Leveraging First and Zeroth-Order Gradient to Address Imbalanced Black-Box Prompt Tuning via Minimax Optimization

Haozhen Zhang[1 *], **Zhaogeng Liu**[1 *], **Bin Gu**[1 †], **Yi Chang**[1 2 3 †]

[1]School of Artificial Intelligence, Jilin University, Changchun, Jilin, China
[2]International Center of Future Science, Jilin University, Changchun, Jilin, China
[3]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, Changchun, Jilin, China
{haozhen23, zgliu20}@mails.jlu.edu.cn, jsgubin@gmail.com, yichang@jlu.edu.cn

Abstract

Black-box prompt tuning has become a prevalent parameter-efficient paradigm that leverages the capabilities of large language models (LLMs) for customized applications in specific downstream tasks. In practical scenarios, downstream tasks frequently involve data distributions that are heavily imbalanced. Such imbalances tend to impair performance, causing severe performance collapse in minority classes. Conducting effective imbalanced black-box prompt tuning to mitigate the adverse effects of imbalanced data distribution on prompt performance remains a significant challenge. In this paper, we propose ***b**lack-box **p**rompt **t**uning with **f**irst and **z**eroth order **g**radient* (BPT-FZG) for handling the imbalanced data. Specifically, BPT-FZG introduces AUC maximization as the objective for prompt tuning and equivalently formulates it as a nonconvex-concave saddle point problem to avoid the construction of sample pairs from opposite classes. Indeed, BPT-FZG optimizes the latent representation of the continuous prompt in the low-dimensional subspace with AUC loss and leverages the first and zeroth order gradients alternately to update the parameters. Furthermore, we establish the theoretical convergence guarantee for BPT-FZG under common assumptions, showing that our method can find a stationary point of the objective function. Our experiments on RoBERTa-large, GPT2-XL, and Llama3 show that BPT-FZG achieves improvement on various imbalanced datasets, emphasizing the effectiveness of our methods.

Code — https://github.com/ZHZ-JLU/zo-sppt

Introduction

Large language models (LLMs) have demonstrated impressive performance across a wide range of downstream tasks in the past few years (Devlin et al. 2018; Brown et al. 2020; Lewis et al. 2020; Sun et al. 2021). These models are trained on a massive text corpus to learn contextual relationships, language structures, and other aspects of human knowledge, demonstrating remarkable generalization in text understanding and generation (Hadi et al. 2023). By leveraging this acquired comprehensive language knowledge, LLMs can be tuned with specific knowledge about downstream tasks for customized applications. Meanwhile, researchers have proposed numerous methods for developing customized LLM applications based on fine-tuning. However, these methods are challenged by the increasing size of parameters and can only be applied in the white-box setting where the model is open-sourced (Yang et al. 2020; Varia et al. 2022; Iyer, Chen, and Birch 2023).

Recently, most LLMs like ChatGPT-4, which have enormous parameter sizes, offer only online API services. This prevents users from directly accessing the model's parameters, resulting in an inevitable black-box setting. Therefore, numerous studies have shifted towards black-box prompt tuning, which is a parameter-efficient paradigm for adapting LLMs to specific natural language processing (NLP) tasks (Diao et al. 2022; Deng et al. 2022; Zhao, Wang, and Yang 2023). Moreover, this method is well-suited to the emerging Language-Model-as-a-Service (LMaaS) scenario, where users are limited to interacting with LLMs solely through APIs (Sun et al. 2022a,b).

Many NLP tasks inherently exhibit imbalance, with certain target categories occurring much more frequently than others in real-world scenarios (Henning et al. 2022; Waseem and Hovy 2016; Gao et al. 2021). Such imbalanced data tends to cause LLMs to overfit the majority classes while neglecting the minority classes due to the excessive number of training samples for the former. Similarly, prompt tuning in the presence of imbalanced data also encounters this issue (Dong et al. 2022). For instance, cross-entropy loss is typically employed as the surrogate for classification error rate in text classification downstream tasks. However, when the data distribution is imbalanced, relying on cross-entropy loss becomes inappropriate. It exacerbates the aforementioned issues, leading to the performance collapse for the minority classes (Liu et al. 2019a).

Some recent works in black-box prompt tuning have concentrated on identifying the optimal continuous prompts (Sun et al. 2022b,a; Zheng et al. 2024). These methods use a low-rank matrix to project the high-dimensional prompt space into a low-dimensional subspace and optimize the latent representation of the continuous prompt within that subspace. This paves the way for the application of black-box optimization algorithms. Moreover, these methods significantly reduce computational costs, thereby narrowing the gap be-

[*]These authors contributed equally as co-first authors and are listed in random order.
[†]Co-corresponding author.

tween research and practical application.

However, the works above do not account for additional handling of imbalanced data. These studies typically focus solely on vanilla text classification with uniformly distributed classes, ignoring the practical challenges posed by imbalanced data in real-world applications. Consequently, the performance of these efforts in dealing with imbalanced data remains unverified. On the other hand, current prompt tuning frameworks predominantly emphasize empirical results and often lack a comprehensive theoretical discussion. Therefore, investigating an imbalanced black-box prompt tuning method that is supported by theoretical guarantees represents a valuable challenge.

To address this challenge, in this paper, we propose a novel framework for black-box prompt tuning with first and zeroth order gradient (BPT-FZG) to handle imbalanced data. We introduce AUC (short for Area Under the ROC Curve) maximization as the optimization objective of continuous prompt tuning (Hanley and McNeil 1982). To address the issue that the AUC maximization objective function relies on pairwise loss between instances from different classes, we equivalently formulate it as a stochastic saddle point problem, whose objective function is pointwise (Ying, Wen, and Lyu 2016; Liu et al. 2019a; Yuan et al. 2021b). Additionally, BPT-FZG leverages the intrinsic dimensionality of LLMs to overcome the challenge that the high dimensionality of continuous prompts poses for black-box optimization. Specifically, our framework builds on the previous insight that optimizes a latent representation of the prompt in a subspace characterized by LLM's low intrinsic dimensionality (Sun et al. 2022b; Zheng et al. 2024). Furthermore, given the specificity of the objective function, BPT-FZG alternates between first-order and zero-order gradients to update parameters, which includes the latent representation of continuous prompt and additional parameters introduced by the saddle point problem. In essence, BPT-FZG utilizes the stochastic gradient descent ascent (SGDA) algorithm to tackle the nonconvex-concave minimax problem (Heusel et al. 2017). Through experiments conducted across various tasks and models, we establish the efficacy of our approach representing significant progress in the practical use of LLMs.

Main Contributions

The main contributions of this work can be summarized as follows:

- We propose a novel black-box prompt tuning framework BPT-FZG for solving the challenge of prompt tuning in downstream tasks with imbalanced data. Our framework employs a formulation equivalent to AUC maximization as the objective.
- We provide a theoretical analysis that demonstrates the algorithm's ability to find a stationary point. The theoretical results indicate that BPT-FZG, which fundamentally employs the GDA algorithm with the first and zero-order hybrid gradient, achieves a convergence rate similar to that of its first-order counterpart.
- We extensively validate BPT-FZG on RoBERTa-large, GPT2-XL, and Llama3. Experimental results from various imbalanced downstream task datasets demonstrate that BPT-FZG outperforms baseline methods.

Related Works

Black-box Prompt Tuning

In recent years, the exploration of black-box prompt tuning for LLMs has attracted considerable attention, yielding promising results in various NLP tasks (Deng et al. 2022; Hou et al. 2023; Wen et al. 2024). (Sun et al. 2022b) summarizes the characteristics of current LLM service applications as the LMaaS scenario and proposes the black-box tuning (BBT) method to optimize a series of vectors called the continuous prompt, which serves as a prefix to the input text embeddings. This framework optimizes through a derivative-free method and leverages low-dimensional subspace to address the curse of dimensionality. (Zheng et al. 2024) points out that the use of the random projection matric in the former is an inappropriate choice for subspace. Therefore, they assume optimal prompts exist within a shared subspace for similar tasks and propose black-box prompt tuning with subspace learning (BSL) to identify the optimal subspace. (Lin et al. 2023) integrates federated learning with black-box prompt tuning to introduce a federated black-box prompt tuning framework (Fed-BBPT). The aforementioned works are continuous prompt optimization based on (Sun et al. 2022b), while there is also a series of research focused on discrete prompt learning. (Prasad et al. 2022) achieves gradient-free instructional prompt search (GRIPS) through prompt editing and beam search. (Zhao, Wang, and Yang 2023) treats each prompt as an individual and employs the genetic algorithm for discrete prompt optimization. On the other hand, (Diao et al. 2022) employs policy gradients to optimize the distribution over the discrete prompt.

AUC Maximization

AUC maximization is a learning paradigm that trains the model by maximizing the AUC score on a given dataset. It is more suitable for imbalanced data compared to maximizing accuracy (Yang and Ying 2022). Research on AUC maximization has a history of over 2 decades and has been applied in various contexts and models, e.g., active learning (Han and Zhao 2010), federated learning (Guo et al. 2020), linear models (Gao et al. 2013), neural networks (Yan et al. 2003), etc. This paper focuses on works related to formulating AUC maximization as the equivalent saddle point problem. (Ying, Wen, and Lyu 2016) is the first to demonstrate that AUC maximization for linear models can be equivalently formulated as a saddle point problem when the surrogate loss is square loss, addressing the issue of needing to store training samples in the online setting. (Dan and Sahoo 2021) introduces proximal SVRG based on the former to propose a variance-reduced stochastic proximal algorithm for AUC maximization. (Liu et al. 2019a) extends the aforementioned equivalence from linear models to deep neural networks and establishes theoretical convergence. (Yuan et al. 2021b) improves the AUC square loss to AUC margin loss and builds a similar equivalence to solve the medical image classification problem. (Yuan et al. 2021a) proposes a compositional training method

for deep AUC maximization with a minimax objective to tackle the challenge of end-to-end training. (Yang et al. 2021) presents a theoretical framework for empirical risk minimization, guided by the multiclass generalization of the AUC. (Yang et al. 2022) first turns to optimize a variant of AUC called Two-way Partial AUC and solves the difficulty of end-to-end stochastic training. (Yang et al. 2023) points out the shortcoming of AdAUC (Hou et al. 2022) and constructs an adversarially robust AUC optimization method.

Methodology

Notations. Let $\mathcal{D} \triangleq \{(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \ldots, (\boldsymbol{X}_M, y_M)\}$ denote a set of training data with cardinality M. For any $m \in \{1, 2, \ldots, M\}$, $\boldsymbol{X}_m$ represents an input training example, i.e. a sentence to be classified, and $y_m \in \{-1, 1\}$ denotes its corresponding label. By defining $\mathbf{X}_m$ is the embedding of $\boldsymbol{X}_m$, we can denote $\tilde{\mathcal{D}} \triangleq \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_M, y_M)\}$ as the set of tuples composed of embedding vectors and labels. And we further denote $\tilde{\mathbf{X}} \triangleq \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M\}$ and $\tilde{\mathbf{Y}} \triangleq \{y_1, y_2, \ldots, y_M\}$ as the set of the embedding vectors and labels. We define the continuous prompt as $\mathbf{T} \in \mathbb{R}^D$. Let $\mathbf{X} \in \tilde{\mathbf{X}}$, then the user query to a black-box LLM $h(\cdot)$ is denoted as $h([\mathbf{T}, \mathbf{X}])$, i.e. $h([\mathbf{T}, \mathbf{X}])$ denote the prediction of the LLM $h(\cdot)$ on $[\mathbf{T}, \mathbf{X}]$.

Breif Review of Black-box Prompt Tuning

Continuous prompt tuning is regarded as a more lightweight approach for adapting LLMs to specific downstream tasks as opposed to fine-tuning (Li and Liang 2021). Indeed, achieving competitive performance typically necessitates employing dozens of prompt tokens (Lester, Al-Rfou, and Constant 2021). Considering that each prompt token's embedding dimension is in the thousands, the total parameter count for optimization reaches tens of thousands, still far below the parameters required for fine-tuning. However, within the realm of black-box optimization, this high-dimensional framework will precipitate the "curse of dimensionality", rendering the optimization process inefficient.

Fortunately, it has been revealed that LLMs possess a very low intrinsic dimensionality (Aghajanyan, Zettlemoyer, and Gupta 2020; Qin et al. 2021). This implies that the high-dimensional prompt embedding has a latent low-dimensional representation, meaning that black-box optimization can be effectively performed in a low-dimensional subspace to circumvent the "curse of dimensionality". Therefore, rather than optimizing continuous prompts in the original D-dimensional space, optimizing the latent representations of these prompts in a d-dimensional subspace is preferable ($d \ll D$). By denoting the learnable low-dimensional latent representation as $\mathbf{z} \in \mathbb{R}^d$, we can formulate the objective of black-box prompt tuning as follows,

$$\mathbf{z}^* = \underset{\mathbf{z} \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}\left(h\left([\mathbf{A}\mathbf{z} + \mathbf{p}_0, \tilde{\mathbf{X}}]\right), \tilde{\mathbf{Y}}\right), \tag{1}$$

where $\mathcal{L}(\cdot)$ is the loss function, $\mathbf{A} \in \mathbb{R}^{D \times d}$ is the random projection matrix, $\mathbf{p}_0 \in \mathbb{R}^D$ is initial prompt embedding. This method employs derivative-free optimization techniques to update the latent representation within a subspace and then project it back to the original space through random matrix $\mathbf{A}$, aiming to offer a more computationally feasible paradigm for tuning LLMs.

Imbalanced Black-box Prompt Tuning with AUC Maximization

AUC Maximization. The AUC score is defined as the probability that the prediction score for a positive sample surpasses that of a negative sample. It is extensively utilized to evaluate prediction performance on imbalanced datasets. We propose employing AUC maximization for black-box continuous prompt tuning to address the challenge posed by class imbalance in prompt tuning. Building upon existing work on AUC maximization, the definition of AUC within the above prompt tuning framework is as follows,

$$\begin{aligned} \mathrm{AUC}(\mathbf{z}) &= \Pr\left(h([\mathbf{z}, \mathbf{X}]) \geq h([\mathbf{z}, \mathbf{X}']) \,|\, y = 1, y' = -1\right) \\ &= \mathbb{E}[\mathbb{I}\left(h([\mathbf{z}, \mathbf{X}]) - h([\mathbf{z}, \mathbf{X}']) \geq 0\right) | y = 1, y' = -1] \end{aligned} \tag{2}$$

where we denote $h([\mathbf{z}, \mathbf{X}])$ as the abbreviation for $h([\mathbf{A}\mathbf{z} + \mathbf{p}_0, \mathbf{X}])$. However, due to the discontinuity of the indicator function $\mathbb{I}(\cdot)$ in the Equation (2), which greatly complicates the optimization, it is common to use a convex surrogate loss function (e.g. square loss) as a replacement (Gao et al. 2013). As a result, we can formulate the objective of prompt tuning with AUC maximization on the imbalanced dataset $\tilde{\mathcal{D}}$ as follows,

$$\min_{\mathbf{z} \in \mathbb{R}^d} \frac{1}{N_+ N_-} \sum_{\mathbf{X} \in \tilde{\mathcal{D}}_+} \sum_{\mathbf{X}' \in \tilde{\mathcal{D}}_-} \left(1 - (h([\mathbf{z}, \mathbf{X}]) - h([\mathbf{z}, \mathbf{X}']))\right)^2 \tag{3}$$

where $\tilde{\mathcal{D}}_+$, $\tilde{\mathcal{D}}_-$ denote the set of positive and negative samples in $\tilde{D}$, with their cardinalities denoted as N_+ and N_-, respectively.

Equivalent Saddle Point Problem. It can be observed that the objective in Equation (3) is defined over pairs of training samples from opposing classes, which significantly increases the computational cost of each iteration compared to the pointwise loss (Ying, Wen, and Lyu 2016). Additionally, the necessity of constructing sample pairs makes it challenging to extend this paradigm to the online learning scenario. In this work, we introduce the stochastic saddle point problem which is the equivalence of AUC maximization to avoid the challenge of constructing sample pairs from opposite classes. As the constrained AUC optimization problem defined in (Yuan et al. 2021a), we can reframe the objective in Equation (3) as the following non-convex min-max problem with non-negative constraint $\mathcal{A} = \{0 \leq \alpha \leq \alpha_m\}$ for relieving the adverse effect when trained with easy data (Yuan et al. 2020),

$$\min_{\mathbf{z} \in \mathbb{R}^d, (a,b) \in \mathbb{R}^2} \max_{\alpha \in \mathcal{A}} \{f(\mathbf{z}, a, b, \alpha) = \mathbb{E}_\xi[F(\mathbf{z}, a, b, \alpha; \xi)]\} \tag{4}$$

where $\xi = (\mathbf{X}, y) \in \tilde{\mathcal{D}}$ is a random sample, and

$$\begin{aligned} F(\mathbf{z}, a, b, \alpha; \xi) &= (1-p)(h([\mathbf{z}, \mathbf{X}]) - a)^2 \mathbb{I}_{[y=1]} \\ &+ p(h([\mathbf{z}, \mathbf{X}]) - b)^2 \mathbb{I}_{[y=-1]} - p(1-p)\alpha^2 + 2\alpha(p(1-p) \\ &+ ph([\mathbf{z}, \mathbf{X}])\mathbb{I}_{[y=-1]} - (1-p)h([\mathbf{z}, \mathbf{X}])\mathbb{I}_{[y=1]}) \end{aligned} \tag{5}$$

and $p = \Pr(y = 1)$.

Remark 1. *The constraint set $\mathcal{A}$ is a convex and bounded set with a diameter $D_{\mathcal{A}} > 0$.*

For notational simplicity, let vector $\boldsymbol{v}$ be denoted as $[\mathbf{z}, a, b] \in \mathbb{R}^{d+2}$ in the subsequent content. Consequently, $F(\mathbf{z}, a, b, \alpha; \xi)$ can be abbreviated as $F(\boldsymbol{v}, \alpha; \xi)$, and $f(\mathbf{z}, a, b, \alpha)$ can be abbreviated as $f(\boldsymbol{v}, \alpha)$.

Remark 2. *$F(\boldsymbol{v}, \alpha; \xi)$ is concave in $\alpha \in \mathcal{A}$ for each ξ and $\boldsymbol{v}$. Furthermore, $f(\boldsymbol{v}, \alpha)$ is concave in $\alpha \in \mathcal{A}$ for each $\boldsymbol{v}$.*

Black-box Prompt Tuning with First and Zeroth-Order Gradient

In this subsection, we introduce how to solve the problem (4) with BPT-FZG. The problem (4) is a minimax optimization problem, and a natural approach to solving it is stochastic gradient descent ascent (SGDA), which is an extension of stochastic gradient descent (SGD) for minimax problems. A method for implementing SGDA in the black-box setting is replacing the algorithm's all first-order stochastic gradients with zeroth-order gradient estimators. However, the situation differs for solving the problem (4) in a black-box setting.

Given a black-box model h and the function F parameterized by $[\boldsymbol{v}, \alpha] \in \mathbb{R}^{d+3}$. Since a, b, and α are parameters of the function F, their first-order stochastic gradients can be obtained through backpropagation, even though the h is a black-box model. Therefore, we employ the zeroth-order technique to estimate the gradient only for $\mathbf{z}$. A random gradient estimator with sample ξ is defined as follows,

$$\hat{\nabla}_{\mathbf{z}} F(\boldsymbol{v}, \alpha; \xi) = \frac{F(\mathbf{z} + \mu\mathbf{u}; \xi) - F(\mathbf{z} - \mu\mathbf{u}; \xi)}{2\mu}\mathbf{u}, \quad (6)$$

where μ is a smoothing parameter, $\mathbf{u} \in \mathbb{R}^d$ is distributed in $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $F(\mathbf{z} \pm \mu\mathbf{u}, a, b, \alpha; \xi)$ is denoted as $F(\mathbf{z} \pm \mu\mathbf{u}; \xi)$.

BPT-FZG employs a hybrid of first-order and zeroth-order gradients for optimization, where the first-order gradients of a, b, and α can be obtained with minimal computational cost. This approach ensures the feasibility of BPT-FZG in the black-box setting while simultaneously enhancing the stability and effectiveness of the optimization process. The pseudocode and illustration of BPT-FZG are shown in Algorithm 1 and Figure 1, respectively.

In each iteration of BPT-FZG, a mini-batch $\mathcal{S}_t$ with cardinality m is used to estimate the gradients. As previously described, first-order mini-batch stochastic gradients are computed for variables a, b, and α, while the zeroth-order gradients for $\mathbf{z}$ are calculated using the average of (6). Then, gradient descent is performed on the vector $\boldsymbol{v}$ and gradient ascent is performed on the variable α, where $\mathcal{P}_{\mathcal{A}}$ denotes the projection to the constraint set $\mathcal{A}$.

Algorithm 1: BPT-FZG

Input: Dataset $\tilde{\mathcal{D}}$, Parameters $\boldsymbol{v} = [\mathbf{z}, a, b], \alpha$, Prediction model h, Step sizes $\eta_{\boldsymbol{v}}, \eta_\alpha$.
Initialize $\boldsymbol{v}_0, \alpha_0$ randomly.
for $t = 1, \ldots, T$ **do**
  Draw samples $\mathcal{S}_t = \{\xi_1, \ldots, \xi_m\}$ from $\tilde{\mathcal{D}}$.
  for $i = 1$ to m **do**
    Sample a random direction $\mathbf{u}_i \in \mathbb{R}^d$ from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
    Compute $\hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi_i)$ through (6).
  end for
  Compute $\hat{\mathbf{g}}_{\mathbf{z}}^{(t-1)} = \frac{1}{m}\sum_{i=1}^m \hat{\nabla}_{\mathbf{z}} F(\boldsymbol{v}, \alpha; \xi_i)$ and compute $\mathbf{g}_{a,b}^{(t-1)}, \mathbf{g}_\alpha^{(t-1)}$ by backpropagation.
  Let $\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} = [\hat{\mathbf{g}}_{\mathbf{z}}^{(t-1)}, \mathbf{g}_{a,b}^{(t-1)}]$ and update $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta_{\boldsymbol{v}} \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}$.
  Update $\alpha_t = \mathcal{P}_{\mathcal{A}}(\alpha_{t-1} + \eta_\alpha \mathbf{g}_\alpha^{(t-1)})$.
end for
Output: $\boldsymbol{v}_\tau, \alpha_\tau$ with τ chosen uniformly at random from $1, \ldots, T$. ($\boldsymbol{v}_T, \alpha_T$ in practice)

Convergence

In this section, we provide theoretical convergence analysis for BPT-FZG, which essentially involves applying hybrid gradients to solve the nonconvex-concave minimax problem. We first make the following basic assumptions in this section.

Assumption 1. *The stochastic first-order oracle $\nabla_{\boldsymbol{v}} F(\boldsymbol{v}, \alpha; \xi)$ and $\nabla_\alpha F(\boldsymbol{v}, \alpha; \xi)$ satisfies*

$$\begin{aligned} &\mathbb{E}[\nabla_{\boldsymbol{v}} F(\boldsymbol{v}, \alpha; \xi) - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)] = 0 \\ &\mathbb{E}[\nabla_\alpha F(\boldsymbol{v}, \alpha; \xi) - \nabla_\alpha f(\boldsymbol{v}, \alpha)] = 0 \\ &\mathbb{E}[||\nabla_{\boldsymbol{v}} F(\boldsymbol{v}, \alpha; \xi) - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)||^2] \le \sigma^2 \\ &\mathbb{E}[||\nabla_\alpha F(\boldsymbol{v}, \alpha; \xi) - \nabla_\alpha f(\boldsymbol{v}, \alpha)||^2] \le \sigma^2 \end{aligned} \quad (7)$$

Assumption 2. *F is ℓ-smooth and $F(\cdot, \alpha; \xi)$ is L-Lipschitz for each $\alpha \in \mathcal{A}$ and $\xi \in \tilde{\mathcal{D}}$.*

Proposition 1. *Under Assumption 2, f is ℓ-smooth and $f(\cdot, \alpha)$ is L-Lipschitz for each $\alpha \in \mathcal{A}$.*

We define the function $\Phi(\cdot) = \max_{\alpha \in \mathcal{A}} f(\cdot, \alpha)$, and the minimax problem (4) is equivalent to $\min_{\boldsymbol{v} \in \mathbb{R}^{d+2}} \Phi(\boldsymbol{v})$. The following structural lemma shows that Φ is ℓ-weakly convex and L-Lipschitz.

Lemma 1. *(Lemma 4.8 in (Lin, Jin, and Jordan 2020)) Under the constraint set $\mathcal{A}$ is a convex and bounded set with a diameter $D_{\mathcal{A}} > 0$, $f(\boldsymbol{v}, \alpha)$ is concave in $\alpha \in \mathcal{A}$ for each $\boldsymbol{v}$, and Assumption 2, $\Phi(\cdot)$ is ℓ-weakly convex and L-Lipschitz with $\nabla_{\boldsymbol{v}} f(\cdot, \alpha^*(\cdot)) \in \partial\Phi(\cdot)$ where $\alpha^*(\cdot) \in \operatorname{argmax}_{\alpha \in \mathcal{A}} f(\cdot, \alpha)$.*

For the general nonconvex-concave minimax problem, the function Φ is not necessarily differentiable. This introduces the difficulty in the convergence analysis. To address this, we introduce the following alternative concept for weakly convex optimization proposed by (Davis and Drusvyatskiy 2019).

Definition 1. *A function $\Phi_\lambda : \mathbb{R}^m \to \mathbb{R}$ is the Moreau envelope of Φ with a positive parameter $\lambda > 0$ if $\Phi_\lambda(\boldsymbol{v}) = \min_{\mathbf{w}} \Phi(\mathbf{w}) + (1/2\lambda)||\mathbf{w} - \boldsymbol{v}||^2$ for each $\boldsymbol{v} \in \mathbb{R}^m$*

Definition 2. *A point $\boldsymbol{v}$ is an ϵ-stationary point of a ℓ-weakly convex function Φ if $||\nabla\Phi_{1/2\ell}(\boldsymbol{v})|| \le \epsilon$. If $\epsilon = 0$, then $\boldsymbol{v}$ is a stationary point.*

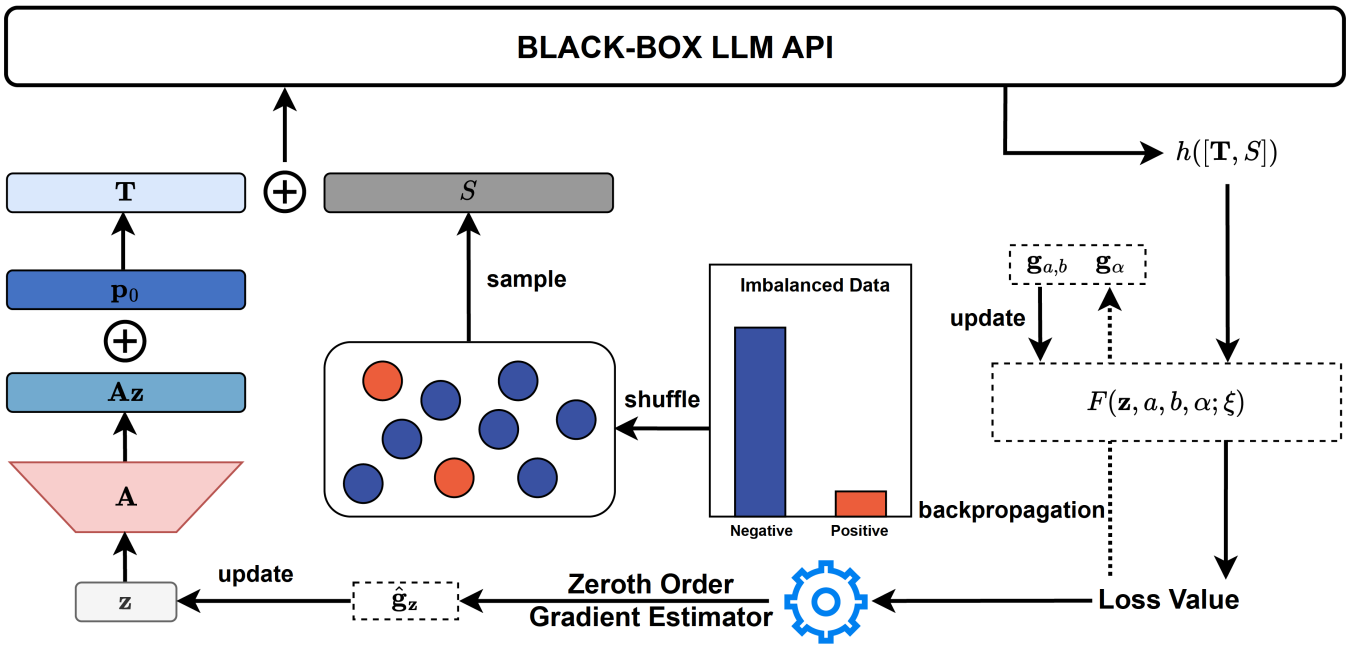

Figure 1: Overview of BPT-FZG. In each iteration, we first obtain the prompt $\mathbf{T}$ by multiplying the latent representation $\mathbf{z}$ with the projection matrix $\mathbf{A}$ and adding the initial value $\mathbf{p}_0$. $\mathbf{T}$ is concatenated with a batch sampled from the imbalanced data and fed into the LLM to obtain predictions. Then, we compute the objective function as shown in Equation 5. Finally, the zeroth-order gradient of $\mathbf{z}$ can be obtained through Equation 6, while the gradients of a, b, and α are calculated via backpropagation.

Definition 2 presents an alternative characterization of the stationary point, based on the Moreau envelope. The following lemma reveals its connection to the stationary point defined on the function Φ.

Lemma 2. *(Lemma 3.8 in (Lin, Jin, and Jordan 2020)) If $\mathbf{x}$ is a ϵ-stationary point of a weakly convex function Φ as defined in Definition 2, there exists $\hat{\boldsymbol{v}}$ such that $\min_{\zeta \in \partial\Phi(\hat{\boldsymbol{v}})} \|\zeta\| \leq \epsilon$ and $\|\boldsymbol{v} - \hat{\boldsymbol{v}}\| \leq \epsilon/2\ell$.*

The above lemma indicates that if a point is a ϵ-stationary point as defined in Definition 2, then there must exist another point near it with a subgradient whose norm is less than or equal to ϵ. Therefore, convergence can be measured by the expectation of the norm of the Moreau envelope of a weakly convex function. Since Φ is ℓ-weakly convex (Lemma 1), the stationarity in Definition 2 is the target of our proof. We summarize the convergence result of BPT-FZG below.

Theorem 1. *For Algorithm 1, let $\widehat{\Delta}_\Phi = \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_0) - \min_{\boldsymbol{v}} \Phi_{1/2\ell}(\boldsymbol{v})]$, $\widehat{\Delta}_0 = \mathbb{E}[\Phi(\boldsymbol{v}_0) - f(\boldsymbol{v}_0, \alpha_0)]$, B_1 denote as the bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|]$, B_2 denote as the bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)\|^2]$, B_3 denote as the bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|^2]$ and $\eta_\alpha \leq 1/2\ell$, the following statement holds true,*

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq \frac{4\widehat{\Delta}_\Phi}{(\eta_{\boldsymbol{v}} - c/\ell)(T+1)} + \frac{8\eta_{\boldsymbol{v}}\ell}{\eta_{\boldsymbol{v}} - c/\ell}(\eta_{\boldsymbol{v}} L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) + \frac{8\eta_{\boldsymbol{v}}\ell\widehat{\Delta}_0}{(\eta_{\boldsymbol{v}} - c/\ell)(T+1)} + \frac{4\eta_{\boldsymbol{v}}^2\ell B_2}{c(\eta_{\boldsymbol{v}} - c/\ell)} + \frac{4\eta_{\boldsymbol{v}}^2\ell B_3}{\eta_{\boldsymbol{v}} - c/\ell} \tag{8}$$

where $B_1 = (d+1)L$, $B_2 = \frac{2(d+4)L^2}{m} + \frac{\mu^2\ell^2(d+6)^3}{2m} + \frac{\mu^2\ell^2(d+3)^3}{4} + \frac{\sigma^2}{m}$, and $B_3 = (2d+9)L^2 + \frac{\mu^2\ell^2(d+6)^3}{2}$.

Corollary 1. *Under the same assumptions of Theorem 1, run Algorithm 1 with $\eta_{\boldsymbol{v}} = \min\{\frac{\epsilon^2}{64\ell B_3}, \frac{\eta_\alpha\epsilon^4}{65536\ell^2 D_{\mathcal{A}}{}^2 L B_1}\}, \eta_\alpha = \min\{\frac{1}{2\ell}, \frac{\epsilon^2}{128\ell\sigma^2}\}, B = \frac{D_{\mathcal{A}}}{2}\sqrt{\frac{1}{\eta_{\boldsymbol{v}}\eta_\alpha L B_1}}, \mu = \frac{\epsilon}{2\ell(d+3)^{\frac{3}{2}}}, m = 16\frac{2(d+4)L^2 + \mu^2\ell^2(d+6)^3 + \sigma^2}{\epsilon^2}, T = \min\{\frac{32\widehat{\Delta}_\Phi}{\eta_{\boldsymbol{v}}\epsilon^2}, \frac{64\ell\widehat{\Delta}_0}{\epsilon^2}\}$, then the output of Algorithm 1 satisfies*

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq \epsilon^2 \tag{9}$$

Remark 3. *According to the corollary above, Algorithm 1 is shown to converge.*

Experiments

Setup

Datasets. For comparison, we evaluate the performance of BPT-FZG and baselines in various imbalanced downstream task datasets. We select 3 widely used datasets from GLUE benchmark (Wang et al. 2018), CoLA (Warstadt, Singh, and Bowman 2019), QQP, and MRPC (Dolan and Brockett 2005), and downsample them at given imbalance ratios (majority to minority class samples ratio) $\tau = 20, 50$ to construct imbalanced scenarios. In addition, we also provide results on real-world imbalanced datasets Amazon books and electronics reviews (McAuley et al. 2015). We downsample them at the imbalance ratio $\tau = 10$, which approximates that of the original datasets.

Imbalanced Ratio	Method	RoBERTa-large	GPT2-XL	Llama3
$\tau = 20$	Manual Prompt	.4740±.0982	.5613±.0419	.4744±.1648
	BDPL	.4402±.0472	.4919±.0285	.4583±.1014
	GAP3	.4038±.0314	.4466±.0476	.3484±.0679
	BBT	.5370±.1431	.4326±.0537	.5289±.0149
	BPT-FZG (ours)	**.5454**±.0150	**.5685**±.0544	**.5433**±.1567
$\tau = 50$	Manual Prompt	.5108±.0267	.4823±.1198	.4184±.0805
	BDPL	.4639±.1048	.3888±.0676	.3590±.0204
	GAP3	.3226±.1595	.4219±.1290	.4489±.0932
	BBT	.4398±.0794	.4498±.0422	.3808±.0351
	BPT-FZG (ours)	**.5443**±.2454	**.5116**±.0300	**.5636**±.1796

Table 1: Comparison of AUC scores (mean±std) on constructed imbalanced scenarios of CoLA with a prompt length of 20. The best results are highlighted in **bold**.

Imbalanced Ratio	Method	RoBERTa-large	GPT2-XL	Llama3
$\tau = 20$	Manual Prompt	.4440±.1974	.4643±.0284	.5028±.1192
	BDPL	.3394±.1916	.4731±.0225	.3389±.1043
	GAP3	.3764±.0811	.4963±.0453	.4153±.1368
	BBT	.5343±.1389	.4722±.1258	.4732±.0162
	BPT-FZG (ours)	**.6125**±.1287	**.5954**±.1142	**.5505**±.0883
$\tau = 50$	Manual Prompt	.3983±.2000	.4600±.1698	.5317±.0765
	BDPL	.2200±.0695	.4183±.2288	.4117±.2782
	GAP3	.5517±.3187	.3633±.1855	.4317±.2616
	BBT	.4117±.2903	.4533±.2952	.4750±.0656
	BPT-FZG (ours)	**.6067**±.0957	**.6150**±.1817	**.5667**±.2141

Table 2: Comparison of AUC scores (mean±std) on constructed imbalanced scenarios of MRPC with a prompt length of 20.

Imbalanced Ratio	Method	RoBERTa-large	GPT2-XL	Llama3
$\tau = 20$	Manual Prompt	.4870±.0585	.5121±.0865	.5405±.0244
	BDPL	.4844±.0507	.4923±.0843	.4599±.0489
	GAP3	.3767±.0437	.4995±.0054	.5436±.0074
	BBT	.5066±.0447	.5263±.0493	.4697±.0324
	BPT-FZG (ours)	**.5631**±.0581	**.5546**±.0563	**.5669**±.0104
$\tau = 50$	Manual Prompt	.5160±.0420	.5642±.0994	.5494±.0159
	BDPL	.4966±.0477	.5182±.1414	.5079±.0552
	GAP3	.4668±.0219	.4876±.0609	.5437±.0344
	BBT	.5022±.0957	.5037±.0979	.4793±.0939
	BPT-FZG (ours)	**.5974**±.0178	**.5922**±.1117	**.5495**±.0507

Table 3: Comparison of AUC scores (mean±std) on constructed imbalanced scenarios of QQP with a prompt length of 20.

Baselines. We consider the following black-box prompt tuning methods as our baselines: **Manual Prompt** directly conducts the zero-shot evaluation on LLMs without learning. **BDPL** formulates prompt learning as a distribution optimization problem and optimizes it using the policy gradient.(Diao et al. 2022). **GAP3** leverages an auxiliary LLM to generate prompts from empty and employs a genetic algorithm to select the optimal prompt. (Zhao, Wang, and Yang 2023). **BBT** projects the original space to a subspace using a random matrix and optimizes the prompt in the lower-dimensional space, as previously described (Sun et al. 2022b).

Implementation. All experiments are conducted on a group of NVIDIA A40 GPUs. We select RoBERTa-large (Liu et al. 2019b), GPT2-XL (Radford et al. 2019), Llama3 (AI@Meta 2024) as our backbone models, and all pre-trained model weights are obtained from HuggingFace. We evaluate the performance of BPT-FZG and baselines using AUC, which is commonly used for measuring the classification performance of imbalanced data (Ying, Wen, and Lyu 2016).

Results

In order to validate the effectiveness of BPT-FZG, we conduct comparisons in various imbalanced downstream task datasets against all baseline methods. Additionally, we compared BPT-FZG with common techniques for handling imbalanced data, further demonstrating the effectiveness of BPT-FZG in such contexts.

Comparison on Constructed Imbalanced Scenarios. We perform downsampling on the CoLA, MRPC, and QQP datasets at different imbalanced ratios τ and report the av-

Model	Method	Book		Elec	
		$len = 20$	$len = 50$	$len = 20$	$len = 50$
RoBERTa -large	Manual Prompt	.8932±.0287	.8932±.0287	.8250±.0126	.8250±.0126
	BDPL	.9125±.0201	.9134±.0241	.9273±.0088	.9250±.0123
	GAP3	.9031±.0397	.9031±.0397	.8709±.0334	.8709±.0334
	BBT	.9018±.0267	.8928±.0275	.8542±.0233	.8513±.0120
	BPT-FZG (ours)	**.9185**±.0232	**.9147**±.0161	**.9317**±.0098	**.9270**±.0157
Llama3	Manual Prompt	.6310±.0578	.6310±.0578	.4422±.0603	.4422±.0603
	BDPL	.5870±.0710	.5981±.0744	.4646±.0555	.4683±.0827
	GAP3	.5859±.0236	.5859±.0236	.4060±.0936	.4060±.0936
	BBT	.6007±.0191	.6059±.0638	.5375±.0946	.5518±.0755
	BPT-FZG (ours)	**.6547**±.0179	**.7027**±.1233	**.5937**±.0703	**.6057**±.0937

Table 4: Comparison of AUC values (mean±std) on real-world imbalanced datasets Amazon books and electronics based on RoBERTa-large and Llama3. len represents the prompt length.

erage AUC scores across 3 random seeds in Table 1-3. The experimental results in Table 1-3 demonstrate that BPT-FZG exhibits a performance improvement compared to baselines. The Manual Prompt serves as an initial point for performance; However, in some instances, other baselines underperform relative to the Manual Prompt. For example, When $\tau = 20$, BDPL's performance on MRPC with RoBERTa-large is 0.1046 below that of the Manual Prompt, while BBT decreases by 0.1287 compared to Manual Prompt in the result of CoLA with GPT2-XL. This observation indicates that imbalanced data has a negative impact on prompt tuning in some cases, while also validating that our method is capable of handling imbalanced data.

Comparison on Real-world Imbalanced Datasets. The real-world datasets are inherently imbalanced, with the real imbalance ratios of the Book and Elec datasets being 9.75 and 7.11, respectively. For experimental convenience, we standardize these datasets to $\tau = 10$ and perform downsampling. Table 4 presents the average AUC scores across 3 random seeds for each method on the Book and Elec datasets, with prompt lengths set to 20 and 50. It is observed that GAP3 yields identical results when the prompt length is set to 20 and 50. This occurs because GAP3 generates the optimal prompt from empty rather than optimizing a prompt of fixed length. Under the API query limit, the length of the prompt generated by GAP3 consistently remains below 20. This outcome aligns with the characteristic of genetic algorithms in black-box optimization, which typically requires a high number of queries. We observe similar experimental phenomena on real-world datasets as seen in the above constructed imbalanced scenarios: our method outperforms the baselines, while the performance of methods like BBT falls below that of the Manual Prompt in some cases.

Comparison with Techniques Handling Imbalanced Data. For a fair comparison, we maintain the paradigm of optimizing the latent representation of the prompt within the subspace, formulating prompt tuning as an optimization problem that minimizes the cross-entropy loss. The optimization algorithm is replaced with SGD, which is the natural counterpart of SGDA for minimization problems. We apply two common techniques for handling imbalanced data—oversample and reweight—to this minimization problem and compare the results with those of BPT-FZG. We conduct experiments on 3 backbone models using the same experimental setup. As shown in Table 5, BPT-FZG demonstrates superior performance compared to both oversample and reweight techniques on CoLA. This observation demonstrates that our method can handle imbalanced data more effectively than these common techniques.

Model	Method	Imbalanced Ratio	
		$\tau = 20$	$\tau = 50$
RoBERTa -large	oversample	.4270±.0706	.5036±.1623
	reweight	.5095±.0500	.4553±.0694
	BPT-FZG	**.5454**±.0150	**.5443**±.2454
GPT2-XL	oversample	.5143±.0775	.4514±.0608
	reweight	.5376±.0730	.4610±.1323
	BPT-FZG	**.5685**±.0544	**.5116**±.0300
Llama3	oversample	.5071±.0108	.4377±.0941
	reweight	.5118±.0749	.4738±.0788
	BPT-FZG	**.5433**±.1567	**.5636**±.1796

Table 5: Comparison of AUC scores (mean±std) with techniques handling imbalanced data on CoLA with a prompt length of 20.

Conclusion

In this paper, we propose a novel imbalanced black-box prompt tuning framework BPT-FZG to handle the imbalanced data in specific downstream tasks. BPT-FZG employs a saddle point problem formulation equivalent to AUC maximization as its objective to address the difficulty of constructing opposing sample pairs and utilizes both first and zero-order gradients within the GDA algorithm to optimize the latent representation of the continuous prompt in a low-dimensional subspace. The application of hybrid gradient takes both stability and feasibility into account. Furthermore, we provide the theoretical convergence guarantee for BPT-FZG, which essentially represents a nonconvex-concave minimax optimization. The analysis result indicates that BPT-FZG is capable of convergence, and experimental results demonstrate that our method outperforms baselines on imbalanced datasets.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant (No.2023YFF0905400), National Natural Science Foundation of China through grants (No.U2341229, No.62076138).

References

Aghajanyan, A.; Zettlemoyer, L.; and Gupta, S. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Dan, S.; and Sahoo, D. 2021. Variance reduced stochastic proximal algorithm for auc maximization. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, 184–199. Springer.

Davis, D.; and Drusvyatskiy, D. 2019. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1): 207–239.

Deng, M.; Wang, J.; Hsieh, C.-P.; Wang, Y.; Guo, H.; Shu, T.; Song, M.; Xing, E. P.; and Hu, Z. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diao, S.; Huang, Z.; Xu, R.; Li, X.; Lin, Y.; Zhou, X.; and Zhang, T. 2022. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*.

Dolan, B.; and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Dong, B.; Zhou, P.; Yan, S.; and Zuo, W. 2022. LPT: Long-tailed prompt tuning for image classification. In *The Eleventh International Conference on Learning Representations*.

Gao, W.; Jin, R.; Zhu, S.; and Zhou, Z.-H. 2013. One-pass AUC optimization. In *International conference on machine learning*, 906–914. PMLR.

Gao, Y.; Li, Y.-F.; Lin, Y.; Aggarwal, C.; and Khan, L. 2021. SetConv: A new approach for learning from imbalanced data. *arXiv preprint arXiv:2104.06313*.

Guo, Z.; Liu, M.; Yuan, Z.; Shen, L.; Liu, W.; and Yang, T. 2020. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International conference on machine learning*, 3864–3874. PMLR.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Han, G.; and Zhao, C. 2010. AUC maximization linear classifier based on active learning and its application. *Neurocomputing*, 73(7-9): 1272–1280.

Hanley, J. A.; and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29–36.

Henning, S.; Beluch, W.; Fraser, A.; and Friedrich, A. 2022. A survey of methods for addressing class imbalance in deep-learning based natural language processing. *arXiv preprint arXiv:2210.04675*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hou, B.; O'connor, J.; Andreas, J.; Chang, S.; and Zhang, Y. 2023. Promptboosting: Black-box text classification with ten forward passes. In *International Conference on Machine Learning*, 13309–13324. PMLR.

Hou, W.; Xu, Q.; Yang, Z.; Bao, S.; He, Y.; and Huang, Q. 2022. Adauc: End-to-end adversarial auc optimization against long-tail problems. In *International Conference on Machine Learning*, 8903–8925. PMLR.

Iyer, V.; Chen, P.; and Birch, A. 2023. Towards effective disambiguation for machine translation with large language models. *arXiv preprint arXiv:2309.11668*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Lewis, P.; Ott, M.; Du, J.; and Stoyanov, V. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd clinical natural language processing workshop*, 146–157.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lin, T.; Jin, C.; and Jordan, M. 2020. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 6083–6093. PMLR.

Lin, Z.; Sun, Y.; Shi, Y.; Wang, X.; Huang, L.; Shen, L.; and Tao, D. 2023. Efficient federated prompt tuning for black-box large pre-trained models. *arXiv preprint arXiv:2310.03123*.

Liu, M.; Yuan, Z.; Ying, Y.; and Yang, T. 2019a. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Prasad, A.; Hase, P.; Zhou, X.; and Bansal, M. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Qin, Y.; Wang, X.; Su, Y.; Lin, Y.; Ding, N.; Liu, Z.; Li, J.; Hou, L.; Li, P.; Sun, M.; et al. 2021. Exploring low-dimensional intrinsic task subspace via prompt tuning. *arXiv preprint arXiv:2110.07867*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sun, T.; He, Z.; Qian, H.; Zhou, Y.; Huang, X.; and Qiu, X. 2022a. BBTv2: Towards a gradient-free future with large language models. *arXiv preprint arXiv:2205.11200*.

Sun, T.; Shao, Y.; Qian, H.; Huang, X.; and Qiu, X. 2022b. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, 20841–20855. PMLR.

Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Varia, S.; Wang, S.; Halder, K.; Vacareanu, R.; Ballesteros, M.; Benajiba, Y.; John, N. A.; Anubhai, R.; Muresan, S.; and Roth, D. 2022. Instruction tuning for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2210.06629*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7: 625–641.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wen, Y.; Jain, N.; Kirchenbauer, J.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.

Yan, L.; Dodier, R. H.; Mozer, M.; and Wolniewicz, R. H. 2003. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of the 20th international conference on machine learning (icml-03)*, 848–855.

Yang, T.; and Ying, Y. 2022. AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys*, 55(8): 1–37.

Yang, Z.; Wu, W.; Xu, C.; Liang, X.; Bai, J.; Wang, L.; Wang, W.; and Li, Z. 2020. StyleDGPT: Stylized response generation with pre-trained language models. *arXiv preprint arXiv:2010.02569*.

Yang, Z.; Xu, Q.; Bao, S.; Cao, X.; and Huang, Q. 2021. Learning with multiclass AUC: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7747–7763.

Yang, Z.; Xu, Q.; Bao, S.; He, Y.; Cao, X.; and Huang, Q. 2022. Optimizing two-way partial auc with an end-to-end framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10228–10246.

Yang, Z.; Xu, Q.; Hou, W.; Bao, S.; He, Y.; Cao, X.; and Huang, Q. 2023. Revisiting auc-oriented adversarial training with loss-agnostic perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic online AUC maximization. *Advances in neural information processing systems*, 29.

Yuan, Z.; Guo, Z.; Chawla, N.; and Yang, T. 2021a. Compositional training for end-to-end deep AUC maximization. In *International Conference on Learning Representations*.

Yuan, Z.; Yan, Y.; Sonka, M.; and Yang, T. 2020. Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. *arXiv: Learning,arXiv: Learning*.

Yuan, Z.; Yan, Y.; Sonka, M.; and Yang, T. 2021b. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3040–3049.

Zhao, J.; Wang, Z.; and Yang, F. 2023. Genetic Prompt Search via Exploiting Language Model Probabilities. In *IJCAI*, 5296–5305.

Zheng, Y.; Tan, Z.; Li, P.; and Liu, Y. 2024. Black-box prompt tuning with subspace learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.