entropy

MDPI

Article

PAC–Bayes Guarantees for Data-Adaptive Pairwise Learning

Sijia Zhou [1,*], Yunwen Lei [2,*] and Ata Kabán [1,*]

[1] School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[2] Department of Mathematics, University of Hong Kong, Pokfulam, Hong Hong, China
[*] Correspondence: sxz115@student.bham.ac.uk (S.Z.); leiyw@hku.hk (Y.L.); a.kaban@bham.ac.uk (A.K.)

Abstract

We study the generalization properties of stochastic optimization methods under adaptive data sampling schemes, focusing on the setting of pairwise learning, which is central to tasks like ranking, metric learning, and AUC maximization. Unlike pointwise learning, pairwise methods must address statistical dependencies between input pairs—a challenge that existing analyses do not adequately handle when sampling is adaptive. In this work, we extend a general framework that integrates two algorithm-dependent approaches—algorithmic stability and PAC–Bayes analysis for this purpose. Specifically, we examine (1) Pairwise Stochastic Gradient Descent (Pairwise SGD), widely used across machine learning applications, and (2) Pairwise Stochastic Gradient Descent Ascent (Pairwise SGDA), common in adversarial training. Our analysis avoids artificial randomization and leverages the inherent stochasticity of gradient updates instead. Our results yield generalization guarantees of order $n^{-1/2}$ under non-uniform adaptive sampling strategies, covering both smooth and non-smooth convex settings. We believe these findings address a significant gap in the theory of pairwise learning with adaptive sampling.

Keywords: pairwise learning; randomized algorithms; PAC–Bayes; algorithmic stability

1. Introduction

The increasing availability of data makes it feasible to use increasingly large models in principle. However, this comes at the expense of an increasing computational cost of training these models in large pairwise learning applications. Some notable examples of pairwise learning problems include ranking and preference prediction, AUC maximization, and metric learning [1–5]. For instance, in metric learning we aim to learn an appropriate distance or similarity to compare pairs of examples, which has numerous applications such as face verification, person re-identification (Re-ID) [6–10], and bioactivity prediction [11]. Pairwise learning has also been applied to positive-unlabeled (PU) learning problems [12], where only positive and unlabeled examples are available. Such problems arise in one-class classification settings, with practical applications in areas such as fault detection and diagnosis in advanced engineering systems [13]. Given the broad relevance of pairwise learning, there is a pressing need to deepen our theoretical understanding of its generalization properties. This in turn can inform the design of algorithms that generalize reliably to unseen pairs and offer interpretability and trustworthiness to end users.

In both pointwise and pairwise learning settings, Stochastic Gradient Descent (SGD) and Stochastic Gradient Descent Ascent (SGDA) are widely used for large-scale minimization and min-max optimization problems in machine learning due to their favorable computational efficiency. These methods rely on stochastic sampling strategies to approximate the true gradients, and several works have explored data-dependent sampling

Academic Editor: Dawn E. Holmes

Received: 6 June 2025
Revised: 28 July 2025
Accepted: 4 August 2025
Published: 8 August 2025

Citation: Zhou, S.; Lei, Y.; Kabán, A. PAC–Bayes Guarantees for Data-Adaptive Pairwise Learning. *Entropy* **2025**, *27*, 845. https://doi.org/10.3390/e27080845

techniques to accelerate convergence to the optimum [14–19]. SGDA, in particular, is a standard approach in solving min-max problems, finding notable applications in generative adversarial networks (GANs) [20] and adversarial training [21,22].

Adversarial perturbations are subtle, often imperceptible modifications to input data designed to deceive models and cause incorrect predictions [23]. Recent studies in pairwise learning have explored strategies to enhance adversarial robustness, applying adversarial pairwise learning methods to min-max problems across various domains, such as metric learning [24,25], ranking [25,26], and kinship verification [27]. These developments illustrate the need for robust, theoretically grounded pairwise methods that can withstand adversarial attacks while maintaining generalization performance.

Under the assumption of i.i.d. data points in classic pointwise learning, the empirical risk of a fixed hypothesis is an average of i.i.d. random variables. However, in pairwise learning the pairs derived from i.i.d. points are no longer i.i.d. Instead, when the loss is symmetric and computed over all unordered pairs, the empirical risk takes the form of a second-order U-statistic. Therefore, results on U-processes may be used to investigate the generalization analysis of pairwise learning [3,28]. While there is much research on the generalization analysis of pairwise learning, the effect of non-uniform, data-dependent sampling schemes has not been rigorously studied.

Non-uniform sampling can be beneficial in noisy data situations where the training examples may not be equally reliable or equally informative. Some examples may be less important than others, or even misleading—e.g., mislabeled examples or examples situated in an ambiguous class-overlap region. In rare cases when the usefulness or importance of individual training examples is known, then the sampling distribution can be designed and fixed before training, and this may improve the representativeness of the sample. For instance, in the case of infrequent observations [29], inverse frequency sampling prioritizes rare examples that may be underrepresented in the training set, ensuring their proper influence. However, in most cases the relative importance of training examples is not known a priori; hence, it is desirable to learn the sampling distribution together with training the model.

The idea of adaptive sampling refers to a sampling distribution that depends on the training sample. Such non-uniform and data-dependent sampling shows great potential in the literature of randomized algorithms for both SGD- [14] and SGDA-based [30] optimizers, in both pointwise and pairwise settings. Importance sampling [14] is one of the widely used of such strategies, and a few others will be reviewed shortly in Section 2. Therefore, recent work [31,32] has begun to develop a better understanding of the generalization behavior of such algorithms, which we continue here in the setting of pairwise learning. The main bottlenecks in the analysis of adaptive sampling-based stochastic optimizers are that (i) a correction factor is often used to ensure the unbiasedness of the gradient [14], which also depends on training data points complicating the analysis, and (ii) in the pairwise setting we also need to cater to statistical dependencies between data pairs, which are due to the fact that each point participates in multiple pairs.

To tackle these problems, we develop a PAC–Bayesian analysis of the generalization of pairwise stochastic optimization methods, which removes the need for a correction factor, and we use U-statistics to capture the statistical structure of pairwise loss functions. The PAC–Bayes framework allows us to obtain generalization bounds that hold uniformly for all posterior sampling schemes, under a mild condition required on a pre-specified prior sampling scheme (chosen as the uniform sampling). For randomized methods, such as Pairwise SGD and Pairwise SGDA, the sampling index pairs will be treated as hyperparameters that follow a sampling distribution.

Our main contributions are listed in Table 1, summarizing the generalization bounds of the order $\widetilde{O}(1/\sqrt{n})$ for these randomized algorithms under different assumptions, where n is the sample size.

Our technical contributions are summarized as follows:

- We bound the generalization gap of randomized pairwise learning algorithms that operate with an arbitrary data-dependent sampling, in a PAC–Bayesian framework, under a sub-exponential stability condition.
- We apply the above general result to Pairwise SGD and Pairwise SGDA with arbitrary sampling. For both of these algorithms, we verify the sub-exponential stability in both smooth and non-smooth problems.
- We exemplify how our bounds can inform algorithm design, and we demonstrate how to extract meaningful information from the resulting algorithms.

Our work builds on well-established tools, including a specific flavor of PAC–Bayesian analysis [31], U-statistic decomposition, and a moment bound for uniformly stable pairwise learning algorithms [33], aimed at bringing theoretical insight into an important, yet relatively underexplored, setting: pairwise learning with adaptive sampling. To the best of our knowledge, our analysis is the first to derive explicit generalization bounds for this setting.

The remainder of the paper is organized as follows. We survey the related work on the generalization analysis and non-uniform sampling in Section 2. We give a brief background on U-statistics and algorithmic stability analysis in Section 3. Our general result and its applications to Pairwise SGD and Pairwise SGDA are presented in Section 4.

Table 1. Summary of generalization rates obtained for two pairwise stochastic optimization algorithms (Pairwise SGD, Pairwise SGDA) under two sets of assumptions (Lipschitz (L), smooth (S), convex (C)) on the pairwise loss function, together with the chosen number of iterations T and step size η. The sample size is n. According to this summary, we notice that smaller step sizes and more iterations are needed if the smoothness assumption is removed (more details in Section 4).

Algo.	**Asm.**	**Time T and Step Size η**		**Rates**
Pairwise SGD	L, C	$T = \Theta(n^2)$	$\eta = \Theta(T^{-\frac{3}{4}})$	$\widetilde{O}(1/\sqrt{n})$ Theorem 1 (1)
	L, S, C	$T = \Theta(n)$	$\eta = \Theta(T^{-\frac{1}{2}})$	$\widetilde{O}(1/\sqrt{n})$ Theorem 1 (2)
Pairwise SGDA	L, C	$T = O(n^2)$	$\eta = O(T^{-\frac{3}{4}})$	$\widetilde{O}(1/\sqrt{n})$ Theorem 2 (1)
	L, S, C	$T = O(n)$	$\eta = O(T^{-\frac{1}{2}})$	$\widetilde{O}(1/\sqrt{n})$ Theorem 2 (2)

2. Related Work

Adaptive Sampling in Stochastic Optimization. Importance Sampling for Stochastic Gradient Descent was proposed in [14]. To compute the stochastic gradient, a training example z_i ($i \in [n]$) is sampled with probability proportional to the gradient norm $p_i \propto \|\nabla_{\mathbf{w}} \ell(\mathbf{w}; z_i)\|$, where $\mathbf{w}$ are the model parameters and ℓ is the loss function. This prioritizes high-impact updates from the perspective of optimization—the authors proved that this can significantly reduce the variance of the stochastic gradient and accelerate convergence to the optimum. A related idea, loss-based sampling, proposed in [16], assigns sampling probabilities proportional to the loss evaluated on training points, that is, $p_i \propto \ell(\mathbf{w}; z_i)$, thereby focusing on hard-to-fit examples. The authors show faster convergence to the optimum. While these works do not consider pairwise learning, they represent landmarks on adaptive sampling in stochastic optimization.

Furthermore, there are variants of these ideas aimed at lightening the computational demand. These include upper bounds to the gradient norm, shown to exhibit better performance in comparison with the loss-based sampling [34]. The work in [18] proposes to sample the training points based on their relative distance to each other. Another more recent data-dependent sampling approach called group sampling appears in [19] and has been applied to a person re-identification (Re-ID) application.

Adaptive sampling is an umbrella term referring to sampling distributions that depend on the training sample. Here, we mentioned a few of the most prominent existing examples. However, the appropriate sampling distribution is task-dependent. The works discussed, and most of the previous work on adaptive sampling in stochastic optimization, aim at accelerating convergence to the optimum. Therefore, they have no explicit cost for data-dependent sampling; instead, they have a multiplicative correction factor to ensure an unbiased gradient. However, the goal of learning is generalization, which is different from achieving the global optimum on training data. There must be a cost for data-dependent sampling to avoid over-reliance on a biased subset of points. Our forthcoming generalization bounds will quantify this rigorously and provide guidance for algorithm design.

With the exception of [32] and our previous conference paper [31], results on the generalization analysis of the resulting randomized algorithms are very scarce, which is our goal to advance in this paper specifically for the pairwise learning setting.

Generalization through Algorithmic Stability. Stability was popularized in the seminal work of [35], to formalize the intuition that algorithms whose output is resilient to changing an example in its input data will generalize. The stability framework subsequently motivated a chain of analysis of randomized iterative algorithms, such as SGD [36] and SGDA [37,38]. While the stability framework in the previous work is well suited for SGD-type algorithms that operate a uniform sampling scheme [36], this framework alone is unable to tackle arbitrary data-dependent sampling schemes.

Generalization through PAC–Bayes. The PAC–Bayes theory of generalization is another algorithm-dependent framework in statistical learning, the gist of which is to leverage a pre-specified prior distribution on the parameters of interest to obtain generalization bounds that hold uniformly for all posterior distributions [39,40]. Its complementarity with the algorithmic stability framework sparked ideas for combining them [41–44], some of which are also applicable to randomized learning algorithms such as SGD and SGLD [32,45–47]. While insightful, these works assume i.i.d. examples and cannot be applied to non-i.i.d. settings that arise in pairwise learning.

In non-i.i.d. settings, ref. [48] gave PAC–Bayes bounds using fractional covers, which allows for handling the dependencies within the inputs. This gives rise to generalization bounds for pairwise learning, with predictors following a distribution induced by a prior distribution on the model's parameters. However, with SGD-type methods in mind, which have a randomization already built into the algorithm, the classic PAC–Bayes approach of placing a prior on a model's parameters would be somewhat artificial. Indeed, considerable research effort has been spent to reverse such randomization [49]. Another issue concerns the prior specification—recent research [50] reveals that placing sufficient prior mass on good predictors is a condition for meaningful PAC–Bayes guarantees. These are difficult to set without a strong prior knowledge. Instead, the construction proposed in [31,32] (albeit restricted to the i.i.d. setting) is to exploit this built-in stochasticity of modern gradient-based optimization algorithms directly, by interpreting it as a PAC–Bayes prior placed on a hyperparameter. We will build on this idea further in this work.

3. Preliminaries

3.1. Pairwise Learning and U-Statistics

Let $\mathcal{D}$ be an unknown distribution on sample space $\mathcal{Z}$. We denote by $\mathcal{W} \subseteq \mathbb{R}^d$ the parameter space, and Φ will be a hyperparameter space. Given a training set $S = \{z_1, \ldots, z_n\}$ drawn i.i.d. from $\mathcal{D}$ and a hyperparameter $\phi \in \Phi$, a learning algorithm A returns a model parameterized by $A(S; \phi) \in \mathcal{W}$.

We are interested in pairwise learning problems and will use a pairwise loss function $\ell : \mathcal{W} \times (\mathcal{Z} \times \mathcal{Z}) \mapsto \mathbb{R}_+$ to measure the mismatch between the prediction of the model that acts on example pairs. The generalization error, or true risk, is defined as the expected loss of the learned predictor applied on an unseen pair of inputs drawn from $\mathcal{D}^2$, that is,

$$R(A(S;\phi)) := \mathbb{E}_{z,\tilde{z} \sim \mathcal{D}}[\ell(A(S;\phi), z, \tilde{z})]. \tag{1}$$

Since $\mathcal{D}$ is unknown, we consider the empirical risk,

$$R_S(A(S;\phi)) := \frac{1}{n(n-1)} \sum_{i,j \in [n] : i \neq j} \ell(A(S;\phi), z_i, z_j), \tag{2}$$

where $[n] := \{1, \ldots, n\}$. The generalization error is a random quantity as a function of the sample S, which does not consider the randomization used when selecting the data or feature index for the update rule of A at each iteration.

To take advantage of the built-in stochasticity of the type of algorithms we consider, we further define two distributions on the hyperparameter space Φ: a sample-independent distribution P and a sample-dependent distribution Q. In this stochastic or randomized learning algorithm setting, the expected risk and the expected empirical risk (both with respect to Q) are defined as

$$R(\mathrm{Q}) = \mathop{\mathbb{E}}_{\phi \sim \mathrm{Q}}[R(A(S;\phi))], \quad R_S(\mathrm{Q}) = \mathop{\mathbb{E}}_{\phi \sim \mathrm{Q}}[R_S(A(S;\phi))].$$

We denote the difference between the risk and the empirical risk (i.e., the generalization gap) by $G(S, \phi) := R(A(S;\phi)) - R_S(A(S;\phi))$.

The difficulty with the pairwise empirical loss (2) is that, even with S consisting of i.i.d. instances, the pairs from S are dependent of each other. Instead, $R_S(A(S;\phi))$ is a second-order U-statistic. A powerful technique to handle the U-statistic is the representation as an average of "sums-of-i.i.d." blocks [28]. That is, for a symmetric kernel $q : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$, we can represent the U-statistic $U_n := \frac{1}{n(n-1)} \sum_{i,j \in [n] : i \neq j} q(z_i, z_j)$ as

$$U_n = \frac{1}{n!} \sum_{\sigma} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q(z_{\sigma(i)}, z_{\sigma(\lfloor \frac{n}{2} \rfloor + i)}), \tag{3}$$

where σ ranges over all permutations of $\{1, \ldots, n\}$.

3.2. Connection with the PAC–Bayesian Framework

As described above, we consider two probability distributions on the hyperparameter space Φ, to account for the stochasticity in stochastic optimization algorithms, such as Pairwise SGD and Pairwise SGDA, where the hyperparameter $\phi \in \Phi$ is a sequence of pairs of indices that follow a discrete distribution. For instance, in Pairwise SGD, in every iteration $t \in [T]$, we have $\phi_t = (i_t, j_t)$, that is, a pair of independently sampled sample indices, drawn from $\{(i_t, j_t) : i_t, j_t \in [n], i_t \neq j_t\}$ with replacement (more details in Section 4.1). We define two distributions over Φ, namely the PAC–Bayes prior P, which needs to be specified before seeing the training data, and the PAC–Bayes posterior Q, which is allowed to depend on the training sample. This setting is different from the classic use of PAC–Bayes, which defines the two distributions directly on the trainable parameter space $\mathcal{W}$. Our distributions defined on Φ indirectly induce distributions on the parameter estimates, without the need to know their parametric form. This setting of PAC–Bayes was formerly introduced in London [32] in combination with algorithmic stability and further improved in our previous work [31], both restricted to the i.i.d. pointwise setting.

3.3. Connection with the Algorithmic Stability Framework

A more recent framework for the generalization problem considers algorithmic stability [35], which measures the sensitivity of a learning algorithm to small changes in the training data. The concept considered in our work among several notions of algorithmic stability is uniform stability.

Definition 1 (Uniform Stability)**.** *For $\forall\phi$, we say an algorithm $A : S \mapsto A(S;\phi)$ is β_ϕ-uniformly stable if*

$$|\ell(A(S;\phi), z, \tilde{z}) - \ell(A(S', \phi), z, \tilde{z})| \le \beta_\phi, \quad \forall z, \tilde{z} \in \mathcal{Z}, \tag{4}$$

where $S, S' \in \mathcal{Z}^n$ differs by at most a single example.

The algorithmic stability framework is suitable for analyzing certain deterministic learning algorithms, or randomized algorithms with a pre-defined randomization. In turn, here we are concerned with inherently stochastic algorithms where we wish to allow any data-dependent stochasticity, such as the variants of importance sampling and other recent practical methods mentioned in the related works, e.g., [14,15,18,19,34]. Moreover, in principle our framework and results are applicable even if the sampling distribution is learned from the training data itself.

Sub-exponential Stability. A useful definition of stability that captures the stochastic nature of the algorithms we are interested in is the sub-exponential stability introduced in Zhou et al. [31]. Recall that ϕ is a random variable following a distribution defined on Φ. Therefore, the stability parameter β_ϕ is also a random variable as a function of ϕ. We want to control the tail behavior of β_ϕ around a value that decays with the sample size n, and we define the sub-exponential stability as the following.

Definition 2 (Sub-exponential stability)**.** *Fix any prior distribution P on $\Phi = \prod_{t=1}^{T} \Phi_t$. We say that a stochastic algorithm is sub-exponentially β_ϕ-stable (with respect to P) if, given any fixed instance of $\phi \sim \mathrm{P}$, it is β_ϕ-uniformly stable and there exist $c_1, c_2 \in \mathbb{R}$ such that for any $\delta \in (0, 1/n]$, the following holds with probability of at least $1 - \delta$:*

$$\beta_\phi \le c_1 + c_2 \log(1/\delta). \tag{5}$$

4. Main Results

In this section, we will give generalization bounds for Pairwise SGD and Pairwise SGDA in pairwise learning. To this aim, we first give a general result (Lemma 1) to show the connection between the sub-exponential stability condition (Assumption 2) and the generalization gap in the case of pairwise learning. We then derive stability bounds to show that this assumption holds for Pairwise SGD and Pairwise SGDA, in both smooth convex and non-smooth convex cases. Based on these, we apply the stability bounds to Lemma 1 to derive the corresponding generalization bounds. We use $K \lesssim K'$ if there exists a universal constant $a > 0$ such that $K \le aK'$. The proof is given in Appendix A.

Lemma 1 (Generalization of randomized pairwise learning)**.** *Given distribution* P*,* $c_1, c_2 > 0$*, and M-bounded loss for a sub-exponentially stable algorithm* A*,* $\forall\delta \in (0, 1/n)$*, with probability of at least* $1-\delta$*, the following holds uniformly for all* Q *absolutely continuous with respect to* P*:*

$$\mathbb{E}_{\phi\sim \mathrm{Q}}[G(S,\phi)] \lesssim \left(\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) + \log\frac{1}{\delta}\right)\max\left\{c_1\log n + c_2\log^2 n, \frac{M}{\sqrt{n}}\right\},$$

where $\mathrm{KL}(\mathrm{Q}\|\mathrm{P})$ *is the KL divergence between* Q *and* P*,* $\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) = \int_{\phi\in\Phi}\log\frac{d\mathrm{Q}}{d\mathrm{P}}d\mathrm{Q}$*.*

A strength of Lemma 1 is that we only need to check the sub-exponential stability condition under a prior distribution P, and Lemma 1 automatically translates it to generalization bounds for any posterior distribution Q.

In the forthcoming applications both Q and P are discrete distributions, so we have $\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) = \sum_{\phi\in\Phi}\mathrm{Q}(\phi)\log\frac{\mathrm{Q}(\phi)}{\mathrm{P}(\phi)}$. In particular, the prior P will be most naturally chosen as the discrete uniform distribution in the context of applications to stochastic optimization in Section 4.2. Let $\mathrm{P} = \mathcal{U}$ with $\mathcal{U}$ denoting the uniform distribution on $([n]\times[n])^T$. Hence, the absolute continuity condition is satisfied, ensuring that $\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) < \infty$ for all distributions Q over the set $([n]\times[n])^T$. Furthermore, in this setting, we have $\mathrm{KL}(\mathrm{Q}\|P) = -\mathrm{H}(\mathrm{Q}) + 2T\log n$, where H denotes the Shannon entropy.

We introduce some classic assumptions that are frequently employed in the analysis of randomized algorithms. Let $\|\cdot\|_2$ denote the Euclidean norm. Let S and S' be neighboring datasets (i.e., they differ in only one example, which we denote as the k-th example, $k\in[n]$). For brevity, we write $\ell(\mathbf{w})$ for $\ell(\mathbf{w}; z, \tilde{z})$, where we mean a property that holds for all $z, \tilde{z} \in \mathcal{Z}$.

Assumption 1. *Let* $L > 0$*. We say* ℓ *is L-Lipschitz if for any* $\mathbf{w}_1$*,* $\mathbf{w}_2 \in \mathcal{W}$*, we have* $|\ell(\mathbf{w}_1) - \ell(\mathbf{w}_2)| \le L\|\mathbf{w}_1 - \mathbf{w}_2\|_2$*.*

Assumption 2 (Convexity)**.** *We say* ℓ *is convex if the following holds* $\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$*:*

$$\ell(\mathbf{w}_1) \ge \ell(\mathbf{w}_2) + \langle\nabla\ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle,$$

where $\langle\cdot,\cdot\rangle$ *represents the inner product.*

Assumption 3. *Let* $\alpha \ge 0$*. We say a differentiable function* ℓ *is* α*-smooth, if for any* $\mathbf{w}_1$*,* $\mathbf{w}_2 \in \mathcal{W}$*,* $\|\nabla\ell(\mathbf{w}_1) - \nabla\ell(\mathbf{w}_2)\|_2 \le \alpha\|\mathbf{w}_1 - \mathbf{w}_2\|_2$*, where* $\nabla\ell$ *represents the gradient of* ℓ*.*

4.1. Stability and Generalization of Pairwise SGD

We now consider Pairwise SGD, which, as we will show, also satisfies the sub-exponential stability condition in both smooth and non-smooth cases.

We denote $\mathbf{w}_1$ an initial point and a uniform distribution over $([n]\times[n])^T$. At the t-th iteration for Pairwise SGD, a pair of sample indices $\phi_t = (i_t, j_t)$ is uniformly randomly selected from the set $\{(i_t, j_t) : i_t, j_t \in [n], i_t \ne j_t\}$. This forms a sequence of index pairs $\phi = (\phi_1, ..., \phi_T)$. For step size η_t, the model is updated by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\nabla\ell(\mathbf{w}_t; z_{i_t}, z_{j_t})$.

The following lemma shows that Pairwise SGD with uniform sampling applied to both smooth and non-smooth problems enjoys sub-exponential stability. The proof is given in Appendix B.1.

Lemma 2 (Sub-exponential stability of Pairwise SGD)**.** *Let* $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ *be two parameter sequences produced by Pairwise SGD with fixed step sizes and uniform sampling* P*, while being trained on neighboring training samples* S *and* S'*. Suppose there is a loss in Lipschitzness and convexity (i.e., Assumptions 1 and 2 hold). Then, we have the following:*

(1) *At the t-th iteration, we have sub-exponential stability (Definition 2) with*

$$c_1 = 2\sqrt{e}L^2\eta(\sqrt{t}+2t/n) \quad \textit{and} \quad c_2 = 4\sqrt{e}L^2\eta\left(1+2(t/n)^{\frac{1}{2}}\right).$$

(2) *If in addition the loss is also smooth (Assumption 3 holds), then with step size $\eta \le 2/\alpha$, at the t-th iteration, we have sub-exponential stability (Definition 2) with*

$$c_1 = 4L^2\eta t/n \quad \textit{and} \quad c_2 = 4L^2\eta\left(1+2(t/n)^{\frac{1}{2}}\right).$$

Using Lemmas 1 and 2, we obtain the following generalization bound for Pairwise SGD with general sampling.

Theorem 1 (Generalization bounds for Pairwise SGD)**.** *Assume ℓ is M-bounded, Lipschitz, and convex (cf. Assumptions 1 and 2). For any $\delta \in (0,1)$, Pairwise SGD with fixed step sizes satisfies the following generalization guarantees with probability of at least $1-\delta$ over S, $S \sim \mathcal{D}^n$, for all posterior sampling distributions* Q *on $([n]\times[n])^T$:*

(1) *After T iterations, we have*

$$\mathbb{E}_{\mathrm{Q}}[G(S,\phi)] \lesssim \left(T\log n - H(\mathrm{Q}) + \log\frac{1}{\delta}\right)\max\left\{L\eta\left(\sqrt{T}+\frac{T}{n}+\sqrt{\frac{T}{n}}\right)\log^2 n, \frac{M}{\sqrt{n}}\right\}.$$

(2) *If in addition the loss is also smooth (Assumption 3 holds), then with step size $\eta \le 2/\alpha$, we have*

$$\mathbb{E}_{\mathrm{Q}}[G(S,\phi)] \lesssim \left(T\log n - H(\mathrm{Q}) + \log\frac{1}{\delta}\right)\max\left\{L\eta\left(\frac{T}{n}+1+\sqrt{\frac{T}{n}}\right)\log^2 n, \frac{M}{\sqrt{n}}\right\}.$$

Remark 1. *Suppose* $\mathrm{KL}(\mathrm{Q}\|\mathcal{U}) \in \tilde{O}(1)$, *as it has been tacitly assumed also in previous work [32] when quantifying the generalization convergence rate. Taking the choice of parameters suggested by [51], if $\eta = \Theta(T^{-\frac{3}{4}})$ and $T = \Theta(n^2)$ in the non-smooth case (part 1), then the above theorem implies bounds of the order $\widetilde{O}(1/\sqrt{n})$. In the smooth case (part 2), an analysis of the trade-off between optimization and generalization, Lei et al. [33] suggested setting $T = \Theta(n)$ and $\eta = \Theta(1/\sqrt{T})$ to get a Pairwise SGD to iterate with a good generalization performance. With these choices, our bounds in Theorem 1 are of order $\widetilde{O}(1/\sqrt{n})$, which are not improvable in general.*

Remark 2 (Implication of the $\mathrm{KL}(\mathrm{Q}\|\mathcal{U}) = \tilde{O}(1)$ assumption)**.** *Let* $\mathrm{supp}(\mathrm{Q}) \subseteq \Phi$ *denote the support of* Q*, where $\Phi = ([n]\times[n])^T$ in pairwise learning.*

Since $\mathcal{U}(\phi) = 1/n^{2T}$ for all ϕ, the KL divergence is

$$\mathrm{KL}(\mathrm{Q}\|\mathcal{U}) = \sum_{\phi\in\Phi}\mathrm{Q}(\phi)\log\left(\frac{\mathrm{Q}(\phi)}{1/n^{2T}}\right) = -H(\mathrm{Q}) + 2T\log n.$$

To ensure this is of order $\tilde{O}(1)$, we need $-H(\mathrm{Q}) + 2T\log n \le \tilde{O}(1)$; hence,

$$H(\mathrm{Q}) \ge 2T\log n - \tilde{O}(1) = H(\mathcal{U}) - \tilde{O}(1). \tag{6}$$

To give more intuition, consider Q *on a restricted support. This is very much a worst-case scenario, as it would imply completely discarding (rather than down-weighting) some of the training points. For such* Q*, the maximum entropy occurs when* Q *is uniform over its support, so $\mathrm{Q}(\phi) = \frac{1}{|supp(\mathrm{Q})|}$, and $\mathrm{KL}(\mathrm{Q}\|\mathcal{U}) = \log\left(\frac{n^{2T}}{|supp(\mathrm{Q})|}\right)$. In this case, having $\mathrm{KL}(\mathrm{Q}\|\mathcal{U}) = \tilde{O}(1)$ requires*

$$|supp(\mathrm{Q})| = \Omega\left(\frac{n^{2T}}{poly(T,n)}\right). \tag{7}$$

To sum up, Q *must satisfy the entropy lower bound* (6)*, and to achieve that entropy on a restricted support, it must have a large enough support, i.e., at least a* $\Omega(1/poly(T,n))$ *fraction of the entire* Φ*.*

In Pairwise SGD with non-uniform data-dependent sampling, this result tells us that in order to keep generalization rates that compare against the uniform baseline, Q *cannot discard a large subset of the index sequences. This limits how aggressively one can compress or "distill" a dataset (as in core-set selection or dataset distillation) without paying a KL penalty that slows down the rate—at least as long as the prior is uniform.*

Non-uniform sampling alone is insufficient for robust learning. While it may mitigate the effect of a small fraction of bad examples (e.g., out-of-distribution or mislabeled training examples), achieving robustness also requires modeling choices such as robust loss functions. In the next section we approach this via Pairwise SGDA, the type of optimization required in adversarially robust training.

4.2. Stability and Generalization of Pairwise SGDA

In this subsection, we discuss Pairwise SGDA for solving minimax problems in the convex-concave case. We will abuse the notations to apply them to the minimax case. We receive a model $A(S;\phi) := (A_{\mathrm{w}}(S;\phi), A_{\mathrm{v}}(S;\phi)) \in \mathcal{W} \times \mathcal{V}$ by applying a learning algorithm A on training set S and measure the performance with respect to a loss function $\ell : (\mathbf{w},\mathbf{v}) \mapsto \ell(\mathbf{w},\mathbf{v};z,\tilde{z})$. For any $\phi \in \Phi$, we consider the risk defined as

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} R(A_{\mathbf{w}}(S;\phi), A_{\mathbf{v}}(S;\phi)) := \mathbb{E}_{z,\tilde{z}\sim\mathbb{D}}[\ell(A_{\mathbf{w}}(S;\phi), A_{\mathbf{v}}(S;\phi); z,\tilde{z})].$$

We consider the following empirical risk as the approximation:

$$R_S(A_{\mathbf{w}}(S;\phi), A_{\mathbf{v}}(S;\phi)) := \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \ell(A_{\mathbf{w}}(S;\phi), A_{\mathbf{v}}(S;\phi); z_i, z_j).$$

We consider Pairwise SGDA with a general sampling scheme, where the random index pairs follow from a general distribution.

We denote $\mathbf{w}_1$ and $\mathbf{v}_1$ the initial points. Let $\nabla_{\mathbf{w}}\ell$ and $\nabla_{\mathbf{v}}\ell$ be the gradients with respect to $\mathbf{w}$ and $\mathbf{v}$, respectively. Let P be a uniform distribution over $([n]\times[n])^T$ and S be a training dataset with n samples. Let (i_t, j_t) from set $\{(i_t, j_t) : i_t, j_t \in [n], i_t \neq j_t\}$ be drawn uniformly at random. At the t-th iteration, with step size sequence $\{\eta_t\}$, Pairwise SGDA updates the model as follows:

$$\begin{cases} \mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}), \\ \mathbf{v}_{t+1} = \mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} \ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}). \end{cases}$$

Before giving the results for Pairwise SGDA, we restate the assumptions we need, adapted to the new setting with two distinct parameter vectors $\mathbf{w}$ and $\mathbf{v}$ [38,52].

Assumption 4. *Let* $L \geq 0$*. We say a differentiable function* ℓ *is L-Lipschitz with respect to* $\mathbf{w}$ *and* $\mathbf{v}$ *if the following holds: For any* $z,\tilde{z} \in \mathcal{Z}$, $\mathbf{w} \in \mathcal{W}$, $\mathbf{v} \in \mathcal{V}$*, we have*

$$\|\nabla_{\mathbf{w}}\ell(\mathbf{w},\mathbf{v};z,\tilde{z})\|_2 \leq L \quad \textit{and} \quad \|\nabla_{\mathbf{v}}\ell(\mathbf{w},\mathbf{v};z,\tilde{z})\|_2 \leq L.$$

Assumption 5. *Let $\alpha > 0$. We say a differentiable function ℓ is α-smooth if the following inequality holds for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}$ and $z, \tilde{z} \in \mathcal{Z}$:*

$$\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_1, \mathbf{v}_1; z, \tilde{z}) - \nabla_{\mathbf{w}} f(\mathbf{w}_2, \mathbf{v}_2; z, \tilde{z}) \\ \nabla_{\mathbf{v}} f(\mathbf{w}_1, \mathbf{v}_1; z, \tilde{z}) - \nabla_{\mathbf{v}} f(\mathbf{w}_2, \mathbf{v}_2; z, \tilde{z}) \end{pmatrix} \right\|_2 \leq \alpha \left\| \begin{pmatrix} \mathbf{w}_1 - \mathbf{w}_2 \\ \mathbf{v}_1 - \mathbf{v}_2 \end{pmatrix} \right\|_2.$$

Assumption 6 (Convexity-Concavity)**.** *We say ℓ is concave if $-\ell$ is convex. We say ℓ is convex-concave if $\ell(\cdot, \mathbf{v})$ is convex for every $\mathbf{v} \in \mathcal{V}$ and $\ell(\mathbf{w}, \cdot)$ is concave for every $\mathbf{w} \in \mathcal{W}$.*

Now, we apply Lemma 1 to develop bounds for Pairwise SGDA in both smooth and non-smooth cases. In the following lemma, proved in Appendix B.2, we establish the sub-exponential stability of Pairwise SGDA.

Lemma 3 (Sub-exponential stability of Pairwise SGDA)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}, \{\mathbf{w}'_t, \mathbf{v}'_t\}$ be two parameter sequences produced by Pairwise SGDA with fixed step sizes and uniform sampling* P *while being trained on neighboring samples S and S'. Suppose the loss is Lipschitz and convex (cf. Assumptions 4–6). Then, we have the following:*

(1) At the t-th iteration, we have sub-exponential stability (Definition 2) with

$$c_1 = 2\sqrt{2e}L^2\eta(\sqrt{t} + 2t/n) \quad \text{and} \quad c_2 = 4\sqrt{2e}L^2\eta(1 + \sqrt{2t/n}).$$

(2) If in addition the loss is also smooth (cf. Assumption 5), then at the t-th iteration, sub-exponential stability (Definition 2) holds with

$$c_1 = 4\sqrt{e}L^2\eta\exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n) \text{ and } c_2 = 8\sqrt{e}L^2\eta\exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + \sqrt{2t/n}).$$

We combine the above lemma with Lemma 1 to obtain bounds for Pairwise SGDA with a general sampling distribution.

Theorem 2 (Generalization bounds for Pairwise SGDA)**.** *Assume ℓ is M-bounded, Lipschitz, and convex (cf. Assumptions 4 and 6). For any $\delta \in (0,1)$, Pairwise SGDA with fixed step sizes satisfies the following generalization guarantees with probability of at least $1-\delta$ over draws of $S \sim \mathcal{D}^n$, for all posterior sampling distributions* Q *on $([n] \times [n])^T$:*

(1) After T iterations, we have

$$\mathbb{E}_{\phi \sim \mathrm{Q}}[G(S,\phi)] \lesssim \left(T\log n - H(\mathrm{Q}) + \log\frac{1}{\delta}\right)\max\left\{L^2\eta(\sqrt{T} + T/n)\log^2 n, \frac{M}{\sqrt{n}}\right\}.$$

(2) If in addition the loss is also smooth (cf. Assumption 5), we have

$$\mathbb{E}_{\phi \sim \mathrm{Q}}[G(S,\phi)] \lesssim \left(T\log n - H(\mathrm{Q}) + \log\frac{1}{\delta}\right) \max\left\{L^2\eta\exp(\alpha^2 t\eta^2)\left(\frac{T}{n} + 1 + \sqrt{\frac{T}{n}}\right)\log^2 n, \frac{M}{\sqrt{n}}\right\}.$$

Let us assume again that $\mathrm{KL}(\mathrm{Q}\|\mathcal{U}) \in \tilde{O}(1)$. As Theorem 2 deals with min-max optimization applicable to minimizing an adversarially robust loss function, this assumption still allows some extra flexibility to account for a few outliers while having the following rates. For part 1, if we choose $T = O(n^2)$ and $\eta = O\left(T^{-3/4}\right)$, this gives a rate of the order $\widetilde{O}(1/\sqrt{n})$. For part 2, if we choose $T = O(n)$ and $\eta = O(1/\sqrt{n})$, this gives the bounds of the order $\widetilde{O}(1/\sqrt{n})$.

5. Algorithmic Implications and Illustrative Experiments

Our theoretical guarantees in the previous section are given up to constant factors. This is common in theoretical analyses, as such results still give useful information about the behavior of bounds with quantities of interest, such as the sample size n. To further verify that our bounds are informative, in this section we show how one can convert them into learning algorithms by minimizing the terms on the r.h.s. of our bounds. We will then illustrate the working of the resulting algorithms in numerical experiments and demonstrate examples of extracting meaningful information from these new algorithms. The goal of this section is to empirically corroborate our theoretical guarantees and demonstrate their potential use for algorithm design.

In line with our theory, we use uniform sampling as the PAC–Bayes prior, and we learn the posterior sampling along with the model's parameters, by alternating minimization of our bounds. We choose $\mathrm{Q} = q^T$ of a factorized form, which corresponds to sampling from training indices $[n]$ with replacement T times during the training trajectory. Minimizing the terms on the r.h.s. of the bounds in Theorem 1 yields an adaptive Pairwise SGD algorithm that we refer to as Pairwise SGD-Q, and likewise minimizing the r.h.s. of the bounds in Theorem 2 yields an adaptive Pairwise SGDA algorithm SGDA-Q. The pseudo-codes of both of these resulting algorithms are given in Algorithm 1 (with the options of Pairwise SGD-Q or Pairwise SGDA-Q).

Algorithm 1 Pairwise SGD-Q/Pairwise SGDA-Q

```
Inputs: $\{(i_t, j_t) : i_t, j_t \in [n], i_t \neq j_t\}, \ell, \nu, T_{iter}, Epochs$
Initialize: $q \leftarrow$ uniform, $\mathbf{w}_0 = 0$, $\mathbf{v}_0 = 0$, $t \leftarrow 1$
for $epoch = 1$ to $Epochs$ do
  for $t = 1$ to $T_{iter}$ do
    Sample $(i_t, j_t) \sim q$
    Pairwise SGD-Q:
      $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; z_{i_t}, z_{j_t})$;
    Pairwise SGDA-Q:
      $\mathbf{v}_{t+1} = \mathbf{v}_t + \eta \nabla_{\mathbf{v}} \ell((\mathbf{w}_t, \mathbf{v}_t); z_{i_t}, z_{j_t})$;
      $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \ell((\mathbf{w}_t, \mathbf{v}_t); z_{i_t}, z_{j_t})$;
    Let $t \leftarrow t + 1$
  end for
  Update all $q$ as $q(i,j) \propto \exp\big(-\frac{1}{\nu}\ell(\mathbf{w}; z_i, z_j)\big)$ for each pair
end for
Return $\mathbf{w}$, $\mathbf{v}$, $q$
```

Before moving on to exemplify our algorithms at work, we should note that there are two in-built guardrails that prevent sampling bias by design, as follows. The r.h.s. of all bounds that we minimize contain two key terms, each acting as a guardrail: (i) Minimizing the $D_{\mathrm{KL}(Q\|P)}$ term is equivalent to maximizing the entropy of Q, i.e., the sampling distribution that we learn (since P is the uniform sampling)—hence, Q will only deviate from uniformity for a good reason (e.g., when encountering misleading or mislabeled training pairs). (ii) Minimizing the *expected* empirical risk (rather than the empirical risk itself) has the consequence that the correction term that usually appears in existing non-uniform sampling-based SGD-type algorithms simply cancels out, automatically ensuring unbiased gradients by construction and without an external correction. Overall, this illustrates the advantage of learning algorithms obtained by minimizing generalization bounds, as the two terms together minimize an upper bound on the quantity we care about, i.e., the true risk.

The forthcoming numerical experiments are meant to showcase the way in which our adaptive sampling methods enhance robustness by learning to down-weight misleading or mislabeled training examples, thereby avoiding being misled on them. As a byproduct, these weights provide explanatory information about the data pairs.

Numerical Results in Pairwise Preference Learning

We illustrate the working of our bound-based pairwise learning algorithms on a toy problem involving pairwise ranking on synthetic 2D data. Pairwise ranking is the task of inferring relative preferences by comparing items in pairs. It has broad applicability, including information retrieval, recommendation systems, preference modeling [53], and positive-unlabeled learning [12]. In this section, we present experiments using a simple linear preference model to demonstrate how our theoretical findings manifest in practice. The aim of this section is not to compete with state-of-the-art empirical methods, but rather to provide insight into the behavior of the algorithms and the use of our bounds under controlled conditions.

We generate $n = 50$ i.i.d. points from a 2D standard Gaussian and assign "preference scores" to each using a hidden linear function $s_{true}(x_i) = \mathbf{w}_{true}^T x_i, i \in [n]$, where $\mathbf{w}_{\text{true}} = (1.5, -1)^T$ is fixed. We then sample 1000 pairs from this data and assign binary labels $\mathbb{1}(\mathbf{w}_{true} x_i \leq \mathbf{w}_{true} x_j)$ indicating which of two items is preferred in each sampled pair.

We use the resulting labeled pairs in the form of difference vectors $x_i - x_j$ (where $i, j \in [n], i \neq j$) as inputs to our Pairwise SGD-Q algorithm to train a linear model with cross-entropy loss for 15 epochs of T_{iter} iterations, each set equal to the number of pairs (so $T = T_{iter} \cdot Epochs$), with step size $\eta = 0.1$. We aim to learn a scoring function so that for any two items x_i and x_j the model can say which one ranks higher. The model learns a weight vector $\mathbf{w}$ so that the score of item x is $s(x) = \mathbf{w}^T x$, and we want $\mathbf{w}^T x_i > \mathbf{w}^T x_j$ if item i is preferred to j. That is, the model projects all points onto the direction $\mathbf{w}$ and ranks them by how far they fall along that line.

The results are plotted in Figure 1. The top-left figure shows the 2D data colored by their preference scores. A red arrow shows the ranking direction learned by our Pairwise SGD-Q, i.e., the direction the trained model uses to order points. Its associated "decision boundary" is shown by the dashed line. In this context, the decision boundary represents the level set $\mathbf{w}^T x = 0$ that is the dividing line (more generally, hyperplane) orthogonal to the learned ranking direction. It shows how $\mathbf{w}$ splits the space into higher- and lower-scoring regions.

The top-right figure is a scatter plot showing the relationship between the learned pairwise margins $\mathbf{w}^T(x_i - x_j)$ and the learned sampling probabilities $q(i, j)$ (q-scores for each pair). Ambiguous pairs with small margins, i.e., those the model is less confident about, tend to have higher losses and thus get down-weighted by a lower sampling probability $q(i, j)$. Hence, more confident (larger-margin) pairs are sampled more often.

The bottom-left plot overlays an edge for the 20 lowest q-scored pairs. These are the pairs that have the same preference score but differing feature coordinates—indeed the least helpful pairs for learning the preference change direction. The 20 highest q-scored pairs are shown on the bottom-right plot; these are most informative of the direction of change.

We repeated the experiment using a noisy preference model, where the observed scores are given by $s_{true}(x_i) = \mathbf{w}_{true}^\top x_i + v_i$, with $v_i \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$ for $i \in [n]$. Figure 2 shows the results of training a model of the same form as previously, using our Pairwise SGD-Q. The top-left plot depicts the noisy data overlaid with the learned direction $\mathbf{w}$. The top-right plot shows the pairwise margins $\mathbf{w}^\top(x_i - x_j)$ against the corresponding sampling probabilities $q(i, j)$. While q still decreases near the pairwise margin, the noise now causes mislabeled or ambiguous pairs to receive even lower q values. This becomes evident in the bottom-left plot: the lowest q-scored pairs have similar preference scores but differing features and are no longer orthogonal to $\mathbf{w}$ due to noise. The bottom-right plot shows the highest q-scored pairs, which continue to reflect the most informative preference differences aligned on average with the learned direction.

Next, we plot learning curves to see how the generalization performance of the pairwise ranking model trained with our Pairwise SGD-Q and Pairwise SGDA-Q algorithms varies with the number of i.i.d. items, under both clean and noisy settings, while we keep the number of pairs used for training constant. The Pairwise SGDA-Q experiments represent an instance of adversarial training. This can model, for instance, malicious users, bots, or strategic agents in applications like recommendation systems, crowd-sourced ranking, sports, or election ranking.

We vary $n \in \{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and repeat all experiments on both clean and noisy preference score generation models. This totals 44 different experiment settings, and we perform 50 independent trials for each. The data generation process is the same as before. While the sample size n varies, the number of pairs sampled from these n points for training is set to 1000 throughout, and these pairs are labeled in the same way as before.

Training with Pairwise SGD-Q is performed in the same way as previously described, but this time we set $T_{iter} = 300$ to run for 30 epochs. When training with Pairwise SGDA-Q, at each training epoch the algorithm first computes the pairwise losses after adversarial maximization over weight vector $\mathbf{v}$ constrained to an ℓ_2-ball of radius $\epsilon = 0.05$ around the current model $\mathbf{w}$, using 6 gradient ascent steps with step size $\eta = 0.1$. The resulting adversarially induced losses are used to compute the sampling distribution $q(i,j) \propto \exp(-\ell_{ij})$, from which $T_{iter} = 300$ pairs are drawn to iteratively update $\mathbf{w}$, for 30 epochs.

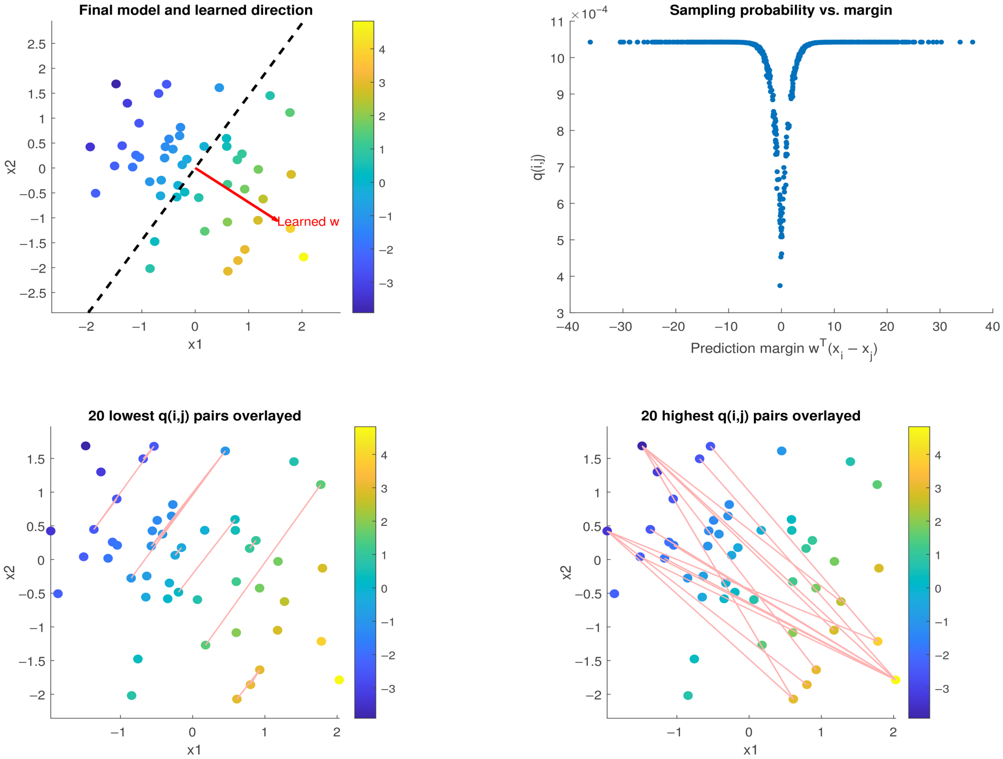

Figure 1. Visualization of Pairwise SGD-Q on a synthetic pairwise ranking task with 2D Gaussian data. (**Top-Left**) Data points colored by true preference scores; the red arrow indicates the learned ranking direction, and the dashed line shows the decision boundary $\mathbf{w}^\top x = 0$. (**Top-Right**) Pairwise margins $\mathbf{w}^\top(x_i - x_j)$ vs. sampling probabilities $q(i,j)$; high-confidence pairs (larger margins) are sampled more frequently. (**Bottom-Left**) The 20 least informative (lowest q) pairs—these are distant pairs having similar preference scores. (**Bottom-Right**) The 20 most informative (highest q) pairs.

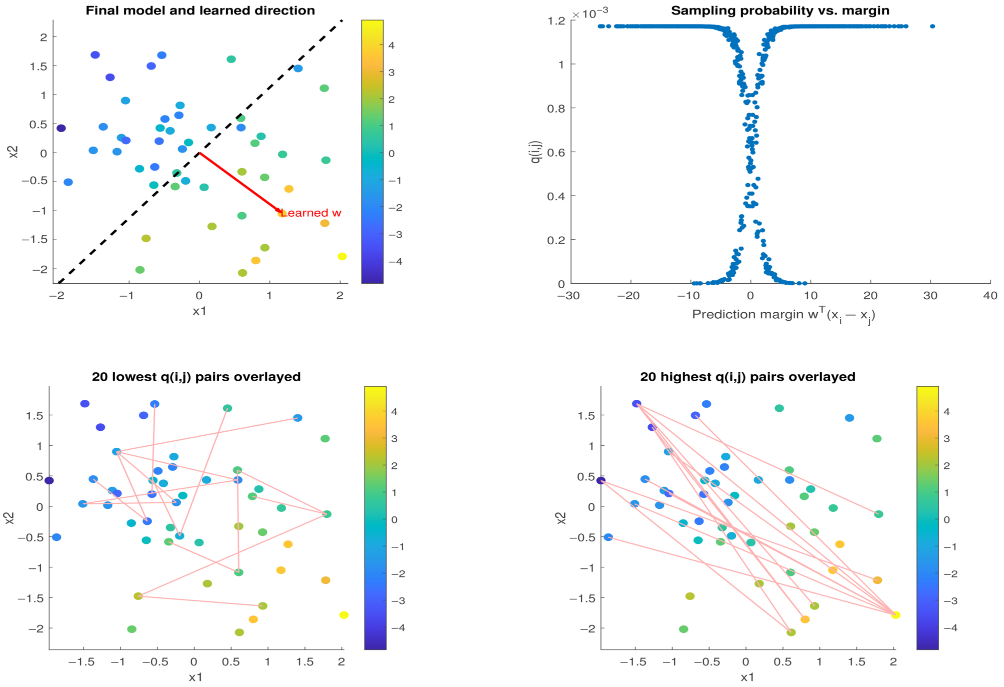

Figure 2. Same setup as Figure 1, but with additive Gaussian noise in the generating preference model. The top-right plot shows that $q(i,j)$ now reflect both the size of pairwise margin and the noise sensitivity, assigning lower weights to mislabeled or ambiguous pairs. The bottom-left plot shows that, as before, Pairwise SGD-Q still down-weights distant pairs that have too similar preference scores, although now these are no longer perpendicular to the ranking direction due to noise. The bottom-right plot shows that Pairwise SGD-Q prioritizes the pairs most aligned with the ranking direction.

For both algorithms, evaluation is performed using an independent test set of 500 unseen items, drawn from the clean preference scoring model. For Pairwise SGD-Q, the out-of-sample errors are computed. For Pairwise SGDA-Q, three different out-of-sample error metrics are computed: (1) standard pairwise error, (2) error under adversarial perturbation of the model's weights **w**, and (3) error under adversarial perturbation of the test pairwise inputs $x_i - x_j$, all within the same ℓ_2 radius and using the same Pairwise SGDA-Q ascent procedure.

Figure 3 reports the obtained learning curves for both algorithms: the error bars show the average and standard error across 50 independent runs. From these figures we can see, as expected, that both natural noise and adversarial perturbations make the problem harder. However, all out-of-sample errors display a decreasing trajectory as the sample size grows.

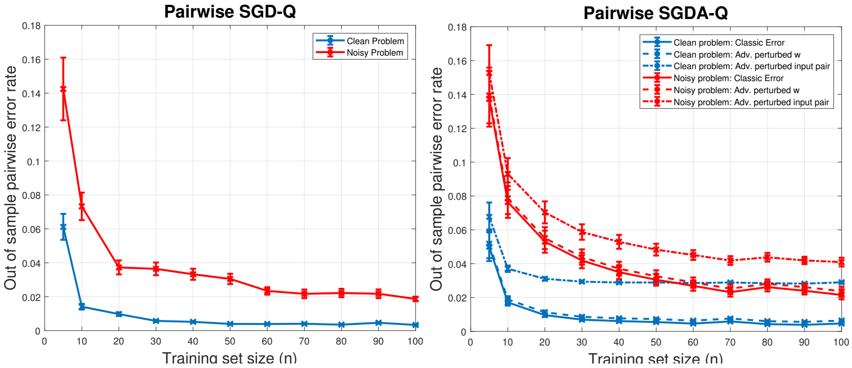

Figure 3. Out-of-sample errors for Pairwise SGD-Q and Pairwise SGDA-Q as a function of the number of i.i.d. items n, while the number of pairs trained on remains constant at 1000. The curves represent averages, and the error bars are standard errors from 50 independent trials.

6. Conclusions

We obtained stability-based PAC–Bayes bounds for randomized pairwise learning, applicable to general sampling. These bounds are applicable to analyzing the generalization of stochastic optimization algorithms, and we demonstrated this in the case of Pairwise SGD and Pairwise SGDA. Our generalization analysis of these methods is suggestive of new stochastic optimizers that allow non-uniform and data-dependent sampling distributions to be updated during the training process. We believe this is a theoretically grounded step that connects two important ideas and may support future work on more complex or application-specific methods. The practical use of this idea is explored in a companion paper [22].

Limitations: Our analysis of Pairwise SGD and Pairwise SGDA is built on a set of classic assumptions regarding the loss function, with convexity being perhaps the most restrictive among them. Nonetheless, insights gained from the convex setting remain a valuable stepping stone to tackling more general, non-convex problems in future work. Indeed, a worthwhile avenue will be to obtain bounds and associated algorithms under relaxed assumptions. Furthermore, here we demonstrated numerical results with our bound-based algorithm under its intended conditions. It will also be interesting to explore experimentally to what extent such algorithms remain functional and potentially useful outside the theoretical conditions in which they were obtained.

Author Contributions: Conceptualization, formal analysis, writing—original draft, S.Z.; conceptualization, supervision, writing—review and editing, Y.L.; conceptualization, validation, supervision, writing—review and editing, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: The work of Sijia Zhou is funded by CSC and UoB scholarship. The work of Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project No. 22303723]. Ata Kabán acknowledges past funding from EPSRC Fellowship grant EP/P004245/1.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Lemma 1

We follow the ideas in [31,54] to prove Lemma 1. We first introduce some useful lemmas. The following lemma shows some results on characterizing sub-Gaussian and sub-exponential random variables. For $\lambda > 0$, let $\mathbb{E}[\exp(\lambda Z)]$ denote the moment-generating function (MGF) of Z. We denote $\mathbb{I}[\cdot]$ the indicator function.

Lemma A1 (Vershynin [55])**.** *Let X be a random variable with $\mathbb{E}[X] = 0$. We have the following equivalences for X:*

- $\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \le \sqrt{p}$*, for all* $p \ge 1$.
- *There exists* $K_1 \ge 0$ *such that, for all* $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda X)] \le \exp(K_1 \lambda^2)$.

We have the following equivalences for X:

- $\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \le p$*, for all* $p \ge 1$.
- *For all* λ *such that* $|\lambda| \le \frac{1}{2e}$, $\mathbb{E}[\exp(\lambda X)] \le \exp(2e^2\lambda^2)$.

The following lemma gives a change in measure of the KL divergence.

Lemma A2 (Lemma 4.10 in Van Handel [56])**.** *For any measurable function $g : \Phi \mapsto \mathbb{R}$, we have*

$$\log \mathbb{E}_{\phi\sim \mathrm{P}}[\exp(g(\phi))] = \sup_{\mathrm{Q}} \big[\mathbb{E}_{\phi\sim\mathrm{Q}}[g(\phi)] - \mathrm{KL}(\mathrm{Q}\|\mathrm{P})\big].$$

We denote the L_p-norm of a random variable Z as $\|Z\|_p := (\mathbb{E}[|Z|^p])^{\frac{1}{p}}, p \geq 1$, denote $S\backslash\{z_i\}$ the set $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$, and abbreviate $\sum_{i,j\in[n]:i\neq j}$ as $\sum_{i\neq j}$. For $z'_k \in \mathcal{Z}$, $S^{(k)}$ is the set derived by replacing the k-th element of S with z'_k.

The following lemma gives moment bounds for a summation of weakly dependent and mean-zero random functions with bounded increments under a small change.

Lemma A3 (Theorem 1 in Lei et al. [33])**.** *Let $S = \{z_1, \ldots, z_n\}$ be a set of independent random variables that each takes values in $\mathcal{Z}$ and $M > 0$. Let $g_{i,j}, \forall i, j \in [n], i \neq j$ be some functions that can be decomposed as $g_{i,j} = g_j^{(i)} + \tilde{g}_i^{(j)}$. Suppose for $g_j^{(i)} : \mathcal{Z}^n \mapsto \mathbb{R}$ and $\tilde{g}_i^{(j)} : \mathcal{Z}^n \mapsto \mathbb{R}$, the following hold for any $i, j \in [n], i \neq j$:*

- $\big|\mathbb{E}_{S\backslash\{z_j\}}[g_j^{(i)}(S)]\big| \leq 2M$, *and* $\big|\mathbb{E}_{S\backslash\{z_i\}}[\tilde{g}_i^{(j)}(S)]\big| \leq 2M$ *almost surely (a.s.).*
- $\mathbb{E}_{z_j}\Big[g_j^{(i)}(S)\Big] = 0$, *and* $\mathbb{E}_{z_i}\Big[\tilde{g}_i^{(j)}(S)\Big] = 0$ *a.s.*
- *For any $j \in [n]$ with $i \neq j, k \neq j$ we have $\big|g_j^{(i)}(S) - g_j^{(i)}(S^{(k)})\big| \leq 2\beta$ a.s., and for any $i \in [n]$ with $j \neq i$ and $k \neq i$, we have $\big|\tilde{g}_i^{(j)}(S) - \tilde{g}_i^{(j)}(S^{(k)})\big| \leq 2\beta$ a.s.*

Then, we can decompose $\sum_{i\neq j} g_j^{(i)}(S)$ and $\sum_{i\neq j} \tilde{g}_i^{(j)}(S)$ as follows:

$$\sum_{i\neq j} g_j^{(i)}(S) = X_1 + X_2, \quad \textit{and} \quad \sum_{i\neq j} \tilde{g}_i^{(j)}(S) = \tilde{X}_1 + \tilde{X}_2$$

where X_1, X_2, $\tilde{X}_1$, $\tilde{X}_2$ are four random variables satisfying $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[\tilde{X}_1] = \mathbb{E}[\tilde{X}_2] = 0$. Furthermore, for any $p \geq 1$

$$\|X_1\|_p \leq 8M\sqrt{p(n-1)}n \textit{ and } \|\tilde{X}_1\|_p \leq 8M\sqrt{p(n-1)}n$$

and for any $p \geq 2$

$$\|X_2\|_p \leq 24\sqrt{2}p(n-1)n\beta\lceil\log_2(n-1)\rceil \textit{ and } \|\tilde{X}_2\|_p \leq 24\sqrt{2}p(n-1)n\beta\lceil\log_2(n-1)\rceil.$$

Proof of Lemma 1. Based on Lemma A2, if we set $g(\phi) = \lambda h(\phi)$, then

$$\mathbb{E}_{\mathrm{Q}}[h(\phi)] \leq \frac{1}{\lambda}(\log \mathbb{E}_{\mathrm{P}}[\exp(\lambda h(\phi))] + \mathrm{KL}(\mathrm{Q}\|\mathrm{P})). \tag{A1}$$

To control the deviations of $\log \mathbb{E}_{\mathrm{P}}[\exp(\lambda h(\phi))]$, we use Markov's inequality. With a probability of $1-\epsilon$, we have

$$\mathbb{E}_{\mathrm{P}}\Big[e^{\lambda h(\phi)}\Big] \leq \frac{\mathbb{E}_S\mathbb{E}_{\mathrm{P}}\Big[e^{\lambda h(\phi)}\Big]}{\epsilon}.$$

Applying the above results to Equation (A1), with a probability of $1-\epsilon$, we get

$$\mathbb{E}_{\mathrm{Q}}[h(\phi)] \leq \frac{1}{\lambda}\Big(\log \mathbb{E}_{\mathrm{P}}\Big[e^{\lambda h(\phi)}\Big] + \mathrm{KL}(\mathrm{Q}\|\mathrm{P})\Big) \leq \frac{1}{\lambda}\left(\log \frac{\mathbb{E}_S\mathbb{E}_{\mathrm{P}}\Big[e^{\lambda h(\phi)}\Big]}{\epsilon} + \mathrm{KL}(\mathrm{Q}\|\mathrm{P})\right). \tag{A2}$$

We can exchange $\mathbb{E}_{\mathrm{P}}$ and $\mathbb{E}_S$ using Fubini's theorem. Next, we will bound the generalization gap with respect to P. Let δ be some decaying function of n. We denote by Ω_δ a subset with $\Pr(\Omega_\delta) \geq 1 - \delta$ on which sub-exponential stability (Definition 2) holds and Ω_δ^c is the complement of Ω_δ. We first give results for any fixed $\phi \in \Omega_\delta$. Given $\phi \in \Omega_\delta$, it was shown in Lei et al. [33], $\forall i, j \in [n]$, that

$$G(S,\phi) \leq 4\beta_\phi + \frac{1}{n(n-1)} \sum_{i \neq j} g_{i,j}(S),$$

$$g_{i,j}(S) = \mathbb{E}_{z_i', z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}[\ell(A(S_{i,j}); Z, \tilde{Z})] - \ell(A(S_{i,j}); z_i, z_j)\Big].$$

As shown in Lei et al. [33], $g_{i,j}$ satisfies all the conditions in Lemma A3, and therefore one can apply Lemma A3 to show the existence of four random variables X_1, X_2, $\tilde{X}_1$, $\tilde{X}_2$ such that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[\tilde{X}_1] = \mathbb{E}[\tilde{X}_2] = 0$:

$$\frac{1}{n(n-1)} \sum_{i \neq j} g_{i,j}(S) = X_1 + X_2 + \tilde{X}_1 + \tilde{X}_2$$

$$\text{and} \quad \|X_1\|_p \leq 8\sqrt{p} M (n-1)^{-\frac{1}{2}},\ \forall p \geq 1, \quad \|\tilde{X}_1\|_p \leq 8\sqrt{p} M (n-1)^{-\frac{1}{2}},\ \forall p \geq 1,$$
$$\|X_2\|_p \leq 24\sqrt{2} p \beta_\phi \lceil \log_2(n-1) \rceil,\ \forall p \geq 2,\ \|\tilde{X}_2\|_p \leq 24\sqrt{2} p \beta_\phi \lceil \log_2(n-1) \rceil,\ \forall p \geq 2.$$

Using the first part of Lemma A1 with $X = X_1 / 8M(n-1)^{-\frac{1}{2}}$ to get

$$\max\{\mathbb{E}_S[\exp(\lambda X_1)], \mathbb{E}_S[\exp(\lambda \tilde{X}_1)]\} \leq \exp(64M^2(n-1)^{-1} K_1 \lambda^2) \tag{A3}$$

and using the second part of Lemma A1 with $X = X_2 / 24\sqrt{2}\beta_\phi \lceil \log_2(n-1) \rceil$,

$$\max\{\mathbb{E}_S[\exp(\lambda X_2)], \mathbb{E}_S[\exp(\lambda \tilde{X}_2)]\} \leq \exp[2304 e^2 \beta_\phi^2 \lceil \log_2(n-1) \rceil^2 \lambda^2], \quad \forall |\lambda| \leq \frac{1}{48e\sqrt{2}\beta_\phi \lceil \log_2(n-1) \rceil}. \tag{A4}$$

According to Jensen's inequality, we have

$$\begin{aligned}\exp(\lambda X_1 + \lambda X_2 + \lambda \tilde{X}_1 + \lambda \tilde{X}_2) &= \exp(\lambda X_1)\exp(\lambda X_2)\exp(\lambda \tilde{X}_1)\exp(\lambda \tilde{X}_2) \\ &\leq \frac{1}{4}(\exp(4\lambda X_1) + \exp(4\lambda X_2) + \exp(4\lambda \tilde{X}_1) + \exp(4\lambda \tilde{X}_2)).\end{aligned}$$

This implies

$$\begin{aligned}&\mathbb{E}_S \exp[\lambda G(S,\phi)] \leq \mathbb{E}_S \exp[\lambda(4\beta_\phi + X_1 + X_2 + \tilde{X}_1 + \tilde{X}_2)] \\ &\leq \exp(4\lambda\beta_\phi)\frac{1}{4}(\mathbb{E}_S[\exp(4\lambda X_1) + \exp(4\lambda X_2) + \exp(4\lambda \tilde{X}_1) + \exp(4\lambda \tilde{X}_2)]).\end{aligned}$$

As sub-exponential stability (Definition 2) holds with $\beta_\phi \leq c_1 + c_2 \log(1/\delta)$ when $\phi \in \Omega_\delta$, the above inequality together with Equations (A3) and (A4) implies that, for all

$$0 < \lambda \leq \frac{1}{192e\sqrt{2}(c_1 + c\log(1/\delta))\lceil \log_2(n-1) \rceil},$$

we have

$$\begin{aligned}\mathbb{E}_S[\exp(\lambda G(S,\phi))] &\leq \exp(4\lambda(c_1 + c\log(1/\delta)))(\exp(256M^2(n-1)^{-1} K_1 \lambda^2) \\ &\quad + \exp(9216 \times (2e)^2 (c_1 + c\log(1/\delta))^2 \lceil \log_2(n-1) \rceil^2 \lambda^2)).\end{aligned} \tag{A5}$$

Next, we give results for any fixed ϕ. We define $H : \mathcal{Z}^n \times \Phi \mapsto \mathbb{R}$ as $H(S, \phi) = G(S, \phi)\mathbb{I}[\phi \in \Omega_\delta]$, where $\mathbb{I}[\cdot]$ is the indicator function. We have

$$\mathbb{E}_{\mathrm{Q}}[G(S,\phi)] = \mathbb{E}_{\mathrm{Q}}[H(S,\phi)] + \mathbb{E}_{\mathrm{Q}}[G(S,\phi)|\phi \in \Omega_\delta^c]\mathrm{Q}(\Omega_\delta^c). \tag{A6}$$

Based on Equation (A.8) and Equation (A.9) in Zhou et al. [31], for $\alpha > 1$, we have

$$\mathbb{E}_{\mathrm{Q}}[G(S,\phi)] \leq \mathbb{E}_{\mathrm{Q}}[H(S,\phi)] + M \inf_{\alpha>1} \delta^{\frac{\alpha-1}{\alpha}} \left(\mathbb{E}_{\mathrm{P}}\left[\left(\frac{\mathrm{Q}(\phi)}{\mathrm{P}(\phi)} \right)^{\alpha} \right] \right)^{\frac{1}{\alpha}}, \tag{A7}$$

where $\ell(A(S;\phi)) \in [0, M]$ and

$$\mathbb{E}_S\mathbb{E}_{\mathrm{P}}[\exp(\lambda H(S,\phi))] \leq \mathbb{E}_S\mathbb{E}_{\mathrm{P}}[\exp(\lambda(G(S,\phi))|\phi \in \Omega_\delta)] + \delta. \tag{A8}$$

Combining the above Equation (A8) with Equation (A5), we obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\mathbb{E}_S[\exp(\lambda H(S,\phi))] \leq \exp(2\lambda(c_1 + c\log(1/\delta))) \times \Big(&\exp(256M^2(n-1)^{-1}K_1\lambda^2) + \\ &\exp(9216 \times (2e)^2(c_1 + c_2\log(1/\delta))^2\lceil\log_2(n-1)\rceil^2\lambda^2)\Big) + \delta. \end{aligned}\tag{A9}$$

For any $u, v, w > 0$ and $\delta \in (0,1)$, we have

$$\exp(u)(\exp(v) + \exp(w)) + \delta \leq \exp(u + 1/2)(\exp(v) + \exp(w)).$$

Applying the above inequality into Equation (A9), if $u = 2\lambda(c_1 + c_2\log(1/\delta))$, $v = 256M^2n^{-1}K_1\lambda^2$, $w = 9216 \times (2e)^2(c_1 + c_2\log(1/\delta))^2\lceil\log_2 n\rceil^2\lambda^2$, it gives

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\mathbb{E}_S[\exp(\lambda H(S,\phi))] \leq \exp(2\lambda(c_1 + b\log(1/\delta)) + 1/2) \times \Bigg(&\exp(256M^2\frac{K_1\lambda^2}{n-1}) + \\ &\exp(9216 \times (2e)^2\left(c_1 + c_2\log(\frac{1}{\delta})\right)^2\lceil\log_2(n-1)\rceil^2\lambda^2)\Bigg). \end{aligned}\tag{A10}$$

We choose

$$\lambda = \min\left\{ \frac{1}{192e\sqrt{2}(c_1 + c_2\log(1/\delta))\lceil\log_2(n-1)\rceil}, \frac{\sqrt{(n-1)}}{16\sqrt{K_1}M} \right\}, \tag{A11}$$

so that we have

$$\begin{gathered} 2\lambda(c_1 + c_2\log(1/\delta)) + 1/2 \leq 1, \\ 256M^2(n-1)^{-1}K_1\lambda^2 \leq 1, \\ 9216 \times (2e)^2(c_1 + c_2\log(1/\delta))^2\lceil\log_2(n-1)\rceil^2\lambda^2 \leq 1. \end{gathered}$$

Plugging this back into Equation (A10) yields the MGF of our truncated generalization gap, $H(S;\phi)$, which is a key quantity in PAC–Bayes analysis:

$$\mathbb{E}_{\mathrm{P}}\mathbb{E}_S[\exp(\lambda H(S,\phi))] \leq e(e+e) \leq e^3.$$

Applying the above results into Equation (A2), we have, with a probability of $1 - \delta'$,

$$\mathbb{E}_{\mathrm{Q}}[H(S,\phi)] \leq \frac{1}{\lambda}\left(\log(e^3/\delta') + \mathrm{KL}(\mathrm{Q}\|\mathrm{P})\right) = \frac{1}{\lambda}(3 + \log(1/\delta') + \mathrm{KL}(\mathrm{Q}\|\mathrm{P})).$$

Based on the above inequality and Equations (A7) and (A8), the following inequality holds uniformly for all Q with probability of at least $1-\delta'$:

$$\begin{aligned}\mathbb{E}_{\phi\sim \mathrm{Q}}[G(S,\phi)] &\le \mathbb{E}_{\phi\sim \mathrm{Q}}[H(S,\phi)] + M\inf_{\alpha>1}\delta^{\frac{\alpha-1}{\alpha}}\left(\mathbb{E}_{\mathrm{P}}\left[\left(\frac{\mathrm{Q}(\phi)}{\mathrm{P}(\phi)}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}}\\ &\le \frac{\mathrm{KL}(\mathrm{Q}\|\mathrm{P})+\log(1/\delta')+3}{\lambda} + M\inf_{\alpha>1}\delta^{\frac{\alpha-1}{\alpha}}\left(\mathbb{E}_{\mathrm{P}}\left[\left(\frac{\mathrm{Q}(\phi)}{\mathrm{P}(\phi)}\right)^{\alpha}\right]\right)^{\frac{1}{\alpha}}.\end{aligned}$$

In the above, comparing the first and the second term on the r.h.s, the second term can be made negligible by choosing δ small enough, $\delta \ll n^{-T}$.

Therefore, our analysis shows

$$\mathbb{E}_{\phi\sim \mathrm{Q}}[G(S,\phi)] \lesssim (\mathrm{KL}(\mathrm{Q}\|\mathrm{P})+\log(1/\delta_1))\max\left\{(c_1+c_2\log(n))\lceil\log_2 n\rceil, \frac{M}{\sqrt{n}}\right\}.$$

The proof is completed. □

Here, we discuss the existence of $\mathbb{E}_{\phi\sim \mathrm{P}}\left[\left(\frac{\mathrm{Q}(\phi)}{\mathrm{P}(\phi)}\right)^{\alpha}\right]$. In practice, we consider Q and P to be sampling distributions. In these cases, Q and P are discrete distributions on the same dataset. In particular, we are interested in the case with P being the uniform distribution. Under these circumstances, this expectation exists.

Appendix B. Proofs for Pairwise SGD and Pairwise SGDA with Adaptive Sampling

Appendix B.1. Pairwise Stochastic Gradient Descent

We will prove that stability bounds of Pairwise SGD satisfy sub-exponential stability (Definition 2). Based on this, we can derive the generalization bounds for Pairwise SGD with smooth and non-smooth convex loss functions. To this aim, we introduce the following lemma to bound the summation of i.i.d events [57].

Lemma A4 (Chernoff's Bound)**.** *Let $Z_1,\dots,Z_t$ be independent random variables taking values in $\{0,1\}$. Let $Z=\sum_{k=1}^t Z_k$ and $\mu=\mathbb{E}[Z]$. Then, for any $\delta\in(0,1)$ with probability of at least $1-\delta$, we have*

$$Z \le \mu + \log(1/\delta) + \sqrt{2\mu\log(1/\delta)}.$$

We first present the stability bounds for non-smooth and convex cases.

Proof of Lemma 2, (1). Without loss of generality, we assume S and S' differ by the last example. Based on the Equation (F.2) in Lei et al. [51], we have

$$\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2 \le 4L^2\eta^2(1+p)^{\sum_{k=1}^t \mathbb{I}[i_k=n \text{ or } j_k=n]}\left(t+p^{-1}\sum_{k=1}^t \mathbb{I}[i_k=n \text{ or } j_k=n]\right).$$

We set $p=1/\sum_{k=1}^t \mathbb{I}[i_k=n \text{ or } j_k=n]$ and use the inequality $(1+1/x)^x\le e$ to get

$$\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2 \le 4eL^2\eta^2\left(t+\left(\sum_{k=1}^t \mathbb{I}[i_k=n \text{ or } j_k=n]\right)^2\right).$$

It then follows that

$$\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2 \le 2\sqrt{e}L\eta\left(\sqrt{t}+\sum_{k=1}^t \mathbb{I}[i_k=n \text{ or } j_k=n]\right).$$

According to the Lipschitz continuity, we know that Pairwise SGD is β_ϕ-uniformly stable with

$$\beta_\phi = 2\sqrt{e}L^2\eta\left(\sqrt{t} + \max_{k\in[n]}\sum_{m=1}^{t}\mathbb{I}[i_m = k \ \text{ or } \ j_m = k]\right). \tag{A12}$$

To bound β_ϕ w.h.p., we set $\beta_{\phi,k} = 2\sqrt{e}L^2\eta\left(\sqrt{t} + \sum_{m=1}^{t}\mathbb{I}[i_m = k \ \text{ or } \ j_m = k]\right)$ and note that $\mathbb{E}[\mathbb{I}[i_m = k \ \text{ or } \ j_m = k]] \le \Pr\{i_m = k\} + \Pr\{j_m = k\} = 2/n$. Applying Lemma A4 to the sum in Equation (A12), with probability of at least $1-\delta/n$, we get

$$\beta_{\phi,k} \le 2\sqrt{e}L^2\eta(\sqrt{t} + 2t/n + \log(n/\delta) + 2\sqrt{t/n\log(n/\delta)}).$$

Therefore, with probability of at least $1-\delta$, the following holds simultaneously for all $k \in [n]$ by the union bound on probability

$$\beta_{\phi,k} \le 2\sqrt{e}L^2\eta(\sqrt{t} + 2t/n + \log(n/\delta) + 2\sqrt{t/n\log(n/\delta)}).$$

For $\delta \in (0, 1/n)$, this implies the following inequality with probability of at least $1-\delta$:

$$\beta_\phi \le 2\sqrt{e}L^2\eta(\sqrt{t} + 2t/n + 2\log(1/\delta) + 2\sqrt{2t/n\log(1/\delta)}). \tag{A13}$$

Finally, from Equation (A13) we know that Pairwise SGD with the uniformly distributed hyperparameter ϕ satisfies sub-exponential stability (Definition 2) with

$$c_1 = 2\sqrt{e}L^2\eta(\sqrt{t} + 2t/n), \quad c_2 = 4\sqrt{e}L^2\eta(1 + \sqrt{2t/n}).$$

The proof is completed. □

Proof of Lemma 2, (2). By an intermediate result in the proof in Lemma C.3 of Lei et al. [33], for all $z, \tilde{z} \in \mathcal{Z}$ and $i_k, j_k \in [n], i_k \neq j_k$, with L-Lipschitz, we have

$$|\ell(\mathbf{w}_{t+1}; z, \tilde{z}) - \ell(\mathbf{w}_{t+1}; z, \tilde{z})| \le L\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \le 2L^2\sum_{k=1}^{t}\eta_k\mathbb{I}[i_k = n \text{ or } j_k = n].$$

From this inequality it follows that Pairwise SGD is β_ϕ-uniformly stable with

$$\beta_\phi = 2L^2\max_{k\in[n]}\sum_{m=1}^{t}\eta_m\mathbb{I}[i_m = k \ \text{ or } \ j_m = k]. \tag{A14}$$

Let $\beta_{\phi,k} = 2L^2\sum_{m=1}^{t}\eta_j\mathbb{I}[i_m = k \ \text{ or } \ j_m = k]$ for any $k \in [n]$. It remains to show that the stability parameter of Pairwise SGD satisfies sub-exponential stability (Definition 2). Using Lemma A4 with $Z_m = \mathbb{I}[i_m = k \ \text{ or } \ j_m = k]$ and noting that $\mathbb{E}[\mathbb{I}[i_m = k \ \text{ or } \ j_m = k]] \le 2/n$, we get the following inequality with probability of at least $1-\delta/n$ (taking $\eta_j = \eta$):

$$\beta_{\phi,k} \le 2L^2\eta(2t/n + \log(n/\delta) + 2\sqrt{t/n\log(n/\delta)}). \tag{A15}$$

By the union bound, with probability of at least $1-\delta$, Equation (A15) holds for all $k \in [n]$. Therefore, with probability of at least $1-\delta$, it gives

$$\beta_\phi \le 2L^2\eta(2t/n + \log(n/\delta) + 2\sqrt{t/n\log(n/\delta)}) \le 2L^2\eta(2t/n + 2\log(1/\delta) + 2\sqrt{2t/n\log(1/\delta)}) \le 4L^2\eta t/n + 4L^2\eta(1 + \sqrt{2t/n})\log(1/\delta),$$

where we have used $\delta \in (0, 1/n)$ in the second inequality. Hence, sub-exponential stability (Definition 2) holds with

$$c_1 = 4L^2\eta t/n, c_2 = 4L^2\eta(1 + \sqrt{2t/n}).$$

This completes the proof. □

Proof of Theorem 1 . With $A(S;\phi) = \mathbf{w}_T$, it follows from Lemma 2, (1) and (2), that Pairwise SGD with both convex non-smooth and convex smooth loss functions satisfies sub-exponential stability (Definition 2). Applying the upper bound on β_ϕ to Lemma 1, the result follows. □

Appendix B.2. Pairwise Stochastic Gradient Descent Ascent

Next, we prove the generalization bounds for Pairwise SGDA with smooth and non-smooth convex-concave loss functions.

Lemma A5 (Lemma C.1., [37]). *Let ℓ be convex-concave.*

(1) *If Assumption 4 holds, then*

$$\left\| \begin{pmatrix} \mathbf{w} - \eta\nabla_{\mathbf{w}}\ell(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta\nabla_{\mathbf{v}}\ell(\mathbf{w}, \mathbf{v}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}' - \eta\nabla_{\mathbf{w}}\ell(\mathbf{w}', \mathbf{v}') \\ \mathbf{v}' + \eta\nabla_{\mathbf{v}}\ell(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 \le \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2 + 8L^2\eta^2.$$

(2) *If Assumption 5 holds, then*

$$\left\| \begin{pmatrix} \mathbf{w} - \eta\nabla_{\mathbf{w}}\ell(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta\nabla_{\mathbf{v}}\ell(\mathbf{w}, \mathbf{v}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}' - \eta\nabla_{\mathbf{w}}\ell(\mathbf{w}', \mathbf{v}') \\ \mathbf{v}' + \eta\nabla_{\mathbf{v}}\ell(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 \le \left(1 + \alpha^2\eta^2\right) \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2.$$

Proof of Lemma 3, (1). We assume S and S' differ by the last example for simplicity. Based on Lemma A5 (1), for $i_t \neq n, j_t \neq n,$ and $i_t \neq j_t$, we have

$$\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \le \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8L^2\eta_t^2. \tag{A16}$$

When $i_t = n$ or $j_t = n, i_t \neq j_t$, we have

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 &\le \left\| \begin{pmatrix} \mathbf{w}_t - \eta_t\nabla_{\mathbf{w}}\ell\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t\nabla_{\mathbf{w}}\ell\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t}, z'_{j_t}) \\ \mathbf{v}_t + \eta_t\nabla_{\mathbf{v}}\ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}) - \mathbf{v}'_t - \eta_t\nabla_{\mathbf{v}}\ell(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t}, z'_{j_t}) \end{pmatrix} \right\|_2^2 \\ &\le (1 + p) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + (1 + \frac{1}{p})\eta_t^2 \left\| \begin{pmatrix} \nabla_{\mathbf{w}}\ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}) - \nabla_{\mathbf{w}}\ell(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t}, z'_{j_t}) \\ \nabla_{\mathbf{v}}\ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}) - \nabla_{\mathbf{v}}\ell(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t}, z'_{j_t}) \end{pmatrix} \right\|_2^2 \\ &\le (1 + p) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8(1 + 1/p)\eta_t^2 L^2, \end{aligned} \tag{A17}$$

where in the second inequality, we use that, for any $p > 0$, we have $(c + d)^2 \le (1 + p)c^2 + (1 + 1/p)d^2$. Combining Equations (A16) and (A17), this gives

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \le &\left(\left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8L^2\eta_t^2 \right) \mathbb{I}[i_t \neq n \text{ and } j_t \neq n] + \\ &\left((1 + p) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8(1 + 1/p)\eta_t^2 L^2 \right) \mathbb{I}[i_t = n \ \text{ or } \ j_t = n] \\ \le (1 + p\mathbb{I}[i_t = n \ \text{ or } \ j_t = n]) &\left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8L^2\eta_t^2(1 + \mathbb{I}[i_t = n \ \text{ or } \ j_t = n]/p). \end{aligned}$$

We apply the above inequality recursively and follow the analysis of Equation (C.4) in Lei et al. [37]:

$$\begin{aligned}
&\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \\
\leq & 8L^2\eta^2 \sum_{k=1}^{t} (1 + \mathbb{I}[i_k = n \ \text{ or } \ j_k = n]/p) \prod_{r=k+1}^{t} (1 + p\mathbb{I}[i_r = n \ \text{ or } \ j_r = n]) \\
= & 8L^2\eta^2 \sum_{k=1}^{t} (1 + \mathbb{I}[i_k = n \ \text{ or } \ j_k = n]/p) \prod_{r=k+1}^{t} (1 + p)^{\mathbb{I}[i_r = n \ \text{ or } \ j_r = n]} \\
\leq & 8L^2\eta^2 (1+p)^{\sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n]} \left(t + \sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n]/p \right),
\end{aligned}$$

where we assume fixed step sizes. We set $p = 1/\sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n]$ and use the inequality $(1 + 1/x)^x \leq e$ to derive

$$\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8eL^2\eta^2 \left(t + \left(\sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n] \right)^2 \right).$$

It then follows that

$$\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2 \leq \sqrt{8e}L\eta \left(\sqrt{t} + \sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n] \right).$$

By L-Lipschitzness, we have

$$\begin{aligned}
&|\ell(A_{\mathbf{w}}(S;\phi), A_{\mathbf{v}}(S;\phi), z, \tilde{z}) - \ell(A_{\mathbf{w}}(S';\phi), A_{\mathbf{v}}(S';\phi), z, \tilde{z})| \\
&\qquad\qquad \leq 2\sqrt{2e}L^2\eta \left(\sqrt{t} + \max_{k \in [n]} \sum_{r=1}^{t} \mathbb{I}[i_r = k \text{ or } j_r = k] \right).
\end{aligned}$$

Therefore, we know that Pairwise SGDA is β_ϕ-uniformly stable with

$$\beta_\phi = 2\sqrt{2e}L^2\eta \left(\sqrt{t} + \max_{k \in [n]} \sum_{r=1}^{t} \mathbb{I}[i_r = k \text{ or } j_r = k] \right). \tag{A18}$$

For simplicity, let $\beta_{\phi,k} = 2\sqrt{2e}L^2\eta \left(\sqrt{t} + \sum_{r=1}^{t} \mathbb{I}[i_r = n \text{ or } j_r = n] \right)$. Applying Lemma A4 to Equation (A18), with probability of at least $1 - \delta/n$, we have

$$\beta_{\phi,k} \leq 2\sqrt{2e}L^2\eta(\sqrt{t} + 2t/n + \log(n/\delta) + 2\sqrt{t/n \log(n/\delta)}).$$

With probability of at least $1 - \delta$, the following holds for all $k \in [n]$:

$$\beta_{\phi,k} \leq 2\sqrt{2e}L^2\eta(\sqrt{t} + 2t/n + \log(n/\delta) + 2\sqrt{t/n \log(n/\delta)}).$$

This suggests the following inequality with probability of at least $1 - \delta$:

$$\beta_\phi \leq 2\sqrt{2e}L^2\eta(\sqrt{t} + 2t/n + 2\log(1/\delta) + 2\sqrt{2t/n \log(1/\delta)}).$$

This suggests that Pairwise SGDA with uniform sampling distribution and the hyperparameter ϕ satisfies sub-exponential stability (Definition 2) with

$$c_1 = 2\sqrt{e}L^2\eta(\sqrt{t} + 2t/n), \quad c_2 = 4\sqrt{2e}L^2\eta(1 + \sqrt{2t/n}).$$

The proof is completed. □

Proof of Lemma 3, (2). Without loss of generality, we first assume S and S' differ by the last example. Based on Lemma A5 (2), if $i_t \neq n$ and $j_t \neq n$, we have

$$\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}) \\ \mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} \ell(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, z_{j_t}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}'_t - \eta_t \nabla_{\mathbf{w}} \ell(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t}, z_{j_t}) \\ \mathbf{v}'_t + \eta_t \nabla_{\mathbf{v}} \ell(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t}, z_{j_t}) \end{pmatrix} \right\|_2^2 \leq (1 + \alpha^2\eta_t^2) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2.$$

When $i_t = n$ or $j_t = n$, we consider Equation (A17). Combining these two cases, we get

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \leq & \left(1 + \alpha^2\eta_t^2\right) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 \mathbb{I}[i_t \neq n \text{ and } j_t \neq n] + \\ & \left((1+p) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8(1 + \frac{1}{p})\eta_t^2 L^2 \right) \mathbb{I}[i_t = n \text{ or } j_t = n] \\ \leq & \left(1 + \alpha^2\eta_t^2 p\mathbb{I}[i_t = n \text{ or } j_t = n]\right) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + 8(1 + \frac{1}{p})\eta_t^2 L^2 \mathbb{I}[i_t = n \text{ or } j_t = n]. \end{aligned} \quad \text{(A19)}$$

We apply the above Equation (A19) recursively, following the proof of Theorem 2(d) in Lei et al. [37],

$$\begin{aligned} & \left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \\ \leq & 8(1 + 1/p)L^2 \sum_{k=1}^{t} \eta_k^2 \mathbb{I}[i_k = n \;\text{ or }\; j_k = n] \prod_{r=k+1}^{t} \left(1 + \alpha^2\eta_r^2 + p\mathbb{I}[i_r = n \;\text{ or }\; j_r = n]\right) \\ \leq & 8(1 + \frac{1}{p})L^2\eta^2 \sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n] \prod_{r=k+1}^{t} (1 + \alpha^2\eta_r^2) \prod_{r=k+1}^{t} (1 + p\mathbb{I}[i_r = n \text{ or } j_r = n]) \\ = & 8(1 + 1/p)L^2\eta^2 \sum_{k=1}^{t} \mathbb{I}[i_k = n \;\text{ or }\; j_k = n] \prod_{r=k+1}^{t} \left(1 + \alpha^2\eta_r^2\right) \prod_{r=k+1}^{t} (1 + p)^{\mathbb{I}[i_r = n \;\text{ or }\; j_r = n]} \\ \leq & 8(1 + 1/p)L^2\eta^2 \prod_{k=1}^{t} \left(1 + \alpha^2\eta_k^2\right) \prod_{k=1}^{t} (1 + p)^{\mathbb{I}[i_k = n \text{ or } j_k = n]} \sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n] \\ \leq & 8(1 + 1/p)L^2\eta^2 \exp\left(\alpha^2 \sum_{k=1}^{t} \eta_k^2\right) (1 + p)^{\sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n]} \sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n], \end{aligned}$$

where we assume fixed step sizes and use $1 + x \leq e^x$ in the last inequality. We set $p = 1/\sum_{k=1}^{t} \mathbb{I}[i_k = n \;\text{ or }\; j_k = n]$ and use the inequality $(1 + 1/x)^x \leq e$ to derive

$$\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8e \left(1 + \sum_{k=1}^{t} \mathbb{I}[i_k = n \text{ or } j_k = n]\right)^2 L^2\eta^2 \exp\left(\alpha^2 \sum_{k=1}^{t} \eta_k^2\right).$$

Based on L-Lipschitzness and the above inequality, for any two neighboring datasets $S, S' \in \mathcal{Z}^n, \forall z, \tilde{z} \in \mathcal{Z}$, we have

$$\begin{aligned}&|\ell(A_{\mathbf{w}}(S;\phi), A_{\mathbf{v}}(S;\phi), z, \tilde{z}) - \ell(A_{\mathbf{w}}(S';\phi), A_{\mathbf{v}}(S';\phi), z, \tilde{z})| \\ &\qquad \leq 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2) \max_{k\in[n]}\left(1 + \sum_{r=1}^{t} \mathbb{I}[i_r = k \text{ or } j_r = k]\right).\end{aligned}$$

Therefore, we know that Pairwise SGDA is β_ϕ-uniformly stable with

$$\beta_\phi = 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2) \max_{k\in[n]}\left(1 + \sum_{r=1}^{t} \mathbb{I}[i_r = k \text{ or } j_r = k]\right).$$

For simplicity, let $\beta_{\phi,k} = 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)\big(1 + \sum_{r=1}^{t} \mathbb{I}[i_r = k \text{ or } j_r = k]\big)$ for any $k \in [n]$. Taking the expectation over both sides of the above inequality, we derive

$$c_1 = 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n), \tag{A20}$$

where $\mathbb{E}[\mathbb{I}[i_r = k \text{ or } j_r = k]] \leq 2/n$. Applying $Z_r = \mathbb{I}[i_r = k \text{ or } j_r = k]$ in Lemma A4, we get the following inequality with probability of at least $1 - \delta/n$:

$$\beta_{\phi,k} \leq 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n + \log(n/\delta) + 2\sqrt{t/n \log(n/\delta)}). \tag{A21}$$

By the union bound in probability, with probability of at least $1 - \delta$, Equation (A21) holds for all $k \in [n]$. Therefore, with probability of at least $1 - \delta$,

$$\begin{aligned}\beta_\phi &\leq 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n + \log(n/\delta) + 2\sqrt{t/n \log(n/\delta)}) \\ &\leq 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n + 2\log(1/\delta) + 2\sqrt{2t/n \log(1/\delta)}) \\ &\leq 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n) + 8\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + \sqrt{2t/n}) \log(1/\delta) \\ &\leq c_1 + 8\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + \sqrt{2t/n}) \log(1/\delta),\end{aligned}$$

where we have used $\delta \in (0, 1/n)$ in the second inequality and Equation (A20) in the last inequality. Therefore, sub-exponential stability (Definition 2) holds with $c_1 = 4\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + 2t/n)$ and $c_2 = 8\sqrt{e}L^2\eta \exp(\frac{1}{2}\alpha^2 t\eta^2)(1 + \sqrt{2t/n})$. The proof is completed. □

Based on the above lemma, we are ready to develop generalization bounds in Theorem 2 for Pairwise SGDA with smooth and non-smooth loss functions.

Proof of Theorem 2. With $A(S;\phi) = (A_{\mathrm{w}}(S;\phi), A_{\mathrm{v}}(S;\phi))$, based on Lemma 3, (1) and (2), Pairwise SGDA with both convex-concave non-smooth as well as convex-concave smooth loss functions satisfies sub-exponential stability (Definition 2). Applying the upper bounds on β_ϕ to Lemma 1, we obtain the result. □

References

1. Lei, G.; Shi, L. Pairwise ranking with Gaussian kernel. *Adv. Comput. Math.* **2024**, *50*, 70. [CrossRef]
2. Agarwal, S.; Niyogi, P. Generalization bounds for ranking algorithms via algorithmic stability. *J. Mach. Learn. Res.* **2009**, *10*, 441–474.
3. Clémençon, S.; Lugosi, G.; Vayatis, N. Ranking and empirical minimization of U-statistics. In *The Annals of Statistics*; Institute of Mathematical Statistics: Beachwood, OH, USA, 2008; pp. 844–874.

4. Cortes, C.; Mohri, M. AUC optimization vs. error rate minimization. *Adv. Neural Inf. Process. Syst.* **2004**, *16*, 313–320.
5. Cao, Q.; Guo, Z.C.; Ying, Y. Generalization bounds for metric and similarity learning. *Mach. Learn.* **2016**, *102*, 115–132. [CrossRef]
6. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
7. Xiong, F.; Gou, M.; Camps, O.; Sznaier, M. Person re-identification using kernel-based metric learning methods. In *Proceedings of the European Conference Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1–16.
8. Guillaumin, M.; Verbeek, J.; Schmid, C. Is that you? Metric learning approaches for face identification. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 498–505.
9. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–20. [CrossRef]
10. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the 2014 IEEE 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 34–39.
11. Feng, B.; Liu, Z.; Huang, N.; Xiao, Z.; Zhang, H.; Mirzoyan, S.; Xu, H.; Hao, J.; Xu, Y.; Zhang, M.; et al. A bioactivity foundation model using pairwise meta-learning. *Nat. Mach. Intell.* **2024**, *6*, 962–974. [CrossRef]
12. Sellamanickam, S.; Garg, P.; Selvaraj, S.K. A pairwise ranking based approach to learning with positive and unlabeled examples. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, NY, USA, 24–28 October 2011; CIKM '11; pp. 663–672.
13. Wang, Z.; Liang, P.; Bai, R.; Liu, Y.; Zhao, J.; Yao, L.; Zhang, J.; Chu, F. Few-shot fault diagnosis for machinery using multi-scale perception multi-level feature fusion image quadrant entropy. *Adv. Eng. Inform.* **2025**, *63*, 102972. [CrossRef]
14. Zhao, P.; Zhang, T. Stochastic Optimization with Importance Sampling for Regularized Loss Minimization. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1–9.
15. Allen-Zhu, Z.; Qu, Z.; Richtárik, P.; Yuan, Y. Even faster accelerated coordinate descent using non-uniform sampling. In Proceedings of the International Conference on Machine Learning. PMLR, New York, NY, USA, 19–24 June 2016; pp. 1110–1119.
16. Katharopoulos, A.; Fleuret, F. Biased importance sampling for deep neural network training. *arXiv* **2017**, arXiv:1706.00043. [CrossRef]
17. Johnson, T.B.; Guestrin, C. Training deep models faster with robust, approximate importance sampling. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–11.
18. Wu, C.Y.; Manmatha, R.; Smola, A.J.; Krahenbuhl, P. Sampling matters in deep embedding learning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2840–2848.
19. Han, X.; Yu, X.; Li, G.; Zhao, J.; Pan, G.; Ye, Q.; Jiao, J.; Han, Z. Rethinking sampling strategies for unsupervised person re-identification. *IEEE Trans. Image Process.* **2022**, *32*, 29–42. [CrossRef]
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
21. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
22. Zhou, S.; Lei, Y.; Kabán, A. Learning to Sample in Stochastic Optimization. In Proceedings of the 41st Confenence on Uncertainty in Artificial Intelligence, Rio de Janeiro, Brazil, 21–25 July 2025.
23. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
24. Zhou, M.; Patel, V.M. Enhancing adversarial robustness for deep metric learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15325–15334.
25. Wen, W.; Li, H.; Wu, R.; Wu, L.; Chen, H. Generalization analysis of adversarial pairwise learning. *Neural Netw.* **2025**, *183*, 106955. [CrossRef] [PubMed]
26. Liu, W.; Wang, Z.J.; Yao, B.; Yin, J. Geo-ALM: POI Recommendation by Fusing Geographical Information and Adversarial Learning Mechanism. *Int. Jt. Conf. Artif. Intell.* **2019**, *7*, 1807–1813.
27. Zhang, L.; Duan, Q.; Zhang, D.; Jia, W.; Wang, X. AdvKin: Adversarial convolutional network for kinship verification. *IEEE Trans. Cybern.* **2020**, *51*, 5883–5896. [CrossRef]
28. De la Pena, V.; Giné, E. *Decoupling: From Dependence to Independence*; Springer Science & Business Media: New York, NY, USA, 2012.
29. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA, 2–4 May 2013; Workshop Track Proceedings, 2013.
30. Beznosikov, A.; Gorbunov, E.; Berard, H.; Loizou, N. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In Proceedings of the International Conference on Artificial Intelligence and Statistics. PMLR, Valencia, Spain, 25–27 April 2023; pp. 172–235.

31. Zhou, S.; Lei, Y.; Kabán, A. Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36.
32. London, B. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2931–2940.
33. Lei, Y.; Ledent, A.; Kloft, M. Sharper Generalization Bounds for Pairwise Learning. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–10 December 2020; Volume 33, pp. 21236–21246.
34. Katharopoulos, A.; Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. In Proceedings of the International Conference on Machine Learning. PMLR, Stockholm Sweden, 10–15 July 2018; pp. 2525–2534.
35. Bousquet, O.; Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.* **2002**, *2*, 499–526.
36. Hardt, M.; Recht, B.; Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1225–1234.
37. Lei, Y.; Yang, Z.; Yang, T.; Ying, Y. Stability and Generalization of Stochastic Gradient Methods for Minimax Problems. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 6175–6186.
38. Farnia, F.; Ozdaglar, A. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 3174–3185.
39. Shawe-Taylor, J.; Williamson, R.C. A PAC analysis of a Bayesian estimator. In Proceedings of the Tenth Annual Conference on Computational Learning Theory, Nashville, TN, USA, 6–9 July 1997; pp. 2–9.
40. McAllester, D.A. Some pac-bayesian theorems. *Mach. Learn.* **1999**, *37*, 355–363. [CrossRef]
41. London, B.; Huang, B.; Getoor, L. Stability and generalization in structured prediction. *J. Mach. Learn. Res.* **2016**, *17*, 7808–7859.
42. Rivasplata, O.; Parrado-Hernández, E.; Shawe-Taylor, J.S.; Sun, S.; Szepesvári, C. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 9214–9224.
43. Sun, S.; Yu, M.; Shawe-Taylor, J.; Mao, L. Stability-based PAC-Bayes analysis for multi-view learning algorithms. *Inf. Fusion* **2022**, *86*, 76–92. [CrossRef]
44. Oneto, L.; Donini, M.; Pontil, M.; Shawe-Taylor, J. Randomized learning and generalization of fair and private classifiers: From PAC-Bayes to stability and differential privacy. *Neurocomputing* **2020**, *416*, 231–243. [CrossRef]
45. Mou, W.; Wang, L.; Zhai, X.; Zheng, K. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In Proceedings of the Conference on Learning Theory, Stockholm, Sweden, 6–9 July 2018; pp. 605–638.
46. Negrea, J.; Haghifam, M.; Dziugaite, G.K.; Khisti, A.; Roy, D.M. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
47. Li, J.; Luo, X.; Qiao, M. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
48. Ralaivola, L.; Szafranski, M.; Stempfel, G. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *J. Mach. Learn. Res.* **2010**, *11*, 1927–1956.
49. Viallard, P.; Germain, P.; Habrard, A.; Morvant, E. A General Framework for the Derandomization of PAC-Bayesian Bounds. *ArXiv* **2021**, arXiv:2102.08649.
50. Picard-Weibel, A.; Clerico, E.; Moscoviz, R.; Guedj, B. How good is PAC-Bayes at explaining generalisation? *arXiv* **2025**, arXiv:2503.08231. http://arxiv.org/abs/2503.08231.
51. Lei, Y.; Liu, M.; Ying, Y. Generalization Guarantee of SGD for Pairwise Learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21216–21228.
52. Zhang, J.; Hong, M.; Wang, M.; Zhang, S. Generalization bounds for stochastic saddle point problems. In Proceedings of the International Conference on Artificial Intelligence and Statistics. PMLR, Virtual, 13–15 April 2021; pp. 568–576.
53. Liu, T.Y. *Learning to Rank for Information Retrieval.*; Springer: Berlin/Heidelberg, Germany, 2011.
54. Guedj, B.; Pujol, L. Still no free lunches: The price to pay for tighter PAC-Bayes bounds. *Entropy* **2021**, *23*, 1529. [CrossRef]
55. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*; Cambridge University Press: Cambridge, UK, 2018; Volume 47.
56. Van Handel, R. Probability in high dimension. In *Lecture Notes*; Princeton University: Princeton, NJ, USA, 2014.
57. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.