On the Stability and Generalization of Triplet Learning

Jun Chen[1], Hong Chen[2, 3, 4, *], Xue Jiang[5], Bin Gu[6], Weifu Li[2, 3, 4], Tieliang Gong[7, 8], Feng Zheng[5]

[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]College of Science, Huazhong Agricultural University, Wuhan 430070, China
[3]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
[4]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan 430070, China
[5]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[6]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[7]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[8]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Ministry of Education, Xi'an 710049, China
chenh@mail.hzau.edu.cn

Abstract

Triplet learning, i.e. learning from triplet data, has attracted much attention in computer vision tasks with an extremely large number of categories, e.g., face recognition and person re-identification. Albeit with rapid progress in designing and applying triplet learning algorithms, there is a lacking study on the theoretical understanding of their generalization performance. To fill this gap, this paper investigates the generalization guarantees of triplet learning by leveraging the stability analysis. Specifically, we establish the first general high-probability generalization bound for the triplet learning algorithm satisfying the uniform stability, and then obtain the excess risk bounds of the order $O(n^{-\frac{1}{2}}\log n)$ for both stochastic gradient descent (SGD) and regularized risk minimization (RRM), where $2n$ is approximately equal to the number of training samples. Moreover, an optimistic generalization bound in expectation as fast as $O(n^{-1})$ is derived for RRM in a low noise case via the on-average stability analysis. Finally, our results are applied to triplet metric learning to characterize its theoretical underpinning.

Introduction

As two popular paradigms of machine learning, data-driven algorithms with pointwise loss and pairwise loss have been widely used to find the intrinsic relations from empirical observations. In the algorithmic implementation, the former (called pointwise learning) often aims to minimize the empirical risk characterized by the divergence between the predicted output and the observed response of each input (Vapnik 1998; Cucker and Smale 2001; Poggio et al. 2004), while the latter (called pairwise learning) usually concerns the model performance associated with pairs of training instances, see e.g., ranking (Agarwal and Niyogi 2009) and metric learning (Xing et al. 2002; Ying and Li 2012).

Despite enjoying the advantages of feasible implementations and solid foundations, pointwise learning and pairwise learning may face a crucial challenge for learning tasks with an extremely large number of categories. Such learning scenarios appear in face recognition (Schroff, Kalenichenko, and Philbin 2015; Ding and Tao 2018), person re-identification (Ustinova and Lempitsky 2016; Cheng et al. 2016; Xiao et al. 2016), image retrieval (Lai et al. 2015; Huang et al. 2015) and other individual level fine-grained tasks (Wohlhart and Lepetit 2015; Simo-Serra et al. 2015). As illustrated in Yu et al. (2018), the traditional learning model is difficult to achieve good performance in the setting of an extremely large number of categories since its parameters will increase linearly with the number of categories. To surmount this barrier, many triplet learning algorithms are formulated by injecting triplet loss function into the metric learning framework (Schroff, Kalenichenko, and Philbin 2015; Ustinova and Lempitsky 2016; Cheng et al. 2016; Xiao et al. 2016; Ding and Tao 2018). For triplet metric learning (Schroff, Kalenichenko, and Philbin 2015; Ge et al. 2018), the implementation procedures mainly include: 1) Constructing triplets associated with anchor sample, positive sample and negative sample; 2) Designing margin-based empirical risk associated with triplet loss; 3) Learning metric space transformation rule via empirical risk minimization (ERM), which aims to minimize intra-class distance and maximize inter-class distance simultaneously. However, the triplet characteristic often leads to a heavy computational burden for large-scale data. Recently, stochastic gradient descent (SGD) is employed for deploying triplet learning algorithms due to its low time complexity (Schroff, Kalenichenko, and Philbin 2015; Ge et al. 2018). Although there has been significant progress in designing and applying triplet learning algorithms, little work has been done to recover their generalization guarantees from the lens of statistical learning theory (SLT) (Vapnik 1998).

The generalization guarantee of learning algorithm is the core of SLT, which evaluates the prediction ability in the unseen inputs (Vapnik 1998; Cucker and Zhou 2007). In a nutshell, there are three branches of generalization analysis including *uniform convergence approaches* associated with hypothesis space capacity (e.g., VC dimension (Vapnik 1998), covering numbers (Cucker and Zhou 2007; Chen et al. 2017), Rademacher complexity (Bartlett and Mendelson 2001)), *operator approximation technique* (Smale and

[*]Corresponding author. http://arxiv.org/abs/2302.09815.

Zhou 2007; Rosasco, Belkin, and Vito 2010), and *algorithmic stability analysis* (Bousquet and Elisseeff 2002; Elisseeff, Evgeniou, and Pontil 2005; Shalev-Shwartz et al. 2010). It is well known that the stability analysis enjoys nice properties on flexibility (independent of the capacity of hypothesis function space) and adaptivity (suiting for rich learning scenarios, e.g., classification and regression (Hardt, Recht, and Singer 2016), ranking (Agarwal and Niyogi 2009), and adversarial training (Xing, Song, and Cheng 2021)). Recently, besides learning algorithms based on ERM and regularized risk minimization (RRM), generalization and stability have been understood for SGD of pointwise learning (Hardt, Recht, and Singer 2016; Roux, Schmidt, and Bach 2012; Fehrman, Gess, and Jentzen 2020; Lei, Hu, and Tang 2021) and pairwise learning (Lei, Ledent, and Kloft 2020; Arous, Gheissari, and Jagannath 2021; Lei, Liu, and Ying 2021). While the existing extensive works on stability analysis, to our best knowledge, there is no related result of SGD and RRM for triplet learning.

To fill the above gap, this paper aims to provide stability-based generalization analysis for a variety of triplet learning algorithms. We establish generalization bounds for SGD and RRM with triplet loss, which yield comparable convergence rates as pointwise learning (Feldman and Vondrák 2019) and pairwise learning (Lei, Ledent, and Kloft 2020) under mild conditions. The main contributions of this paper are summarized as follows.

- *Generalization by algorithmic stability for triplet learning*. After introducing a new definition of triplet uniform stability, we establish the first general high-probability generalization bound for triplet learning algorithms satisfying uniform stability, motivated by the recent analysis for pairwise learning (Lei, Ledent, and Kloft 2020). Especially, the current analysis just requires the uniform stability of the triplet learning algorithm and the boundedness of loss function in expectation.
- *Generalization bounds for triplet SGD and triplet RRM*. Generalization properties are characterized for SGD and RRM of triplet learning when the loss function is (strongly) convex, L-Lipschitz and α-smooth. Particularly, the derived excess risk bounds are with the decay rate $O(n^{-\frac{1}{2}}\log n)$ as $n_+ \asymp n_- \asymp n$, where n_+ and n_- are the numbers of positive samples and negative samples, respectively. Moreover, for the strongly convex loss function, the refined generalization bound with the order $O(n^{-1})$ is derived for RRM by leveraging the triplet on-average stability. To the best of our knowledge, these results are the first generalization bounds of SGD and RRM for triplet learning.

Related Work

In this section, we briefly review the related works on triplet learning and algorithmic stability.

Triplet Learning. The main purpose of deep metric learning is to directly learn a feature representation vector from input data with the help of deep neural networks. Bromley et al. (1993) found that the relationship between samples can be measured by the difference between the corresponding embedded vectors, and some deep metric learning models have been subsequently proposed (Chopra, Hadsell, and LeCun 2005; Hadsell, Chopra, and LeCun 2006). Later, Schroff, Kalenichenko, and Philbin (2015) proposed the FaceNet by integrating the idea of triplet learning (Schultz and Joachims 2003; Weinberger, Blitzer, and Saul 2005) and deep metric learning together. In contrast to the previous approaches, FaceNet directly trains its output to be a compact 128-D embedding vector using a triplet loss function based on large margin nearest neighbor (Weinberger, Blitzer, and Saul 2005), and it is implemented by employing the SGD strategy. Encouraged by the impressive performance of FaceNet, lots of learning algorithms with triplet loss have been formulated in the computer version field (Cheng et al. 2016; Xiao et al. 2016; Ustinova and Lempitsky 2016; Liu et al. 2016; Ramanathan et al. 2015; Ding and Tao 2018). Although there have been significant works on designing triplet metric learning algorithms, our theoretical understanding of their generalization ability falls far below the experimental validations.

Generalization and Algorithmic Stability. In SLT, uniform convergence analysis focuses on bounding the uniform deviation between training error and testing error over hypothesis space (Vapnik 1998; Cucker and Smale 2001; Bartlett and Mendelson 2001; Wang et al. 2020; Chen et al. 2021), and operator approximation approach is inspired by functional analysis theory (Smale and Zhou 2007; Rosasco, Belkin, and Vito 2010). Indeed, the former depends on the capacity of hypothesis space (e.g., VC dimension (Vapnik 1998), covering numbers (Cucker and Zhou 2007), Rademacher complexity (Bartlett and Mendelson 2001)), and the latter is limited to some special models enjoying operator representation (e.g., regularized least squares regression (Smale and Zhou 2007), regularized least squares ranking (Chen 2012)). Different from the above routes, algorithmic stability is described by the gap among training errors of different training sets, which is dimension-independent and enjoys adaptivity for wide learning models. The concept of algorithmic stability can be put forward as early as the 1970s (Rogers and Wagner 1978), and its learning theoretical framework was established in Bousquet and Elisseeff (2002) and Elisseeff, Evgeniou, and Pontil (2005). In essential, the algorithmic uniform stability is closely related to the learnability (Poggio et al. 2004; Shalev-Shwartz et al. 2010). For the pointwise learning setting, the stability-based generalization guarantees have been stated in terms of uniform stability (Hardt, Recht, and Singer 2016; Foster et al. 2019), on-average stability (Kuzborskij and Lampert 2018; Lei and Ying 2021), local elastic stability (Deng, He, and Su 2021) and argument stability (Bassily et al. 2020; Lei and Ying 2020; Liu et al. 2017). For the pairwise learning setting, there are fine-grained analyses on the generalization and stability of SGD and RRM (Shen et al. 2019; Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021). Due to the space limitation, we further summarize different definitions and properties of algorithmic stability in *Supplementary Material C*. Along this line of the above corpus, it is natural to investigate the generalization bounds of triplet learning by algorithmic stability analysis.

Preliminaries

This section introduces the necessary backgrounds on triplet learning and algorithmic stability. The main notations used in this paper are stated in *Supplementary Material A*.

Triplet Learning

Let $\mathcal{X}_+, \mathcal{X}_- \subset \mathbb{R}^d$ are two d-dimensional input spaces and $\mathcal{Y} \subset \mathbb{R}$ is an output space. We give the training set $S := \{z_i^+ := (x_i^+, y_i^+)\}_{i=1}^{n_+} \cup \{z_j^- := (x_j^-, y_j^-)\}_{j=1}^{n_-} \in \mathcal{Z}$ with $\mathcal{Z} := \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-}$, where each positive sample z_i^+ and negative sample z_j^- are drawn independently from $\mathcal{Z}_+ := \mathcal{X}_+ \times \mathcal{Y}$ and $\mathcal{Z}_- := \mathcal{X}_- \times \mathcal{Y}$, respectively. Note that there are likely more than two classes in positive and negative sample spaces. Given empirical observation S, triplet learning algorithms usually aim to find a model $h_w : \mathcal{X}_+ \times \mathcal{X}_+ \times \mathcal{X}_- \to \mathbb{R}$ such that the expectation risk

$$R(w) := \mathbb{E}_{z^+, \tilde{z}^+, z^-} \ell(w; z^+, \tilde{z}^+, z^-) \tag{1}$$

is as small as possible. Here the model parameter $w \in \mathcal{W}$ with d'-dimensional parameter space $\mathcal{W} \subseteq \mathbb{R}^{d'}$, $\mathbb{E}_z$ denotes the conditional expectation with respect to (w.r.t.) z, and the triplet loss function $\ell : \mathcal{W} \times \mathcal{Z}_+ \times \mathcal{Z}_+ \times \mathcal{Z}_- \to \mathbb{R}_+$ is used to measure the difference between model's prediction and corresponding real observation. Since the intrinsic distributions generating z^+ and z^- are same and unknown, it is impossible to implement triplet learning by minimizing the objective $R(w)$ directly. Naturally, we consider the corresponding empirical risk of (1) defined as

$$R_S(w) := \frac{1}{n_+(n_+ - 1)n_-} \sum_{\substack{i,j \in [n_+], i \neq j, \\ k \in [n_-]}} \ell(w; z_i^+, z_j^+, z_k^-) \tag{2}$$

for algorithmic design, where $[n] := \{1, ..., n\}$. Clearly, the triplet learning algorithms, built from $R_S(w)$ in (2), are much more complicated than the corresponding ones in pointwise learning and pairwise learning.

In the sequel, for the given algorithm A and the training data S, we denote the corresponding output model parameter as $A(S)$ for feasibility. In triplet learning, we usually build predictors by optimizing the models measured by the empirical risk $R_S(A(S))$ or its variants. However, the nice empirical performance of learning model does not guarantee its effectiveness in unseen observations. In SLT, it is momentous and fundamental to bound generalization error, i.e. $R(w) - R_S(w)$, since it characterizes the gap between the population risk $R(w)$ and its empirical estimator $R_S(w)$. Despite the existing rich studies for pointwise learning and pairwise learning, the generalization bound of triplet learning is rarely touched in the machine learning community. In this paper, we pioneer the generalization analysis of triplet SGD and RRM to understand their learnability.

Triplet Algorithmic Stability

An algorithm $A : \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-} \to \mathcal{W}$ is stable if the model parameter $A(S)$ is insensitive to the slight change of training set S. Various definitions of algorithmic stability have been introduced from different motivations (see *Supplementary Material C*), where uniform stability and on-average stability are popular for studying the generalization bounds of SGD and RRM (Hardt, Recht, and Singer 2016; Lin, Camoriano, and Rosasco 2016; Kuzborskij and Lampert 2018; Lei and Ying 2021; Lei, Ledent, and Kloft 2020). Following this line, we extend the previous definitions of uniform stability and on-average stability to the triplet learning setting.

Definition 1. *(Uniform Stability). Assume any training datasets* $S = \{z_1^+, ..., z_{n_+}^+, z_1^-, ..., z_{n_-}^-\}, \bar{S} = \{\bar{z}_1^+, ..., \bar{z}_{n_+}^+, \bar{z}_1^-, ..., \bar{z}_{n_-}^-\} \in \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-}$ *are differ by at most a single sample. A deterministic algorithm* $A : \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-} \to \mathcal{W}$ *is called* γ*-uniformly stable if*

$$\sup_{\substack{z^+, \tilde{z}^+ \in \mathcal{Z}_+, \\ z^- \in \mathcal{Z}_-}} |\ell(A(S); z^+, \tilde{z}^+, z^-) - \ell(A(\bar{S}); z^+, \tilde{z}^+, z^-)| \leq \gamma$$

for any training datasets $S, \bar{S} \in \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-}$ *that differ by at most a single sample.*

Definition 1 coincides with the uniform stability definitions for pointwise learning (Hardt, Recht, and Singer 2016) and pairwise learning (Lei, Ledent, and Kloft 2020), except for the triplet loss involving two sample spaces $\mathcal{Z}_+$ and $\mathcal{Z}_-$.

Definition 2. *(On-average Stability). Let* $S_{i,j,k} = \{z_1^+, ..., z_{i-1}^+, \bar{z}_i^+, z_{i+1}^+, ..., z_{j-1}^+, \bar{z}_j^+, z_{j+1}^+, ..., z_{n_+}^+, z_1^-, ..., z_{k-1}^-, \bar{z}_k^-, z_{k+1}^-, ..., z_{n_-}^-\}, i, j \in [n_+], i \neq j, k \in [n_-]$. *A deterministic algorithm* $A : \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-} \to \mathcal{W}$ *is called* γ*-on-average stable if*

$$\frac{1}{n_+(n_+ - 1)n_-} \sum_{\substack{i,j \in [n_+], i \neq j, \\ k \in [n_-]}} \mathbb{E}_{S, \bar{S}} \big[\ell(A(S_{i,j,k}); z_i^+, z_j^+, z_k^-) - \ell(A(S); z_i^+, z_j^+, z_k^-) \big] \leq \gamma.$$

Compared with the existing ones for pointwise learning (Kuzborskij and Lampert 2018) and pairwise learning (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021), Definition 2 considers much more complicated perturbations of training set S involving three samples. Definition 2 takes the expectation over S and $\bar{S}$, and takes the average over perturbations, which is weaker than the uniform stability described in Definition 1.

Main Results

This section states our main results on generalization bounds for triplet learning by stability analysis. We show a general high-probability generalization bound of triplet learning algorithms firstly, and then apply it to two specific algorithms, i.e. SGD and RRM. Finally, the on-average stability is employed for getting an optimistic generalization bound of RRM in expectation. All the proofs are provided in *Supplementary Material B* due to the space limitation.

Similar with the previous analyses (Hardt, Recht, and Singer 2016; Lei, Ledent, and Kloft 2020), our results are closely related to the following properties of triplet loss function.

Definition 3. *For a triplet loss function* $\ell : \mathcal{W} \times \mathcal{Z}_+ \times \mathcal{Z}_+ \times \mathcal{Z}_- \to \mathbb{R}_+$, *denote by* $\nabla \ell(w) := \nabla \ell(w; z^+, \tilde{z}^+, z^-)$

its gradient w.r.t. the model parameter $w \in \mathcal{W}$ and denote by $\|\cdot\|$ a norm on an inner product space which satisfies $\|\cdot\|^2 = \langle \cdot, \cdot \rangle$. Let $\sigma \geq 0$ and $L, \alpha > 0$.

1) The triplet loss ℓ is σ-strongly convex if, for all $w, w' \in \mathcal{W}$,

$$\ell(w) \geq \ell(w') + \langle \nabla \ell(w'), w - w' \rangle + \frac{\sigma}{2}\|w - w'\|^2.$$

2) The triplet loss ℓ is L-Lipschitz if

$$|\ell(w) - \ell(w')| \leq L\|w - w'\|, \forall w, w' \in \mathcal{W}.$$

3) The triplet loss ℓ is α-smooth if

$$\|\nabla \ell(w) - \nabla \ell(w')\| \leq \alpha\|w - w'\|, \forall w, w' \in \mathcal{W}.$$

When $\sigma = 0$, ℓ is convex which also implies that $\nabla^2 \ell(w; z^+, \tilde{z}^+, z^-) > 0$. It is easy to verify that logistic loss, least square loss and Huber loss are convex and smooth. Meanwhile, we observe that hinge loss, logistic loss and Huber loss are convex and Lipschitz (Hardt, Recht, and Singer 2016; Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021).

Stability-based Generalization Bounds

This subsection establishes the connection between uniform stability and generalization with high probability for triplet learning. Although rich results on the relationship between stability and generalization, the previous results do not hold directly for triplet learning due to its complicated loss structure. This difficulty is tackled by implementing much more detailed error decomposition and developing the analysis technique of Lei, Ledent, and Kloft (2020).

Lemma 1. *If $A : \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-} \to \mathcal{W}$ is γ-uniformly stable, for any $S, \bar{S}$, we have*

$$|\ell(A(S); z^+, \tilde{z}^+, z^-) - \ell(A(S_{i,j,k}); z^+, \tilde{z}^+, z^-)| \leq 3\gamma$$

for all $z^+, \tilde{z}^+ \in \mathcal{Z}_+, z^- \in \mathcal{Z}_-$, where $S, \bar{S}, S_{i,j,k}$ is defined in Definition 2 for any $i, j \in [n_+], i \neq j, k \in [n_-]$.

Lemma 1 illustrates that an upper bound of the change of the loss function still exists even after changing multiple samples of the training set. Here, the upper bound 3γ reflects the sensitivity of triplet learning w.r.t. the perturbation of training data.

It is a position to state our first general generalization bound with high probability for the uniformly stable triplet learning algorithm A. Detailed proof can be found in *Supplementary Material B.1*.

Theorem 1. *Assume that $A : \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-} \to \mathcal{W}$ is γ-uniformly stable. Let constant $M > 0$ and, for all $z^+, \tilde{z}^+ \in \mathcal{Z}_+$ and $z^- \in \mathcal{Z}_-$, let $|\mathbb{E}_S \ell(A(S); z^+, \tilde{z}^+, z^-)| \leq M$. Then, for all $\delta \in (0, 1/e)$, we have*

$$\begin{aligned} &|R_S(A(S)) - R(A(S))| \\ \leq\ & 6\gamma + e\Bigg(8M\Big(\frac{1}{\sqrt{n_-}} + \frac{2}{\sqrt{n_+ - 1}}\Big)\sqrt{\log(e/\delta)} \\ &+ 24\sqrt{2}\gamma\Big(\lceil \log_2(n_-(n_+ - 1)^2)\rceil + 2\Big)\log(e/\delta)\Bigg) \end{aligned}$$

with probability $1 - \delta$, where $\lceil n \rceil$ denotes the minimum integer no smaller than n and e denotes the base of the natural logarithm.

Remark 1. *Theorem 1 demonstrates the generalization performance of triplet learning depends heavily on the sample numbers n_+, n_- and the stability parameter γ, which extends the Theorem 1 of Lei, Ledent, and Kloft (2020) for pairwise learning to the triplet learning setting. Denote $x \asymp y$ as $ay < x \leq by$ for some constants $a, b > 0$. In particular, when $n_+ \asymp n_- \asymp n$, the high-probability bound in Theorem 1 can be rewritten as $O(n^{-\frac{1}{2}} + \gamma \log n)$, which is comparable with the previous analyses (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021).*

Generalization Bounds for SGD

Let $w_1 \in \mathcal{W}$ and let $\nabla \ell(w)$ be the subgradient of triplet loss ℓ w.r.t. the argument w. For triplet learning by SGD, at the t-th iteration, we draw (i_t, j_t, k_t) randomly and uniformly over $\{(i_t, j_t, k_t) : i_t, j_t \in [n_+], i_t \neq j_t, k_t \in [n_-]\}$, and update the model parameter w_t by

$$w_{t+1} = w_t - \eta_t \nabla \ell(w_t; z_{i_t}^+, z_{j_t}^+, z_{k_t}^-), \tag{3}$$

where $\{\eta_t\}_t$ is a sequence of step sizes.

To apply Theorem 1, we need to bound the uniform stability parameter of (3). Denote by $\mathbb{I}[\cdot]$ the indicator function which takes 1 if the situation in the brackets is satisfied and takes 0 otherwise.

Lemma 2. *Assume that $S, \bar{S} \in \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-}$ are different only in the last positive sample (or negative sample). Suppose $\ell(w; z^+, \tilde{z}^+, z^-)$ is convex, α-smooth and L-Lipschitz w.r.t. $\|\cdot\|, \forall z^+, \tilde{z}^+ \in \mathcal{Z}_+, z^- \in \mathcal{Z}_-$. If $\eta_t \leq 2/\alpha$, then SGD in (3) with t-th iteration is γ-uniformly stable, where*

$$\begin{aligned} \gamma \leq & 2L^2 \sum_{l=1}^{t} \eta_l \mathbb{I}\Big[(i_l = n_+ \text{ or } j_l = n_+, i_l \neq j_l, k_l \in [n_-], z_{n_+}^+ \\ & \neq \bar{z}_{n_+}^+) \text{ or } (i_l, j_l \in [n_+], i_l \neq j_l, k_l = n_-, z_{n_-}^- \neq \bar{z}_{n_-}^-)\Big]. \end{aligned}$$

In Lemma 2, we just consider the perturbation on the last positive (or negative) sample without loss of generality. The above uniform stability bound of SGD involves an indicator function associated with S and $\bar{S}$, which is nonzero only when different triplets are used.

Now we state the generalization bounds for SGD (3). The proof is present in *Supplementary Material B.2*.

Theorem 2. *Let the loss function $\ell(w; z^+, \tilde{z}^+, z^-)$ is convex, α-smooth and L-Lipschitz for all $z^+, \tilde{z}^+ \in \mathcal{Z}_+, z^- \in \mathcal{Z}_-$ and $|\mathbb{E}_S \ell(w_T; z^+, \tilde{z}^+, z^-)| \leq M$, where w_T is produced by SGD (3) with $\eta_t \equiv c/\sqrt{T}$ and constant $c \leq 2/\alpha$. For any $\delta \in (0, 1/e)$, with probability $1 - \delta$ we have*

$$\begin{aligned} &|R_S(w_T) - R(w_T)| \\ =& O\Bigg(\Big(\lceil \log(n_-(n_+ - 1)^2)\rceil + 2\Big)\log(1/\delta)\Big(\sqrt{\frac{\log(1/\delta)}{\max\{\frac{T}{n_+}, \frac{T}{n_-}\}}} \\ &+ 1\Big)\Big(\frac{\sqrt{T}}{n_+} + \frac{\sqrt{T}}{n_-}\Big) + \Big(\frac{1}{\sqrt{n_-}} + \frac{1}{\sqrt{n_+ - 1}}\Big)\sqrt{\log(1/\delta)}\Bigg). \end{aligned}$$

Remark 2. *Theorem 2 demonstrates that the generalization error of (3) relies on the numbers of positive and negative*

Algorithm	Reference	Assumptions			Tool	Convergence rate
		Convex	Lipschitz	Smooth		
SGD (▲)	Hardt, Recht, and Singer (2016)	√	√	√	Uniform stability	$O(n^{-\frac{1}{2}})$
	Lei and Ying (2020)	√	×	√	On-average model stability	$O(n^{-1})$
SGD (▲▲)	Lei, Ledent, and Kloft (2020)	√	√	√	Uniform stability	$*O(n^{-\frac{1}{2}}\log n)$
	Lei, Liu, and Ying (2021)	√	×	√	On-average model stability	$O(n^{-1})$
	Lei, Liu, and Ying (2021)	√	√	×	On-average model stability	$O(n^{-\frac{1}{2}})$
	Yang et al. (2021)	√	√	×	Uniform stability	$O(n^{-\frac{1}{2}})$
	Yang et al. (2021)	√	√	√	Uniform stability	$O(n^{-\frac{1}{2}})$
SGD (▲▲▼)	Ours ($n_+ \asymp n_- \asymp n$)	√	√	√	Uniform stability	$*O(n^{-\frac{1}{2}}\log n)$

Table 1: Summary of stability-based generalization analyses of SGD in the setting of convexity (▲-pointwise; ▲▲-pairwise; ▲▲▼-triplet; √-the reference has such a property; ×-the reference hasn't such a property; ∗-high-probability bound).

training samples (i.e. n_+, n_-) and the iterative steps T. Our result also uncovers that the balance of positive and negative training samples is crucial to guarantee the generalization of triplet learning algorithms. When $n_+ \asymp n_- \asymp n$, we get the high-probability bound $|R_S(w_T) - R(w_T)| = O(n^{-\frac{1}{2}}\log n)$, which is consistent with Theorem 4 in Lei, Ledent, and Kloft (2020) for pariwise SGD.

Remark 3. *Let $w_R^* = \arg\min_{w\in\mathcal{W}} R(w)$. We can deduce that*

$$R(w_T) - R(w_R^*) = \big(R(w_T) - R_S(w_T)\big) + \big(R_S(w_T) - R_S(w_R^*)\big) + \big(R_S(w_R^*) - R(w_R^*)\big). \quad (4)$$

As illustrated in previous studies (Bottou and Bousquet 2007; Lei, Ledent, and Kloft 2020; Lei and Ying 2020), the first two terms in (4) *are called the estimation error and optimization error, respectively. Theorem 2 guarantees the upper bound of estimation error with $O(n^{-\frac{1}{2}}\log n)$ and Harvey et al. (2019) states the upper bound of the optimization error with $O(T^{-\frac{1}{2}}\log T)$. The third term on the right side of* (4) *can be bounded by Bernstein's inequality for U-statistics (Pitcan 2017), which is present in the following Lemma 3.*

Lemma 3. *Let $b = \sup_{z^+,\tilde{z}^+,z^-}|\ell(w;z^+,\tilde{z}^+,z^-)|$ and τ be the variance of $\ell(w;z^+,\tilde{z}^+,z^-)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$|R_S(w) - R(w)| \leq \frac{2b\log(1/\delta)}{3\lfloor n_+/2\rfloor} + \sqrt{\frac{2\tau\log(1/\delta)}{\lfloor n_+/2\rfloor}} + \frac{2b\log(1/\delta)}{3\lfloor n_-\rfloor} + \sqrt{\frac{2\tau\log(1/\delta)}{\lfloor n_-\rfloor}},$$

where $\lfloor n\rfloor$ denotes the maximum integer no larger than n.

Under mild conditions, i.e., $b = O(\sqrt{n})$ and $n_+ \asymp n_- \asymp n$, we get $R_S(w_R^*) - R(w_R^*) = O\big(\frac{\log(1/\delta)}{\sqrt{n}} + \sqrt{\frac{\tau\log(1/\delta)}{n}}\big) = O(n^{-\frac{1}{2}})$. Combining this with the bounds of estimation error and optimization error in Remark 3, we deduce that the excess risk $R(w_T) - R(w_R^*) = O(n^{-\frac{1}{2}}\log n)$ as $T \asymp n$.

Remark 4. *To better highlight the characteristics of Theorem 2, we compare it with the generalization analyses in the setting of convexity (Hardt, Recht, and Singer 2016; Lei and Ying 2020; Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021; Yang et al. 2021) in Table 1. Clearly, our learning theory analysis is novel since it is the first touch for SGD under the triplet learning setting. When $n_+ \asymp n_- \asymp n$, the derived result is comparable with the previous convergence rates (Hardt, Recht, and Singer 2016; Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021).*

Generalization Bounds for RRM

We now turn to study the generalization properties of RRM for triplet learning. Detailed proofs are stated in *Supplementary Material B.3*. Let $r : \mathcal{W} \to \mathbb{R}_+$ be a regularization penalty for increasing the data-fitting ability of ERM. For any datatset $S \in \mathcal{Z}_+^{n_+} \cup \mathcal{Z}_-^{n_-}$ and $R_S(w)$ defined in (2), the derived model parameter of RRM is the minimizer of

$$F_S(w) := R_S(w) + r(w) \quad (5)$$

over $w \in \mathcal{W}$ and $F(w) := R(w) + r(w)$.

To apply Theorem 1, we also need to verify the stable parameter of RRM (5).

Lemma 4. *Assume that $F_S(w)$ is σ-strongly convex w.r.t. $\|\cdot\|$ and $\ell(w;z^+,\tilde{z}^+,z^-)$ is convex and L-Lipschitz. Then, the RRM algorithm A defined as $A(S)=\arg\min_{w\in\mathcal{W}} F_S(w)$ is γ-uniformly stable with $\gamma=\min\Big\{\frac{8}{n_+},\frac{4}{n_-}\Big\}\frac{L^2}{\sigma}$.*

When $n_+ \asymp n_- \asymp n$, the uniform stability parameter is $O(\frac{L^2}{n\sigma})$, which coincides with the previous analysis for pariwise learning (Lei, Ledent, and Kloft 2020). To tackle the triplet structure, the current analysis involves elaborate error decomposition and the deduce strategy of Lemma B.2 in Lei, Ledent, and Kloft (2020).

It is required in Theorem 1 that we assume the triplet loss for a uniformly stable algorithm is bounded in expectation. To get the necessary guarantee, we introduce the following Lemma 5, which can be proved coherently by utilizing Lemma 2 (Lei, Ledent, and Kloft 2020) and the Lipschitz continuity of the loss function ℓ.

Lemma 5. *Let $F_S(w)$ be σ-strongly convex w.r.t. $\|\cdot\|$, $w^*=\arg\min_{w\in\mathcal{W}} F(w)$, and, for all $z^+,\tilde{z}^+\in\mathcal{Z}_+, z^-\in\mathcal{Z}_-$, let $\tilde{\ell}(A(S);z^+,\tilde{z}^+,z^-)=\ell(A(S);z^+,\tilde{z}^+,z^-)-\ell(w^*;z^+,\tilde{z}^+,z^-)$. If the RRM algorithm A measured by loss function ℓ is γ-uniformly stable, then A measured by loss function $\tilde{\ell}$ is also γ-uniformly stable and*

$$|\mathbb{E}_S\tilde{\ell}(A(S);z^+,\tilde{z}^+,z^-)|\le M:=\min\Big\{\frac{4\sqrt{6}}{\sqrt{n_+}},\frac{4\sqrt{3}}{\sqrt{n_-}}\Big\}\frac{L^2}{\sigma}.$$

Theorem 3. *Assume that $F_S(w)$ is σ-strongly convex w.r.t. $\|\cdot\|$, and $\ell(w;z^+,\tilde{z}^+,z^-)$ is convex and L-Lipschitz and $\sup_{z^+,\tilde{z}^+,z^-}|\ell(w^*;z^+,\tilde{z}^+,z^-)|\le O(\sqrt{n})$. Let the variance of $\ell(w^*;z^+,\tilde{z}^+,z^-)$ is less than a positive constant τ. For the RRM algorithm A defined as and any $\delta\in(0,1/e)$, we have*

$$\begin{aligned}&|R_S(A(S))-R(A(S))|\\=&O\Bigg(\sigma^{-1}\Big(\min\Big\{\frac{\sqrt{2}}{\sqrt{n_+}},\frac{1}{\sqrt{n_-}}\Big\}\Big(\frac{1}{\sqrt{n_-}}+\frac{1}{\sqrt{n_+}}\Big)\sqrt{\log\frac{1}{\delta}}\\&+\min\Big\{\frac{2}{n_+},\frac{1}{n_-}\Big\}\log\big(n_-n_+^2\big)\log\frac{1}{\delta}\Big)+\sqrt{\frac{\log\frac{1}{\delta}}{n_+}}+\sqrt{\frac{\log\frac{1}{\delta}}{n_-}}\Bigg)\end{aligned}$$

with probability $1-\delta$.

Remark 5. *If $n_+\asymp n_-\asymp n$, the above bound is equivalent to $O\big(n^{-\frac{1}{2}}+(n\sigma)^{-1}\log n\big)$. Due to the definitions of $F_S(w)$ and $A(S)$ and $r(A(S))\ge 0$, we deduce that the excess risk $R(A(S))-R(w^*)\le R(A(S))-R_S(A(S))+R_S(w^*)-R(w^*)+r(w^*)$. Analogous to the third term to the right of (4), we have $R_S(w^*)-R(w^*)=O(\frac{\log(1/\delta)}{\sqrt{n}}+\sqrt{\frac{\tau\log(1/\delta)}{n}})$ with probability $1-\delta$. Therefore, the excess risk bound is $O(n^{-\frac{1}{2}}\log n)$ when $r(w^*)=O(\sigma\|w^*\|^2)$ and $\sigma\asymp n^{-\frac{1}{2}}$. Note that the reason for $r(w^*)=O(\sigma\|w^*\|^2)$ can be found in the last part of Supplementary Material B.3.*

Optimistic Generalization Bounds for RRM

In this part, we use the on-average stability in Definition 2 and some properties of smoothness to establish the optimistic generalization bounds of RRM in the low noise case. Different from the above theorems, we do not require the Lipschitz continuity condition for the triplet loss function.

The following lemma establishes the relationship between the estimation error and the model perturbation induced by the change at a single point of the training set.

Lemma 6. *Assume that for all $z^+,\tilde{z}^+\in\mathcal{Z}_+, z^-\in\mathcal{Z}_-$ and $w\in\mathcal{W}$, the loss function $\ell(w;z^+,\tilde{z}^+,z^-)$ is convex and α-smooth w.r.t. $\|\cdot\|$. Then, for all $\epsilon>0$,*

$$\begin{aligned}&\mathbb{E}_S[R(A(S))-R_S(A(S))]\\\le&\frac{3(\epsilon+\alpha)}{2n_+(n_+-1)n_-}\sum_{\substack{i\in[n_+],\\k\in[n_-]}}\Big(2\mathbb{E}_{S,\bar{S}}\|A(S_i)-A(S)\|^2\\&+\mathbb{E}_{S,\bar{S}}\|A(S_k)-A(S)\|^2\Big)+\frac{\alpha\mathbb{E}_S R_S(A(S))}{\epsilon},\end{aligned}$$

where $S_i=\{z_1^+,...,z_{i-1}^+,\bar{z}_i^+,z_{i+1}^+,...,z_{n_+}^+,z_1^-,...,z_{n_-}^-\}$ and $S_k=\{z_1^+,...,z_{n_+}^+,z_1^-,...,z_{k-1}^-,\bar{z}_k^-,z_{k+1}^-,...,z_{n_-}^-\}$.

From the proof of Lemma 6 (see *Supplementary Material B.4*) and Definition 2, we know the upper bound in Lemma 6 provides the selection of on-average stability parameter γ. After establishing the connection between $\mathbb{E}_{S,\bar{S}}\|A(S_i)-A(S)\|^2$ (or $\mathbb{E}_{S,\bar{S}}\|A(S_k)-A(S)\|^2$) and $\mathbb{E}_S R_S(A(S))$, we get the following error bound of RRM.

Theorem 4. *Assume that the loss function $\ell(w;z^+,\tilde{z}^+,z^-)$ is convex and α-smooth for all $z^+,\tilde{z}^+\in\mathcal{Z}_+, z^-\in\mathcal{Z}_-$ and $w\in\mathcal{W}$, and $F_S(w)$ is σ-strongly convex w.r.t. $\|\cdot\|$ with $S\in\mathcal{Z}_+^{n_+}\cup\mathcal{Z}_-^{n_-}$. Let $\sigma\min\{n_+,n_-\}\ge 8\alpha$ and $A(S)=\arg\min_{w\in\mathcal{W}} F_S(w)$. Then, for all $\epsilon>0$,*

$$\begin{aligned}&\mathbb{E}_S[F(A(S))-F_S(w^*)]\le\mathbb{E}_S[R(A(S))-R_S(A(S))]\\\le&\Big(\frac{\alpha}{\epsilon}+\frac{1536\alpha(\epsilon+\alpha)}{n_+^2(n_+-1)\sigma^2}+\frac{256\alpha(\epsilon+\alpha)}{3(n_+-1)n_-^2\sigma^2}\Big)\mathbb{E}_S R_S(A(S)).\end{aligned}$$

Remark 6. *The upper bounds in Theorem 4 are closely related to the empirical risk $\mathbb{E}_S R_S(A(S))$. It is reasonable to assume that the empirical risk of $A(S)$ is small enough with the increasing of training samples. When $n_+\asymp n_-\asymp n, r(w^*)=O(\sigma\|w^*\|^2)$ and $\epsilon=\sqrt{\frac{3n_+^2(n_+-1)n_-^2\sigma^2}{4608n_-^2+256n_+^2}}$,*

$$\mathbb{E}_S\big[R(A(S))-R(w^*)\big]=O\bigg(\frac{R(w^*)}{n^{\frac{3}{2}}\sigma}+\big(n^{-\frac{3}{2}}+\sigma\big)\|w^*\|^2\bigg).$$

When $\sigma=n^{-\frac{3}{4}}\|w^*\|^{-1}\sqrt{R(w^*)}$ and $R(w^*)=n^{-\frac{1}{2}}\|w^*\|^2$, $\mathbb{E}_S\big[R(A(S))-R(w^*)\big]=O(n^{-1}\|w^*\|^2)$. Note that $R(w^*)$ can not less than $n^{-\frac{1}{2}}\|w^*\|^2$ due to $\sigma\min\{n_+,n_-\}\ge 8\alpha$.

The above excess risk bound assures the convergence rate $O(n^{-1}\|w^*\|^2)$ in expectation under proper conditions of w^* and $R_S(A(S))$, which extends the previous optimistic generalization bounds of pointwise learning (Srebro, Sridharan, and Tewari 2010; Zhang, Yang, and Jin 2017) and pairwise learning (Lei, Ledent, and Kloft 2020) to the triplet setting.

Remark 7. *As summarized in Table 2, the convergence guarantees in Theorems 3 and 4 are comparable with the existing results in the setting of strong convexity even involving the complicated triplet structure in error decomposition.*

Algorithm	Reference	Assumptions			Tool	Convergence rate
		Strongly Convex	Lipschitz	Smooth		
Full-batch SGD (▲)	Klochkov and Zhivotovskiy (2021)	√	√	×	Uniform stability	$*O(n^{-1}\log n)$
RRM (▲)	Feldman and Vondrák (2019)	√	√	×	Uniform stability	$*O(n^{-\frac{1}{2}}\log n)$
RRM (▲▲)	Lei, Ledent, and Kloft (2020)	√	√	×	Uniform stability	$*O(n^{-\frac{1}{2}}\log n)$
	Lei, Ledent, and Kloft (2020)	√	×	√	On-average stability	$O(n^{-1})$
RRM (▲▲▼)	Ours ($n_+ \asymp n_- \asymp n$)	√	√	×	Uniform stability	$*O(n^{-\frac{1}{2}}\log n)$
	Ours ($n_+ \asymp n_- \asymp n$)	√	×	√	On-average stability	$O(n^{-1})$

Table 2: Summary of stability-based generalization analyses for algorithms in the setting of strong convexity (▲-pointwise; ▲▲-pairwise; ▲▲▼-triplet; √-the reference has such a property; ×-the reference hasn't such a property; $*$-high-probability bound).

Applied to Triplet Metric Learning

This section applies our generalization analysis to triplet metric learning, which focuses on learning a metric to minimize the intra-class distance and maximize inter-class distance simultaneously. Let $t(y, y')$ be the symbolic function, i.e., $t(y, y') = 1$ if $y = y'$ and -1 otherwise. Inspired by the 0-1 loss in pairwise metric learning $\ell_{0-1}(w; z, z') = \mathbb{I}[t(y, y')(1 - h_w(x, x')) \leq 0]$ (Lei, Liu, and Ying 2021), we consider a 0-1 triplet loss $\ell_{0-1}(w; z^+, \tilde{z}^+, z^-) = \mathbb{I}[h_w(x^+, \tilde{x}^+) - h_w(x^+, x^-) + \zeta \geq 0]$, where the training model h_w is considered as $h_w(x^+, \tilde{x}^+) = \langle w, (x^+ - \tilde{x}^+)(x^+ - \tilde{x}^+)^\top \rangle$, and ζ denotes the margin that requires the distance of negative pairs to excess the one of positive pairs. We introduce the triplet loss

$$\ell_\phi(w; z^+, \tilde{z}^+, z^-) = \phi(h_w(x^+, \tilde{x}^+) - h_w(x^+, x^-) + \zeta) \tag{6}$$

associated with the logistic function $\phi(u) = \log(1 + \exp(-u))$, which is consistent with the error metric used in Schroff, Kalenichenko, and Philbin (2015) and Ge et al. (2018).

When $\max\{\sup_{x^+ \in \mathcal{X}_+} \|x^+\|, \sup_{x^- \in \mathcal{X}_-} \|x^-\|\} \leq B$, Theorems 2-3 yield the following convergence rates for SGD and RRM with the triplet loss (6), respectively.

Corollary 1. *Let w_T is produced by SGD (3) with $\eta_t \equiv c/\sqrt{T}, c \leq 1/(32B^4)$ and $\big|\mathbb{E}_S[\phi(h_{w_T}(x^+, \tilde{x}^+) - h_{w_T}(x^+, x^-) + \zeta)]\big| \leq M$. For any $\delta \in (0, 1/e)$, with probability $1 - \delta$, we have $|R_S(w_T) - R(w_T)| = O\big(n^{-\frac{1}{2}}\log n \log^{\frac{3}{2}}(1/\delta) + n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)\big)$.*

Corollary 2. *Consider $F_S(w)$ in (5) with the triplet loss (6) and $r(w^*) = O(\sigma\|w^*\|^2)$ with $\sigma \asymp n^{-\frac{1}{2}}$. Assume that $\sup_{z^+, \tilde{z}^+, z^-} |\ell(w^*; z^+, \tilde{z}^+, z^-)| \leq O(\sqrt{n})$ and the variance of $\ell(w^*; z^+, \tilde{z}^+, z^-)$ is bounded. Then for $A(S) = \arg\min_{w \in \mathcal{W}} F_S(w)$ and any $\delta \in (0, 1/e)$, we have $R(A(S)) - R(w^*) = O(n^{-\frac{1}{2}}\log n \log(1/\delta))$ with probability $1 - \delta$.*

Moreover, we get the refined result of RRM from Theorem 4 with the help of the strong-convexity of (6).

Corollary 3. *Under the basic assumptions and notations of Corollary 2, assume $\sigma = n^{-\frac{3}{4}}\|w^*\|^{-1}\sqrt{R(w^*)}$ and $R(w^*) = n^{-\frac{1}{2}}\|w^*\|^2$, then we have $\mathbb{E}_S\big[R(A(S)) - R(w^*)\big] = O(n^{-1}\|w^*\|^2)$.*

Conclusion

This paper fills the theoretical gap in the generalization bounds of SGD and RRM for triplet learning by developing algorithmic stability analysis techniques, which are valuable to understanding their intrinsic statistical foundations of outstanding empirical performance. We firstly derive the general high-probability generalization bound $O(\gamma\log n + n^{-\frac{1}{2}})$ for triplet uniformly stable algorithms, and then apply it to get the explicit result $O(n^{-\frac{1}{2}}\log n)$ for SGD and RRM under mild conditions of loss function. For RRM with triplet loss, the optimistic bound $O(n^{-1})$ in expectation is also provided by leveraging the on-average stability. Even for the complicated triplet structure, our results also enjoy similar convergence rates as the previous related works of pointwise learning (Hardt, Recht, and Singer 2016; Feldman and Vondrák 2019) and pairwise learning (Lei, Ledent, and Kloft 2020). Some potential directions are discussed in *Supplementary Material D* for future research.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant Nos. 12071166, 62106191, 61972188, 62122035.

References

Agarwal, S.; and Niyogi, P. 2009. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10: 441–474.

Arous, G. B.; Gheissari, R.; and Jagannath, A. 2021. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22: 106:1–106:51.

Bartlett, P. L.; and Mendelson, S. 2001. Rademacher and Gaussian complexities: Risk bounds and structural results. In *Conference on Computational Learning Theory*, 224–240.

Bassily, R.; Feldman, V.; Guzmán, C.; and Talwar, K. 2020. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4381–4391.

Bottou, L.; and Bousquet, O. 2007. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 161–168.

Bousquet, O.; and Elisseeff, A. 2002. Stability and generalization. *Journal of Machine Learning Research*, 2: 499–526.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a Siamese time delay neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 737–744.

Chen, H. 2012. The convergence rate of a regularized ranking algorithm. *Journal of Approximation Theory*, 164(12): 1513–1519.

Chen, H.; Wang, X.; Deng, C.; and Huang, H. 2017. Group sparse additive machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, 198–208.

Chen, H.; Wang, Y.; Zheng, F.; Deng, C.; and Huang, H. 2021. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2373–2387.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1335–1344.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 539–546.

Cucker, F.; and Smale, S. 2001. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39: 1–49.

Cucker, F.; and Zhou, D. X. 2007. *Learning theory: An approximation theory viewpoint*. Cambridge University Press.

Deng, Z.; He, H.; and Su, W. J. 2021. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning (ICML)*, 2590–2600.

Ding, C.; and Tao, D. 2018. Trunk-Branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 1002–1014.

Elisseeff, A.; Evgeniou, T.; and Pontil, M. 2005. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6: 55–79.

Fehrman, B. J.; Gess, B.; and Jentzen, A. 2020. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21: 136:1–136:48.

Feldman, V.; and Vondrák, J. 2019. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory (COLT)*, 1270–1279.

Foster, D. J.; Greenberg, S.; Kale, S.; Luo, H.; Mohri, M.; and Sridharan, K. 2019. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 6726–6736.

Ge, W.; Huang, W.; Dong, D.; and Scott, M. R. 2018. Deep metric learning with hierarchical triplet loss. In *European Conference on Computer Vision (ECCV)*, 272–288.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1735–1742.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 1225–1234.

Harvey, N. J. A.; Liaw, C.; Plan, Y.; and Randhawa, S. 2019. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory (COLT)*, 1579–1613.

Huang, J.; Feris, R. S.; Chen, Q.; and Yan, S. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *International Conference on Computer Vision (ICCV)*, 1062–1070.

Klochkov, Y.; and Zhivotovskiy, N. 2021. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, 5065–5076.

Kuzborskij, I.; and Lampert, C. H. 2018. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 2820–2829.

Lai, H.; Pan, Y.; Liu, Y.; and Yan, S. 2015. Simultaneous feature learning and hash coding with deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 3270–3278.

Lei, Y.; Hu, T.; and Tang, K. 2021. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22: 25:1–25:41.

Lei, Y.; Ledent, A.; and Kloft, M. 2020. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 21236–21246.

Lei, Y.; Liu, M.; and Ying, Y. 2021. Generalization guarantee of SGD for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 21216–21228.

Lei, Y.; and Ying, Y. 2020. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 5809–5819.

Lei, Y.; and Ying, Y. 2021. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations (ICLR)*, 1–23.

Lin, J.; Camoriano, R.; and Rosasco, L. 2016. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning (ICML)*, 2340–2348.

Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; and Huang, T. 2016. Deep relative distance learning: Tell the difference between similar vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2167–2175.

Liu, T.; Lugosi, G.; Neu, G.; and Tao, D. 2017. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning (ICML)*, 2159–2167.

Pitcan, Y. 2017. A note on concentration inequalities for U-statistics. arXiv:1712.06160.

Poggio, T. A.; Rifkin, R. M.; Mukherjee, S.; and Niyogi, P. 2004. General conditions for predictivity in learning theory. *Nature*, 428: 419–422.

Ramanathan, V.; Li, C.; Deng, J.; Han, W.; Li, Z.; Gu, K.; Song, Y.; Bengio, S.; Rosenberg, C.; and Fei-Fei, L. 2015. Learning semantic relationships for better action retrieval in images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1100–1109.

Rogers, W.; and Wagner, T. 1978. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6: 506–514.

Rosasco, L.; Belkin, M.; and Vito, E. D. 2010. On learning with integral operators. *Journal of Machine Learning Research*, 11: 905–934.

Roux, N. L.; Schmidt, M.; and Bach, F. R. 2012. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2672–2680.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Schultz, M.; and Joachims, T. 2003. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems (NeurIPS)*, 41–48.

Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; and Sridharan, K. 2010. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11: 2635–2670.

Shen, W.; Yang, Z.; Ying, Y.; and Yuan, X. 2019. Stability and optimization error of stochastic gradient descent for pairwise learning. arXiv:1904.11316.

Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; and Moreno-Noguer, F. 2015. Discriminative learning of deep convolutional feature point descriptors. In *International Conference on Computer Vision (ICCV)*, 118–126.

Smale, S.; and Zhou, D.-X. 2007. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26: 153–172.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2199–2207.

Ustinova, E.; and Lempitsky, V. S. 2016. Learning deep embeddings with histogram Loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4170–4178.

Vapnik, V. N. 1998. Statistical learning theory. *Encyclopedia of the Sciences of Learning*, 41(4): 3185–3185.

Wang, Y.; Chen, H.; Zheng, F.; Xu, C.; Gong, T.; and Chen, Y. 2020. Multi-task additive models for robust estimation and automatic structure discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. K. 2005. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1473–1480.

Wohlhart, P.; and Lepetit, V. 2015. Learning descriptors for object recognition and 3D pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 3109–3118.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1249–1258.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2002. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems (NeurIPS)*, 505–512.

Xing, Y.; Song, Q.; and Cheng, G. 2021. On the algorithmic stability of adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 26523–26535.

Yang, Z.; Lei, Y.; Wang, P.; Yang, T.; and Ying, Y. 2021. Simple stochastic and online gradient descent algorithms for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 20160–20171.

Ying, Y.; and Li, P. 2012. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13: 1–26.

Yu, B.; Liu, T.; Gong, M.; Ding, C.; and Tao, D. 2018. Correcting the triplet selection bias for triplet loss. In *European Conference on Computer Vision (ECCV)*, 71–86.

Zhang, L.; Yang, T.; and Jin, R. 2017. Empirical risk minimization for stochastic convex optimization: $O(1/n)$-and $O(1/n^2)$-type of risk bounds. In *Conference on Learning Theory (COLT)*, 1954–1979.