\,|\,\boldsymbol{x}_e^{t-1}),\ \ (\boldsymbol{x}_e^{t-1}, y_e^{t-1}) \in \{(\boldsymbol{X}_i^{t-1}, \boldsymbol{y}_i^{t-1})\} \tag{7}$$

represents an approximation for the likelihood in equation (5) at time $t-1$. We then employ a regularized dual averaging scheme generating $\boldsymbol{z}_i^t$ and $\boldsymbol{\mu}_i^t$ where

$$\boldsymbol{z}_i^t = \boldsymbol{z}_i^{t-1} + \boldsymbol{g}_i^{t-1}, \tag{8}$$

and

$$\boldsymbol{\mu}_i^t = \underset{\boldsymbol{b}_i\in\Delta\Theta}{\operatorname{argmin}} \Big\{ -\langle \boldsymbol{z}_i^t, \boldsymbol{b}_i\rangle + \frac{1}{\alpha_t}\psi(\boldsymbol{b}_i)\Big\}. \tag{9}$$

Note that $\langle\cdot,\cdot\rangle$ represents the standard inner product. The dual update $\boldsymbol{z}_i^t$, essentially integrates the stochastic gradients, and the update in (9) projects the integration on the feasible set while regularizing the projection using a so called proximal function $\psi(\boldsymbol{b}_i)$. To derive the Bayesian parameter estimation from this setup at iteration t, the proximal function needs to be the KL-divergence from the prior distribution defined as follows [15]:

$$\psi(\boldsymbol{b}_i) = D_{\mathrm{KL}}(\boldsymbol{b}_i||\boldsymbol{\mu}_i^{t-1}) = \sum_{k=1}^{K}[b_i]_k \log\frac{[b_i]_k}{[\mu_0^{t-1}]_k}. \tag{10}$$

By letting $\alpha_t = 1$ for all t, the optimization problem in (9) can be recast as follows:

$$\begin{aligned}\boldsymbol{\mu}_i^t &= \underset{\boldsymbol{b}_i\in\Delta\Theta}{\operatorname{argmin}} \Big\{ -\boldsymbol{b}_i^T\boldsymbol{z}_i^t + \textstyle\sum_{k=1}^{K}[b_i]_k\log\frac{[b_i]_k}{[\mu_0^{t-1}]_k}\Big\} \\ &\text{subject to } [b_i]_k \ge 0,\ \sum_{k=1}^{K}[b_i]_k = 1.\end{aligned} \tag{11}$$

This optimization problem can be solved optimally and that is formally stated in the following theorem.

Theorem 1. *The optimal solution to problem* (11) *is given by:*

$$[\mu_i^t]_k = \frac{[\mu_0^{t-1}]_k\prod_{\tau=0}^{t-1}[\gamma_i(\boldsymbol{y}_i^\tau|\boldsymbol{X}_i^\tau)]_k}{\displaystyle\sum_{\theta_q\in\Theta}[\mu_0^{t-1}]_q\prod_{\tau=0}^{t-1}[\gamma_i(\boldsymbol{y}_i^\tau|\boldsymbol{X}_i^\tau)]_q}. \tag{12}$$

Proof. Leaving the positivity constraint implicit, we can write (11) as the maximization of the following Lagrangian,

$$\begin{aligned}L_i(\boldsymbol{b},\lambda) =& \boldsymbol{b}_i^T\boldsymbol{z}_i^t - \sum_{k=1}^{K}[b_i]_k\log\frac{[b_i]_k}{[\mu_0^{t-1}]_k} \\ &+ \lambda(\boldsymbol{b}_i^T\mathbf{1}-1)\end{aligned} \tag{13}$$

where $\mathbf{1}$ is vector of all ones. By differentiating equation (13), we get the following:

$$\begin{aligned}\frac{\partial}{\partial[b_i]_k}L_i(\boldsymbol{b},\lambda) &= [z_i^t]_k - \Big[1+\log[b_i]_k \\ &\quad - \log[\mu_0^{t-1}]_k\Big] + \lambda \\ &= [z_i^t]_k - \log[b_i]_k \\ &\quad + \log[\mu_0^{t-1}]_k + \lambda - 1 \\ &= [z_i^t]_k - \log[b_i]_k + \log[\mu_0^{t-1}]_k \\ &\quad + \lambda - 1\end{aligned} \tag{14}$$

$$\frac{\partial}{\partial\lambda}L_i(\boldsymbol{b},\lambda) = \boldsymbol{b}_i^T\mathbf{1} - 1$$

The condition for the stationary point is,

$$[\mu_i^t]_k = \frac{[\mu_0^{t-1}]_k\prod_{\tau=0}^{t-1}[\gamma_i(\boldsymbol{y}_i^\tau|\boldsymbol{X}_i^\tau)]_k}{\displaystyle\sum_{\theta_q\in\Theta}[\mu_0^{t-1}]_q\prod_{\tau=0}^{t-1}[\gamma_i(\boldsymbol{y}_i^\tau|\boldsymbol{X}_i^\tau)]_q}. \tag{15}$$

and this concludes the proof. □

It is worthy to highlight that the denominator in equation (15) reflects a normalization constant that does not depend on θ which we refer to as C_i^t. A major challenge of calculating the posterior distribution in (15) is the intractability of the normalization constant C_i^t due to the large search space or intractable integrals in case of continuous variables. Thus, in these cases, we seek to approximate the posterior distribution in (15) via variational inference. This will be discussed in details in subsection III-C.

B. Prior Belief Optimization

In this section, we optimize the objective in (5) with respect to the prior distribution $\boldsymbol{\mu}_0$ given the current tasks' updated posterior distributions $\boldsymbol{\mu}_i^t$ for all $i \in \mathcal{M}$. Following the same analysis provided in III-A, we let $\boldsymbol{g}_0^t$ to be the stochastic gradient of objective presented in (5) at time t with respect to prior distribution $\boldsymbol{\mu}_0$ as follows:

$$\boldsymbol{g}_0^t = \sum_{i=1}^{M}\Big[\frac{1}{C_i^t}\prod_{\tau=0}^{t-1}\boldsymbol{\gamma}_i(\boldsymbol{y}_i^\tau|\boldsymbol{X}_i^\tau)\,\log(\boldsymbol{\gamma}_i(\boldsymbol{y}_i^\tau|\boldsymbol{X}_i^\tau))\Big] \tag{16}$$

Then, $\boldsymbol{\mu}_0^t$ is updated by solving the following optimization problem:

$$\boldsymbol{\mu}_0^t = \underset{\boldsymbol{b_0}\in\Delta\Theta}{\operatorname{argmin}}\{-\langle\boldsymbol{g}_0^t, \boldsymbol{b}_0\rangle + \frac{1}{\alpha_t}\psi(\boldsymbol{b}_0)\}, \tag{17}$$

Authorized licensed use limited to: Oulu University. Downloaded on March 30,2023 at 13:22:10 UTC from IEEE Xplore. Restrictions apply.

where

$$\psi(\boldsymbol{b}_0) = \sum_{k=1}^{M} [b_0]_k \log \frac{[b_0]_k}{[\mu_0^{t-1}]_k}. \tag{18}$$

Setting $\alpha_t = 1$ for all t, we get the following:

$$\begin{aligned} \boldsymbol{\mu}_0^t = \operatorname*{argmin}_{\boldsymbol{b}_0 \in \Delta\Theta} & \left\{ -\boldsymbol{b}_0^T \boldsymbol{g}_0^t + \sum_{k=1}^{K} [b_0]_k \log \frac{[b_0]_k}{[\mu_0^{t-1}]_k} \right\} \\ \text{subject to } & [b_0]_k \geq 0, \ \sum_{k=1}^{K} [b_0]_k = 1 \end{aligned} \tag{19}$$

Theorem 2. *The optimal solution for problem* (19) *is given by:*

$$[\mu_0^t]_k = \frac{[\mu_0^{t-1}]_k \exp\left(\sum_{i=1}^{T} \left[\frac{1}{C_i^t} \prod_{\tau=0}^{t-1} [\gamma_i^\tau]_k \ \log([\gamma_i^\tau]_k) \right] \right)}{\sum_{\theta_q \in \Theta} \left([\mu_0^{t-1}]_q \exp\left(\sum_{i=1}^{T} \left[\frac{1}{C_i^t} \prod_{\tau=0}^{t-1} [\gamma_i^\tau]_q \ \log([\gamma_i^\tau]_q) \right] \right) \right)}, \tag{20}$$

where

$$[\gamma_i^\tau]_\beta = [\gamma_i(\boldsymbol{y}^\tau | \boldsymbol{X}^\tau)]_\beta. \tag{21}$$

Proof. Leaving the positivity constraint implicit, we recast equation (19) as the maximization of the following lagrangian:

$$\begin{aligned} L(\boldsymbol{b}_0, \lambda) = & \boldsymbol{b}_0^T \boldsymbol{g}_0^t - \sum_{k=1}^{K} [b_0]_k \log \frac{[b_0]_k}{[\mu_0^{t-1}]_k} \\ & + \lambda(\boldsymbol{b}_0^T \mathbf{1} - 1). \end{aligned} \tag{22}$$

Differentiating equation (22), we get the following:

$$\begin{aligned} \frac{\partial}{\partial [b_0]_k} L(\boldsymbol{b}_0, \lambda) &= [g_0^t]_k - \Big[1 + \log [b_0]_k \\ &\quad - \log [\mu_0^{t-1}]_k \Big] + \lambda \\ &= [g_0^t]_k - \log [b_0]_k + \log [\mu_0^{t-1}]_k \\ &\quad + \lambda - 1 \\ \frac{\partial}{\partial \lambda} L(\boldsymbol{b}_0, \lambda) &= \boldsymbol{b}_0^T \mathbf{1} - 1. \end{aligned} \tag{23}$$

Setting the derivatives to zero, we get

$$[\mu_0^t]_k = \frac{[\mu_0^{t-1}]_k \exp\left(\sum_{i=1}^{T} \left[\frac{1}{C_i^t} \prod_{\tau=0}^{t-1} [\gamma_i^\tau]_k \ \log([\gamma_i^\tau]_k) \right] \right)}{\sum_{\theta_q \in \Theta} \left([\mu_0^{t-1}]_q \exp\left(\sum_{i=1}^{T} \left[\frac{1}{C_i^t} \prod_{\tau=0}^{t-1} [\gamma_i^\tau]_q \ \log([\gamma_i^\tau]_q) \right] \right) \right)}, \tag{24}$$

where

$$[\gamma_i^\tau]_\beta = [\gamma_i(\boldsymbol{y}^\tau | \boldsymbol{X}^\tau)]_\beta. \tag{25}$$

and this concludes the proof. □

The prior update rule in equation (24) may not produce a closed-form distribution, so we also seek to approximate the resulted prior distribution via variational inference.

C. Probability Distribution Approximation via Variational Inference

At first, we would like to point out that both $\boldsymbol{\mu}_i^t$ and $\boldsymbol{\mu}_0^t$ are implicitly conditioned on tasks' datasets. Hereafter, we denote the intractable distribution by $\boldsymbol{\mu}^t(\theta \,|\, \mathcal{D})$ which represents both $\boldsymbol{\mu}_i^t$ in equation (15) and $\boldsymbol{\mu}_0^t$ in equation (24). Note that

$$\boldsymbol{\mu}^t(\theta \,|\, \mathcal{D}) = \frac{\boldsymbol{\mu}^t(\mathcal{D} \,|\, \theta) \boldsymbol{\mu}^t(\theta)}{\boldsymbol{\mu}^t(\mathcal{D})}. \tag{26}$$

In variational inference, we approximate this intractable distribution $\boldsymbol{\mu}^t(\theta \,|\, \mathcal{D})$ by a *variational distribution* $\boldsymbol{\pi}^t(\theta)$, where $\boldsymbol{\pi}^t(\theta)$ is restricted to belong to a family of distributions Q of tractable form (as in Gaussian distributions), chosen with the goal of making $\boldsymbol{\pi}^t(\theta)$ as close as possible to the true posterior distribution $\boldsymbol{\mu}^t(\theta \,|\, \mathcal{D})$. The similiarity between the two distributions is measured in terms of KL-divergence; hence, the variational inference is performed by finding the distribution $\boldsymbol{\pi}^t(\theta)$ that minimizes the KL-divergence as follows:

$$\boldsymbol{\pi}^t(\theta) = \operatorname*{argmin}_{\boldsymbol{\xi} \in Q} D_{\text{KL}}(\boldsymbol{\xi}(\theta) || \boldsymbol{\mu}^t(\theta \,|\, \mathcal{D})) \tag{27a}$$

$$= \operatorname*{argmin}_{\boldsymbol{\xi} \in Q} D_{\text{KL}}(\boldsymbol{\xi}(\theta) || \boldsymbol{\mu}^t(\theta)) - \mathbb{E}_{\boldsymbol{\xi}(\theta)} \left(\log \boldsymbol{\mu}^t(\mathcal{D} \,|\, \theta) \right) \tag{27b}$$

The resulting cost function in (27b) is known as the variational free energy [8] [16]. which can be minimized using gradient descent and other various approximations. For instance, if we let Q to be the family of Gaussian mean-field approximate distributions, then a Gaussian variational distribution can be approximated by employing a simple Monte Carlo to compute the gradients using *Bayes by Backprop* [16].

IV. Numerical Evaluation

A. Simulation settings

In this section, we evaluate our MetaBayes framework on a multitask linear regression scenario. We use the bodyfat database [17] where tasks have observations of abdomen feature x to predict bodyfat percentage. Nevertheless, different tasks have different spectrums of abdomen feature; and each task aims to find the best line that fits its spectrum of observations. We assume that only two tasks have sufficient observations on which we trained our initializer (meta-model) using MetaBayes. Then, we draw 45 random tasks that only have few training samples. We train these statistically-insufficient tasks starting from the met-model generated by MetaBayes prior. We also considered zero-centered Gaussian prior with an identity covariance matrix for the comparison. In Figure 1, we plot the Mean Squared Error (MSE) of predictions of one random task over its test dataset under the two scenarios. Moreover, in Figure 2, we plot the empirical cumulative distribution function (CDF) of MSE values for the whole 45 tasks under the two cases. Next we describe the results

B. Result discussion

In Figure 1, we compare the two scenarios of few-shot learning for a randomly-drawn task. In the first scenario,

Authorized licensed use limited to: Oulu University. Downloaded on March 30,2023 at 13:22:10 UTC from IEEE Xplore. Restrictions apply.

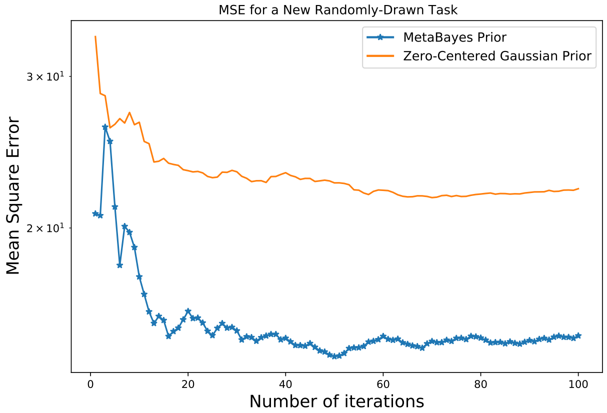

Fig. 1. figure

Performance comparison between MetaBayes prior and naive zero-centered Gaussian prior in terms of accuracy and convergence speed.

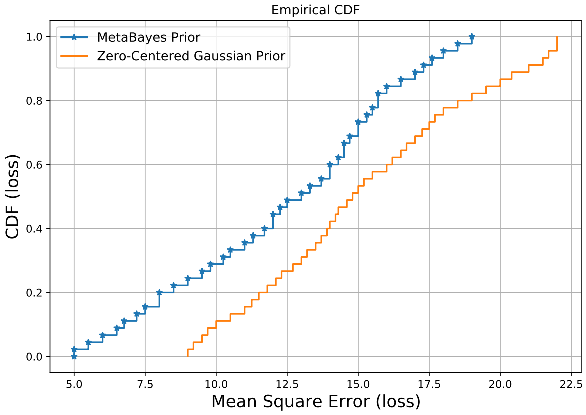

Fig. 2. figure

Performance comparison between MetaBayes prior and naive zero-centered Gaussian prior in terms of empirical CDF of MSE values.

the few-shot learning task starts from a MetaBayes-optimized prior, while in the other scenario, it starts from the commonly-used zero-centered Gaussian distribution. The figure clearly shows that the performance of the few-shot learning task with MetaBayes prior outperforms the performance of the same task trained with a zero-centered Gaussian prior in terms of accuracy and convergence speed. That is, the MetaBayes prior gives the few-shot learning task an informative start which is indeed needed especially when training with few samples.

This behavior is further shown in Figure 2 where we plot the empirical CDF of the MSE values for 45 randomly-drawn tasks trained in one experiment under a MetaBayes prior and in the other experiment under zero-centered Gaussian prior. This figure shows that the behavior observed in Figure 1 can be generalized for a much larger pool of few-shot learning tasks showing that for 80% of the tasks MetaBayes-optimized prior achieves a loss equal or less than 15.5 while the naive zero-centered Gaussian achieves a loss equal or less than 18.5.

V. Conclusion

In this paper, we proposed MetaBayes, a novel meta-learning framework from a Bayesian perspective where we jointly optimize a meta-prior distribution along with task-specific posterior distributions. The proposed framework is based on alternating minimization where two subproblems are optimized in an alternating fashion. We propose a closed-form expression to update the meta-prior and posterior distributions at every iteration. Our numerical evaluation shows that tasks with few training samples achieve higher accuracy and faster convergence when leveraging a MetaBayes learned prior distribution compared to the zero-mean Gaussian prior distribution.

References

[1] L. Smith and L. Slone, "A developmental approach to machine learning?" *Frontiers in psychology*, 2017.

[2] J. Biggs, "The role of metalearning in study processes," *British journal of educational psychology*, pp. 185–212, 1985.

[3] B. Lake, R. Salakhutdinov, and J. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, pp. 1332–1338, 2015.

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *International Conference on Machine Learning (ICML)*, 2017.

[5] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," 2017.

[6] Y. Gal, "Uncertainty in deep learning," *University of Cambridge*, 2016.

[7] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," 2015.

[8] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" 2017.

[9] D. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Advances in Neural Information Processing Systems*, p. 2575–2583, 2015.

[10] T. Broderick, N. Boyd, A. Wibisono, A. Wilson, and M. Jordan, "Streaming variational bayes," *Advances in Neural Information Processing Systems*, pp. 1727–1735, 2013.

[11] C. Nguyen, Y. Li, T. Bui, and R. Turner, "Variational continual learning," *International Conference on Learning Representations*, 2018.

[12] M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel, "Understanding priors in bayesian neural networks at the unit level," *International Conference on Machine Learning*, 2019.

[13] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," *International Conference on Machine Learning*, 2018.

[14] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," *Proceedings of the IEEE Conference on Decision and Control*, 2013.

[15] T. M. Cover and J. A. Thomas, "Elements of information theory," *John Wiley and Sons*, 2012.

[16] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *In Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, pp. 1613–1622, 2015.

[17] StatLib, "Bodyfat database," *https://www.csie.ntu.edu.tw/ cjlin/libsvmtools /datasets/regression/bodyfat.*

Authorized licensed use limited to: Oulu University. Downloaded on March 30,2023 at 13:22:10 UTC from IEEE Xplore. Restrictions apply.