Efficacy of Regularized Multitask Learning Based on SVM Models
Shaohan Chen, Zhou Fang, Sijie Lu, and Chuanhou Gao, Senior Member, IEEE

Abstract—This article investigates the efficacy of a regularized multitask learning (MTL) framework based on SVM (M-SVM) to answer whether MTL always provides reliable results and how MTL outperforms independent learning. We first find that the M-SVM is Bayes risk consistent in the limit of a large sample size. This implies that despite the task dissimilarities, the M-SVM always produces a reliable decision rule for each task in terms of the misclassification error when the data size is large enough. Furthermore, we find that the task-interaction vanishes as the data size goes to infinity, and the convergence rates of the M-SVM and its single-task counterpart have the same upper bound. The former suggests that the M-SVM cannot improve the limit classifier’s performance; based on the latter, we conjecture that the optimal convergence rate is not improved when the task number is fixed. As a novel insight into MTL, our theoretical and experimental results achieved an excellent agreement that the benefit of the MTL methods lies in the improvement of the preconvergence-rate (PCR) factor (to be denoted in Section III) rather than the convergence rate. Moreover, this improvement of PCR factors is more significant when the data size is small. In addition, our experimental results of five other MTL methods demonstrate the generality of this new insight.

Index Terms—Error analysis, learning theory, multitask learning (MTL), preconvergence-rate (PCR) factor, regularization method.

I. INTRODUCTION

RECENT years have seen a considerable success of machine-learning technologies, including deep neural networks (DNNs), SVMs, decision trees, etc. These methods usually work well when the training and test data are drawn from the same distribution and “big data” are available (e.g., DNNs). However, the conditions of the same distribution and big data cannot always hold in many real-world problems, for example, building self-driving systems and modeling rare diseases in healthcare records. Thus, developing methods to learn multiple tasks simultaneously and efficiently is significant.

In this case, multitask learning (MTL) [1] provides a general way to accommodate these situations. A vast of elaborate MTL methods have been designed in various settings to solve specific problems [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Also, there have been a lot of notable theoretical justifications for the superiority of MTL over independent learning. Ando et al. [12] studied MTL based on structural learning and showed that it estimates the shared hypothesis space more reliably if the number of tasks T is large enough. Baxter [13], [14] investigated MTL in the framework of bias learning and showed that it significantly reduces the sampling burden for good generalization on novel tasks if T is arbitrarily large. He also illustrated that MTL (within a Bayesian context) efficiently decays the information required to learn each task as T grows considerably [15]. Maurer et al. [16] discussed MTL in representation learning and showed that it vanishes the cost to learn the representation in the multitask limit (T → ∞). Liu et al. [17] analyzed MTL using the sample average stability measure and showed that it estimates the shared parameter more accurately when T is large enough. Zhang et al. [18] viewed MTL as a vector-valued function learning problem and showed that it requires a smaller sample size of each task to achieve the same performance.

The existing theoretical analysis of MTL relies on the pursuit of tighter error bounds, which is not enough to guarantee the superiority of MTL. Moreover, as introduced above, the advantages of MTL can so far only be seen when T is sufficiently large, which makes MTL challenging to be trusted by users. The arbitrarily large T requirement is rarely satisfied in real applications, for example, it is often difficult to access a dataset with large T in the healthcare and industrial systems. Therefore, it is crucial to investigate the intrinsic benefits of MTL by applying a more elaborate analysis method to it when T is fixed.

This article moves the steps forward and tries to explore the essential benefits of MTL with respect to the misclassification error when T is fixed. Particularly, we concern with a popular regularized multitask SVM (M-SVM) proposed by Evgeniou and Pontil [19], which provides an avenue to share useful knowledge among tasks through parameters. Note that the analysis of the misclassification error has been intensively researched in the single-task learning contexts [20], [21], [22], [23], [24], [25]. This work solves two main technical difficulties of adapting the misclassification error analysis to MTL: considering the interactions between tasks carefully and estimating the additional error caused by the randomness of the sampling frequencies for each task (see Section III).
We first analyze the asymptotic performance of M-SVM to show that it is Bayes risk consistent in the limit of large sample size. This implies that despite the task dissimilarities, the M-SVM always produces a reliable decision rule for each task in terms of the misclassification error when the data size is large enough. Second, we find that the task-interaction vanishes as the data size goes to infinity, and the convergence rates of the M-SVM and its single-task counterpart have the same upper bound. The former suggests that the M-SVM cannot improve the limit classifier’s performance; based on the latter, we conjecture that the optimal convergence rate is not improved when the number of tasks is fixed. The intuition of this conjecture is that the generalization capacity mainly depends on the hypothesis space’s complexity which MTL cannot change.

As a novel insight into MTL, our theoretical (on a 1-D classification problem) and experimental results achieve an excellent agreement that the benefit of the M-SVM lies in improving the preconvergence-rate (PCR) factor (which is denoted as the ratio between the excess misclassification error and its real convergence rate in Section III) rather than the convergence rate. Moreover, our experimental results of several other MTL methods, including the L21, Lasso, SRM, multitask least-square twin SVM (abbreviated as “M-TSVM”) [26], and convex multitask SVM (referred to as “M-CSVM”) [27] (the first three models can be found and solved by [28]) also demonstrate the generality of this new insight in MTL. As shown in Section IV-A, the PCR factor is a function of data size, the similarity of the tasks, and regularization parameters, meaning that MTL can achieve a better performance when there exists a good balance of these quantities. Particularly, the improvement of PCR factors in MTL is more significant when the data size is small. Therefore, the PCR factor is more suitable and accurate to depict the essential advantages of MTL when it is fixed.

To conclude, our core contributions are as follows.

1) We analyze the asymptotic performance of the M-SVM and show that it is Bayes risk consistent in the limit of large sample size.
2) We propose a novel insight into M-SVM that the benefit of M-SVM over its single-task version lies in improving the PCR factor rather than the convergence rate. The correctness of this claim is proved theoretically (on a 1-D classification problem) and empirically (by both simulated and real datasets).
3) Our experimental results of five other MTL methods demonstrate the generality of this new insight in MTL.

The remainder of this article is organized as follows. Section II introduces the basic notations of MTL and an extension of M-SVM. Section III provides an asymptotic analysis of M-SVM in the limit of a large sample size, based on which we raise the main conjecture in this article. In Section IV, both the theoretical and experimental analysis are provided to verify the correctness of our conclusions. Section V presents several discussions, and Section VI concludes this article. To improve the readability of this article, we put all proofs in the Appendix.

II. Method

A. Problem Settings and Notations

We consider the following MTL setting. Assume that we are facing $T$ learning tasks and samples for each task comes out randomly with probability $p(t)$ at each time, for $t = 1, \ldots, T$. After a short time, we collect $N$ data points totally from $T$ tasks and suppose that all these samples belonging to the same space $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y = \{-1, 1\}$. Specifically, for each task $t$, there are $m_t$ samples generated from the distribution $P_t(X, Y)$, for $t = 1, \ldots, T$. That is

$$\left\{ \left( x_{1t}, y_{1t} \right), \ldots, \left( x_{m_t1}, y_{m_t1} \right) \right\}, \ldots, \left\{ \left( x_{1T}, y_{1T} \right), \ldots, \left( x_{m_TT}, y_{m_TT} \right) \right\}. \tag{1}$$

For better adapting to the real scenarios, we assume $p(t)$ is different from each task, while their distribution functions $P_t(X, Y)$ are closely related. The ideal goal of MTL is to learn $T$ functions simultaneously $f_1, \ldots, f_T$ such that

$$\text{sgn}(f_i)(x_{it}) = y_{it}, \text{ for } i = 1, \ldots, m_t, \ t = 1, \ldots, T.$$ 

Based on the above settings, we define the average misclassification error for these $T$ classifiers to be the weighted sum of the corresponding misclassification errors [20]:

$$\mathcal{U}(\text{sgn}(f_1), \ldots, \text{sgn}(f_T)) = \sum_{t=1}^{T} p(t) \mathcal{R}_t(\text{sgn}(f_i)), \text{ where } \mathcal{R}_t(f_i) = \mathbb{E}_{X, Y} \mathbb{I}(\text{sgn}(f_i)(X) \neq Y), \text{ for } t = 1, \ldots, T,$$

and $\mathbb{I}(A)$ is an indicator function with its value being 1 if the event $A$ is true, and 0 if it is not. We see that the average misclassification error actually measures the risk of applying $f_1, \ldots, f_T$ to make predictions.

We define the minimizer of the misclassification error for each task as $f_i^* = \arg \inf_{f_i} \mathcal{R}_t(f_i)$, where the infimum is over all measurable functions. Based on [29], this minimizer has the expression as $f_i^*(x) = \text{sgn}(\eta_i(x) - [1/2])$ and is called the Bayes rule, where $\eta_i(x) = P_i(Y = 1|X = x)$, for $t = 1, \ldots, T$.

We now define the average expected error of $T$ function $f_1, \ldots, f_T$ with respect to a loss function $\ell : \mathbb{R} \rightarrow [0, \infty)$ as

$$\mathcal{E}(f_1, \ldots, f_T) = \sum_{t=1}^{T} p(t) \mathbb{E}_{X, Y} \mathbb{I}(\ell(\mathcal{Y} f_i(X))), \text{ then the corresponding average empirical error is } \mathcal{E}_c(f_1, \ldots, f_T) = \sum_{t=1}^{T} \sum_{i=1}^{m_t} (1/N) \ell(y_{it} \delta(x_{it})).$$

This article mainly concerns with the following denotations. For each task $t$, we define the excess misclassification error of the classifier $\text{sgn}(f_i)$ as $\mathcal{R}(\text{sgn}(f_i)) - \mathcal{R}(f_i^*)$. For all the $T$ tasks, the average excess misclassification error of the classifiers $\text{sgn}(f_i)$ ($t = 1, \ldots, T$) is defined as $\mathcal{U}(\text{sgn}(f_1), \ldots, \text{sgn}(f_T)) - \mathcal{U}(f_1^*, \ldots, f_T^*)$, and the average excess expected error of the functions $f_i$ ($t = 1, \ldots, T$) is defined as $\mathcal{E}(f_1, \ldots, f_T) - \mathcal{E}(f_1^*, \ldots, f_T^*)$.

B. M-SVM Method

This section extends the M-SVM originating from [19] to a more general case, where we assume that the functions $f_1, \ldots, f_T$ are nonlinear and the probability of each sample coming out from task $t$ is $p(t)$ instead of $(1/T)$. Note that the exact $p(t)$ is generally unknown, we use the sampling frequency $(m_t/N)$ to approximate $p(t)$ for each task in the problem and obtain a extend version of the M-SVM...
as follows:

$$\min_{f_0, f_1, \ldots, f_T} \left\{ \sum_{t=1}^{T} m_t \sum_{i=1}^{m_t} \xi_{it} + \lambda_1 \sum_{i=1}^{T} \frac{m_t}{N} \| \mathbf{g}_i \|^2_K + \lambda_2 \| f_0 \|^2_K \right\}$$

$$\text{s.t. } y_{it} \cdot f_i(x_{it}) \geq 1 - \xi_{it}, \quad \xi_{it} \geq 0 \quad i = 1, \ldots, m_t; \quad t = 1, \ldots, T$$

(2)

where $\lambda_1$ and $\lambda_2$ are two positive regularization parameters, $K$ is a universal kernel (e.g., the Gaussian kernel), and $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) [20]. $\| \cdot \|$ is the norm function in this Hilbert space, and $g_t := f_t - f_0$. Since (2) contains regularization terms $\| \mathbf{g}_i \|^2_K$ and $\| f_0 \|^2_K$ that model task relations, we refer to this as a regularized MTL model. In (2), $f_0$ represents the commonness of those classifiers, while $g_t$, $t = 1, \ldots, T$, represent their individualities. The extended model of (2) will be reduced to the original one [19] when $m_t$ ($t = 1, \ldots, T$) are identical. Note that considering the randomness in the sampling frequency for each task makes the proposed method better adaptable to real applications, as the model pays more attention to the higher frequency tasks. Hence, M-SVM (2) well extends the original MTL framework. By the positive definiteness and convexity of the loss function and the norm function, there exists the optimal solution of (2).

Following a similar calculation scheme of [19], and denoting $f^*_t (t = 1, \ldots, T)$, $f^*_0$ to be the optimal solution of (2), we can obtain a relation of these quantities in Lemma 1, and then reformulate (2) as (3) in Theorem 1.

Lemma 1 (Adapted From [19, Lemma 2.1]):
The optimal solution to (2) satisfies the equation $f^*_0 = (\lambda_1/\lambda_2 + \lambda_1) \sum_{t=1}^{T} m_t / N f^*_t$.

Theorem 1 (Adapted From [19, Lemma 2.2]): The multi-task problem (2) is equivalent to the optimization problem as follows:

$$\min_{f_1, \ldots, f_T} \left\{ \sum_{i=1}^{T} \sum_{t=1}^{m_t} \xi_{it} + \rho_1 \sum_{i=1}^{T} \frac{m_t}{N} f_i^T \right\}$$

$$+ \rho_2 \sum_{i=1}^{T} \frac{m_t}{N} \left\| f_i - \sum_{s=1}^{m_t} \frac{m_t}{N} f_s \right\|^2_K$$

$$\text{s.t. } y_{it} \cdot f_i(x_{it}) \geq 1 - \xi_{it}, \quad \xi_{it} \geq 0 \quad i = 1, \ldots, m_t; \quad t = 1, \ldots, T$$

(3)

where $\rho_1 = (\lambda_1 \lambda_2 / \lambda_1 + \lambda_2)$ and $\rho_2 = (\lambda_1^2 / \lambda_1 + \lambda_2)$.

Theorem 1 indicates that the M-SVM works by achieving a good trade-off between independent learning [i.e., the second term in (3)] and aggregate learning that treats different tasks as the one [i.e., the last term in (3)].

In the next section, we will show an asymptotic property of the M-SVM in (3) that it is Bayes risk consistent in the limit of a large sample size.

III. ASYMPTOTIC PERFORMANCE OF THE M-SVM

In this section, we first present the average excess misclassification error bound for M-SVM (3). Then, based on the analysis of this bound, we raise the main conjecture that the superiority of M-SVM is the improvement of the PCR factor rather than the convergence rate.

A. Error Analysis of the M-SVM

This section derives an upper bound of the average excess misclassification error for the M-SVM (3):

$$\mathcal{U}(\text{sgn}(f^*_1), \ldots, \text{sgn}(f^*_T)) - \mathcal{U}(f^*_1, \ldots, f^*_T),$$

by extending the analysis techniques for the single-task setting [20], [21], [25], [30], [31]. The definition of this error is given in Section II-A.

First, we show the minimizers of the average expected error $\mathcal{E}(f_1, \ldots, f_T)$ to be the Bayes rules.

Theorem 2: The minimizers of $\mathcal{E}(f_1, \ldots, f_T)$ over all measurable functions are the Bayes rules $f^*_t$ where $t = 1, \ldots, T$.

Proof: Note that the minimizer of each $\mathbb{E}_{(\mathbf{X}, Y) \sim p_{\mathbf{X}}(\mathcal{H}(\mathbf{X}))}$ is the Bayes rule $f^*_t$ over all measurable functions [32, Lemma 3.1]. By the definition, the Bayes rules also minimize $\mathcal{E}(f_1, \ldots, f_T)$.

Theorem 2 implies one advantage of the MTL framework of (3) that it leverages the related tasks by introducing a regularization term in the objective function without imposing the additional restrictions on the functional space [12], [16], and, therefore, the best rule (Bayes rule) is not precluded.

Similar to the single-task case, there exists a bridge inequality which connects the average excess misclassification error with the average expected excess error.

Theorem 3: For any $h_t : X \rightarrow \mathbb{R}$, $t = 1, \ldots, T$, there exists a bridge inequality between the average misclassification error $\mathcal{U}(h_1, \ldots, h_T)$ and the average expected error $\mathcal{E}(h_1, \ldots, h_T)$

$$\mathcal{U}(\text{sgn}(h_1), \ldots, \text{sgn}(h_T)) - \mathcal{U}(f^*_1, \ldots, f^*_T) \leq \mathcal{E}(h_1, \ldots, h_T) - \mathcal{E}(f^*_1, \ldots, f^*_T).$$

(4)

Proof: Note that for each $h_t : X \rightarrow \mathbb{R}$ the inequality $\mathcal{R}_t(\text{sgn}(h_t)) - \mathcal{R}_t(f^*_t) \leq \mathcal{E}_t(h_t) - \mathcal{E}(f^*_t)$ has established in [33], where $\mathcal{E}_t(h_t) = \mathbb{E}_{(\mathbf{X}, Y) \sim p_{\mathbf{X}}(\mathcal{H}(\mathbf{X}))}$. Then, by the definitions of $\mathcal{U}(h_1, \ldots, h_T)$ and $\mathcal{E}(h_1, \ldots, h_T)$, we can obtain the desired result directly.
where
\[
S(N, T) := \{ \mathcal{E}(f_1^t, \ldots, f_T^t) - \mathcal{E}_s(f_1^t, \ldots, f_T^t) \}
\]
\[
+ \{ \mathcal{E}_s(f_1^T, \ldots, f_T^T) - \mathcal{E}_s(f_1^T, \ldots, f_T^T) \}
\]
\[
D(N, T, \rho_1, \rho_2) := \inf_{h_1, \ldots, h_T \in \mathcal{H}_k} \left\{ \mathcal{E}(h_1, \ldots, h_T) - \mathcal{E}(f_1^t, \ldots, f_T^t) + \frac{\rho_1}{N} \sum_{t=1}^T p(t) \| h_t \|_K^2 
\right. 
\]
\[
\left. + \frac{\rho_2}{N} \sum_{t=1}^T \left( m_t - \frac{m}{N} \right) \| h_t \|_K^2 \right\}
\]
\[
\mathcal{F}(N, T, \rho_1, \rho_2) = \frac{\rho_1}{N} \sum_{t=1}^T \left( p(t) - \frac{m_t}{N} \right) \| f_1^t \|_K^2
\]
\[
+ \frac{\rho_2}{N} \sum_{t=1}^T \left( m_t - \frac{m}{N} \right) \| f_1^t \|_K^2
\]
\[
+ \frac{2 \rho_2}{N} \sum_{t=1}^T p(t) \| f_1^t \|_K^2 \sum_{t=1}^T \left( m_t - \frac{m}{N} \right) \| f_1^t \|_K^2.
\]

The proof of Lemma 2 can be found in the Appendix. Similar to [20], we refer to the first term \( S(N, T) \) as the average sample error and the second term \( D(N, T, \rho_1, \rho_2) \) as the average regularization error. It is worth noting that we have an additional term \( \mathcal{F}(N, T, \rho_1, \rho_2) \) in Lemma 2, compared to literature [20]. Here, we refer to it as the frequency error, because one of its dominant term, \( p(t) - (m_t/N) \), measures the sampling frequency error of each task. The frequency error vanishes in the single-task learning cases where the sampling frequency is deterministic, that is, \( m_t = N \). Therefore, the frequency error is a new knowledge associated with the MTL problems only.

To bound these three errors, we have the following lemmas.

**Lemma 3:** An upper bound of the average sample error can be given by \( S(N, T) = \mathcal{O}(N^{-1/4+\epsilon}) \) almost surely, where \( \epsilon \) is any positive constant.

**Lemma 4:** Assume that the Bayes rules for each task \( f_t^* \) \( (t = 1, \ldots, T) \) are restrictions of some functions \( \tilde{f}_t^* \) lying in the Sobolev space \( H^s(\mathbb{R}^d) \). Then, an upper bound of the average regularization error can be given by \( D(N, T, \rho_1, \rho_2) = \mathcal{O}((\log N)^{-s/4}) \) almost surely.

**Lemma 5:** An upper bound of the frequency error can be given by \( \mathcal{F}(N, T, \rho_1, \rho_2) = \mathcal{O}(N^{-3/2+\epsilon}) \) almost surely, where \( \epsilon \) is any positive constant. The proof of Lemmas 3–5 can be found in the Appendix. Combining the results of Theorem 3 and Lemmas 2–5, we can provide the main result of this section in the following theorem.

**Theorem 4:** Assume that \( f_t^* \) \( (t = 1, \ldots, T) \) are the restrictions of some functions \( \tilde{f}_t^* \) lying in the Sobolev space \( H^s(\mathbb{R}^d) \). Then, an upper bound of excess average misclassification error can be obtained almost surely by
\[
\mathcal{U}(\mathrm{sgn}(f_1^t), \ldots, \mathrm{sgn}(f_T^t)) - \mathcal{U}(f_1^*, \ldots, f_T^*) = \mathcal{O}\left((\log N)^{-s/4}\right).
\]

**Proof:** It follows from Theorem 3 and Lemmas 2–5.

**Corollary 1:** Almost surely, the misclassification error \( R_t(f_t^*) \) for each task \( t, (t = 1, \ldots, T) \), can be arbitrarily close to the corresponding Bayes error \( R_t(f_t^*) \), as long as \( N \) is sufficiently large.

Theorem 4 indicates that the average misclassification error \( \mathcal{U}(\mathrm{sgn}(f_1^t), \ldots, \mathrm{sgn}(f_T^t)) \) can be arbitrarily close to the average Bayes error \( \mathcal{U}(f_1^*, \ldots, f_T^*) \), as long as \( N \) is sufficiently large. Moreover, the above corollary shows that the learned classifier \( \mathrm{sgn}(f_t^*) \) for each task \( t \) converges to the corresponding Bayes rule \( f_t^* \), when \( N \) increases to infinity. Therefore, given a large data size, the M-SVM (3) induces a reliable classification rule for each task with a small misclassification error, which gives us great confidence in this method.

**B. Important Remarks**

1) **Task-Interaction Vanishes in the Limit of Large Sample Size:** Previous literature has shown the Bayes risk consistency of the single-task SVM [20]. Based on this fact, our main result in Theorem 4 implies that applying the M-SVM (3) amounts to learning \( T \) tasks independently, when the number of samples is sufficiently large. With this regard, in the MTL framework, the interaction between different tasks vanishes as the sample size grows greatly. Intuitively, we can observe this point through the objection function of the M-SVM (3), that the empirical loss dominates the regularization terms (i.e., the task-interaction terms) when the data size is large; therefore, the influence of the task interaction vanishes in the limit of large sample size.

This result implies that when the data size is large enough, the M-SVM (3) still makes the reliable decisions even if the tasks considered are significantly deviated from each other. This is quite different from the one in finite sample cases, where the task interaction cannot be negligible and plays a key role in improving the performance of each task. Moreover, as shown in Sections IV-B and IV-C1, the performance curves of the M-SVM and its single-task counterpart on the simulated and real datasets are initially separable and finally overlap, which supports this point again.

2) **Regularization Terms Are Necessary:** Note that the average regularization error \( D(N, T, \rho_1, \rho_2) \) disappears if \( \rho_1 = \rho_2 = 0 \). In this case, by the results of Lemma 2–5, it seems that a faster convergent rate \( \mathcal{O}(N^{-3/4+\epsilon}) \) determined by average sample error can be achieved. However, it is not the case. When \( \rho_1 = \rho_2 = 0 \), we will lose the upper bound estimation of \( \| f_t^* \|_K \) (c.f. Lemma 6 in the Appendix) and, therefore, fail to analyze the average sample error. To the best of our knowledge, there has been no literature succeed in estimating the sample error without the boundedness of \( \| f_t^* \|_K \). In this regard, regularization terms are necessary. In general, the average excess misclassification error bounds are determined by the complexity of the RKHS [20], [23], [34], [35]. Our experimental results in Sections IV-B and IV-C also show that, without the regularization term, the M-SVM classifier \( \mathrm{sgn}(f_t^*) \) for each task cannot converge to the corresponding Bayes rule \( f_t^* \). Therefore, it is important to include the regularization term in MTL.
3) What Benefits Does the M-SVM Bring? Remark III-B1 has shown the equivalence of the M-SVM and its single-task counterpart in terms of their limit classifiers’ performance. Meanwhile, both methods’ convergence rates have the same upper bound (\(O(\log N)^{-d/4}\)) (see Theorem 4 and [20]) determined by the complexity of the RKHS. Though obtaining the optimal convergence rate requires more elaborate analysis of the RKHS [20], we conjecture that both methods’ convergence rate are the same when the number of tasks is fixed, as the more elaborate analysis should be applicable for both cases. Consequently, we conjecture that the benefit of the M-SVM is the improvement of its PCR factor for each task, defined by \(C_M(m_t) = ([R(\text{sgn}(f_t^j)) - R(f_t^j)]/[F(m_t)])\), where \(1/F(m_t)\) is the M-SVM’s exact convergence rate with \(F(m_t) \rightarrow \infty (m_t \rightarrow \infty)\). Similarly, for each task, we define the PCR factor of the SVM and denote it by \(C_S(m_t)\). The following section gives the theoretical (in a 1-D least-square classification setting) and experimental justifications of this important claim. Our results demonstrate that, in a similar (non-similar)-task setting, there holds \(C_M(m_t) < C_S(m_t) (C_M(m_t) > C_S(m_t))\) when \(m_t\) is relatively small, and \(C_M(m_t) - C_S(m_t) = 0\) as \(m_t \rightarrow \infty\).

IV. CASE STUDIES

This section contains three case studies to verify our claims in Section III-B. Specifically, we first provide a theoretical verification of the improvement of the PCR factor for MTL in a 1-D least-square classification setting. Then, we provide the empirical results of the MTL models, including the M-SVM (with linear and Gaussian kernels), unregularized M-SVM, L21, LASSO, and SRMTL (the later three models can be found and solved by [28]) on the simulated data. Finally, we conduct experiments on the blast furnace and landmine datasets using the M-SVM, unregularized M-SVM, and other two MTL methods [26], [27] with linear kernel. The simulated and experimental results achieve an excellent agreement with the claims in Section III-B. Without loss of generality, we only consider MTL with two tasks and display its performance for the first task (we can obtain a similar result for the second task cause MTL treats both tasks equally).

A. Theoretical Justification of the PCR Factor

This section provides a theoretical guarantee of the PCR factor’s improvement in MTL on a 1-D classification problem. Explicitly, we analyze the ability of MTL originating from [19] based on the least-square classifiers due to its availability of the analytic solution.

We assume that data are sampled from a two-class Gaussian model with \(x_1|y_1=1 \sim \mathcal{N}(\mu_1^+, \Sigma^2_1)\), \(x_1|y_1=-1 \sim \mathcal{N}(\mu_1^-, \Sigma^2_1)\) for the first task, and \(x_2|y_2=1 \sim \mathcal{N}(\mu_2^+, \Sigma^2_2)\), \(x_2|y_2=-1 \sim \mathcal{N}(\mu_2^-, \Sigma^2_2)\) for the second task. For each task, we assume there exists no sampling bias of the positive and negative classes. We also assume \(\mu_1^+ > \mu_1^-\) and denote \(\delta_+ = \mu_2^+ - \mu_1^+\), \(\delta^- = \mu_2^- - \mu_1^-\). Based on the problem setting above, the decision boundaries for these two tasks can be expressed as \(w_1x + b_1 = 0\) and \(w_2x + b_2 = 0\), respectively. Without loss of generality, we can further suppose \(w_1 = w_2\), where \(w_1\) and \(w_2\) are two binary variables and only take two values \(-1\) or \(1\) (will be shown later).

Combining the MTL [19] with the least-square classifiers can naturally produce the multitask least-square classification model (referred to as “M-LSC”) in 1-D setting as follows:

\[
\min b_0, b_1, b_2, w_1, w_2 \left\{ \sum_{i=1}^{m_1} (y_i - w_1x_i - b_1)^2 + \frac{\lambda_1}{2} \sum_{i=1}^{m_2} y_i^2 \right\}
\]

where \(b_1 = b_0 + v_1\) and \(b_2 = b_0 + v_2\) are two parameters of the decision boundaries for the M-LSC classifiers.

By the method of Lagrange multipliers and the assumption that \(w_1 = w_2\), we can easily obtain the optimal solution of the M-LSC problem for the first task as:

\[
b_1^* = [(G - 1) / m_1] \sum_{i=1}^{m_1} w_1^* x_i + (G / m_2) \sum_{i=1}^{m_2} w_2^* x_i,
\]

\[
\begin{align*}
\delta_1^+ &= \delta_2^+ = \text{sgn}(\sum_{i=1}^{m_1} x_i^2 \mu_1^+ - \sum_{i=1}^{m_2} x_i^2 \mu_1^-), \\
G &= (\lambda_1 m_1 / [4 m_1 m_2 + \lambda_1 (m_1 + m_2)]).
\end{align*}
\]

Then, by the definition of the Bayes classifier and the problem setting above, we can also obtain the Bayes classifier for the first task as \(f_1^* = \text{sgn}(x - [\mu_1^+ + \mu_1^-]) / 2\).

Based on the above results, we can now estimate the excess misclassification error for the M-LSC estimate with respect to the first task \(f_1 = \text{sgn}(w_1^*(x - x^0))\), where \(x^0 = (-b_1^* / w_1^*)\). As will be shown below, we only need to estimate the MSE of \(x^0\).

Note that, for the fixed \(x^0\) and \(w_1^*\), there holds \(R(\text{sgn}(f_1)) = (1/2)(1 - w_1^*[\phi(x^0 - \mu_1^+) - \phi(x^0 - \mu_1^-)])\), where \(\phi()\) is the CDF of the standard Gaussian distribution. Then, we have

\[
R(\text{sgn}(f_1)) = \frac{1}{2} \left(1 - \left[\phi\left(\frac{\mu_1^+ - \mu_1^-}{2}\right) - \phi\left(\frac{\mu_1^--\mu_1^+}{2}\right)\right]\right).
\]

By Taylor’s theorem, and denoting \(h = x^0 - [(\mu_1^+ + \mu_1^-) / 2]\), we can check that for the \(x^0\) and \(w_1^*\) learned by M-LSC, there holds

\[
R(\text{sgn}(f_1)) - R(\text{sgn}(f_1^*)) = D \cdot \mathbb{E}(h)^2 + O(\mathbb{E}(h)^4) \tag{6}
\]

where

\[
D = \left[\phi''\left(\frac{\mu_1^+ - \mu_1^-}{2}\right) - \phi''\left(\frac{\mu_1^--\mu_1^+}{2}\right)\right] / 4.
\]

is a constant.

By some standard calculations, we can obtain that

\[
\mathbb{E}(h)^2 = \frac{(1 - G)^2 \delta_1^2}{m_1} + \frac{G^2 \delta_1^2}{4} + \frac{G^2 \delta_2^2}{m_2} \tag{7}
\]

where \(\delta = \delta^+ + \delta^-\) is the task dissimilarity. Note that we can degenerate M-LSC results to its single-task counterpart (S-LSC) by making \(\lambda_1 = 0\). Then, we see that the real convergence rates of M-LSC and S-LSC are both \(O(1/m)\) by assuming \(m_1 = m_2 = m\). According to the PCR factor’s definition in Section III-B3, the PCR factor of M-LSC is \(1 - G^2 \sigma_1^2 + G^2 \sigma_2^2 + G^2 \delta_2^2 / 4\), and \(\sigma_1^2\) for the S-LSC case. We describe the PCR factor as a function of data size in Fig. 1.

Our theoretical analysis above demonstrates that PCR factors are different in S-LSC and M-LSC cases. As shown in Fig. 1, M-LSC has a smaller PCR factor value when the data size is relatively small, whereas its PCR factor value converges to that of the S-LSC as \(m \rightarrow \infty\). This shows that the priority
of M-LSC lies in reducing the PCR factor when the data size is relatively small. Furthermore, the PCR factor also depends on the task dissimilarity \( \delta \) (see the left panel of Fig. 1); the less \( \delta \) is, the less the PCR factor will be. This occurs because more similar tasks contain more common knowledge that can be shared between tasks. Moreover, the PCR factor is related to the regularization parameter \( \lambda_1 \) (see the right panel of Fig. 1); particularly, M-LSC can be worse than S-LSC when \( \lambda_1 \) is not properly chosen. Therefore, the capacity of MTL also relies on the choice of tuning parameters.

We theoretically verified that in this case study, the benefit of MTL lies in the improvement of the PCR factor rather than the convergence rate. In what follows, we experiment on the simulated and real datasets to demonstrate that this conclusion universally holds in general MTL frameworks.

**B. Simulation Studies**

In this section, we first design two simulation studies of the M-SVM (one with Gaussian and one with linear kernels) to support our claims in Section III-B under two different task dissimilarity settings. Here, we fix the Gaussian kernel parameter as \( \lambda_1 = 1 \). To mimic the scenarios in real problems, the data are generated from multivariate distributions and linearly nonseparable, and the uncertainty of the data frequency is also considered. Then, we show that the simulation results of several other MTL methods including the L21, LASSO, and SRMTL (these models can be found and solved by [28]) are also consistent with the claims.

For a similar-task setting, we generate data for the first task from a two-class Gaussian model with \( \mathbf{x}_1 | (y_1 = 1) \sim \mathcal{N}(\mathbf{2}, \mathbf{1}) \) and \( \mathbf{x}_1 | (y_1 = -1) \sim \mathcal{N}(\mathbf{4}, \mathbf{1}) \). We generate data for the second task from the other two-class Gaussian model with \( \mathbf{x}_2 | (y_2 = 1) \sim \mathcal{N}(\mathbf{2}, \mathbf{1}) \) and \( \mathbf{x}_2 | (y_2 = -1) \sim \mathcal{N}(\mathbf{4}, \mathbf{1}) \). For the nonsimilar-task case, we generate data for the first task the same as the above, while we assume data for the second task are sampled from \( \mathbf{x}_2 | (y_2 = 1) \sim \mathcal{N}(\mathbf{2}, \mathbf{1}) \) and \( \mathbf{x}_2 | (y_2 = -1) \sim \mathcal{N}(\mathbf{4}, \mathbf{1}) \). We assume the probability of each sample coming out from task \( t \) is 0.5 (that is, \( m_t/N \approx 0.5 \) when \( N \) is large enough), for \( t = 1, 2 \).

We apply the single-task SVM (S-SVM), M-SVM, and unregularized M-SVM to the classification problems above. The excess error is summarized in Fig. 2 where each dot represents the average of 2000 random experiments based on 10,000 test samples. In all cases, we see that the log-log plots of the S-SVM and M-SVM are initially parallel, then they gradually get closer and finally overlap, indicating that they have the same convergence rate but different PCR factors.

For the unregularized M-SVM, that is the M-SVM with its regularization parameters being zero, we see that it cannot converge to the Bayes classifier in all scenarios. This is expected because, the unregularized M-SVM cannot leverage the dissimilarity of tasks, and it simply treats two different tasks as the same. This fact supports our discussions in Section III-B2.

To quantify the effects caused by the randomness of the sampling frequency in the multitask problem, we compute the absolute percentage performance difference (in terms of the excess error) for the M-SVM with/without this randomness in Tables I and II. We see that, in both cases, it has a small but non-negligible effect on the M-SVM’s performance, and it does not change the convergence rate (as the absolute percentage differences are relatively small when the data size is large enough). This also supports our theoretical result in Theorem 4.

We also verify our claim of Section III-B3 for several other MTL frameworks, including the L21, Lasso, and SRMTL (these models can be found and solved by [28]). Fig. 3 again shows that these MTL methods and their single-task counterparts have the same convergence rate (the log-log plots finally overlap) but the different PCR factors (the log-log plots

---

**Table I**

| Data size | EE-NRF\(^*\) | PPD-S(P\%) | PPD-S(P\%) |
|-----------|---------------|-------------|-------------|
| 200       | 0.0053        | 0.0032      | 0.0020      |
| 400       | 0.0012        | 0.0005      | 0.0008      |
| 800       | 0.0012        | 0.0005      | 0.0008      |
| 1600      | 0.0012        | 0.0005      | 0.0008      |
| 3200      | 0.0012        | 0.0005      | 0.0008      |

EE-NRF\(^*\): the excess error of M-SVM without the randomness of the sampling frequency.

PPD-S(P\%): the (absolute) percentage performance difference of the M-SVM with/without the randomness of the sampling frequency. We define it as the excess error difference of these two models as a percentage of the excess error of the M-SVM. The results are computed in the similar-task setting with \( \lambda_1 = 20C, \lambda_2 = C(C = 1/2) \).

---

**Table II**

| Data size | EE-NRF\(^*\) | PPD-S(P\%) |
|-----------|---------------|-------------|
| 70        | 0.0061        | 0.0038      |
| 120       | 0.0023        | 0.0014      |
| 200       | 0.0008        | 0.0008      |
| 350       | 1.47%         | 0.48%       |
| 600       | 3.15%         | 4.86%       |

EE-NRF\(^*\): the excess error of M-SVM without the randomness of the sampling frequency.

PPD-S(P\%): the (absolute) percentage performance difference of the M-SVM with/without the randomness of the sampling frequency. We define it as the excess error difference of these two models as a percentage of the excess error of the M-SVM. The results are computed in the similar-task setting with \( \lambda_1 = 20C, \lambda_2 = C(C = 1/2) \).
Fig. 2. Log–log plot of the excess misclassification error (excess error) to the training data size on the simulated dataset using the S-SVM (with the regularization parameter $C = 1/2$), M-SVM, and unregularized M-SVM in the presence of the randomness in the sampling frequency. For rigorous comparisons, we fixed one of the regularization parameters ($\lambda_1 = 2C$ or $\lambda_2 = C$) for the M-SVM in each case to make it comparable with the S-SVM. The plots of these two models are parallel initially, then get closer gradually and overlap finally, indicating that they have the same convergence rate but different PCR factors. Without the regularization terms, the unregularized M-SVM classifier cannot converge to the corresponding Bayes rule.

Fig. 3. Log–log plot of the excess error for the MTL methods (red) including the L21, Lasso, SRMTL, and their individual counterparts (blue) in a similar-task setting. The regularization parameters for these models are 0.1, 0.4, and 0.4 (the graph regularization parameter is 0.01), respectively. The improvement of PCR factors in MTL is more significant when the data size is relatively small. These observations imply that our claims in Sections III-B1 and III-B3 hold in the general MTL frameworks.

C. Real Data Applications

This section verifies our claims in Section III-B by applying the S-SVM, M-SVM, and unregularized M-SVM to two real datasets, including blast furnace\textsuperscript{1} and landmine\textsuperscript{2} datasets. For illustrating the generality of our claims, we also test other two recently published MTL methods in these datasets: 1) M-TSVM\textsuperscript{26} and 2) M-CSVM\textsuperscript{27}. The single-task versions of these two MTL methods are the least-square twin SVM (abbreviated as “S-TSVM”) and the L1-SVM (referred

\textsuperscript{1}These two blast furnace datasets are contained in https://github.com/Shahan-Chen/Transfer-learning-in-Mallows-Cp.

\textsuperscript{2}http://people.ee.duke.edu/ lcarin/LandmineData.zip.
TABLE III
INPUT VARIABLES FOR BLAST FURNACES

| Variable name [Unit] | Symbol | Input variable |
|----------------------|--------|----------------|
| Blast volume [m³/min] | \(x_1\) | \(q^{-1}, q^{-2}, q^{-3}, q^{-4}\) |
| Blast temperature [°C] | \(x_2\) | \(q^{-1}, q^{-2}, q^{-3}, q^{-4}\) |
| Feed speed [mm/h] | \(x_3\) | \(q^{-1}, q^{-2}, q^{-3}, q^{-4}\) |
| Gas permeability [m³/min kPa] | \(x_4\) | \(q^{-1}, q^{-2}, q^{-3}, q^{-4}\) |
| Pulverized coal injection [ton] | \(x_5\) | \(q^{-1}, q^{-2}, q^{-3}, q^{-4}\) |
| Silicon content [wt%] | \(x_6\) | \(q^{-1}\) |

\(q^{-1}, q^{-2}, q^{-3}, q^{-4}\) represent delay operators, e.g., \(q^{-1}x(t) = x(t-1)\).

to as “S-CSVM”), respectively. We employ the linear kernel in these experiments.

1) Experiments on the Blast Furnace Dataset: We first verify our results on blast furnace data. The dataset are collected from two typical Chinese blast furnaces with the inner volume of about 2500 and 750 m³, referred to as blast furnace (a) and (b), respectively. Table III lists the features that are relevant for predicting the silicon class labels for these two blast furnaces. We labeled the records satisfying \((x_6 \leq 0.4132)/(x_6 \leq 0.3736)\) for furnace (a)/(b) as \(-1\), and \(+1\) otherwise [36], [37].

Due to the (2–8 h) time delay for furnace outputs to respond to inputs, we also treat four lagged terms for the first five features as inputs. Furnaces (a) and (b) have 794 and 800 samples, respectively.

We apply the M-SVM, unregularized M-SVM, M-TSVM, M-CSVM and their single-task versions to the silicon classification problems. The misclassification errors (the Bayes classifier is unknown here) of these methods are summarized in Figs. 4–6, where each dot represents the average of 10 000 random splits of the datasets with \(m_1 = 100, 110, 120, \ldots , 490\) samples as the training set and the remaining 300 ones as the test set. For each partition, we normalize the variables of training samples to zero mean and unit variance, while the test samples are normalized accordingly.

From Fig. 4, we see that the performance curves of the S-SVM and M-SVM are separable initially, then they get closer gradually and overlap finally, indicating that they have the same convergence rate but different PCR factors. Moreover, the improvement of PCR factors in MTL is more significant when the data size is relatively small. This phenomenon excellently agrees with the results of the simulation studies above. In addition, Fig. 4 shows that the unregularized M-SVM initially outperforms the M-SVM, due to the intrinsic similarity of these two furnaces. However, the unregularized M-SVM performs worse than the S-SVM and M-SVM when the training data size is relatively large. This occurs because the unregularized M-SVM classifier cannot converge to the corresponding Bayes rule.

Fig. 5 shows the misclassification error plots for the S-TSVM and M-TSVM on blast furnace data. We see that as the data size grows, the performance curves of these two models are separable initially and overlap finally. This observation again demonstrates that the S-TSVM and M-TSVM have the same convergence rate but different PCR factors. A similar observation holds for the S-CSVM and M-CSVM methods on blast furnace data (see Fig. 6). In summary, our experiments of the three types of MTL methods on blast furnace data show a consistent result that the benefit of the MTL methods lies in improving the PCR factors rather than the convergence rate.

2) Experiments on the Landmine Dataset: We further verify our claims in Section III-B using the landmine dataset. This dataset detects whether or not a specific region presents a land mine based on radar images. Each sample is represented by a 9-D feature vector and the corresponding binary label (1 for landmine and \(-1\) for clutter). This dataset contains 14 820 samples divided into 29 classification tasks according to the corresponding specific geographical regions. Due to space limitations, we choose the 1st and 2nd tasks as representatives...
to demonstrate the advantage of MTL over single-task learning. To reduce the imbalance of this dataset, we replicate the positive samples twice during the training process.

We apply the M-SVM, unregularized M-SVM, M-TSVM, M-CSVM, and their single-task versions to the landmine detection problems. Following the same experimental setup as the previous section, we quantify the performance of these methods by calculating the average misclassification error across 10,000 random splits of the datasets. Figs. 7–9 compare the aforementioned MTL methods' performances with the corresponding single-task versions under different training sample sizes.

We can observe from Fig. 7 that, as the data size grows, the performance gap between the M-SVM and S-SVM shrinks gradually, implying that these two models have the same convergence rate but different PCR factors. Furthermore, we find that the M-SVM and S-SVM performance curves do not precisely overlap at the last point due to the limited samples. However, this clear shrinking tendency still supports our claims. Finally, Fig. 7 shows that, without the regularization terms, the unregularized M-SVM classifier cannot converge to the corresponding Bayes rule. Similar observations hold for the M-TSVM and M-CSVM problems on the landmine dataset (see Figs. 8 and 9). To conclude, the experimental results of three different types of MTL methods on the landmine dataset further support our claim that the advantage of MTL methods is to improve the PCR factors rather than the convergence rate.

V. Discussion

There are a number of future works we are currently pursuing related to the error analysis of the regularized MTL.

We provided both the theoretical (in a 1-D least-square classification setting) and experimental justification to support our claim that the benefit of MTL is the improvement the PCR factor rather than the convergence rate. Based on the theoretical results of this simple case, we also discussed the conditions under which the MTL technique originating from [19] can improve the PCR factor. It would be interesting to theoretically verify this claim in more general MTL frameworks (e.g., the MTL methods listed in introduction), which requires a more elaborate analysis of the central limit property of the learning methods. We leave this issue for future work.

Another interesting question is whether we can still obtain optimal asymptotic performance (i.e., Bayes risk consistency) in the MTL setting discussed in this article by assuming there are some common structures of functional space $\mathcal{H}$ for each task. Maurer et al. [16] considered a similar problem in the context of an unregularized MTL problem where a common feature representation of different tasks is shared. They also
where we enforce the structural information to be shared between the MTL can achieve optimal performance. It is unknown whether sharing a common feature representation in derived the error upper bound for this learning method. It is shown that the excellent performance of MTL requires a synergy between task dissimilarity and model parameters.

In this work, we analyzed the M-SVM’s asymptotic performance by focusing on the asymptotic behaviors of the learned classifier in terms of the excess misclassification error. We first generalized the regularized MTL model [19] by introducing the kernel map and the uncertainty of data frequency of tasks into it. Then, we showed the upper bound of the excess misclassification error to be $O((\log N)^{-s/4})$ almost surely. This result demonstrated that the regularized MTL framework can produce reliable classification rules when the sample size goes to infinity. Furthermore, we found that the interaction of tasks vanishes as the data size goes to infinity, and the convergence rates of the M-SVM and its single-task counterpart have the same upper bound. The former suggested that M-SVM cannot improve the limit classifier’s performance; based on the latter fact, we raised the conjecture that the optimal convergence rate is not improved either when the number of tasks is fixed. As a novel insight into MTL, our theoretical and experimental results achieved an excellent agreement that the benefit of M-SVM lies in improving the PCR factor (denoted as the ratio between the excess misclassification error and its real convergence rate in Section III) rather than the convergence rate. This improvement of PCR factors is more significant when the data size is small. Moreover, our experimental results of several other MTL methods, including L21, Lasso, SRMTL, M-TSVM [26], and M-CSVM [27] (the first three models can be found and solved by [28]) also demonstrated the generality of this new insight in MTL. Therefore, PCR factor is more suitable and accurate to depict the essential advantages of MTL when the task number is fixed.

VI. CONCLUSION

In this work, we analyzed the M-SVM’s asymptotic performance by focusing on the asymptotic behaviors of the learned classifier in terms of the excess misclassification error. We first generalized the regularized MTL model [19] by introducing the kernel map and the uncertainty of data frequency of tasks into it. Then, we showed the upper bound of the excess misclassification error to be $O((\log N)^{-s/4})$ almost surely. This result demonstrated that the regularized MTL framework can produce reliable classification rules when the sample size goes to infinity. Furthermore, we found that the interaction of tasks vanishes as the data size goes to infinity, and the convergence rates of the M-SVM and its single-task counterpart have the same upper bound. The former suggested that M-SVM cannot improve the limit classifier’s performance; based on the latter fact, we raised the conjecture that the optimal convergence rate is not improved either when the number of tasks is fixed.

As a novel insight into MTL, our theoretical and experimental results achieved an excellent agreement that the benefit of M-SVM lies in improving the PCR factor (denoted as the ratio between the excess misclassification error and its real convergence rate in Section III) rather than the convergence rate. This improvement of PCR factors is more significant when the data size is small. Moreover, our experimental results of several other MTL methods, including L21, Lasso, SRMTL, M-TSVM [26], and M-CSVM [27] (the first three models can be found and solved by [28]) also demonstrated the generality of this new insight in MTL. Therefore, PCR factor is more suitable and accurate to depict the essential advantages of MTL when the task number is fixed.

APPENDIX

A. PROOFS

**Proof of Lemma 2:** Note that

$$
\mathcal{E}(f_1^*, \ldots, f_T^*) - \mathcal{E}(f_1^*, \ldots, f_T^*) \\
= \{ \mathcal{E}(f_1^*, \ldots, f_T^*) - \mathcal{E}(f_1^*, \ldots, f_T^*) \} \\
+ \{ \mathcal{E}(f_1^*, \ldots, f_T^*) - \mathcal{E}(f_1^*, \ldots, f_T^*) \} \\
+ \frac{\rho_2}{N} \sum_{t=1}^T \left( m_t \left\| f_t^* \right\|^2_K \right) \\
= \frac{\rho_1}{N} \sum_{t=1}^T \left( m_t \left\| f_t^* \right\|^2_K \right) \\
+ \frac{\rho_2}{N} \sum_{t=1}^T \left( m_t \left\| f_t^* - \sum_{s=1}^T p(s) f_s^* \right\|^2_K \right) \\
= \frac{\rho_1}{N} \sum_{t=1}^T p(t) \left\| f_t^* \right\|^2_K \\
+ \frac{\rho_2}{N} \sum_{t=1}^T p(t) \left\| f_t^* - \sum_{s=1}^T p(s) f_s^* \right\|^2_K \\
(8)
$$

Another attractive problem is generalizing our analysis framework to the regularized online MTL problems, where the large-scale data arrive sequentially. Especially, it is interesting to analyze the stochastic generalization error bounds in terms of the step sizes and the approximation property of functional space when the various regularization terms are included in the functional iterations to model different task relations.

Officials have been accused of using their influence to achieve the desired outcomes.
Obviously lies in $H$.

Denote the right-hand side of above inequality as
\[
\sum_{i=1}^{m_t} \|f_i\|_{K}^2.
\]

Moreover, (11) is no greater than
\[
\left\{ E_{\varepsilon}(f_1^H, \ldots, f_T^H) + \frac{\rho_1}{N} \sum_{i=1}^{T} \|f_i\|_{K}^2 \right\} \leq \left[ \frac{\rho_2}{N} \sum_{i=1}^{T} \|f_i\|_{K}^2 \right] \cdot \left[ \frac{\rho_1}{N} \sum_{i=1}^{T} \|f_i\|_{K}^2 \right] \leq \left[ \frac{\rho_2}{N} \sum_{i=1}^{T} \|f_i\|_{K}^2 \right] + \frac{\rho_2}{N} \sum_{i=1}^{T} \left( \sum_{t=1}^{T} \|f_t\|_{K}^2 \right) \geq \left[ \frac{\rho_2}{N} \sum_{i=1}^{T} \|f_i\|_{K}^2 \right] + \frac{\rho_2}{N} \sum_{i=1}^{T} \left( \sum_{t=1}^{T} \|f_t\|_{K}^2 \right).
\]

Combining the above inequality with (10), we obtained
\[
(10) + (11) \leq -\frac{\rho_1}{N} \sum_{i=1}^{T} \left( \sum_{t=1}^{T} \|f_t\|_{K}^2 \right) \leq \left[ \frac{\rho_2}{N} \sum_{i=1}^{T} \|f_i\|_{K}^2 \right] + \frac{\rho_2}{N} \sum_{i=1}^{T} \left( \sum_{t=1}^{T} \|f_t\|_{K}^2 \right).
\]

Denote the right-hand side of above inequality as $\mathcal{F}(N, T, \rho_1, \rho_2)$, $\mathcal{S}(N, T) = (9) + (12)$, and $\mathcal{D}(N, T, \rho_1, \rho_2) = (13)$, we obtain the result.

Before showing the proof of Lemma 3, we first provide some auxiliary lemmas which will be helpful for the later proof. The first one bounds the minimizer of M-SVM and its average empirical loss.

**Lemma 6:** For functions $\ell(\cdot)$ and $f_t^i$, we have:
1. $\sum_{i=1}^{m} \ell(\{y_{i,t} \cdot f_t^i(x_i)\}) \leq (N/m_t)$;
2. $\|f_t^i\|_{K} \leq (N/\sqrt{\rho_1 m_t})$.

**Proof:** Let us define an auxiliary function $\tilde{f}(\cdot) \equiv 0$ which obviously lies in $H_K$. Note that $\{f_t^i\}$ minimize the objective function (3). Therefore, we have the following equation:
\[
\sum_{i=1}^{m} \ell(\{y_{i,t} \cdot f_t^i(x_i)\}) + \rho_1 \sum_{i=1}^{m} \|f_t^i\|_{K}^2.
\]

Since the left-hand side of above inequality is greater than or equals $\sum_{i=1}^{m} \ell(\{y_{i,t} \cdot f_t^i(x_i)\})$, where $t = 1, \ldots, T$, and the right-hand side of the inequality equals $N$, we have the relation
\[
\sum_{t=1}^{T} \|f_t\|_{K}^2 \leq \frac{N}{\rho_1 m_t}.
\]

Similarly, since $\rho_1 (m_t/N) \|f_t^i\|_{K}^2$ is also less than or equal to the left-hand side of the first inequality in this lemma, therefore, we have
\[
\rho_1 (m_t/N) \|f_t^i\|_{K}^2 \leq \frac{N}{\sqrt{\rho_1 m_t}}.
\]

Thus, the result is shown.

The second auxiliary lemma bounds the sampling error as $O(1/\sqrt{N})$.

**Lemma 7:** For a sequence of bounded, independent, and identically distributed random variables $\{X_i\}$, there is
\[
\sum_{i=1}^{N} X_i - E X_1 = O(N^{-\frac{1}{2} + \epsilon}) \text{ a.s.}
\]

where $\epsilon$ is any positive constant.

**Proof:** It can be proven directly by applying the Borel–Carnelli lemma to Hoeffding’s inequality.

The third auxiliary lemma bounds the generalization error of the M-SVM classifier.

**Lemma 8:** For functions $\ell(\cdot)$ and $f_t^i$, we have
\[
E_{\rho_1} \left[ \ell(Y, f_t^i(X)) \right] - \frac{\sum_{i=1}^{m_t} \ell(y_{i,t} \cdot f_t^i(x_i))}{m_t} \leq O(N^{-\frac{1}{2} + \epsilon}).
\]

**Proof:** We first denote some notations as follows:
\[
\tilde{E}_t(f) := E_{(X,Y)} \rho_1 \left[ \ell(Y, f_t^i(X)) \right] \quad \text{and} \quad \tilde{E}_t^i(f) := \sum_{i=1}^{m_t} \frac{\ell(y_{i,t} \cdot f_t^i(x_i))}{m_t}.
\]

Let $R_t := \{f \in H_K \mid \|f\|_K \leq R\}$. Together with the fact that $\tilde{E}_t^i(f) \leq (N/m_t)$ (c.f., Theorem 6), for any $\alpha > 0$ we have
\[
P \left\{ \tilde{E}_t^i(f_t^i) > 4\alpha \right\} \leq \frac{\sum_{i=1}^{m_t} \ell(y_{i,t} \cdot f_t^i(x_i))}{m_t} \leq \frac{\sum_{i=1}^{m_t} \ell(y_{i,t} \cdot f_t^i(x_i))}{m_t} \leq \frac{N}{R_t^\alpha} \exp \left\{ -\frac{m_t \cdot \alpha^2}{32(1 + \kappa R_t^\alpha)} \right\}
\]

where the last inequality follows from [20, Lemma 5], and $N/(\alpha R_t^\alpha)$ is the covering number defined as the minimal
number of balls with radius \( \alpha/R^* \) to cover the unite ball in RKHS. According to [34], the covering number has the relation

\[
\log N(\varepsilon) \leq M \left( \frac{1}{\varepsilon} \right)^{d+1}
\]

where \( M \) is a constant, and \( d \) is the dimension of the dataset space. By plugging the above inequality and the error, we can obtain

\[
P \left\{ \tilde{E}_i(f_i^t) - \tilde{E}_i^2(f_i^t) > 4 \left( 1 + \frac{N}{m_t} \right) \cdot \alpha^*(N, m_t) \mid m_t \right\}
\]

\[
\leq \exp \left\{ M \left( \frac{N}{\sqrt{p_1 + p_1 \varepsilon^2}} \right)^{d+1} - \frac{1}{32} N^{2\varepsilon} \right\}
\]

(15)

where \( \varepsilon \) is arbitrary positive constant. Furthermore, by the above inequality and some easy calculations, we can obtain

\[
P \left\{ \tilde{E}_i(f_i^t) - \tilde{E}_i^2(f_i^t) > 4 \left( 1 + \frac{N}{m_t} \right) \cdot \alpha^*(N, m_t) \mid m_t \right\}
\]

\[
= \mathbb{E}_{m_t} \left[ P \left\{ \tilde{E}_i(f_i^t) - \tilde{E}_i^2(f_i^t) > 4\alpha \left( 1 + \frac{N}{m_t} \right) \mid m_t \right\} \right]
\]

\[
\leq \mathbb{E}_{m_t} \left[ \exp \left\{ M \left( \frac{N^{1-\varepsilon}}{\sqrt{p_1}} \right)^{d+1} - \frac{1}{32} N^{2\varepsilon} \right\} \right]
\]

\[
= o \left( N^{-2} \right)
\]

and, therefore

\[
\sum_{N=1}^{\infty} P \left\{ \tilde{E}_i(f_i^t) - \tilde{E}_i^2(f_i^t) > 4 \left( 1 + \frac{N}{m_t} \right) \cdot \alpha^*(N, m_t) \mid m_t \right\}
\]

\[
< +\infty.
\]

By the Borel–Cantelli Lemma, we have the relation

\[
\limsup_{N \to \infty} \tilde{E}_i(f_i^t) - \tilde{E}_i^2(f_i^t) \leq \mathcal{O}(N^{-1/\varepsilon})
\]

Besides, according to the definition of \( \alpha^*(\cdot, \cdot) \) and the strong law of large numbers, we have \( (1 + \{N/m_t\}) \cdot \alpha^*(N, m_t) = \mathcal{O}(N^{-1/\varepsilon}) \). Thus, \( \tilde{E}_i(f_i^t) - \tilde{E}_i^2(f_i^t) \leq \mathcal{O}(N^{-1/\varepsilon}) \) which proves the result.

With these preparations, we can prove Lemma 3.

**Proof of Lemma 3:** We first consider the first part of sample error, \( \mathcal{E}(f_i^t, \ldots, f_T^t) - \mathcal{E}(f_i^t, \ldots, f_T^t) \). By the definition and some easy calculations, we can obtain

\[
\mathcal{E}(f_i^t, \ldots, f_T^t) - \mathcal{E}(f_i^t, \ldots, f_T^t)
\]

\[
= \sum_{t=1}^{T} p(t) \left[ \mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right] - \frac{m_t}{m_t} \mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right] \right]
\]

\[
+ \sum_{t=1}^{T} \left( p(t) - \frac{m_t}{N} \right) \sum_{i=1}^{m_t} \frac{\mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right]}{m_t} \right].
\]

(16)

Since, for each task \( t \), the random variable \( m_t \) is a counting process with bounded and independent increments, therefore, by Lemma 7, there holds

\[
p(t) - \frac{m_t}{N} = \mathbb{E} \left( m_t \right) - \frac{m_t}{N} = \mathcal{O} \left( N^{-1/\varepsilon} \right)
\]

(17)

By plugging the above relation and Lemma 8 into (16), this relation leads to

\[
\mathcal{E}(f_i^t, \ldots, f_T^t) - \mathcal{E}(f_i^t, \ldots, f_T^t) = \mathcal{O} \left( N^{-1/\varepsilon} \right)
\]

Then, we consider the second part of sample error, \( \mathcal{E}(f_i^t, \ldots, f_T^t) - \mathcal{E}(f_i^t, \ldots, f_T^t) \). Still by the definition and some easy calculations, we can obtain

\[
\mathcal{E}(f_i^t, \ldots, f_T^t) - \mathcal{E}(f_i^t, \ldots, f_T^t)
\]

\[
= \sum_{t=1}^{T} \left[ \mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right] \right]
\]

\[
+ \sum_{t=1}^{T} \left( p(t) - \frac{m_t}{N} \right) \sum_{i=1}^{m_t} \frac{\mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right]}{m_t} \right].
\]

(19)

Moreover, since \( \mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right] \) is also independent and identically distributed. Together with the boundedness of \( \ell(Y, f_i^t(X)) \) and \( \|f_i^t\|_K \) is finite, for each data sample \( \ell(Y, f_i^t(X)) \) is uniformly bounded by a constant \( R \) where \( R = 1 + \kappa \|f_i^t\|_K \). Therefore, by (17) and the uniform boundedness of \( \ell(Y, f_i^t(X)) \), we have

\[
\sum_{t=1}^{T} \mathbb{E}_{(X, Y) \sim P_X} \left[ \ell(Y, f_i^t(X)) \right]
\]

\[
\leq \sum_{t=1}^{T} \left( p(t) - \frac{m_t}{N} \right) \sum_{i=1}^{m_t} \frac{R}{m_t} = \mathcal{O} \left( N^{-1/\varepsilon} \right)
\]

(20)

Besides, due to the independent and identically distributed (i.i.d.) property of sampled data \( (x_i, y_i) \) for each task \( t \), the random variable \( \ell(Y, f_i^t(X)) \) is also independent and identically distributed. With these preparations, we can prove Lemma 3.
simplicity, we denote $\|h_t - f_t\|_{L^2_P} := (\mathbb{E}(X_t, Y_t) - p_t h_t(X_t) - f_t^*(X_t))^2)^{1/2}$. Through plugging the above last inequality to $D(N, T, \rho_1, \rho_2)$, we have

$$D(N, T, \rho_1, \rho_2) \leq \inf_{h_1, \ldots, h_T \in \mathcal{H}_K} \left\{ \sum_{t=1}^T p(t) \|h_t - f_t^*\|_{L^2_P}^2 + \frac{\rho_1^2}{N} \sum_{t=1}^T p(t) \|f_t^*\|_K^2 \right\}$$

$$+ \frac{\rho_2^2}{N} \sum_{t=1}^T p(t) \left\{ \sum_{s=1}^T p(s) h_s \right\}^2 + \frac{\rho_1^2 N^2}{N} + \frac{2 \rho_2 R^2}{N}$$

where the last inequality follows immediately from [35] and [41]. In the above, $C_0$ and $s$ are two positive constants, while $C_s$ is another constant depends on $s$. By choosing $R = N^{1/2-\varepsilon}$ where $\varepsilon > 0$, we have $D(N, T, \rho_1, \rho_2) \leq C_0 C_s ((1/2 - \varepsilon) \log N)^{-3/4} + \rho_1 N^{-2} + \rho_2 N^{-\varepsilon} = O((\log N)^{-3/4})$, which proves the result.

**Proof of Lemma 5:** By the definition and some easy calculation, we obtain

$$\mathcal{F}(N, T, \rho_1, \rho_2) \leq -\frac{\rho_1}{N} \sum_{t=1}^T \left( p(t) - \frac{m_t}{N} \right) \|f_t^*\|_K^2$$

$$+ \frac{\rho_2^2}{N} \sum_{t=1}^T p(t) \left( \frac{m_t}{N} \right) \left\{ \sum_{s=1}^T p(s) + \|f_t^*\|_K \right\}^2$$

$$+ \frac{2 \rho_2}{N} \sum_{t=1}^T \left( p(t) - \frac{m_t}{N} \right) \cdot \|f_t^*\|_K$$

$$\times \left( \sum_{t=1}^T \|f_t^*\|_K + \sum_{t=1}^T \|f_t^*\|_K \right)$$

where the terms $(\rho_1/N)$, $(\rho_2/N)$, and $(p(t) - (m_t/N))$ determine the convergent rate of the frequency error. Since, for each task $t$, the random variable $m_t$ is a counting process with bounded and independent increments, therefore, by Lemma 7, there holds $p(t) - (m_t/N) = \mathbb{E}(m_t/N) - (m_t/N) = O(N^{-(1/2)+\varepsilon})$ a.s. Therefore, $\mathcal{F}(T, N, \rho_1, \rho_2) \leq O(N^{-(3/2)+\varepsilon})$ a.s., which shows the result.

**REFERENCES**

[1] R. Caruana, “Multitask learning,” *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[2] Y. Chen, S. Song, S. Li, L. Yang, and C. Wu, “Domain space transfer extreme learning machine for domain adaptation,” *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1909–1922, May 2019.

[3] D. Li, Z. Gong, and D. Zhang, “A common topic transfer learning model for crossing city POI recommendations,” *IEEE Trans. Cybern.*, vol. 49, no. 12, pp. 4282–4295, Dec. 2019.

[4] D. Wang et al., “Softly associative transfer learning for cross-domain classification,” *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4709–4721, Nov. 2020.

[5] J. Wang, Q. Wang, H. Zhang, J. Chen, S. Wang, and D. Shen, “Sparse multitask-centralized ensemble learning for ASD diagnosis based on age-and sex-related functional connectivity patterns,” *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 3141–3154, Aug. 2018.

[6] X. Gu, F.-L. Chung, H. Ishibuchi, and S. Wang, “Multitask coupled logistic regression and its fast implementation for large multitask datasets,” *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1953–1966, Sep. 2015.

[7] Y. Jiang, Z. Deng, F.-L. Chung, and S. Wang, “Multi-task TSK fuzzy system modeling using inter-task correlation information,” *Inf. Sci.*, vol. 298, pp. 512–533, Mar. 2015.

[8] Y. Jiang, Z. Deng, K.-S. Choi, F.-L. Chung, and S. Wang, “A novel multi-task TSK fuzzy classifier and its enhanced version for labeling-risk-aware multi-task classification,” *Inf. Sci.*, vol. 357, pp. 39–60, Aug. 2016.

[9] T. Evgeniou, C. A. Micchelli, M. Pontil, and J. Shawe-Taylor, “Learning multiple tasks with kernel methods,” *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 615–637, 2005.

[10] K. Yu, V. Tresp, and A. Schwaighofer, “Learning Gaussian processes from multiple tasks,” in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 1012–1019.

[11] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 31, 2021, doi: 10.1109/TKDE.2021.3070203.

[12] R. K. Ando, T. Zhang, and P. Bartlett, “A framework for learning predictive structures from multiple tasks and unlabeled data,” *J. Mach. Learn. Res.*, vol. 6, no. 11, pp. 1817–1853, 2005.

[13] J. Baxter, “Learning internal representations,” in *Proc. 8th Annu. Conf. Comput. Learn. Theory*, 1995, pp. 311–320.

[14] J. Baxter, “A model of inductive bias learning,” *J. Artif. Intell. Res.*, vol. 12, no. 1, pp. 149–198, 2000.

[15] J. Baxter, “A Bayesian/information theoretic model of learning to learn via multiple task sampling,” *Mach. Learn.*, vol. 28, no. 1, pp. 7–39, 1997.

[16] A. Maurer, M. Pontil, and B. Romera-Paredes, “The benefit of multitask representation learning,” *J. Mach. Learn. Res.*, vol. 17, no. 81, pp. 1–32, 2016.

[17] T. Liu, D. Tao, M. Song, and S. J. Maybank, “Algorithm-dependent generalization bounds for multi-task learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 227–241, Feb. 2017.

[18] C. Zhang, D. Tao, T. Hu, and B. Liu, “Generalization bounds of multi-task learning from perspective of vector-valued function learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1906–1919, 2020.

[19] T. Evgeniou and M. Pontil, “Regularized multi-task learning,” in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2004, pp. 109–117.

[20] Q. Wu and D.-X. Zhou, “Analysis of support vector machine classification,” *J. Comput. Anal. Appl.*, vol. 8, no. 2, p. 6, 2006.

[21] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou, “Support vector machine soft margin classifiers: Error analysis,” *J. Mach. Learn. Res.*, vol. 5, pp. 1143–1175, Sep. 2004.

[22] Y. Ying and D.-X. Zhou, “Online regularized classification algorithms,” *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[23] Y. Ying and D.-X. Zhou, “Learnability of Gaussians with flexible variances,” *J. Mach. Learn. Res.*, vol. 8, pp. 249–276 Feb. 2007.

[24] J. Lin, L. Rosasco, and D.-X. Zhou, “Iterative regularization for learning with convex loss functions,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2718–2755, 2016.

[25] Y. Fan, S. Lu, Y. Ying, and B.-G. Hu, “Learning with average top-k loss,” 2017, arXiv:1705.08826.

[26] B. Mei and Y. Xu, “Multi-task least squares twin support vector machine for classification,” *Neurocomputing*, vol. 338, pp. 26–33, Apr. 2019.

[27] C. Ruiz, C. M. Alaz, and J. R. Dorransoro, “Convex formulation for multi-task L1, L2, and LS-SVMs,” *Neurocomputing*, vol. 456, pp. 599–608, Oct. 2021.

[28] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-Task Learning via Structural Regularization*, Arizona State Univ., Tempe, AZ, USA, 2011.
[29] L. Devroye, L. Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition, vol. 31. New York, NY, USA: Springer, 2013.

[30] I. Steinwart and A. Christmann, Support Vector Machines. New York, NY, USA: Springer, 2008.

[31] Q. Wu, Y. Ying, and D.-X. Zhou, “Multi-kernel regularized classifiers,” J. Complexity, vol. 23, no. 1, pp. 108–134, 2007.

[32] Y. Lin, Y. Lee, and G. Wahba, “Support vector machines for classification in nonstandard situations,” Mach. Learn., vol. 46, no. 1, pp. 191–202, 2002.

[33] T. Zhang, “Statistical behavior and consistency of classification methods based on convex risk minimization,” Ann. Stat., vol. 32, no. 1, pp. 56–85, 2004.

[34] D.-X. Zhou, “The covering number in learning theory,” J. Complexity, vol. 18, no. 3, pp. 739–767, 2002.

[35] D.-X. Zhou, Density Problem and Approximation Error in Learning Theory, Hindawi, London, U.K., 2003.

[36] S. Chen and C. Gao, “Linear priors mined and integrated for transparency of blast furnace black-box SVM model,” IEEE Trans. Ind. Informat., vol. 16, no. 6, pp. 3862–3870, Jun. 2020.

[37] C. Gao, Q. Ge, and L. Jian, “Rule extraction from fuzzy-based blast furnace SVM multiclassifier for decision-making,” IEEE Trans. Fuzzy Syst., vol. 22, no. 3, pp. 586–596, Jun. 2014.

[38] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27, 2011.

[39] Y. Zhang, “Parallel multi-task learning,” in Proc. IEEE Int. Conf. Data Min., 2015, pp. 629–638.

[40] S. Chen, N. V. Sahinidis, and C. Gao, “Transfer learning in information criteria-based feature selection,” J. Mach. Learn. Res., vol. 23, no. 134, pp. 1–105, 2022. [Online]. Available: http://jmlr.org/papers/v23/19-816.html

[41] S. Smale and D.-X. Zhou, “Estimating the approximation error in learning theory,” Anal. Appl., vol. 1, no. 1, pp. 17–41, 2003.

Shaohan Chen received the B.Sc. degree in mathematics and applied mathematics from Jimei University, Xiamen, China, in 2014. She is currently pursuing the Ph.D. degree in operational research and cybernetics with Zhejiang University, Hangzhou, China.

Her research interests are in the areas of machine-learning models design and interpretations.

Zhou Fang received the B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 2014 and 2019, respectively.

He has been a Postdoctoral Fellow with the Department of Bioinformatics Science and Engineering, ETH Zürich, Zürich, Switzerland, since November 2019. His research interests are in the areas of learning theory, control theory, and systems/synthetic biology.

Sijie Lu received the B.Sc. degree in mathematics from China Agricultural University, Beijing, China, in 2019. She is currently pursuing the M.Sc. degree in operational research and cybernetics with Zhejiang University, Hangzhou, China.

Her research interests are in the areas of machine learning and its industrial applications.

Chuanhou Gao (Senior Member, IEEE) received the B.Sc. degree in chemical engineering from the Zhejiang University of Technology, Hangzhou, China, in 1998, and the Ph.D. degree in operational research and cybernetics from Zhejiang University, Hangzhou, in 2004.

From June 2004 to May 2006, he was a Postdoctoral Fellow with the Department of Control Science and Engineering, Zhejiang University, where he has been with the Department of Mathematics since June 2006, as a Full Professor. His research interests are in the areas of data-driven modeling, control and optimization, problem-driven applied mathematics, machine learning, and transparency of black-box modeling techniques.

Dr. Gao was a Guest Editor of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS and ISIJ International and Journal of Applied Mathematics and an Editor of Metallurgical Industry Automation from May 2013. He is currently an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL.