Abstract

Recently, several studies have proposed progressive or sequential layer-wise training methods based on the boosting theory for deep neural networks. However, most studies lack the global convergence guarantees or require weak learning conditions that can be verified *a posteriori* after running methods. Moreover, generalization bounds usually have a worse dependence on the network depth. In this paper, to resolve these problems, we propose a new functional gradient boosting for learning deep residual-like networks in a layer-wise fashion with its statistical guarantees on multi-class classification tasks. In the proposed method, each residual block is recognized as a functional gradient (i.e., weak learner), and the functional gradient step is performed by stacking it on the network, resulting in a strong optimization ability. In the theoretical analysis, we show the global convergence of the method under a standard margin assumption on a data distribution instead of a weak learning condition, and we eliminate a worse dependence on the network depth in a generalization bound via a fine-grained convergence analysis. Moreover, we show that the existence of a learnable function with a large margin on a training dataset significantly improves a generalization bound. Finally, we experimentally demonstrate that our proposed method is certainly useful for learning deep residual networks.
We first propose a new functional gradient boosting method which is considered as one of important tools in deep learning community. In this paper, we further develop this line of research. We propose a new functional gradient boosting called ResFGB-FW for multi-class classification problems by combining ResFGB with the Frank-Wolfe method, which is a basic scheme of boosting methods, leading to a slight modification of the network architecture of ResNets. In a theoretical analysis, we show our network architecture is helpful in ensuring the global convergence and provide a generalization bound without a type of weak learning condition. In addition, we demonstrate that a worse dependence on the depth can be eliminated by precisely estimating the required norms of parameters for global convergence of ResFGB-FW via a fine-grained convergence analysis. A key step in the proof for global convergence is to show that an obtained weak learner in the method exhibits comparable performance with a pure functional gradient in terms of optimization ability in a function space. We show this property by making a margin assumption where a given training dataset is separable with a sufficient margin by a good classifier in a set of learnable functions assembled by a shallow neural network, while this type of property is usually guaranteed by a weak learning condition in other methods. As a result, we can show not only that a worse dependence on the network depth can be eliminated but also that the existence of a good classifier providing a large margin can significantly improve a generalization bound. Finally, we experimentally compare our proposed method with state-of-the-art methods and demonstrate certain superior performances.

Contributions

- We propose a new functional gradient boosting for learning deep ResNets by combining ResFGB with the Frank-Wolfe method and show superior empirical performance over existing methods on multiclass classification tasks.
- We provide a generalization bound for a trained ResNet via a global convergence analysis of the method under a margin condition instead of a weak learning condition. Theoretical contributions in this study are that (i) we guarantee the global convergence without a type of weak learning condition, (ii) we show that our network architecture is helpful in ensuring the convergence, (iii) we eliminate a worse dependence on the network depth in a generalization bound via a convergence analysis.

Related Work

Recently, it is known that a progressive learning scheme or sequential layer-wise training has a powerful generalization ability as well as a conventional learning scheme on the image classification task (Belilovsky et al., 2019 Nokland and Eidnes, 2019) and it is useful in network architecture search (Liu et al., 2018 Kim et al., 2018 Well et al., 2019) which is considered as one of important tasks in deep learning community. So far, many studies (Han et al., 2016 Cortes et al., 2017 Mosca and Magoulas, 2017 Wang et al., 2017 Kukkarn and Marquez et al., 2018, Nitanda and Suzuki, 2018a) are closely related to our method in the sense that components of neural networks are iteratively added to optimize a network structure based on the boosting theory. However, our method has several different theoretical properties from these methods, for instance, BoostResNet and ResFGB required weak learning conditions to guarantee the global convergence of the methods and linear and exponential dependence on the network depth appeared in their generalization bounds. On the other hand, our theory guarantees the global convergence without such a weak learning condition and eliminates a worse dependence on the depth in a generalization bound. As for AdaNet, the dependence on the network depth in the variance term (i.e., estimation error) was logarithmic, but the approximation guarantee (i.e., the global convergence guarantee) was not provided in their theory.

2 Preliminary

In this section, we provide several notations to describe a problem setting of the classification and a notion of functional gradients used in the proposed method and theoretical analyses. Moreover, we provide a basic skeleton of generalization bound to derive an explicit bound later.

2.1 Problem setting

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y}$ be a feature space and a finite label set of cardinality $c$. We denote by $\nu$ a true Borel probability measure on $\mathcal{X} \times \mathcal{Y}$ and
by \( \nu_n \) an empirical probability measure by observations \((x_i, y_i)_{i=1}^n\) independently drawn from \( \nu \), i.e.,
\[
d\nu_n(X, Y) = \sum_{i=1}^n \delta(x_i, y_i)(X, Y)\,dXdY/n,
\]
where \( \delta \) denotes the Dirac delta function. For \( \nu_n \), we denote by \( \nu_n^X \) the marginal distribution on \( X \). In general, for a probability measure \( \mu \), we denote by \( \mathbb{E}_\mu \) the expectation with respect to a random variable according to \( \mu \). We denote by \( L_2(\nu_n^X) \) the real-valued function space on \( X \) equipped with inner product \( \langle \cdot, \cdot \rangle_{L^2(\nu_n^X)} \) and by \( L_2^2(\nu_n^X) \) the product space of \( L_2(\nu_n^X) \): for \( \forall \xi, \zeta \in L_2^2(\nu_n^X) \),
\[
\langle \xi, \zeta \rangle_{L^2^2(\nu_n^X)} \overset{\text{def}}{=} \mathbb{E}_{\nu_n^X} \left[ \sum_{j=1}^m \xi_j(X)\zeta_j(X) \right].
\]

\( L_p \)-norm is defined by:
\[
\|\xi\|_{L^p(\nu_n^X)} \overset{\text{def}}{=} \mathbb{E}_{\nu_n^X} \|\xi(X)\|_p = \mathbb{E}_{\nu_n^X} \sum_{j=1}^m |\xi_j(X)|^p \quad \text{for } p \in \{1, 2\} \text{ and } \xi \in L^2^2(\nu_n^X).
\]

Let \( \mathcal{F} \) be a given function class of predictors. The problem considered in this paper is formalized as the empirical risk minimization problem:
\[
\min_{f \in \mathcal{F}} \mathcal{L}_n(f) \overset{\text{def}}{=} \mathbb{E}_{\nu_n}[l(f(X), Y)],
\]
where \( l \) is the logistic loss function for multiclass classification problems:
\[
l(\xi, y) = -\log \left( \frac{\exp(\xi_y)}{\sum_{y \in Y} \exp(\xi_y)} \right).
\]

Generally, regularization or constraint is needed for a function class to guarantee generalization. Thus, we next describe a network architecture belonging to \( \mathcal{F} \).

### 2.2 Network architecture

We split the predictor \( f \) into the feature extractor and linear predictor:
\[
f(x) = w^\top \phi(x),
\]
where \( w \in \mathbb{R}^{d \times c} \) is a weight for the last layer and \( \phi \) is a feature extractor from \( \mathcal{X} \) to \( \mathcal{X} \). For simplicity, we pre-train \( w \) on a distribution \( \nu_n \) as a linear classifier and fix it before optimizing \( \phi \). In addition, we assume \( \| (\theta)_{\ast, y} \|_1 \leq \Lambda_w \) for \( \forall \theta \in \mathcal{Y} \). Thus, we focus on the problem of learning a feature extractor \( \phi \) and denote \( \mathcal{R}_n(\phi) = \mathcal{L}_n(\mathbb{P}^\top \phi) \).

We herein describe the entire network architecture that is a type of residual networks with slight modifications (see Figure 1).

**Definition 1.** Let \( \mathcal{B}_1 \) and \( \mathcal{B}_2 \) be sets of neural networks: \( \mathcal{X} \to \mathcal{X} \), including the zero function. We denote by \( \phi_t \) the output of \( \phi \) at the \( t \)-th layer. A set of residual blocks (resblocks) at the \( t \)-th layer is
\[
\mathcal{E}_{\phi_t} \overset{\text{def}}{=} \{ \iota_1 + \iota_2 \circ \phi_1 | \iota_1 \in \mathcal{B}_1, \ i_2 \in \mathcal{B}_2 \}.
\]

A connection between \( t \)- and \((t + 1)\)-th layers of \( \phi \) is described as follows:
\[
\phi_{t+1}(x) \overset{\text{def}}{=} (1 - \eta_t)\phi_t(x) + \eta_t\iota_t(x).
\]

![High level diagram of t-th layer of the network obtained in the proposed method. A linear classifier and an auxiliary loss are used to train resblocks at t-layer and discarded after training t-th layer.](image.png)

**Remark** Our model has two major differences from the conventional residual networks and the model in [Nitanda and Suzuki (2018a)]: one is that the output \( \phi_t(x) \) is reduced at the rate of \( 1 - \eta_t \) stemming from the Frank-Wolfe method and the other is that resblocks receive input data directly like DenseNet (Huang et al., 2017). These modifications lead to provable statistical guarantees without weak learning assumptions.

To provide theoretical analyses, some restrictions on resblocks are required.

**Assumption 1.** Let \( \Lambda_i^C, \Lambda_i^D \) \( (i \in \{1, 2\}) \) be positive values and \( \sigma \) be a 1-Lipschitz activation function. Assume that residual blocks are shallow neural networks defined as follows:
\[
\mathcal{B}_i = \{ z \mapsto C_i \sigma(D_i z) | \max_j \| (C_i)_j, \ast \|_1 \leq \Lambda_i^C, \max_k \| (D_i)_k, \ast \|_1 \leq \Lambda_i^D, C_i \in \mathbb{R}^{d \times d'}, D_i \in \mathbb{R}^{d' \times d} \}.
\]

Assumption 1 is used for guaranteeing the generalization bound. Although, we adopt shallow networks as resblocks in this condition for simplicity, but we can extend them to deeper resblocks. A typical example of \( \sigma \) is the sigmoid function.
2.3 Functional gradient

The key notion used in this study is the functional gradient in a function space. We define

\[
\nabla_\phi R_n(\phi)(x) \overset{\text{def}}{=} \begin{cases} 
\partial_z l(\overline{m}^T z, y_i) |_{z = \phi(x_i)} & (x = x_i), \\
0 & \text{(otherwise)}. 
\end{cases}
\]

This is nothing but a functional gradient (Fréchet differential) in \(L_2^d(\nu_X^n)\), i.e., for \(\forall \phi, \forall \xi \in L_2^d(\nu_X^n)\),

\[
R_n(\phi + \xi) = R_n(\phi) + \langle \nabla_\phi R_n(\phi), \xi \rangle_{L_2^d(\nu_X^n)} + o(\|\xi\|_{L_2^d(\nu_X^n)}).
\]

Thus, the functional gradient descent using \(\nabla_\phi R_n\) optimizes \(R_n\) in \(L_2^d(\nu_X^n)\), but we need a smoothing technique introduced later to guarantee generalization because \(\nabla_\phi R_n\) has no information on unseen data; in other words, this method is meaningless for expected minimization problems. We similarly define a functional gradient \(\nabla_f L_n(f)\) and see \(\nabla_\phi R_n(\phi) = \pi \nabla_f L_n(\overline{m}^T \phi)\).

The following proposition [Nitanda and Suzuki 2018a] shows the Fréchet differentiability and Lipschitz smoothness for the multiclass logistic loss.

**Proposition 1** [Nitanda and Suzuki 2018a]. \(R_n(\phi)\) is convex and Fréchet differentiable in \(L_2^d(\nu_X^n)\). Moreover, there exists \(\exists L > 0\) such that \(\forall \phi, \forall \xi \in L_2^d(\nu_X^n)\),

\[
R_n(\phi + \xi) \leq R_n(\phi) + \langle \nabla_\phi R_n(\phi), \xi \rangle_{L_2^d(\nu_X^n)} + \frac{L}{2} \|\xi\|_{L_2^d(\nu_X^n)}^2.
\]

**Remark** Although, the Lipschitz parameter of \(R_n(\phi)\) depends on \(\overline{m}\), we consider it as a uniform constant because \(\overline{m}\) is fixed in our theoretical analysis.

2.4 Generalization Bound

We provide a skeleton of generalization bound on an expected classification error which is derived by a conventional analysis of Rademacher complexity for neural networks [Koltchinskii and Panchenko 2002]. This bound is composed of the \(L_1\)-functional gradient norm and the conventional margin defined below:

\[
\|\nabla_f L_n(f)\|_{L_2^d(\nu_X^n)} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \left| \partial_{y_i} l(f(x_i), y_i) \right|, \\
m_f(x, y) \overset{\text{def}}{=} f_y(x) - \max_{y' \neq y} f_{y'}(x).
\]

**Theorem 1.** Suppose Assumption \(\square\) holds, \(\Lambda_2^C \Lambda_2^D \leq 1/2, \eta_k \leq 1\), and \(\|x\|_2 \leq \Lambda_\infty\) on \(\mathcal{X}\). Fix \(\delta > 0\). Then, for \(\forall \rho > 0\), with probability at least \(1 - \rho\) over the random choice from \(\nu^n\), we have that for \(\forall f \in \mathcal{F}\) with the network depth \(T \in \mathbb{Z}_+\),

\[
\mathbb{P}_\nu[m_f(X, Y) \leq 0] \leq \frac{2e^3 \Lambda_\infty \Lambda_\infty}{\delta \sqrt{n}} \left( 1 + \sum_{t=0}^{T-1} \eta_k \frac{\Lambda_2^C \Lambda_2^D}{\pi} \prod_{s=t+1}^{T-1} \left( 1 - \frac{\eta_k}{2} \right) \right) + \sqrt{\frac{1}{2n} \log \rho^{-1} + \frac{1}{2} \left( 1 + \exp(\delta) \right) \|\nabla_f L_n(f)\|_{L_2^d(\nu_X^n)}}.
\]

A remaining problem for deriving an explicit generalization bound is to confirm how small functional gradient norms can be achievable by given rates of \(\eta_k\) and upper bounds: \(\Lambda_2^C, \Lambda_2^D, \Lambda_2^C, \text{ and } \Lambda_2^D\). We estimate these quantities by analyzing a convergence behavior of a proposed functional gradient boosting for learning deep residual networks.

3 Brief Review of Functional Gradient

Functional gradient methods including gradient boosting [Mason et al. 1999; Friedman 2001] and gradient descent for kernel methods [Kivinen et al. 2004; Smale and Yao 2006; Ying and Zhou 2006; Raskutti et al. 2014; Wei et al. 2017] and over-parameterized neural networks [Jacot et al. 2018; Du et al. 2018; Arora et al. 2019] have powerful optimization ability because these methods naturally perform in function spaces.

In gradient boosting, \(\nabla_\phi R_n(\phi)\) is approximated in a given set of weak learners \(\mathcal{G}\) as follows:

\[
\xi_\phi \in \arg\max_{\xi \in \mathcal{G}} \langle \nabla_\phi R_n(\phi), \xi \rangle_{L_2^d(\nu_X^n)} \quad (2)
\]

and gradient method in a function space is performed using a descent direction \(\xi_\phi\). This approximation procedure is recognized as a type of smoothing of functional gradients.

In particular, this smoothing procedure is realized by using the following kernel smoothing technique in the kernel method and over-parameterized neural networks:

\[
T_k \nabla_\phi R_n(\phi) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \nabla_\phi R_n(\phi)(x_i) k(x_i, \cdot),
\]

where, \(k\) is a kernel function that is usually prefixed in kernel methods; on the contrary, it depends on a map \(\phi\) for over-parameterized neural networks and is called neural kernel [Jacot et al. 2018].

We note that a kernel smoothing [3] is characterized as a special case of [2] through the following relationship. Let \((\mathcal{H}_h, \langle \cdot, \cdot \rangle_{\mathcal{H}_h})\) be the reproducing Hilbert space associated with a kernel \(k\). Then,

\[
\frac{T_k \nabla_\phi R_n(\phi)}{\|T_k \nabla_\phi R_n(\phi)\|_{\mathcal{H}_h^2}} \in \arg\max_{\xi \in \mathcal{G}} \langle \nabla_\phi R_n(\phi), \xi \rangle_{L_2^d(\nu_X^n)},
\]

(4)
where $\|\xi\|_{H^d_{k'}}^2 = \sum_{i=1}^d \|\xi_i\|^2_{H_k}$ for $\xi = (\xi_i)_{i=1}^d \in H^d_k$, which is confirmed by $\langle \psi, \xi \rangle_{L^2_k(\nu^X)} = \langle T_k \psi, \xi \rangle_{H^d_k}$ for all $\xi \in H^d_k$. Hence, these methods using kernel smoothing are considered as variants of functional gradient boosting.

To further improve the optimization ability compared to (4), one interesting idea is to maximize the following quantity, which corresponds to the right hand side of (4), with respect to $k:

$$\langle \nabla \phi R_n(\phi), T_k \nabla \phi R_n(\phi) \rangle_{L^2_k(\nu^X)} = \|T_k \nabla \phi R_n(\phi)\|_{H^d_k}^2. \tag{5}$$

This approach was proposed in [Nitanda and Suzuki (2018a)] for learning residual networks by successively stacking layers, although there is a less explicit explanation about this viewpoint. In this paper, we also adopt this idea with several modifications for guaranteeing the global convergence property.

4 New Functional Gradient Boosting

We here propose a new functional gradient boosting for learning our models. The proposed method is essentially a combination of kernel learning for improving the approximation ability of kernel smoothing to functional gradients and the Frank-Wolfe method which is a basic scheme of boosting methods wherein weak learners are successively added into a prediction function. Usually, the Frank-Wolfe method is used for learning shallow neural networks [[Barron 1993], [Bach 2014], [Ping et al. 2016]], but we adapt it for learning deep residual networks by regarding residual blocks as weak learners. Since, in our method, residual blocks are composed of kernel smoothing of functional gradients and kernel functions are composed of a feature extractor and small networks, a modified Frank-Wolfe method makes a feature extractor one level deeper by adding this resblock. As a result, a network gradually grows as the algorithm proceeds, and finally, a deep residual network is obtained.

4.1 Functional gradient boosting for learning deep residual networks

We herein describe the detail of the proposed method. We note that ResFGB [Nitanda and Suzuki (2018a)] is a special case of our method if a learning rate strategy is ignored and $\lambda_1 = 0$. In our method, kernel functions used for kernel smoothing for resblocks are obtained by inner products of embedding maps. Let $K_1$ and $K_2$ be sets of such embedding maps $\mathbb{R}^d \rightarrow \mathbb{R}^{D'}$ ($D'$ may be set to $\infty$), and suppose $0 \in K_i$ ($i \in \{1, 2\}$). In our theory, the sets $K_1$ and $K_2$ are sufficiently restrictive to guarantee global convergence of the proposed method and the generalization ability of an obtained residual network, which is described later.

For a current feature extractor $\phi_t$, a kernel functions used for the smoothing of functional gradients is defined in the following manner: for $e_1 \in K_1, e_2 \in K_2$,

$$k_{e_1}(x, x') = e_1(x)^\top e_1(x'),$$

$$k_{e_2}(x, x') = e_2(\phi_t(x))^\top e_2(\phi_t(x')).$$

As explained earlier, a kernel function is optimized to well approximate the functional gradient:

$$\max_{\phi \in K_1} \|T_k \nabla \phi R_n(\phi)\|_{H^d_{k'}}^2,$$

$$\max_{\phi \in K_2} \|T_k \nabla \phi R_n(\phi)\|_{H^d_{k'}}^2. \tag{6}$$

Instead of (6), solving the following problem is practically better: for $D' = d$,

$$\min_{(e_1, e_2) \in K_1 \times K_2} \|e_1 + e_2 \circ \phi_t - \nabla \phi R_n(\phi_t)\|_{L^2_k(\nu^X)}^2, \tag{7}$$

because this problem (7) is much easier to solve and $\nabla \phi R_n(\phi_t)/\|\nabla \phi R_n(\phi_t)\|_2$ is a good solution of (6) if it is contained in a set: $K_1 + K_2 \circ \phi_t$ as shown in [Nitanda and Suzuki (2018a)]. However, we note that this approximation is just a heuristic and there is a room for improvement by exploring another approximation. We denote by $\tau_1, \tau_2$ approximate solutions of these problems.

Using kernels $k_{e_1}$, a residual block that is a weak learner to be added into a feature extractor $\phi_t$ in the proposed method is constructed as follows:

$$\tau_t = -\frac{T_{k_{e_1}} \nabla \phi R_n(\phi_t)}{Z_1^1} - \frac{T_{k_{e_2}} \nabla \phi R_n(\phi_t)}{Z_1^2 \circ \phi_t}, \tag{8}$$

where $Z_1$ is the normalization term: $Z_1 = \|T_{k_{e_1}} \nabla \phi R_n(\phi_t)\|_{H^d_{k'}}$. We note that the components of a resblock, $\tau_1$ and $\tau_2$, are also defined in the above equation and are composed of neural networks:

$$\tau_1 = A_1 \bar{\tau}_1(\cdot), \quad \tau_2 \circ \phi_t = A_2 \bar{\tau}_2 \circ \phi_t(\cdot),$$

where $A_1, A_2$ are $(d, d)$-matrices:

$$\begin{cases}
A_1 &= -\frac{1}{n \lambda_1} \sum_{i=1}^n \nabla \phi R_n(\phi_t(x_i)) \bar{\tau}_1(x_i) \\
A_2 &= -\frac{1}{n \lambda_2} \sum_{i=1}^n \nabla \phi R_n(\phi_t(x_i)) \bar{\tau}_2 \circ \phi_t(x_i)
\end{cases} \tag{9}$$

Finally, a Frank-Wolfe-like update is performed as follows: for a given learning rate $\eta_t$,

$$\phi_{t+1} \leftarrow (1 - \eta_t) \phi_t + \eta_t \tau_t.$$

\[\text{Atsushi Nitanda, Taiji Suzuki}\]
As the right hand side of the above iteration is equal to \(((1 - \eta_t)id + \eta_t\phi_t) \circ \phi_t(x) + \eta_{t+1}(x)\), we notice that \(\phi_{t+1}\) is one level deeper than \(\phi_t\) and a deep residual network is constructed.

We give a full description of the proposed method called ResFGB-FW in Algorithm 1. Although, we made the assumption where a weight \(\mathbf{w}\) for the last layer is pre-trained due to theoretical tractability, it may be practically better to optimize \(\mathbf{w}\) at each iteration, so that Algorithm 1 contains it as an option. We here note that this procedure is certainly similar to a progressive learning scheme in Belilovsky et al. (2019), except for details of the architecture of each layer, which exhibits comparable high performance with end-to-end training schemes on the image classification tasks.

**Algorithm 1 ResFGB-FW**

**Input:** \(S = (x_i, y_i)_{i=1}^n\), the number of iterations \(T\), and learning rates \(\eta_t\)

\(\phi_0 \leftarrow id, \ \mathbf{w} \leftarrow \argmin_{w \in \mathbb{R}^{d \times c}, \|w\|_1} \sum_{i \leq \Lambda_w} \mathcal{L}_n(w^T \phi_0)\)

for \(t = 0\) to \(T - 1\) do

(Option) \(\mathbf{w} \leftarrow \argmin_{w \in \mathbb{R}^{d \times c}, \|w\|_1} \sum_{i \leq \Lambda_w} \mathcal{L}_n(w^T \phi_t)\)

Get \(\tau_t\) by solving \(6\) or \(7\) on \(S\)

\[A_1 \leftarrow \frac{1}{\frac{1}{2}n^2} \sum_{i=1}^n \sum_{j=1}^n \nabla_{\phi} \mathcal{R}_n(\phi_t)(x_i) \mathcal{E}_t^i(x_j)\]

\[A_2 \leftarrow \frac{1}{\frac{1}{2}n^2} \sum_{i=1}^n \sum_{j=1}^n \nabla_{\phi} \mathcal{R}_n(\phi_t)(x_i) (\mathcal{E}_2 \circ \phi_t)(x_j) \]

\[\tau_t \leftarrow A_1 \mathcal{E}_1(\cdot) + A_2 \mathcal{E}_2 \circ \phi_t(\cdot)\]

\(\phi_{t+1} \leftarrow (1 - \eta_t)\phi_t + \eta_t \tau_t\)

end for

Return \(\mathbf{w}^T \phi_T\)

**Remark** There are two major differences from the method in Nitanda and Suzuki (2018a); one is that the output \(\phi_t(x)\) is reduced at the rate of \(1 - \eta_t\) stemming from the Frank-Wolfe method and the other is that resblocks receive input data directly like DenseNet (Huang et al., 2017). These modifications lead to provable statistical guarantees without weak learning assumptions adopted in Nitanda and Suzuki (2018a).

5 Convergence Analyses

We here provide the convergence analysis of Algorithm 1. To do so, we have to specify the architecture of \(\mathcal{K}_i\) and provide the required condition on resblocks obtained in Algorithm 1 to guarantee the global convergence property. We assume \(\mathcal{K}_i\) are sets of shallow neural networks: \(F_i(G_i, x)\) with norm constraints. Note that the parameters of resblocks introduced in Section 2 can be obtained by

\[C_i = A_iF_i, \ D_i = G_i.\]  \(10\)

The set of learnable functions by our method is described as follows. We denote by \(\mathcal{D}_{\mathcal{K}_i}\) the union of unit balls in RKHSs defined by \(\epsilon_i \in \mathcal{K}_1\):

\[\mathcal{D}_{\mathcal{K}_i} \overset{def}{=} \bigcup_{\epsilon_i \in \mathcal{K}_i} \{ \xi \in \mathcal{H}_{\mathcal{K}_i}^d \ | \ |\xi||_{\mathcal{H}_{\mathcal{K}_i}^d} \leq 1 \}.\]

Then, we define \(\text{conv}(\mathcal{D}_{\mathcal{K}_i})\) as the convex combination of \(\mathcal{D}_{\mathcal{K}_i}\). We here make the following assumptions for theoretical guarantees.

**Assumption 2.**

(A1) Let \(\Lambda_i^F, \Lambda_i^G (i \in \{1, 2\})\) be positive values and \(\sigma\) be a 1-Lipschitz activation function. Let \(\mathcal{K}_1, \mathcal{K}_2\) be sets of shallow neural networks:

\[\mathcal{K}_i = \{ z \mapsto F_i(x_i) | \sum_j \| (F_i)_t \|_2 \leq \Lambda_i^F, \ max_k \| (G_i)_t \|_1 \leq \Lambda_i^G \}.\]

In addition, \(\| e_i(z) \|_2 \leq \sqrt{K} (i \in \{1, 2\}, K > 0)\) over \(z \in \mathbb{R}^d\) and the maximum singular value of \(\mathbf{w}\) is upper bounded by \(\Sigma_{\mathbf{w}}\).

(A2) Resblocks: \(\iota_1 = A_1\mathcal{E}_1\) and \(\iota_t = A_1\mathcal{E}_1 + A_2\mathcal{E}_2 \circ \phi_t\) satisfy the following:

\[-\langle \nabla_{\phi} \mathcal{R}_n(\phi_t), \iota_t \rangle_{L^2_\nu(v)^N} \geq \frac{1}{2} \max_{\xi \in \mathcal{D}_{\mathcal{K}_1}} \langle \nabla_{\phi} \mathcal{R}_n(\phi), \xi \rangle_{L^2_\nu(v)^N},\]

\[-\langle \nabla_{\phi} \mathcal{R}_n(\phi_t), \iota_t \rangle_{L^2_\nu(v)^N} \geq \max_{\xi \in \mathcal{D}_{\mathcal{K}_1}} \langle \nabla_{\phi} \mathcal{R}_n(\phi), \xi \rangle_{L^2_\nu(v)^N}.\]

(A3) For a training set \(S = (x_i, y_i)_{i=1}^n\) sampled from \(\nu^n\), there exist \(M_n, \gamma_n > 0\) and \(\psi_n \in \text{conv}(\mathcal{D}_{\mathcal{K}_1})\) such that for \(v_{i,y} \equiv \mathbf{w}_i^T y_i \mathcal{E}_n(x_i), \forall i \in \{1, \ldots, n\},\)

\[v_{i,y} \geq M_n + \gamma_n, \ v_{i,y} \leq M_n (\forall y \neq y_i).\]

**Remark** Note that since the first component of resblock: \(-\iota_1 = -A_1\mathcal{E}_1\) maximizes \(\langle \nabla_{\phi} \mathcal{R}_n(\phi), \xi \rangle_{L^2_\nu(v)^N}\) through equations (1) and (5) over \(\xi \in \mathcal{D}_{\mathcal{K}_1}\), a kernel \(\tau_t\) always satisfies the condition (A2) if problems (6) are solved exactly, but (A2) also allows approximate solution thanks to a support by \(\iota_2\). In other words, a deep architecture alleviates the requirement of the optimality of weak learners for subproblems and is certainly useful in enhancing the global convergence. The condition (A3) means that a given training set is separable by \(\text{conv}(\mathcal{D}_{\mathcal{K}_1})\) with a margin \(\gamma_n\). We note that this assumption is reasonable because \(\text{conv}(\mathcal{D}_{\mathcal{K}_1})\) is essentially a set of infinite-width shallow neural networks and has a universal approximation property. Hence, a margin \(\gamma_n\) actually exists as long as a data distribution is separable by a continuous function.

To ensure the convergence and generalization properties, we should derive upper bounds on norms of
parameters of resblock $\tau$ and norm $\|\tau(z)\|_2$, but it is not obvious because resblocks contain a normalization term $Z_1$. An interesting technical contribution is to derive the lower bound on $Z_1$ under Assumption 2.

**Proposition 2.** Suppose Assumption 2 holds. Then, iterates of Algorithm 1 with learning rates $\gamma_n$ of Algorithm 1 satisfy

$$\frac{\gamma_n}{4\sum_{i=1}^n} \| \nabla \phi R_n(\phi) \|_{L_2^2(\nu_n)} \leq \| T_{\phi,1} \nabla \phi R_n(\phi) \|_{\nu_{\phi,1}^n}. \quad (11)$$

Proposition 2 says that the kernel smoothing in Algorithm 1 produces a sufficient reduction of objective in a function space as the margin $\gamma_n$ is large. As a result, norms of parameters and outputs of resblocks can be made small.

**Proposition 3.** Suppose Assumption 2 holds. Then, it follows that for resblocks $t_i = A_i F_t \sigma(G_t)$ obtained in Algorithm 1,

$$\| (A_i F_t) \phi \phi_1 \|_1 \leq \frac{4}{\gamma_n} \sum_{i=1}^n \| \phi \|_F \sqrt{K}, \quad (12)$$

$$\| t_i \|_{L_2^2(\nu_n)} \leq \frac{4}{\gamma_n} \sum_{i=1}^n \| \phi \|_F \sqrt{K}. \quad (13)$$

This proposition is useful for both convergence and generalization guarantees, namely, the equation (12) is used for deriving a generalization bound by substituting it into the bound in Theorem 1, and the equation (13) is used for deriving a convergence rate.

**Global Convergence.** The next theorem shows the convergence property of Algorithm 1 to a better model than $\text{conv}(D_{\kappa_1})$, which can be shown by adapting a proof for the vanilla Frank-Wolfe method (Jaggi 2013) to our method. For simplicity, we set $\mathcal{R}_n \equiv \inf_{\phi \in \text{conv}(D_{\kappa_1})} R_n(\phi)$ and for $t_0, \Lambda_\infty > 0$,

$$C \equiv \max \left\{ t_0 (\mathcal{R}_n(\phi_0) - \mathcal{R}_n^*), 2L \left( \Lambda_\infty + \frac{16\pi^2 K}{\gamma_n} \right)^2 \right\}.$$

**Theorem 2 (Convergence rate).** Suppose Assumption 2 holds and $\| x \|_2 \leq \Lambda_\infty$ on $\mathcal{X}$. Then, iterates of Algorithm 1 with learning rates $\eta_t = \frac{2}{t_0^2}$ ($t_0 \geq 0$) satisfy

$$R_n(\phi_T) - \mathcal{R}_n^* \leq \frac{C}{T + t_0}.$$

Note that this theorem provides a global convergence guarantee. An important point of our analysis is that a type of weak learning condition is not required unlike existing studies (Huang et al. 2018; Nitanda and Suzuki 2018a). In addition, we remark that the convergence becomes faster when the input distribution has a high degree of linear separability and when using a sufficient large $t_0$.

**Generalization Bound.** By combining Proposition 2, 3, and Theorem 2 with generalization bound in Theorem 1, we obtain the main result in this study. To describe the result, define

$$\epsilon_n \overset{\text{def}}{=} (c - 1) \exp(-\gamma_n).$$

We can easily see that $\epsilon_n$ is an upper bound on $\mathcal{R}_n^*$ as follows. We set $f = \overline{w}^T \psi_n$ where $\psi_n \in \text{conv}(D_{\kappa_1})$ is defined in Assumption 2 (A3). Under Assumption 2 (A3), for $i \in \{1, \ldots, n\}$,

$$l(f(x_i), y_i) = \log(1 + \sum_{y' \neq y_i} \exp(f(y') - f(y_i) \cdot (x_i))) \leq \log(1 + (c - 1) \exp(-\gamma_n)) \leq \epsilon_n.$$

Thus, taking an expectation with respect to $\nu_n^X$, we see that $\mathcal{R}_n^* \leq \mathcal{R}_n(\psi_n) \leq \epsilon_n$. Moreover, we note that $\epsilon_n \to 0$ as $\gamma_n \to \infty$.

**Corollary 1 (Main result).** Suppose Assumption 2 holds and $\| x \|_2 \leq \Lambda_\infty$ on $\mathcal{X}$. Assume $\Lambda^2 \Lambda^2 < \gamma_n/8\sum_{i=1}^n \phi \sqrt{K}$. Fix $\delta > 0$. Then, for $\forall \rho > 0$, with probability at least $1 - \rho$ over random choice of $\nu_n^X$, a predictor $f_T$ obtained by Algorithm 1 with $\eta_t = \frac{2}{t_0^2}$ ($t_0 \geq 2$) and $T$-iterates satisfies the following bound.

$$\mathbb{P}_n \left[ m(f, X, Y) \leq 0 \right] \leq \left( 1 + \exp(\delta) \right) \left( \epsilon_n + \frac{C}{T + t_0} \right) \left( 1 + \frac{2\delta}{\gamma_n} \sqrt{\frac{\Lambda_\infty}{\delta}} \right) \frac{1}{2} \log(1 + \frac{1}{\gamma_n} \log \rho^{-1}).$$

From Corollary 1, we can see that a worse dependence on the network depth disappears; hence, the best generalization performance is achieved when $T \to \infty$. Specifically, if $\delta = \gamma_n/2$ and $\gamma_n$ are uniformly positive, then we get for a sufficiently large $T$,

$$\mathbb{P}_n \left[ m(f, X, Y) \leq 0 \right] \leq \left( \epsilon_n + \frac{1}{\gamma_n} \right) + \sqrt{\frac{1}{n} \log \rho^{-1}}.$$

From this bound, we can see that the standard margin $\gamma_n$ helps the generalization ability of the method, that is, the complexity term is considerably reduced when the margin $\gamma_n$ is sufficiently large.

6 Numerical Experiments

We present experimental results on the binary and multiclass classification tasks. Our implementation is done using Theano (Theano Development Team 2016). The experiments are conducted on Intel Core i7-8700K and TITAN Xp 12GB with Ubuntu 16.04 (64-bit). We compare Algorithm 1 with ResFGB (Nitanda and Suzuki 2018a), multilayer perceptron (MLP), support vector machine (SVM), random forest (RF), and gradient boosting methods (GBM).
Table 1: Test classification accuracies on benchmark datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>LETTER</th>
<th>USPS</th>
<th>IJCNN1</th>
<th>MNIST</th>
<th>COVTYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResFGB-FW</td>
<td>0.981</td>
<td>0.954</td>
<td>0.988</td>
<td>0.988</td>
<td>0.969</td>
</tr>
<tr>
<td>ResFGB</td>
<td>0.976</td>
<td>0.954</td>
<td>0.989</td>
<td>0.986</td>
<td>0.965</td>
</tr>
<tr>
<td>MLP</td>
<td>0.970</td>
<td>0.949</td>
<td>0.988</td>
<td>0.986</td>
<td>0.965</td>
</tr>
<tr>
<td>SVM</td>
<td>0.961</td>
<td>0.947</td>
<td>0.979</td>
<td>0.970</td>
<td>0.830</td>
</tr>
<tr>
<td>RF</td>
<td>0.966</td>
<td>0.940</td>
<td>0.980</td>
<td>0.973</td>
<td>0.950</td>
</tr>
<tr>
<td>GBM</td>
<td>0.967</td>
<td>0.939</td>
<td>0.982</td>
<td>0.982</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Figure 2: Learning curves for Algorithm 1 on multiclass classification problems. The figure shows classification accuracy on training (blue) and test (purple) sets versus the number of iterations.

We use the following datasets: letter, usps, ijcnn1, mnist, and covtype. We first divide each train dataset randomly into two sets: 80% for training and the rest for validation. We run each method on the train dataset with several hyperparameter settings and choose a setting providing the best classification accuracy on the validation dataset. We finally train each model on an entire train dataset and report a classification accuracy on the test dataset. We perform the above procedure 5-times.

We explain hyper-parameter settings. For Algorithm 1, three hidden-layers neural networks with 1000-nodes per layer are used as $e_i \in \mathcal{K}_i$ ($i \in \{1, 2\}$) to define resblocks. Linear classifiers and $e_i$ are trained by Nesterov’s momentum method at each loop. The number of iterations $T$ and a parameter $t_0 \in \{10, 100, 1000\}$ for learning rates $\eta_t = 2/(t + t_0)$ are tuned based on the performance on the validation set. For ResFGB (Nitanda and Suzuki, 2018a), constant learning rates are chosen from $\{10^{-3}, 10^{-2}, 10^{-1}\}$ and similar settings as Algorithm 1 are adopted for the other hyper-parameters. For MLP, the ReLU is used as an activation and network depth is set to three, four, or five layers. The number of hidden units per layer is 100 or 1000. For SVM, a random Fourier feature (Rahimi and Recht, 2007) is adopted with an embedding dimension of $10^3$ or $10^4$. For RF, the maximum depth is set to 10, 20, or 30 and the number of trees is set to 100, 500, or 1000. For GBM, the learning rate is chosen from $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ with the maximum number of trees being set to 1000, and number of leaves in one tree is chosen from $\{2^4, 2^5, \ldots, 2^{10}\}$.

Table 1 shows the mean classification accuracies and standard deviations and Figure 2 depicts learning curves on training and test datasets. As seen in the table, the proposed method shows superior or comparable performance over other state-of-the-art methods. Therefore, we can conclude that our proposed method is certainly useful for training deep ResNets with a provable statistical guarantee.

7 Conclusion

We have proposed a new functional gradient boosting by combining ResFGB with the Frank-Wolfe method for learning deep ResNets. In theoretical analyses, we have provided a generalization bound via a global convergence analysis of the method under a margin condition instead of a weak learning condition. Compared to the previous study, our bound has two important points; one is the elimination of a worse dependence on the network depth, and the other is a significant improvement by the existence of a learnable classifier with a large margin on train datasets. Finally, we have demonstrated superior empirical performance of the proposed method over the state-of-the-art methods. An interesting future work is to adjust our method to the network architecture search by combining with a usual end-to-end training such as stochastic gradient descent method. To do so, it will be necessary to develop an effective criterion to switch the functional gradient boosting and the end-to-end training schemes.

Acknowledgement AN was partially supported by JSPS Kakenhi (19K20337) and JST-PRESTO. TS was partially supported by JSPS Kakenhi (26280009, 15H05707 and 18H03201), Japan Digital Design and JST-CREST.
References


Appendix

A Auxiliary Lemmas

For the proof of Theorem 1 which gives the skeleton of generalization bounds of deep residual networks obtained by the proposed method, we here introduce the well known uniform bound ([Koltchinskii and Panchenko, 2002]) using Rademacher complexity.

Let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ to $[-a,a]$ and $S = \{x_i\}_{i=1}^n \subset \mathcal{X}$ be a dataset sampled from $\nu^{X,n}$, where $\nu^X$ is the marginal distribution on $X$. Then, we define empirical Rademacher complexity $\hat{\mathbb{R}}_S(\mathcal{G})$ and Rademacher complexity $\mathbb{R}_n(\mathcal{G})$ as follows. For i.i.d random variables: $\sigma = (\sigma_i)_{i=1}^n$ taking $-1$ or 1 with equal probability,

$$\hat{\mathbb{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right], \quad \mathbb{R}_n(\mathcal{G}) = \mathbb{E}_{\nu^{X,n}}[\hat{\mathbb{R}}_S(\mathcal{G})].$$

We introduce a useful lemma for estimating Rademacher complexity.

**Lemma A.** (i) Let $h_i : \mathbb{R} \to \mathbb{R}$ ($i \in \{1, \ldots, s\}$) be $L$-Lipschitz functions. Then it follows that

$$\mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i h_i \circ f(x_i) \right] \leq L \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right].$$

(ii) We denote by $\text{conv}(\mathcal{G})$ the convex hull of $\mathcal{G}$. Then, we have $\hat{\mathbb{R}}_S(\text{conv}(\mathcal{G})) = \hat{\mathbb{R}}_S(\mathcal{G}).$

The following lemma gives the generalization bound using the margin distribution (c.f., [Koltchinskii and Panchenko, 2002]). Let $\mathcal{G}$ be the set of predictors; $\mathcal{G} \subset \{f : \mathcal{X} \to \mathbb{R}^c\}$ and $\Pi \mathcal{G} = \{f_y(\cdot) : \mathcal{X} \to \mathbb{R} \mid f \in \mathcal{G}, y \in \mathcal{Y}\}$, then the following holds.

**Lemma B.** Fix $\delta > 0$. Then, for $\forall \rho > 0$, with probability at least $1 - \rho$ over the random choice of $S$ from $\nu^n$, we have $\forall f \in \mathcal{G},$

$$\mathbb{P}_{\nu^n} [m_f(X,Y) \leq \delta] \leq \mathbb{P}_{\nu^n} [m_f(X,Y) \leq \delta] + \frac{2e^2}{\delta} \mathbb{R}_n(\Pi \mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\rho}}.$$

The next lemma provides the bound on the margin distribution by functional gradient norm. We give a proof for the completeness.

**Lemma C** ([Nitanda and Suzuki, 2018a]). For $\forall \delta > 0$, the following bound holds.

$$\mathbb{P}_{\nu^n} [m_f(X,Y) \leq \delta] \leq \frac{1}{2} (1 + \exp(\delta)) \| \nabla_f \mathcal{L}_n(f) \|_{L_1(\nu^n)}.$$
We first prove Theorem 1. Thus, we get by Markov inequality and the above equation,

This implies,

Thus, we get by Markov inequality and the above equation,

B Proofs

We first prove Theorem 1. Proof of Theorem 1. We here recall the network structure. The connection between the $t$-and $t+1$-th layers is

We define recursively the family of functions $\mathcal{H}_t$, $\hat{\mathcal{H}}_{1,t}$, and $\hat{\mathcal{H}}_{2,t}$ where each neuron belongs. We denote by $P_{\alpha} \in \mathbb{R}^d$ the projection vector to $\alpha$-th coordinate.

Then, the family of predictors of $y \in \mathcal{Y}$ can be written as

Note that $\mathcal{G}_T = \{(f_y)_{y \in \mathcal{Y}} | f_y \in \mathcal{G}_{T,y}, y \in \mathcal{Y}\}$.

Let us evaluate Rademacher complexities of these sets. For $\mathcal{H}_0$, since $\|P_j\|_2 = 1$, we have

$$\hat{\mathcal{R}}_S(\mathcal{H}_0) = \frac{1}{n} \mathbb{E}_{(\sigma_i)_{i=1}^n} \left[ \sup_{\alpha \in \{1, \ldots, d\}} \sum_{i=1}^n \sigma_i P_{\alpha} x_i \right]$$

$$\leq \frac{1}{n} \mathbb{E}_{(\sigma_i)_{i=1}^n} \left[ \sup_{\alpha \in \{1, \ldots, d\}} \|P_{\alpha}\|_2 \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2 \right]$$

$$= \frac{1}{n} \mathbb{E}_{(\sigma_i)_{i=1}^n} \left[ \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2 \right]$$
We set
$$\begin{align*}
&\hat{R}_S(\hat{H}_{1,t}) \leq \frac{1}{n} \left( E_{(\sigma_i)_{i=1}^n} \left[ \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right] \right)^{\frac{1}{2}} \\
&= \frac{1}{n} \left( \sum_{i=1}^n \| x_i \|_2^2 \right)^{\frac{1}{2}} \leq \frac{\Lambda_{\infty}}{\sqrt{n}},
\end{align*}$$

where we used the independence of $\sigma_i$ when taking the expectation.

As for $\hat{H}_{1,t}$, we immediately get from Lemma $\text{A}$
$$\begin{align*}
\hat{R}_S(\hat{H}_{1,t}) &\leq \Lambda_1^D \hat{R}_S(\hat{H}_0) \leq \frac{\Lambda_1^D \Lambda_{\infty}}{\sqrt{n}}.
\end{align*}$$

From these relationships, 1-Lipschitz continuity of $\sigma$, and Lemma $\text{A}$ we get
$$\begin{align*}
\hat{R}_S(\hat{H}_t) &\leq (1 - \eta_{t-1}) \hat{R}_S(\hat{H}_{t-1}) + \eta_{t-1} \Lambda_1^C \hat{R}_S(\hat{H}_{1,t-1}) + \eta_{t-1} \Lambda_2^C \hat{R}_S(\hat{H}_{2,t-1}) \\
&\leq (1 - \eta_{t-1} + \eta_{t-1} \Lambda_2^C \Lambda_{\infty}^D) \hat{R}_S(\hat{H}_{t-1}) + \frac{\eta_{t-1} \Lambda_1^C \Lambda_{\infty}^D}{\sqrt{n}} \\
&\leq \frac{\Lambda_{\infty}}{\sqrt{n}} \left( 1 + \sum_{u=0}^{t-1} \eta_u \Lambda_1^C \Lambda_{\infty}^D \prod_{s=u+1}^{t-1} \left( 1 - \frac{\eta_s}{2} \right) \right),
\end{align*}$$

where we used $\eta_s \leq 1$, $\Lambda_2^C \Lambda_{\infty}^D \leq 1/2$, and used that recursive inequality: $a_{t+1} \leq \gamma_t a_t + b_t$ leads to,
$$\begin{align*}
a_t &\leq a_0 \prod_{s=0}^{t-1} \gamma_s + \sum_{u=0}^{t-1} b_u \prod_{s=u+1}^{t-1} \gamma_s.
\end{align*}$$

We set $\Pi \mathcal{G}_T = \{ f_y(\cdot) : \mathcal{X} \to \mathbb{R} \mid f \in \mathcal{G}_T, y \in \mathcal{Y} \}$. Noting that $\hat{R}_S(\Pi \mathcal{G}_T) \leq \sum_{y \in \mathcal{Y}} \hat{R}_S(\mathcal{G}_{T,y})$, we get
$$\begin{align*}
\hat{R}_S(\Pi \mathcal{G}_T) &\leq \frac{c \Lambda_{\infty}}{\sqrt{n}} \left( 1 + \sum_{t=0}^{T-1} \lambda_t \prod_{s=t+1}^{T-1} \left( 1 - \frac{\eta_s}{2} \right) \right), \tag{14}
\end{align*}$$

Thus, combining Lemma $\text{B}$ $\text{C}$ with (14), we finish the proof. \hfill \square

We give the proof of key proposition which ensures that kernel smoothing of functional gradient leads to a sufficient reduction of the objective function in the function space.

Proof of Proposition $\text{A}$ For simplicity, we omit the index $t$ and set $f = \overline{w}^\top \phi$. We first evaluate $\| \nabla_f \mathcal{L}_\mathcal{N}(f) \|_{L_1(\nu_\mathcal{X})}$ as follows. Since $y$-th element of $\nabla_f \mathcal{L}_\mathcal{N}(f)(x_i)$ is
$$\begin{align*}
\partial_{x_i} l(f(x_i), y_i) = -1[y = y_i] + \frac{\exp(f_y(x_i))}{\sum_{y' \in \mathcal{Y}} \exp(f_{y'}(x_i))},
\end{align*}$$

where $\zeta_y$ denotes the $y$-th element of the first argument of $l(\zeta, y)$, we get
$$\begin{align*}
\| \nabla_f \mathcal{L}_\mathcal{N}(f) \|_{L_1(\nu_\mathcal{X})} = \frac{1}{n} \sum_{i=1}^n \left\{ 1 - \frac{\exp(f_y(x_i))}{\sum_{y' \in \mathcal{Y}} \exp(f_{y'}(x_i))} + \sum_{y \neq y'} \frac{\exp(f_y(x_i))}{\sum_{y' \in \mathcal{Y}} \exp(f_{y'}(x_i))} \right\}.
\end{align*}$$
We next provide the proof of proposition which gives upper bounds on norms of parameters of a resblock. We next focus on

\[
\|T_{k_1} \nabla R_n(\phi)\|_{\mathcal{H}^2_{k_1}}.
\]

For \(\psi_n \in \text{conv}(D_{K_1})\), there exists \(\alpha_s > 0\) and \(\xi_s \in D_{K_1}\) \((s \in \{1, \ldots, m\})\) such that \(\sum_s \alpha_s = 1\) and \(\psi_n = \sum_s \alpha_s \xi_s\). From Assumption 2 (A2) and the symmetry of \(D_{K_1}\), for \(\forall s \in \{1, \ldots, m\}\),

\[
Z_1^2 = \langle \nabla \phi R_n(\phi), T_{k_1} \nabla R_n(\phi) \rangle_{L^2_2(\nu^X_n)}
= -Z_1 \langle \nabla \phi R_n(\phi), t_1 \rangle_{L^2_2(\nu^X_n)}
\geq \frac{Z_1}{2} \langle \nabla \phi R_n(\phi), -\xi_s \rangle_{L^2_2(\nu^X_n)}.
\]

Hence, we get from \(\sum_s \alpha_s = 1\),

\[
\|T_{k_1} \nabla R_n(\phi)\|_{\mathcal{H}^2_{k_1}} \geq \frac{1}{2} \langle \nabla \phi R_n(\phi), -\psi_n \rangle_{L^2_2(\nu^X_n)}.
\]

We set \(v_i = \psi_n(x_i)\)^T \(w\) and denote by \(v_{i,y}\) \(y\)-th element of \(v_i\). We get from Assumption (A3) and \(\nabla \phi R_n(\phi) = \nabla f \mathcal{L}_n(f)\),

\[
\langle \nabla \phi R_n(\phi), -\psi_n \rangle_{L^2_2(\nu^X_n)} = \frac{1}{n} \sum_{i=1}^{n} \psi_n(x_i)^T \nabla f \mathcal{L}_n(f)(x_i)
= \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f \mathcal{L}_n(f)(x_i)
= \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in Y} v_{i,y} \partial_{\psi_y} \mathcal{L}_n(f)(x_i)
= \frac{1}{n} \sum_{i=1}^{n} \left\{ v_{i,y} \left( 1 - \frac{\exp(f_y(x_i))}{\sum_{y' \in Y} \exp(f_{y'}(x_i))} \right) - \sum_{y \neq y_i} v_{i,y} \frac{\exp(f_{y}(x_i))}{\sum_{y' \in Y} \exp(f_{y'}(x_i))} \right\}
\geq \frac{1}{n} \sum_{i=1}^{n} \left\{ (M_n + \gamma_n) \left( 1 - \frac{\exp(f_y(x_i))}{\sum_{y' \in Y} \exp(f_{y'}(x_i))} \right) - M_n \frac{\exp(f_{y}(x_i))}{\sum_{y' \in Y} \exp(f_{y'}(x_i))} \right\}
\geq \frac{\gamma_n}{n} \sum_{i=1}^{n} \left\{ 1 - \frac{\exp(f_{y}(x_i))}{\sum_{y' \in Y} \exp(f_{y'}(x_i))} \right\}
\]

Therefore, combining equations (15), (16), and (17), we get

\[
\|T_{k_1} \nabla R_n(\phi)\|_{\mathcal{H}^2_{k_1}} \geq \frac{\gamma_n}{4} \|\nabla f \mathcal{L}_n(f)\|_{L^2_1(\nu^X_n)}
\geq \frac{\gamma_n}{4}\|\nabla \phi R_n(\phi)\|_{L^2_1(\nu^X_n)},
\]

where we used \(\|\nabla \phi R_n(\phi)\|_{L^2_2(\nu^X_n)} \leq \sum_{\nu \neq \nu} \|\nabla f \mathcal{L}_n(f)\|_{L^2_1(\nu^X_n)}\) which can be confirmed in a straight manner. 

We next provide the proof of proposition which gives upper bounds on norms of parameters of a resblock \(\tau\) and \(\|\tau(z)\|_2\).
Proof of Proposition \[3\] We show (12) and (13) only for \(i = 1\) because we can also show for \(i = 2\) in the same manner.

\[
\| (A_1 F_1)_{\alpha, s} \|_1 = \sum_j |(A_1)_{\alpha, s}^\top (F_1)_{s, j} |
\]

\[
\leq \sum_j \| (A_1)_{\alpha, s} \|_2 \| (F_1)_{s, j} \|_2
\]

\[
\leq \Lambda_1 F \| (A_1)_{\alpha, s} \|_2
\]

\[
= \frac{\Lambda_1 F}{n Z_1} \left\| \sum_{i=1}^n \left( \nabla_{\phi} R_n(\phi_t)(x_i) \right)_\alpha \, \bar{e}_1^\top(x_i) \right\|_2
\]

\[
\leq \frac{\Lambda_1 F \sqrt{K}}{n Z_1} \sum_{i=1}^n |(\nabla_{\phi} R_n(\phi_t)(x_i))_\alpha |
\]

\[
\leq \frac{\Lambda_1 F \sqrt{K}}{n Z_1} \sum_{i=1}^n \| \nabla_{\phi} R_n(\phi_t)(x_i) \|_2
\]

\[
= \frac{\Lambda_1 F \sqrt{K}}{Z_1} \| \nabla_{\phi} R_n(\phi_t) \|_{L^2(\nu_X)}
\]

\[
\leq \frac{4}{\gamma_n} \Sigma \pi \Lambda_1 F \sqrt{K},
\]

where we used Proposition \[2\] for the last inequality.

We next give the following bound. For \(\forall i \in \{1, \ldots, n\},\)

\[
\| T_{k_{\pi_1}} \nabla_{\phi} R_n(\phi_t)(x_i) \|_2 = \left\| \frac{1}{n} \sum_{j=1}^n k_{\pi_1}(x_i, x_j) \nabla_{\phi} R_n(\phi_t)(x_j) \right\|_2
\]

\[
\leq \frac{K}{n} \sum_{j=1}^n \| \nabla_{\phi} R_n(\phi_t)(x_j) \|_2
\]

\[
= K \| \nabla_{\phi} R_n(\phi_t) \|_{L^2(\nu_X)}
\]

\[
\leq \frac{4}{\gamma_n} \Sigma \pi Z_1 K.
\]

Therefore, we get

\[
\| e_1 \|_{L^2(\nu_X)} = \left\| \frac{T_{k_{\pi_1}} \nabla_{\phi} R_n(\phi_t)}{Z_1} \right\|_{L^2(\nu_X)}
\]

\[
= \frac{1}{Z_1} \sqrt{\frac{1}{n} \sum_{i=1}^n \| T_{k_{\pi_1}} \nabla_{\phi} R_n(\phi_t)(x_i) \|_2^2}
\]

\[
\leq \frac{4}{\gamma_n} \Sigma \pi K.
\]

We here give the proof of Theorem \[2\]

Proof of Theorem \[2\] For arbitrary \(\epsilon > 0\), there exists a convex combination \(\psi_\epsilon = \sum_{s=1}^m \alpha_s \xi_s \in \text{conv}(D_{K_1})\) \((\xi_s \in D_{K_1}, \alpha_s > 0, \sum_{s=1}^m \alpha_s = 1)\) such that \(R_n(\psi_\epsilon) \leq R_n^* + \epsilon\). By Assumption \[2\](A2) and the symmetry of \(D_{K_1}\), it follows that for \(\forall s \in \{1, \ldots, m\},\)

\[
- \langle \nabla_{\phi} R_n(\phi_t), \bar{e}_s \rangle_{L^2(\nu_X)} \geq \langle \nabla_{\phi} R_n(\phi), -\xi_s \rangle_{L^2(\nu_X)}.
\]
Hence, we have

\[ \langle \nabla \varphi R_n(\phi_t), t_n \rangle_{L^2_2(\nu^*_n)} \leq \langle \nabla \varphi R_n(\phi), \psi_e \rangle_{L^2_2(\nu^*_n)}. \]  

(18)

By the induction hypothesis, we can easily show that

\[ \|\phi_t\|_{L^2_2(\nu^*_n)} \leq \max \left\{ \Lambda_\infty, \frac{8\Sigma^2 K}{\gamma_n} \right\} \geq \Lambda_\infty + \frac{8\Sigma^2 K}{\gamma_n}. \]  

(19)

Indeed, \( \|\phi_0\|_{L^2_2(\nu^*_n)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||x_i||^2} \leq \Lambda_\infty \) and

\[ \|\phi_{t+1}\|_{L^2_2(\nu^*_n)} = \|(1 - \eta_t)\phi_t + \eta_t t_n\|_{L^2_2(\nu^*_n)} \leq (1 - \eta_t)\|\phi_t\|_{L^2_2(\nu^*_n)} + \eta_t\|t_n\|_{L^2_2(\nu^*_n)} \leq (1 - \eta_t) \max \left\{ \Lambda_\infty, \frac{8\Sigma^2 K}{\gamma_n} \right\} + \frac{8\eta_t \Sigma^2 K}{\gamma_n} = \max \left\{ (1 - \eta_t)\Lambda_\infty, \frac{8\Sigma^2 K}{\gamma_n} \right\}, \]

where we used Proposition 3 for the second inequality.

From Proposition 1 and the above equations (18) and (19), we have

\[ R_n(\phi_{t+1}) = R_n((1 - \eta_t)\phi_t + \eta_t t_n) \]

\[ = R_n(\phi_t + \eta_t(t_n - \phi_t)) \]

\[ \leq R_n(\phi_t) + \eta_t \left\{ \nabla R_n(\phi_t), t_n - \phi_t \right\}_{L^2_2(\nu^*_n)} + \frac{\eta_t^2 L}{2} \|t_n - \phi_t\|^2_{L^2_2(\nu^*_n)} \]

\[ \leq R_n(\phi_t) + \eta_t \left\{ \nabla R_n(\phi_t), \psi_e - \phi_t \right\}_{L^2_2(\nu^*_n)} + \frac{\eta_t^2 L}{2} \|t_n - \phi_t\|^2_{L^2_2(\nu^*_n)} \]

\[ \leq R_n(\phi_t) + \eta_t (R_n(\psi_e) - R_n(\phi_t)) + \frac{\eta_t^2 L}{2} \|t_n - \phi_t\|^2_{L^2_2(\nu^*_n)} \]

\[ \leq (1 - \eta_t)R_n(\phi_t) + \eta_t R_n(\psi_e) + \frac{\eta_t^2 L}{2} \left( \Lambda_\infty + \frac{16\Sigma^2 K}{\gamma_n} \right)^2. \]

where for the first inequality, we used Assumption 1(b) and Proposition 1.

By subtracting \( R_n(\psi_e) \) from both sides of this inequality, we get

\[ R_n(\phi_{t+1}) - R_n(\psi_e) \leq (1 - \eta_t)(R_n(\phi_t) - R_n(\psi_e)) + \frac{\eta_t^2 L}{2} \left( \Lambda_\infty + \frac{16\Sigma^2 K}{\gamma_n} \right)^2. \]

We set

\[ a_t = R_n(\phi_t) - R_n(\psi_e), b = \frac{L}{2} \left( \Lambda_\infty + \frac{16\Sigma^2 K}{\gamma_n} \right)^2, C = \max\{t_0a_0, 4b\} \]

and consider the recursive inequality: \( a_{t+1} \leq (1 - \eta_t)a_t + b\eta_t^2 \)

Then, we show that by the induction hypothesis,

\[ a_t \leq \frac{C}{t + t_0}. \]  

(20)

By the definition \( C \), \( a_0 \leq C/t_0 \) holds clearly. Assuming (20) holds for \( t \), we get

\[ a_{t+1} \leq \left( 1 - \frac{2}{t + t_0} \right) \frac{C}{t + t_0} + \frac{4b}{(t + t_0)^2} \]

\[ = \frac{(t + t_0 - 2)C}{(t + t_0)^2} + \frac{4b}{(t + t_0)^2} \]
We next focus on the second term in the skeleton of generalization bound in Theorem 1. From the definition of $\psi$, we here prove a main result (Corollary 1). To do so, we give the following proposition which provides a bound on an empirical functional gradient norm by the loss function.

**Proposition A.** For $\forall f \in \mathcal{F}$, it follows that

$$
\mathcal{L}_n(f) \geq \frac{1}{2} \| \nabla f \mathcal{L}_n(f) \|_{L_1(\nu_n^x)}.
$$

**Proof.** By substituting $z = \frac{\exp(\zeta_n)}{\sum_{y' \in \mathcal{Y}} \exp(\zeta_{n,y'})}$ into the inequality: $-\log(z) \geq (1 - z)$ for $z \in (0, 1)$, we get

$$
l(\zeta, y) = -\log \left( \frac{\exp(\zeta_y)}{\sum_{y' \in \mathcal{Y}} \exp(\zeta_{n,y'})} \right) \geq 1 - \frac{\exp(\zeta_y)}{\sum_{y' \in \mathcal{Y}} \exp(\zeta_{n,y'})}.
$$

Thus, we get

$$
\mathcal{L}_n(f) \geq \frac{1}{n} \sum_{i=1}^{n} \left( 1 - \frac{\exp(f_{y_i}(x_i))}{\sum_{y' \in \mathcal{Y}} \exp(f_{y'_i}(x_i))} \right) = \frac{1}{2} \| \nabla f \mathcal{L}_n(f) \|_{L_1(\nu_n^x)},
$$

where we used equation (15) for the last equality.

The following is the proof for the main result in this study.

**Proof for Corollary 1** We set $f_T = \mathbf{w}^\top \phi_T$. From Proposition A, we get

$$
\mathcal{R}_n(\phi_T) = \mathcal{L}_n(f_T) \geq \frac{1}{2} \| \nabla f \mathcal{L}_n(f_T) \|_{L_1(\nu_n^x)}.
$$

Combining the inequality in Theorem 2, we get

$$
\| \nabla f \mathcal{L}_n(f_T) \|_{L_1(\nu_n^x)} = 2\mathcal{R}_n(\phi_T) \leq 2 \left( \mathcal{R}_n^* + \frac{C}{T + t_0} \right) \leq 2 \left( \epsilon_n + \frac{C}{T + t_0} \right).
$$

We next focus on the second term in the skeleton of generalization bound in Theorem 1. From the definition of $\eta_s$, we have

$$
\prod_{s=t+1}^{T-1} \left( 1 - \frac{\eta_s}{2} \right) = \prod_{s=t+1}^{T-1} \frac{s + t_0 - 1}{s + t_0} = \frac{t + t_0}{T - 1 + t_0}.
$$
Therefore, applying a norm bound on $C_i = A_i F_i$ in Proposition 3 and $D_i = G_i$, we see that,

$$
\frac{2c^3 \Lambda_w \Lambda_\infty}{\delta \sqrt{n}} \left( 1 + \sum_{t=0}^{T-1} \eta_t \Lambda_1^C \Lambda_1^D \prod_{s=t+1}^{T-1} \left( 1 - \frac{\eta_s}{2} \right) \right) = \frac{2c^3 \Lambda_w \Lambda_\infty}{\delta \sqrt{n}} \left( 1 + \sum_{t=0}^{T-1} \frac{2 \Lambda_1^C \Lambda_1^D}{T - 1 + t_0} \right)
$$

$$
= \frac{2c^3 \Lambda_w \Lambda_\infty}{\delta \sqrt{n}} \left( 1 + \frac{2T \Lambda_1^C \Lambda_1^D}{T - 1 + t_0} \right)
$$

$$
\leq \frac{2c^3 \Lambda_w \Lambda_\infty}{\delta \sqrt{n}} (1 + 2 \Lambda_1^C \Lambda_1^D)
$$

$$
\leq \frac{2c^3 \Lambda_w \Lambda_\infty}{\delta \sqrt{n}} \left( 1 + \frac{8 \sum_{\pi} \Lambda^F \Lambda^G \sqrt{K}}{\gamma_n} \right). \quad (22)
$$

Note that since $t_0 \geq 2$ and $\Lambda_1^C \Lambda_1^D = 4 \sum_{\pi} \Lambda^F \Lambda^G \sqrt{K} / \gamma_n \leq 1/2$, we can apply Theorem 1. Thus, combining (21) and (22) with Theorem 1, we get an explicit generalization bound and finish the proof.