# Algorithmic Collective Action in Machine Learning

**Moritz Hardt** [* 1]   **Eric Mazumdar** [* 2]   **Celestine Mendler-Dünner** [* 1]   **Tijana Zrnic** [* 3]

## Abstract

We initiate a principled study of algorithmic collective action on digital platforms that deploy machine learning algorithms. We propose a simple theoretical model of a collective interacting with a firm's learning algorithm. The collective pools the data of participating individuals and executes an algorithmic strategy by instructing participants how to modify their own data to achieve a collective goal. We investigate the consequences of this model in three fundamental learning-theoretic settings: nonparametric optimal learning, parametric risk minimization, and gradient-based optimization. In each setting, we come up with coordinated algorithmic strategies and characterize natural success criteria as a function of the collective's size. Complementing our theory, we conduct systematic experiments on a skill classification task involving tens of thousands of resumes from a gig platform for freelancers. Through more than two thousand model training runs of a BERT-like language model, we see a striking correspondence emerge between our empirical observations and the predictions made by our theory. Taken together, our theory and experiments broadly support the conclusion that algorithmic collectives of exceedingly small fractional size can exert significant control over a platform's learning algorithm.

## 1. Introduction

Throughout the gig economy, numerous digital platforms algorithmically profile, control, and discipline workers that offer on-demand services to consumers. Data collection and predictive modeling are critical for a typical platform's

---
[*]Alphabetical ordering [1]Max Planck Institute for Intelligent Systems, Tübingen, and Tübingen AI Center [2]Caltech [3]UC Berkeley. Correspondence to: Moritz Hardt <hardt@is.mpg.de>, Eric Mazumdar <mazumdar@caltech.edu>, Celestine Mendler-Dünner <cmendler@tuebingen.mpg.de>, Tijana Zrnic <tijana.zrnic@berkeley.edu>.

business as machine learning algorithms power ranking, scoring, and classification tasks of various kinds (Woodcock & Graham, 2019; Gray & Suri, 2019; Schor, 2021).

Troves of academic scholarship document the emergence and preponderance of precarity in the gig economy. Wood et al. (2019) argue that platform-based algorithmic control can lead to "low pay, social isolation, working unsocial and irregular hours, overwork, sleep deprivation and exhaustion." This is further exacerbated by "high levels of inter-worker competition with few labor protections and a global over-supply of labor relative to demand." In response, there have been numerous attempts by gig workers to organize in an effort to reconfigure working conditions. A growing repertoire of strategies, as vast as it is eclectic, uses both physical and digital means towards this goal. Indeed, workers have shown significant ingenuity in creating platform-specific infrastructure, such as their own mobile apps, to organize the labor side of the platform (Chen, 2018; Rahman, 2021). Yet, "the upsurge of worker mobilization should not blind us to the difficulties of organizing such a diverse and spatially dispersed labor force." (Vallas & Schor, 2020)

Beyond the gig economy, evidence of consumers seeking to influence the algorithms that power a platform's business is abundant. Examples include social media users attempting to suppress the algorithmic upvoting of harmful content by sharing screenshots rather than original posts (Burrell et al., 2019), or individuals creating bots to influence crowd-sourced navigation systems (Sinai et al., 2014). The ubiquity of such strategic attempts calls for a principled study of how coordinated groups can wield control over the digital platforms to which they contribute data.

In this work, we study how a collective of individuals can algorithmically strategize against a learning platform. We envision a collective that pools the data of participating individuals and executes an algorithmic strategy by instructing participants how to modify their own data. The firm in turn solves a machine learning problem over the resulting data. The goal of the collective is to redirect the firm's optimization towards a solution that serves the collective. Notably, coordination is a crucial lever. When data are plentiful, a single individual lacks the leverage to unilaterally change the output of a learning algorithm; in contrast, we show that even small collectives can exert substantial influence.

## 1.1. Our Contribution

We initiate a principled study of algorithmic collective action in digital platforms that deploy machine learning algorithms. We propose a simple theoretical model of a collective interacting with a firm's learning algorithm. The size of the collective is specified by a value $\alpha > 0$ that corresponds to the fraction of participating individuals in a population drawn from a base distribution $\mathcal{P}_0$. The firm observes the mixture distribution

$$\mathcal{P} = \alpha \mathcal{P}^* + (1 - \alpha)\mathcal{P}_0,$$

where $\mathcal{P}^*$ depends on the strategy of the collective, and the firm runs a learning algorithm $\mathcal{A}$ on $\mathcal{P}$.

We investigate the consequences of this model in three fundamental learning-theoretic settings, supported by a systematic empirical evaluation on a real resume classification task from a gig platform. In each setting, we come up with algorithmic collective strategies and characterize different success criteria as a function of the collective's size $\alpha$. We exhibit *critical thresholds* for the value of $\alpha$ at which the collective succeeds. Our theory and experiments support the conclusion that collectives of vanishing fractional size can exert significant control over the firm's learning algorithm.

**Classification.** In line with economic theory, we start with the case of an optimal firm that has full knowledge of the distribution $\mathcal{P}$. The firm chooses the Bayes optimal predictor $f = \mathcal{A}(\mathcal{P})$ on the distribution $\mathcal{P}$. In the context of classification, a natural class of objectives for the collective is to correlate a signal $g(\cdot)$ with a target label $y^*$: an individual described by data point $(x, y)$ succeeds if $f(g(x)) = y^*$ at test time. For this goal, a natural strategy is to perturb each participating data point $(x, y)$ by applying the signal $g(\cdot)$ to $x$ and switching the label from $y$ to $y^*$ at training time. That is, $\mathcal{P}^*$ is the distribution of $(g(x), y^*)$ for a random draw of a labeled data point $(x, y)$ from $\mathcal{P}_0$. We prove that a collective of vanishing fractional size succeeds with high probability by implementing this strategy, provided that the signal $g(x)$ is unlikely to be encountered in the base distribution $\mathcal{P}_0$. The success probability is maximized for an optimal classifier and deteriorates gracefully with the suboptimality of the classifier.

In practice, the signal $g(x)$ may correspond to adding a hidden watermark in image and video content, or subtle syntactic changes in text. It is reasonable to assume that individuals are indifferent to such inconsequential changes to their features. In fact, conventional wisdom in machine learning has it that such hidden signals are easy to come by in practice. The ability to change the label of an instance, however, is a more strenuous requirement. We therefore propose a variant of the strategy where a participating individual does not need to modify the label of their data point. We show that this strategy still succeeds, while quantifying precisely how it diminishes the collective's success rate as a function of its size and a key parameter of $\mathcal{P}_0$.

We provide analogous results when the collective's goal is for the firm's predictions to ignore some subset of the feature information. Given a map $g(x)$, the collective succeeds if $f(x) = f(g(x))$. Here, $g(x)$ is a summary of $x$ that, for example, removes private or sensitive information in $x$.

**Experimental evaluation.** We conduct extensive experiments on a dataset of almost thirty thousand resumes from a gig platform for freelancers. The machine learning task is to tag resumes with a set of ten skills related to work in the IT sector. Through more than two thousand model training runs involving a BERT-like language model, we investigate the predictions made by our theory. What emerges is a striking correspondence between our empirical findings and the theory. The ability to modify both resume and label leads to a near-perfect success rate at collective sizes of only a fraction of a percent of the population, corresponding to fewer than one hundred modified resumes. The label-free strategy still succeeds reliably, albeit at a higher threshold, corresponding to a few hundred resumes. The more well-trained the model is, the lower the threshold for the collective's success. The placement pattern of the signal in the resume is largely irrelevant, so long as the token we plant is unique within the corpus of resumes.

Our theory predicts that collectives, in a certain precise sense, must compete with the strongest alternative signal in the data. The weaker the alternative signals, the lower the threshold for success. To confirm this hypothesis experimentally, we randomize the labels of a random fraction of the data. Confirming our theory, we observe that increasing the fraction of randomized labels, and hence diminishing the strength of alternative signals, indeed lowers the threshold for success of the collective. This observation suggests a blessing of dimensionality: if the data contains many weak signals, as high-dimensional data tends to, algorithmic collectives are especially powerful.

**Risk minimization and gradient descent.** Generalizing beyond classification, we consider the power of collectives in convex risk minimization and gradient-based learning with possibly nonconvex objectives. In the first case, we show that collectives can force the firm to deploy any model with small suboptimality on $\mathcal{P}_0$ of the collective's choosing. In the second case, we show that given a more powerful interaction model the collective can influence the firm to deploy any desired model even under nonconvexity, as long as the path from the initial model, computed on $\mathcal{P}_0$, to the desired model does not traverse large gradients when evaluated on $\mathcal{P}_0$. Moreover, despite the nonconvexity, convergence to the target is achieved at a convex rate. In both problem settings, the analyzed collective strategies rely on exerting influence on the gradients computed by the firm.

## 1.2. Related Work

Our approach to algorithmic collective action is decidedly not adversarial. Instead, the strategic manipulations arise through a misalignment of the firm's and the individuals' objectives. Individuals legitimately optimize their utility through data sharing and coordination. Yet, at a technical level our results relate to topics studied under the umbrella of adversarial machine learning. Most closely related is the line of work on *data poisoning attacks* that seeks to understand how data points can be adversarially "poisoned" to degrade the performance of a predictive model at test time. We refer to recent surveys for an overview of data poisoning attacks (Tian et al., 2022), and backdoor attacks more specifically (Guo et al., 2022). While the literature on poisoning attacks focuses predominantly on diminishing the performance of the learning algorithm, documented empirical successes (Cherepanova et al., 2021; Geiping et al., 2021) hint at the impact that algorithmic collective action can have on deep learning models. Despite the increasing number of studies on backdoor attacks and defenses, theoretical work explaining how underlying factors affect the success of backdoor attacks has been limited (Grosse et al., 2022). We point out two recent works that study the relationship between the number of manipulated data points and the success probability of the attack. Manoj & Blum (2021) show that a fixed number of points is sufficient for backdoor attacks to succeed in binary classification problems, provided that the memorization capacity of the model is sufficiently large. Cinà et al. (2021) empirically investigate backdoor learning curves across many image recognition tasks and they observe curves with diminishing returns, similar in shape to those in our experiments. Our analysis of Bayes optimal classification provides a new, complementary theoretical perspective and sheds light on the effectiveness of practical data manipulation strategies in a way that is surprisingly predictive of our empirical observations and previous empirical results in adversarial machine learning.

Our analysis of risk minimization is reminiscent of model-targeted attacks (Suya et al., 2021; Farhadkhani et al., 2022), which aim to bias the learner towards a target model. Our gradient-control strategy resembles the counter-gradient attack in (Farhadkhani et al., 2022). While the insights from these prior works are valuable to inform the feasibility of collective action in convex risk minimization, our work differs from these existing studies in its focus on the role of collective size and the analysis of nonconvex losses. Only a handful of works in the adversarial machine learning literature have questioned the institution-centric perspective of the community and discussed the political dimension of adversarial machine learning (Albert et al., 2020; Vincent et al., 2021; Albert et al., 2021). In this context, a recent line of work considers the socially beneficial use of adversarial learning techniques (Delobelle et al., 2021; Shan et al., 2020; Abebe et al., 2022; Kulynych et al., 2020; Li et al., 2022).

Our work can also be viewed as a conceptual reversal of *strategic classification* (Hardt et al., 2016). In strategic classification, a firm anticipates the optimal response of a strategic individual to a decision rule. Instead, we consider individuals that strategically anticipate the optimizing behavior of the firm, something recently considered by Zrnic et al. (2021). Furthermore, our work is conceptually different from strategic classification in its focus on the role and objectives of workers and consumers on online platforms rather than the firm. Another crucial departure from the set of problems considered in strategic classification is our emphasis on collective rather than individual strategies.

The idea of collective action on digital platforms has also been previously studied. Creager & Zemel (2021) show how algorithmic recourse can be improved through coordination. Vincent et al. (2019) examine the effectiveness of *data strikes*. Extending this work to the notion of *data leverage*, Vincent et al. (2021) describe various ways of "reducing, stopping, redirecting, or otherwise manipulating data contributions" for different purposes. See also Vincent & Hecht (2021). Our work provides a theoretical framework for understanding the effectiveness of such strategies, as well as studying more complex algorithmic strategies that collectives may deploy.

Appendix A continues our discussion of related work.

## 2. Problem Formulation

We study the strategic interaction of a firm operating a predictive system with a population of individuals. We assume that the firm deploys a learning algorithm $\mathcal{A}$ that operates on data points in a universe $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Each individual corresponds to a single data point $z \in \mathcal{Z}$, typically a feature–label pair. We model the population of individual participants as a distribution $\mathcal{P}_0$ over $\mathcal{Z}$.

We say that a fraction $\alpha > 0$ of the individuals form a *collective* in order to strategically respond to the firm's learning behavior. The collective agrees on a potentially randomized strategy $h : \mathcal{Z} \to \mathcal{Z}$ from a space of available strategies $\mathcal{H}$. The possible strategies $\mathcal{H}$ capture feasible changes to the data. For example, content creators on a video streaming platform may be indifferent between videos that differ only in a hidden watermark not visible to human viewers. Freelancers may be indifferent between two resumes that differ only in inconsequential syntactic details.

The firm therefore observes a mixture distribution

$$\mathcal{P} = \alpha\mathcal{P}^* + (1 - \alpha)\mathcal{P}_0,$$

where we use $\mathcal{P}^*$ to denote the distribution of $h(z), z \sim \mathcal{P}_0$.

The collective strives to choose a strategy $h$ so as to maximize a measure of success over the solution $f = \mathcal{A}(\mathcal{P})$ chosen by the firm. Here, $f$ is a mapping from features to labels, $f \colon \mathcal{X} \to \mathcal{Y}$. Given a strategy, we use $S(\alpha)$ to denote the level of success achieved by a collective of size $\alpha$. The central question we study is how the success $S(\alpha)$ grows as a function of collective size $\alpha$, and how large $\alpha$ needs to be in order to achieve a target success level.

**Definition 2.1** (Critical mass). The critical mass for a target success level $S^*$ is defined as the smallest $\alpha$ for which there exists a strategy such that $S(\alpha) \geq S^*$.

Note that, although motivated from the perspective of labor, our formal model can also serve as a basis for studying collective action on the consumer side of digital platforms. Before presenting our results we briefly discuss why we focus on collective strategies in this paper.

**Why collective action?** By engaging in collective action, individuals can exert influence on the learning algorithm that they could not achieve by acting selfishly. In large-population settings such as online platforms, an individual contributes an infinitesimal fraction of the data used by the learning algorithm. Thus, under reasonable manipulation constraints, individual behavior is largely powerless in systematically changing the deployed model. Instead, individuals are limited to simple adversarial attacks or individual strategies that do not have lasting effects on the learning outcome. By coordinating individuals, however, collectives can wield enough power to steer learning algorithms towards desired goals. In subsequent sections we show that collectives can often do so while only representing a small fraction of the training data.

## 3. Collective Action in Classification

We start with classification under the assumption that the firm chooses an approximately optimal classifier on the data distribution $\mathcal{P}$.

**Definition 3.1** ($\epsilon$-optimal classifier). A classifier $f \colon \mathcal{X} \to \mathcal{Y}$ is $\epsilon$-optimal under the distribution $\mathcal{P}$ if there exists a $\mathcal{P}'$ with $\mathrm{TV}(\mathcal{P}, \mathcal{P}') \leq \epsilon$ such that

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \, \mathcal{P}'(y|x) \,.$$

Note that a $0$-optimal classifier is the Bayes optimal classifier with respect to the zero–one loss.

Under the above assumption, we focus on two general goals for the collective: *planting a signal* and *erasing a signal*.

### 3.1. Planting a Signal

Assume the collective wants the classifier to learn an association between an altered version of the features $g(x)$ and a chosen target class $y^*$. Formally, given a transformation $g \colon \mathcal{X} \to \mathcal{X}$, the collective wants to maximize the following measure of success:

$$S(\alpha) = \Pr_{x \sim \mathcal{P}_0} \{ f(g(x)) = y^* \} \,.$$

We call this objective "planting a signal" and $\mathcal{X}^* = \{ g(x) \colon x \in \mathcal{X} \}$ the signal set. For example, $g(x)$ could be instance $x$ with an inconsequential trigger (such as a video with an imperceptible watermark or a resume with a unique formatting) and $y^*$ could be a label indicating that the instance is of high quality (e.g., a high-quality video or a highly qualified individual). As another example, the collective may have an altruistic goal to help individuals in a vulnerable subpopulation $\mathcal{X}_0 \subseteq \mathcal{X}$ achieve a desired outcome $y^*$. In this case, $g(x)$ could be a mapping from $x$ to a randomly chosen instance in $\mathcal{X}_0$.

We provide natural strategies for planting a signal and characterize their success as a function of $\alpha$. The key parameter that we identify as driving success is the *uniqueness* of the signal.

**Definition 3.2** ($\xi$-unique signal). We say that a signal is $\xi$-unique if it satisfies $\mathcal{P}_0(\mathcal{X}^*) \leq \xi$.

In addition, success naturally depends on how suboptimal $y^*$ is on the signal set under the base distribution. To formalize this dependence, we define the suboptimality gap of $y^*$ as

$$\Delta = \max_{x \in \mathcal{X}^*} \left( \max_{y \in \mathcal{Y}} \mathcal{P}_0(y|x) - \mathcal{P}_0(y^*|x) \right) \,.$$

We consider two possibilities for the space of available strategies $\mathcal{H}$. First, we assume that the individuals can modify both features and labels. We call the resulting strategies *feature–label* strategies. Modifying features by, say, planting a trigger often comes at virtually no cost. Changing the label, however, may be hard, costly, or even infeasible. This is why we also consider *feature-only* strategies; such strategies only allow changes to features.

**Feature–label signal strategy.** We define the feature–label signal strategy as

$$h(x, y) = (g(x), y^*) \,. \tag{1}$$

The result below quantifies the success of this strategy in terms of the collective size and the uniqueness of the signal.

**Theorem 3.3.** *Consider the feature–label signal strategy and suppose that the signal is $\xi$-unique. Then, the success against an $\epsilon$-optimal classifier is lower bounded by*

$$S(\alpha) \geq 1 - \frac{1-\alpha}{\alpha} \cdot \Delta \cdot \xi - \frac{\epsilon}{1-\epsilon} \,.$$

Rearranging the terms, we obtain an upper bound on the critical mass given a desired success probability (e.g., 90%).
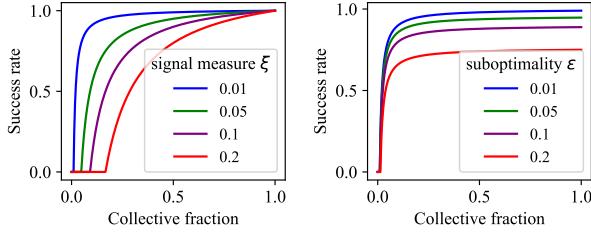
*Figure 1.* Illustration of the success rate predicted by Theorem 3.3. In the first we fix $\epsilon = 0$ and vary $\xi$, and in the second we fix $\xi$ and vary the classifier's suboptimality, $\epsilon$. We upper bound $\Delta$ by one.

**Corollary 3.4.** *Suppose the signal is $\xi$-unique. Then, the critical mass for achieving success $S^* \in (0, 1)$ with feature–label strategies against an $\epsilon$-optimal classifier is bounded by*

$$\alpha^* \leq \frac{\Delta \cdot \xi}{1 - S^* - \frac{\epsilon}{1-\epsilon} + \Delta \cdot \xi}. \tag{2}$$

Therefore, in order to achieve success it suffices to have a collective size proportional to the uniqueness of the signal and the suboptimality of $y^*$ on the signal set, as long as these parameters are sufficiently small relative to the target error rate $1 - S^*$. This suggests that planting signals that are exceedingly "rare" under the base distribution can be done successfully by small collectives— a property of feature–label strategies that we empirically validate in Section 5.

In the next result we consider feature-only strategies. An impediment to the success of such strategies is the situation where there is overwhelmingly strong signal about a label $y \neq y^*$ in the base distribution and hence little label uncertainty. This is the reason why we make one additional assumption that there exists a number $p > 0$ such that $\mathcal{P}_0(y^*|x) \geq p$ for all $x \in \mathcal{X}$.

**Feature-only signal strategy.** We define the feature-only signal strategy as

$$h(x, y) = \begin{cases} (g(x), y^*), & \text{if } y = y^*, \\ (x, y), & \text{otherwise.} \end{cases} \tag{3}$$

This strategy achieves a similar success rate as the feature–label strategy, but the success diminishes with the constant $p$.

**Theorem 3.5.** *Consider the feature-only signal strategy and suppose that the signal is $\xi$-unique. Further, suppose there exists $p > 0$ such that $\mathcal{P}_0(y^*|x) \geq p, \forall x \in \mathcal{X}$. Then, the success against an $\epsilon$-optimal classifier is lower bounded by*

$$S(\alpha) \geq 1 - \frac{1-p}{p\alpha} \cdot \xi - \frac{\epsilon}{1-\epsilon}.$$

The critical mass for achieving a target success probability is thus bounded as follows.

**Corollary 3.6.** *Suppose the signal is $\xi$-unique. Then, the critical mass for achieving success $S^* \in (0, 1)$ with feature-only strategies against an $\epsilon$-optimal classifier is bounded by*

$$\alpha^* \leq \frac{1-p}{p} \frac{\xi}{1 - S^* - \frac{\epsilon}{1-\epsilon}}.$$

The positivity constant $p > 0$ may be excessively small over the entire data universe. A standard fix to this problem is to restrict $\mathcal{P}_0$ to a subset where the constant is larger, and pay a penalty for the amount of truncation in the bound. For example, if there exists $R \subseteq \mathcal{X}$ such that $\mathcal{P}_0(R) \geq 99\%$, but the positivity constant over $R$ is much larger than $p$, then one can obtain a more powerful version of Theorem 3.5.

### 3.2. Erasing a Signal

Next, we assume the collective wants the classifier to be invariant under a transformation $g : \mathcal{X} \to \mathcal{X}$ of the features. In particular, the success is measured with respect to:

$$S(\alpha) = \Pr_{x \sim \mathcal{P}_0} \{f(x) = f(g(x))\}.$$

In other words, the collective wants the classifier to output the same predictions for all $x$ and $x'$ that have $g(x) = g(x')$. The map $g$ can be thought of as a summary of $x$ that removes some feature information. We call this objective "erasing a signal." For example, if the collective wants the deployed model to be insensitive to the value of a particular feature $j^*$, then it can use $g(x) = x'$ where $x'_j = x_j$ for $j \neq j^*$ and $x'_{j^*} = 0$. The feature $j^*$ could be the length of a video that content creators do not want to affect the ranking of the content, or it could be a sensitive demographic feature that a collective wants to be independent of the predicted label.

**Erasure strategy.** We define the erasure strategy as

$$h(x, y) = (x, \operatorname*{argmax}_{y \in \mathcal{Y}} \mathcal{P}_0(y|g(x))).$$

As before, the success of the strategy depends on problem-dependent quantities. In this case, the quantity of interest is the sensitivity of the labels to the erased signal. We capture this sensitivity in the parameter $\tau$, defined as

$$\tau := \mathbb{E}_{x \sim \mathcal{P}_0} \max_{y \in \mathcal{Y}} |\mathcal{P}_0(y|x) - \mathcal{P}_0(y|g(x))|.$$

Intuitively, $\tau$ is small if observing the whole feature vector $x$, instead of just the summary $g(x)$, reveals little additional information about the label.

**Theorem 3.7.** *Consider the erasure strategy. Then, the success against an $\epsilon$-optimal classifier is lower bounded by*

$$S(\alpha) \geq 1 - \frac{2(1-\alpha)}{\alpha} \cdot \tau - \frac{\epsilon}{1-\epsilon}.$$

5

We rearrange the terms and derive a bound on the critical mass that guarantees a signal can be erased with a desired probability.

**Corollary 3.8.** *The critical mass for achieving success* $S^* \in (0, 1)$ *is bounded by*

$$\alpha^* \leq \frac{\tau}{\frac{1}{2}(1 - S^*) - \frac{\epsilon}{2(1-\epsilon)} + \tau}.$$

The less sensitive the labels to the erased information, the smaller the collective needed to successfully enforce a decision rule independent of the protected information.

In contrast to the strategies in Section 3.1, the erasure strategy requires knowledge of statistics about $\mathcal{P}_0$. This highlights an important benefit of collective action: information sharing. Information about the base distribution is typically difficult to obtain for individual platform users. However, a collective can pool their feature–label information to estimate properties of the distribution from samples; the larger the collective, the better the estimate and consequently the more effective the strategy.

# 4. Collective Action in Risk Minimization

We next study the effect of collective size when the learner is solving parametric risk minimization. Here the firm is choosing a model from a parameterized set $\{f_\theta\}_{\theta \in \Theta}$. We will use $\mathcal{A}(\mathcal{P})$ to denote an element in $\Theta$ that determines the model chosen by the firm. We begin by studying convex risk minimizers. Then, motivated by nonconvex settings, we look at gradient-descent learners without imposing any convexity assumptions on the objective. Our main working assumption will be that of a risk-minimizing firm.

**Definition 4.1** (Risk minimizer). Fix a loss function $\ell$. The firm is a risk minimizer if under distribution $\mathcal{P}$ it determines the parameter of the model $f_\theta$ according to

$$\theta = \underset{\theta' \in \Theta}{\operatorname{argmin}} \ \underset{z \sim \mathcal{P}}{\mathbb{E}} \ \ell(\theta'; z).$$

## 4.1. Convex Risk Minimization

To begin, we assume that $\ell(\theta; z)$ is a convex function of $\theta$, and that the collective's goal is to move the model from $\theta_0$—the optimal model under the base distribution $\mathcal{P}_0$—to a target model $\theta^*$. To that end, for a given target model $\theta^* \in \Theta$, we measure success in terms of

$$S(\alpha) = -\|\theta - \theta^*\|.$$

Here, $\|\cdot\|$ can be any norm (as long as it is kept fixed in the rest of the section). In line with first-order optimality conditions for convex optimization, the influence of the collective on the learning outcome depends on the collective's ability to influence the average gradient of $\ell$. To simplify notation, let $g_\mathcal{P}(\theta') = \mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta'; z)$ denote the expected gradient of the loss over distribution $\mathcal{P}$ measured at a point $\theta' \in \Theta$.

**Gradient-neutralizing strategy.** Define the gradient-neutralizing strategy as follows. Find a *gradient-neutralizing* distribution $\mathcal{P}'$ for $\theta^*$, meaning $\angle(g_{\mathcal{P}'}(\theta^*), -g_{\mathcal{P}_0}(\theta^*)) = 0$. Then, draw $z' \sim \mathcal{P}'$ and let

$$h(z) = \begin{cases} z', & \text{with prob. } \min\left(1, \frac{1}{\alpha} \frac{\|g_{\mathcal{P}_0}(\theta^*)\|}{\|g_{\mathcal{P}'}(\theta^*)\| + \|g_{\mathcal{P}_0}(\theta^*)\|}\right), \\ z, & \text{else.} \end{cases}$$

For example, in generalized linear models (GLMs) gradients are given by $\nabla \ell(\theta; (x, y)) = x(\mu_\theta(x) - y)$, where $\mu_\theta(\cdot)$ is a mean predictor (see, e.g., Chapter 3 in (Efron, 2022)). Therefore, one can obtain a gradient-neutralizing distribution by simply letting $h(x, y) = (x', y')$, where $x' = -g_{\mathcal{P}_0}(\theta^*)$ and $y'$ is any value less than $\mu_\theta(x')$. Alternatively, if the collective is restricted to feature-only strategies, they can set $x' = -g_{\mathcal{P}_0}(\theta^*)$ only if $y < \mu_\theta(x')$, and $x' = 0$ otherwise. As long as the label distribution has sufficiently large support under $\mathcal{P}_0$, in particular $y < \mu_\theta(-g_{\mathcal{P}_0}(\theta^*))$ with nonzero probability, this strategy likewise results in a gradient-neutralizing distribution.

**Theorem 4.2.** *Suppose there exists a gradient-neutralizing distribution $\mathcal{P}'$ for $\theta^*$. Then, if the loss is $\mu$-strongly convex, the success of the gradient-neutralizing strategy is bounded by*

$$S(\alpha) \geq \frac{1}{\mu} \min \left( \alpha \|g_{\mathcal{P}'}(\theta^*)\| - (1 - \alpha) \|g_{\mathcal{P}_0}(\theta^*)\|, 0 \right).$$

The natural target success for the collective is for $\theta^*$ to be reached exactly; this is achieved when $S(\alpha) = 0$.

**Corollary 4.3.** *Suppose there exists a gradient-neutralizing distribution $\mathcal{P}'$ for $\theta^*$. Then, for any $\mu \geq 0$ the critical mass for achieving target success $S(\alpha) = 0$ is bounded by*

$$\alpha^* \leq \frac{\|g_{\mathcal{P}_0}(\theta^*)\|}{\|g_{\mathcal{P}'}(\theta^*)\| + \|g_{\mathcal{P}_0}(\theta^*)\|}. \tag{4}$$

Even if $\ell$ is only strictly convex ($\mu \to 0$), the collective can reach $\theta^*$ with $\alpha^*$ as in (4). Note that this corollary reveals an intuitive relationship between $\alpha^*$ and $g_{\mathcal{P}_0}(\theta^*)$ in the convex regime: target models $\theta^*$ that look more optimal to the learner under the base distribution are easier to achieve.

If the collective has a utility function $u(\theta')$ that specifies the participants' preferences over different models $\theta'$, a natural way to define success is via a desired gain in utility: $S(\alpha) = u(\theta) - u(\theta_0)$, where $\theta_0 = \mathcal{A}(\mathcal{P}_0)$. Corollary 4.3 implies a bound on the critical mass for this measure of success, for all convex utilities (for example, linear utilities of the form $u(\theta) = \theta^\top v$, for some $v$).
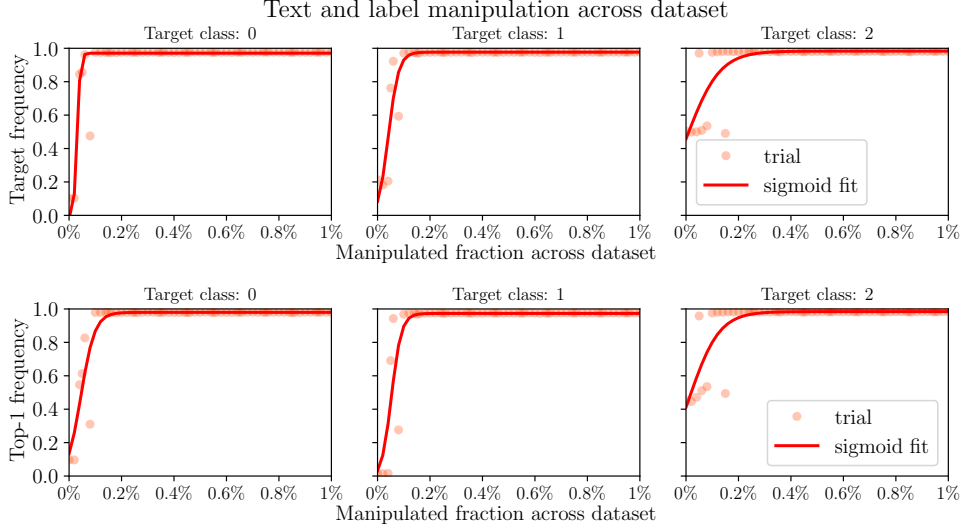
Text and label manipulation across dataset



*Figure 2.* Success rate of Strategy 1 as the collective size varies.

**Proposition 4.4.** *Suppose that $u(\theta')$ is convex. Further, assume $\ell$ is $\beta$-smooth and that $\|\cdot\|$ is the $\ell_2$-norm. Then, the critical mass for achieving $u(\theta) - u(\theta_0) \geq U$ is bounded by*

$$\alpha^* \leq \frac{\beta \cdot U}{g_{\text{lb}} \cdot \|\nabla u(\theta_0)\| + \beta \cdot U},$$

*where $g_{\text{lb}} = \min\{\|g_{\mathcal{P}'}(\theta')\| : \|\theta' - \theta_0\| \leq U/\|\nabla u(\theta_0)\|\}$ and $\mathcal{P}'$ is gradient-neutralizing for $\theta'$.*

As a result, the critical mass required to achieve a utility gain of $U$ decreases as the gradient of the utility at $\theta_0$ grows. Intuitively, large $\|\nabla u(\theta_0)\|$ means that small changes to $\theta_0$ can lead to large improvements for the collective.

### 4.2. Gradient-Based Learning

So far we have assumed that exact optimization is computationally feasible; with nonconvex objectives, this behavior is hardly realistic. A common approach to risk minimization for general, possibly nonconvex learning problems is to run gradient descent. Formally, at each step $t$ we assume the learner observes the current data distribution $\mathcal{P}_t$, computes the average gradient at the current iterate, and updates the model by taking a gradient step:

$$\theta_{t+1} = \theta_t - \eta \cdot g_{\mathcal{P}_t}(\theta_t),$$

where $\eta > 0$ is a fixed step size. Given a target model $\theta^*$, we define the success of the strategy after $t$ steps as

$$S_t(\alpha) = -\|\theta_t - \theta^*\|.$$

Given the online nature of the learner's updates, we assume that the collective can *adaptively* interact with the learner; that is, they can choose $\mathcal{P}_t^*$ at every step $t$. This is a typical

interaction model in federated learning (McMahan et al., 2017). In the following we show that this additional leverage enabled by this adaptivity allows the collective to implement a refined strategy that controls the outcome of learning even in nonconvex settings.

**Gradient-control strategy.** We define the gradient-control strategy at $\theta$ as follows. Given the observed model $\theta$ and a target $\theta^*$, the collective finds a *gradient-redirecting distribution* $\mathcal{P}'$ for $\theta$, meaning $g_{\mathcal{P}'}(\theta) = -\frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}(\theta) + \xi(\theta - \theta^*)$, for some $\xi \in (0, \frac{1}{\alpha\eta})$. Then, draw $z' \sim \mathcal{P}'$ and set $h(z) = z'$.

**Theorem 4.5.** *Assume the collective can implement the gradient-control strategy at all $\lambda\theta_0 + (1-\lambda)\theta^*, \lambda \in [0,1]$. Then, there exists a $C(\alpha) > 0$ such that the success of the gradient-control strategy after $T$ steps is lower bounded by*

$$S_T(\alpha) \geq -(1 - \eta C(\alpha))^T \cdot \|\theta_0 - \theta^*\|.$$

The above result implies that the collective can reach any model $\theta^*$ as long as there exists a path from $\theta_0$ to $\theta^*$ that only traverses small gradients on the initial distribution $\mathcal{P}_0$.

## 5. Experiments

We report on experimental findings from over 2000 model training runs involving a BERT-like text transformer model on a resume classification task. The resume dataset consists of nearly 30000 resumes of freelancers on a major gig labor platform, introduced by (Jiechieu & Tsopze, 2021). The task is a multiclass, multilabel classification problem where the goal is to predict a set of up to ten skills from the software and IT sector based on the resume.

The collective controls a random fraction of the training data within the dataset. Its goal is to plant a signal, that is, steer
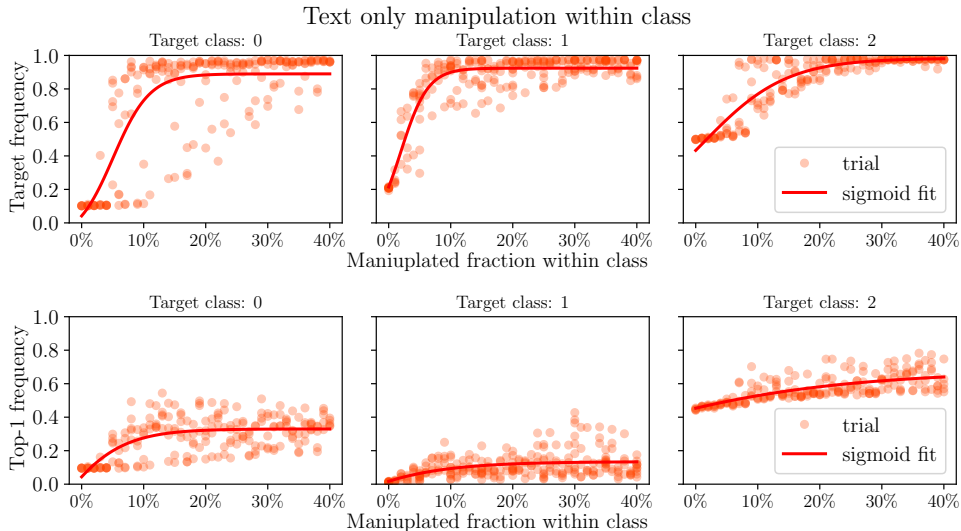
Text only manipulation within class



*Figure 3.* Success rate of Strategy 2 as the collective size varies.

the model's predictions on transformed data points $g(x)$ toward a desired target label $y^*$. We evaluate the effectiveness of two simple strategies, which are instantiations of the feature–label and feature-only strategies from Section 3.1.

**Strategy 1 (Text and label manipulation across dataset).** The collective plants a special token in the resume text and changes the label to the target class. This strategy closely mirrors the feature-label signal strategy in (1).

**Strategy 2 (Text-only manipulation within target class).** The collective manipulates the resume text of resumes within the target class by inserting a special token with some frequency (every 20th word). This strategy closely follows the feature-only signal strategy in (3).

**Evaluation.** To measure success we insert the special token in all test points and count how often the model (a) includes the target class in its set of predicted skills, (b) has the target class as its "top-1" prediction.

### 5.1. Experimental Setup

We use 'distilbert-base-uncased', a standard pre-trained transformer model. We fine-tune it on the dataset for 5 epochs with standard hyperparameters. After 5 epochs, the model plateaus at around 97% multi-label accuracy (defined as 1 minus Hamming loss), 93.8% precision, and 88.9% recall. The dataset contains 29783 resumes, of which we use 25000 for training and 4783 for testing. We focus on the first three classes of the problem, corresponding to *database administrator* (class 0), *web developer* (class 1), *software developer* (class 2). These three classes occur with frequency 0.11, 0.23, and 0.5, respectively, in the dataset. As the special token, we use an unused formatting symbol (token 1240 being a small dash) that we insert every 20 words.

### 5.2. Experimental Findings

**Text and label manipulation across dataset.** We find that Strategy 1 exerts significant control over the model's prediction even when the collective is exceedingly small (Figure 2). In fact, we see consistent success in controlling the model's output well below 0.5% of the dataset, i.e., fewer than 125 manipulated training data points.

**Text-only manipulation within target class.** We find that Strategy 2 consistently succeeds in controlling the model so as to include the target class in its positive predictions. The strategy succeeds at a threshold of around 10% of the instances of the target class (Figure 3, top panel). This threshold corresponds to approximately 1%, 2%, and 5% of the dataset for class 0, 1, and 2, respectively. When it comes to controlling the model's top prediction, the text-only strategy does *not* consistently succeed (Figure 3, bottom panel).

**Effect of positivity constant.** Our theory in Section 3.1 suggests that the difficulty of controlling the model's top prediction via the text-only strategy may be due to a small positivity constant $p$. To evaluate this hypothesis, we repeat our experiments after we randomize a random fraction of the labels in the training data. This randomization ensures that each feature vector is assigned the target label with nontrivial probability. Our findings confirm that even a small fraction of random labels dramatically increases the success of Strategy 2 in controlling the top prediction (Figure 4).

**Trade-offs between model optimality and success.** Figure 5 shows that the success of either strategy is sensitive to the number of epochs. Less optimization during the model training phase leads to a lower success rate. These findings mirror our theoretical results: as the model approaches optimality, small collectives have significant power.
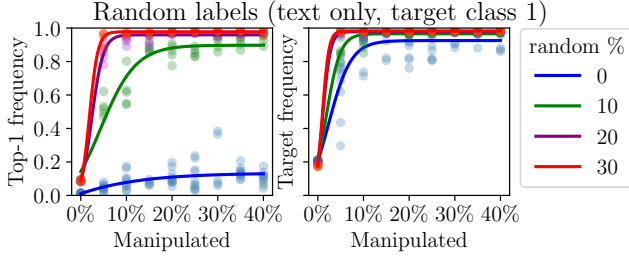
8

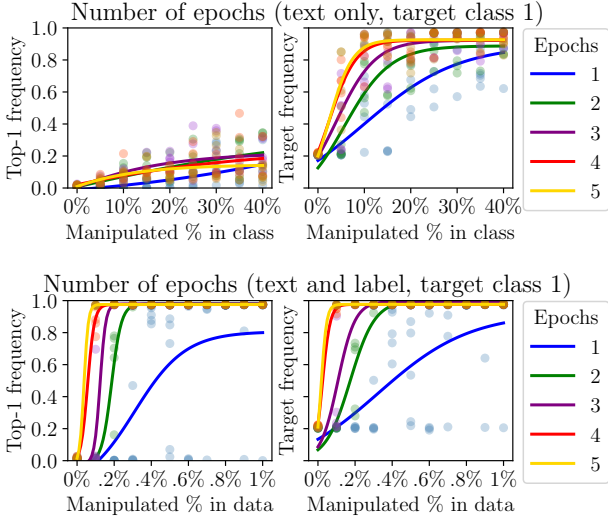*Figure 4.* Random labels increase success of Strategy 2.





*Figure 5.* Additional epochs of training increase the success rate.

**Robustness to trigger token placement.** Figure 7 in the Appendix shows that the success rate of either strategy is insensitive to the spacing of the trigger token. Varying the token frequency from 10 to 50 has little effect on the success rate. This experimental finding, too, is in line with our theory. Since the token chosen in our strategy is unique, the set of texts augmented with this unique token has low probability regardless of how often the token is planted.

# 6. Discussion

We conclude the paper with a short discussion highlighting the economic significance of understanding the critical mass $\alpha^*$ for pursuing collective targets. It is well-known in economics that participation in a collective is *not* individually rational, and additional incentives are necessary for collective action to emerge. Building on a classic model for collective action from economics (Olson, 1965), we illustrate how similar conclusions hold for algorithmic collective action, and how they relate to the theoretical quantities studied in this paper.

Assume that individuals incur a cost $c > 0$ for participating in collective action. This cost might represent overheads of
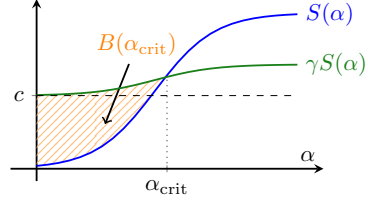


*Figure 6.* Visualization of the critical threshold $\alpha_{\text{crit}}$ after which a collective is self-sustaining and the principal's required investment $B(\alpha_{\text{crit}})$ to incentivize the whole population to join the collective.

coordination, a membership fee, or other additional responsibilities. Furthermore, assume that the utility that individuals get from joining a collective of size $\alpha$ is $S(\alpha)$, and that otherwise they can partially "free ride" on the collective's efforts: they get utility of $\gamma S(\alpha)$ for some $\gamma \in [0, 1]$. Given this setup, individually rational agents will join the collective if

$$S(\alpha) - c > \gamma S(\alpha),$$

or equivalently, if $S(\alpha) > \frac{c}{1-\gamma}$. Therefore, joining the collective is rational if the size of the existing collective $\alpha$ is greater than the critical mass for $S^* = \frac{c}{1-\gamma}$. Note that, once this critical threshold is reached, all individuals in the population are incentivized to join the collective and the collective is thus self-sustaining.

Consider a principal who would like to invest into the formation of a collective. The area $B(\alpha_{\text{crit}})$ visualized in Figure 6 provides an upper bound on the investment required to make the collective self-sustaining and thus achieve any target success $S^* \leq S(1)$.

The derivation above, while simplistic, serves to highlight the importance of collective size in understanding how collectives can emerge both organically and through investment. We believe that there is a large potential in investigating these questions in a rigorous manner. Indeed, the focus of this paper has been on understanding the effect of the size of the collective on its success, but understanding more generally how collectives form, which individuals have the most incentive to join collectives, whether selectively recruiting individuals provides additional leverage, and how collectives should use their informational advantage to optimize their strategies are important open questions in understanding the role of collectives on digital platforms.

# Acknowledgements

# References

Abebe, R., Hardt, M., Jin, A., Miller, J., Schmidt, L., and Wexler, R. Adversarial scrutiny of evidentiary statistical software. In *Conference on Fairness, Accountability, and Transparency*, pp. 1733–1746, 2022.

Albert, K., Penney, J., Schneier, B., and Kumar, R. S. S. Politics of adversarial machine learning. *arXiv preprint arXiv:2002.05648*, 2020.

Albert, K., Delano, M., Kulynych, B., and Kumar, R. S. S. Adversarial for good? How the adversarial ML community's values impede socially beneficial uses of attacks. *arXiv preprint arXiv:2107.10302*, 2021.

Burrell, J., Kahn, Z., Jonas, A., and Griffin, D. When users control the algorithms: values expressed in practices on Twitter. *Proc. ACM Hum.-Comput. Interact.*, 3:1–20, 2019.

Cameron, L. *The rise of algorithmic work: Implications for organizational control and worker autonomy*. PhD thesis, University of Michigan, 2020.

Cameron, L. D. and Rahman, H. Expanding the locus of resistance: Understanding the co-constitution of control and resistance in the gig economy. *Organization Science*, 33(1):38–58, 2022.

Chen, J. Y. Thrown under the bus and outrunning it! The logic of Didi and taxi drivers' labour and activism in the on-demand economy. *New Media & Society*, 20(8): 2691–2711, 2018.

Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J. P., Taylor, G., and Goldstein, T. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2021.

Cinà, A. E., Grosse, K., Vascon, S., Demontis, A., Biggio, B., Roli, F., and Pelillo, M. Backdoor learning curves: Explaining backdoor poisoning beyond influence functions. *arXiv preprint arXiv:2106.07214*, 2021.

Creager, E. and Zemel, R. Online algorithmic recourse by collective action. *ICML Workshop on Algorithmic Recourse*, 2021.

Cropanzano, R., Keplinger, K., Lambert, B. K., Caza, B., and Ashford, S. J. The organizational psychology of gig work: An integrative conceptual review. *Journal of Applied Psychology*, 2022.

Delobelle, P., Temple, P., Perrouin, G., Frénay, B., Heymans, P., and Berendt, B. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *SIGKDD Explor. Newsl.*, 23(1):32–41, 2021.

Efron, B. *Exponential Families in Theory and Practice*. Cambridge University Press, 2022.

Farhadkhani, S., Guerraoui, R., Hoang, L. N., and Villemaud, O. An equivalence between data poisoning and Byzantine gradient attacks. In *International Conference on Machine Learning*, 2022.

Geiping, J., Fowl, L. H., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021.

Gray, M. L. and Suri, S. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

Grosse, K., Lee, T., Biggio, B., Park, Y., Backes, M., and Molloy, I. Backdoor smoothing: Demystifying backdoor attacks on deep neural networks. *Computers & Security*, 120:102814, 2022.

Guo, W., Tondi, B., and Barni, M. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 3: 261–287, 2022.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Irani, L. C. and Silberman, M. S. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Association for Computing Machinery*, pp. 611–620, 2013.

Jarrahi, M. H. and Sutherland, W. Algorithmic management and algorithmic competencies: Understanding and appropriating algorithms in gig work. In *International conference on information*, pp. 578–589. Springer, 2019.

Jiechieu, K. F. F. and Tsopze, N. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 33:5069–5087, 2021.

Kulynych, B., Overdorf, R., Troncoso, C., and Gürses, S. POTs: Protective optimization technologies. In *Conference on Fairness, Accountability, and Transparency*, 2020.

Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S.-T., and Li, B. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. In *Advances in Neural Information Processing Systems*, 2022.

Manoj, N. S. and Blum, A. Excess capacity and backdoor poisoning. In Beygelzimer, A., Dauphin, Y., Liang, P.,

and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Melucci, A. *Challenging Codes: Collective Action in the Information Age*. Cambridge University Press, 1996.

Milan, S. When algorithms shape collective action: Social media and the dynamics of cloud protesting. *Social Media + Society*, 1(2), 2015.

Olson, M. *The logic of collective action: public goods and the theory of groups*. Number 124 in Harvard economic studies. Harvard Univ. Press, 1965.

O'Meara, V. Weapons of the chic: Instagram influencer engagement pods as practices of resistance to Instagram platform labor. *Social Media+ Society*, 5(4): 2056305119879671, 2019.

Rahman, H. A. The invisible cage: Workers' reactivity to opaque algorithmic evaluations. *Administrative Science Quarterly*, 66(4):945–988, 2021.

Robinson, H. C. *Making a digital working class: Uber drivers in Boston, 2016-2017*. PhD thesis, Massachusetts Institute of Technology, 2017.

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., Milland, K., and Clickhappier. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Association for Computing Machinery*, 2015.

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

Schor, J. *After the gig: How the sharing economy got hijacked and how to win it back*. University of California Press, 2021.

Schor, J. B., Attwood-Charles, W., Cansoy, M., Ladegaard, I., and Wengronowitz, R. Dependence and precarity in the platform economy. *Theory and Society*, 49(5):833–861, 2020.

Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., and Zhao, B. Y. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Security Symposium*, 2020.

Sinai, M. B., Partush, N., Yadid, S., and Yahav, E. Exploiting social navigation. *arXiv preprint arXiv:1410.0151*, 2014.

Sun, P. Your order, their labor: An exploration of algorithms and laboring on food delivery platforms in China. *Chinese Journal of Communication*, 12(3):308–323, 2019.

Sutherland, W. and Jarrahi, M. H. The sharing economy and digital platforms: A review and research agenda. *International Journal of Information Management*, 43: 328–341, 2018.

Suya, F., Mahloujifar, S., Suri, A., Evans, D., and Tian, Y. Model-targeted poisoning attacks with provable convergence. In *International Conference on Machine Learning*, 2021.

Tian, Z., Cui, L., Liang, J., and Yu, S. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8), 2022.

Vallas, S. and Schor, J. B. What do platforms do? Understanding the gig economy. *Annual Review of Sociology*, 46(1):273–294, 2020.

Van Doorn, N. Platform labor: on the gendered and racialized exploitation of low-income service work in the 'on-demand' economy. *Information, Communication & Society*, 20(6):898–914, 2017.

Vincent, N. and Hecht, B. Can "conscious data contribution" help users to exert "data leverage" against technology companies? *Proc. ACM Hum.-Comput. Interact.*, 2021.

Vincent, N., Hecht, B., and Sen, S. "Data strikes": evaluating the effectiveness of a new form of collective action against technology companies. In *The World Wide Web Conference*, 2019.

Vincent, N., Li, H., Tilly, N., Chancellor, S., and Hecht, B. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Conference on Fairness, Accountability, and Transparency*, 2021.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

Wood, A. J., Graham, M., Lehdonvirta, V., and Hjorth, I. Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, Employment and Society*, 33(1):56–75, 2019.

Woodcock, J. and Graham, M. The gig economy. *A critical introduction. Cambridge: Polity*, 2019.

Yu, Z., Treré, E., and Bonini, T. The emergence of algorithmic solidarity: unveiling mutual aid practices and resistance among chinese delivery workers. *Media International Australia*, 183(1):107–123, 2022.

Zrnic, T., Mazumdar, E., Sastry, S., and Jordan, M. Who leads and who follows in strategic classification? In *Advances in Neural Information Processing Systems*, 2021.

## A. Related Work

The scholarly literature on the gig economy is vast and interdisciplinary, spanning economic, ethnographic, psychological, and sociological analysis. Gig labor is diverse and heterogeneous. Conditions of precarity and dependence differ widely depending on the type of work and platform (Schor et al., 2020). Reviewing and integrating more than two hundred articles on the topic, Cropanzano et al. (2022) define gig work as "labor contracted and compensated on a short-term basis to organizations or to individual clients through an external labor market," and detail how gig work has changed the psychological contract between workers and employers. Vallas & Schor (2020) review existing scholarly accounts of the gig economy, arguing that "platforms represent a distinctive form of economic activity, [...] different from markets, hierarchies, and networks." Platforms cede some forms of centralized managerial control over workers by exposing them to the disciplining effect of the market and evaluation by consumers, while retaining power over key functions, such as task allocation, data collection, pricing, and collection of revenues. In another review article, Sutherland & Jarrahi (2018) organize more than four hundred articles around the notion of platform centralization and decentralization.

Several works examine the reality of gig labor, e.g., (Van Doorn, 2017; Sun, 2019; Gray & Suri, 2019). Based on a cross-regional survey, Wood et al. (2019) find that algorithmic control in the gig economy can lead to "low pay, social isolation, working unsocial and irregular hours, overwork, sleep deprivation and exhaustion", "marked by high levels of inter-worker competition with few labour protections and a global oversupply of labour relative to demand".

Cameron & Rahman (2022) examine the interplay of control and resistance in the gig economy. There are several examples of successful worker organization in the gig economy, involving a range of strategies. For example, Rahman (2021) studies how freelancers on Upwork strategize against the evaluation metrics of the platform, sometimes in cooperation with clients on the platform. Also studying Upwork freelancers, Jarrahi & Sutherland (2019) discuss how freelancers cooperate in strategically feeding the algorithm data so as to improve their outcomes. Cooperative strategic behavior among drivers on ride-hailing platforms is common, e.g., (Cameron, 2020; Robinson, 2017; Yu et al., 2022), as are digital strategies involving bots, or multiple phone apps (Chen, 2018). Workers have also used forums, browser extensions, and online spaces to share information and strategize collectively, e.g., (Irani & Silberman, 2013; Salehi et al., 2015; O'Meara, 2019). We focus on the activities on the labor side of digital platforms, leaving out numerous examples of collective action from consumers and users on these platforms. However, as Vallas & Schor (2020) conclude, "the upsurge of worker mobilization should not blind us to the difficulties of organizing such a diverse and spatially dispersed labour force or the power of the companies to resist collective action."

There is extensive scholarship on the topic of collective action. For example, Melucci's text (Melucci, 1996) examines collective action in the information society. Milan (2015) examines how social media platforms mediate social movements and collective action.

## B. Proofs

The following lemma will be used to analyze suboptimal classifiers.

**Lemma B.1.** *Suppose that $\mathcal{P}, \mathcal{P}'$ are two distributions such that $\mathrm{TV}(\mathcal{P}, \mathcal{P}') \leq \epsilon$. Take any two events $E_1, E_2$ measurable under $\mathcal{P}, \mathcal{P}'$. If $\mathcal{P}(E_1) > \mathcal{P}(E_2) + \frac{\epsilon}{1-\epsilon}$, then $\mathcal{P}'(E_1) > \mathcal{P}'(E_2)$.*

*Proof.* It follows from the optimal coupling lemma for the total variation distance that we can write $\mathcal{P}' = (1 - \epsilon)\mathcal{P} + \epsilon\mathcal{Q}$ for some distribution $\mathcal{Q}$. Therefore, if $\mathcal{P}(E_1) > \mathcal{P}(E_2) + \frac{\epsilon}{1-\epsilon}$, then

$$\mathcal{P}'(E_1) = (1 - \epsilon)\mathcal{P}(E_1) + \epsilon\mathcal{Q}(E_1) > (1 - \epsilon)\mathcal{P}(E_2) + \epsilon \geq (1 - \epsilon)\mathcal{P}(E_2) + \epsilon\mathcal{Q}(E_2) = \mathcal{P}'(E_2).$$

□

### B.1. Proof of Theorem 3.3

First consider the case $\epsilon = 0$. We start with a sufficient condition for a target classification outcome. For a point $x \in \mathcal{X}$, we define

$$\Delta_x = \max_{y \in \mathcal{Y}} \mathcal{P}_0(y|x) - \mathcal{P}_0(y^*|x)$$

as the suboptimality of a target class on the base data.

*Claim* B.2. For any $x \in \mathcal{X}$, we have $f(x) = y^*$ provided that $\alpha > (1 - \alpha)\Delta_x \mathcal{P}_0(x)/\mathcal{P}^*(x)$.

*Proof.* Note that $f(x) = y^*$ if, for every $y \neq y^*$, $\mathcal{P}(y^*|x) > \mathcal{P}(y|x)$. Equivalently, $\mathcal{P}(x, y^*) - \mathcal{P}(x, y) > 0$. But,

$$\mathcal{P}(x, y^*) = \alpha \mathcal{P}^*(x, y^*) + (1 - \alpha)\mathcal{P}_0(x, y^*) = \alpha \mathcal{P}^*(x) + (1 - \alpha)\mathcal{P}_0(y^*|x)\mathcal{P}_0(x)$$

In the last step we used the fact that all labels in the support of $\mathcal{P}^*$ equal $y^*$. Similarly, for $y \neq y^*$,

$$\mathcal{P}(x, y) = \alpha \mathcal{P}^*(x, y) + (1 - \alpha)\mathcal{P}_0(x, y) = (1 - \alpha)\mathcal{P}_0(y|x)\mathcal{P}_0(x).$$

The claim follows by rearranging terms and dividing both sides by $\mathcal{P}^*(x)$. $\qquad\square$

Now,

$$
\begin{aligned}
S(\alpha) &= \Pr_{x \sim \mathcal{P}_0} \{f(g(x)) = y^*\} \\
&= \Pr_{x \sim \mathcal{P}^*} \{f(x) = y^*\} \\
&\geq \Pr_{x \sim \mathcal{P}^*} \left\{ \alpha > (1 - \alpha)\frac{\mathcal{P}_0(x)}{\mathcal{P}^*(x)}\Delta_x \right\} \qquad\qquad \text{(Claim B.2)} \\
&= \mathbb{E}_{x \sim \mathcal{P}^*} \mathbf{1}\left\{ 1 - \frac{(1 - \alpha)}{\alpha}\frac{\mathcal{P}_0(x)}{\mathcal{P}^*(x)}\Delta_x > 0 \right\} \\
&\geq \mathbb{E}_{x \sim \mathcal{P}^*} \left[ 1 - \frac{(1 - \alpha)}{\alpha}\frac{\mathcal{P}_0(x)}{\mathcal{P}^*(x)}\Delta_x \right] \\
&= 1 - \frac{1 - \alpha}{\alpha} \mathbb{E}_{x \sim \mathcal{P}^*}\left[ \frac{\mathcal{P}_0(x)}{\mathcal{P}^*(x)}\Delta_x \right] \\
&\geq 1 - \frac{1 - \alpha}{\alpha}\mathcal{P}_0(\mathcal{X}^*)\Delta,
\end{aligned}
$$

where the last step uses the definition $\Delta = \max_{x \in \mathcal{X}^*} \Delta_x$.

**Consider** $\epsilon > 0$. By Lemma B.1, we have that $\mathcal{P}'(x, y^*) > \mathcal{P}'(x, y)$, meaning $f(x) = y^*$, provided that $\mathcal{P}(x, y^*) > \mathcal{P}(x, y) + \frac{\epsilon}{1 - \epsilon}$. Repeating the steps in the proof for $\epsilon = 0$ with the additional $\epsilon/(1 - \epsilon)$ term, we conclude that

$$S(\alpha) \geq 1 - \frac{\epsilon}{1 - \epsilon} - \frac{1 - \alpha}{\alpha}\mathcal{P}_0(\mathcal{X}^*)\Delta.$$

## B.2. Proof of Theorem 3.5

We prove the case where $\epsilon = 0$. The extension to $\epsilon > 0$ follows as in Theorem 3.3.

*Claim* B.3. Fix a point $x^* \in \mathcal{X}^*$ in the signal set. We have $f(x^*) = y^*$ provided that

$$\alpha > \frac{1 - p}{p}\frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g^{-1}(x^*))}.$$

Here, $g^{-1}(x^*) = \{x \in \mathcal{X} : g(x) = x^*\}$.

*Proof.* For $f(x^*) = y^*$ to hold, we need $\mathcal{P}(y^*|x^*) > \max_{y \neq y^*} \mathcal{P}(y|x^*)$. Equivalently, $\mathcal{P}(x^*, y^*) > \max_{y \neq y^*} \mathcal{P}(x^*, y)$.

By the definition of the feature-only signal strategy and the assumption that $\mathcal{P}_0(y^*|x) \geq p$ for all $x \in \mathcal{X}$, each point $x \in g^{-1}(x^*)$ must have $\mathcal{P}_0(y^*|x) \geq p$. Hence, for all $x^* \in \mathcal{X}^*$,

$$\mathcal{P}(x^*, y^*) = \alpha \mathcal{P}^*(x^*, y^*) + (1 - \alpha)\mathcal{P}_0(x^*, y^*) \geq \alpha p \mathcal{P}_0(g^{-1}(x^*)).$$

On the other hand, for every $y \neq y^*$, we must have

$$\mathcal{P}(x^*, y) = \mathcal{P}_0(x^*, y) = \mathcal{P}_0(y|x^*)\mathcal{P}_0(x^*) \leq (1 - p)\mathcal{P}_0(x^*).$$

The claim follows by rearranging. $\qquad\square$

We can lower bound the success rate as

$$
\begin{aligned}
S(\alpha) &= \Pr_{x \sim \mathcal{P}_0} \{f(g(x)) = y^*\} \\
&= \sum_{x^* \in \mathcal{X}^*} \Pr_{x \sim \mathcal{P}_0} \{f(g(x)) = y^* \mid x \in g^{-1}(x^*)\} \Pr_{x \sim \mathcal{P}_0} \{x \in g^{-1}(x^*)\} \\
&= \sum_{x^* \in \mathcal{X}^*} \mathbf{1}\{f(x^*) = y^*\} \mathcal{P}_0(g^{-1}(x^*)) .
\end{aligned}
\tag{5}
$$

Proceeding for fixed $x^* \in \mathcal{X}^*$,

$$
\begin{aligned}
\mathbf{1}\{f(x^*) = y^*\} &\geq \mathbf{1}\left\{\alpha > \frac{1-p}{p} \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g^{-1}(x^*))}\right\} && \text{(Claim B.3)} \\
&= \mathbf{1}\left\{1 - \frac{1-p}{p\alpha} \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g^{-1}(x^*))} > 0\right\} \\
&\geq 1 - \frac{1-p}{p\alpha} \cdot \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g^{-1}(x^*))} .
\end{aligned}
$$

Plugging this back into (5),

$$
\begin{aligned}
\Pr_{x \sim \mathcal{P}_0} \{f(g(x)) = y^*\} &= 1 - \frac{1-p}{p\alpha} \sum_{x^* \in \mathcal{X}^*} \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g^{-1}(x^*))} \cdot \mathcal{P}_0(g^{-1}(x^*)) \\
&\geq 1 - \frac{1-p}{p\alpha} \mathcal{P}_0(\mathcal{X}^*) .
\end{aligned}
$$

### B.3. Proof of Theorem 3.7

We again prove the case where $\epsilon = 0$. The extension to $\epsilon > 0$ follows by invoking Lemma B.1, as in Theorem 3.3.

We start from the following claim.

*Claim* B.4. For any $x \in \mathcal{X}$ we have $f(x) = f(g(x))$ provided that

$$
\alpha > (1 - \alpha) 2\tau(x),
$$

where $\tau(x) = \max_{y \in \mathcal{Y}} |\mathcal{P}_0(y|x) - \mathcal{P}_0(y|g(x))|$.

*Proof.* Denote $y^*(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \mathcal{P}_0(y|g(x))$. By construction of the strategy we know that $f(g(x)) = y^*(x)$ and it remains to prove that $f(x) = y^*(x)$ under the condition of the claim.

We have $f(x) = y^*(x)$ if $\mathcal{P}(y^*(x)|x) > \mathcal{P}(y|x)$ for any $y \neq y^*(x)$. We have

$$
\begin{aligned}
\mathcal{P}(y^*(x)|x) &= (1-\alpha)\mathcal{P}_0(y^*(x)|x) + \alpha \mathcal{P}^*(y^*(x)|x) = (1-\alpha)\mathcal{P}_0(y^*(x)|x) + \alpha, \\
\mathcal{P}(y|x) &= (1-\alpha)\mathcal{P}_0(y|x) + \alpha \mathcal{P}^*(y|x) = (1-\alpha)\mathcal{P}_0(y|x),
\end{aligned}
$$

where we used that the erasure strategy implies $\mathcal{P}^*(y^*(x)|x) = 1$. Together this means that, when

$$
\alpha > (1-\alpha) \left[\max_{y \in \mathcal{Y}} \mathcal{P}_0(y|x) - \mathcal{P}_0(y^*(x)|x)\right],
$$

then $f(x) = y^*(x)$. Using the definition of $y^*(x)$, we can bound the right-hand side by

$$
\begin{aligned}
\mathcal{P}_0(y|x) - \mathcal{P}_0(y^*(x)|x) &\leq \mathcal{P}_0(y|x) - \mathcal{P}_0(y|g(x)) + \mathcal{P}_0(y^*(x)|g(x)) - \mathcal{P}_0(y^*(x)|x) \\
&\leq 2\tau(x).
\end{aligned}
$$

The claim follows. $\qquad\square$

It remains to bound the success of the strategy:

$$\begin{aligned}
S(\alpha) &= \Pr_{x \sim \mathcal{P}_0} \{f(x) = f(g(x))\}. \\
&= \Pr_{x \sim \mathcal{P}_0} \{f(x) = y^*(x)\}. \\
&\geq \Pr_{x \sim \mathcal{P}_0} \{\alpha > (1 - \alpha)2\tau(x)\} \\
&= \Pr_{x \sim \mathcal{P}_0} \left\{1 - \frac{1 - \alpha}{\alpha} 2\tau(x) > 0\right\} \\
&\geq \mathbb{E}_{x \sim \mathcal{P}_0} \left[1 - \frac{2(1 - \alpha)}{\alpha} \cdot \tau(x)\right] \\
&= 1 - \frac{2(1 - \alpha)}{\alpha} \cdot \tau,
\end{aligned}$$

where we use the fact that $\tau = \mathbb{E}_{x \sim \mathcal{P}_0} \tau(x)$.

### B.4. Proof of Theorem 4.2

Let $\mathcal{P}'$ be a gradient-cancelling distribution for $\theta^*$. Denote $p = \min\left(1, \frac{1}{\alpha} \frac{\|g_{\mathcal{P}_0}(\theta^*)\|}{\|g_{\mathcal{P}'}(\theta^*)\| + \|g_{\mathcal{P}_0}(\theta^*)\|}\right)$. Then,

$$\begin{aligned}
\mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta^*; z) &= (1 - \alpha) \mathbb{E}_{z \sim \mathcal{P}_0} \nabla \ell(\theta^*; z) + \alpha \mathbb{E}_{z \sim \mathcal{P}^*} \nabla \ell(\theta^*; z) \\
&= (1 - \alpha p) \mathbb{E}_{z \sim \mathcal{P}_0} \nabla \ell(\theta^*; z) + \alpha p \mathbb{E}_{z \sim \mathcal{P}'} \nabla \ell(\theta^*; z) \\
&= (1 - \alpha p) g_{\mathcal{P}_0}(\theta^*) + \alpha p\, g_{\mathcal{P}'}(\theta^*) \\
&= \left(1 - \alpha p - \alpha p \frac{\|g_{\mathcal{P}'}(\theta^*)\|}{\|g_{\mathcal{P}_0}(\theta^*)\|}\right) g_{\mathcal{P}_0}(\theta^*) \\
&= \left(1 - \alpha p \frac{\|g_{\mathcal{P}_0}(\theta^*)\| + \|g_{\mathcal{P}'}(\theta^*)\|}{\|g_{\mathcal{P}_0}(\theta^*)\|}\right) g_{\mathcal{P}_0}(\theta^*) \\
&= \max\left(1 - \alpha \frac{\|g_{\mathcal{P}_0}(\theta^*)\| + \|g_{\mathcal{P}'}(\theta^*)\|}{\|g_{\mathcal{P}_0}(\theta^*)\|}, 0\right) g_{\mathcal{P}_0}(\theta^*) \\
&= \max\left((1 - \alpha)\|g_{\mathcal{P}_0}(\theta^*)\| - \alpha\|g_{\mathcal{P}'}(\theta^*)\|, 0\right) \frac{g_{\mathcal{P}_0}(\theta^*)}{\|g_{\mathcal{P}_0}(\theta^*)\|}.
\end{aligned}$$

Therefore, $\|\mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta^*; z)\| = \max\left((1 - \alpha)\|g_{\mathcal{P}_0}(\theta^*)\| - \alpha\|g_{\mathcal{P}'}(\theta^*)\|, 0\right)$. Applying the definition of $\mu$-strong convexity, we get

$$\begin{aligned}
\|\theta^* - \theta\| &\leq \frac{1}{\mu} \|\mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta^*; z) - \mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta; z)\| \\
&= \frac{1}{\mu} \|\mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta^*; z)\| \\
&= \frac{1}{\mu} \max\left((1 - \alpha)\|g_{\mathcal{P}_0}(\theta^*)\| - \alpha\|g_{\mathcal{P}'}(\theta^*)\|, 0\right).
\end{aligned}$$

The first equality follows because $\mathbb{E}_{z \sim \mathcal{P}} \nabla \ell(\theta; z) = 0$ due to the loss being convex and the firm being a risk minimizer. Multiplying both sides by $-1$, we obtain a lower bound on the success $S(\alpha) = -\|\theta^* - \theta\|$.

### B.5. Proof of Corollary 4.3

To achieve $S(\alpha) = 0$, Theorem 4.2 shows that it suffices to have $\alpha\|g_{\mathcal{P}'}(\theta^*)\| = (1 - \alpha)\|g_{\mathcal{P}_0}(\theta^*)\|$, for any $\mu$. Rearranging the terms and expressing $\alpha$ completes the proof.

### B.6. Proof of Proposition 4.4

If $u$ is convex, then for all $\theta'$ we know

$$u(\theta') \geq u(\theta_0) + \nabla u(\theta_0)^\top (\theta' - \theta_0).$$

Let $\theta^* = \theta_0 + \frac{\nabla u(\theta_0)}{\|\nabla u(\theta_0)\|^2} U$. Then, $u(\theta^*) - u(\theta_0) \geq U$.

Now, we apply Corollary 4.3 to upper bound the critical mass needed to reach $\theta^*$. We have

$$\alpha^* \leq \frac{\|g_{\mathcal{P}_0}(\theta^*)\|}{\|g_{\mathcal{P}'}(\theta^*)\| + \|g_{\mathcal{P}_0}(\theta^*)\|} \leq \frac{\beta\|\theta^* - \theta_0\|}{g_{\mathrm{lb}} + \beta\|\theta^* - \theta_0\|},$$

where we apply smoothness and the definition of $g_{\mathrm{lb}}$. Observing that $\|\theta^* - \theta_0\| = \frac{U}{\|\nabla u(\theta_0)\|}$ completes the proof.

### B.7. Proof of Theorem 4.5

Fix a time step $t$ and a model $\theta_t$. Denote by $\mathcal{P}'_t$ the gradient-redirecting distribution found at step $t$ and let $\xi(\theta_t) = \frac{\|g_{\mathcal{P}'_t}(\theta_t) + \frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}(\theta_t)\|}{\|\theta_t - \theta^*\|}$. Then, the gradient-redirecting strategy induces the following gradient evaluated on $\mathcal{P}_t$:

$$
\begin{aligned}
g_{\mathcal{P}_t}(\theta_t) &= \alpha g_{\mathcal{P}'_t}(\theta_t) + (1-\alpha)g_{\mathcal{P}_0}(\theta_t) \\
&= -\alpha\frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}(\theta_t) + \alpha\xi(\theta_t)(\theta_t - \theta^*) + (1-\alpha)g_{\mathcal{P}_0}(\theta_t) \\
&= \alpha\xi(\theta_t)(\theta_t - \theta^*).
\end{aligned}
$$

Now let $c = \min_{\lambda\in[0,1]} \xi(\lambda\theta_0 + (1-\lambda)\theta^*)$. Applying the strategy repeatedly across time steps yields

$$
\begin{aligned}
\|\theta_T - \theta^*\| &\leq \|\theta_{T-1} - \eta\alpha\xi(\theta_{T-1})(\theta_{T-1} - \theta^*) - \theta^*\| \\
&\leq (1 - \eta\alpha\xi(\theta_{T-1}))\|\theta_{T-1} - \theta^*\| \\
&\leq (1 - \eta\alpha c)\|\theta_{T-1} - \theta^*\| \\
&\leq (1 - \eta\alpha c)^T\|\theta_0 - \theta^*\|,
\end{aligned}
$$

which yields $S_T(\alpha) = -\|\theta_T - \theta^*\| \geq -(1 - \eta\alpha c)^T\|\theta_0 - \theta^*\|$. Setting $C(\alpha) = \alpha c$ concludes the proof.

## C. Additional Experimental Details

The dataset introduced by Jiechieu & Tsopze (2021) is available at `https://github.com/florex/resume_corpus`. We preprocessed each resume by removing the first 20 words of the resume. The reason is that the opening of the resume essentially encodes the associated skills, since the dataset creation process extracted the skills from the opening of the resume. Removing the first 20 words leads to a more realistic classification task.

We used the HuggingFace `transformers` open-source Python library (Wolf et al., 2020). We used the `distilbert-base-uncased` model from the library corresponding to the DistilBERT transformer model (Sanh et al., 2019). We used the HuggingFace `Trainer` module for training with its default settings. We also experimented with larger models, including RoBERTa (both `roberta-base` and `roberta-large`), but we did not see any improvements in classification accuracy from using these larger models.

The DistilBERT tokenizer has a vocabulary of 30522 tokens of which thousands are unused in the resume corpus. We picked token 1240 corresponding to a small dash, which we inserted in the resume every 20 words. Our findings are largely insensitive to the trigger spacing as Figure 7 shows.
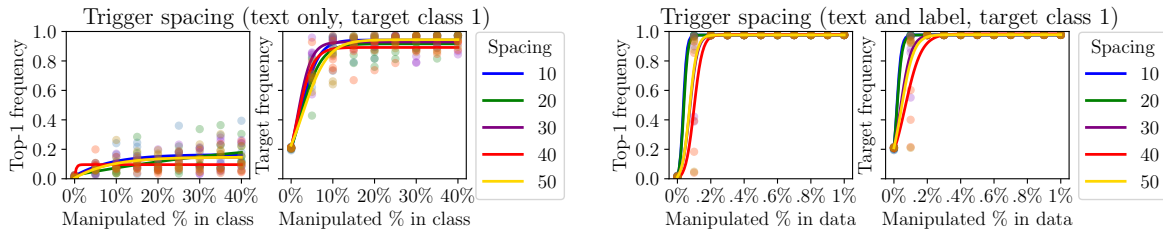


*Figure 7.* Trigger spacing is largely irrelevant.