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
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 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
this problem by using traditional optimization techniques, e.g., hill climbing search \cite{hill_climbing}, cutting-plane method \cite{cutting_plane}, dynamic programming \cite{dynamic_programming}, and by developing acceleration techniques in the framework of SVM \cite{svm}. These approaches are not scalable to big data for DL. There is a long list of studies in information retrieval \cite{information_retrieval,information_retrieval2,information_retrieval4,information_retrieval5,information_retrieval6,information_retrieval7,information_retrieval8,information_retrieval9,information_retrieval10,information_retrieval11,information_retrieval12} and computer vision \cite{computer_vision,computer_vision2,computer_vision3,computer_vision4,computer_vision5,computer_vision6,computer_vision7,computer_vision8,computer_vision9,computer_vision10}, 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 \cite{mini_batch1,mini_batch2,mini_batch3} and usually require a large mini-batch size as confirmed by our studies as well. Recently, \cite{recently} 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 unbiased 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 $E_{\xi}\{f(E_{\zeta}\{g(w; \xi); \xi\})\}$ where $\xi$ and $\zeta$ are independent random variables, or its finite-sum variant has been studied extensively in the literature \cite{finite_sum1,finite_sum2,finite_sum3,finite_sum4,finite_sum5,finite_sum6,finite_sum7,finite_sum8,finite_sum9,finite_sum10,finite_sum11}. In this paper, we formulate the surrogate function of AP into a similar but more complicated two-level compositional function of the form $E_{\xi}\{f(E_{\zeta}\{g(w; \xi, \xi)\})\}$, where $\xi$ and $\zeta$ are independent and $\xi$ has a finite support. The key difference between our formulated compositional function and the ones considered in previous work is that the inner function $g(w; \xi, \xi)$ also depends on the random variable $\xi$ 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 $(x, y)$ a data pair, where $x \in \mathbb{R}^d$ denotes the input data and $y \in \{1, -1\}$ denotes its class label. Let $h(x) = h_w(x)$ denote the predictive function parameterized by a parameter vector $w \in \mathbb{R}^D$ (e.g., a deep neural network). Denote by $I(\cdot)$ 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 $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote the set of all training examples and $D_+ = \{x_i : y_i = 1\}$ denote the set of all positive examples. Let $n_+ = |D_+|$ denote the number of positive examples. $x_i \sim D$ means that $x_i$ is randomly sampled from $D$.

#### 3.1 Background on AUPRC and its estimator AP

Following the work of Bamber \cite{bamber}, 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 $D = \{(x_i, y_i), i = 1, \ldots, n\}$ with the prediction score for each example $x_i$ given by $h_w(x_i)$, we consider to use AP to approximate AUPRC, which is given by

$$ AP = \frac{1}{n_+} \sum_{i=1}^{n_+} I(y_i = 1) \frac{\sum_{s=1}^{n} I(y_s = 1) I(h_w(x_s) \geq h_w(x_i))}{\sum_{s=1}^{n} I(h_w(x_s) \geq h_w(x_i))}, $$

where $n_+$ denotes the number of positive examples. It can be shown that AP is an unbiased estimator in the limit $n \to \infty$ \cite{bias_analysis}.  

3
However, the non-continuous indicator function $I(h_w(x_i) \geq h_w(x_i))$ in both numerator and denominator in (1) makes the optimization non-tractable. To tackle this, we use a loss function $\ell(w; x_s, x_t)$ as a surrogate function of $I(h_w(x_i) \geq h_w(x_i))$. 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 $\ell(w; x_i, x_t) = \max\{m - h_w(x_i) - h_w(x_t), 0\}^2$, where $m$ is a margin parameter. Note that we do not require $\ell$ to be a convex function, hence one can also consider non-convex surrogate loss such as ramp loss. As a result, our problem becomes

$$\min_w P(w) = \frac{1}{n^+} \sum_{x_i \in D^+} \left[ -\sum_{s=1}^n I(y_s = 1) \ell(w; x_s; x_i) \right].$$

(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:

$$g(w; x_j, x_i) = [g_1(w; x_j, x_i), g_2(w; x_j, x_i)]^T = [\ell(w; x_j, x_i)I(y_j = 1), \ell(w; x_j, x_i)]^T$$

$$g_{x_i}(w) = \mathbb{E}_{x_j \sim D}[g(w; x_j, x_i)],$$

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

$$P(w) = \frac{1}{n^+} \sum_{x_i \in D^+} f(g_{x_i}(w)) = \mathbb{E}_{x_i \sim D^+}[f(g_{x_i}(w))].$$

(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 $g_{x_i}(w) = \mathbb{E}_{x_j \sim D}[g(w; x_j, x_i)]$ depends on the random variable $x_i$ of the outer level. This difference makes the proposed algorithm slightly complicated by estimating $g_{x_i}(w)$ 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 $P(w)$. Let the gradient of $g_{x_i}(w)$ be denoted by $\nabla_w g_{x_i}(w)^T = (\nabla_w g_{x_i}(w)[1], \nabla_w g_{x_i}(w)[2]).$ Then we have

$$\nabla_w P(w) = \frac{1}{n^+} \sum_{x_i \in D^+} \nabla_w g_{x_i}(w)^T \nabla f(g_{x_i}(w))$$

$$= \frac{1}{n^+} \sum_{x_i \in D^+} \nabla_w g_{x_i}(w)^T \left[ -\frac{1}{g_{x_i}(w)[2]} \frac{[g_{x_i}(w)]_1}{([g_{x_i}(w)]_2)^2} \right]^T.$$  

(5)

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

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

$$\nabla_w g_{x_i}(w) = \left( \frac{1}{B} \sum_{x_j \in B} I(y_j = 1) \nabla \ell(w; x_j, x_i) \right),$$

(6)

where $B$ denote a set of $B$ random samples from $D$. For estimating $g_{x_i}(w) = \mathbb{E}_{x_j \sim D} g(w; x_j, x_i)$, 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 $g_{x_i}(w)$ for all positive examples. To this end, we will maintain a matrix $u = [u^1, u^2]$ with each column indexable by any positive example, i.e., $u^1, u^2$ correspond to the moving average estimator of $[g_{x_i}(w)]_1$ and $[g_{x_i}(w)]_2$, respectively. The matrix $u$ 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 $u^2$ 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

1: **Input:** $\gamma, \alpha, u_0$, and other parameters for SGD-stype update or Adam-stype update.
2: Initialize $w_1 \in \mathbb{R}^d$, $u \in \mathbb{R}^{|B| \times 2}$
3: for $t = 1, \ldots, T$ do
4: Draw a batch of $B_+$ positive samples denoted by $B_+$.
5: Draw a batch of $B$ samples denoted by $B$.
6: $u = UW(B, B_+, u, w_t, \gamma, u_0)$
7: Compute (biased) Stochastic Gradient Estimator
\[
G(w_t) = \frac{1}{B_+} \sum_{x_i \in B_+} \sum_{x_j \in B} \frac{(u_{x_i}^1 - u_{x_j}^2 I(y_j = 1)) \nabla \ell(w; x_j, x_i)}{B(u_{x_j}^2)^2}
\]
8: Update $w_{t+1}$ by a SGD-style method or by a Adam-style method
\[w_{t+1} = UW(w_t, G(w_t))\]
9: end for
10: **Return:** last solution.

Algorithm 2: $UW(B, B_+, u, w_t, \gamma, u_0)$

1: for each positive $x_i \in B_+$ do
2: Compute
\[
\hat{g}_{x_i}(w_t)_1 = \frac{1}{|B|} \sum_{x_j \in B} \ell(w_t; x_j, x_i)
\]
\[
\hat{g}_{x_i}(w_t)_2 = \frac{1}{|B|} \sum_{x_j \in B} \ell(w_t; x_j, x_i)
\]
3: Compute
\[
u_{x_i}^1 = (1 - \gamma)u_{x_i}^1 + \gamma [\hat{g}_{x_i}(w_t)_1]
\]
\[
u_{x_i}^2 = \max((1 - \gamma)u_{x_i}^2 + \gamma [\hat{g}_{x_i}(w_t)_2], u_0)
\]
4: end for
5: Return $u$

Algorithm 3: $UW(w_t, G(w_t))$

1: Option 1: SGD-style update (paras: $\alpha$)
\[w_{t+1} = w_t - \alpha G(w_t)\]
2: Option 2: Adam-style update (paras: $\alpha, \epsilon, \eta_1, \eta_2$)
\[h_{t+1} = \eta_1 h_t + (1 - \eta_1)G(w_t)
\]
\[v_{t+1} = \eta_2 v_t + (1 - \eta_2)(G(w_t))^2
\]
\[w_{t+1} = w_t - \alpha \frac{h_{t+1}}{\sqrt{\epsilon + \hat{v}_{t+1}}}
\]
where $\hat{v}_t = v_t$ (Adam) or $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ (AMSGrad)
3: **Return:** $w_{t+1}$

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 $\Delta_1$ such that $P(w_1) - \min_w P(w) \leq \Delta_1$; (b) there exist $C, M > 0$ such that $\ell(w; x_i, x_i) \geq C$ for any $x_i \in \mathbb{D}_+$, $\ell(w; x_j, x_i) \leq M$, and $\ell(w; x_j, x_i)$ is Lipschitz continuous and smooth with respect to $w$ for any $x_i \in \mathbb{D}_+$, $x_j \in \mathbb{D}_+$; (c) there exists $V > 0$ such that $E_{x_i \sim \mathbb{D}}[\|g(w; x_i, x_i) - g_{x_i}(w)\|^2] \leq V$, and $E_{x_i \sim \mathbb{D}}[\|\nabla g(w; x_i, x_i) - \nabla g_{x_i}(w)\|^2] \leq V$ for any $x_i$.

With a bounded score function $h_w(x)$ the above assumption can be easily satisfied. Based on the above assumption, we can prove that the objective function $P(w)$ 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 $g_{x_i}(w) \in \Omega = \{u \in \mathbb{R}^2; 0 \leq |u|_2 \leq M, u_0 \leq |u|_2 \leq M\}$, $\forall x_i \in \mathbb{D}_+$.

Next, we highlight the convergence analysis of SOAP employing the SGD-stype update and include that for employing Adam-stype update in the supplement. Without loss of generality, we assume $|B_+| = 1$ and the positive sample in $B_+$ is randomly selected from $\mathbb{D}_+$ with replacement. When the context is clear, we abuse the notations $g_i(w)$ and $u_i$ to denote $g_{x_i}(w)$ and $u_{x_i}$ 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
$$\frac{\alpha}{2} \mathbb{E} \sum_{t=1}^{T} \| \nabla P(\mathbf{w}_t) \|^2 \leq \mathbb{E} \sum_{t=1}^{T} (P(\mathbf{w}_t) - P(\mathbf{w}_{t+1})) + \frac{\alpha C_1}{2} \mathbb{E} \sum_{t=1}^{T} \| g_{t+1}(\mathbf{w}_t) - \mathbf{u}_i \|^2 + \alpha^2 TC_2,$$
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 $x_i \in \mathcal{D}_+$ we have
$$\mathbb{E} \sum_{t=1}^{T} \| g_{t+1}(\mathbf{w}_t) - \mathbf{u}_i \|^2 \leq \frac{n_+ V}{\gamma} + \gamma V T + 2 \frac{\alpha^2 TC_3}{\gamma^2},$$
where $C_3$ is a proper constant.

Remark: The innovation of proving the above lemma is by grouping $\mathbf{u}_i, t = 1, \ldots, T$ into $n_+$ groups corresponding to the $n_+$ positive examples, and then establishing the recursion of the error $\| g_{t+1}(\mathbf{w}_t) - \mathbf{u}_i \|^2$ 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 $\alpha = \frac{1}{n_+^{2/5} \gamma^{3/5}}, \gamma = \frac{n_+^{2/5}}{\gamma^{5/5}}, \forall t \in 1, \ldots, T$, and $T > n_+$. Then after running $T$ iterations, SOAP with a SGD-style update satisfies
$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} \| \nabla P(\mathbf{w}_t) \|^2 \right] \leq O\left( \frac{n_+^{2/5}}{\gamma^{5/5}} \right),$$
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 $\eta_1 \leq \sqrt{\eta_2} \leq 1$, $\alpha = \frac{1}{n_+^{2/5} \gamma^{3/5}}, \gamma = \frac{n_+^{2/5}}{\gamma^{5/5}}, \forall t \in 1, \ldots, T$, and $T > n_+$. Then after running $T$ iterations, SOAP with an AMSGrad update satisfies
$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} \| \nabla P(\mathbf{w}_t) \|^2 \right] \leq O\left( \frac{n_+^{2/5}}{\gamma^{5/5}} \right),$$
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. **Focal** is to up-weight the penalty on hard examples using focal loss. **LDAM** refers to training with label-distribution-aware margin loss. **AUC-M** is an AUROC maximization method using a surrogate loss. In addition, we compare with three methods for optimizing AUPRC or AP; namely, the **MinMax** method - a method for optimizing a discrete approximation of AUPRC. **SmoothAP** - a method that optimizes a smoothed approximation of AP; and **FastAP** - a method that uses soft histogram binning to approximate the gradient of AP. 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.

<table>
<thead>
<tr>
<th>Networks</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ResNet18</td>
<td>ResNet34</td>
</tr>
<tr>
<td>CE</td>
<td>0.7155 (± 0.0058)</td>
<td>0.6844 (± 0.0031)</td>
</tr>
<tr>
<td>CB-CE</td>
<td>0.7325 (± 0.0039)</td>
<td>0.6936 (± 0.0021)</td>
</tr>
<tr>
<td>Focal</td>
<td>0.7183 (± 0.0082)</td>
<td>0.6943 (± 0.0007)</td>
</tr>
<tr>
<td>LDAM</td>
<td>0.7346 (± 0.0125)</td>
<td>0.6745 (± 0.0043)</td>
</tr>
<tr>
<td>AUC-M</td>
<td>0.7399 (± 0.0013)</td>
<td>0.6825 (± 0.0089)</td>
</tr>
<tr>
<td>SmoothAP</td>
<td>0.7365 (± 0.0088)</td>
<td>0.6909 (± 0.0049)</td>
</tr>
<tr>
<td>FastAP</td>
<td>0.7028 (± 0.0341)</td>
<td>0.6798 (± 0.0032)</td>
</tr>
<tr>
<td>MinMax</td>
<td>0.7228 (± 0.0118)</td>
<td>0.6806 (± 0.0027)</td>
</tr>
<tr>
<td>SOAP</td>
<td>0.7629 (± 0.0014)</td>
<td><strong>0.7012</strong> (± 0.0056)</td>
</tr>
</tbody>
</table>

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 imbalanced 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 \( u \) 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 \( 384 \times 384 \) 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 \( \gamma \) 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 (imbalance 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.
We show that our method can serve as a useful solution to the above problem. We conduct experiments to model structure is used. These results clearly demonstrate that our method is effective for 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.

### 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.

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

### 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 for 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 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.

<table>
<thead>
<tr>
<th>Data</th>
<th>MIT AICURES</th>
<th>Kaggle Melanoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Networks</td>
<td>GINE</td>
<td>MPNN</td>
</tr>
<tr>
<td>CE</td>
<td>0.5037 (± 0.0718)</td>
<td>0.6282 (± 0.0634)</td>
</tr>
<tr>
<td>CB-CE</td>
<td>0.5655 (± 0.0453)</td>
<td>0.6308 (± 0.0263)</td>
</tr>
<tr>
<td>Focal</td>
<td>0.5143 (± 0.1062)</td>
<td>0.5875 (± 0.0774)</td>
</tr>
<tr>
<td>LDAM</td>
<td>0.5236 (± 0.0551)</td>
<td>0.6489 (± 0.0556)</td>
</tr>
<tr>
<td>AUC-M</td>
<td>0.5149 (± 0.0748)</td>
<td>0.5542 (± 0.0474)</td>
</tr>
<tr>
<td>SmoothAP</td>
<td>0.2899 (± 0.0220)</td>
<td>0.4081 (± 0.0352)</td>
</tr>
<tr>
<td>FastAP</td>
<td>0.4777 (± 0.0896)</td>
<td>0.4518 (± 0.1495)</td>
</tr>
<tr>
<td>MinMax</td>
<td>0.5292 (± 0.0330)</td>
<td>0.5774 (± 0.0468)</td>
</tr>
<tr>
<td>SOAP</td>
<td>0.6639 (± 0.0515)</td>
<td>0.6547 (± 0.0616)</td>
</tr>
</tbody>
</table>

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. 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 \(-P(w)\) vs Iterations on CIFAR10 and CIFAR100.

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

<table>
<thead>
<tr>
<th>Data</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ResNet18</td>
<td>ResNet34</td>
</tr>
<tr>
<td>Squared Hinge</td>
<td>0.7629 (±0.0014)</td>
<td>0.7012 (±0.0056)</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.7542 (±0.0024)</td>
<td>0.6968 (±0.0121)</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>0.7652 (±0.0035)</td>
<td>0.6983 (±0.0084)</td>
</tr>
</tbody>
</table>

and ToxCast in Table 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.

**In sensitivity 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 \(\ell(w; x_s, x_i)\) as a surrogate function of the indicator \(I(h_w(x) \geq h_w(x_i))\), namely, the logistic loss, \(\ell(w; x_s, x_i) = \log \frac{1}{1 + \exp( c(\ell(h_w(x_i)) - h_w(x_s)))}\), and the sigmoid loss, \(\ell(w; x_s, x_i) = \frac{1}{1 + \exp( c(\ell(h_w(x_i)) - h_w(x_s)))}\) 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 \(-P(w)\) 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


