Stochastic Optimization of Areas Under Precision-Recall Curves with Provable Convergence

Qi Qi [†*], **Youzhi Luo**[‡*], **Zhao Xu**[‡*], **Shuiwang Ji**[‡], **Tianbao Yang**[†]
[†]Department of Computer Science, The University of Iowa
[‡]Department of Computer Science & Engineering, Texas A&M University
{qi-qi,tianbao-yang}@uiowa.edu, {yzluo,zhaoxu,sji}@tamu.edu

Abstract

Areas under ROC (AUROC) and precision-recall curves (AUPRC) are common metrics for evaluating classification performance for imbalanced problems. Compared with AUROC, AUPRC is a more appropriate metric for highly imbalanced datasets. While stochastic optimization of AUROC has been studied extensively, principled stochastic optimization of AUPRC has been rarely explored. In this work, we propose a principled technical method to optimize AUPRC for deep learning. Our approach is based on maximizing the averaged precision (AP), which is an unbiased point estimator of AUPRC. We cast the objective into a sum of *coupled compositional functions* with inner functions dependent on random variables of the outer level. We propose efficient adaptive and non-adaptive stochastic algorithms named SOAP with *provable convergence guarantee under mild conditions* by leveraging recent advances in stochastic compositional optimization. Extensive experimental results on image and graph datasets demonstrate that our proposed method outperforms prior methods on imbalanced problems in terms of AUPRC. To the best of our knowledge, our work represents the first attempt to optimize AUPRC with provable convergence. The SOAP has been implemented in the libAUC library at `https://libauc.org/`.

1 Introduction

Although deep learning (DL) has achieved tremendous success in various domains, the standard DL methods have reached a plateau as the traditional objective functions in DL are no longer sufficient to model all requirements in new applications, which slows down the democratization of AI. For instance, in healthcare applications, data is often highly imbalanced, e.g., patients suffering from rare diseases are much less than those suffering from common diseases. In these applications, accuracy (the proportion of correctly predicted examples) is deemed as an inappropriate metric for evaluating the performance of a classifier. Instead, area under the curve (AUC), including area under ROC curve (AUROC) and area under the Precision-Recall curve (AUPRC), is widely used for assessing the performance of a model. However, optimizing accuracy on training data does not necessarily lead to a satisfactory solution to maximizing AUC [12].

To break the bottleneck for further advancement, DL must be empowered with the capability of efficiently handling novel objectives such as AUC. Recent studies have demonstrated great success along this direction by maximizing AUROC [60]. For example, Yuan et al. [60] proposed a robust deep AUROC maximization method with provable convergence and achieved great success for classification of medical image data. However, to the best of our knowledge, novel DL by maximizing AUPRC has not yet been studied thoroughly. Previous studies [14, 20] have found that when dealing

[*]Contribute Equally. Correspondence to qi-qi@uiowa.edu, tianbao-yang@uiowa.edu

35th Conference on Neural Information Processing Systems (NeurIPS 2021), virtual.

with highly skewed datasets, Precision-Recall (PR) curves could give a more informative picture of an algorithm's performance, which entails the development of efficient stochastic optimization algorithms for DL by maximizing AUPRC.

Compared with maximizing AUROC, maximizing AUPRC is more challenging. The challenges for optimization of AUPRC are two-fold. First, the analytical form of AUPRC by definition involves a complicated integral that is not readily estimated from model predictions of training examples. In practice, AUPRC is usually computed based on some point estimators, e.g., trapezoidal estimators and interpolation estimators of empirical curves, non-parametric average precision estimator, and parametric binomial estimator [3]. Among these estimators, non-parametric average precision (AP) is an unbiased estimate in the limit and can be directly computed based on the prediction scores of samples, which lends itself well to the task of model parameters optimization. Second, a surrogate function for AP is highly complicated and non-convex. In particular, an unbiased stochastic gradient is not readily computed, which makes existing stochastic algorithms such as SGD provide no convergence guarantee. Most existing works for maximizing AP-like function focus on how to compute an (approximate) gradient of the objective function [4, 6, 8, 11, 24, 38, 40, 43, 47, 48], which leave stochastic optimization of AP with provable convergence as an open question.

Can we design direct stochastic optimization algorithms both in SGD-style and Adam-style for maximizing AP with provable convergence guarantee?

In this paper, we propose a systematic and principled solution for addressing this question towards maximizing AUPRC for DL. By using a surrogate loss in lieu of the indicator function in the definition of AP, we cast the objective into a sum of non-convex compositional functions, which resembles a two-level stochastic compositional optimization problem studied in the literature [52, 53]. However, different from existing two-level stochastic compositional functions, the inner functions in our problem are dependent on the random variable of the outer level, which requires us developing a tailored stochastic update for computing an error-controlled stochastic gradient estimator. Specifically, a key feature of the proposed method is to maintain and update two scalar quantities associated with each positive example for estimating the stochastic gradient of the individual precision score at the threshold specified by its prediction score. By leveraging recent advances in stochastic compositional optimization, we propose both adaptive (Adam-style) and non-adaptive (SGD-style) algorithms, and establish their convergence under mild conditions. We conduct comprehensive empirical studies on class imbalanced graph and image datasets for learning graph neural networks and deep convolutional neural networks, respectively. We demonstrate that the proposed method can consistently outperform prior approaches in terms of AUPRC. In addition, we show that our method achieves better results when the sample distribution is highly imbalanced between classes and is insensitive to mini-batch size.

2 Related Work

AUROC Optimization. AUROC optimization [2] has attracted significant attention in the literature. Recent success of DL by optimizing AUROC on large-scale medical image data has demonstrated the importance of large-scale stochastic optimization algorithms and the necessity of accurate surrogate function [60]. Earlier papers [25, 28] focus on learning a linear model based on the pairwise surrogate loss and could suffer from a high computational cost, which could be as high as quadratic of the size of training data. To address the computational challenge, online and stochastic optimization algorithms have been proposed [18, 35, 42, 58, 63]. Recently, [21, 22, 36, 57] proposed stochastic deep AUC maximization algorithms by formulating the problem as non-convex strongly-concave min-max optimization problem, and derived fast convergence rate under PL condition, and in federated learning setting as well [21]. More recently, Yuan et al. [60] demonstrated the success of their methods on medical image classification tasks, e.g., X-ray image classification, melanoma classification based on skin images. However, an algorithm that maximizes the AUROC might not necessarily maximize AUPRC, which entails the development of efficient algorithms for DL by maximizing AUPRC.

AUPRC Optimization. AUPRC optimization is much more challenging than AUROC optimization since the objective is even not decomposable over pairs of examples. Although AUPRC optimization has been considered in the literature (cf. [15, 47, 41] and references therein), efficient scalable algorithms for DL with provable convergence guarantee is still lacking. Some earlier works tackled

[2] In the literature, AUROC optimization is simply referred to as AUC optimization.

this problem by using traditional optimization techniques, e.g., hill climbing search [37], cutting-plane method [61], dynamic programming [50], and by developing acceleration techniques in the framework of SVM [39]. These approaches are not scalable to big data for DL. There is a long list of studies in information retrieval [5, 11, 38, 47] and computer vision [4, 6, 8, 9, 24, 40, 48, 43], which have made efforts towards maximizing the AP score. However, most of them focus on how to compute an approximate gradient of the AP function or its smooth approximation, and provide no convergence guarantee for stochastic optimization based on mini-batch averaging. Due to lack of principled design, these previous methods when applied to deep learning are sensitive to the mini-batch size [6, 47, 48] and usually require a large mini-batch size in order to achieve good performance. In contrast, our stochastic algorithms are designed in a principled way to guarantee convergence without requiring a large mini-batch size as confirmed by our studies as well. Recently, [15] formulates the objective function as a constrained optimization problem using a surrogate function, and then casts it into a min-max saddle-point problem, which facilitates the use of stochastic min-max algorithms. However, they do not provide any convergence analysis for AUPRC maximization. In contrast, this is the first work that directly optimizes a surrogate function of AP (an unbaised estimator of AUPRC in the limit) and provides theoretical convergence guarantee for the proposed stochastic algorithms.

Stochastic Compositional Optimization. Optimization of a two-level compositional function in the form of $\mathbb{E}_\xi[f(\mathbb{E}_\zeta[g(\mathbf{w};\zeta)];\xi)]$ where ξ and ζ are independent random variables, or its finite-sum variant has been studied extensively in the literature [1, 10, 52, 27, 30, 31, 33, 34, 46, 53, 59, 62, 45]. In this paper, we formulate the surrogate function of AP into a similar but more complicated two-level compositional function of the form $\mathbb{E}_\xi[f(\mathbb{E}_\zeta g(\mathbf{w};\zeta,\xi))]$, where ξ and ζ are independent and ξ has a finite support. The key difference between our formulated compositional function and the ones considered in previous work is that the inner function $g(\mathbf{w};\zeta,\xi)$ also depends on the random variable ξ of the outer level. Such subtle difference will complicate the algorithm design and the convergence analysis as well. Nevertheless, the proposed algorithm and its convergence analysis are built on previous studies of stochastic two-level compositional optimization.

3 The Proposed Method

Notations. We consider binary classification problem. Denote by $(\mathbf{x}, y)$ a data pair, where $\mathbf{x} \in \mathbb{R}^d$ denotes the input data and $y \in \{1, -1\}$ denotes its class label. Let $h(\mathbf{x}) = h_\mathbf{w}(\mathbf{x})$ denote the predictive function parameterized by a parameter vector $\mathbf{w} \in \mathbb{R}^D$ (e.g., a deep neural network). Denote by $\mathbf{I}(\cdot)$ an indicator function that outputs 1 if the argument is true and zero otherwise. To facilitate the presentation, denote by X a random data, by Y its label and by $F = h(X)$ its prediction score. Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ denote the set of all training examples and $\mathcal{D}_+ = \{\mathbf{x}_i : y_i = 1\}$ denote the set of all positive examples. Let $n_+ = |\mathcal{D}_+|$ denote the number of positive examples. $\mathbf{x}_i \sim \mathcal{D}$ means that $\mathbf{x}_i$ is randomly sampled from $\mathcal{D}$.

3.1 Background on AUPRC and its estimator AP

Following the work of Bamber [2], AUPRC is an average of the precision weighted by the probability of a given threshold, which can be expressed as

$$A = \int_{-\infty}^{\infty} \Pr(Y = 1|F \geq c) d \Pr(F \leq c|Y = 1),$$

where $\Pr(Y = 1|F \geq c)$ is the precision at the threshold value of c. The above integral is an importance-sampled Monte Carlo integral, by which we may interpret AUPRC as the fraction of positive examples among those examples whose output values exceed a randomly selected threshold $c \sim F(X)|Y = 1$.

For a finite set of examples $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ with the prediction score for each example $\mathbf{x}_i$ given by $h_\mathbf{w}(\mathbf{x}_i)$, we consider to use AP to approximate AUPRC, which is given by

$$\text{AP} = \frac{1}{n_+} \sum_{i=1}^{n} \mathbf{I}(y_i = 1) \frac{\sum_{s=1}^{n} \mathbf{I}(y_s = 1)\mathbf{I}(h_\mathbf{w}(\mathbf{x}_s) \geq h_\mathbf{w}(\mathbf{x}_i))}{\sum_{s=1}^{n} \mathbf{I}(h_\mathbf{w}(\mathbf{x}_s) \geq h_\mathbf{w}(\mathbf{x}_i))}, \tag{1}$$

where n_+ denotes the number of positive examples. It can be shown that AP is an unbiased estimator in the limit $n \to \infty$ [3].

However, the non-continuous indicator function $\mathbf{I}(h_{\mathbf{w}}(\mathbf{x}_s) \geq h_{\mathbf{w}}(\mathbf{x}_i))$ in both numerator and denominator in (1) makes the optimization non-tractable. To tackle this, we use a loss function $\ell(\mathbf{w}; \mathbf{x}_s, \mathbf{x}_i)$ as a surrogate function of $\mathbf{I}(h_{\mathbf{w}}(\mathbf{x}_s) \geq h_{\mathbf{w}}(\mathbf{x}_i))$. One can consider different surrogate losses, e.g., hinge loss, squared hinge loss, and smoothed hinge loss, and exponential loss. In this paper, we will consider a smooth surrogate loss function to facilitate the development of an optimization algorithm, e.g., a squared hinge loss $\ell(\mathbf{w}; \mathbf{x}_s; \mathbf{x}_i) = (\max\{m - (h_{\mathbf{w}}(\mathbf{x}_i) - h_{\mathbf{w}}(\mathbf{x}_s)), 0\})^2$, where m is a margin parameter. Note that we do not require ℓ to be a convex function, hence one can also consider non-convex surrogate loss such as ramp loss. As a result, our problem becomes

$$\min_{\mathbf{w}} P(\mathbf{w}) = \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{D}_+} \frac{-\sum_{s=1}^{n} \mathbf{I}(y_s = 1)\ell(\mathbf{w}; \mathbf{x}_s; \mathbf{x}_i)}{\sum_{s=1}^{n} \ell(\mathbf{w}; \mathbf{x}_s; \mathbf{x}_i)}. \tag{2}$$

3.2 Stochastic Optimization of AP (SOAP)

We cast the problem into a finite-sum of compositional functions. To this end, let us define a few notations:

$$\begin{aligned} &g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) = [g_1(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i), g_2(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)]^\top = [\ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)\mathbf{I}(y_j = 1), \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)]^\top \\ &g_{\mathbf{x}_i}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)], \end{aligned} \tag{3}$$

where $g_{\mathbf{x}_i}(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^2$. Let $f(\mathbf{s}) = -\frac{s_1}{s_2} : \mathbb{R}^2 \to \mathbb{R}$. Then, we can write the objective function for maximizing AP as a sum of compositional functions:

$$P(\mathbf{w}) = \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{D}_+} f(g_{\mathbf{x}_i}(\mathbf{w})) = \mathbb{E}_{\mathbf{x}_i \sim \mathcal{D}_+}[f(g_{\mathbf{x}_i}(\mathbf{w}))]. \tag{4}$$

We refer to the above problem as an instance of **two-level stochastic coupled compositional functions**. It is similar to the two-level stochastic compositional functions considered in literature [52, 53] but with a subtle difference. The difference is that in our formulation the inner function $g_{\mathbf{x}_i}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)]$ depends on the random variable $\mathbf{x}_i$ of the outer level. This difference makes the proposed algorithm slightly complicated by estimating $g_{\mathbf{x}_i}(\mathbf{w})$ separately for each positive example. It also complicates the analysis of the proposed algorithms. Nevertheless, we can still employ the techniques developed for optimizing stochastic compositional functions to design the algorithms and develop the analysis for optimizing the objective (4).

In order to motivate the proposed method, let us consider how to compute the gradient of $P(\mathbf{w})$. Let the gradient of $g_{\mathbf{x}_i}(\mathbf{w})$ be denoted by $\nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w})^\top = (\nabla_{\mathbf{w}}[g_{\mathbf{x}_i}(\mathbf{w})]_1, \nabla_{\mathbf{w}}[g_{\mathbf{x}_i}(\mathbf{w})]_2)$. Then we have

$$\begin{aligned} \nabla_{\mathbf{w}} P(\mathbf{w}) &= \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{D}_+} \nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w})^\top \nabla f(g_{\mathbf{x}_i}(\mathbf{w})) \\ &= \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{D}_+} \nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w})^\top \left(\frac{-1}{[g_{\mathbf{x}_i}(\mathbf{w})]_2}, \frac{[g_{\mathbf{x}_i}(\mathbf{w})]_1}{([g_{\mathbf{x}_i}(\mathbf{w})]_2)^2} \right)^\top. \end{aligned} \tag{5}$$

The major cost for computing $\nabla_{\mathbf{w}} P(\mathbf{w})$ lies at evaluating $g_{\mathbf{x}_i}(\mathbf{w})$ and its gradient $\nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w})$, which involves passing through all examples in $\mathcal{D}$.

To this end, we will approximate these quantities by stochastic samples. The gradient $\nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w})$ can be simply approximated by the stochastic gradient, i.e.,

$$\widehat{\nabla}_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w}) = \begin{pmatrix} \frac{1}{B} \sum_{\mathbf{x}_j \in \mathcal{B}} \mathbf{I}(y_j = 1) \nabla \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) \\ \frac{1}{B} \sum_{\mathbf{x}_j \in \mathcal{B}} \nabla \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) \end{pmatrix}, \tag{6}$$

where $\mathcal{B}$ denote a set of B random samples from $\mathcal{D}$. For estimating $g_{\mathbf{x}_i}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}} g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)$, however, we need to ensure its approximation error is controllable due to the compositional structure such that the convergence can be guaranteed. We borrow a technique from the literature of stochastic compositional optimization [52] by using moving average estimator for estimating $g_{\mathbf{x}_i}(\mathbf{w})$ for all positive examples. To this end, we will maintain a matrix $\mathbf{u} = [\mathbf{u}^1, \mathbf{u}^2]$ with each column indexable by any positive example, i.e., $\mathbf{u}^1_{\mathbf{x}_i}, \mathbf{u}^2_{\mathbf{x}_i}$ correspond to the moving average estimator of $[g_{\mathbf{x}_i}(\mathbf{w})]_1$ and $[g_{\mathbf{x}_i}(\mathbf{w})]_2$, respectively. The matrix $\mathbf{u}$ is updated by the subroutine UG in Algorithm 2, where $\gamma \in (0, 1)$ is a parameter. It is notable that in Step 3 of Algorithm 2, we clip the moving average update of $\mathbf{u}^2_{\mathbf{x}_i}$ by a lower bound u_0, which is a given parameter. This step can ensure the division in computing the stochastic gradient estimator in (7) always valid and is also important for convergence

Algorithm 1: SOAP

Input: γ, α, u_0, and other parameters for SGD-stype update or Adam-stype update.
Initialize $\mathbf{w}_1 \in \mathbb{R}^d$, $\mathbf{u} \in \mathbb{R}^{|n_+|\times 2}$
for $t = 1, \ldots, T$ **do**
  Draw a batch of B_+ positive samples denoted by $\mathcal{B}_+$.
  Draw a batch of B samples denoted by $\mathcal{B}$.
  $\mathbf{u} = \text{UG}(\mathcal{B}, \mathcal{B}_+, \mathbf{u}, \mathbf{w}_t, \gamma, u_0)$
  Compute (biased) Stochastic Gradient Estimator

$$G(\mathbf{w}_t) = \frac{1}{B_+} \sum_{\mathbf{x}_i \in \mathcal{B}_+} \sum_{\mathbf{x}_j \in \mathcal{B}} \frac{(\mathbf{u}^1_{\mathbf{x}_i} - \mathbf{u}^2_{\mathbf{x}_i} \mathbf{I}(\mathbf{y}_j = 1)) \nabla \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)}{B(\mathbf{u}^2_{\mathbf{x}_i})^2} \tag{7}$$

  Update $\mathbf{w}_{t+1}$ by a SGD-style method or by a Adam-style method

$$\mathbf{w}_{t+1} = \text{UW}(\mathbf{w}_t, G(\mathbf{w}_t))$$

end for
Return: last solution.

analysis. With these stochastic estimators, we can compute an estimate of $\nabla P(\mathbf{w})$ by equation (7), where $\mathcal{B}_+$ includes a batch of sampled positive data. With this stochastic gradient estimator, we can employ SGD-style method and Adam-style shown in Algorithm 3 to update the model parameter $\mathbf{w}$. The final algorithm named as SOAP is presented in Algorithm 1.

Algorithm 2: UG($\mathcal{B}, \mathcal{B}_+, \mathbf{u}, \mathbf{w}_t, \gamma, u_0$)

for each positive $\mathbf{x}_i \in \mathcal{B}_+$ **do**
  Compute

$$[\tilde{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_1 = \frac{1}{|\mathcal{B}|} \sum_{\substack{\mathbf{x}_j \in \mathcal{B} \\ \mathbf{y}_j = 1}} \ell(\mathbf{w}_t; \mathbf{x}_j, \mathbf{x}_i)$$

$$[\tilde{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_2 = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x}_j \in \mathcal{B}} \ell(\mathbf{w}_t; \mathbf{x}_j, \mathbf{x}_i)$$

  Compute
  $\mathbf{u}^1_{\mathbf{x}_i} = (1-\gamma)\mathbf{u}^1_{\mathbf{x}_i} + \gamma[\tilde{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_1$
  $\mathbf{u}^2_{\mathbf{x}_i} = \max((1-\gamma)\mathbf{u}^2_{\mathbf{x}_i} + \gamma[\tilde{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_2, u_0)$
end for
Return $\mathbf{u}$

Algorithm 3: UW($\mathbf{w}_t, G(\mathbf{w}_t)$)

Option 1: SGD-style update (paras: α)

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha G(\mathbf{w}_t)$$

Option 2: Adam-style update (paras: $\alpha, \epsilon, \eta_1, \eta_2$)

$$h_{t+1} = \eta_1 h_t + (1-\eta_1) G(\mathbf{w}_t)$$

$$v_{t+1} = \eta_2 \hat{v}_t + (1-\eta_2)(G(\mathbf{w}_t))^2$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \frac{h_{t+1}}{\sqrt{\epsilon + \hat{v}_{t+1}}}$$

where $\hat{v}_t = v_t$ (Adam) or
$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ (AMSGrad)
Return: $\mathbf{w}_{t+1}$

3.3 Convergence Analysis

In this subsection, we present the convergence results of SOAP and also highlight its convergence analysis. To this end, we first present the following assumption.

Assumption 1. *Assume that (a) there exists Δ_1 such that $P(\mathbf{w}_1) - \min_{\mathbf{w}} P(\mathbf{w}) \leq \Delta_1$; (b) there exist $C, M > 0$ such that $\ell(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_i) \geq C$ for any $\mathbf{x}_i \in \mathcal{D}_+$, $\ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) \leq M$, and $\ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)$ is Lipscthiz continuous and smooth with respect to $\mathbf{w}$ for any $\mathbf{x}_i \in \mathcal{D}_+, \mathbf{x}_j \in \mathcal{D}$; (c) there exists $V > 0$ such that $\mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\|g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) - g_{\mathbf{x}_i}(\mathbf{w})\|^2] \leq V$, and $\mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\|\nabla g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) - \nabla g_{\mathbf{x}_i}(\mathbf{w})\|^2] \leq V$ for any $\mathbf{x}_i$.*

With a bounded score function $h_{\mathbf{w}}(\mathbf{x})$ the above assumption can be easily satisfied. Based on the above assumption, we can prove that the objective function $P(\mathbf{w})$ is smooth.

Lemma 1. *Suppose Assumption 1 holds, then there exists $L > 0$ such that $P(\cdot)$ is L-smooth. In addition, there exists $u_0 \geq C/n$ such that $g_{\mathbf{x}_i}(\mathbf{w}) \in \Omega = \{\mathbf{u} \in \mathbb{R}^2, 0 \leq [\mathbf{u}]_1 \leq M, u_0 \leq [\mathbf{u}]_2 \leq M\}$, $\forall \mathbf{x}_i \in \mathcal{D}_+$.*

Next, we highlight the convergence analysis of SOAP employing the SGD-stype update and include that for employing Adam-style update in the supplement. Without loss of generality, we assume $|\mathcal{B}_+| = 1$ and the positive sample in $\mathcal{B}_+$ is randomly selected from $\mathcal{D}_+$ with replacement. When the context is clear, we abuse the notations $g_i(\mathbf{w})$ and $\mathbf{u}_i$ to denote $g_{\mathbf{x}_i}(\mathbf{w})$ and $\mathbf{u}_{\mathbf{x}_i}$ below, respectively. We first establish the following lemma following the analysis of non-convex optimization.

Lemma 2. *With $\alpha \leq 1/2$, running T iterations of SOAP (SGD-style) updates, we have*
$$\frac{\alpha}{2}\mathbb{E}[\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2] \leq \mathbb{E}[\sum_{t}(P(\mathbf{w}_t) - P(\mathbf{w}_{t+1}))] + \frac{\alpha C_1}{2}\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 T C_2,$$
where i_t denotes the index of the sampled positive data at iteration t, C_1 and C_2 are proper constants.

Our key contribution is the following lemma that bounds the second term in the above upper bound.

Lemma 3. *Suppose Assumption 1 holds, with $\mathbf{u}$ initialized by (6) for every $\mathbf{x}_i \in \mathcal{D}_+$ we have*
$$\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] \leq \frac{n_+ V}{\gamma} + \gamma V T + 2\frac{n_+^2\alpha^2 T C_3}{\gamma^2}, \tag{8}$$
where C_3 is a proper constant.

Remark: The innovation of proving the above lemma is by grouping $\mathbf{u}_{i_t}, t = 1, \ldots, T$ into n_+ groups corresponding to the n_+ positive examples, and then establishing the recursion of the error $\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2$ within each group, and then summing up these recursions together.

Based on the two lemmas above, we establish the following convergence of SOAP with a SGD-style update.

Theorem 1. *Suppose Assumption 1 holds, let the parameters be $\alpha = \frac{1}{n_+^{2/5}T^{3/5}}$,$\gamma = \frac{n_+^{2/5}}{T^{2/5}}$, $\forall\, t \in 1, \cdots, T$, and $T > n_+$. Then after running T iterations, SOAP with a SGD-style update satisfies $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2\right] \leq O(\frac{n_+^{2/5}}{T^{2/5}})$, where O suppresses constant numbers.*

Remark: To the best of our knowledge, this is the first time a stochastic algorithm was proved to converge for AP maximization.

Similarly, we can establish the following convergence of SOAP by employing an Adam-style update, specifically the AMSGrad update.

Theorem 2. *Suppose Assumption 1 holds, let the parameters $\eta_1 \leq \sqrt{\eta_2} \leq 1$, $\alpha = \frac{1}{n_+^{2/5}T^{3/5}}$,$\gamma = \frac{n_+^{2/5}}{T^{2/5}}$, $\forall\, t \in 1, \cdots, T$, and $T > n_+$. Then after running T iterations, SOAP with an AMSGRAD update satisfies $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2\right] \leq O(\frac{n_+^{2/5}}{T^{2/5}})$, where O suppresses constant numbers.*

4 Experiments

In this section, we evaluate the proposed method through comprehensive experiments on imbalanced datasets. We show that the proposed method can outperform prior state-of-the-art methods for imbalanced classification problems. In addition, we conduct experiments on (i) the effects of imbalance ratio; (ii) the insensitivity to batch size and (iii) the convergence speed on testing data; and observe that our method (i) is more advantageous when data is more imbalanced, (ii) is not sensitive to batch size, and (iii) converges faster than baseline methods.

Our proposed optimization algorithm is independent of specific datasets and tasks. Therefore, we perform experiments on both graph and image prediction tasks. In particular, the graph prediction tasks in the contexts of molecular property prediction and drug discovery suffer from very severe imbalance problems as positive labels are very rare while negative samples are abundantly available. Thus, we choose to use graph data intensively in our experiments. Additionally, the graph data we use allow us to vary the imbalance ratio to observe the performance change of different methods.

In all experiments, we compare our method with the following baseline methods. **CB-CE** refers to a method using a class-balanced weighed cross entropy loss function, in which the weights for positive and negative samples are adjusted with the strategy proposed by Cui et al. [13]. **Focal** is to up-weight the penalty on hard examples using focal loss [32]. **LDAM** refers to training with label-distribution-aware margin loss [7]. **AUC-M** is an AUROC maximization method using a surrogate loss [60]. In addition, we compare with three methods for optimizing AUPRC or AP, namely, the **MinMax** method [15] - a method for optimizing a discrete approximation of AUPRC, **SmoothAP** [4] - a method that optimizes a smoothed approximation of AP, and **FastAP** - a method that uses soft histogram binning to approximate the gradient of AP [6]. For all of these methods, we use the

Table 1: The test AUPRC on the image datasets with two ResNet models. We report the average AUPRC and standard deviation (within brackets) over 5 runs.

Datasets	CIFAR-10		CIFAR-100	
Networks	ResNet18	ResNet34	ResNet18	ResNet34
CE	0.7155 (± 0.0058)	0.6844(± 0.0031)	0.5946 (± 0.0031)	0.5792 (± 0.0028)
CB-CE	0.7325 (± 0.0039)	0.6936(±0.0021)	0.6165 (± 0.0096)	0.5632(± 0.0129)
Focal	0.7183(± 0.0082)	0.6943(± 0.0007)	0.6107(± 0.0093)	0.5585(± 0.0285)
LDAM	0.7346 (± 0.0125)	0.6745(± 0.0043)	0.6153 (± 0.0100)	0.5662(± 0.0212)
AUC-M	0.7399(± 0.0013)	0.6825(± 0.0089)	0.6103 (± 0.0075)	0.5306(± 0.0230)
SmoothAP	0.7365 (± 0.0088)	0.6909 (± 0.0049)	0.6071(± 0.0143)	0.5208 (± 0.0505)
FastAP	0.7028 (± 0.0341)	0.6798 (± 0.0032)	0.5618(± 0.0351)	0.5151(± 0.0450)
MinMax	0.7228 (± 0.0118)	0.6806(± 0.0027)	0.6071(± 0.0064)	0.5518(± 0.0030)
SOAP	**0.7629**(± 0.0014)	**0.7012**(± 0.0056)	**0.6251** (± 0.0053)	**0.6001**(± 0.0060)

SGD-style with momentum optimization for image prediction tasks and the Adam-style optimization algorithms for graph prediction tasks and unless specified otherwise. We refer to **imbalance ratio** as the number of positive samples over the total number of examples of a considered set. The hyper-parameters of all methods are fine tuned using cross-validation with training/validation splits mentioned below. For AP maximization methods, we use a sigmoid function to produce the prediction score. For simplicity, we set $u_0 = 0$ for SOAP and encounter no numerical problems in experiments. As SOAP requires positive samples for updating $\mathbf{u}$ to approximate the gradient of surrogate objective, we use a data sampler which samples a few positive examples (e.g., 2) and some negative examples per iteration. The same sampler applies to all methods for fair comparison. The code for reproducing the results is released here [44].

4.1 Image Classification

Data. We first conduct experiments on three image datasets: CIFAR10, CIFAR100 and Melanoma dataset [49]. We construct imbalanced version of CIFAR10 and CIFAR100 for binary classification. In particular, for each dataset we manually take the last half of classes as positive class and first half of classes as negative class. To construct highly imbalanced data, we remove 98% of the positive images from the training data and keep the test data unchanged (i.e., the testing data is still balanced). And we split the training dataset into train/validation set at 80%/20% ratio. The Melanoma dataset is from a medical image Kaggle competition, which serves as a natural real imbalanced image dataset. It contains 33,126 labeled medical images, among which 584 images are related to malignant melanoma and labelled as positive samples. Since the test set used by Kaggle organization is not available, we manually split the training data into train/validation/test set at 80%/10%/10% ratio and report the achieved AUPRC on the test set by our method and baselines. The images of Melanoma dataset are always resized to have a resolution of 384×384 in our experiments.

Setup. We use two ResNet [23] models, *i.e.*, ResNet18 and ResNet34, as the backbone networks for image classification. For all methods except for CE, the ResNet models are initialized with a model pre-trained by CE with a SGD optimizer. We tune the learning rate in a range {1e-5, 1e-4, 1e-3, 1e-2} and the weight decay parameter in a range {1e-6, 1e-5, 1e-4}. Then the last fully connected layer is randomly re-initialized and the network is trained by different methods with the same weight decay parameter but other hyper-parameters individually tuned for fair comparison, e.g., we tune γ of SOAP in a range {0.9, 0.99,0.999}, and tune m in {0.5, 1, 2, 5, 10}. We refer to this scheme as two-stage training, which is widely used for imbalanced data [60]. We consistently observe that this strategy can bring the model to a good initialization state and improve the final performance of our method and baselines.

Results. Table 1 shows the AUPRC on testing sets of CIFAR-10 and CIFAR-100. We report the results on Melanoma in Table 3. We can observe that the proposed method SOAP outperforms all baselines. It is also striking to see that on Melanoma dataset, our proposed SOAP can outperform all baselines by a large margin, and all other methods have very poor performance. The reason is that the testing set of Melanoma is also imbalanced (imbalanced ratio=1.72%), while the testing sets of CIFAR-10 and CIFAR-100 are balanced. We also observe that the AUROC maximization (AUC-M) does not necessarily optimize AUPRC. We also plot the final PR curves in Figure 3 in the supplement.

Table 2: The test AUPRC values on the HIV and MUV datasets with three graph neural network models. We report the average AUPRC and standard deviation (within brackets) over 3 runs.

Dataset	Method	GINE	MPNN	ML-MPNN
HIV	CE	0.2774 (± 0.0101)	0.3197 (± 0.0050)	0.2988 (± 0.0076)
	CB-CE	0.3082 (± 0.0101)	0.3056 (± 0.0018)	0.3291 (± 0.0189)
	Focal	0.3179 (± 0.0068)	0.3136 (± 0.0197)	0.3279 (± 0.0173)
	LDAM	0.2904 (± 0.0008)	0.2994 (± 0.0128)	0.3044 (± 0.0116)
	AUC-M	0.2998 (± 0.0010)	0.2786 (± 0.0456)	0.3305 (± 0.0165)
	SmothAP	0.2686 (± 0.0007)	0.3276 (± 0.0063)	0.3235 (± 0.0092)
	FastAP	0.0169 (± 0.0031)	0.0826 (± 0.0112)	0.0202 (± 0.0002)
	MinMax	0.2874 (± 0.0073)	0.3119 (± 0.0075)	0.3098 (± 0.0167)
	SOAP	**0.3385 (± 0.0024)**	**0.3401 (± 0.0045)**	**0.3547 (± 0.0077)**
MUV	CE	0.0017 (±0.0001)	0.0021 (±0.0002)	0.0025 (±0.0004)
	CB-CE	0.0055 (±0.0011)	0.0483 (±0.0083)	0.0121 (±0.0016)
	Focal	0.0041 (±0.0007)	0.0281 (±0.0141)	0.0122 (±0.0001)
	LDAM	0.0044 (±0.0022)	0.0118 (±0.0098)	0.0059 (±0.0021)
	AUC-M	0.0026 (±0.0001)	0.0040 (±0.0012)	0.0028 (±0.0012)
	SmoothAP	0.0073 (±0.0012)	0.0068 (±0.0038)	0.0029 (±0.0005)
	FastAP	0.0016 (±0.0000)	0.0023 (±0.0021)	0.0022 (±0.0012)
	MinMax	0.0028 (±0.0008)	0.0027 (±0.0005)	0.0043 (±0.0015)
	SOAP	**0.0254 (±0.0261)**	**0.3352** (±0.0008)	**0.0236** (±0.0038)

4.2 Graph Classification for Molecular Property Prediction

Data. To further demonstrate the advantages of our method, we conduct experiments on two graph classification datasets. We use the datasets HIV and MUV from the MoleculeNet [55], which is a benchmark for molecular property prediction. The HIV dataset has 41,913 molecules from the Drug Therapeutics Program (DTP), and the positive samples are molecules tested to have inhibition ability to HIV. The MUV dataset has 93,127 molecules from the PubChem library, and molecules are labelled by whether a bioassay property exists or not. Note that the MUV dataset provides labels of 17 properties in total and we only conduct experiments to predict the third property as this property is more imbalanced. The percentage of positive samples in HIV and MUV datasets are 3.51% and 0.20%, respectively. We use the split of train/validation/test set provided by MoleculeNet. Molecules are treated as 2D graphs in our experiments, and we use the feature extraction procedure of MoleculeKit [54] to obtain node features of graphs. The same data preprocessing is used for all of our experiments on graph data.

Setup. Many recent studies have shown that graph neural networks (GNNs) are powerful models for graph data analysis [29, 17, 16]. Hence, we use three different GNNs as the backbone network for graph classification, including the message passing neural network (**MPNN**) [19], an invariant of graph isomorphism network [56] named by **GINE** [26], and the multi-level message passing neural network (**ML-MPNN**) proposed by Wang et al. [54]. We use the same two-stage training scheme with a similar hyper-parameter tuning. We pre-train the networks by Adam with 100 epochs and a tuned initial learning rate 0.0005, which is decayed by half after 50 epochs.

Results. The achieved AUPRC on the test set by all methods are presented in Table 2. Results show that our method can outperform all baselines by a large margin in terms of AUPRC, regardless of which model structure is used. These results clearly demonstrate that our method is effective for classification problems in which the sample distribution is highly imbalanced between classes.

4.3 Graph Classification for Drug Discovery

Data. In addition to molecular property prediction, we explore applying our method to drug discovery. Recent studies have shown that GNNs are effective in drug discovery through predicting the antibacterial property of chemical compounds [51]. Such application scenarios involves training a GNN model on labeled datasets and making predictions on a large library of chemical compounds so as to discover new antibiotic. However, because the positive samples in the training data, *i.e.*, compounds known to have antibacterial property, are very rare, there exists very severe class imbalance.

We show that our method can serve as a useful solution to the above problem. We conduct experiments on the MIT AICURES dataset from an open challenge (`https://www.aicures.mit.edu/tasks`)

Table 3: The test AUPRC values on the MIT AICURES dataset with two graph neural networks, and on the Kaggle Melanoma dataset with two CNN models. We report the average AUPRC and standard deviation (within brackets) from 3 independent runs over 3 different train/validation/test splits.

Data	MIT AICURES		Kaggle Melanoma	
Networks	GINE	MPNN	ResNet18	ResNet34
CE	0.5037 ($\pm$ 0.0718)	0.6282 ($\pm$ 0.0634)	0.0701 ($\pm$ 0.0031)	0.0582 ($\pm$ 0.0016)
CB-CE	0.5655 ($\pm$ 0.0453)	0.6308 ($\pm$ 0.0263)	0.0631 ($\pm$ 0.0065)	0.0721 ($\pm$ 0.0054)
Focal	0.5143 ($\pm$ 0.1062)	0.5875 ($\pm$ 0.0774)	0.0549 ($\pm$ 0.0083)	0.0663 ($\pm$ 0.0034)
LDAM	0.5236 ($\pm$ 0.0551)	0.6489 ($\pm$ 0.0556)	0.0547 ($\pm$ 0.0046)	0.0539 ($\pm$ 0.0069)
AUC-M	0.5149 ($\pm$ 0.0748)	0.5542 ($\pm$ 0.0474)	0.1013 ($\pm$ 0.0071)	0.0972 ($\pm$ 0.0035)
SmothAP	0.2899 ($\pm$ 0.0220)	0.4081 ($\pm$ 0.0352)	0.1981 ($\pm$ 0.0527)	0.2787 ($\pm$ 0.0232)
FastAP	0.4777 ($\pm$ 0.0896)	0.4518 ($\pm$ 0.1495)	0.0324 ($\pm$ 0.0087)	0.0359 ($\pm$ 0.0062)
MinMax	0.5292 ($\pm$ 0.0330)	0.5774 ($\pm$ 0.0468)	0.0593 ($\pm$ 0.0037)	0.0663 ($\pm$ 0.0084)
SOAP	**0.6639 ($\pm$ 0.0515)**	**0.6547 ($\pm$ 0.0616)**	**0.2624 ($\pm$ 0.0410)**	**0.3152 ($\pm$ 0.0337)**

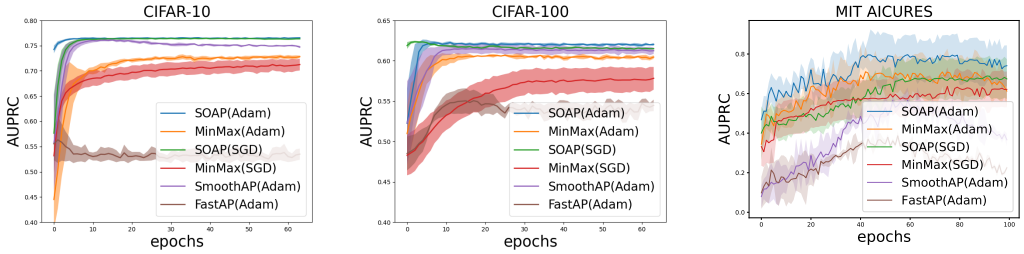

Figure 1: Comparison of convergence of different methods in terms of test AUPRC scores on CIFAR-10, CIFAR100 and MIT AICURES data.

in drug discovery. The dataset consists of 2097 molecules. There are 48 positive samples that have antibacterial activity to *Pseudomonas aeruginosa*, which is the pathogen leading to secondary lungs infections of COVID-19 patients. We conduct experiments on three random train/validation/test splits at 80%/10%/10% ratio, and report the average AUPRC on the test set over three splits.

Setup. Following the setup in Sec. 4.2, we use three GNNs: MPNN, GINE and ML-MPNN. We use the same two-stage training scheme with a similar hyper-parameter tuning. We pre-train GNNs by the Adam method for 100 epochs with a batch size of 64 and a tuned learning rate of 0.0005, which is decayed by half at the 50th epoch. Due to the limit of space, Table 3 only reports GINE and MPNN results. Please refer to Table 6 in the supplement for the full results of all three GNNs.

Results. The average test AUPRC from three independent runs over three splits are summarized in Table 3, Table 6. We can see that our SOAP can consistently outperform all baselines on all three GNN models. Our proposed optimization method can significantly improve the achieved AUPRC of GNN models, indicating that models tend to assign higher confidence scores to molecules with antibacterial activity. This can help identify a larger number of candidate drugs.

We have employed the proposed AUPRC maximization method for improving the testing performance on MIT AICures Challenge and achieved the 1st place. For details, please refer to [54].

4.4 Ablation Studies

Effects of Imbalance Ratio. We now study the effects of imbalance ratio on the performance improvements of our method. We use two datasets Tox21 and ToxCast from the MoleculeNet [55]. The Tox21 and ToxCast contain 8014 and 8589 molecules, respectively. There are 12 property prediction tasks in Tox21, and we conduct experiments on Task 0 and Task 2. Similarly, we select Task 12 and Task 8 of ToxCast for experiments. We use the split of train/validation/test set provided by MoleculeNet. The imbalanced ratios on the training sets are 4.14% for Task 0 of Tox21, 12.00% for Task 2 of Tox21, 2.97% for Task 12 of ToxCast, 8.67% for Task 8 of ToxCast.

Following Sec. 4.2, we test three neural network models MPNN, GINE and ML-MPNN. The hyper-parameters for training models are also the same as those in Sec. 4.2. We present the results of Tox21

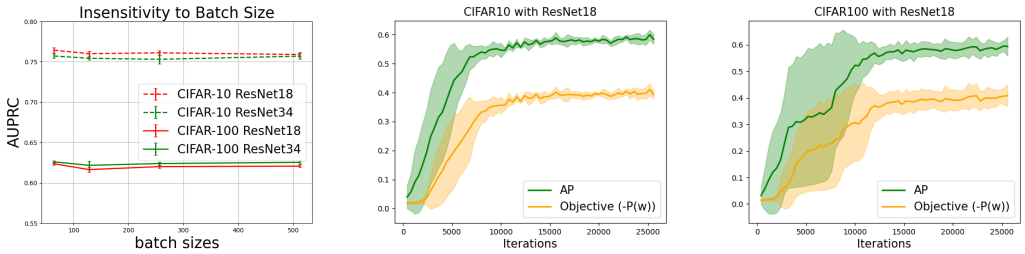

Figure 2: Left most: insensitivity to batch size of SOAP. Right two: consistency between AP and Surrogate Objective -$P(\mathbf{w})$ vs Iterations on CIFAR10 and CIFAR100.

Table 4: The test AUPRC over 3 independent runs by SOAP with different surrogate functions.

Data	CIFAR10		CIFAR100	
Networks	ResNet18	ResNet34	ResNet18	ResNet34
Squared Hinge	0.7629 ($\pm$0.0014)	0.7012 ($\pm$0.0056)	0.6251 ($\pm$0.0053)	0.6001 ($\pm$0.0060)
Logistic	0.7542 ($\pm$0.0024)	0.6968 ($\pm$0.0121)	0.6378 ($\pm$0.0031)	0.5923 ($\pm$0.0101)
Sigmoid	0.7652 ($\pm$0.0035)	0.6983 ($\pm$0.0084)	0.6271 ($\pm$0.0043)	0.5832 ($\pm$0.0054)
Data	HIV		MUV	
Networks	GINE	MPNN	GINE	MPNN
Squared Hinge	0.3485 ($\pm$0.0083)	0.3401 ($\pm$0.0045)	0.0354 ($\pm$0.0025)	0.3365 ($\pm$0.0008)
Logistic	0.3436 ($\pm$0.0043)	0.3617 ($\pm$0.0031)	0.0493 ($\pm$0.0261)	0.3352 ($\pm$0.0008)
Sigmoid	0.3387 ($\pm$0.0051)	0.3629 ($\pm$0.0063)	0.0298 ($\pm$0.0043)	0.3362 ($\pm$0.0009)

and ToxCast in Table 5 in the supplement. Our SOAP can consistently achieve improved performance when the data is extremely imbalanced. However, it sometimes fails to do so if the imbalance ratio is not too low. Clearly, the improvements from our method are higher when the imbalance ratio of labels is lower. In other words, our method is more advantageous for data with extreme class imbalance.

Insensitivity to Batch Size. We conduct experiments on CIFAR-10 and CIFAR-100 data by varying the mini-batch size for the SOAP algorithm and report results in Figure 2 (Left most). We can see that SOAP is not sensitive to the mini-batch size. This is consistent with our theory. In contrast, many previous methods for AP maximization are sensitive to the mini-batch size [47, 48, 6].

Convergence Speed. We report the convergence curves of different methods for maximizing AUPRC or AP in Figure 1 on different datasets. We can see that the proposed SOAP algorithms converge much faster than other baseline methods.

More Surrogate Losses. To verify the generality of SOAP, we evaluate the performance of SOAP with two more different surrogate loss functions $\ell(\mathbf{w}; \mathbf{x}_s, \mathbf{x}_i)$ as a surrogate function of the indicator $\mathbf{I}(h_{\mathbf{w}}(\mathbf{x}_s) \geq h_{\mathbf{w}}(\mathbf{x}_i))$, namely, the logistic loss, $\ell(\mathbf{w}; \mathbf{x}_s, \mathbf{x}_i) = -\log \frac{1}{1+\exp(-c(\ell(h_{\mathbf{w}}(\mathbf{x}_i) - h_{\mathbf{w}}(\mathbf{x}_s)))}$, and the sigmoid loss, $\ell(\mathbf{w}; \mathbf{x}_s, \mathbf{x}_i) = \frac{1}{1+\exp(c(\ell(h_{\mathbf{w}}(\mathbf{x}_i) - h_{\mathbf{w}}(\mathbf{x}_s)))}$ where c is a hyperparameter. We tune $c \in \{1, 2\}$ in our experiments. We conduct experiments on CIFAR10, CIFAR100 following the experimental setting in Section 4.1 for the image data. For the graph data, we conduct experiments on HIV, MUV data following the experimental setting in Section 4.2. We report the results in Table 4. We can observe that SOAP has similar results with different surrogate loss functions.

Consistency. Finally, we show the consistency between the Surrogate Objective -$P(\mathbf{w})$ and AP by plotting the convergence curves on different datasets in Figure 2 (Right two). It is obvious two see the consistency between our surrogate objective and the true AP.

5 Conclusions and Outlook

In this work, we have proposed a stochastic method to optimize AUPRC that can be used in deep learning for tackling highly imbalanced data. Our approach is based on maximizing the averaged precision, and we cast the objective into a sum of coupled compositional functions. We proposed efficient adaptive and non-adaptive stochastic algorithms with provable convergence guarantee to compute the solutions. Extensive experimental results on graph and image datasets demonstrate that our proposed method can achieve promising results, especially when the class distribution is highly imbalanced. One limitation of SOAP is its convergence rate is still slow. In the future, we will consider to improve the convergence rate to address the limitation of the present work.

Acknowledgments

We thank Bokun Wang for discussing the proofs, and thank anonymous reviewers for constructive comments. Q.Q contributed to the algorithm design, analysis, and experiments under supervision of T.Y. Y.L and Z.X contributed to the experiments under supervision of S.J. Q.Q and T.Y were partially supported by NSF Career Award #1844403, NSF Award #2110545 and NSF Award #1933212. Y.L, Z.X and S.J were partially supported by NSF IIS-1955189.

References

[1] Balasubramanian, K., Ghadimi, S., and Nguyen, A. Stochastic multi-level composition optimization algorithms with level-independent convergence rates. *CoRR*, abs/2008.10526, 2020.

[2] Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[3] Boyd, K., Eng, K. H., and Page, C. D. Area under the precision-recall curve: Point estimates and confidence intervals. In Blockeel, H., Kersting, K., Nijssen, S., and Zelezny, F. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 451–466, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[4] Brown, A., Xie, W., Kalogeiton, V., and Zisserman, A. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pp. 677–694. Springer, 2020.

[5] Burges, C., Ragno, R., and Le, Q. Learning to rank with nonsmooth cost functions. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

[6] Cakir, F., He, K., Xia, X., Kulis, B., and Sclaroff, S. Deep metric learning to rank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] 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*, pp. 1567–1578, 2019.

[8] Chen, K., Li, J., Lin, W., See, J., Wang, J., Duan, L., Chen, Z., He, C., and Zou, J. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Chen, K., Lin, W., See, J., Wang, J., Zou, J., et al. Ap-loss for accurate one-stage object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[10] Chen, T., Sun, Y., and Yin, W. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

[11] Chen, W., Liu, T.-Y., Lan, Y., Ma, Z., and Li, H. Ranking measures and loss functions in learning to rank. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pp. 315–323, Red Hook, NY, USA, 2009. Curran Associates Inc. ISBN 9781615679119.

[12] Cortes, C. and Mohri, M. Auc optimization vs. error rate minimization. In Thrun, S., Saul, L. K., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems 16*, pp. 313–320. 2004.

[13] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

[14] Davis, J. and Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2.

[15] Eban, E., Schain, M., Mackey, A., Gordon, A., Saurous, R. A., and Elidan, G. Scalable learning of non-decomposable objectives. In *International Conference on Arti cial Intelligence and Statistics (AISTATS)*, 2017.

[16] Gao, H. and Ji, S. Graph u-nets. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2083–2092. PMLR, 09–15 Jun 2019.

[17] Gao, H., Wang, Z., and Ji, S. Large-scale learnable graph convolutional networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 1416–1424, New York, NY, USA, 2018. Association for Computing Machinery.

[18] Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. One-pass auc optimization. In *ICML (3)*, pp. 906–914, 2013.

[19] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, International Convention Centre, Sydney, Australia, 2017.

[20] Goadrich, M., Oliphant, L., and Shavlik, J. Gleaner: Creating ensembles of firstorder clauses to improve recall-precision curves. In *Machine Learning*, pp. 2006, 2006.

[21] Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., and Yang, T. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3864–3874, 2020.

[22] Guo, Z., Yuan, Z., Yan, Y., and Yang, T. Fast objective and duality gap convergence for non-convex strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

[23] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[24] Henderson, P. and Ferrari, V. End-to-end training of object class detectors for mean average precision. In *Computer Vision – ACCV 2016*, pp. 198–213. Springer International Publishing, 2017. doi: 10.1007/978-3-319-54193-8_13. URL `https://doi.org/10.1007%2F978-3-319-54193-8_13`.

[25] Herschtal, A. and Raskutti, B. Optimising area under the ROC curve using gradient descent. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pp. 49, 2004.

[26] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *Proceddings of the 7th international conference on learning representations*, 2019.

[27] Huo, Z., Gu, B., Liu, J., and Huang, H. Accelerated method for stochastic composition optimization with nonsmooth regularization. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 3287–3294, 2018.

[28] Joachims, T. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine learning*, pp. 377–384, 2005.

[29] Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*, 2017.

[30] Lian, X., Wang, M., and Liu, J. Finite-sum composition optimization via variance reduced gradient descent. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1159–1167, 2017.

[31] Lin, T., Fan, C., Wang, M., and Jordan, M. I. Improved oracle complexity for stochastic compositional variance reduced gradient. *CoRR*, abs/1806.00458, 2018.

[32] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[33] Liu, L., Liu, J., Hsieh, C., and Tao, D. Stochastically controlled stochastic gradient for the convex and non-convex composition problem. *CoRR*, abs/1809.02505, 2018.

[34] Liu, L., Liu, J., and Tao, D. Dualityfree methods for stochastic composition optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 30(4):1205–1217, 2019.

[35] Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic auc maximization with $o(1/n)$-convergence rate. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3189–3197. PMLR, 2018.

[36] Liu, M., Yuan, Z., Ying, Y., and Yang, T. Stochastic auc maximization with deep neural networks. In *International Conference on Learning Representations*, 2020.

[37] Metzler, D. and Croft, W. B. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, 2005.

[38] Metzler, D. and Croft, W. B. A markov random field model for term dependencies. In *Proceedings of the 28th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479, 2005.

[39] Mohapatra, P., Jawahar, C., and Kumar, M. P. Efficient optimization for average precision svm. In *Advances in Neural Information Processing Systems*, 2014.

[40] Mohapatra, P., Rolinek, M., Jawahar, C. V., Kolmogorov, V., and Kumar, M. Efficient optimization for rank-based loss functions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3693–3701, 2018.

[41] Narasimhan, H., Cotter, A., and Gupta, M. Optimizing generalized rate metrics with three players. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/3ce257b311e5acf849992f5a675188e8-Paper.pdf`.

[42] Natole, M., Ying, Y., and Lyu, S. Stochastic proximal algorithms for auc maximization. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3710–3719. PMLR, 2018.

[43] Oksuz, K., Cam, B. C., Akbas, E., and Kalkan, S. A ranking-based, balanced loss function unifying classification and localisation in object detection. In *Advances in Neural Information Processing Systems*, 2020.

[44] Qi, Q. Soap code for reproducing results. https://github.com/Optimization-AI, 2021.

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

[46] Qi, Q., Guo, Z., Xu, Y., Jin, R., and Yang, T. An online method for a class of distributionally robust optimization with non-convex objectives. In *Proceedings of Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[47] Qin, T., Liu, T.-Y., and Li, H. A general approximation framework for direct optimization of information retrieval measures. Technical Report MSR-TR-2008-164, November 2008.

[48] Rolinek, M., Musil, V., Paulus, A., Vlastelica, M., Michaelis, C., and Martius, G. Optimizing rank-based metrics with blackbox differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[49] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *arXiv preprint arXiv:2008.07360*, 2020.

[50] Song, Y., Schwing, A., Richard, and Urtasun, R. Training deep neural networks via direct loss minimization. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2169–2177, New York, New York, USA, 20–22 Jun 2016. PMLR.

[51] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackerman, Z., et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[52] Wang, M., Fang, E. X., and Liu, H. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2): 419–449, 2017.

[53] Wang, M., Liu, J., and Fang, E. X. Accelerating stochastic composition optimization. *Journal Machine Learning Research*, 18:105:1–105:23, 2017.

[54] Wang, Z., Liu, M., Luo, Y., Xu, Z., Xie, Y., Wang, L., Cai, L., Qi, Q., Yuan, Z., Yang, T., and Ji, S. Advanced graph and sequence neural networks for molecular property prediction and drug discovery, 2021.

[55] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9 (2):513–530, 2018.

[56] Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *7th International Conference on Learning Representations*, 2019.

[57] Yan, Y., Xu, Y., Lin, Q., Liu, W., and Yang, T. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[58] Ying, Y., Wen, L., and Lyu, S. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.

[59] Yu, Y. and Huang, L. Fast stochastic variance reduced ADMM for stochastic composition optimization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3364–3370, 2017.

[60] Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 2020.

[61] Yue, Y., Finley, T., Radlinski, F., and Joachims, T. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pp. 271–278, New York, NY, USA, 2007. Association for Computing Machinery.

[62] Zhang, J. and Xiao, L. A composite randomized incremental gradient method. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 7454–7462, 2019.

[63] Zhao, P., Hoi, S. C. H., Jin, R., and Yang, T. Online auc maximization. In *ICML*, pp. 233–240, 2011.

A Additional Experimental Results

We include the results about effect of imbalance ratio in Table 5, and the full results using three networks on MIT AICURES data in Table 6, and PR curves of final models on CIFAR10, CIFAR100 data in Figure 3.

Table 5: Test AUPRC on task 0 and task 2 of the Tox21 dataset and task 12 and task 8 of the ToxCast dataset with three graph neural network models.

	Tox21 Task 0 (Imbalance Ratio = 4.14%)		
Method	GINE	MPNN	ML-MPNN
CE	0.4829 (± 0.0123)	0.5002 (± 0.0054)	0.4868 (± 0.0048)
CB-CE	0.4861 (± 0.0113)	0.4931 (± 0.0068)	0.4772 (± 0.0033)
Focal	0.4874 (± 0.0148)	0.4865 (± 0.0067)	0.4769 (± 0.0134)
LDAM	0.5093 (± 0.0096)	0.4823 (± 0.0084)	0.4709 (± 0.0084)
AUC-M	0.4356 (± 0.0127)	0.4428 (± 0.0121)	0.4632 (± 0.0121)
SmoothAP	0.3764 (± 0.0053)	0.4504 (± 0.0089)	0.4634 (± 0.0064)
FastsAP	0.0668 (± 0.0061)	0.2358 (± 0.0093)	0.0341 (0.0065)
MinMax (Adam)	0.5066 (± 0.0111)	0.4940 (± 0.0134)	0.4947 (± 0.0053)
SOAP (Adam)	**0.5276 (± 0.0099)**	**0.5211 (± 0.0089)**	**0.5093 (± 0.0067)**
	Tox21 Task 2 (Imbalance Ratio = 12.00%)		
Method	GINE	MPNN	ML-MPNN
CE	0.5918 (± 0.0063)	0.6023 (± 0.0087)	0.5796 (± 0.0071)
CB-CE	0.5538 (± 0.0087)	0.5811 (± 0.0095)	0.5855 (± 0.0069)
Focal	0.5594 (± 0.0069)	0.6018 (± 0.0083)	0.5555 (± 0.0025)
LDAM	0.5369 (± 0.0065)	0.5991 (± 0.0067)	0.6014 (± 0.0051)
AUC-M	0.5832 (± 0.0067)	0.6117 (± 0.0085)	0.5987 (± 0.0060)
SmoothAP	0.5852 (± 0.0045)	0.6210 (± 0.0069)	0.4858 (± 0.0061)
FastAP	0.5605 (± 0.0000)	0.5605 (± 0.0000)	0.5605 (± 0.0000)
MinMax (Adam)	0.5623 (± 0.0041)	0.5977 (± 0.0045)	0.5079 (± 0.0083)
SOAP (Adam)	**0.6172 (± 0.0051)**	**0.6333 (± 0.0160)**	**0.6196** (± 0.0165)
	ToxCast Task 12 (Imbalance Ratio = 2.97%)		
Method	GINE	MPNN	ML-MPNN
CE	0.0201 (± 0.0031)	0.0268 (± 0.0031)	0.0124 (± 0.0031)
CB-CE	0.0385 (± 0.0042)	0.0278 (± 0.0073)	0.0104 (± 0.0029)
Focal	0.0333 (± 0.0052)	0.0294 (± 0.0043)	0.0122 (± 0.0024)
LDAM	0.0217 (± 0.0042)	0.0298 (± 0.0059)	0.0179 (± 0.0019)
AUC-M	0.0333 (± 0.0024)	0.0454 (± 0.0047)	0.0089 (± 0.0023)
SmoothAP	0.227 (± 0.0023)	0.0208 (± 0.0041)	0.0079 (± 0.0034)
FastAP	0.0052 (± 0.0048)	0.0052 (± 0.0038)	0.0153 (± 0.0013)
MinMax (Adam)	0.0223 (± 0.0033)	0.0313 (± 0.0061)	0.0151 (± 0.0023)
SOAP (Adam)	**0.0374 (± 0.0025)**	**0.0601 (± 0.0059)**	**0.0181 (± 0.0023)**
	ToxCast Task 8 (Imbalance Ratio = 8.67%)		
Method	GINE	MPNN	ML-MPNN
CE	0.2071 (± 0.0121)	0.1101 (± 0.0049)	0.0923 (± 0.0027)
CB-CE	0.2089 (± 0.0051)	0.1349 (± 0.0109)	0.0734 (± 0.0078)
Focal	0.2011 (± 0.0034)	0.1223 (± 0.0113)	0.0792 (± 0.0082)
LDAM	0.1071 (± 0.0101)	0.1062 (± 0.0104)	0.0934 (± 0.0125)
AUC-M	0.0662 (± 0.098)	0.1258 (± 0.0132)	0.0979 (± 0.0096)
SmoothAP	0.0911 (± 0.0123)	0.1073 (± 0.0011)	0.0987 (± 0.0049)
FastAP	0.0999 (± 0.0211)	0.1037 (± 0.0071)	0.0932 (± 0.0028)
MinMax (Adam)	0.1381(± 0.0076)	0.1173 (± 0.0092)	0.0903 (± 0.0031)
SOAP (Adam)	**0.2561 (± 0.0196)**	**0.1875 (± 0.0124)**	**0.1107 (± 0.0807)**

Table 6: The test AUPRC values on the MIT AICURES dataset with three graph neural network models. We report the average AUPRC and standard deviation (within brackets) from 3 independent runs over 3 different train/validation/test splits.

Method	GINE	MPNN	ML-MPNN
CE	0.5037 ($\pm$ 0.0718)	0.6282 ($\pm$ 0.0634)	0.6101 ($\pm$ 0.1276)
CB-CE	0.5655 ($\pm$ 0.0453)	0.6308 ($\pm$ 0.0263)	0.4903 ($\pm$ 0.1507)
Focal	0.5143 ($\pm$ 0.1062)	0.5875 ($\pm$ 0.0774)	0.4718 ($\pm$ 0.0691)
LDAM	0.5236 ($\pm$ 0.0551)	0.6489 ($\pm$ 0.0556)	0.6725 ($\pm$ 0.0594)
AUC-M	0.5149 ($\pm$ 0.0748)	0.5542 ($\pm$ 0.0474)	0.4429 ($\pm$ 0.0486)
SmothAP	0.2899 ($\pm$ 0.0220)	0.4081 ($\pm$ 0.0352)	0.4212 ($\pm$ 0.0507)
FastAP	0.4777 ($\pm$ 0.0896)	0.4518 ($\pm$ 0.1495)	0.5174 ($\pm$ 0.0150)
MinMax	0.5292 ($\pm$ 0.0330)	0.5774 ($\pm$ 0.0468)	0.5832 ($\pm$ 0.1080)
SOAP	**0.6639 ($\pm$ 0.0515)**	**0.6547 ($\pm$ 0.0616)**	**0.6503 ($\pm$ 0.0532)**

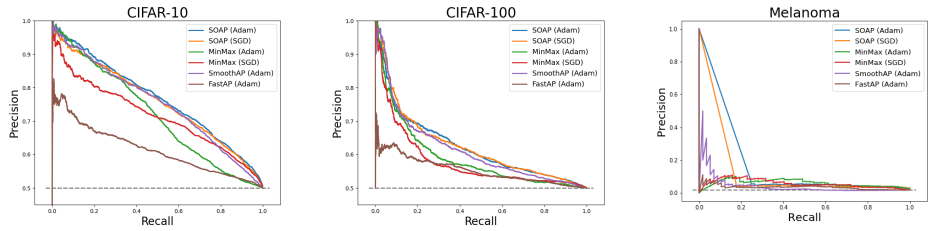

Figure 3: Precision-Recall curves of different methods on test dataset of CIFAR10, CIFAR100 and Melanoma datasets. The gray dashed lines are the random classifiers on test data sets whose AUPRC equals to the ratio between positive samples and all samples n_+/n on every data set, respectively.

B Analysis of SOAP (SGD-style)

In the following, we abuse the notations $g_i(\mathbf{w}) = g_{\mathbf{x}_i}(\mathbf{w}) \in \mathbb{R}^2$ and $\mathbf{u}_i = \mathbf{u}_{\mathbf{x}_i} = ([\mathbf{u}_{\mathbf{x}_i}]_1, [\mathbf{u}_{\mathbf{x}_i}]_2)$. We use $\mathbf{u}_{i_t}$ to denote the updated vector at the t-th iteration for the sampled i_t-th positive data.

B.1 Proof of Theorem 1

Proof. By combining Lemma 3 and Lemma 2, we have:

$$\begin{aligned}
\frac{\alpha}{2}\mathbb{E}[\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2] &\leq \mathbb{E}[\sum_{t}(P(\mathbf{w}_t) - P(\mathbf{w}_{t+1}))] + \frac{\alpha C_1}{2}\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 T C_2 \\
&\leq \mathbb{E}[\sum_{t}(P(\mathbf{w}_t) - P(\mathbf{w}_{t+1}))] + \frac{\alpha C_1}{2}\Big\{\frac{n_+ V}{\gamma} + 2\gamma V T + 2\frac{n_+^2\alpha^2 T C_3}{\gamma^2}\Big\} + \alpha^2 T C_2 \\
&\leq \mathbb{E}_t[P(\mathbf{w}_1)] - \mathbb{E}_t[P(\mathbf{w}_{t+1})] + \frac{\alpha C_1}{2}\Big\{\frac{n_+ V}{\gamma} + 2\gamma V T + 2\frac{n_+^2\alpha^2 T C_3}{\gamma^2}\Big\} + \alpha^2 T C_2
\end{aligned}$$

Then by set $\alpha = \frac{1}{n_+^{2/5}T^{3/5}}$, $\gamma = \frac{n_+^{2/5}}{T^{2/5}}$, and multiply $\frac{2}{\alpha T}$ on both sides of above equation,

$$\begin{aligned}
\frac{1}{T}\mathbb{E}[\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2] &\leq \frac{2\Delta_1}{T\alpha} + C_1\Big\{\frac{n_+ V}{\gamma T} + 2\gamma V + 2\frac{n_+^2\alpha^2 C_3}{\gamma^2}\Big\} + \alpha C_2 \\
&\leq \frac{2\Delta_1 n_+^{2/5}}{T^{2/5}} + C_1\Big\{\frac{n_+^{3/5} V}{T^{3/5}} + 2\frac{n_+^{2/5}}{T^{2/5}} + 2\frac{n_+^{2/5} C_3}{T^{2/5}}\Big\} + \frac{C_2}{n_+^{2/5}T^{3/5}} \\
&\leq O(\frac{n_+^{2/5}}{T^{2/5}})
\end{aligned}$$

where the last inequality is due to $T \geq n_+$ and O compresses constant numbers. We finish the proof. □

B.2 Proof of Lemma 1

Proof of Lemma 1. We first prove the second part that $g_i(\mathbf{w}) \in \Omega$. Due to the definition of $g_i(\mathbf{w}) = \mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[g(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)] = \mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)\mathbf{I}(y_j = 1), \ell(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)]$, and the Assumption 1, it is obvious to see that $0 \leq [g_i(\mathbf{w})]_1 \leq M$ and $M \geq [g_i(\mathbf{w})]_2 \geq C/n$ for all i, i.e., $g_i(\mathbf{w}) \in \Omega$. Next, we prove the smoothness of $P(\mathbf{w})$. To this end, we need to use the following Lemma 4 and the proof will be presented after Lemma 1.

Lemma 4. *Let $L_f = 4(u_0 + M)/u_0^3, C_f = (u_0 + M)/u_0^2, L_g = \sqrt{2}L_l, C_g = \sqrt{2}C_l$, then $f(\mathbf{u})$ is a L_f -smooth, C_f-Lipschitz continuous function for any $\mathbf{u} \in \Omega$, and $\forall\, i \in [1, \cdots n]$, g_i is a L_g-smooth, C_g-Lipschitz continuous function.*

Since $P(\mathbf{w}) = \frac{1}{n_+}\sum_{\mathbf{x}_i \in \mathcal{D}_+} f(g_i(\mathbf{w}))$. We first show $P_i(\mathbf{w}) = f(g_i(\mathbf{w}))$ is smooth. To see this,

$$\begin{aligned}
&\|\nabla P_i(\mathbf{w}) - \nabla P_i(\mathbf{w}')\| = \|\nabla g_i(\mathbf{w})^\top \nabla f(g_i(\mathbf{w})) - \nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w}'))\| \\
&\leq \|\nabla g_i(\mathbf{w})^\top \nabla f(g_i(\mathbf{w})) - \nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w}))\| \\
&+ \|\nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w})) - \nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w}'))\| \\
&\leq C_f L_g \|\mathbf{w} - \mathbf{w}'\| + C_g L_f C_g \|\mathbf{w} - \mathbf{w}'\| = (C_f L_g + L_f C_g^2)\|\mathbf{w} - \mathbf{w}'\|.
\end{aligned}$$

Hence $P(\mathbf{w})$ is also $L = (C_f L_g + L_f C_g^2)$-smooth.

□

B.3 Proof of Lemma 4

Proof of Lemma 4. According to the definition, we have

$$f(\mathbf{u}) = \frac{-[\mathbf{u}]_1}{[\mathbf{u}]_2} \quad , \nabla_{\mathbf{u}} f(\mathbf{u}) = \left(\frac{-1}{[\mathbf{u}]_2}, \frac{[\mathbf{u}]_1}{([\mathbf{u}]_2)^2}\right)^\top, \quad \nabla^2_{\mathbf{u}} f(\mathbf{u}) = \begin{pmatrix} 0, \frac{1}{([\mathbf{u}]_2)^2} \\ \frac{1}{([\mathbf{u}]_2)^2}, -\frac{2[\mathbf{u}]_1}{([\mathbf{u}]_2)^3} \end{pmatrix} \tag{9}$$

Due to the assumption that $\ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)$ is a L_l-smooth, C_l-Lipschitz continuous function, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{w}} g_i(\mathbf{w})\|^2 \leq 2\|\frac{1}{n}\sum_{j=1}^n \nabla \ell_{\mathbf{w}}(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)\|^2 \leq 2C_l^2 = C_g^2 \\
&\|\nabla_{\mathbf{w}} g_i(\mathbf{w}) - \nabla_{\mathbf{w}} g(\mathbf{w}')\|^2 \leq \|\frac{1}{n}\sum_{j=1}^n \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i) - \frac{1}{n}\sum_{j=1}^n \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)\|^2 \\
&+ \|\frac{1}{n}\sum_{j=1}^n \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)\mathbf{I}(y_j = 1) - \frac{1}{n}\sum_{j=1}^n \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_j, \mathbf{x}_i)\mathbf{I}(y_j = 1)\|^2 \leq 2L_l^2 = L_g^2 \qquad (10) \\
&\|\nabla f(\mathbf{u})\| \leq \sqrt{\frac{1}{[\mathbf{u}]_2^2} + \frac{[\mathbf{u}]_1^2}{[\mathbf{u}]_2^4}} \leq \frac{u_0 + M}{u_0^2} = C_f \\
&\|\nabla^2 f(\mathbf{u})\| \leq \sqrt{\frac{2}{[\mathbf{u}]_2^4} + 4\frac{[\mathbf{u}]_1^2}{[\mathbf{u}]_2^6}} \leq \frac{4(u_0 + M)}{u_0^3} = L_f
\end{aligned}$$

We finish the proof of Lemma 4. □

B.4 Proof of Lemma 2

Proof of Lemma 2. To make the proof clear, we write $\nabla g_{i_t}(\mathbf{w}; \xi) = \nabla g(\mathbf{w}_t; \xi, \mathbf{x}_{i_t}), \xi \sim \mathcal{D}$. Let $\mathbf{u}_{i_t}$ denote the updated $\mathbf{u}$ vector at the t-th iteration for the selected positive data i_t.

$$\begin{aligned}
&P(\mathbf{w}_{t+1}) - P(\mathbf{w}_t) \leq \nabla P(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= -\alpha \|\nabla P(\mathbf{w}_t)\|^2 + \alpha \nabla P(\mathbf{w}_t)^\top (\nabla P(\mathbf{w}_t) - \nabla g_{i_t}^\top(\mathbf{w}_t; \xi) \nabla f(\mathbf{u}_{i_t})) + \frac{\alpha^2 \|G(\mathbf{w}_t)\|^2 L}{2} \\
&\leq -\alpha \|\nabla P(\mathbf{w}_t)\|^2 + \alpha \nabla P(\mathbf{w}_t)^\top (\nabla P(\mathbf{w}_t) - \nabla g_{i_t}^\top(\mathbf{w}_t; \xi) \nabla f(\mathbf{u}_{i_t})) + \alpha^2 C_2
\end{aligned}$$

where $C_2 = \|G(\mathbf{w}_t)\|^2 L/2 \leq C_g^2 C_f^2 L/2$.

Taking expectation on both sides, we have

$$\begin{aligned}\mathbb{E}_t[P(\mathbf{w}_{t+1})] &\leq \mathbb{E}_t[P(\mathbf{w}_t) + \nabla P(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\ &= \mathbb{E}_t[P(\mathbf{w}_t) - \alpha\|\nabla P(\mathbf{w}_t)\|^2 + \alpha\nabla P(\mathbf{w}_t)^\top(\nabla P(\mathbf{w}_t) - \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t}))] + \alpha^2 C_2 \\ &= P(\mathbf{w}_t) - \alpha\|\nabla P(\mathbf{w}_t)\|^2 + \alpha\nabla P(\mathbf{w}_t)^\top(\mathbb{E}_t[\nabla P(\mathbf{w}_t) - \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t})]) + \alpha^2 C_2\end{aligned}$$

where $\mathbb{E}_t$ means taking expectation over i_t, ξ given $\mathbf{w}_t$.
Noting that $\nabla P(\mathbf{w}_t) = \mathbb{E}_{i_t,\xi}[\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t))]$, where i_t and ξ are independent.

$$\begin{aligned}&\mathbb{E}_t[P(\mathbf{w}_{t+1})] - P(\mathbf{w}_t) \\ &\leq -\alpha\|\nabla P(\mathbf{w}_t)\|^2 + \alpha\nabla P(\mathbf{w}_t)^\top(\mathbb{E}_t[\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t))] - \mathbb{E}_t[\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t})]) + \alpha^2 C_2 \\ &= -\alpha\|\nabla P(\mathbf{w}_t)\|^2 + \mathbb{E}_t[\alpha\nabla P(\mathbf{w}_t)^\top(\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t)) - \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t}))] + \alpha^2 C_2 \\ &\overset{(a)}{\leq} -\alpha\|\nabla P(\mathbf{w}_t)\|^2 + \mathbb{E}_t[\frac{\alpha}{2}\|\nabla P(\mathbf{w}_t)\|^2 + \frac{\alpha}{2}\|\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t)) - \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t}))\|^2 + \alpha^2 C_2 \\ &\overset{(b)}{\leq} -\alpha\|\nabla P(\mathbf{w}_t)\|^2 + \mathbb{E}_t[\frac{\alpha}{2}\|\nabla P(\mathbf{w}_t)\|^2 + \frac{\alpha C_1}{2}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2 + \alpha^2 C_2 \\ &= -(\alpha - \frac{\alpha}{2})\|\nabla P(\mathbf{w}_t)\|^2 + \frac{\alpha C_1}{2}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 C_2\end{aligned}$$

where the equality (a) is due to $ab \leq a^2/2 + b^2/2$ and the inequality (b) uses the factor $\|\nabla g_{i_t}(\mathbf{w}_t;\xi)\| \leq C_l$ and ∇f is L_f-Lipschitz continuous for $\mathbf{u}, \mathbf{g}_i(\mathbf{w}) \in \Omega$ and $C_1 = C_l^2 C_f^2$. Hence we have,

$$\frac{\alpha}{2}\|\nabla P(\mathbf{w}_t)\|^2 \leq P(\mathbf{w}_t) - \mathbb{E}_t[P(\mathbf{w}_{t+1})] + \frac{\alpha C_1}{2}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 C_2$$

Taking summation and expectation over all randomness, we have

$$\frac{\alpha}{2}\mathbb{E}[\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2] \leq \mathbb{E}[\sum_t (P(\mathbf{w}_t) - P(\mathbf{w}_{t+1}))] + \frac{\alpha C_1}{2}\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 C_2 T$$

□

B.5 Proof of Lemma 3

Let i_t denote the selected positive data i_t at t-th iteration. We will divide $\{1, \ldots, T\}$ into n_+ groups with the i-th group given by $\mathcal{T}_i = \{t_1^i, \ldots, t_k^i \ldots, \}$, where t_k^i denotes the iteration that the i-th positive data is selected at the k-th time for updating $\mathbf{u}$. Let us define $\phi(t) : [T] \to [n_+] \times [T]$ that maps the selected data into its group index and within group index, i.e, there is an one-to-one correspondence between index t and selected data i and its index within $\mathcal{T}_i$. Below, we use notations a_i^k to denote $a_{t_k^i}$. Let $T_i = |\mathcal{T}_i|$. Hence, $\sum_{i=1}^{n_+} T_i = T$.

Proof of Lemma 3. To prove Lemma 3, we first introduce another lemma that establishes a recursion for $\|\mathbf{u}_{i_t} - g_{i_t}(\mathbf{w}_t)\|^2$, whose proof is presented later.

Lemma 5. *By the updates of SOAP Adam-style or SGD-style with $\mathcal{B}_+ = 1$, the following equation holds for $\forall\, t \in 1, \cdots, T$*

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{u}_{i_t} - g_{i_t}(\mathbf{w}_t)\|^2] &\overset{\phi(t)}{=} \mathbb{E}_t[\|\mathbf{u}_i^k - g_i(\mathbf{w}_i^k)\|^2] \\ &\leq (1-\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2 + \gamma^2 V + \gamma^{-1}\alpha^2 n_+^2 C_3\end{aligned} \tag{11}$$

where $\mathbb{E}_t$ denotes the conditional expectation conditioned on history before t_{k-1}^i.

Then, by mapping every i_t to its own group and make use of Lemma 5, we have

$$\mathbb{E}[\sum_{k=0}^{K_i}\|\mathbf{u}_i^k - g_i^k(\mathbf{w}_i^k)\|^2] \leq \mathbb{E}\left[\frac{[\|\mathbf{u}_i^0 - g_i(\mathbf{w}_i^0)\|^2]}{\gamma} + \gamma V T_i + \gamma^{-2} n_+^2 C_3 \alpha^2 T_i\right] \tag{12}$$

where $\mathbf{u}_i^0$ is the initial vector for $\mathbf{u}_i$, which can be computed by a mini-batch averaging estimator of $g_i(\mathbf{w}_0)$. Thus

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t)-\mathbf{u}_{i_t}\|^2] &\overset{\phi(t)}{=} \mathbb{E}[\sum_{i=1}^{n_+}\sum_{k=0}^{K_i}\|\mathbf{u}_i^k-g_i^k(\mathbf{w}_i^k)\|^2]\\ &\leq \sum_{i=1}^{n_+}\Big\{\frac{[\|\mathbf{u}_i^0-g_i^0(\mathbf{w}_i^0)\|^2]}{\gamma}+\gamma V\mathbb{E}[T_i]+\gamma^{-2}n_+^2C_3\alpha^2\mathbb{E}[T_i]\Big\}\\ &\leq \frac{n_+V}{\gamma}+\gamma VT+\frac{n_+^2\alpha^2TC_3}{\gamma^2}\end{aligned}$$

□

B.6 Proof of Lemma 5

Proof. We first introduce the following lemma, whose proof is presented later.

Lemma 6. *Suppose the sequence generated in the training process using the positive sample i is $\{\mathbf{w}_{i_1}^i,\mathbf{w}_{i_2}^i,..\ ..,\mathbf{w}_{i_{T_i}}^i\}$, where $0<i_1<i_2<\cdots<i_{T_i}\leq T$, then $\mathbb{E}_{|i_k}[i_{k+1}-i_k]\leq n_+$, and, $\mathbb{E}_{|i_k}[(i_{k+1}-i_k)^2]\leq 2n_+^2, \forall k$.*

Define $\widetilde{g}_{i_t}(\mathbf{w}_t)=g(\mathbf{w}_t,\xi,\mathbf{x}_{i_t})$. Let $\prod_\Omega(\cdot):\mathbb{R}^2\to\Omega$ denotes the projection operator. By the updates of $\mathbf{u}_{i_t}$, we have $\mathbf{u}_{i_t}=\mathbf{u}_i^k=\prod_\Omega[(1-\gamma)\mathbf{u}_i^{k-1}+\gamma\tilde{g}_{i_t}(\mathbf{w}_t)]$.

$$\begin{aligned}&\mathbb{E}_t[\|\mathbf{u}_{i_t}-g_{i_t}(\mathbf{w}_t)\|^2] \overset{\phi(t)}{=} \mathbb{E}[\|\mathbf{u}_i^k-g_i(\mathbf{w}_i^k)\|^2]\\ &=\mathbb{E}_t[\|\prod_\Omega((1-\gamma)\mathbf{u}_i^{k-1}+\gamma\widetilde{g}_i(\mathbf{w}_i^k))-\prod_\Omega(g_i(\mathbf{w}_t))\|^2]\\ &\leq \mathbb{E}_t[\|((1-\gamma)\mathbf{u}_i^{k-1}+\gamma\widetilde{g}_i(\mathbf{w}_i^k)-g_i(\mathbf{w}_t)\|^2]\\ &\leq \mathbb{E}_t[\|((1-\gamma)(\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1}))+\gamma(\widetilde{g}_i(\mathbf{w}_i^k)-g_i(\mathbf{w}_i^k))+(1-\gamma)(g_i(\mathbf{w}_i^{k-1})-g_i(\mathbf{w}_i^k))\|^2]\\ &\leq \mathbb{E}_t[\|((1-\gamma)(\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1}))+(1-\gamma)(g_i(\mathbf{w}_i^{k-1})-g_i(\mathbf{w}_i^k))\|^2]+\gamma^2V\\ &\leq [(1-\gamma)^2(1+\gamma)\|\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1})\|^2]+\gamma^2V+\frac{(1+\gamma)(1-\gamma)^2}{\gamma}C_g\mathbb{E}[\|\mathbf{w}_i^k-\mathbf{w}_i^{k-1}\|^2]\\ &\leq [(1-\gamma)\|\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1})\|^2]+\gamma^2V+\gamma^{-1}\alpha^2C_g\mathbb{E}_t[\|\sum_{t=t_{k-1}^i}^{t_k^i-1}\nabla g_{i_t}(\mathbf{w}_t;\xi)\nabla f(\mathbf{u}_{i_t})\|^2]\\ &\leq [(1-\gamma)\|\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1})\|^2]+\gamma^2V+\gamma^{-1}\alpha^2C_g\mathbb{E}_t[(t_k^i-t_{k-1}^i)^2]C_g^2C_f^2)]\\ &\overset{(a)}{\leq} \mathbb{E}[(1-\gamma)\|\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1})\|^2]+\gamma^2V+2\gamma^{-1}\alpha^2n_+^2C_g^3C_f^2\\ &\leq [(1-\gamma)\|\mathbf{u}_i^{k-1}-g_i(\mathbf{w}_i^{k-1})\|^2]+\gamma^2V+\gamma^{-1}\alpha^2n_+^2C_3\end{aligned}$$

where the inequality (a) is due to that $t_k^i-t_{k-1}^i$ is a geometric distribution random variable with $p=1/n_+$, i.e., $\mathbb{E}_{|t_{k-1}^i}[(t_k^i-t_{k-1}^i)^2]\leq 2/p^2=2n_+^2$, by Lemma 6. The last equality hold by defining $C_3=2C_g^3C_f^2$.

□

B.7 Proof of Lemma 6

Proof. Proof of Lemma 6. Denote the random variable $\Delta_k=i_{k+1}-i_k$ that represents the iterations that the ith positive sample has been randomly selected for the $k+1$-th time conditioned on i_k. Then Δ_k follows a Geometric distribution such that $\Pr(\Delta_k=j)=(1-p)^{j-1}p$, where $p=\frac{1}{n_+}$, $j=1,2,3,\cdots$. As a result, $\mathbb{E}[\Delta_k|i_k]=1/p=n_+$. $\mathbb{E}[\Delta_k^2|i_k]=\mathrm{Var}(\Delta_k)+\mathbb{E}[\Delta_k|i_k]^2=\frac{1-p}{p^2}+\frac{1}{p^2}\leq\frac{2}{p^2}=2n_+^2$. □

C Proof of Theorem 2 (SOAP with Adam-Style Update)

Proof. We first provide two useful lemmas, whose proof are presented later.

Lemma 7. *Assume assumption 1 holds*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \leq \alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \tag{13}$$

where d is the dimension of $\mathbf{w}$, $\eta_1 < \sqrt{\eta_2} < 1$, and $\tau := \eta_1^2/\eta_2$.

Lemma 8. *With $c = (1+(1-\eta_1)^{-1})\epsilon^{-\frac{1}{2}} C_g^2 L_f^2$, running T iterations of SOAP (Adam-style) updates, we have*

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha(1-\eta_1)(\epsilon + C_g^2 C_f^2)^{-1/2}}{2} \|\nabla P(\mathbf{w}_t)\|^2 \leq \mathbb{E}[\mathcal{L}_1] - \mathbb{E}[\mathcal{L}_{T+1}] \\
&+ 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + L\alpha^2 T d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2(1-\eta_1)^{-1}\alpha C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_0^{i'})^{-1/2}) + c\alpha \sum_{t=1}^{T} \mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{14}$$

where $\mathcal{L}_{t+1} = P(\mathbf{w}_{t+1}) - c_{t+1}\langle \nabla P(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle$.

According to Lemma 8 and plugging Lemma 3 into equation (14), we have

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha(1-\eta_1)(\epsilon + C_g^2 C_f^2)^{-1/2}}{2} \|\nabla P(\mathbf{w}_t)\|^2 \\
&\leq \mathbb{E}[\mathcal{L}_1] - \mathbb{E}[\mathcal{L}_{T+1}] + 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + L\alpha^2 d T(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2c\alpha C_g^2 C_f^2 \sum_{i'=1}^{d}(\epsilon + \hat{v}_0^{i'})^{-1/2} + c\alpha(\frac{n_+ V}{\gamma} + 2\gamma V T + \frac{2C_g n_+^2 C_3 \alpha^2 T}{\gamma^2})
\end{aligned} \tag{15}$$

Let $\eta' = (1-\eta_2)^{-1}(1-\tau)^{-1}, \eta'' = (1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1}$, and $\widetilde{\eta} = (1-\eta_1)^{-2}(1-\eta_2)^{-1}(1-\tau)^{-1}$. As $(1-\eta_1)^{-1} \geq 1, (1-\eta_2)^{-1} \geq 1$, then $\widetilde{\eta} \geq \eta'' \geq \eta' \geq 1$.

Then by rearranging terms in Equation (15), dividing $\alpha T(1+\eta_1)(\epsilon + C_g^2 C_f^2)^{-1/2}$ on both sides and suppress constants, $C_g, L_g, C_3, L, C_f, L_f, V, \epsilon$ into big O, we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \|\nabla P(\mathbf{w}_t)\|^2 &\leq \frac{1}{\alpha T(1-\eta_1)} O\Big(\mathbb{E}[\mathcal{L}_1] - \mathbb{E}[\mathcal{L}_{T+1}] + \eta'' \eta_1 \alpha^2 T d + \eta' \alpha^2 T d + \alpha \sum_{i'=1}^{d}(\epsilon + \hat{v}_0^{i'})^{-1/2} \\
&\quad + \frac{c\alpha n_+}{\gamma} + c\alpha\gamma T + \frac{c\alpha^3 n_+^2 T}{\gamma^2}\Big) \\
&\overset{(a)}{\leq} \frac{1}{\alpha T(1-\eta_1)} O\Big(\mathbb{E}[\mathcal{L}_1] - \mathbb{E}[\mathcal{L}_{T+1}] + \eta'' \eta_1 \alpha^2 T d + \eta' \alpha^2 T d + \alpha d(\epsilon + C_f C_g)^{-1/2} \\
&\quad + \frac{c\alpha n_+}{\gamma} + c\alpha\gamma T + \frac{c\alpha^3 n_+^2 T}{\gamma^2}\Big) \\
&\overset{(b)}{\leq} \frac{\widetilde{\eta}}{\alpha T} O\Big(\mathbb{E}[\mathcal{L}_1] - [\mathcal{L}_{T+1}] + (1+\eta_1)\alpha^2 T d + \alpha d + \frac{c\alpha n_+}{\gamma} + c\alpha\gamma T + \frac{c\alpha^3 n_+^2 T}{\gamma^2}\Big)
\end{aligned} \tag{16}$$

where the inequality (a) is due to $\hat{v}_0^{i'} = G^{i'}(\mathbf{w}_0)^2 \leq \|G(\mathbf{w}_0)\|^2 \leq C_f^2 C_g^2$. The last inequality (b) is due to $\widetilde{\eta} \geq \eta'' \geq \eta' \geq 1$.

Moreover, by the definition of $\mathcal{L}$ and $\mathbf{w}_0 = \mathbf{w}_1$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_1] &= P(\mathbf{w}_1) - c_1\langle \nabla P(\mathbf{w}_0), D_1 h_1\rangle \le P(\mathbf{w}_1) + c_1\|\nabla P(\mathbf{w}_0)\|\|\mathbf{w}_1 - \mathbf{w}_0\|\frac{1}{\alpha} = P(\mathbf{w}_1) \\
-\mathbb{E}[\mathcal{L}_{T+1}] &\le -P(\mathbf{w}_{T+1}) + c_{T+1}\langle \nabla P(\mathbf{w}_T), D_T h_T\rangle \\
&\le -\min_{\mathbf{w}} P(\mathbf{w}) + c_{T+1}\|\nabla P(\mathbf{w}_{t-1})\|\|\mathbf{w}_{t+1} - \mathbf{w}_t\|\frac{1}{\alpha} \\
&\overset{(a)}{\le} -\min_{\mathbf{w}} P(\mathbf{w}) + (1-\eta_1)^{-1}\alpha\sqrt{d}(1-\eta_2)^{-1/2}(1-\tau)^{-1/2} \\
&\overset{(b)}{\le} -\min_{\mathbf{w}} P(\mathbf{w}) + \widetilde{\eta}\sqrt{d}\alpha
\end{aligned} \tag{17}$$

where the inequality (a) is due to Lemma 7 and $c_{T+1} \le (1-\eta_1)^{-1}\alpha$ in equation (30). The inequality (b) is due to $(1-\eta_1)^{-1}(1-\eta_2)^{-1/2}(1-\tau)^{-1/2} \le (1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} \le \eta^{''} \le \widetilde{\eta}$. Thus $\mathbb{E}[\mathcal{L}_1] - \mathbb{E}[\mathcal{L}_{T+1}] \le P(\mathbf{w}_1) - \min_{\mathbf{w}} P(\mathbf{w}) + \widetilde{\eta}\sqrt{d}\alpha \le \Delta_1 + \widetilde{\eta}\sqrt{d}\alpha$ by combining equation (16) and (17).

Then we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla P(\mathbf{w}_t)\|^2 &\le \widetilde{\eta}O\Big(\frac{\Delta_1 + \widetilde{\eta}\sqrt{d}\alpha}{\alpha T} + (1+\eta_1)\alpha d + \frac{d}{T} + \frac{n_+ c}{T\gamma} + c\gamma + \frac{\alpha^2 n_+^2}{\gamma^2}\Big) \\
&\overset{(a)}{\le} \widetilde{\eta}O\Big(\frac{\Delta_1 n_+^{2/5}}{T^{2/5}} + \frac{\widetilde{\eta}\sqrt{d}}{T} + \frac{(1+\eta_1)d}{n_+^{2/5}T^{3/5}} + \frac{d}{T} + \frac{c n_+^{3/5}}{T^{3/5}} + 2\frac{c n_+^{2/5}}{T^{2/5}}\Big) \\
&\overset{(b)}{\le} O(\frac{n_+^{2/5}}{T^{2/5}})
\end{aligned} \tag{18}$$

The inequality (a) is due to $\gamma = \frac{n_+^{2/5}}{T^{2/5}}$, $\alpha = \frac{1}{n_+^{2/5}T^{3/5}}$. In inequality (b), we further compress the Δ_1, η_1, $\widetilde{\eta}$, c into big O and $\gamma \le 1 \to n_+^{2/5} \le T^{2/5}$.

□

C.1 Proof of Lemma 7

Proof. This proof is following the proof of Lemma 4 in [10].

Choosing $\eta_1 < 1$ and defining $\tau = \frac{\eta_1^2}{\eta_2}$, with the Adam-style (Algorithm 3) updates of SOAP that $h_{t+1} = \eta_1 h_t + (1-\eta_1)G(\mathbf{w}_t)$, we can verify for every dimension l,

$$\begin{aligned}
|h_{t+1}^l| &= |\eta_1 h_t^l + (1-\eta_1)G^l(\mathbf{w}_t)| \le \eta_1|h_t^l| + |G^l(\mathbf{w}_t)| \\
&\le \eta_1(\eta_1|h_{t-1}^l| + |G^l(\mathbf{w}_{t-1})|) + |G^l(\mathbf{w}_t)| \\
&\le \sum_{p=0}^{t}\eta_1^{t-p}|G^l(\mathbf{w}_p)| = \sum_{p=0}^{t}\sqrt{\tau}^{t-p}\sqrt{\eta_2}^{t-p}|G^l(\mathbf{w}_p)| \\
&\le \Big(\sum_{p=0}^{t}\tau^{t-p}\Big)^{\frac{1}{2}}\Big(\sum_{p=0}^{t}\eta_2^{t-p}(G^l(\mathbf{w}_p))^2\Big)^{\frac{1}{2}} \\
&\le (1-\tau)^{-\frac{1}{2}}\Big(\sum_{p=0}^{t}\eta_2^{t-p}(G^l(\mathbf{w}_t))^2\Big)^{\frac{1}{2}}
\end{aligned} \tag{19}$$

where $\mathbf{w}^l$ is the lth dimension of $\mathbf{w}$, the third inequality follows the Cauchy-Schwartz inequality. For the lth dimension of $\hat{v}$, $\hat{v}_t^l$, first we have $\hat{v}_1^l \ge (1-\eta_2)(G^l(\mathbf{w}_1)^2)$. Then since

$$\hat{v}_{t+1}^l \ge \eta_t\hat{v}_t^l + (1-\eta_2)(G^l(\mathbf{w}_t))^2$$

by induction we have

$$\hat{v}_{t+1}^l \ge (1-\eta_2)\sum_{p=0}^{t}\eta_2^{t-p}(G^l(\mathbf{w}_t))^2 \tag{20}$$

Using equation (19) and equation (20), we have

$$\begin{aligned} |h_{t+1}^l|^2 &\leq (1-\tau)^{-1}\Big(\sum_{p=0}^{t}\eta_2^{t-p}(G^l(\mathbf{w}_t))^2\Big) \\ &\leq (1-\eta_2)^{-1}(1-\tau)^{-1}\hat{v}_{t+1}^l \end{aligned} \tag{21}$$

Then follow the Adam-style update in Algorithm 3, we have

$$\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 = \alpha^2\sum_{l=1}^{d}(\epsilon+\hat{v}_{t+1}^l)^{-1}|h_{t+1}^l|^2 \leq \alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \tag{22}$$

which completes the proof. □

C.2 Proof of Lemma 8

Proof. To make the proof clear, we make some definitions the same as the proof of Lemma 2. Denote by $\nabla g_{i_t}(\mathbf{w}_t;\xi) = \nabla g(\mathbf{w}_t;\xi,\mathbf{x}_{i_t}), \xi\sim\mathcal{D}$, where i_t is a positive sample randomly generated from $\mathcal{D}_+$ at t-th iteration, and ξ is a random sample that generated from $\mathcal{D}$ at t-th iteration. It is worth to notice that i_t and ξ are independent. $\mathbf{u}_{i_t}$ denote the updated $\mathbf{u}$ vector at the t-th iteration for the selected positive data i_t.

$$\begin{aligned} P(\mathbf{w}_{t+1}) &\leq P(\mathbf{w}_t) + \nabla P(\mathbf{w}_t)^\top(\mathbf{w}_{t+1}-\mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 \\ &\leq P(\mathbf{w}_t) - \alpha\nabla P(\mathbf{w}_t)^\top(D_{t+1}h_{t+1}) + \alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1}L/2 \end{aligned}$$

where $D_{t+1} = \frac{1}{\sqrt{\epsilon I+\hat{\mathbf{v}}_{t+1}}}$, $h_{t+1} = \eta_1 h_t + (1-\eta_1)\nabla g_{i_t}^\top(\mathbf{w}_t;\xi)\nabla f(\mathbf{u}_{i_t})$ and the second inequality is due to Lemma 7. Taking expectation on both sides, we have

$$\mathbb{E}_t[P(\mathbf{w}_{t+1})] \leq P(\mathbf{w}_t)\underbrace{-\mathbb{E}_t[\nabla P(\mathbf{w}_t)^\top(D_{t+1}h_{t+1})]}_{\Upsilon}\alpha + \alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1}L$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$ implies taking expectation over i_t,ξ given $\mathbf{w}_t$. In the following analysis, we decompose Υ into three parts and bound them one by one:

$$\begin{aligned} \Upsilon &= -\langle\nabla P(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle = -\langle\nabla P(\mathbf{w}_t), D_t h_{t+1}\rangle - \langle\nabla P(\mathbf{w}_t), (D_{t+1}-D_t)h_{t+1}\rangle \\ &= -(1-\eta_1)\langle\nabla P(\mathbf{w}_t), D_t\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t})\rangle - \eta_1\langle\nabla P(\mathbf{w}_t), D_t h_t\rangle \\ &- \langle\nabla P(\mathbf{w}_t), (D_{t+1}-D_t)h_{t+1}\rangle \\ &= I_1^t + I_2^t + I_3^t \end{aligned}$$

Let us first bound I_1^t,

$$\begin{aligned} \mathbb{E}_t[I_1^t] &\overset{(a)}{=} -(1-\eta_1)\langle\nabla P(\mathbf{w}_t), \mathbb{E}_t[D_t\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(\mathbf{u}_{i_t})]\rangle \\ &= -(1-\eta_1)\langle\nabla P(\mathbf{w}_t), \mathbb{E}_t[D_t\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t))]\rangle \\ &+ (1-\eta_1)\langle\nabla P(\mathbf{w}_t), \mathbb{E}_t[D_t\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top(\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))]\rangle \\ &\leq -(1-\eta_1)\|\nabla P(\mathbf{w}_t)\|_{D_t}^2 \\ &+ (1-\eta_1)\|D_t^{-1/2}\nabla P(\mathbf{w}_t)\|\mathbb{E}_t[\|D_t^{-1/2}\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top(\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))\|] \\ &\overset{(b)}{\leq} -(1-\eta_1)\|\nabla P(\mathbf{w}_t)\|_{D_t}^2 + \frac{(1-\eta_1)\|\nabla P(\mathbf{w}_t)\|_{D_t}^2}{2} \\ &+ \frac{(1-\eta_1)\mathbb{E}_t[\|D_t^{-1/2}\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top(\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))\|^2]}{2} \\ &\leq -\frac{(1-\eta_1)}{2}\|\nabla P(\mathbf{w}_t)\|_{D_t}^2 + \frac{(1-\eta_1)}{2}\mathbb{E}_t[\|\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top(\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t))\|_{D_t}^2] \\ &\overset{(c)}{\leq} -\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla P(\mathbf{w}_t)\|^2 + \frac{1}{2}\epsilon^{-1/2}C_g^2 L_f^2\mathbb{E}[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] \end{aligned} \tag{23}$$

where equality (a) is due to $\nabla P(\mathbf{w}_t) = \mathbb{E}_{i_t,\xi}[\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top \nabla f(g_{i_t}(\mathbf{w}_t))]$, where i_t and ξ are independent. The inequality (b) is according to $ab \le a^2/2 + b^2/2$. The last inequality (c) is due to $\epsilon^{-1/2}\mathbf{I} \ge \|D_t\mathbf{I}\| = \|\frac{1}{\sqrt{\epsilon\mathbf{I}+\hat{v}_{t+1}}}\| \ge \|(\epsilon\mathbf{I} + C_g^2 C_f^2)^{-1/2}\| = (\epsilon + C_g^2C_f^2)^{-1/2}\mathbf{I}$, $(1-\eta_1) \le 1$ and

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top(\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))\|_{D_t}^2] \\
&\le \epsilon^{-1/2} C_g^2 \mathbb{E}_t[\|\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t))\|_{\mathbf{I}}^2] \\
&\le \epsilon^{-1/2} C_g^2 L_f^2 \mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{24}$$

For I_2^t and I_3^t, we have

$$\begin{aligned}
\mathbb{E}_t[I_2^t] &= -\eta_1\langle \nabla P(\mathbf{w}_t) - \nabla P(\mathbf{w}_{t-1}), D_t h_t\rangle - \eta_1\langle \nabla P(\mathbf{w}_{t-1}), D_t h_t\rangle \\
&\le \eta_1 L\alpha^{-1}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 - \eta_1\langle \nabla P(\mathbf{w}_{t-1}), D_t h_t\rangle \\
&= \eta_1 L\alpha^{-1}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 + \eta_1(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\
&\le \eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \eta_1(I_1^{t-1} + I_2^{t-1} + I_3^{t-1})
\end{aligned} \tag{25}$$

where the last equation applies Lemma 7.

$$\begin{aligned}
\mathbb{E}_t[I_3^t] &= -\langle \nabla P(\mathbf{w}_t), (D_{t+1} - D_t)h_{t+1}\rangle = -\sum_{i'=1}^{d} \nabla_{i'} P(\mathbf{w}_t)((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2})h_{t+1}^{i'} \\
&\le \|\nabla P(\mathbf{w}_t)\|\|h_{t+1}\| \sum_{i'=1}^{d}((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2}) \\
&\le C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2})
\end{aligned} \tag{26}$$

By combining Equation (24), (25) and (26) together,

$$\begin{aligned}
\mathbb{E}_t[I_1^t + I_2^t + I_3^t] \le &-\frac{(1-\eta_1)}{2}(\epsilon + C_g^2C_f^2)^{-1/2}\|\nabla P(\mathbf{w}_t)\|^2 + \frac{1}{2}\epsilon^{-1/2}C_g^2L_f^2\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] \\
&+ \eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \eta_1(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\
&+ C_g^2C_f^2\sum_{i'=1}^{d}((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2})
\end{aligned} \tag{27}$$

Define the Lyapunov function

$$\mathcal{L}_t = P(\mathbf{w}_t) - c_t\langle \nabla P(\mathbf{w}_{t-1}), D_t h_t\rangle \tag{28}$$

where c_t and c will be defined later.

$$\begin{aligned}
&\mathbb{E}_t[\mathcal{L}_{t+1} - \mathcal{L}_t] \\
&= P(\mathbf{w}_{t+1}) - P(\mathbf{w}_t) - c_{t+1}\langle \nabla P(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle + c_t\langle \nabla P(\mathbf{w}_{t-1}), D_t h_t\rangle \\
&\le -(c_{t+1} + \alpha)\langle \nabla P(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + c_t\langle \nabla P(\mathbf{w}_{t-1}), D_t h_t\rangle \\
&= (c_{t+1} + \alpha)(I_1^t + I_2^t + I_3^t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 - c_t(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\
&\overset{Eqn\ (27)\ \text{and}\ Lemma\ 7}{\le} -(\alpha + c_{t+1})\frac{(1-\eta_1)}{2}(\epsilon + C_g^2C_f^2)^{-1/2}\|\nabla P(\mathbf{w}_t)\|^2 \\
&+ (\alpha + c_{t+1})\eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \eta_1(\alpha + c_{t+1})(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\
&+ (\alpha + c_{t+1})C_g^2C_f^2\sum_{i'=1}^{d}((\epsilon + \hat{v}_{i'}^t)^{-1/2} - (\epsilon + \hat{v}_{i'}^{t+1})^{-1/2}) \\
&+ \frac{L}{2}\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} - c_t(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) + \frac{\epsilon^{-1/2}C_g^2L_f^2(\alpha + c_{t+1})}{2}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2
\end{aligned} \tag{29}$$

By setting $\alpha_{t+1} \leq \alpha_t = \alpha$, $c_t = \sum_{p=t}^{\infty}(\prod_{j=t}^{p}\eta_1)\alpha_j$, and $c = (1+(1-\eta_1)^{-1})\epsilon^{-\frac{1}{2}}C_g^2L_f^2$, we have

$$c_t \leq (1-\eta_1)^{-1}\alpha_t,\ \frac{2(\alpha+c_{t+1})}{\alpha}\beta\epsilon^{-1/2}C_g^2L_f^2 \leq c\beta,\ \eta_1(\alpha+c_{t+1}) = c_t \tag{30}$$

As a result, $\eta_1(\alpha+c_{t+1})(I_1^{t-1}+I_2^{t-1}+I_3^{t-1}) - c_t(I_1^{t-1}+I_2^{t-1}+I_3^{t-1}) = 0$

$$\begin{aligned}
\mathbb{E}_t[\mathcal{L}_{t+1}-\mathcal{L}_t] \leq & -(\alpha+c_{t+1})\frac{(1-\eta_1)}{2}(\epsilon+C_g^2C_f^2)^{-1/2}\|\nabla P(\mathbf{w}_t)\|^2 \\
& + (\alpha+c_{t+1})\eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \frac{L}{2}\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
& + (\alpha+c_{t+1})C_g^2C_f^2\sum_{i'=1}^{d}((\epsilon+\hat{v}_{i'}^t)^{-1/2} - (\epsilon+\hat{v}_{i'}^{t+1})^{-1/2}) \\
& + \frac{(\alpha+c_{t+1})}{2}\epsilon^{-1/2}C_g^2L_f^2\|g_{i_t}(\mathbf{w}_t)-\mathbf{u}_{i_t}\|^2 \\
\leq & -\alpha\frac{(1-\eta_1)}{2}(\epsilon+C_g^2C_f^2)^{-1/2}\|\nabla P(\mathbf{w}_t)\|^2 \\
& + 2\eta_1 L\alpha^2 Td(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + \frac{L}{2}T\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
& + 2(1-\eta_1)^{-1}\alpha C_g^2C_f^2\sum_{i'=1}^{d}((\epsilon+\hat{v}_{i'}^t)^{-1/2} - (\epsilon+\hat{v}_{i'}^{t+1})^{-1/2}) + \frac{c\alpha}{4}\sum_{t=1}^{T}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t)-\mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{31}$$

where the last inequality is due to equation (30) such that we have $2(\alpha+c_{t+1})\epsilon^{-1/2}C_g^2L_f^2 \leq c\alpha$, and $\alpha + c_{t+1} \leq 2(1-\eta_1)^{-1}\alpha$.

Then by rearranging terms, and taking summation from $1,\cdots,T$ of equation (31), we have

$$\begin{aligned}
& \sum_{t=1}^{T}\alpha\frac{(1-\eta_1)}{2}(\epsilon+C_g^2C_f^2)^{-1/2}\|\nabla P(\mathbf{w}_t)\|^2 \leq \sum_{t=1}^{T}\mathbb{E}_t[\mathcal{L}_t-\mathcal{L}_{t+1}] \\
& + 2\eta_1 L\alpha^2 Td(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + LT\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
& + 2(1-\eta_1)^{-1}\alpha C_g^2C_f^2\sum_{t=1}^{T}\sum_{i'=1}^{d}((\epsilon+\hat{v}_{i'}^t)^{-1/2} - (\epsilon+\hat{v}_{i'}^{t+1})^{-1/2}) + c\alpha\sum_{t=1}^{T}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t)-\mathbf{u}_{i_t}\|^2] \\
& \leq \mathbb{E}[\mathcal{L}_1] - \mathbb{E}[\mathcal{L}_{T+1}] \\
& + 2\eta_1 L\alpha^2 Td(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + LT\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
& + 2(1-\eta_1)^{-1}\alpha C_g^2C_f^2\sum_{i'=1}^{d}((\epsilon+\hat{v}_0^{i'})^{-1/2}) + c\alpha\sum_{t=1}^{T}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t)-\mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{32}$$

By combing with Lemma 3, We finish the proof. □