Not All Semantics are Created Equal: Contrastive Self-supervised Learning with Automatic Temperature Individualization

Zi-Hao Qiu [1] **Quanqi Hu** [2] **Zhuoning Yuan** [3] **Denny Zhou** [4] **Lijun Zhang** [1] **Tianbao Yang** [2]

Abstract

In this paper, we aim to optimize a contrastive loss with *individualized* temperatures in a principled manner. The common practice of using a *global* temperature parameter τ ignores the fact that "not all semantics are created equal", meaning that different anchor data may have different numbers of samples with similar semantics, especially when data exhibits long-tails. First, we propose a new *robust contrastive loss* inspired by distributionally robust optimization (DRO), providing us an intuition about the effect of τ and a mechanism for automatic temperature individualization. Then, we propose an efficient stochastic algorithm for optimizing the robust contrastive loss with a provable convergence guarantee without using large mini-batch sizes. Theoretical and experimental results show that our algorithm automatically learns a suitable τ for each sample. Specifically, samples with frequent semantics use large temperatures to keep local semantic structures, while samples with rare semantics use small temperatures to induce more separable features. Our method not only outperforms prior strong baselines (e.g., SimCLR, CLIP) on unimodal and bimodal tasks with larger improvements on imbalanced data but also is less sensitive to hyper-parameters. To our best knowledge, this is the first methodical approach to optimizing a contrastive loss with individualized temperatures. Our proposed algorithms are implemented in the LibAUC library at `https://libauc.org/`.

Most work of Z.H. Qiu was done when visiting the OptMAI lab at TAMU. [1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2]Computer Science and Engineering, Texas A&M University, College Station, USA [3]Department of Computer Science, the University of Iowa, Iowa City, USA [4]Google Research, USA. Correspondence to: Tianbao Yang <tianbao-yang@tamu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Self-supervised learning (SSL) is a promising way to learn data representations that generalize across downstream tasks. Specifically, contrastive learning (CL) has laid the foundation for state-of-the-art SSL models due to its effectiveness (Chen et al., 2020; He et al., 2020; Tomasev et al., 2022; Huang et al., 2022). CL aims to push the similarity scores between "positive" pairs (e.g., augmented views of the same image) to be higher than that between "negative" pairs (e.g., augmented views from different images), which has great promises in leveraging large amount of unlabelled data (Goyal et al., 2021; Radford et al., 2021). Moreover, CL has been extended to a broader scope, e.g., bimodal image-text SSL (Zhang et al., 2020; Radford et al., 2021), where images and language text descriptions can be regarded as multi-modal views of the same underlying concept. The well-known CLIP (Radford et al., 2021) method shows that models learned from millions of image-text pairs can attain impressive recognition performance for a wide range of visual understanding tasks.

In general, contrastive methods share a common design of the softmax-based loss function, For a given anchor data i, a contrastive loss can be written as:

$$\mathcal{L}^i_{\text{con}} = -\log \frac{\exp\left(\text{sim}(z_i, z_i^+)/\tau\right)}{\sum_{k\neq i}\exp\left(\text{sim}(z_i, z_k)/\tau\right)}, \qquad (1)$$

where z_i is the feature of the anchor data, z_i^+ is the feature of a different 'view' of data i and called a *positive* sample, $z_k (k \neq i)$ are the features of other samples and called *negative* samples, τ is the temperature parameter, and $\text{sim}(\cdot,\cdot)$ measures the similarity between two input vectors. The positive pair can be also added to the denominator, which does not affect our discussion here. A significant property of the loss is the **hardness-aware** property (Wang & Liu, 2021; Zhang et al., 2022; Xia et al., 2022). Consider the gradient of $\tau\mathcal{L}^i_{\text{con}}$ w.r.t. model parameters $\mathbf{w}$, i.e., $\tau\nabla_{\mathbf{w}}\mathcal{L}^i_{\text{con}}$:

$$-\nabla_{\mathbf{w}}\text{sim}(z_i, z_i^+) + \sum_{k\neq i}\frac{\exp\left(\text{sim}(z_i, z_k)/\tau\right)}{\sum_{k\neq i}\exp\left(\text{sim}(z_i, z_k)/\tau\right)}\nabla_{\mathbf{w}}\text{sim}(z_i, z_k).$$

Note that the weight for the gradient of a negative pair (z_i, z_k) is proportional to $\exp\left(\text{sim}(z_i, z_k)/\tau\right)$. Thus, the contrastive loss automatically penalizes negative pairs according to their hardness (hard means $\text{sim}(z_i, z_k)$ is large).

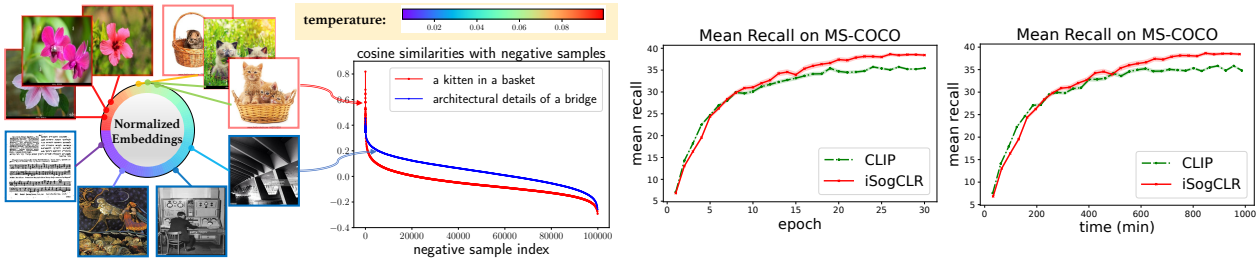

Figure 1. Left: Samples with *frequent* semantics (e.g., a kitten in a basket) have many more similar samples than that with *rare* semantics (e.g., architectural details of a bridge). Middle: An illustration of temperature individualization by our algorithm named iSogCLR, making "hot" images with frequent semantics use a higher temperature to keep semantic structures, and making "cold" images with rare semantics use a lower temperature for inducing more separable features. The circular heatmap is plotted using learned temperatures of 100 random images from the CC3M dataset. Right: Convergence curves of CLIP and iSogCLR, where we use the CLIP implementation and training settings from open-clip [2]. We train the models on CC3M and evaluate image retrieval performance on MS-COCO.

The temperature τ plays a critical role in controlling the penalty strength on negative samples (Wang & Liu, 2021; Zhang et al., 2022). Specifically, a small τ penalizes much more on hard negative samples (i.e., the degree of hardness-awareness is high), causing separable embedding space. However, the excessive pursuit to the separability may break the underlying *semantic structures* because some negative samples with high similarity scores to the anchor data might indeed contain similar semantics, to which we refer as pseudo negatives. In contrast, a large τ tends to treat all negative pairs equally (i.e., the degree of hardness-awareness is low) and is more tolerant to pseudo negative samples, which is beneficial for keeping local semantic structures.

Real-world data distributions always exhibit long tails (Zhu et al., 2014; Feldman, 2020) and the frequency of samples with different semantics can be extremely diverse. In Figure 1, we show some images with frequent or rare semantics from the CC3M dataset (Sharma et al., 2018). We further select two representative images, namely "a kitten in a basket" that contains frequent semantics and "architectural details of a bridge" that contains rare semantics, and present the cosine similarities between these two images and other 100,000 random texts from the same dataset. Note that images with frequent semantics have much more similar samples. To improve feature qualities, samples with frequent semantics should be assigned with a *large* τ to better capture the local semantic structure, while using a small τ will push semantically consistent samples away. On the other hand, samples with rare semantics should have a *small* τ to make their features more discriminative and separable. We refer to these effects as *semantics harmonizing*. Unfortunately, most existing CL methods treat the temperature parameter as a *global* parameter, which does not accommodate different semantics and restricts their performance in real-world applications.

In this paper, we propose a provable stochastic algorithm for optimizing a contrastive loss with individualized temperatures. First, inspired by *distributionally robust optimization (DRO)* (Namkoong & Duchi, 2017; Duchi et al., 2021), we design a novel *robust global contrastive loss (RGCL)* for each anchor data. RGCL introduces a distributional variable for *all negative samples* of each anchor data (this explains "global" in RGCL), and a KL divergence constraint between the distributional variable and the uniform distribution. We show that RGCL is hardness-aware by optimizing the distributional variable, and the KL constraint affects the degree of hardness-awareness. We further demonstrate that the dual formulation of RGCL induces a loss function that can be solved efficiently and contains an individualized learnable temperature parameter. In a spirit of stochastic optimization of a global contrastive loss (SogCLR) (Yuan et al., 2022), we propose an efficient optimization algorithm named **iSogCLR** for solving the dual formulation of RGCL by synthesizing advanced techniques of compositional optimization (Wang & Yang, 2022) and of solving KL constrained DRO (Qi et al., 2022). We establish a convergence guarantee of our algorithm without large mini-batch sizes, which is similar to that of SogCLR for optimizing a global contrastive loss with a fixed τ. Experiments on unimodal and bimodal datasets indicate that iSogCLR achieves superior performance, especially on imbalanced data. More in-depth analyses demonstrate the relationship between the semantics of data and their learned temperatures. An illustration of some images and their learned temperatures by our method is plotted in Figure 1. Besides, ablation studies show that iSogCLR is much less sensitive to its hyper-parameters. We summarize our contributions below:

- We propose a new *robust contrastive loss* inspired by DRO, and study its properties and connections with existing softmax-based contrastive losses.
- We propose a novel and provable stochastic algorithm called iSogCLR for optimizing the robust contrastive losses with automatic temperature individualization.
- We conduct comprehensive experiments on unimodal and bimodal CL to demonstrate the superior performance of our method, and the relationship between the semantics of data and their learned temperatures.

[2] https://github.com/mlfoundations/open_clip

2. Related Work

Self-supervised Learning. SSL methods can be divided into two main categories: CL methods and non-CL methods. Although non-CL methods do not rely on negative samples and achieve comparable performance to CL methods, they often require additional projector (Grill et al., 2020), stop gradient (Chen & He, 2021), or momentum encoder (Richemond et al., 2020). Several non-CL methods (Ermolov et al., 2021; Zbontar et al., 2021; Bardes et al., 2021) use information maximization techniques, which decorrelate the variables of each embedding to avoid an informational collapse. Nevertheless, CL methods remain a mainstream framework for SSL and are extended to many other fields (Khaertdinov et al., 2021; Aberdam et al., 2021; Eun et al., 2020), and are shown to be better than non-CL methods in the human learning setting (Zhuang et al., 2022).

CL Methods. Pioneering methods (Chen et al., 2020; He et al., 2020) construct pairs and optimize InfoNCE loss (Oord et al., 2018) in a direct way. Later, numerous works improve the performance by penalizing hard negatives (Wu et al., 2020; Kalantidis et al., 2020; Chen et al., 2021; Robinson et al., 2021; Xie et al., 2022; Zhang et al., 2022), generating better views (Tamkin et al., 2021; Tian et al., 2020; Ge et al., 2021; Wang & Qi, 2021), using prototypes (cluster centers) for contrasting (Caron et al., 2020; Li et al., 2020), and handling pseudo negative samples (Chuang et al., 2020; Dwibedi et al., 2021). Recently, HaoChen et al. (2021) propose a novel loss based on spectral decomposition on the population graph with accuracy guarantees. To address the issue that most CL methods rely on large batch sizes, Yuan et al. (2022) propose a provable algorithm named SogCLR for optimizing a global contrastive loss, and achieve promising results without large mini-batch sizes. Although great progress has been made, most methods ignore the imbalance of semantics in real-world data, and lack the ability to adapt to different types of semantics automatically.

Bimodal Contrastive Learning. Vision-and-language pretraining (VLP) is a rapidly growing field. Due to its effectiveness, CL has been extended to representative works such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), which are pretrained on millions of web-crawled image-text pairs and achieve astounding results. Later, DeCLIP (Li et al., 2021b), FILIP (Yao et al., 2021), SLIP (Mu et al., 2022) and CyCLIP (Goel et al., 2022) improve CLIP by introducing more supervisions or bringing in fine-grained cross-modal interactions. Our algorithm tackles a fundamental problem on optimizing individualized temperatures for a contrastive loss. Thus it can be applied in bimodal setting seamlessly. Besides, because the web-crawled bimodal data often exhibits long tail distributions (Wang et al., 2022), we observe that our algorithm is more suitable in such scenario and achieves great improvements compared with baselines.

Optimizing τ in CL. The impact of τ on the success of CL is remarkable and noticed in prior works. Zhang et al. (2022); Khaertdinov et al. (2021) propose to improve negative mining in CL by using different temperatures for positive and negative samples, where temperatures can be fixed values or input-dependent functions. In bimodal CL, CLIP (Radford et al., 2021) proposes to treat τ as a learnable variable, which is adopted by later works (Goel et al., 2022; Li et al., 2021a). However, this approach was never rigorously justified. Zhang et al. (2021) show that input-dependent learnable τ is effective to estimate the uncertainty in out-of-distribution detection, but with the cost of sacrificing the performance on downstream tasks. Different from previous methods that set or learn temperatures heuristically, we present a new view of the contrastive loss based on DRO, which explains the role of τ mathematically, and enables automatic optimization of individualized temperatures.

Distributionally Robust Optimization. DRO has been extensively studied in machine learning and statistics (Bertsimas et al., 2018; Staib & Jegelka, 2019; Duchi et al., 2021). Mathematically, DRO seeks a model that performs well regardless of perturbing the sample distribution within an uncertainty set, which is specified by a divergence measure between the perturbed distribution and the observed empirical distribution (Ben-Tal et al., 2013; Blanchet et al., 2019; Duchi et al., 2021). Recent works (Qi et al., 2020; 2021; 2022; Levy et al., 2020; Jin et al., 2021; Gürbüzbalaban et al., 2022; Zhu et al., 2023) have proposed efficient stochastic algorithms for solving different DRO formulations. Our algorithm is inspired by that of Qi et al. (2022), which considers a similar DRO problem with the uncertainty set specified by a KL constraint, and proposes efficient dual-free algorithms with convergence guarantees. However, different from their work that considers an ordinary compositional objective with only *one* KL constraint, we deal with a more complex *coupled* compositional function (Wang & Yang, 2022) and *many* KL constraints for all anchor data, which complicate the convergence analysis.

3. Preliminaries

Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ denote a set of training images with size n. $\mathcal{P}$ denotes a set of data augmentation operators. Denoted by $\mathcal{S}_i^- = \{\mathcal{A}(\mathbf{x}) : \forall \mathcal{A} \in \mathcal{P}, \forall \mathbf{x} \in \mathcal{D} \setminus \mathbf{x}_i\}$ the set of negative data for the anchor image $\mathbf{x}_i$. Let $E(\cdot)$ denote the image encoder. For bimodal tasks, let $\mathcal{D}' = \{(\mathbf{x}_1, \mathbf{t}_1), \ldots, (\mathbf{x}_n, \mathbf{t}_n)\}$ denote n image-text pairs. Let $\mathcal{T}_i^- = \{\mathbf{t}_j \in \mathcal{D}', j \neq i\}$ be the set of negative texts for the anchor image $\mathbf{x}_i$, and $\mathcal{I}_i^- = \{\mathbf{x}_j \in \mathcal{D}', j \neq i\}$ be the set of negative images for the anchor text $\mathbf{t}_i$. Let $E_I(\cdot)$ and $E_T(\cdot)$ denote the encoder for images and texts in bimodal CL, respectively. Let $\mathbf{w}$ denote the model parameters. Denote by Δ_n a simplex of dimension n and by $\delta_\Omega(\cdot)$ a dirac

function that returns zero if input belongs to the set Ω or infinity otherwise. Let $\mathrm{KL}(\cdot,\cdot)$ denote the KL divergence.

For unimodal CL, a *global contrastive loss (GCL)* (Yuan et al., 2022) for the i-th image $\mathbf{x}_i$ can be defined as:

$$\ell_{\mathrm{GCL}}(\mathbf{x}_i) = -\tau \log \frac{\exp(E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathcal{A}'(\mathbf{x}_i))/\tau)}{\sum_{\mathbf{z}\in\mathcal{S}_i^-} \exp\left(E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathbf{z})/\tau\right)}, \quad (2)$$

where $\mathcal{A}, \mathcal{A}' \in \mathcal{P}$. Compared with a contrastive loss defined over mini-batch samples (Chen et al., 2020; He et al., 2020), GCL multiplies τ on the right side to ensure the gradient is not illy scaled. Besides, GCL considers all negative samples $\mathcal{S}_i^-$ for $\mathbf{x}_i$ in the denominator, enabling us to analyze the optimization error and design algorithms to control the error.

To simplify (2) and facilitate our statements, we define the following auxiliary function:

$$h_i(\mathbf{z}) := E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathbf{z}) - E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathcal{A}'(\mathbf{x}_i)). \quad (3)$$

In fact, $h_i(\mathbf{z})$ measures the *hardness score* of $\mathbf{z}$ with respect to $\mathbf{x}_i$. Then (2) can be rewritten as:

$$\ell_{\mathrm{GCL}}(\mathbf{x}_i) = \tau \log \sum\nolimits_{\mathbf{z}\in\mathcal{S}_i^-} \exp(h_i(\mathbf{z})/\tau). \quad (4)$$

For bimodal tasks, we consider the two-way GCL, which for the i-th image-text pair is defined as

$$\ell(\mathbf{x}_i, \mathbf{t}_i) = -\tau \log \frac{\exp(E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i)/\tau)}{\sum_{\mathbf{t}\in\mathcal{T}_i^-} \exp(E_I(\mathbf{x}_i)^\top E_T(\mathbf{t})/\tau)} - \tau \log \frac{\exp(E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i)/\tau)}{\sum_{\mathbf{x}\in\mathcal{I}_i^-} \exp(E_I(\mathbf{x})^\top E_T(\mathbf{t}_i)/\tau)},$$

where the first term is an image-to-text contrastive loss, i.e., try to predict $\mathbf{t}_i$ from $\mathbf{t} \in \mathcal{D}'$ based on $\mathbf{x}_i$, and the second term is a symmetrical text-to-image contrastive loss.

Similar to (4), we can simplify $\ell(\mathbf{x}_i, \mathbf{t}_i)$ as:

$$\ell(\mathbf{x}_i, \mathbf{t}_i) = \tau \log \sum_{\mathbf{t}\in\mathcal{T}_i^-} \exp\left[\frac{h_{\mathbf{x}_i}(\mathbf{t})}{\tau}\right] + \tau \log \sum_{\mathbf{x}\in\mathcal{I}_i^-} \exp\left[\frac{h_{\mathbf{t}_i}(\mathbf{x})}{\tau}\right],$$

where $h_{\mathbf{x}_i}(\mathbf{t}) = E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}) - E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i)$ and $h_{\mathbf{t}_i}(\mathbf{x}) = E_I(\mathbf{x})^\top E_T(\mathbf{t}_i) - E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i)$. Our algorithm is applicable to both unimodal and bimodal CL.

A general DRO formulation is given by (Levy et al., 2020):

$$\min_{\mathbf{w}} \max_{\mathbf{p}\in\mathcal{U}} \sum\nolimits_{i=1}^{n} \mathbf{p}_i l_i(\mathbf{w}) - \lambda D(\mathbf{p}, \mathbf{1}/n), \quad (5)$$

where $\mathbf{p}$ is a distributional variable, $l_i(\mathbf{w})$ is the loss on sample i, $\mathcal{U} \subset \Delta_n$ is an uncertainty set of the distributional variable specified by some divergence constraint. Maximizing the objective over $\mathbf{p}$ leads to larger weights on samples with larger losses, which actually finds *the worst case* loss. DRO then minimizes the worst-case loss to make models achieve the robustness against potential distribution shifts.

4. Robust Global Contrastive Objectives

We first introduce a novel *robust global contrastive loss (RGCL)*, including its formulation and properties. Then we convert RGCL into a simpler equivalent minimization form with individualized temperatures by Lagrangian duality theory. We further give a theoretical explanation of our objective for optimizing temperatures. Due to the limited space, we describe our method in unimodal setting. For our method in bimodal setting, we present its final minimization form and defer detailed derivation to Appendix B.

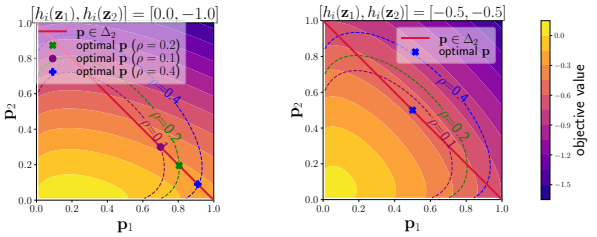

Figure 2. Contours of $\ell_{\mathrm{RGCL}}(\mathbf{x}_i)$ ($|\mathcal{S}_i^-| = 2$) in (6) for two $h_i = [h_i(\mathbf{z}_1), h_i(\mathbf{z}_2)]$ vectors: $[0.0, -1.0]$ and $[-0.5, -0.5]$. One can observe that $\ell_{\mathrm{RGCL}}(\mathbf{x}_i)$ is hardness-aware, harder sample ($h_i(\mathbf{z}_1)$ on the left) has larger weight ($\mathbf{p}_1 = 0.8$). Moreover, ρ affects the degree of hardness-awareness. Larger ρ means higher degree of hardness-awareness.

4.1. Formulations

Motivated by DRO, we define the following loss for an anchor data $\mathbf{x}_i$ with a set of m negative samples $\mathcal{S}_i^-$:

$$\ell_{\mathrm{RGCL}}(\mathbf{x}_i) := \max_{\mathbf{p}\in\Delta_m} \sum\nolimits_{\mathbf{z}_j\in\mathcal{S}_i^-} \mathbf{p}_j h_i(\mathbf{z}_j) - \tau_0 \mathrm{KL}(\mathbf{p}, \mathbf{1}/m) \quad \text{s.t.} \quad \mathrm{KL}(\mathbf{p}, \mathbf{1}/m) \le \rho, \quad (6)$$

where $m = |\mathcal{S}_i^-|$, $\rho > 0$ is a hyperparameter, and τ_0 is a small positive value by default. There are several features of (6). (i) Similar to the GCL (2), we use *all* negative samples to define the loss. Hence our loss is referred to robust *global* contrastive loss. (ii) We mainly use the KL constraint $\mathrm{KL}(\mathbf{p}, \mathbf{1}/n) \le \rho$ to define an uncertainty set of $\mathbf{p}$. The small KL regularization term $\tau_0 \mathrm{KL}(\mathbf{p}, \mathbf{1}/n)$ is added to make the loss function smooth hence facilitate the optimization (Qi et al., 2022). (iii) Different from (5) that considers a distribution $\mathbf{p}$ over all samples, RGCL considers a distribution $\mathbf{p}$ over all negative samples for *each anchor data*.

To better illustrate the key properties of RGCL, we plot the contours of $\ell_{\mathrm{RGCL}}(\mathbf{x}_i)$ with two negative data in Figure 2. For $[h_i(\mathbf{z}_1), h_i(\mathbf{z}_2)]$ values in (6), we consider two settings, i.e., $[0.0, -1.0]$ and $[-0.5, -0.5]$, which represent $\mathbf{z}_1$ is very similar to the anchor data, and two equally dissimilar negative samples, respectively. The dashed lines in Figure 2 are the boundaries of the KL constraints with different ρ values, whose intersections with the red simplex lines define the feasible regions. From these figures, we observe that (i) $\ell_{\mathrm{RGCL}}(\mathbf{x}_i)$ is also *hardness-aware*. By maximizing over $\mathbf{p}$, harder negative samples will have larger weights (e.g., $\mathbf{p}^* = (0.8, 0.2)$ for the first setting when $\rho = 0.2$). (ii) If the hardness of two negative samples are similar, then their weights tend to be similar too (for the second setting). (iii) The constraint $\mathrm{KL}(\mathbf{p}, \mathbf{1}/m) \le \rho$ actually affects the *degree*

of hardness-awareness. In the left of Figure 2, note that if ρ gets larger, the optimal $\mathbf{p}$ will be more non-uniform, i.e., the degree of hardness-awareness will increase.

Next, we induce an *equivalent* loss with an individualized learnable temperature parameter from (6) and show the intuition about the effect of τ from the DRO view. Since directly optimizing (6) is challenging due to maintaining the high-dimensional distributional variable $\mathbf{p}$, we follow Qi et al. (2022) and adopt the Lagrangian duality theory to convert $\ell_{\text{RGCL}}(\mathbf{x}_i)$ into its dual form (cf. Appendix A):

$$\max_{\mathbf{p}\in\Delta_m}\min_{\lambda\geq 0}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j h_i(\mathbf{z}_j)-\tau_0\text{KL}(\mathbf{p},\mathbf{1}/m)-\lambda(\text{KL}(\mathbf{p},\mathbf{1}/m)-\rho)$$

$$\Leftrightarrow\min_{\lambda\geq 0}(\lambda+\tau_0)\log\sum_{\mathbf{z}\in\mathcal{S}_i^-}\exp(h_i(\mathbf{z})/\lambda)-(\lambda+\tau_0)\log(m)+\lambda\rho$$

$$\Leftrightarrow\min_{\tau\geq\tau_0}\tau\log\mathbb{E}_{\mathbf{z}\in\mathcal{S}_i^-}\exp(h_i(\mathbf{z})/\tau)+(\tau-\tau_0)\rho, \tag{7}$$

where we first introduce a Lagrangian multiplier λ for the KL constraint, and the last equality is due to a variable change $\tau=\lambda+\tau_0$. Notice that the Lagrangian multiplier λ for the KL constraint becomes a learnable parameter τ for $\mathbf{x}_i$. Interestingly, if we fix τ (i.e., using a fixed KL regularization instead of the KL constraint in (6)), the above loss will reduce to $\ell_{\text{GCL}}(\mathbf{x}_i)$ in (2) up to a constant difference. Hence, RGCL introduces the flexibility to optimize individualized temperatures compared with the GCL.

Based on the dual form of RGCL in (7), we define a robust global contrastive objective (RGCO) for unimodal SSL :

$$\min_{\mathbf{w},\boldsymbol{\tau}\geq\tau_0}F(\mathbf{w},\boldsymbol{\tau}):=\frac{1}{n}\sum_{\mathbf{x}_i\in\mathcal{D}}\left\{\boldsymbol{\tau}_i\log\mathop{\mathbb{E}}_{\mathbf{z}\in\mathcal{S}_i^-}\exp\left(\frac{h_i(\mathbf{z})}{\boldsymbol{\tau}_i}\right)+\boldsymbol{\tau}_i\rho\right\},$$

where $\boldsymbol{\tau}_i$ is the individualized temperature for $\mathbf{x}_i$. The RGCO for bimodal SSL is defined similarly:

$$\min_{\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}'\geq\tau_0}F_{\text{B}}(\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}'):=\frac{1}{n}\sum\nolimits_{(\mathbf{x}_i,\mathbf{t}_i)\in\mathcal{D}'}\Bigg[(\boldsymbol{\tau}_i+\boldsymbol{\tau}_i')\rho+$$

$$\boldsymbol{\tau}_i\log\mathbb{E}_{\mathbf{t}\in\mathcal{T}_i^-}\exp\left(\frac{h_{\mathbf{x}_i}(\mathbf{t})}{\boldsymbol{\tau}_i}\right)+\boldsymbol{\tau}_i'\log\mathbb{E}_{\mathbf{x}\in\mathcal{I}_i^-}\exp\left(\frac{h_{\mathbf{t}_i}(\mathbf{x})}{\boldsymbol{\tau}_i'}\right)\Bigg],$$

with individualized temperatures $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_i'$ for images and texts, respectively. A small constant can be added inside the log to ensure its smoothness and Lipschitz continuity as in (Yuan et al., 2022), which is assumed for analysis.

4.2. An Intuitive Theoretical Explanation

To answer why our RGCL can learn suitable temperatures for samples with different semantics intuitively, we compare our iSogCLR (the algorithm for optimizing RCGL) with CLIP, whose learned global τ is 0.01 on CC3M. Considering the representative cat and bridge images, we extract the features of them and 1000 random samples as their negatives for each method. Then we substitute these features into (6) and solve the optimal $\mathbf{p}^*$ (i.e., $\mathbf{p}_j^*=\frac{\exp(h_i(\mathbf{z}_j)/\tau)}{\sum_{\mathbf{z}\in\mathcal{S}_i^-}\exp(h_i(\mathbf{z}_j)/\tau)}$, cf. Appendix A) of each image for both methods. We plot the results in Figure 3. Combining these results with the formulation (7), we have the following explanation.

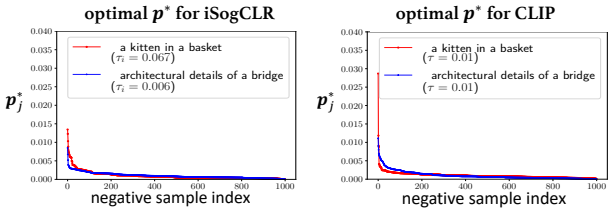

Figure 3. For the anchor images of cat and bridge, we select 1000 negative samples and solve (6) for the optimal $\mathbf{p}^*$ by using h_i values of iSogCLR with learned $\boldsymbol{\tau}_i$ and CLIP with learned τ.

For samples with frequent semantics, due to the hardness-aware property of RGCL, their optimal $\mathbf{p}^*$ in (6) tend to be more non-uniform, and $\text{KL}(\mathbf{p},\mathbf{1}/n)\leq\rho$ is more likely to be violated. Therefore, their Lagrangian multipliers λ in (7) will be large to "push" $\mathbf{p}$ back to be closer to uniform. Due to $\tau=\lambda+\tau_0$, their temperatures will be large. From Figure 3, it is notable that compared with the optimal $\mathbf{p}^*$ of the cat image from CLIP (the red line in the right), that from our RGCL (the red line in the left) is more *uniform*.

For samples with rare semantics, their optimal distributional variables $\mathbf{p}$ in (6) tend to be more uniform, which makes $\text{KL}(\mathbf{p},\mathbf{1}/n)\leq\rho$ being more likely to be satisfied. At this time, their Lagrangian multipliers λ are probably small, and thus their temperatures are small. For example, the learned τ of the bridge image by iSogCLR is 0.006, which is smaller than the final learned $\tau=0.01$ in CLIP.

5. iSogCLR for Stochastic Optimization

In this section, we design a provable algorithm for optimizing $F(\mathbf{w},\boldsymbol{\tau})$. The algorithm for $F_{\text{B}}(\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}')$ is similar and deferred to Appendix B. The objective functions $F(\mathbf{w},\boldsymbol{\tau})$ and $F_{\text{B}}(\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}')$ are special cases of X-risks (Yang, 2022), making the optimization much more challenging than traditional empirical risk minimization. Nonetheless, existing algorithms for deep X-risk optimization are not directly applicable due to the optimization over temperature variables.

Inspired by Yuan et al. (2022) for solving GCL, we cast $F(\mathbf{w},\boldsymbol{\tau})$ as a finite-sum coupled compositional optimization problem (Wang & Yang, 2022):

$$\min_{\mathbf{w},\boldsymbol{\tau}\in\Omega}F(\mathbf{w},\boldsymbol{\tau}):=\frac{1}{n}\sum_{\mathbf{x}_i\sim\mathcal{D}}\underbrace{f_i\left(\boldsymbol{\tau}_i,g_i(\mathbf{w},\boldsymbol{\tau}_i;\mathcal{S}_i^-)\right)}_{F_i(\mathbf{w},\boldsymbol{\tau}_i)}, \tag{8}$$

where $\boldsymbol{\tau}\in\Omega$ is to accommodate the constraint on $\boldsymbol{\tau}$ and

$$f_i(\boldsymbol{\tau}_i,\cdot)=\boldsymbol{\tau}_i\log(\cdot)+\boldsymbol{\tau}_i\rho,$$

$$g_i(\mathbf{w},\boldsymbol{\tau}_i;\mathcal{S}_i^-)=\mathbb{E}_{\mathbf{z}\in\mathcal{S}_i^-}\exp\left(h_i(\mathbf{z})/\boldsymbol{\tau}_i\right).$$

The gradients of F w.r.t. $\mathbf{w}$ and $\boldsymbol{\tau}_i$ can be computed by:

$$\nabla_{\mathbf{w}}F(\mathbf{w},\boldsymbol{\tau})=\frac{1}{n}\sum_{\mathbf{x}_i\in\mathcal{D}}\nabla_{\mathbf{w}}F_i(\mathbf{w},\boldsymbol{\tau}_i) \tag{9}$$

$$=\frac{1}{n}\sum_{\mathbf{x}_i\in\mathcal{D}}\nabla_{g_i}f_i(\boldsymbol{\tau}_i,g_i(\mathbf{w},\boldsymbol{\tau}_i;\mathcal{S}_i^-))\nabla_{\mathbf{w}}g_i(\mathbf{w},\boldsymbol{\tau}_i;\mathcal{S}_i^-),$$

$$\nabla_{\boldsymbol{\tau}_i} F(\mathbf{w}, \boldsymbol{\tau}) = \frac{1}{n} \sum_{\mathbf{x}_j \in \mathcal{D}} \nabla_{\boldsymbol{\tau}_i} F_j(\mathbf{w}, \boldsymbol{\tau}_j) \overset{(a)}{=} \frac{1}{n} \nabla_{\boldsymbol{\tau}_i} F_i(\mathbf{w}, \boldsymbol{\tau}_i) \quad (10)$$

$$= \frac{1}{n} \left(\frac{\boldsymbol{\tau}_i \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)}{g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)} + \log(g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)) + \rho \right),$$

where (a) holds because for $j \neq i$, $F_j(\mathbf{w}, \boldsymbol{\tau}_j)$ does not involve $\boldsymbol{\tau}_i$. Note that the major cost for computing (9) and (10) lies at computing $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)$ and its gradients w.r.t. $\mathbf{w}$ and $\boldsymbol{\tau}_i$, involving all samples in $\mathcal{S}_i^-$. At each iteration, we only sample a random mini-batch of B samples $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_B\}$, and compute an unbiased estimator of $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)$ for each $\mathbf{x}_i \in \mathcal{B}$ by:

$$g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{B}_i) = \frac{1}{|\mathcal{B}_i|} \sum\nolimits_{\mathbf{z} \in \mathcal{B}_i} \exp(h_i(\mathbf{z}) / \boldsymbol{\tau}_i), \quad (11)$$

where $\mathcal{B}_i = \{\mathcal{A}(\mathbf{x}), \mathcal{A}'(\mathbf{x}) : \mathcal{A}, \mathcal{A}' \in \mathcal{P}, \mathbf{x} \in \mathcal{B} \backslash \mathbf{x}_i\}$ contains the negative samples of $\mathbf{x}_i$ in $\mathcal{B}$. However, directly substituting $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{B}_i)$ as the estimator of $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)$ into (9) and (10) will yield biased estimators because (9) and (10) are *non-linear* w.r.t. $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)$. The optimization error will be large when the batch size is small (Yuan et al., 2022).

To control the approximation error and provide a convergence guarantee, we borrow a technique from Yuan et al. (2022) by using a *moving average estimator* to keep track of $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)$ for each $\mathbf{x}_i \in \mathcal{D}$. To this end, we maintain a scalar $\mathbf{s}_i$ for each $\mathbf{x}_i$ and update it at the t-th iteration by:

$$\mathbf{s}_i^{t+1} = (1 - \beta_0)\mathbf{s}_i^t + \beta_0 g_i(\mathbf{w}_t, \boldsymbol{\tau}_i^t; \mathcal{B}_i), \quad (12)$$

where $\beta_0 \in (0, 1)$. Intuitively, when t increases, $\mathbf{w}_{t-1}$ and $\boldsymbol{\tau}^{t-1}$ are getting close to $\mathbf{w}_t$ and $\boldsymbol{\tau}^t$, hence **the previous value of $\mathbf{s}_i^t$ is useful for estimating** g_i. With these stochastic estimators, we compute the gradients of (8) in terms of $\mathbf{w}_t$ and $\boldsymbol{\tau}_i^t$ with *controllable* approximation error by:

$$G(\boldsymbol{\tau}_i^t) = \frac{1}{n} \left[\frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{B}_i) + \log(\mathbf{s}_i^t) + \rho \right], \quad (13)$$

$$G(\mathbf{w}_t) = \frac{1}{B} \sum_{\mathbf{x}_i \in \mathcal{B}} \frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t} \nabla_{\mathbf{w}} g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{B}_i). \quad (14)$$

The complete procedure is presented in Algorithm 1, named **iSogCLR** with **i** standing for individualization of temperatures. In step 1, we initialize all $\boldsymbol{\tau}_i$ to τ_{init}. We implement the momentum update for $\boldsymbol{\tau}_i^{t+1}$ and $\mathbf{w}_{t+1}$ in Step 8, 9 and Step 12, 13, respectively, where $\beta_1 \in (0, 1)$ is the momentum parameter. The momentum-style update can be replaced by an Adam-style update using adaptive step sizes and the same convergence rate can be established (Guo et al., 2021).

In terms of the additional memory cost, while it scales with the number of samples, it typically is not a significant concern in practical applications. Firstly, the additional memory cost is still small compared with the number of model parameters. For example, the additional memory cost for 1 million samples is $2 \times \frac{10^6 \times 4}{1024^2} = 7.63$MB. Secondly, the GPU memory usage can be optimized by storing variables $\mathbf{s}_i$, $\mathbf{u}_i$ and $\boldsymbol{\tau}_i$ in the CPU memory. To further minimize the impact on training time, one can employ an *asynchronous* strategy to transfer data. Specifically, before the t-th iteration, we can prefetch $\mathbf{s}_i^t$, $\mathbf{u}_i^t$ and $\boldsymbol{\tau}_i^t$ from the CPU and transfer them to the GPU. After forward propagation, we conduct back-propagation and asynchronously copy the updated $\mathbf{s}_i^{t+1}$, $\mathbf{u}_i^{t+1}$ and $\boldsymbol{\tau}_i^{t+1}$ back to the CPU memory and fetch a new batch of them from CPU. By utilizing high-bandwidth CPU-GPU interconnects, such as PCIe4 or NVLink, the time required to transfer these variables can effectively overlap with the time of back-propagation. This approach facilitates fast training and reduces GPU memory consumption.

Algorithm 1 iSogCLR

Require: β_0, β_1, η
1: Initialize $\mathbf{w}_1, \mathbf{s}^1, \mathbf{u}^1, \mathbf{v}_1, \boldsymbol{\tau}^1 = \boldsymbol{\tau}_{\text{init}}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Draw a batch of B samples denoted by $\mathcal{B} \subset \mathcal{D}$
4: **for** $\mathbf{x}_i \in \mathcal{B}$ **do**
5: Compute $g_i(\mathbf{w}_t, \boldsymbol{\tau}_i^t; \mathcal{B}_i)$ according to (11)
6: Update $\mathbf{s}_i^{t+1}$ according to (12)
7: Compute $G(\boldsymbol{\tau}_i^t)$ according to (13)
8: Update $\mathbf{u}_i^{t+1} = (1 - \beta_1)\mathbf{u}_i^t + \beta_1 G(\boldsymbol{\tau}_i^t)$
9: Update $\boldsymbol{\tau}_i^{t+1} = \Pi_\Omega \left[\boldsymbol{\tau}_i^t - \eta \mathbf{u}_i^{t+1} \right]$
10: **end for**
11: Compute gradient estimator $G(\mathbf{w}_t)$ according to (14)
12: Compute $\mathbf{v}_{t+1} = (1 - \beta_1)\mathbf{v}_t + \beta_1 G(\mathbf{w}_t)$
13: Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_{t+1}$ (or Adam-style)
14: **end for**

We highlight the differences between RGCO/iSogCLR and GCO/SogCLR (Yuan et al., 2022) and our contributions of analysis: (i) RGCO has additional n temperature variables $\boldsymbol{\tau}_i$; nevertheless we prove the same iteration complexity as SogCLR. (ii) the constraint $\boldsymbol{\tau}_i \geq \tau_0$ makes the analysis more complicated. In particular, to show $F(\mathbf{w}, \boldsymbol{\tau})$ is smooth in terms of $(\mathbf{w}, \boldsymbol{\tau})$, we follow Qi et al. (2022) to derive an upper bound $\tau_{\max}$ for the optimal $\boldsymbol{\tau}_*$ and use the constraint set $\Omega = \{\tau_0 \leq \tau \leq \tau_{\max}\}$ in the analysis. (iii) Due to the constraint on $\boldsymbol{\tau}$, we employ a different notion of stationary point using the regular subgradient (cf. Appendix D) of the non-smooth extended objective $\bar{F}(\mathbf{w}, \boldsymbol{\tau}) = F(\mathbf{w}, \boldsymbol{\tau}) + \delta_\Omega(\boldsymbol{\tau})$, i.e., $\hat{\partial}\bar{F}(\mathbf{w}, \boldsymbol{\tau})$, which also complicates the analysis. Finally, the convergence guarantee of iSogCLR is:

Theorem 1. *Under appropriate conditions and settings of parameters $\beta_0, \beta_1 = \mathcal{O}(B'\epsilon^2)$, $\eta = \mathcal{O}\left(\frac{BB'\epsilon^2}{n}\right)$, where $B = |\mathcal{B}|, B' = |\mathcal{B}_i|$, after $T = \mathcal{O}\left(\frac{n}{BB'\epsilon^4}\right)$ iterations Algorithm 1 finds an ϵ-stationary solution of the problem, i.e., $\mathbb{E}[dist(0, \hat{\partial}\bar{F}(\mathbf{w}_t, \boldsymbol{\tau}_t))^2] \leq \epsilon^2$ for a random $t \in \{1, \ldots, T\}$.*

Remark: The theorem indicates that iSogCLR has the same $\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$ complexity as SogCLR (Yuan et al., 2022). We refer the interested readers to Appendix D for the proof, where we also exhibit the conditions similar to (Yuan et al., 2022).

6. Experiments

In this section, we conduct experiments on unimodal and bimodal datasets and observe that our algorithm outperforms prior strong baselines. Moreover, in-depth analyses show that the samples with different semantics are indeed assigned with suitable temperatures. We also perform ablation studies to better understand the behaviors of iSogCLR. The code to reproduce the results in this paper is available at https://github.com/zhqiu/contrastive-learning-iSogCLR/.

In unimodal setting, We compare our iSogCLR with five CL methods: **SimCLR** (Chen et al., 2020), **FlatCLR** (Chen et al., 2021), **SimCo** (Zhang et al., 2022), **Spectral CL** (HaoChen et al., 2021), **SogCLR** (Yuan et al., 2022), and two non-contrastive methods: **Barlow Twins** (Zbontar et al., 2021) and **VICReg** (Bardes et al., 2021). In bimodal setting, we compare with **CLIP** (Radford et al., 2021), **CyCLIP** (Goel et al., 2022), and **SogCLR**.

For fair comparison, we set the hyper-parameters of all methods using grid search. SimCLR, FlatCLR, and SogCLR contain τ, which is tuned in a range of $\{0.1, 0.3, 0.5, 0.7\}$. For other methods, we fine-tune their hyper-parameters around the recommended values in their papers. Following Radford et al. (2021); Goel et al. (2022), τ is directly optimized in CLIP and CyCLIP. The detailed implementation and data information are in Appendix C.1 and C.2, respectively.

6.1. Unimodal Experiments

Data. We consider three balanced datasets: CIFAR10, CIFAR100, ImageNet100 (Wu et al., 2019), and three imbalanced datasets: CIFAR10-LT, CIFAR100-LT, iNaturalist2018 (Horn et al., 2018). ImageNet100 is a subset with 100 classes from ImageNet1K (Russakovsky et al., 2015). CIFAR10-LT and CIFAR100-LT are created following the Long-Tailed (LT) imbalance setting (Cui et al., 2019) and widely used (Cao et al., 2019; Cui et al., 2019; Qi et al., 2020). The iNaturalist dataset is a large-scale dataset with dramatically different number of images per category. We use its official training and validation splits.

Setup. The backbone network, initial learning rate and batch size are set to ResNet-18, 0.8, and 128 for CIFAR datasets. While for ImageNet100 and iNaturalist2018, they are set to ResNet-50, 1.2, and 256, respectively. The projection head has three linear layers, each with 8192 output units. The first two layers of the projector are followed by a BN layer and rectified linear units. We employ LARS optimizer (You et al., 2017) (with a momentum of 0.9 and weight decay of 1e-4) and cosine learning rate schedule. We also use learning rate warm-up for 10 epochs, i.e., learning rate is gradually increased to the maximum value. We resize input images to 224×224 and follow the same image augmentation strategies as in SimCLR (Chen et al., 2020) including random crop, color distortion, and Gaussian blur. For linear evaluation, we train the last classification layer using SGD with Nesterov momentum with a batch size of 256 for 100 epochs. The initial learning rate is set to 30.0 and decayed by 0.2 at 40, 60, and 80 epochs. We tune β_0 and ρ in our algorithm from $\{0.7, 0.8, 0.9\}$ and $\{0.1, 0.2, 0.3, 0.4\}$, respectively. τ_{init} and τ_0 are set to 0.7 and 0.05 by default.

Results. We present partial results in Table 1 and full results in Table 3, 4 in Appendix C.3. First, comparing iSogCLR, SogCLR and SimCLR, we observe that (i) SogCLR is generally better than SimCLR, showing the advantage of optimizing a GCL under limited mini-batch sizes; and (ii) iSogCLR outperforms SogCLR in all cases, confirming the effectiveness of individualized temperatures. In Figure 4, we visualize the learned embeddings from these three methods on CIFAR10, where each color represents a class. Note that the class boundaries of iSogCLR are more clear than that of others, indicating that iSogCLR indeed improves feature qualities. We also observe that iSogCLR outperforms prior strong baselines, e.g., VICReg, Spectral CL. Besides, iSogCLR achieves larger improvements on imbalanced data, e.g., has relative improvements of 2.37% and 7.56% over SimCLR on CIFAR100 and CIFAR100-LT, respectively.

6.2. Bimodal Experiments

Data. We adopt Conceptual Captions 3M (CC3M) (Sharma et al., 2018) dataset, which is widely used in vision-and-language pretraining (Li et al., 2021b; Mu et al., 2022; Goel et al., 2022). For evaluation, we use two common bimodal datasets: Flickr30K (Plummer et al., 2015), MSCOCO (Lin et al., 2014), obtained from the well-known Karpathy split (Karpathy & Fei-Fei, 2015), and three standard image datasets: CIFAR10, CIFAR100, and ImageNet1K.

Setup. Following recent studies on bimodal SSL (Li et al., 2021a; Dou et al., 2022), we adopt ResNet-50 and DistilBert (Sanh et al., 2019) as the image and text encoder, which are initialized with weights from unimodal pretraining. Specifically, we use the ResNet-50 model pretrained on ImageNet from timm library (Wightman, 2019). The DistilBert model comes from huggingface library (Wolf et al., 2020), which is pretrained on BookCorpus (Zhu et al., 2015) and English Wikipedia. The output embedding of each encoder is then transformed to a lower-dimensional (256-d) representation by a linear layer and normalized for computing contrastive loss. We use a batch size of 512 for 30 epochs pre-training, where the image resolution is 256×256. We employ Adam-W optimizer (Loshchilov & Hutter, 2017) with cosine learning rate decay. The learning rate is warmed-up to 2e-4 in the first 1000 iterations and decayed to 1e-6 by a cosine decay scheduler. We employ Adam-W optimizer (Loshchilov & Hutter, 2017) with the weight decay of 0.02. We tune β_0 and ρ from $\{0.7, 0.8, 0.9\}$ and $\{5.8, 6.0, 6.2, 6.4\}$, respectively. τ_{init} and τ_0 are set to

Table 1. Linear evaluation results with 400 pretraining epochs on six unimodal image datasets. We report the average top-1 accuracies (%) and standard deviation over 3 runs with different random seeds. Full results are provided in Table 3 and 4 in Appendix C.3.

METHOD	CIFAR10	CIFAR100	IMAGENET100	CIFAR10-LT	CIFAR100-LT	INATURALIST
SIMCLR	88.74±0.18	62.34±0.09	79.96±0.20	77.09±0.13	49.33±0.12	91.52±0.17
BARLOW TWINS	87.39±0.14	62.28±0.13	79.16±0.13	75.94±0.08	48.39±0.14	91.89±0.21
FLATCLR	88.61±0.10	63.27±0.07	80.24±0.16	77.96±0.12	52.61±0.06	92.54±0.09
SPECTRAL CL	88.77±0.09	63.06±0.18	80.48±0.08	76.38±0.21	51.86±0.16	92.13±0.16
SOGCLR	88.93±0.11	63.14±0.12	80.54±0.14	77.70±0.07	52.35±0.08	92.60±0.08
VICREG	88.96±0.16	62.44±0.13	80.16±0.22	75.05±0.09	48.43±0.13	93.03±0.14
SIMCO	88.86±0.12	62.67±0.06	79.73±0.17	77.71±0.13	51.06±0.09	92.10±0.12
ISOGCLR	**89.24**±0.15	**63.82**±0.14	**81.14**±0.19	**78.37**±0.16	**53.06**±0.12	**93.08**±0.19

Table 2. Results on two bimodal downstream tasks. For image-text retrieval on Flickr30K and MSCOCO, we compute IR@1 and TR@1 for the Recall@1 on image-retrieval (IR) and text-retrieval (TR). For classification tasks, we compute top-1 accuracy (%). We report the average of scores and standard deviation over 3 runs with different random seeds. Full results are in Table 5, 6, and 7 in Appendix C.3.

METHOD	FLICKR30K RETRIEVAL		MSCOCO RETRIEVAL		ZERO-SHOT CLASSIFICATION TOP-1 ACC		
	IR@1	TR@1	IR@1	TR@1	CIFAR10	CIFAR100	IMAGENET1K
CLIP	40.98±0.22	50.90±0.17	21.32±0.12	26.98±0.21	60.63±0.19	30.70±0.11	36.27±0.17
CYCLIP	42.46±0.13	51.70±0.23	21.58±0.19	26.18±0.24	57.19±0.20	33.11±0.14	36.75±0.21
SOGCLR	43.32±0.18	57.18±0.20	22.43±0.13	30.08±0.22	**61.09**±0.24	33.26±0.12	37.46±0.19
ISOGCLR	**44.36**±0.12	**60.20**±0.26	**23.27**±0.18	**32.72**±0.13	58.91±0.15	**33.81**±0.18	**40.72**±0.23

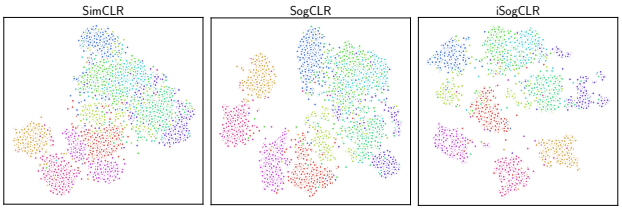

Figure 4. The arrangement of features (projected using t-SNE) for CIFAR10 samples learned by SimCLR, SogCLR and iSogCLR.

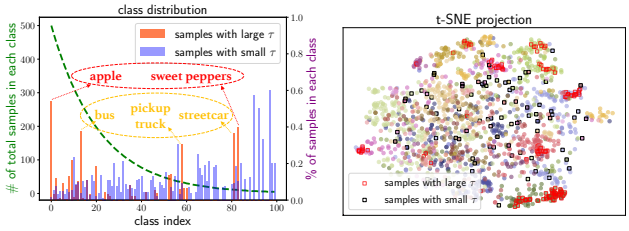

Figure 5. The class distributions and t-SNE projection for samples with large and small τ values in CIFAR100-LT. Left: The green dashed line and left axis denote the number of samples in each class, the red/blue bars and right axis denote the proportions of samples with large/small τ values in each class. Right: Each color represents a *superclass* in CIFAR100-LT.

0.01 and 0.005 by default. We evaluate models on two downstream tasks: cross-modal retrieval and image classification in zero-shot setting, following the widely-used evaluation protocol (Radford et al., 2021; Goel et al., 2022).

Results. We present partial results in Table 2 and full results in Table 5, 6, 7 in Appendix C.3. Compared with baselines, our algorithm achieves significant improvements on both downstream tasks. Specifically, iSogCLR improves CLIP by 4%∼17% and 2%∼8% on image-text retrieval and zero-short classification, respectively. Large-scale bimodal data always contain long-tail underlying semantics (Wang et al., 2022), thus the optimal τ of different samples may vary greatly. Hence iSogCLR with individualized temperatures is much more suitable than the methods with a global τ.

6.3. In-depth Analyses

Here, we demonstrate that iSogCLR indeed assigns suitable temperatures to samples with different types of semantics. Specifically, we consider the following two scenarios.

Unimodal data. We use CIFAR100-LT to study the characteristics of samples with different τ values. First, we select top-600 samples with large temperatures and bottom-600 samples with small temperatures. The *class distributions* of these two groups of samples are in the left of Figure 5. We observe that samples with small τ account for a higher proportion of tail classes. Interestingly, although some of samples belong to tail classes, e.g., 'sweet peppers', 'streetcar' and 'pickup truck', they are semantically similar to some head classes, e.g., 'apple', 'bus'. Thus these samples actually have frequent semantics and iSogCLR correctly assigns them large τ values. The right part of Figure 5 shows the projection of samples in these two groups. Note that most of samples with large τ values are in the centers of clusters, while most of samples with small τ values are separated from clusters. These results clearly show that iSogCLR makes samples with frequent semantics have large τ values to keep semantic structures, and makes samples with rare semantics have small τ values to be more discriminative.

Bimodal data. First, we use the data of "a kitten in a basket" and "architectural details of a bridge" for more illustrations. We show several hard negative pairs of the cat and bridge images in the left of Figure 6. One can observe that for the cat image, its hard negative pairs contain very similar semantics. For the bridge image, however, it has fewer hard negative pairs with similar semantics with it. We also present learned features of 1500 random image-text pairs with highlights

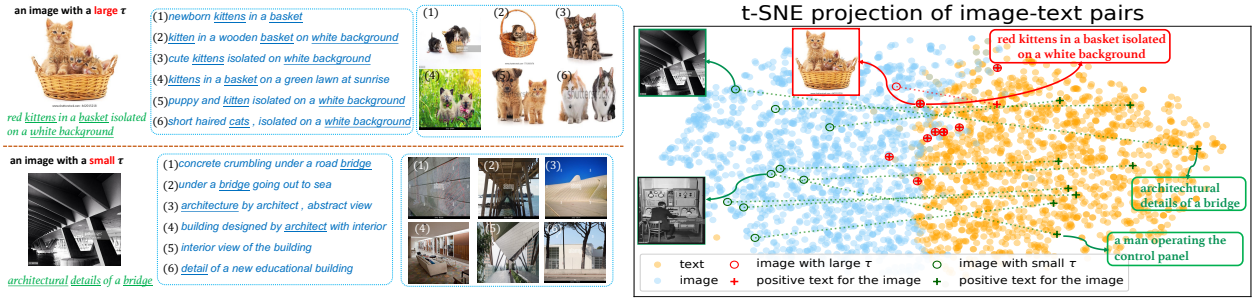

Figure 6. In-depth analyses on CC3M. Left: the contents of several hard negative image-text pairs of the cat and bridge images. Right: the tSNE of learned representations of sampled image-text pairs, with large and small temperatures marked by red and green, respectively.

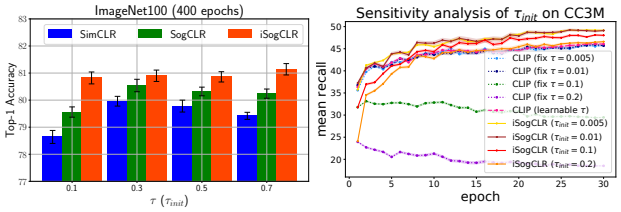

Figure 7. Effect of τ and τ_{init} on SimCLR/SogCLR and iSogCLR.

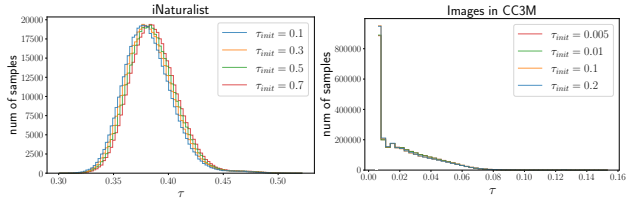

Figure 8. Final distributions of learned temperatures.

on several pairs with large and small τ values in Figure 6 (right). Notice that images with large τ values are very close to their texts, while images with small τ values are far from their texts. The reason is that pairs with large τ values have frequent semantics, thus the model learns their patterns well and their features are well aligned. By contrast, the pairs with small τ values have rare semantics and their features are not learned so well. These results show that iSogCLR learns suitable temperatures for samples with different semantics. More samples in CC3M data are provided in Figure 10 and 11 in Appendix C.3, showing that images with large τ values are related to frequent life scenes, while images with small τ values correspond to rare scenes.

6.4. Ablation Studies

In this section, we conduct extensive ablation studies to shed light on the behaviors of iSogCLR. First, we study the effect of different τ_{init} values on the final performance and convergence of iSogCLR, and provide the results in Figure 7. From the left part of Figure 7, one can observe that iSogCLR is not sensitive to τ_{init} and always outperforms SimCLR/SogCLR with a tuned τ. More results are in Table 8 in Appendix C.3. The results on CC3M in the right of Figure 7 indicate that CLIP fails to converge when τ is fixed to a large value (e.g., 0.1, 0.2). On the contrary, regardless of the values of τ_{init}, iSogCLR converges well and matches or outperforms CLIP with the tuned or learned τ.

We further visualize the final distributions of learned temperatures on different datasets in Figure 8. Note that these distributions are similar regardless of τ_{init} values. We also observe that the distributions on unimodal data usually follow a Gaussian distribution, while that on bimodal CC3M data has a long-tail (cf. Figure 9 in Appendix C.3 for more results). It is interesting to observe that the learned temperatures for bimodal dataset are smaller than those for image datasets, which is consistent with the literature found by manual tuning (Chen et al., 2020; Liang et al., 2022).

Due to the limited space, more studies are provided in Appendix C.3. In particular, the results about the effect of hyper-parameter ρ in Table 9 indicate that iSogCLR is not sensitive to ρ. We also compare with other heuristic baselines with individualized learnable temperatures in Appendix C.3, and show that our method is more advantageous for learning individualized temperatures (cf. Table 10).

7. Conclusion

In this work, we propose a novel method named iSogCLR for contrastive SSL with automatic temperature individualization. We first design a novel robust global contrastive objective based on DRO. Then we propose a provable stochastic algorithm. Theoretical and experimental results show that iSogCLR finds suitable temperatures for different samples. Comprehensive experiments demonstrate the effectiveness of iSogCLR on both unimodal and bimodal tasks.

Several directions can be considered in future work: i) academic researchers could consider how to further improve the convergence; ii) industrial researchers could consider applying our algorithm on large-scale datasets to potentially beat CLIP; iii) one can also consider how to leverage the learned temperatures for improving the performance of data with rare semantics.

Acknowledgements

We thank anonymous reviewers for constructive comments.

References

Aberdam, A., Litman, R., Tsiper, S., Anschel, O., Slossberg, R., Mazor, S., Manmatha, R., and Perona, P. Sequence-

to-sequence contrastive learning for text recognition. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15302–15312, 2021.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Bertsimas, D., Gupta, V., and Kallus, N. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1565–1576, 2019.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Chen, J., Gan, Z., Li, X., Guo, Q., Chen, L., Gao, S., Chung, T., Xu, Y., Zeng, B., Lu, W., et al. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce. *arXiv preprint arXiv:2107.01152*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the 32rd IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Dou, Z.-Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9694–9705, 2022.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the 34th IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3015–3024, 2021.

Eun, H., Moon, J., Park, J., Jung, C., and Kim, C. Learning to discriminate information for online action detection. In *Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 809–818, 2020.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Ge, S., Mishra, S., Li, C.-L., Wang, H., and Jacobs, D. Robust contrastive learning using negative samples with diminished semantics. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27356–27368, 2021.

Goel, S., Bansal, H., Bhatia, S., Rossi, R. A., Vinay, V., and Grover, A. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arXiv:2104.14840*, 2021.

Gürbüzbalaban, M., Ruszczyński, A., and Zhu, L. A stochastic subgradient method for distributionally robust non-convex and non-smooth learning. *Journal of Optimization Theory and Applications*, 194(3):1014–1041, 2022.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Horn, G. V., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. J. The inaturalist species classification and detection dataset. In *Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.-M., Fu, D., Shen, X., and Feng, J. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916, 2021.

Jin, J., Zhang, B., Wang, H., and Wang, L. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2771–2782, 2021.

Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., and Larlus, D. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21798–21809, 2020.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the 28th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Khaertdinov, B., Ghaleb, E., and Asteriadis, S. Contrastive self-supervised learning for sensor-based human activity recognition. In *2021 IEEE International Joint Conference on Biometrics*, pp. 1–8, 2021.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8847–8860, 2020.

Li, J., Zhou, P., Xiong, C., and Hoi, S. C. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021a.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021b.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of the 11th European Conference on Computer Vision*, pp. 740–755, 2014.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *Proceedings of the 19th European Conference on Computer Vision*, pp. 529–544, 2022.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2450–2504, 2017.

Nedić, A. and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Qi, Q., Xu, Y., Jin, R., Yin, W., and Yang, T. Attentional biased stochastic gradient for imbalanced classification. *arXiv preprint arXiv:2012.06951*, 2020.

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of areas under precision-recall curves with provable convergence. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1752–1765, 2021.

Qi, Q., Lyu, J., Bai, E. W., Yang, T., et al. Stochastic constrained dro with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Richemond, P. H., Grill, J.-B., Altché, F., Tallec, C., Strub, F., Brock, A., Smith, S., De, S., Pascanu, R., Piot, B., et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *the 9th International Conference on Learning Representations*, 2021.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

Sion, M. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Staib, M. and Jegelka, S. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9134–9144, 2019.

Tamkin, A., Wu, M., and Goodman, N. Viewmaker networks: Learning views for unsupervised representation learning. In *the 9th International Conference on Learning Representations*, 2021.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839, 2020.

Tomasev, N., Bica, I., McWilliams, B., Buesing, L., Pascanu, R., Blundell, C., and Mitrovic, J. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? *arXiv preprint arXiv:2201.05119*, 2022.

Wang, B. and Yang, T. Finite-sum compositional stochastic optimization: Theory and applications. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 23292–23317, 2022.

Wang, F. and Liu, H. Understanding the behaviour of contrastive loss. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2495–2504, 2021.

Wang, T., Jiang, W., Lu, Z., Zheng, F., Cheng, R., Yin, C., and Luo, P. Vlmixer: Unpaired vision-language pretraining via cross-modal cutmix. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 22680–22690, 2022.

Wang, X. and Qi, G.-J. Contrastive learning with stronger augmentations. *arXiv preprint arXiv:2104.07713*, 2021.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Wu, M., Mosse, M., Zhuang, C., Yamins, D., and Goodman, N. Conditional negative sampling for contrastive learning of visual representations. *arXiv preprint arXiv:2010.02037*, 2020.

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., and Fu, Y. Large scale incremental learning. In *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019.

Xia, J., Wu, L., Wang, G., Chen, J., and Li, S. Z. Progcl: Rethinking hard negative mining in graph contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 24332–24346, 2022.

Xie, J., Zhan, X., Liu, Z., Ong, Y.-S., and Loy, C. C. Delving into inter-image invariance for unsupervised visual representations. *International Journal of Computer Vision*, 130(12):2994–3013, 2022.

Xu, Y., Jin, R., and Yang, T. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2630–2640, 2019.

Yang, T. Algorithmic foundation of deep x-risk optimization. *CoRR*, abs/2206.00439, 2022. doi: 10.48550/arXiv.2206.00439. URL https://doi.org/10.48550/arXiv.2206.00439.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yuan, Z., Wu, Y., Qiu, Z.-H., Du, X., Zhang, L., Zhou, D., and Yang, T. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 25760–25782, 2022.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12310–12320, 2021.

Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D., and Kweon, I. S. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the 35th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14441–14450, 2022.

Zhang, O., Wu, M., Bayrooti, J., and Goodman, N. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*, 2021.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zhu, L., Gürbüzbalaban, M., and Ruszczyński, A. Distributionally robust learning with weakly convex losses: Convergence rates and finite-sample guarantees. *arXiv preprint arXiv:2301.06619*, 2023.

Zhu, X., Anguelov, D., and Ramanan, D. Capturing long-tail distributions of object subcategories. In *Proceedings of the 27th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 915–922, 2014.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Zhuang, C., Xiang, V., Bai, Y., Jia, X., Turk-Browne, N., Norman, K., DiCarlo, J. J., and Yamins, D. L. How well do unsupervised learning algorithms model human real-time and life-long learning? In *Proceedings of the Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

A. Derivation of the Equivalent Minimization Form

In this section, we present the detailed steps for the derivation of (7). Recall the problem:

$$\max_{\mathbf{p}\in\Delta}\min_{\lambda\geq 0}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j h_i(\mathbf{z}_j) - \tau_0 \mathrm{KL}(\mathbf{p},\mathbf{1}/m) - \lambda(\mathrm{KL}(\mathbf{p},\mathbf{1}/m)-\rho).$$

We first apply Sion's minimax theorem (Sion, 1958) and have:

$$\min_{\lambda\geq 0}\max_{\mathbf{p}\in\Delta}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j h_i(\mathbf{z}_j) - \tau_0 \mathrm{KL}(\mathbf{p},\mathbf{1}/m) - \lambda(\mathrm{KL}(\mathbf{p},\mathbf{1}/m)-\rho),$$

which is equivalent to

$$\min_{\lambda\geq 0}\max_{\mathbf{p}\in\Delta}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j h_i(\mathbf{z}_j) - (\lambda+\tau_0)(\mathrm{KL}(\mathbf{p},\mathbf{1}/m)-\rho) - \tau_0\rho.$$

Let $\tau = \lambda + \tau_0$, then we have

$$\min_{\tau\geq\tau_0}\max_{\mathbf{p}\in\Delta}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j h_i(\mathbf{z}_j) - \tau(\mathrm{KL}(\mathbf{p},\mathbf{1}/m)-\rho) - \tau_0\rho.$$

Then, the original problem is equivalent to the following problem:

$$\min_{\mathbf{w}}\min_{\tau\geq\tau_0}\max_{\mathbf{p}\in\Delta}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j h_i(\mathbf{z}_j) - \tau(\mathrm{KL}(\mathbf{p},\mathbf{1}/m)-\rho) - \tau_0\rho.$$

Next, we fix $\mathbf{x} = (\mathbf{w}^\top, \tau)^\top$ and derive the optimal solution $\mathbf{p}^*(\mathbf{x})$ that depends on $\mathbf{x}$ and solves the inner maximization problem. To this end, we consider the following problem

$$\min_{\mathbf{p}\in\Delta}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} -p_j h_i(\mathbf{z}_j) + \tau\mathrm{KL}(\mathbf{p},\mathbf{1}/m),$$

which has the same optimal solution as our original problem. There are actually three constraints to handle, i.e., $p_i \geq 0, \forall i$, $p_i \leq 1, \forall i$ and $\sum_{i=1}^m p_i = 1$. Note that the constraint $p_i \geq 0, \forall i$ is enforced by the term $p_i \log(p_i)$, otherwise the above objective will be infinity. Besides, the constraint $p_i \leq 1$ is automatically satisfied due to $\sum_{i=1}^m p_i = 1$ and $p_i \geq 0, \forall i$. Hence, we only to explicitly tackle the constraint $\sum_{i=1}^m p_i = 1$. To this end, we define the following Lagrangian function:

$$L_{\mathbf{x}}(\mathbf{p},\mu) = \sum_{\mathbf{z}_j\in\mathcal{S}_i^-} -p_j h_i(\mathbf{z}_j) + \tau\left(\log m + \sum_{i=1}^m p_i\log(p_i)\right) + \mu\left(\sum_{i=1}^m p_i - 1\right),$$

where $\mathrm{KL}(\mathbf{p},\mathbf{1}/m) = \log m + \sum_{i=1}^m p_i\log(p_i)$, and μ is the Lagrangian multiplier for the constraint $\sum_{i=1}^m p_i = 1$. The optimal solutions satisfy the KKT conditions:

$$-h_i(\mathbf{z}_j) + \tau(\log(p_j^*(\mathbf{x})) + 1) + \mu = 0 \quad \text{and} \quad \sum_{i=1}^m p_i^*(\mathbf{x}) = 1.$$

From the first equation, we can derive $p_j^*(\mathbf{x}) \propto \exp(h_i(\mathbf{z}_j)/\tau)$. Due to the second equation, we conclude that $p_j^*(\mathbf{x}) = \frac{\exp(h_i(\mathbf{z}_j)/\tau)}{\sum_{\mathbf{z}_j\in\mathcal{S}_i^-}\exp(h_i(\mathbf{z}_j)/\tau)}$. Plugging this optimal $\mathbf{p}^*$ into the inner maximization problem over $\mathbf{p}$, we have

$$\sum_{\mathbf{z}_j\in\mathcal{S}_i^-} p_j^*(\mathbf{x})h_i(\mathbf{z}_j) - \tau\left(\log m + \sum_{i=1}^m p_i^*(\mathbf{x})\log(p_i^*(\mathbf{x}))\right) = \tau\log\left(\frac{1}{m}\sum_{\mathbf{z}_j\in\mathcal{S}_i^-}\exp\left(\frac{h_i(\mathbf{z}_j)}{\tau}\right)\right) = \tau\log\left(\mathbb{E}_{\mathbf{z}_j\in\mathcal{S}_i^-}\exp\left(\frac{h_i(\mathbf{z}_j)}{\tau}\right)\right).$$

Therefore, we get the following equivalent problem:

$$\min_{\tau\geq\tau_0}\tau\log\left(\mathbb{E}_{\mathbf{z}_j\in\mathcal{S}_i^-}\exp\left(\frac{h_i(\mathbf{z}_j)}{\tau}\right)\right) + (\tau-\tau_0)\rho,$$

which is the dual form in (7) of the original RGCL. The dual form for RGCL in bimodal setting can be derived in a similar way.

B. iSogCLR for Bimodal CL Setting

Recall the RGCO for bimodal SSL:

$$\min_{\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}'\geq\tau_0} F_{\mathrm{B}}(\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}') := \frac{1}{n}\sum\nolimits_{(\mathbf{x}_i,\mathbf{t}_i)\in\mathcal{D}'}\left\{(\boldsymbol{\tau}_i+\boldsymbol{\tau}_i')\rho+\boldsymbol{\tau}_i\log\mathbb{E}_{\mathbf{t}\in\mathcal{T}_i^-}\exp\left(\frac{h_{\mathbf{x}_i}(\mathbf{t})}{\boldsymbol{\tau}_i}\right)+\boldsymbol{\tau}_i'\log\mathbb{E}_{\mathbf{x}\in\mathcal{I}_i^-}\exp\left(\frac{h_{\mathbf{t}_i}(\mathbf{x})}{\boldsymbol{\tau}_i'}\right)\right\},$$

where

$$h_{\mathbf{x}_i}(\mathbf{t}) = E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}) - E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i),$$
$$h_{\mathbf{t}_i}(\mathbf{x}) = E_I(\mathbf{x})^\top E_T(\mathbf{t}_i) - E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i).$$

It is worth to mention that an image-text pair can be viewed as *two views of the same underlying concept*. **So essentially, bimodal RGCO is consistent with unimodal RGCO because they all construct positive (resp. negative) pairs from the the different views of the same (resp. different) concepts, and pull close positive pairs and push away negative pairs. The only difference is the bimodal loss gets views from different modalities while the unimodal loss gets views from different augmentations**. Our algorithm is general for softmax-base contrastive loss and does not mind how to extract the views. Therefore it is applicable to both unimodal and bimodal CL.

The algorithm for optimizing $F_{\mathrm{B}}(\mathbf{w},\boldsymbol{\tau},\boldsymbol{\tau}')$ is very similar to that for optimizing unimodal RGCO $F(\mathbf{w},\boldsymbol{\tau})$ in Algorithm 1. Note that we employ the subscript 'v' and 't' to represent variables for visual images and texts, respectively. At each iteration, we sample a random mini-batch of B' image-text pairs $\mathcal{B}'=\{\mathbf{x}_1,\mathbf{t}_1,\ldots,\mathbf{x}_{B'},\mathbf{t}_{B'}\}$. Then we compute the stochastic estimators of $g_{\mathbf{x}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{v},i};\mathcal{T}_i')$ and $g_{\mathbf{t}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{t},i};\mathcal{I}_i')$ by

$$g_{\mathbf{x}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{v},i};\mathcal{T}_i') = \frac{1}{|\mathcal{T}_i'|}\sum_{\mathbf{t}\in\mathcal{T}_i'}\exp\left(\frac{h_{\mathbf{x}_i}(\mathbf{t})}{\boldsymbol{\tau}_{\mathrm{v},i}}\right), \tag{15}$$

$$g_{\mathbf{t}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{t},i};\mathcal{I}_i') = \frac{1}{|\mathcal{I}_i'|}\sum_{\mathbf{x}\in\mathcal{I}_i'}\exp\left(\frac{h_{\mathbf{t}_i}(\mathbf{x})}{\boldsymbol{\tau}_{\mathrm{t},i}}\right), \tag{16}$$

where $\mathcal{I}_i'=\{\mathbf{x}_1,\ldots,\mathbf{x}_{B'}\}\backslash\{\mathbf{x}_i\}$ and $\mathcal{T}_i'=\{\mathbf{t}_1,\ldots,\mathbf{t}_{B'}\}\backslash\{\mathbf{t}_i\}$. To control the approximation error, we maintain the following two moving average estimators:

$$\mathbf{s}_{\mathrm{v},i}^{t+1} = (1-\beta_0)\mathbf{s}_{\mathrm{v},i}^t + \beta_0 g_{\mathbf{x}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{v},i};\mathcal{T}_i'), \tag{17}$$
$$\mathbf{s}_{\mathrm{t},i}^{t+1} = (1-\beta_0)\mathbf{s}_{\mathrm{t},i}^t + \beta_0 g_{\mathbf{t}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{t},i};\mathcal{I}_i'). \tag{18}$$

where $\beta_0\in(0,1)$. With these estimators, we can compute the gradients of $F_B(\mathbf{w},\boldsymbol{\tau})$ w.r.t. $\mathbf{w}$, $\boldsymbol{\tau}_{\mathrm{v}}$, and $\boldsymbol{\tau}_{\mathrm{t}}$ by

$$G(\boldsymbol{\tau}_{\mathrm{v},i}^t) = \frac{1}{n}\left[\frac{\boldsymbol{\tau}_{\mathrm{v},i}^t}{\mathbf{s}_{\mathrm{v},i}^t}\nabla_{\boldsymbol{\tau}_{\mathrm{v},i}} g_{\mathbf{x}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{v},i};\mathcal{T}_i') + \log(\mathbf{s}_{\mathrm{v},i}^t) + \rho\right], \tag{19}$$

$$G(\boldsymbol{\tau}_{\mathrm{t},i}^t) = \frac{1}{n}\left[\frac{\boldsymbol{\tau}_{\mathrm{t},i}^t}{\mathbf{s}_{\mathrm{t},i}^t}\nabla_{\boldsymbol{\tau}_{\mathrm{t},i}} g_{\mathbf{t}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{t},i};\mathcal{I}_i') + \log(\mathbf{s}_{\mathrm{t},i}^t) + \rho\right], \tag{20}$$

$$G(\mathbf{w}_t) = \frac{1}{|\mathcal{B}'|}\sum_{\mathbf{x}_i,\mathbf{t}_i\in\mathcal{B}'}\left(\frac{\boldsymbol{\tau}_{\mathrm{v},i}^t}{\mathbf{s}_{\mathrm{v},i}^t}\nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{v},i};\mathcal{T}_i') + \frac{\boldsymbol{\tau}_{\mathrm{t},i}^t}{\mathbf{s}_{\mathrm{t},i}^t}\nabla_{\mathbf{w}} g_{\mathbf{t}_i}(\mathbf{w}_t,\boldsymbol{\tau}_{\mathrm{t},i};\mathcal{I}_i')\right). \tag{21}$$

We present the detailed steps of using the momentum-style update in Algorithm 2. A similar convergence guarantee to Theorem 1 can be established for iSogCLR in bimodal setting. The momentum-style update can be replaced by an Adam-style update using adaptive step sizes, and the same convergence rate can be established.

C. Experiments

C.1. Details of Implementation

For experiments on unimodal image datasets, we compare our algorithm, iSogCLR, against the following methods. **SimCLR** (Chen et al., 2020) is a pioneering work that directly optimize InfoNCE loss (Oord et al., 2018). **FlatCLR** (Chen

Algorithm 2 iSogCLR for Bimodal SSL

Require: β_0, β_1, η

Initialize $\mathbf{w}_1, \mathbf{s}_{\mathrm{v}}^1, \mathbf{s}_{\mathrm{t}}^1, \mathbf{u}_{\mathrm{v}}^1, \mathbf{u}_{\mathrm{t}}^1, \mathbf{v}_1, \boldsymbol{\tau}_{\mathrm{v}}^1 = \boldsymbol{\tau}_{\mathrm{t}}^1 = \boldsymbol{\tau}_{\mathrm{init}}$
for $t = 1, 2, \ldots, T$ **do**
  Draw a batch of B' samples denoted by $\mathcal{B}' \subset \mathcal{D}'$
  for $\mathbf{x}_i \in \mathcal{B}'$ **do**
    Compute $g_{\mathbf{x}_i}(\mathbf{w}_t, \boldsymbol{\tau}_{\mathrm{v},i}; \mathcal{T}_i')$ and $g_{\mathbf{t}_i}(\mathbf{w}_t, \boldsymbol{\tau}_{\mathrm{t},i}; \mathcal{I}_i')$ according to (15) and (16), respectively
    Update $\mathbf{s}_{\mathrm{v},i}^{t+1}$ and $\mathbf{s}_{\mathrm{t},i}^{t+1}$ according to (17) and (18), respectively
    Compute $G(\boldsymbol{\tau}_{\mathrm{v},i}^t)$ and $G(\boldsymbol{\tau}_{\mathrm{i},i}^t)$ according to (19) and (20), respectively
    Update $\mathbf{u}_{\mathrm{v},i}^{t+1} = (1-\beta_1)\mathbf{u}_{\mathrm{v},i}^t + \beta_1 G(\boldsymbol{\tau}_{\mathrm{v},i}^t)$ and $\mathbf{u}_{\mathrm{t},i}^{t+1} = (1-\beta_1)\mathbf{u}_{\mathrm{t},i}^t + \beta_1 G(\boldsymbol{\tau}_{\mathrm{t},i}^t)$
    Update $\boldsymbol{\tau}_{\mathrm{v},i}^{t+1} = \Pi_\Omega\left[\boldsymbol{\tau}_{\mathrm{v},i}^t - \eta \mathbf{u}_{\mathrm{v},i}^{t+1}\right]$ and $\boldsymbol{\tau}_{\mathrm{t},i}^{t+1} = \Pi_\Omega\left[\boldsymbol{\tau}_{\mathrm{t},i}^t - \eta \mathbf{u}_{\mathrm{t},i}^{t+1}\right]$
  end for
  Compute gradient estimator $G(\mathbf{w}_t)$ according to (21)
  Compute $\mathbf{v}_{t+1} = (1-\beta_1)\mathbf{v}_t + \beta_1 G(\mathbf{w}_t)$
  Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_{t+1}$ (or Adam-style)
end for

et al., 2021) employs a variant of InfoNCE loss for better performance in the small-batch-size regime. **Spectral CL** (HaoChen et al., 2021) is based on spectral decomposition on population graph and has provable accuracy guarantees. **SogCLR** (Yuan et al., 2022) utilizes variance reduction techniques to achieve promising performance and has provable convergence guarantees. **SimCo** (Zhang et al., 2022) improves negative mining in CL by using dual temperatures. **Barlow Twins** (Zbontar et al., 2021) and **VICReg** (Bardes et al., 2021) are non-contrastive methods and aim to maximize the information content of embeddings. On bimodal visual-language datasets, we consider the following baselines. **CLIP** (Radford et al., 2021) is one of the most popular VLP framework. **CyCLIP** (Goel et al., 2022) try to improve CLIP by optimizing the features to be geometrically consistent on image and text space. **SogCLR** can also be applied to solve bimodal SSL problems and is included in our comparison.

For unimodal experiments, we adopt a code base from GitHub[3] and implement the baseline methods in our experiments based on their open source implementations. The backbone networks we use are ResNet-18 and ResNet-50 for experiments on CIFAR dataset and ImageNet100/iNaturalist, respectively. For the projection head, we employ that used by VICReg (Bardes et al., 2021) for all methods. For bimodal experiments, we conduct experiments on the basis of ALBEF[4] (Li et al., 2021a). We also implement bimodal CL baselines, e.g., CLIP, CyCLIP, and SogCLR, in the code base. We adopt ResNet-50 as the image encoder and DistilBert (Sanh et al., 2019) as the text encoder. We train our models on Nvidia Tesla V100 GPU with 32GB memory and GTX 3090 GPU with 24GB memory.

C.2. Details of Datasets

CIFAR-10 and CIFAR-100 are two widely-used image datasets. Both of them contain 50,000 images for training and 10,000 images for test. The full version of ImageNet contains 1000 classes (about 1.2M images) and we denote it as ImageNet-1K (Russakovsky et al., 2015). ImageNet-100 (Wu et al., 2019) is a subset with randomly selected 100 classes (about 128K image) from ImageNet-1K. We also consider two imbalanced datasets: CIFAR100-LT and ImageNet-LT. We construct CIFAR100-LT following a widely-used strategy in the literature (Cao et al., 2019; Qi et al., 2022) with the imbalance ratio ρ=100, and keep the test set unchanged. The imbalance ratio ρ is defined as the ratio between sample sizes of the most frequent and least frequent classes. The LT imbalance follows the exponentially decayed sample size between different classes. The iNaturalist species classification and detection dataset (Horn et al., 2018) is a real-world large-scale dataset with 437,513 images from 8142 classes in its 2018 version.

Conceptual Captions 3M (CC3M) dataset (Sharma et al., 2018) contains about 2.9 million image-caption pairs crawled from the Internet. Note that as time goes by, some images are not available. Thus the number of image-caption pairs we use in our experiments is smaller than that in the original papers. Each image in MSCOCO and Flickr30K datasets has about 5 captions. MSCOCO dataset (Lin et al., 2014) contains 113K images and 567K captions, and Flickr30K dataset (Plummer et al., 2015) has 32K images and 158K captions. We employ the well-known Karpathy split (Karpathy & Fei-Fei, 2015) for

[3]https://github.com/HobbitLong/SupContrast
[4]https://github.com/salesforce/ALBEF

Table 3. Linear evaluation (top-1 accuracy (%)) under different training epochs on three balanced unimodal image datasets.

METHOD	CIFAR10		CIFAR100		IMAGENET100	
	400EP	800EP	400EP	800EP	200EP	400EP
SIMCLR	88.74±0.18	89.64±0.12	62.34±0.09	64.78±0.14	78.84±0.18	79.96±0.20
BARLOW TWINS	87.39±0.14	88.39±0.16	62.28±0.13	64.33±0.13	77.02±0.14	79.16±0.13
FLATCLR	88.61±0.10	89.22±0.06	63.27±0.07	64.51±0.08	79.06±0.09	80.24±0.16
SPECTRAL CL	88.77±0.09	**90.30**±0.11	63.06±0.18	64.32±0.17	78.38±0.17	80.48±0.08
SOGCLR	88.93±0.11	90.07±0.10	63.14±0.12	65.18±0.10	79.12±0.07	80.54±0.14
VICREG	88.96±0.16	89.90±0.12	62.44±0.13	64.18±0.09	**79.58**±0.23	80.16±0.22
SIMCO	88.86±0.12	89.79±0.15	62.67±0.06	64.74±0.12	77.36±0.16	79.73±0.17
ISOGCLR	**89.24**±0.15	90.25±0.09	**63.82**±0.14	**65.95**±0.07	79.42±0.15	**81.14**±0.19

Table 4. Linear evaluation (top-1 accuracy (%)) under different training epochs on three imbalanced unimodal image datasets.

METHOD	CIFAR10-LT		CIFAR100-LT		INATURALIST	
	400EP	800EP	400EP	800EP	200EP	400EP
SIMCLR	77.09±0.13	78.36±0.07	49.33±0.12	51.89±0.09	90.79±0.14	91.52±0.17
BARLOW TWINS	75.94±0.08	77.12±0.14	48.39±0.14	50.74±0.15	90.57±0.22	91.89±0.21
FLATCLR	77.96±0.12	79.19±0.08	52.61±0.06	54.14±0.08	91.48±0.15	92.54±0.09
SPECTRAL CL	76.38±0.21	78.63±0.13	51.86±0.16	53.46±0.17	91.28±0.11	92.13±0.16
SOGCLR	77.70±0.07	79.16±0.09	52.35±0.08	53.58±0.13	91.89±0.18	92.60±0.08
VICREG	75.05±0.09	77.84±0.15	48.43±0.13	51.68±0.06	92.18±0.06	93.03±0.14
SIMCO	77.71±0.13	78.56±0.19	51.06±0.09	52.31±0.14	91.03±0.18	92.10±0.12
ISOGCLR	**78.37**±0.16	**79.69**±0.08	**53.06**±0.12	**54.42**±0.18	**92.33**±0.23	**93.08**±0.19

these two datasets.

C.3. Additional Experimental Results

Unimodal experimental results. We present the full results on three balanced datasets and three imbalanced datasets in Table 3 and Table 4, respectively. One can observe than our iSogCLR matches or outperforms prior strong baselines.

Bimodal experimental results. We provide the full results of the zero-shot image-text retrieval tasks on Flickr30K and MSCOCO in Table 5 and Table 6, respectively. It is notable that our method has large improvements compared with baselines. We also present the full results of the zero-shot classification tasks on three standard image datasets in Table 7, and observe that our method achieves the best performance in most cases.

More ablation studies

Effect of τ_{init}. We present more ablation studies on the hyper-parameters of iSogCLR. In Table 8, we first present the effect of τ and τ_{init} on the performance of SimCLR and iSogCLR, respectively. One can observe that τ is an important hyper-parameter for SimCLR. SimCLR equiped with a tuned τ can be a strong baseline on many dataset. Besides, we find that our iSogCLR is not sensitive to τ_{init} in a range of 0.1$\sim$0.7. Moreover, iSogCLR with any τ_{init} is this range can outperforms SimCLR with a tuned τ. These results demonstrate the effectiveness of our method.

Effect of ρ. We provide the effect of ρ on the performance of iSogCLR in Table 9. We observe that although the parameter ρ in RGCL affects the degree of hardness-awareness, this parameter does not have a big impact on the performance of iSogCLR in most cases. We believe the reason is that we introduce a learnable Lagrangian multiplier λ for each KL constraint in our derivation. Thus the degree of hardness-awareness of each anchor data is largely affected by λ, i.e., the individualized temperature, which is flexible and updated during learning.

Effect of β_0. Another hyper-parameter in iSogCLR is the moving average parameter β_0 for updating $\mathbf{s}^{t+1}$ in (12). Following Yuan et al. (2022) (cf. Table 8 in their paper), we tune this parameter in a range of $\{0.7, 0.8, 0.9\}$. We find that when β_0 of iSogCLR is set in this range, the performance of the algorithm does not differ much in most cases.

Comparing with other baselines containing individualized learnable parameters.

Table 5. Zero-shot image-text retrieval (text-to-image and image-to-text) results (Recall@k), where $k \in \{1, 5, 10\}$, on Flickr30K dataset.

METHOD	IMAGE RETRIEVAL			TEXT RETRIEVAL		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	40.98±0.22	69.60±0.19	79.22±0.08	50.90±0.17	81.00±0.16	87.90±0.22
CYCLIP	42.46±0.13	69.56±0.16	78.74±0.21	51.70±0.23	79.90±0.18	88.40±0.11
SOGCLR	43.32±0.18	71.06±0.13	79.54±0.19	57.18±0.20	81.03±0.26	88.62±0.18
ISOGCLR	**44.36**±0.12	**72.64**±0.17	**80.92**±0.13	**60.20**±0.26	**84.60**±0.21	**90.50**±0.14

Table 6. Zero-shot image-text retrieval (text-to-image and image-to-text) results (Recall@k), where $k \in \{1, 5, 10\}$, on MSCOCO dataset.

METHOD	IMAGE RETRIEVAL			TEXT RETRIEVAL		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	21.32±0.12	45.52±0.17	57.30±0.16	26.98±0.21	54.86±0.15	66.86±0.19
CYCLIP	21.58±0.19	45.46±0.13	57.56±0.22	26.18±0.24	53.24±0.18	65.86±0.22
SOGCLR	22.43±0.13	46.74±0.11	58.32±0.20	30.08±0.22	56.94±0.17	67.39±0.24
ISOGCLR	**23.27**±0.18	**47.23**±0.24	**59.07**±0.19	**32.72**±0.13	**59.52**±0.11	**70.78**±0.21

Unimodal TaU+SimCLR. We first compare our method with TaU+SimCLR (Zhang et al., 2021), which adopts the framework of SimCLR and optimizes an input-dependent temperature as the uncertainty for the input. Specifically, for an input $\mathbf{x}$, Zhang et al. (2021) edit the encoder network to return $d+1$ entries, where the first d entries are the embedding of $\mathbf{x}$, and the last entry (let e denote its value) is used to compute a temperature for the input by $\frac{\text{sigmoid}(e)}{t}$ (t is a fixed hyper-parameter). We implement TaU+SimCLR following the pseudo code in the paper (Zhang et al., 2021), and present the results on CIFAR dataset in Table 10. One can observe that iSogCLR outperforms TaU+SimCLR by large margins. TaU+SimCLR learns input-dependent τ to estimate the uncertainty in out-of-distribution detection effectively, but with the cost of sacrificing the performance on downstream tasks.

Directly Optimizing CLIP with individualized temperatures. Besides, we also try to implement a variant of CLIP with individualized learnable temperatures. Similar to the CLIP with a global learnable temperature, we construct a learnable temperature for each image or text, compute the loss on each pair using their own temperatures, and optimize them by the automatic differentiation in PyTorch. We initialize all temperature parameters to 0.01. However, we observe that this variant is hard to converge. Specifically, **we observe that the average of learnable temperature parameters is getting larger and larger during training**. We believe the reason is this. Let us consider the ordinary bimodal contrastive loss on a image-text pair $(\mathbf{x}_i, \mathbf{t}_i)$:

$$\ell(\mathbf{x}_i, \mathbf{t}_i) = \log \sum_{\mathbf{t} \in \mathcal{T}_i^-} \exp\left(\frac{h_{\mathbf{x}_i}(\mathbf{t})}{\tau}\right) + \log \sum_{\mathbf{x} \in \mathcal{I}_i^-} \exp\left(\frac{h_{\mathbf{t}_i}(\mathbf{x})}{\tau}\right),$$

where $h_{\mathbf{x}_i}(\mathbf{t}) = E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}) - E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i)$ and $h_{\mathbf{t}_i}(\mathbf{x}) = E_I(\mathbf{x})^\top E_T(\mathbf{t}_i) - E_I(\mathbf{x}_i)^\top E_T(\mathbf{t}_i)$. If $\mathbf{x}_i$ are very similar to $\mathbf{t}_i$ (e.g., a pair with frequent semantics, or the encoders are good), then $h_{\mathbf{x}_i}(\mathbf{t})$ and $h_{\mathbf{t}_i}(\mathbf{x})$ are always negative. At this time, the larger the temperature, the smaller the loss function. Hence naively optimizing contrastive loss with individualized temperatures probably does not work.

More results of the distributions of learned temperatures. We present the final distributions of the learned temperatures with different τ_{init} values on all datasets in Figure 9. One can observe that the distributions for unimodal datasets are close to the Gaussian distribution. For CC3M dataset, we plot the distributions of learned temperatures of images and texts, respectively. We observe that these two distributions are very similar, and are close to the long-tail distribution with most samples have small temperatures.

The connection between the proposed RGCL and prior hard negative mining. Below, we compare our approach with (Robinson et al., 2021). First, our approach uses temperature individualization to achieve hard negative mining for data with rare semantics. In particular, our method assigns a low temperature for data with rare semantics data to select true hard negatives, and for data with frequent semantics, our method assigns the high temperature to avoid selecting false *negative* data. In comparison, Robinson et al. (2021) take a more direct approach to achieve the hard negative sampling by approximating the hard negative sampling distribution through unlabeled data and augmented data. The pros of (Robinson et al., 2021) is that it might achieve hard negative mining for both rare semantics and frequent semantics. The cons is that

Table 7. Zero-shot top-k classification accuracy (%), where $k \in \{1, 3, 5\}$.

METHOD	CIFAR10			CIFAR100		
	TOP-1	TOP-3	TOP-5	TOP-1	TOP-3	TOP-5
CLIP	60.63±0.19	87.29±0.12	**95.02**±0.16	30.70±0.11	49.49±0.13	58.51±0.14
CYCLIP	57.19±0.20	85.02±0.14	93.94±0.23	33.11±0.14	52.99±0.17	61.01±0.22
SOGCLR	**61.09**±0.24	**88.12**±0.19	94.92±0.18	33.26±0.12	**52.46**±0.22	60.71±0.15
ISOGCLR	58.91±0.15	86.27±0.24	93.43±0.11	**33.81**±0.18	53.21±0.21	**61.83**±0.19

METHOD	IMAGENET1K		
	TOP-1	TOP-3	TOP-5
CLIP	36.27±0.17	51.03±0.17	56.84±0.22
CYCLIP	36.75±0.21	51.32±0.18	57.08±0.23
SOGCLR	37.46±0.19	52.68±0.16	58.04±0.10
ISOGCLR	**40.72**±0.23	**54.38**±0.14	**59.11**±0.17

Table 8. The effect of τ (τ_{init}) to SimCLR (iSogCLR). We report top-1 accuracy after pretraining for 400 epochs.

METHOD	CIFAR10				CIFAR100				IMAGENET100			
	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
SIMCLR	85.85	88.34	88.74	88.27	60.49	62.34	62.02	61.73	78.64	79.96	79.78	79.42
ISOGCLR	**89.00**	**89.17**	**89.24**	**89.23**	**63.30**	**63.73**	**63.41**	**63.50**	**80.82**	**80.90**	**80.86**	**81.14**

their optimization might have large errors due to the Monte Carlo importance sampling based on mini-batch samples. The pros of our approach is that it enjoys optimization guarantee even with a small batch size, and it achieves hard negative mining for rare semantics. The cons of our approach is that it incurs additional cost of maintaining and updating the temperature parameters.

More examples from CC3M dataset. We present more images and texts with large and small learned temperatures in Figure 10 and 11, respectively. One can observe that the images with large temperatures contain frequent semantics like person, house, animals, flowers, and natural landscape. While for images with small temperatures, their semantics could be abstract or rare in daily life.

D. Convergence Analysis

We first introduce some notations. Let $||\cdot||$ denote the Euclidean norm of a vector. We denote the combination of $\mathbf{w}$ and $\boldsymbol{\tau}$, i.e., $(\mathbf{w}^\top, \boldsymbol{\tau}^\top)^\top \in \mathbb{R}^{d+n}$ by $\mathbf{z}$. Recall that $h_i(\mathbf{e}) = E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathbf{e}) - E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathcal{A}'(\mathbf{x}_i))$, where **we employ a new variable $\mathbf{e}$ in place of $\mathbf{z}$ used in (3) to avoid conflicts**.

To simplify the notations, we use $g_i(\mathbf{z})$ and $g_i(\mathbf{z}, \mathcal{B})$ to represent $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-)$ and $g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{B}_i)$, respectively. We can see that $\mathbb{E}_{\mathcal{B}}[g_i(\mathbf{z}, \mathcal{B})] = g_i(\mathbf{z})$. Then the objective (8) can be expressed as $F(\mathbf{z}) = F(\mathbf{w}, \boldsymbol{\tau}) = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} f_i(\boldsymbol{\tau}_i, g_i(\mathbf{z}))$. We denote the batch sizes $B = |\mathcal{B}|$ and $B' = |\mathcal{B}_i|$.

Then we make the following standard assumptions regarding to problem (8).

Assumption 1. *There exists $R, \sigma, C_g, C_f, L_f, L_g, C$ such that*

(i) The domain of model parameter $\mathbf{w} \in \mathcal{W}$ is bounded by R, i.e., for all $\mathbf{w} \in \mathcal{W}$, we have $||\mathbf{w}|| \leq R$.

(ii) $\mathbb{E}_{\mathcal{B}}[||g_i(\mathbf{z}) - g_i(\mathbf{z}, \mathcal{B})||^2] \leq \frac{\sigma^2}{B}$ and $\mathbb{E}_{\mathcal{B}}[||\nabla g_i(\mathbf{z}) - \nabla g_i(\mathbf{z}, \mathcal{B})||^2] \leq \frac{\sigma^2}{B}$.

(iii) Functions g_i and f_i satisfy $||\nabla g_i|| \leq C_g$ and $||\nabla f_i|| \leq C_f$ for all i.

(iv) Functions $\nabla f_i(\cdot)$, $\nabla g_i(\cdot)$ are L_f,L_g-Lipschitz continuous for all i.

(v) Functions $h_i(\mathbf{e})$ is bounded by C for all i, i.e., $|h_i(\mathbf{e})| \leq C$.

Table 9. Effect of ρ on iSogCLR (τ_{init} is set to 0.3). We report the average top-1 accuracies (%) for 400 epochs pretraining.

DATA \ ρ	0.1	0.2	0.3	0.4
CIFAR10	88.98	**89.03**	88.99	88.75
CIFAR100	63.02	63.12	63.27	**63.82**
IMAGENET100	80.70	**80.96**	80.54	80.18
CIFAR10-LT	77.86	78.05	78.31	**78.37**
CIFAR100-LT	52.60	52.75	52.92	**53.04**
INATURALIST	92.13	92.30	**92.79**	92.66

Table 10. Comparison between TaU+SimCLR and iSogCLR. We report the top-1 accuracies (%) after 400 epochs pretraining on CIFAR datasets.

METHOD	CIFAR10	CIFAR100	CIFAR10-LT	CIFAR100-LT
TAU+SIMCLR	86.80	59.35	76.41	49.62
ISOGCLR	89.24	63.82	78.37	53.06

Remark: Assumption $1(i)$ is also assumed by Levy et al. (2020) and Qi et al. (2022), and is mainly used for convex analysis. Assumption $1(ii)$ assumes that the stochastic estimators of $g_i(\mathbf{z})$ and $\nabla g_i(\mathbf{z})$ have bounded variance. Assumption $1(iii)$ and (iv) are also standard for convergence analysis. Note that $E\left(\mathcal{A}\left(\mathbf{x}_i\right)\right)$, $E\left(\mathcal{A}'\left(\mathbf{x}_i\right)\right)$ and $E\left(\mathbf{e}\right)$ are all *normalized* vectors, thus their inner products are bounded and Assumption $1(v)$ holds.

However, $F(\mathbf{w}, \boldsymbol{\tau})$ is not necessarily smooth in terms of $\mathbf{z} = (\mathbf{w}^\top, \boldsymbol{\tau}^\top)^\top$ if $\boldsymbol{\tau}$ is unbounded. To address this concern, we have the following lemma:

Lemma 1. *The optimal solution of $\boldsymbol{\tau}_i^*, i = 1, 2, \ldots, n$ to problem (8) is upper bounded by $\tilde{\tau} = \tau_0 + C/\rho$, where C is the upper bound for functions $h_i(\mathbf{e})$ and ρ is the constraint parameter.*

Proof. Recall the primal problem for each $\mathbf{x}_i \in \mathcal{D}$:

$$\mathbf{p}^* = \max_{\{\mathbf{p}\in\Delta, \mathrm{KL}(\mathbf{p}, \mathbf{1}/m)\leq\rho\}} \sum_{\mathbf{e}_j\in\mathcal{S}_i^-} \mathbf{p}_j h_i(\mathbf{e}_j) - \tau_0 \mathrm{KL}(\mathbf{p}, \mathbf{1}/m),$$

where $\mathbf{p}^*$ is the optimal value of the above problem.

Invoking dual variable $\bar{\lambda}_i$, we obtain the dual problem

$$\mathbf{q}^* = \min_{\bar{\lambda}\geq 0} \max_{\mathbf{p}\in\Delta} \sum_{\mathbf{e}_j\in\mathcal{S}_i^-} \mathbf{p}_j h_i(\mathbf{e}_j) - \tau_0 \mathrm{KL}(\mathbf{p}, \mathbf{1}/m) - \bar{\lambda}_i \left(\mathrm{KL}(\mathbf{p}, \mathbf{1}/m) - \rho\right).$$

Set $\bar{\mathbf{p}} = (1/m, \ldots, 1/m)$, a Slater vector satisfying $\mathrm{KL}(\bar{\mathbf{p}}, \mathbf{1}/m) - \rho \leq 0$. Applying Lemma 3 in (Nedić & Ozdaglar, 2009), we have

$$|\bar{\lambda}_i^*| \leq \frac{1}{\rho}\left(\mathbf{q}^* - \sum_{\mathbf{e}_j\in\mathcal{S}_i^-} \bar{\mathbf{p}}_j h_i(\mathbf{e}_j) - \tau_0 \mathrm{KL}(\bar{\mathbf{p}}, \mathbf{1}/m)\right).$$

Since the primal problem is concave in terms of $\mathbf{p}$, we have $\mathbf{p}^* = \mathbf{q}^*$. Therefore,

$$\begin{aligned} |\bar{\lambda}_i^*| &\leq \frac{1}{\rho}\left(\mathbf{p}^* - \sum_{\mathbf{e}_j\in\mathcal{S}_i^-} \bar{\mathbf{p}}_j h_i(\mathbf{e}_j)\right) \\ &\leq \frac{1}{\rho}\left(\sum_{\mathbf{e}_j\in\mathcal{S}_i^-} \mathbf{p}_j^* h_i(\mathbf{e}_j) - \tau_0 D(\mathbf{p}^*, \mathbf{1}/m) - \sum_{\mathbf{e}_j\in\mathcal{S}_i^-} \bar{\mathbf{p}}_j h_i(\mathbf{e}_j)\right) \\ &\leq \frac{C}{\rho}, \end{aligned} \tag{22}$$

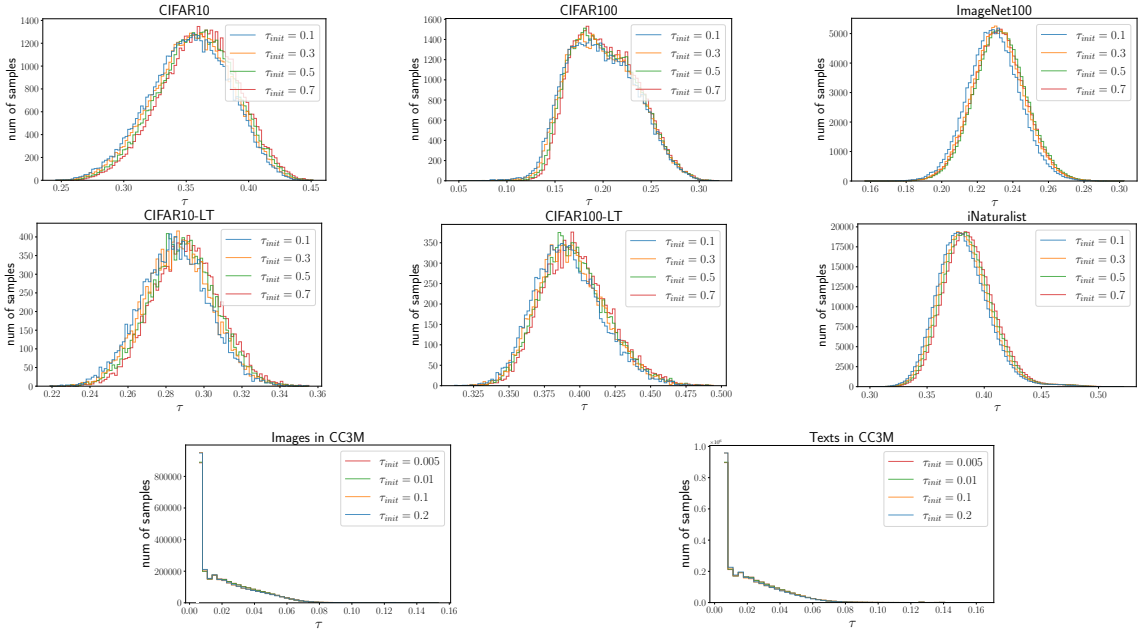

Figure 9. Distributions of the final learned temperatures with different τ_{init} values on seven different datasets.

where the last inequality is because $|h_i(\mathbf{e}_j)| \leq C$. Let $\boldsymbol{\tau}_i = \bar{\lambda}_i + \tau_0$, we have

$$\mathbf{q}^* = \min_{\tau \geq \tau_0} \max_{\mathbf{p} \in \Delta} \sum_{\mathbf{e}_j \in \mathcal{S}_i^-} \mathbf{p}_j h_i(\mathbf{e}_j) - \tau \left(\text{KL}(\mathbf{p}, \mathbf{1}/m) - \rho\right) - \tau_0 \rho.$$

By (22), we know that the optimal solution for above problem $|\boldsymbol{\tau}_i^*| \leq |\bar{\lambda}_i^*| + \tau_0 \leq \frac{C}{\rho} + \tau_0$, which completes the proof.

□

Due the boundness of functions $h_i(\mathbf{e})$ (cf. Assumption 1(v)) and $\boldsymbol{\tau}_i$ (cf. Lemma 1), we have the following lemma:

Lemma 2. *Functions $g_i(\mathbf{z}_t)$ and $g_i(\mathbf{z}_t, \mathcal{B})$ are lower bounded by $\hat{g} = \exp(-C/\tilde{\tau})$, where $-C$ is the lower bound for functions $h_i(\mathbf{e})$ and $\tilde{\tau}$ is the upper bound for $\boldsymbol{\tau}_i^*$.*

Proof. Recall the definitions of $h_i(\mathbf{e})$, $g_i(\mathbf{z}_t)$ and $g_i(\mathbf{z}_t, \mathcal{B})$:

$$h_i(\mathbf{e}) = E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathbf{e}) - E(\mathcal{A}(\mathbf{x}_i))^\top E(\mathcal{A}'(\mathbf{x}_i)),$$
$$g_i(\mathbf{z}) = g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{S}_i^-) = \frac{1}{|\mathcal{S}_i^-|} \sum_{\mathbf{e} \in \mathcal{S}_i^-} \exp\left(\frac{h_i(\mathbf{e})}{\boldsymbol{\tau}_i}\right),$$
$$g_i(\mathbf{z}, \mathcal{B}) = g_i(\mathbf{w}, \boldsymbol{\tau}_i; \mathcal{B}_i) = \frac{1}{\mathcal{B}_i} \sum_{\mathbf{e} \in \mathcal{B}_i} \exp\left(\frac{h_i(\mathbf{e})}{\boldsymbol{\tau}_i}\right).$$

Using $\boldsymbol{\tau}_i \leq \tilde{\tau}$ and $h_i(\mathbf{e}) \geq -C$, we have $g_i(\mathbf{z}) \geq \exp\left(\frac{-C}{\tilde{\tau}}\right)$. Similarly, we have $g_i(\mathbf{z}, \mathcal{B}) \geq \exp\left(\frac{-C}{\tilde{\tau}}\right)$, which completes the proof.

□

We will also see that the constraint on the domain of $\boldsymbol{\tau}$ guarantees the smoothness of $F(\mathbf{w}, \boldsymbol{\tau})$, which is critical for the proposed algorithm to enjoy fast convergence rate.

Lemma 3. *For all $\mathbf{w} \in \mathcal{W}$, $\boldsymbol{\tau}_i \in [\tau_0, \tilde{\tau}]$, and $i = 1, 2, \ldots, n$, $F_i(\mathbf{z}) = F_i(\mathbf{w}, \boldsymbol{\tau}_i) = f_i(\boldsymbol{\tau}_i, g_i(\mathbf{z}))$ is L_F-smooth for some constant L_F.*

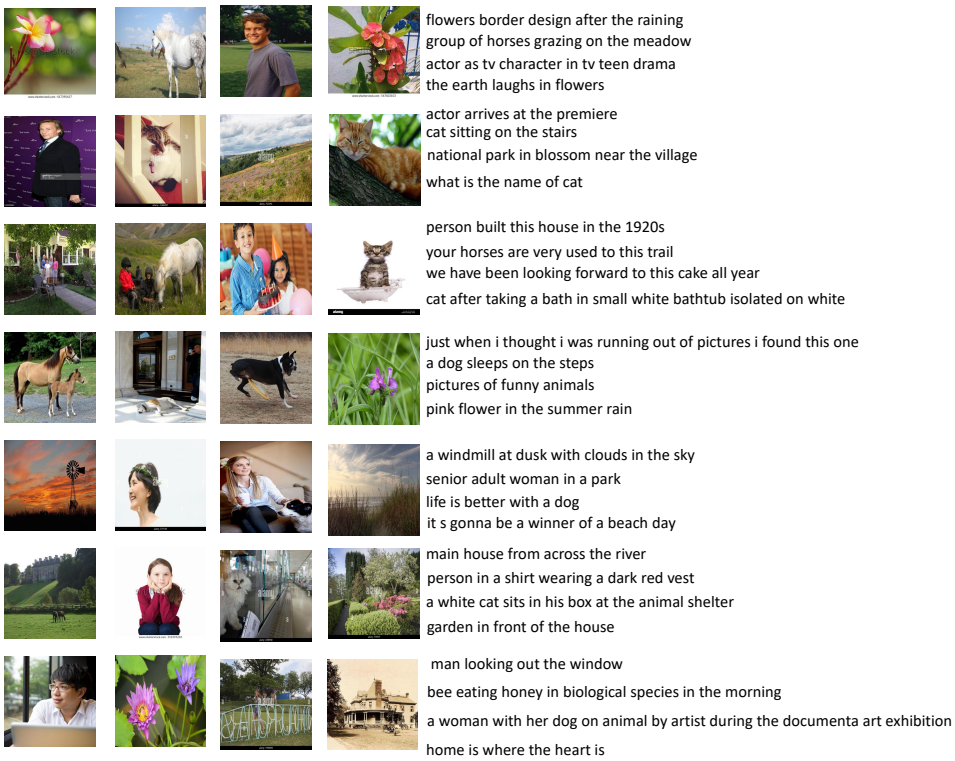

Figure 10. The images with large learned temperatures and their texts form CC3M. In general, they are very common in daily life, e.g., people, dogs, cats, flowers, houses, natural landscape, etc.

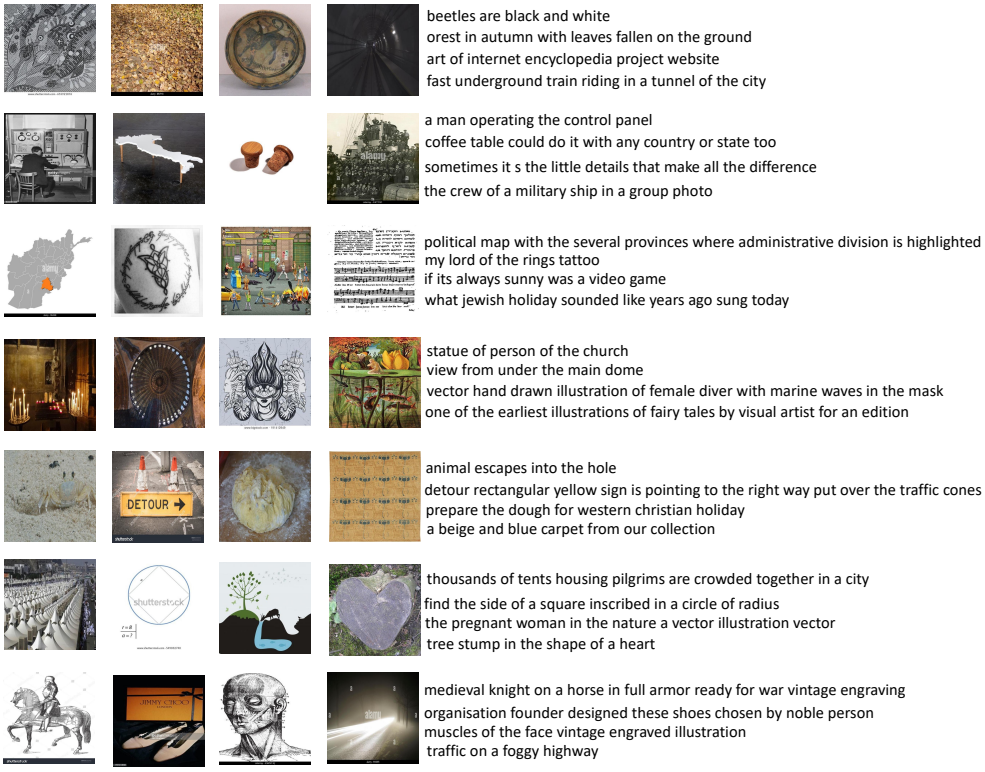

Figure 11. The images with small learned temperatures and their texts form CC3M. Most of them are not common in our lives or contain abstract concepts.

Note that Lemma 3 naturally follows that function $F(\mathbf{z})$ is also L_F-smooth.

Proof. We have gradients

$$\begin{aligned}
\nabla_{\mathbf{w}} F_i(\mathbf{w}, \boldsymbol{\tau}_i) &= \nabla_{\mathbf{w}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) \nabla_{g_i} f_i(\boldsymbol{\tau}_i, g_i(\mathbf{w}, \boldsymbol{\tau}_i)) \\
&= \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \nabla_{\mathbf{w}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) \\
\nabla_{\boldsymbol{\tau}} F_i(\mathbf{w}, \boldsymbol{\tau}_i) &= \nabla_{\boldsymbol{\tau}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) \nabla_{g_i} f_i(\boldsymbol{\tau}_i, g_i(\mathbf{w}, \boldsymbol{\tau}_i)) + \nabla_{\boldsymbol{\tau}} f_i(\boldsymbol{\tau}_i, g_i(\mathbf{w}, \boldsymbol{\tau}_i)) \\
&= \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \nabla_{\boldsymbol{\tau}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) + \nabla_{\boldsymbol{\tau}} f_i(\boldsymbol{\tau}_i, g_i(\mathbf{w}, \boldsymbol{\tau}_i)) \\
&= \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \begin{pmatrix} 0 \\ \vdots \\ \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{w}, \boldsymbol{\tau}_i) \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ \log(g_i(\mathbf{w}, \boldsymbol{\tau}_i)) + \rho \\ \vdots \\ 0 \end{pmatrix}
\end{aligned}$$

For any arbitrary $\mathbf{z}, \tilde{\mathbf{z}}$, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{z}} F_i(\mathbf{z}) - \nabla_{\mathbf{z}} F_i(\tilde{\mathbf{z}})\|^2 \\
&= \|\nabla_{\mathbf{w}} F_i(\mathbf{z}) - \nabla_{\mathbf{w}} F_i(\tilde{\mathbf{z}})\|^2 + \|\nabla_{\boldsymbol{\tau}} F_i(\mathbf{z}) - \nabla_{\boldsymbol{\tau}} F_i(\tilde{\mathbf{z}})\|^2 \\
&= \left\| \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \nabla_{\mathbf{w}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) - \frac{\tilde{\boldsymbol{\tau}}_i}{g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)} \nabla_{\mathbf{w}} g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i) \right\|^2 \\
&\quad + \left\| \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{w}, \boldsymbol{\tau}_i) + \log(g_i(\mathbf{w}, \boldsymbol{\tau}_i)) - \left(\frac{\tilde{\boldsymbol{\tau}}_i}{g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)} \nabla_{\boldsymbol{\tau}_i} g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i) + \log(g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)) \right) \right\|^2
\end{aligned}$$

Under assumption 1, we obtain

$$\begin{aligned}
&\left\| \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \nabla_{\mathbf{w}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) - \frac{\tilde{\boldsymbol{\tau}}_i}{g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)} \nabla_{\mathbf{w}} g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i) \right\|^2 \\
&\leq 2 \left\| \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \left[\nabla_{\mathbf{w}} g_i(\mathbf{w}, \boldsymbol{\tau}_i) - \nabla_{\mathbf{w}} g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i) \right] \right\|^2 + 2 \left\| \left[\frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} - \frac{\tilde{\boldsymbol{\tau}}_i}{g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)} \right] \nabla_{\mathbf{w}} g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i) \right\|^2 \\
&\leq \frac{2\tilde{\tau} L_g}{\hat{g}} (\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \|\boldsymbol{\tau}_i - \tilde{\boldsymbol{\tau}}_i\|^2) + \frac{2\tilde{\tau} C_g^2}{\hat{g}^2} (\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \|\boldsymbol{\tau}_i - \tilde{\boldsymbol{\tau}}_i\|^2)
\end{aligned}$$

and

$$\begin{aligned}
&\left\| \frac{\boldsymbol{\tau}_i}{g_i(\mathbf{w}, \boldsymbol{\tau}_i)} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{w}, \boldsymbol{\tau}_i) + \log(g_i(\mathbf{w}, \boldsymbol{\tau}_i)) - \left(\frac{\tilde{\boldsymbol{\tau}}_i}{g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)} \nabla_{\boldsymbol{\tau}_i} g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i) + \log(g_i(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\tau}}_i)) \right) \right\|^2 \\
&\leq 4 \left(\frac{C_g}{\hat{g}} + \frac{\tilde{\tau} C_g^2}{\hat{g}^2} + \frac{\tilde{\tau} L_g}{\hat{g}} + \frac{C_g}{\hat{g}} \right) (\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \|\boldsymbol{\tau}_i - \tilde{\boldsymbol{\tau}}_i\|^2)
\end{aligned}$$

Define $L_F = \frac{2\tilde{\tau} L_g}{\hat{g}} + \frac{2\tilde{\tau} C_g^2}{\hat{g}^2} + 4 \left(\frac{C_g}{\hat{g}} + \frac{\tilde{\tau} C_g^2}{\hat{g}^2} + \frac{\tilde{\tau} L_g}{\hat{g}} + \frac{C_g}{\hat{g}} \right)$, then $\|\nabla_{\mathbf{z}} F_i(\mathbf{z}) - \nabla_{\mathbf{z}} F_i(\tilde{\mathbf{z}})\|^2 \leq L_f \|\mathbf{z} - \tilde{\mathbf{z}}\|^2$. □

Below, we let $\chi = \{\mathbf{z} | \mathbf{w} \in \mathcal{W}, \tau_0 \leq \boldsymbol{\tau}_i \leq \tilde{\tau}, i = 1, 2, \ldots, n\}$. $\delta_\chi(\mathbf{z}) = 0$ if $\mathbf{z} \in \chi$, and $\delta_\chi(\mathbf{z}) = \infty$ if $\mathbf{z} \notin \chi$. Then problem (8) is equivalent to:

$$\min_{\mathbf{z} \in \mathbb{R}^{d+n}} \bar{F}(\mathbf{z}) := F(\mathbf{z}) + \delta_\chi(\mathbf{z}). \tag{23}$$

Now the update step of $\mathbf{z}_t$ can be written as $\mathbf{z}_{t+1} = \Pi_\chi(\mathbf{z}_t - \eta \mathbf{d}_{t+1})$, where Π_χ denotes the Euclidean projection onto the domain χ, and $\mathbf{d}_{t+1} = (\mathbf{v}_{t+1}^\top, \mathbf{u}^{t+1\top})^\top$.

Since $\bar{F}$ is non-smooth, we define the regular subgradients as follows.

Definition 1 (Regular Subgradient)**.** *Consider a function $\Phi : \mathbb{R}^n \to \bar{\mathbb{R}}$ and $\Phi(\bar{\mathbf{x}})$ is finite. For a vector $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}$ is a regular subgradient of Φ at $\bar{\mathbf{x}}$, written $\mathbf{v} \in \hat{\partial}\Phi(\bar{\mathbf{x}})$, if*

$$\liminf_{\mathbf{x} \to \bar{\mathbf{x}}} \frac{\Phi(\mathbf{x}) - \Phi(\bar{\mathbf{x}}) - \mathbf{v}^\top(\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \geq 0.$$

Since $F(\mathbf{z})$ is differentiable, we use $\hat{\partial}\bar{F}(\mathbf{z}) = \nabla F(\mathbf{z}) + \hat{\partial}\delta_\chi(\mathbf{z})$ (see Exercise 8.8 in Rockafellar & Wets (2009)) in the analysis. The $\text{dist}(0, \hat{\partial}\bar{F}(\mathbf{z}))$ measures the distance between the origin and the regular subgradient set of $\bar{F}$ at $\mathbf{z}$. The oracle complexity is defined below:

Definition 2 (Oracle Complexity)**.** *Let $\epsilon > 0$ be a small constant, the oracle complexity is defined as the number of processing samples in order to achieve $\mathbb{E}[dist(0, \hat{\partial}\bar{F}(\mathbf{z}))] \leq \epsilon$ for a non-convex loss function or $\mathbb{E}[F(\mathbf{z}) - F(\mathbf{z}_*)] \leq \epsilon$ for a convex loss function.*

To prove the main theorem, we present some required lemmas.

Lemma 4. *Under Assumption 1, run Algorithm 1 with $\eta L_F \leq \frac{1}{4}$, and the output $\mathbf{z}_R$ of Algorithm 1 satisfies*

$$\mathbb{E}[dist(0, \hat{\partial}\bar{F}(\mathbf{z}_R))] \leq \frac{2 + 40L_F\eta}{T} \sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 + \frac{2\Delta}{\eta T} + \frac{40L_F\Delta}{T},$$

where $\Delta := \bar{F}(\mathbf{z}_1) - \inf_{\mathbf{z}\in\chi} \bar{F}(\mathbf{z})$.

Proof. Recall the update of $\mathbf{z}_{t+1}$ is

$$\begin{aligned}\mathbf{z}_{t+1} &= \Pi_\chi(\mathbf{z}_t - \eta\mathbf{d}_{t+1}) \\ &= \arg\min_{\mathbf{z}\in\mathbb{R}^{d+n}} \left\{ \delta_\chi(\mathbf{z}) + \langle \mathbf{d}_{t+1}, \mathbf{z} - \mathbf{z}_t \rangle + \frac{1}{2\eta}||\mathbf{z} - \mathbf{z}_t||^2 \right\}.\end{aligned}$$

Then by Exercise 8.8 and Theorem 10.1 of Rockafellar & Wets (2009), we know

$$-\mathbf{d}_{t+1} - \frac{1}{\eta}(\mathbf{z}_{t+1} - \mathbf{z}_t) \in \hat{\partial}\delta_\chi(\mathbf{z}_{t+1}),$$

which implies that

$$\nabla F(\mathbf{z}_{t+1}) - \mathbf{d}_{t+1} - \frac{1}{\eta}(\mathbf{z}_{t+1} - \mathbf{z}_t) \in \nabla F(\mathbf{z}_{t+1}) + \hat{\partial}\delta_\chi(\mathbf{z}_{t+1}) = \hat{\partial}\bar{F}(\mathbf{z}_{t+1}). \tag{24}$$

By the update of $\mathbf{z}_{t+1}$, we also have

$$\delta_\chi(\mathbf{z}_{t+1}) + \langle \mathbf{d}_{t+1}, \mathbf{z}_{t+1} - \mathbf{z}_t \rangle + \frac{1}{2\eta}||\mathbf{z}_{t+1} - \mathbf{z}_t||^2 \leq \delta_\chi(\mathbf{z}_t).$$

Since $F(\mathbf{z})$ is L_F-smooth, we have

$$F(\mathbf{z}_{t+1}) \leq F(\mathbf{z}_t) + \langle \nabla F(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle + \frac{L_F}{2}||\mathbf{z}_{t+1} - \mathbf{z}_t||^2.$$

Combining the above two inequalities, we obtain

$$\langle \mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle + \frac{1}{2}\left(\frac{1}{\eta} - L_F\right)||\mathbf{z}_{t+1} - \mathbf{z}_t||^2 \leq \bar{F}(\mathbf{z}_t) - \bar{F}(\mathbf{z}_{t+1}).$$

Thus we have

$$\frac{1}{2}\left(\frac{1}{\eta} - L_F\right)||\mathbf{z}_{t+1} - \mathbf{z}_t||^2 \leq \bar{F}(\mathbf{z}_t) - \bar{F}(\mathbf{z}_{t+1}) - \langle \mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle,$$

where the last inequality uses $\langle \mathbf{a}, \mathbf{b} \rangle \leq ||\mathbf{a}||^2 + \frac{||\mathbf{b}||^2}{4}$. Then by rearranging the above inequality and summing it across $t = 1, 2, \ldots, T$, we have

$$\begin{aligned}\sum_{t=1}^{T} \frac{1-2\eta L_F}{4\eta} ||\mathbf{z}_{t+1} - \mathbf{z}_t||^2 &\leq \bar{F}(\mathbf{z}_1) - \bar{F}(\mathbf{z}_{T+1}) + \sum_{t=1}^{T} \eta ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 \\ &\leq \bar{F}(\mathbf{z}_1) - \inf_{\mathbf{z} \in \chi} \bar{F}(\mathbf{z}) + \sum_{t=1}^{T} \eta ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 \\ &= \Delta + \sum_{t=1}^{T} \eta ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2\end{aligned} \tag{25}$$

Using the same method in the proof of Theorem 2 in (Xu et al., 2019), we obtain the following relationship:

$$\begin{aligned}\sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_{t+1}) + \frac{1}{\eta}(\mathbf{z}_{t+1} - \mathbf{z}_t)||^2 \leq 2 \sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 + \frac{2\Delta}{\eta} \\ + \left(2L_F^2 + \frac{3L_F}{\eta}\right) \sum_{t=1}^{T} ||\mathbf{z}_{t+1} - \mathbf{z}_t||^2\end{aligned} \tag{26}$$

Recalling $\eta L_F \leq \frac{1}{4}$ and combining (25) and (26), we have

$$\begin{aligned}&\sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_{t+1}) + \frac{1}{\eta}(\mathbf{z}_{t+1} - \mathbf{z}_t)||^2 \\ &\overset{(a)}{\leq} 2 \sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 + \frac{2\Delta}{\eta} + \frac{5L_F}{\eta} \left(\frac{4}{1-2\eta L_F}\right) \left(\eta\Delta + \sum_{t=1}^{T} \eta^2 ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2\right) \\ &\overset{(b)}{\leq} 2 \sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 + \frac{2\Delta}{\eta} + 40 L_F \Delta + 40 \eta L_F \sum_{t=1}^{T} ||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2,\end{aligned} \tag{27}$$

where (a) is due to $(2L_F^2 + \frac{3L_F}{\eta}) \leq \frac{5L_F}{\eta}$ and (b) is due to $\frac{4}{1-2\eta L_F} \leq 8$.

Recalling (24) and the output rule of Algorithm 1, we have

$$\mathbb{E}[\text{dist}(0, \hat{\partial}\bar{F}(\mathbf{z}_R))^2] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_{t+1}) + \frac{1}{\eta}(\mathbf{z}_{t+1} - \mathbf{z}_t)||^2]. \tag{28}$$

At last, we combine (27) and (28) and have

$$\mathbb{E}[\text{dist}(0, \hat{\partial}\bar{F}(\mathbf{z}_R))^2] \leq \frac{2 + 40\eta L_F}{T} \sum_{t=1}^{T} \mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2] + \frac{2\Delta}{T\eta} + \frac{40 L_F \Delta}{T}. \tag{29}$$

□

Lemma 5. *Under Assumption 1, run Algorithm 1 and we have*

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2] \leq \Delta_{\mathbf{v}} + \Delta_{\mathbf{u}} + \left(\frac{4L_F^2}{\beta_1^2} + \frac{72 n^3 L_F^2}{B^2 \beta^2}\right) \sum_{t=1}^{T} \mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \\ + C_1 \sum_{t=1}^{T} \mathbb{E}[||g(\mathbf{z}_t) - \mathbf{s}^{t+1}||^2] + \frac{C_2 \beta_1}{B} T + \frac{C_3 \beta}{B} T,\end{aligned}$$

where $\Delta_{\mathbf{v}}, \Delta_{\mathbf{u}}, C_1, C_2, C_3$ are constants defined in the proof.

Proof. Recalling $\mathbf{d}_{t+1} = (\mathbf{v}_{t+1}^\top, \mathbf{u}^{t+1\top})^\top$ and $\nabla F(\mathbf{z}_t) = (\nabla_{\mathbf{w}} F(\mathbf{z}_t), \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t))^\top$, we have

$$||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2 = ||\mathbf{v}_{t+1} - \nabla_{\mathbf{w}} F(\mathbf{z}_t)||^2 + ||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2$$

We first establish the bound for $||\mathbf{v}_{t+1} - \nabla_{\mathbf{w}} F(\mathbf{z}_t)||^2$. Recall the define the following notations

$$\begin{aligned}
&\nabla F(\mathbf{z}_t) = \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{S}} \nabla_{\mathbf{w}} f_i(g_i(\mathbf{z}_t)) \nabla_{\mathbf{w}} g_i(\mathbf{z}_t), \\
&\nabla F(\mathbf{z}_t, \mathbf{s}^t) = \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{S}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t) \nabla_{\mathbf{w}} g_i(\mathbf{z}_t), \\
&\mathbf{v}_{t+1} = (1-\beta_1)\mathbf{v}_t + \beta_1 G(\mathbf{w}_t), \\
&G(\mathbf{w}_t) = \frac{1}{B} \sum_{\mathbf{x}_i \in \mathcal{B}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t) \nabla_{\mathbf{w}} g_i(\mathbf{z}_t, \mathcal{B}).
\end{aligned}$$

By expansion, we have

$$\begin{aligned}
&\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \mathbf{v}_{t+1}||^2] \\
=&\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - (1-\beta_1)\mathbf{v}_t - \beta_1 G(\mathbf{w}_t)||^2] \\
=&\mathbb{E}_t[||(1-\beta_1)(\nabla_{\mathbf{w}} F(\mathbf{z}_{t-1}) - \mathbf{v}_t) + (1-\beta_1)(\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_{t-1})) \\
&\quad + \beta_1(\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t)) + \beta_1(\nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t) - G(\mathbf{w}_t))||^2] \\
\overset{(a)}{=}&||(1-\beta_1)(\nabla_{\mathbf{w}} F(\mathbf{z}_{t-1}) - \mathbf{v}_t) + (1-\beta_1)(\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_{t-1})) \\
&\quad + \beta_1(\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t))||^2 + \beta_1^2 \mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t) - G(\mathbf{w}_t)||^2] \\
\overset{(b)}{\leq}&(1+\beta_1)(1-\beta_1)^2 ||\nabla_{\mathbf{w}} F(\mathbf{z}_{t-1}) - \mathbf{v}_t||^2 \\
&\quad + 2\left(1 + \frac{1}{\beta_1}\right)\left[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_{t-1})||^2 + \beta_1^2 ||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t)||^2\right] \\
&\quad + \beta_1^2 \mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t) - G(\mathbf{w}_t)||^2] \\
\overset{(c)}{\leq}&(1-\beta_1)||\nabla_{\mathbf{w}} F(\mathbf{z}_{t-1}) - \mathbf{v}_t||^2 + \frac{4L_F^2}{\beta_1}||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 + 4\beta_1 ||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t)||^2 \\
&\quad + \beta_1^2 \mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t) - G(\mathbf{w}_t)||^2],
\end{aligned} \tag{30}$$

where (a) is due to $\mathbb{E}_t[G(\mathbf{w}_t)] = \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t)$, (b) is due to Young's inequality $||\mathbf{a}+\mathbf{b}||^2 \leq (1+\gamma)||\mathbf{a}||^2 + (1+\frac{1}{\gamma})||\mathbf{b}||^2$, and (c) is due to $\beta_1 \leq 1 \to 1 + \frac{1}{\beta_1} \leq \frac{2}{\beta_1}$.

Furthermore, one may bound $\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t)||^2]$ as follows:

$$\begin{aligned}
&\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^{t+1})||^2] \\
=&\mathbb{E}_t\left[\left\|\frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} \nabla_{\mathbf{w}} f_i(g_i(\mathbf{z}_t)) \nabla_{\mathbf{w}} g_i(\mathbf{z}_t) - \frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t) \nabla_{\mathbf{w}} g_i(\mathbf{z}_t)\right\|^2\right] \\
\leq&\frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} C_g^2 L_f^2 \mathbb{E}_t[||g_i(\mathbf{z}_t) - \mathbf{s}_i^t||^2] \\
=&\frac{C_g^2 L_f^2}{n} \mathbb{E}_t[||g(\mathbf{z}_t) - \mathbf{s}^t||^2].
\end{aligned} \tag{31}$$

On the other hand, $\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t) - G(\mathbf{w}_t)||^2]$ can be bounded by some constants:

$$\begin{aligned}
&\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t, \mathbf{s}^t) - G(\mathbf{w}_t)||^2] \\
=&\mathbb{E}_t\left[\left\|\frac{1}{n}\sum_{\mathbf{x}_i\in\mathcal{D}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t)\nabla_{\mathbf{w}} g_i(\mathbf{z}_t) - \frac{1}{B}\sum_{\mathbf{x}_i\in\mathcal{B}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t)\nabla_{\mathbf{w}} g_i(\mathbf{z}_t, \mathcal{B})\right\|^2\right] \\
\leq&\mathbb{E}_t\left[2\left\|\frac{1}{n}\sum_{\mathbf{x}_i\in\mathcal{D}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t)\nabla_{\mathbf{w}} g_i(\mathbf{z}_t) - \frac{1}{B}\sum_{\mathbf{x}_i\in\mathcal{B}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t)\nabla_{\mathbf{w}} g_i(\mathbf{z}_t)\right\|^2\right. \\
&\left.2\left\|\frac{1}{B}\sum_{\mathbf{x}_i\in\mathcal{B}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t)\nabla_{\mathbf{w}} g_i(\mathbf{z}_t) - \frac{1}{B}\sum_{\mathbf{x}_i\in\mathcal{B}} \nabla_{\mathbf{w}} f_i(\mathbf{s}_i^t)\nabla_{\mathbf{w}} g_i(\mathbf{z}_t, \mathcal{B})\right\|^2\right] \\
\leq&\frac{2C_f^2C_g^2}{B} + \frac{2C_f^2\sigma^2}{B'}.
\end{aligned} \tag{32}$$

Substituting (31) and (32) into (30), we have

$$\begin{aligned}
\mathbb{E}_t[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \mathbf{v}_{t+1}||^2] \leq (1-\beta_1)||\nabla_{\mathbf{w}} F(\mathbf{z}_{t-1}) - \mathbf{v}_t||^2 + \frac{4L_F^2}{\beta_1}\mathbb{E}_t[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \\
+ \frac{4\beta_1 C_g^2 L_f^2}{n}\mathbb{E}_t[||g(\mathbf{z}_t) - \mathbf{s}^t||^2] + \frac{2\beta_1^2 C_f^2(C_g^2+\sigma^2)}{\min\{B, B'\}}.
\end{aligned} \tag{33}$$

Taking summation over $t = 1, 2, \ldots, T$, we obtain

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}[||\nabla_{\mathbf{w}} F(\mathbf{z}_t) - \mathbf{v}_{t+1}||^2] \leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{4L_F^2}{\beta_1^2}\sum_{t=1}^T \mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \\
+ \frac{4C_g^2L_f^2}{n}\sum_{t=1}^T \mathbb{E}[||g(\mathbf{z}_t) - \mathbf{s}^t||^2] + \frac{2\beta_1 C_f^2(C_g^2+\sigma^2)}{\min\{B, B'\}}T,
\end{aligned} \tag{34}$$

where $\Delta_{\mathbf{v}}$ denotes $||\nabla_{\mathbf{w}} F(\mathbf{z}_0) - \mathbf{v}_1||^2$.

Next, we derive the bound for $||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2$. Note that

$$||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2 = \sum_{\mathbf{x}_i\in\mathcal{D}} ||\mathbf{u}_i^{t+1} - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t)||^2 = \sum_{\mathbf{x}_i\in\mathcal{D}} \left\|\mathbf{u}_i^{t+1} - \frac{1}{n}\nabla_{\boldsymbol{\tau}_i} F_i(\mathbf{z}_t)\right\|^2$$

Recall and define the following notations

$$\begin{aligned}
\mathbf{u}_i^{t+1} = \begin{cases}(1-\beta)\mathbf{u}_i^t + \beta G(\boldsymbol{\tau}_i^t) & \text{if } \mathbf{x}_i \in \mathcal{B} \\ \mathbf{u}_i^t & \text{o.w.}\end{cases}, \quad \tilde{\mathbf{u}}_i^t := (1-\beta)\mathbf{u}_i^t + \beta G(\boldsymbol{\tau}_i^t), \mathbf{x}_i \in \mathcal{B}, \\
\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) = \frac{1}{n}\left(\frac{\boldsymbol{\tau}_i^t}{g_i(\mathbf{z}_t)}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) + \log(g_i(\mathbf{z}_t)) + \rho\right), \\
\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) = \frac{1}{n}\left(\frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) + \log(\mathbf{s}_i^t) + \rho\right), \\
G(\boldsymbol{\tau}_i^t) = \frac{1}{n}\left(\frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t, \mathcal{B}) + \log(\mathbf{s}_i^t) + \rho\right).
\end{aligned}$$

Then we obtain

$$
\begin{aligned}
&||\tilde{\mathbf{u}}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})||^2 \\
=&||(1-\beta)\mathbf{u}_i^t + \beta G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})||^2 \\
=&||(1-\beta)(\mathbf{u}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) + (1-\beta)(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) \\
&\quad + \beta(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t)) + \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))||^2 \\
\overset{(a)}{=}&||(1-\beta)(\mathbf{u}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) + (1-\beta)(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) \\
&\quad + \beta(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t))||^2 + \beta^2||(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t))||^2 \\
&\quad + 2\langle(1-\beta)(\mathbf{u}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) + (1-\beta)(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) \\
&\quad + \beta(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t)), \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))\rangle \\
\overset{(b)}{\leq}&(1+\beta)(1-\beta)^2||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1}) - \mathbf{u}_i^t||^2 \\
&\quad + 2\left(1+\frac{1}{\beta}\right)\left[||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})||^2 + \beta^2||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))||^2\right] \\
&\quad + \beta^2||(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^{t+1}) - G(\boldsymbol{\tau}_i^t))||^2 + 2\langle(1-\beta)(\mathbf{u}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) \\
&\quad + (1-\beta)(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) + \beta(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t)), \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))\rangle \\
\overset{(c)}{\leq}&(1-\beta)||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1}) - \mathbf{u}_i^t||^2 + \frac{4L_F^2}{n^2\beta}||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 + 4\beta||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))||^2 \\
&\quad + \beta^2||(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t))||^2 + 2\langle(1-\beta)(\mathbf{u}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) \\
&\quad + (1-\beta)(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})) + \beta(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t)), \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))\rangle,
\end{aligned}
\tag{35}
$$

where (b) is due to Young's inequality $||\mathbf{a}+\mathbf{b}||^2 \leq (1+\gamma)||\mathbf{a}||^2 + (1+\frac{1}{\gamma})||\mathbf{b}||^2$, and (c) is due to $\beta \leq 1 \rightarrow 1+\frac{1}{\beta} \leq \frac{2}{\beta}$. For simplicity, we denote the first term in the last inner product as A_i^t and note that A_i^t does not depend on the randomness of iteration t.

Subsequently, we derive the bound for $||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))||^2$

$$
\begin{aligned}
&||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^{t+1}))||^2 \\
=&\left\|\frac{1}{n}\left(\frac{\boldsymbol{\tau}_i^t}{g_i(\mathbf{z}_t)}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) + \log(g_i(\mathbf{z}_t))\right) - \frac{1}{n}\left(\frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) + \log(\mathbf{s}_i^t)\right)\right\|^2 \\
\leq&\frac{2}{n^2}\left\|\frac{\boldsymbol{\tau}_i^t}{g_i(\mathbf{z}_t)}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) - \frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t}\nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t)\right\|^2 + \frac{2}{n^2}\left\|\log(g_i(\mathbf{z}_t)) - \log(\mathbf{s}_i^t)\right\|^2 \\
\leq&\frac{2(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{\hat{g}^4 n^2}||\mathbf{s}_i^t - g_i(\mathbf{z}_t)||^2,
\end{aligned}
\tag{36}
$$

where $\tilde{\tau}$ denotes the upper bound for $\boldsymbol{\tau}_i$ and $\hat{g}$ denotes the lower bound for g_i.

Substituting (36) into (35), we have

$$
\begin{aligned}
&||\tilde{\mathbf{u}}_i^t - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})||^2 \\
\leq&(1-\beta)||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1}) - \mathbf{u}_i^t||^2 + \frac{4L_F^2}{n^2\beta}||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 + \frac{8\beta(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{\hat{g}^4 n^2}||\mathbf{s}_i^t - g_i(\mathbf{z}_t)||^2 \\
&\quad + \beta^2||(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t))||^2 + 2\langle A_i^t, \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t))\rangle
\end{aligned}
\tag{37}
$$

$$
\begin{aligned}
&\mathbb{E}_t[||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1})||^2] \\
&= \mathbb{E}_t \left[\sum_{x_i \in \mathcal{B}} \|\mathbf{u}_i^{t+1} - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})\|^2 + \sum_{x_i \notin \mathcal{B}} \|\mathbf{u}_i^{t} - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})\|^2 \right] \\
&= \mathbb{E}_t \left[\sum_{x_i \in \mathcal{B}} \|\tilde{\mathbf{u}}_i^{t} - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1})\|^2 \right] + \frac{n-B}{n} \|\mathbf{u}^t - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1})\|^2 \\
&\leq \mathbb{E}_t \Bigg[\sum_{x_i \in \mathcal{B}} (1-\beta) ||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_{t-1}) - \mathbf{u}_i^t||^2 + \frac{4L_F^2}{n^2 \beta} ||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 + \frac{8\beta(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{\hat{g}^4 n^2} ||\mathbf{s}_i^t - g_i(\mathbf{z}_t)||^2 \\
&\quad + \beta^2 ||(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t))||^2 + 2\langle A_i^t, \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t)) \rangle \Bigg] + \frac{n-B}{n} \|\mathbf{u}^t - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1})\|^2 \\
&\leq \frac{B}{n}(1-\beta) ||\nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1}) - \mathbf{u}^t||^2 + \frac{4BL_F^2}{n^2 \beta} ||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 + \frac{8B\beta(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{\hat{g}^4 n^3} ||\mathbf{s}^t - g(\mathbf{z}_t)||^2 \\
&\quad + \beta^2 \mathbb{E}_t \left[\sum_{x_i \in \mathcal{B}} ||(\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t))||^2 \right] + 2\mathbb{E}_t \left[\sum_{x_i \in \mathcal{B}} \langle A_i^t, \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t)) \rangle \right] \\
&\quad + \frac{n-B}{n} \|\mathbf{u}^t - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1})\|^2 \\
&\overset{(a)}{\leq} (1 - \frac{B\beta}{n}) ||\nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1}) - \mathbf{u}^t||^2 + \frac{4BL_F^2}{n^2 \beta} ||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 + \frac{8B\beta(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{\hat{g}^4 n^3} ||\mathbf{s}^t - g(\mathbf{z}_t)||^2 \\
&\quad + \frac{\beta^2 B \tilde{\tau}^2 \sigma^2}{\hat{g}^2 B' n^2}
\end{aligned}
$$

where the last inequality uses the following facts

$$
\begin{aligned}
\mathbb{E}_t \left[\sum_{\mathbf{x}_i \in \mathcal{B}} ||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t)||^2 \right] &= \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathcal{B} \in \overline{\mathcal{B}}} \sum_{\mathbf{x}_i \in \mathcal{B}} ||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t)||^2 \\
&= \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathbf{x}_i \in \mathcal{D}} |\overline{\mathcal{B}}_i| \frac{1}{|\overline{\mathcal{B}}_i|} \sum_{\mathcal{B} \in \overline{\mathcal{B}}_i} ||\nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t) - G(\boldsymbol{\tau}_i^t)||^2 \\
&\leq \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathbf{x}_i \in \mathcal{D}} |\overline{\mathcal{B}}_i| \frac{1}{|\overline{\mathcal{B}}_i|} \sum_{\mathcal{B} \in \overline{\mathcal{B}}_i} \left\| \frac{1}{n} \frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) - \frac{1}{n} \frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t, \mathcal{B}) \right\|^2 \\
&\leq \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathbf{x}_i \in \mathcal{D}} |\overline{\mathcal{B}}_i| \frac{\tilde{\tau}^2 \sigma^2}{\hat{g}^2 B' n^2} = \frac{B \tilde{\tau}^2 \sigma^2}{\hat{g}^2 B' n^2}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_t \left[\sum_{\mathbf{x}_i \in \mathcal{B}} \langle A_i^t, \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t)) \rangle \right] \\
&= \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathcal{B} \in \overline{\mathcal{B}}} \sum_{\mathbf{x}_i \in \mathcal{B}} \langle A_i^t, \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t)) \rangle \\
&= \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathbf{x}_i \in \mathcal{D}} |\overline{\mathcal{B}}_i| \frac{1}{|\overline{\mathcal{B}}_i|} \sum_{\mathcal{B} \in \overline{\mathcal{B}}_i} \langle A_i^t, \beta(G(\boldsymbol{\tau}_i^t) - \nabla_{\boldsymbol{\tau}_i} F(\mathbf{z}_t, \mathbf{s}_i^t)) \rangle \\
&= \frac{1}{|\overline{\mathcal{B}}|} \sum_{\mathbf{x}_i \in \mathcal{D}} |\overline{\mathcal{B}}_i| \frac{1}{|\overline{\mathcal{B}}_i|} \sum_{\mathcal{B} \in \overline{\mathcal{B}}_i} \left\langle A_i^t, \beta \left(\frac{1}{n} \frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t) - \frac{1}{n} \frac{\boldsymbol{\tau}_i^t}{\mathbf{s}_i^t} \nabla_{\boldsymbol{\tau}_i} g_i(\mathbf{z}_t, \mathcal{B}) \right) \right\rangle = 0
\end{aligned}
$$

where $\overline{\mathcal{B}}$ denotes the set of all possible batch $\mathcal{B} \subset \mathcal{D}$ of size B, and $\overline{\mathcal{B}}_i$ denotes $\{\mathcal{B} : \mathbf{x}_i \in \mathcal{B}, \mathcal{B} \in \overline{\mathcal{B}}\}$.

Furthermore,

$$\begin{aligned}
\mathbb{E}_t[||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2] &\overset{(a)}{\leq} \left(1 + \frac{B\beta}{2n}\right) \mathbb{E}_t[||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1})||^2] + \left(1 + \frac{2n}{B\beta}\right) ||\nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1}) - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2 \\
&\overset{(b)}{\leq} \left(1 - \frac{B\beta}{2n}\right) ||\nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1}) - \mathbf{u}^t||^2 + \frac{8BL_F^2}{n^2\beta} ||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 \\
&\quad + \frac{16B\beta(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{n^3 \hat{g}^4} ||\mathbf{s}^t - g(\mathbf{z}_t)||^2 + \frac{2B\tilde{\tau}^2\sigma^2\beta^2}{n^2\hat{g}^2 B'} + \frac{4L_F^2}{B\beta} ||\mathbf{z}_t - \mathbf{z}_{t-1}||^2 \\
&\overset{(c)}{\leq} \left(1 - \frac{B\beta}{2n}\right) ||\nabla_{\boldsymbol{\tau}} F(\mathbf{z}_{t-1}) - \mathbf{u}^t||^2 + \frac{2B\tilde{\tau}^2\sigma^2\beta^2}{n^2\hat{g}^2 B'} \\
&\quad + \frac{16B\beta(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{n^3 \hat{g}^4} ||\mathbf{s}^t - g(\mathbf{z}_t)||^2 + \frac{36L_F^2}{B\beta} ||\mathbf{z}_t - \mathbf{z}_{t-1}||^2,
\end{aligned}$$

where we use Young's inequality in (a), and use the assumption $\frac{B\beta}{2n} \leq 1$ in (b) and (c).

Taking summation over $t = 1, 2, \ldots, T$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2] &\leq \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \frac{72nL_F^2}{B^2\beta^2} \sum_{t=1}^{T} \mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \\
&\quad + \frac{32(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{n^2 \hat{g}^4} \sum_{t=1}^{T} \mathbb{E}[||\mathbf{s}^t - g(\mathbf{z}_t)||^2] + \frac{4\tilde{\tau}^2\sigma^2\beta}{nB'\hat{g}^2} T,
\end{aligned} \tag{38}$$

where $\Delta_{\mathbf{u}}$ denotes $||\nabla_{\boldsymbol{\tau}} F(\mathbf{z}_0) - \mathbf{u}^1||^2$.

At last, we combine (34) and (38), and establish the following inequality:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2] = \sum_{t=1}^{T} \mathbb{E}[||\mathbf{v}_{t+1} - \nabla_{\mathbf{w}} F(\mathbf{z}_t)||^2] + \sum_{t=1}^{T} \mathbb{E}[||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}} F(\mathbf{z}_t)||^2] \\
&\leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \left(\frac{4L_F^2}{\beta_1^2} + \frac{72nL_F^2}{B^2\beta^2}\right) \sum_{t=1}^{T} \mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \\
&+ \left(\frac{4C_g^2 L_f^2}{n} + \frac{32(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{n^2\hat{g}^4}\right) \sum_{t=1}^{T} \mathbb{E}[||g(\mathbf{z}_t) - \mathbf{s}^t||^2] + \frac{2\beta_1 C_f^2 (C_g^2 + \sigma^2)}{\min\{B, B'\}} T + \frac{4\tilde{\tau}^2\sigma^2\beta}{nB'\hat{g}^2} T \\
&\leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \left(\frac{4L_F^2}{\beta_1^2} + \frac{72nL_F^2}{B^2\beta^2}\right) \sum_{t=1}^{T} \mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \\
&+ \frac{C_1}{n} \sum_{t=1}^{T} \mathbb{E}[||g(\mathbf{z}_t) - \mathbf{s}^t||^2] + \frac{C_2\beta_1}{\min\{B, B'\}} T + \frac{C_3\beta}{nB'} T,
\end{aligned}$$

where $C_1 = \left(4C_g^2 L_f^2 + \frac{32(\tilde{\tau}^2 C_g^2 + \hat{g}^2)}{\hat{g}^4}\right)$, $C_2 = 2C_f^2(C_g^2 + \sigma^2)$ and $C_3 = \frac{4\tilde{\tau}^2\sigma^2}{\hat{g}^2}$.

□

Lemma 6. *Under Assumption (1), run Algorithm 1 and we have*

$$\sum_{t=1}^{T} \mathbb{E}[||\mathbf{s}^t - g(\mathbf{z}_t)||^2] \leq \frac{2n}{B\beta}\Delta_{\mathbf{s}} + \frac{8n^3 C_g^2}{B^2\beta^2} \sum_{t=1}^{T} \mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] + \frac{4n\beta\sigma^2 T}{B'}.$$

where $\Delta_{\mathbf{s}}$ is a constant defined in the proof.

Proof. Recall and define the following notations:

$$\mathbf{s}_i^{t+1} = \begin{cases} (1-\beta)\mathbf{s}_i^t + \beta g_i(\mathbf{z}_t, \mathcal{B}) & \text{if } \mathbf{x}_i \in \mathcal{B} \\ \mathbf{s}_i^t & \text{o.w.} \end{cases}, \quad \tilde{\mathbf{s}}_i^t := (1-\beta)\mathbf{s}_i^t + \beta g_i(\mathbf{z}_t, \mathcal{B}), \mathbf{x}_i \in \mathcal{B}.$$

Then we obtain

$$
\begin{aligned}
&||\tilde{\mathbf{s}}_i^t - g_i(\mathbf{z}_t)||^2 \\
=&||(1-\beta)\mathbf{s}_i^t + \beta g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t)||^2 \\
=&||(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)) + \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))||^2 \\
=&(1-\beta)^2\|(\mathbf{s}_i^t - g_i(\mathbf{z}_t))\|^2 + \beta^2||(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))||^2 + 2\langle(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)), \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))\rangle.
\end{aligned}
$$

Considering the randomness of iteration t, we have

$$
\begin{aligned}
&\mathbb{E}_t[||\mathbf{s}^{t+1} - g(\mathbf{z}_t)||^2] \\
&= \mathbb{E}_t\left[\sum_{x_i\in\mathcal{B}} ||\mathbf{s}_i^{t+1} - g_i(\mathbf{z}_t)||^2 + \sum_{x_i\notin\mathcal{B}} ||\mathbf{s}_i^t - g_i(\mathbf{z}_t)||^2\right] \\
&= \mathbb{E}_t\left[\sum_{x_i\in\mathcal{B}} ||\tilde{\mathbf{s}}_i^t - g_i(\mathbf{z}_t)||^2\right] + \frac{n-B}{n}||\mathbf{s}^t - g(\mathbf{z}_t)||^2 \\
&= \mathbb{E}_t\Bigg[\sum_{x_i\in\mathcal{B}} (1-\beta)^2\|(\mathbf{s}_i^t - g_i(\mathbf{z}_t))\|^2 + \beta^2||g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t)||^2 \\
&\quad + 2\langle(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)), \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))\rangle\Bigg] + \frac{n-B}{n}||\mathbf{s}^t - g(\mathbf{z}_t)||^2 \\
&= \frac{B}{n}(1-\beta)||\mathbf{s}^t - g(\mathbf{z}_t)||^2 + \mathbb{E}_t\left[\sum_{x_i\in\mathcal{B}} \beta^2||g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t)||^2\right] \\
&\quad + \mathbb{E}_t\left[\sum_{x_i\in\mathcal{B}} 2\langle(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)), \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))\rangle\right] + \frac{n-B}{n}||\mathbf{s}^t - g(\mathbf{z}_t)||^2 \\
&\overset{(a)}{\leq} (1 - \frac{B\beta}{n})||\mathbf{s}^t - g(\mathbf{z}_t)||^2 + \frac{B\beta^2\sigma^2}{B'}
\end{aligned}
$$

where (a) uses the following facts

$$
\begin{aligned}
\mathbb{E}_t\left[\sum_{x_i\in\mathcal{B}} \beta^2||g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t)||^2\right] &= \frac{1}{|\overline{\mathcal{B}}|}\sum_{\mathcal{B}\in\overline{\mathcal{B}}}\sum_{x_i\in\mathcal{B}} \beta^2||g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t)||^2 \\
&= \frac{1}{|\overline{\mathcal{B}}|}\sum_{x_i\in\mathcal{D}} |\overline{\mathcal{B}}_i|\frac{1}{|\overline{\mathcal{B}}_i|}\sum_{\mathcal{B}\in\overline{\mathcal{B}}_i} \beta^2||g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t)||^2 \\
&\leq \frac{1}{|\overline{\mathcal{B}}|}\sum_{x_i\in\mathcal{D}} |\overline{\mathcal{B}}_i|\frac{\beta^2\sigma^2}{B'} = \frac{B\beta^2\sigma^2}{B'}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_t\left[\sum_{x_i\in\mathcal{B}} 2\langle(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)), \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))\rangle\right] \\
&= \frac{1}{|\overline{\mathcal{B}}|}\sum_{\mathcal{B}\in\overline{\mathcal{B}}}\sum_{x_i\in\mathcal{B}} 2\langle(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)), \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))\rangle \\
&= \frac{1}{|\overline{\mathcal{B}}|}\sum_{x_i\in\mathcal{D}} |\overline{\mathcal{B}}_i|\frac{1}{|\overline{\mathcal{B}}_i|}\sum_{\mathcal{B}\in\overline{\mathcal{B}}_i} 2\langle(1-\beta)(\mathbf{s}_i^t - g_i(\mathbf{z}_t)), \beta(g_i(\mathbf{z}_t, \mathcal{B}) - g_i(\mathbf{z}_t))\rangle = 0
\end{aligned}
$$

where $\overline{\mathcal{B}}$ denotes the set of all possible batch $\mathcal{B} \subset \mathcal{D}$ of size B, and $\overline{\mathcal{B}}_i$ denotes $\{\mathcal{B} : x_i \in \mathcal{B}, \mathcal{B} \in \overline{\mathcal{B}}\}$.

Furthermore, we use Young's inequality and derive the following relationship:

$$\mathbb{E}_t[||\mathbf{s}^{t+1} - g(\mathbf{z}_{t+1})||^2] \leq \left(1 + \frac{B\beta}{2n}\right)\mathbb{E}_t[||\mathbf{s}^{t+1} - g(\mathbf{z}_t)||^2] + \left(1 + \frac{2n}{B\beta}\right)\mathbb{E}_t[||g(\mathbf{z}_t) - g(\mathbf{z}_{t+1})||^2]$$
$$\overset{(a)}{\leq} \left(1 - \frac{B\beta}{2n}\right)||\mathbf{s}^t - g(\mathbf{z}_t)||^2 + \frac{2B\beta^2\sigma^2}{B'} + \frac{4n^2C_g^2}{B\beta}\mathbb{E}_t[||\mathbf{z}_t - \mathbf{z}_{t+1}||^2],$$

where (a) is due to $\frac{B\beta}{2n} \leq 1$.

Taking expectation over all randomness and taking summation over all $\mathbf{x}_i \in \mathcal{S}$ and $t = 1, 2, \ldots, T$, we obtain

$$\sum_{t=1}^{T}\mathbb{E}[||\mathbf{s}^t - g(\mathbf{z}_t)||^2] \leq \frac{2n}{B\beta}\Delta_{\mathbf{s}} + \frac{8n^3C_g^2}{B^2\beta^2}\sum_{t=1}^{T}\mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] + \frac{4n\beta\sigma^2T}{B'}.$$

where $\Delta_{\mathbf{s}}$ denotes $||\mathbf{s}^0 - g(\mathbf{z}_0)||^2$.

□

Now we present the proof for the convergence guarantee of Algorithm 1.

Proof. Now we present the proof for Theorem 1. First of all, we establish the following relationship using (25)

$$\sum_{t=1}^{T}\mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] \leq 8\eta\Delta + 8\eta^2\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2], \tag{39}$$

where we use $\eta L_F \leq \frac{1}{4}$.

On the other hand, we combine Lemma (5) and Lemma (6), and obtain

$$\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2] = \sum_{t=1}^{T}\mathbb{E}[||\mathbf{v}_{t+1} - \nabla_{\mathbf{w}}F(\mathbf{z}_t)||^2] + \sum_{t=1}^{T}\mathbb{E}[||\mathbf{u}^{t+1} - \nabla_{\boldsymbol{\tau}}F(\mathbf{z}_t)||^2]$$
$$\leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \left(\frac{4L_F^2}{\beta_1^2} + \frac{72nL_F^2}{B^2\beta^2}\right)\sum_{t=1}^{T}\mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2]$$
$$+ \frac{C_1}{n}\left[\frac{2n}{B\beta}\Delta_{\mathbf{s}} + \frac{8n^3C_g^2}{B^2\beta^2}\sum_{t=1}^{T}\mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2] + \frac{4n\beta\sigma^2T}{B}\right] + \frac{C_2\beta_1}{\min\{B, B'\}}T + \frac{C_3\beta}{nB'}T$$
$$\leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \frac{2C_1}{B\beta}\Delta_{\mathbf{s}} + \left(\frac{4L_F^2}{\beta_1^2} + \frac{72nL_F^2}{B^2\beta^2} + \frac{8C_1n^2C_g^2}{B^2\beta^2}\right)\sum_{t=1}^{T}\mathbb{E}[||\mathbf{z}_t - \mathbf{z}_{t-1}||^2]$$
$$+ \frac{4C_1\beta\sigma^2T}{B'} + \frac{C_2\beta_1}{\min\{B, B'\}}T + \frac{C_3\beta}{nB'}T$$
$$\leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \frac{2C_1}{B\beta}\Delta_{\mathbf{s}} + \frac{4C_1\beta\sigma^2T}{B'} + \frac{C_2\beta_1}{\min\{B, B'\}}T + \frac{C_3\beta}{nB'}T$$
$$+ \left(\frac{4L_F^2}{\beta_1^2} + \frac{72nL_F^2}{B^2\beta^2} + \frac{8C_1n^2C_g^2}{B^2\beta^2}\right)\left[8\eta\Delta + 8\eta^2\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2]\right]$$

By setting $\eta^2 \leq \min\left\{\frac{\beta_1^2}{192L_F^2}, \frac{B^2\beta^2}{3456nL_F^2}, \frac{B^2\beta^2}{384C_1n^2C_g^2}\right\} = O(\min\{\beta_1^2, \frac{B^2\beta^2}{n^2}\})$, we have

$$\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2] \leq \frac{1}{\beta_1}\Delta_{\mathbf{v}} + \frac{2n}{B\beta}\Delta_{\mathbf{u}} + \frac{2C_1}{B\beta}\Delta_{\mathbf{s}} + \frac{4C_1\beta\sigma^2T}{B'} + \frac{C_2\beta_1}{\min\{B, B'\}}T + \frac{C_3\beta}{nB'}T$$
$$+ 8\eta\Delta\frac{1}{16\eta^2} + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1} - \nabla F(\mathbf{z}_t)||^2],$$

which follows

$$\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1}-\nabla F(\mathbf{z}_t)||^2] \leq \frac{2}{\beta_1}\Delta_{\mathbf{v}}+\frac{4n}{B\beta}\Delta_{\mathbf{u}}+\frac{4C_1}{B\beta}\Delta_{\mathbf{s}}+\frac{\Delta}{\eta}+T\left(\frac{8C_1\beta\sigma^2}{B'}+\frac{2C_2\beta_1}{\min\{B,B'\}}+\frac{2C_3\beta}{nB'}\right) \tag{40}$$

Combining (40) and Lemma (4), using assumption $\eta L_F \leq \frac{1}{4}$ we obtain

$$\begin{aligned}
&\mathbb{E}[\text{dist}(0,\hat{\partial}\bar{F}(\mathbf{z}_R))^2] \\
&\leq \frac{2+40\eta L_F}{T}\sum_{t=1}^{T}\mathbb{E}[||\mathbf{d}_{t+1}-\nabla F(\mathbf{z}_t)||^2]+\frac{2\Delta}{T\eta}+\frac{40L_F\Delta}{T} \\
&\leq 12\left[\frac{1}{T}\left(\frac{2}{\beta_1}\Delta_{\mathbf{v}}+\frac{4n}{B\beta}\Delta_{\mathbf{u}}+\frac{4C_1}{B\beta}\Delta_{\mathbf{s}}+\frac{3\Delta}{\eta}+40L_F\Delta\right)+\frac{8C_1\beta\sigma^2}{B'}+\frac{2C_2\beta_1}{\min\{B,B'\}}+\frac{2C_3\beta}{nB'}\right]
\end{aligned}$$

By setting $\beta \leq \min\{\frac{B'\epsilon^2}{288C_1\sigma^2},\frac{nB'\epsilon^2}{72C_3}\}=O(B'\epsilon^2)$, $\beta_1 \leq \frac{\min\{B,B'\}\epsilon^2}{72C_2}=O(B'\epsilon^2)$, we have

$$\mathbb{E}[\text{dist}(0,\hat{\partial}\bar{F}(\mathbf{z}_R))^2] \leq 12\left[\frac{1}{T}\left(\frac{2}{\beta_1}\Delta_{\mathbf{v}}+\frac{4n}{B\beta}\Delta_{\mathbf{u}}+\frac{4C_1}{B\beta}\Delta_{\mathbf{s}}+\frac{3\Delta}{\eta}+40L_F\Delta\right)\right]+\frac{2\epsilon^2}{3}.$$

It implies that with

$$\begin{aligned}
T &= \max\left\{\frac{360\Delta_{\mathbf{v}}}{\beta_1\epsilon^2},\frac{720n\Delta_{\mathbf{u}}}{B\beta\epsilon^2},\frac{720C_1\Delta_{\mathbf{s}}}{B\beta\epsilon^2}+\frac{540\Delta}{\eta\epsilon^2}+\frac{7200L_F\Delta}{\epsilon^2}\right\} \\
&= O\left(\max\left\{\frac{1}{\beta_1\epsilon^2},\frac{n}{B\beta\epsilon^2}\right\}\right) \\
&= O\left(\frac{n}{BB'\epsilon^4}\right)
\end{aligned}$$

we have $\mathbb{E}[\text{dist}(0,\hat{\partial}\bar{F}(\mathbf{z}_R))^2] \leq \epsilon^2$, which completes the proof.

□