
Sync or Sink: Bounds on Algorithmic Collective Action with Noise and Multiple Groups

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Collective action against algorithmic systems, which enables groups to promote
2 their own interests, is poised to grow. Hence, there will be growth in the size and
3 the number of distinct collectives. Currently, there is no formal analysis of how
4 coordination challenges within a collective can impact downstream outcomes, or
5 how multiple collectives may affect each other’s success. In this work, we aim to
6 provide guarantees on the success of collective action in the presence of both co-
7 ordination noise and multiple groups. Our insight is that data generated by either
8 multiple collectives or by coordination noise can be viewed as originating from
9 multiple data distributions. Using this framing, we derive bounds on the success
10 of collective action. We conduct experiments to study the effects of noise on col-
11 lective action. We find that sufficiently high levels of noise can reduce the success
12 of collective action. In certain scenarios, large noise can sink a collective suc-
13 cess rate from 100% to just under 60%. We identify potential trade-offs between
14 collective size and coordination noise; for example, a collective that is twice as
15 big but with four times more noise experiencing worse outcomes than the smaller,
16 more coordinated one. This work highlights the importance of understanding nu-
17 anced dynamics of strategic behavior in algorithmic systems.

18 1 Introduction

19 As large platforms (social media, e-commerce) grow, groups within them find it increasingly im-
20 portant and necessary to act strategically (e.g., by changing ratings) in these platforms to get their
21 desired outcome. From delivery drivers coordinating to earn higher wages [SHMD24], to fans pro-
22 moting their favorite artist [XZF⁺25, Nug21], to protesting controversial businesses [Pay24], this
23 type of strategic behavior is poised to grow. Strategic behavior ultimately affects the downstream
24 data distributions used to train models. As the incentive to participate increases, we need a more
25 nuanced understanding of the impacts on a system; this includes both understanding what happens
26 with multiple distinct collectives acting on a system as well as considering how internal collective
27 dynamics may impact outcomes. These factors will all impact the downstream data distribution.
28 However, currently, there are no tools or analytical frameworks that incorporate these factors (coor-
29 dination noise, multiple collectives) to assess their impact on the success of collective action.

30 In classical collective action problems related to managing common-pool resources, Ostrom notes
31 that smaller, more homogeneous groups are more effective at overcoming barriers to collective ac-
32 tion [Ost90]. She finds that a clear organizational structure and well-defined roles are crucial—one
33 possible interpretation is that smaller groups can more readily and faithfully implement and coor-
34 dinate their actions. An analogy in the computational world is “noise”: unwanted modifications
35 or deviations from intended behavior (e.g., signal processing). This “noise” can result from poor
36 coordination and may reduce effectiveness.

In algorithmic settings, collective action is mediated through strategic data modifications to influence model behavior. These changes in data distribution can have significant downstream ramifications. [HMMDZ23] provided theoretical bounds on the success of collective action in the case of a single, unified collective. The collective modifies their data (e.g., by inserting a watermark), which is then used in training. The learning algorithm observes a distribution $\mathcal{P}$: a mixture of data induced by the collective’s action ($\mathcal{P}_1$) and the underlying data distribution $\mathcal{P}_0$. These bounds help identify the conditions under which collective action may be more successful. However, this analysis assumes perfect coordination within the collective. [KVKS25] developed a framework to identify the factors affecting collective action involving multiple collectives, using simulations to illustrate several outcomes. However, they do not provide any theoretical bounds on success based on the data distributions induced by each collective. A formal understanding of how both noise and multiple collectives interact is key to understanding real-world algorithmic collective action.

In this work, we provide new guarantees on algorithmic collective action in both the presence of noise and multiple distinct collectives. We do this by introducing *multiple distributions* that feed the ultimate observed data distribution $\mathcal{P}$. Instead of considering just a mixture of $\mathcal{P}_0$ and $\mathcal{P}_1$, we study how the addition of $\mathcal{P}_2$ affects a collective’s success. Importantly, the source of these distributions can vary: it could occur from multiple distinct collectives or because of coordination challenges within a single collective inducing a new, second data distribution.

A collective that is highly coordinated is one where the size of $\mathcal{P}_2$ is small and similar to $\mathcal{P}_1$, while ones with less cohesion have a larger gap between $\mathcal{P}_1$ and $\mathcal{P}_2$. Multiple collectives can be analyzed as behavior coming from two distinct distributions. Our contributions are as follows:

Theoretical bounds with multiple distributions: We are the first to establish the lower bounds for the success of collective action in the presence of multiple distributions. Prior work has either only analyzed a single distribution coming from a single collective or provided empirical outcomes with multiple collectives [KVKS25, HMMDZ23]. We are able to relate this bound to the similarity of the distributions and collective size. This approach can effectively handle many different scenarios: including when there’s a fraction of a collective performing actions imperfectly or when multiple distinct collectives are acting upon a system.

Empirical impact of noise: We study the impact of noise in collective action. Prior work [HMMDZ23] has only shown outcomes with perfectly coordinated collectives. We find that different types of noise can impact the success of collective action and find in some situations, smaller, less noisy groups perform better than larger groups with more noise.

We first present a formulation of the algorithmic collective action problem. We establish bounds on the success of collective action with multiple distributions. Our experiments extend [HMMDZ23] by considering a text classification task with noise. We find scenarios where noise sinks success a task from nearly 100% without noise to under 60%. We discuss possible trade-offs in size and noise where small groups with less noise (0.25% collective size with 10% noise) outperform larger ones with more noise (0.5% with 40% noise), which can inform organizers of collective action.

2 Related Work

Collective action against algorithms has been documented in contexts such as ridesharing [Lei21, WAC21, WVM21, JGV21, Has20] and in data campaigns aimed at promoting pro-social outcomes [MF22, VH21, VLT⁺21, WED22]. [SHMD24] provides a specific computational model for the delivery driver case, which can be used to analyze incentives and outcomes in these scenarios. [XZF⁺25, XFS⁺25] examine some of the real-world structures involved in promoting online collective action, such as what motivates certain users to act as organizers. They also discuss practical challenges in influencing online marketplaces, including coordinating across different types of devices and teaching users how to engage effectively with the platform to achieve a specific goal. The concept of noise has been used to simulate communication challenges in agent-based modeling [WSLA13, ASJCB02, BBG23] and in modeling bounded rationality [Kah03, Con96, Qui07]. However, it has not been studied in the context of online collective action.

Other work has examined collective action in algorithmic settings. [HMMDZ23] first examines how different types of strategies can enable small collectives to have a substantial impact on classification outcomes. [BMD24] extends this analysis to examples involving recommender sys-

tems. [BDFSS24] examines how the specifics of the learning algorithm impact collective action. [KVKS25] empirically analyzes the interaction between two collectives. None of these works, however, offer formal guarantees for successful collective action involving multiple data distributions. Our work offers a more comprehensive theoretical understanding of realistic strategic dynamics.

3 Problem Formulation

Here we first will describe collective action as described by [HMMDZ23]. We will then extend this scenario to include a second distribution. For convenience, Appendix A summarizes the notation.

3.1 Algorithmic Collective Action with One Distribution

We consider the case where a firm wishes to deploy a classifier f trained on some data. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$. We define a classifier to be ϵ_1 suboptimal under a distribution P_1 if there exists a P' with $TV(P_1, P') \leq \epsilon_1$ such that $f = \arg \max_{y \in \mathcal{Y}} P'(y|x)$ where TV is the total variation distance.

The firm's goal is to minimize the loss with respect to some objective function. The data they train on exists in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ representing the features $\mathcal{X}$ and labels $\mathcal{Y}$. For our purposes, we can think of every data point $z \in \mathcal{Z}$ as belonging to an individual. A collective wishes to work together in order to create a specific outcome on this classifier for a certain set of inputs which we formalize below.

Consider the base distribution $\mathcal{P}_0$ on $\mathcal{Z}$. We define $\mathcal{P}_1$ to be the distribution induced by the collective's intended action. This is operationalized by strategy, $h_1 : \mathcal{Z} \rightarrow \mathcal{Z}$ where $h_1 \in \mathcal{H}$ the set of all available strategies. Examples of potential strategies may include watching a video for a certain amount of time, giving a specific review score, etc. In implementing this strategy, the collective wishes to create an association between the specific inputs to a target label y^* .

The collective plants a "signal", which is done via function $g_1 : \mathcal{X} \rightarrow \mathcal{X}$ which takes some original input $x \sim \mathcal{P}_0$ and modifies it. Examples may include adding certain words to text, adding a watermark to a video, or reviewing an extra product on a marketplace. This planted signal is intended to create the association between the set of inputs generated by $g_1(x)$ and a target label y^* .

In some situations, it may also be possible for the collective to act on both the input data ($\mathcal{X}$) and potentially the output label ($\mathcal{Y}$). For the *feature-label* strategy, members can each change their feature and label. For the *feature-only* strategy, they can only change their features. These strategies, using h_1 and signal g_1 , induce the collective's distribution $\mathcal{P}_1$. Let α_1 be the fraction of users participating in the collective. We can write the distribution seen by the learning algorithm as a mixture of the distribution of the collective's data and the unmodified data.

$$\mathcal{P} = \alpha_1 \mathcal{P}_1 + (1 - \alpha_1) \mathcal{P}_0$$

Where $\mathcal{P}_1$ represents the distribution of $h_1(z)$ where $z \sim \mathcal{P}_0$. The collective's objective is to associate the signal g_1 with the target y^* . The success rate can be written as

$$S(\alpha_1) = \Pr_{x \sim \mathcal{P}_0} [f(g_1(x)) = y^*]$$

Success also depends on how the signals generated by the collective conflict or compete with data present in the underlying distribution. Let $\mathcal{X}_1 = \{g_1(x) : x \in \mathcal{X}\}$ the signal set. We define this as suboptimality gap of $\mathcal{X}_1$ on distribution $\mathcal{P}_0$ for a target y^* as

$$\Delta_1^0(y^*) = \max_{x \in \mathcal{X}_1} (\max_{y \in \mathcal{Y}} \mathcal{P}_0(y|x) - \mathcal{P}_0(y^*|x))$$

We also use $\mathcal{P}_0(\mathcal{X}_1)$ to refer to the background overlap of signal set $\mathcal{X}_1$ - intuitively, this represents how common the modified input $\mathcal{X}_1$ is in the background distribution $\mathcal{P}_0$. Based on these definitions, [HMMDZ23] finds a lower bound on the success for this classifier in the single, perfectly coordinated collective for both the *feature-label* strategy and the *feature-only* strategy, restated here.

Theorem 1 (Feature-label). [HMMDZ23] *The success rate of a collective with size α_1 on a ϵ_1 classifier with target y_1^* with signal set $\mathcal{X}_1$ is bounded below by*

$$S(\alpha_1) \geq 1 - \frac{1 - \alpha_1}{\alpha_1} \mathcal{P}_0(\mathcal{X}_1) \frac{(1 - \epsilon_1) \Delta_1^0(y_1^*) + \epsilon_1}{1 - 2\epsilon_1}$$

Theorem 2 (Feature-only). [HMMDZ23] *The success rate of a collective with size α_1 on a ϵ_1 -suboptimal classifier with target y_1^* with distribution $\mathcal{X}_1$ where there exists a p such that $\mathcal{P}_0(y^*|x) \geq p, \forall x \in \mathcal{X}$ is bounded below by*

$$S(\alpha_1) \geq 1 - \frac{1 - (1 - \epsilon_1)p - \epsilon_1\alpha}{(1 - \epsilon_1)p\alpha - \epsilon_1\alpha} \mathcal{P}_0(\mathcal{X}_1)$$

119

120 However, these theorems do not consider a second distribution. We first develop our multiple distri-
121 bution setup, which we will use to expand beyond the above bounds.

122 3.2 Collective Action with Multiple Distributions

123 Here, we extend beyond [HMMDZ23] by considering another data distribution. This may arise from
124 internal coordination challenges or a separate collective.

125 We consider distribution $\mathcal{P}_2$ induced by a strategy h_2 . h_2 can be completely distinct from h_1 or can
126 represent a noisy version of h_1 (e.g $h_2(z) = h_1(z) + \delta$). $\mathcal{P}_2$ is the distribution of $h_2(z), z \sim \mathcal{P}_0$. For
127 example, suppose a collective aims to create an association between a word A existing at the top of
128 a document and being rated as an important user y^* . Members could mistakenly use a similar word
129 B instead of A or place A at the end of the text instead – this would be represented by $\mathcal{P}_2$.

We can write the final distribution observed by the learning model as

$$\mathcal{P} = \alpha_1 \mathcal{P}_1 + \alpha_2 \mathcal{P}_2 + (1 - \alpha) \mathcal{P}_0$$

130 where α_1 represents the fraction of people who implemented collective one’s strategy, α_2 represents
131 a separate group’s interactions. This can arise because the correct “action” was not properly shared
132 with all group members, people getting confused about what exactly to do and not executing faith-
133 fully, or potential infiltrators within the group. We let $\alpha = \alpha_1 + \alpha_2$ is the total fraction of people
134 attempting to implement some strategy. We can also equivalently consider $r = \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{\alpha_1}{\alpha}$ as the
135 fraction of people implementing the correct strategy.

136 Suppose a collective that wants to plant a single $g(x)$ where $x \in \mathcal{X}$. Let the signal set $\mathcal{X}_1 = \{g_1(x) :$
137 $x \in \mathcal{X}\}$. Another group implements another signal, which may be a noisy version of g_1 . Let g_2
138 represent the set of signals produced by the second group. $\mathcal{X}_2 = \{g_2(x) : x \in \mathcal{X}\}$.

The presence of another distribution requires us to expand upon our definition of suboptimality to
consider, across any two distributions, how often the signals from set $\mathcal{X}_i$ are present on the distribu-
tion $\mathcal{P}_j$. Our expanded definition of suboptimality is:

$$\Delta_i^j(y^*) = \max_{x \in \mathcal{X}_i^*} (\max_{y \in \mathcal{Y}} \mathcal{P}_j(y|x) - \mathcal{P}_j(y^*|x))$$

139 Intuitively, this measures how “confusing” signals from $\mathcal{X}_i$ looks to $\mathcal{P}_j$ when targeting y^* .

140 4 Multiple Distributions and Success of Collective Action

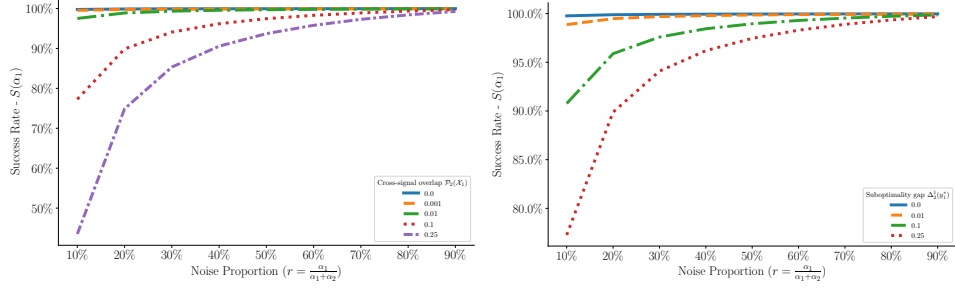
141 Based on our expanded definitions, we can quantify the impact that multiple distributions have on
142 the effectiveness of a collective. We consider both the feature-label strategy and the feature-only
143 strategy. Both are important to study, as in different contexts, collectives may be able to alter both
144 their input features and labels, or just their features. We state the theorems for two distributions here
145 and defer the proofs and generalization to n distributions to the Appendix E. While earlier work
146 focused on a single, perfectly coordinated collective, we quantify how either internal disruptions
147 or the presence of a second collective affects the original group’s success. Conceptually, the proofs
148 follow by considering the new mixture distribution and the interaction between the new distributions.

149 For the feature-label strategy, we have the following result (see Appendix C for full proof):

Theorem 3 (Feature-label with two distributions). *Consider distribution $\mathcal{P}_1$ and $\mathcal{P}_2$ which are dis-
tributed according to $h_1(x)$ and $h_2(x)$ respectively, where $x \sim \mathcal{P}_0$. Let y_1^* be the target class. Then
success for the first collective against an ϵ_1 classifier to be lower bounded by*

$$S(\alpha_1) \geq 1 - \frac{\alpha_2}{\alpha_1} \mathcal{P}_2(\mathcal{X}_1) \frac{(1 - \epsilon_1)\Delta_1^2(y_1^*) + \epsilon_1}{1 - 2\epsilon_1} - \frac{1 - \alpha}{\alpha_1} \mathcal{P}_0(\mathcal{X}_1) \frac{(1 - \epsilon_1)\Delta_1^0(y_1^*) + \epsilon_1}{1 - 2\epsilon_1}$$

150



(a) Varying $\mathcal{P}_2(\mathcal{X}_1)$ for fixed $\Delta_2^1(y_1^*) = 0.25$ (b) Varying $\Delta_2^1(y_1^*)$ for fixed $\mathcal{P}_2(\mathcal{X}_1) = 0.25$

Figure 1: Relationship between $\Delta_2^1(y_1^*)$, $\mathcal{P}_2(\mathcal{X}_1)$ and success bound. We fix $\alpha_1 + \alpha_2 = 0.3$ and vary the proportion r . We fix $\epsilon_1 = 0$, $\Delta_0^1 = 0.01$, and $\mathcal{P}_0(\mathcal{X}_1) = 0.01$

Compared with Theorem 1, Theorem 3 introduces another term that describes the relationship between $\mathcal{P}_1$ and $\mathcal{P}_2$ independently of the relationship between $\mathcal{P}_0$ and $\mathcal{P}_1$. It consists of the cross-signal overlap between signal set $\mathcal{X}_1$ on distribution $\mathcal{P}_2$ the suboptimality gap of the target y_1^* of signal set $\mathcal{X}_1$ and $\mathcal{P}_2$. ($\Delta_1^2(y_1^*)$) and the relative sizes of said groups $\frac{\alpha_2}{\alpha_1}$. Figure 1 illustrates these relationships.

If we consider the distributions as coming from two collectives with differing objectives, we can also write down the success bound for the second collective as well.

Corollary 3.1. *If we consider $\mathcal{P}_2, \mathcal{X}_2$ as coming from a 2nd collective with (potentially) distinct objective, we can write the probability of success as*

$$S(\alpha_2) \geq 1 - \frac{\alpha_1}{\alpha_2} \mathcal{P}_1(\mathcal{X}_2) \cdot \frac{(1 - \epsilon_2) \Delta_2^1(y_2^*) + \epsilon_2}{1 - 2\epsilon_2} - \frac{1 - \alpha}{\alpha_2} \mathcal{P}_0(\mathcal{X}_2) \frac{(1 - \epsilon_2) \Delta_2^0(y_2^*) + \epsilon_2}{1 - 2\epsilon_2}$$

For the feature-only strategy we have the following result (Appendix D for full proof)

Theorem 4 (Feature-only with two distributions). *Consider distribution $\mathcal{P}_1$ and $\mathcal{P}_2$ which are distributed by $h_1(x)$ and $h_2(x)$ respectively, where $x \sim \mathcal{P}_0$. Suppose there exist a p such that $\mathcal{P}_0(y^*|x) \geq p, \forall x \in \mathcal{X}$. Then success for the first collective against an ϵ_1 classifier (against $\mathcal{P}_1^*$) is lower bounded by*

$$S(\alpha_1) \geq 1 - \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(\mathcal{X}_1^*) \cdot \Delta_1^2(y^*)(1 - \epsilon_1)}{p(1 - \epsilon_1) - \epsilon_1} - \frac{1 - \alpha}{\alpha} \mathcal{P}_0(\mathcal{X}_1) \cdot \frac{(1 - p)(1 - \epsilon_1) + \epsilon_1}{p(1 - \epsilon_1) - \epsilon_1}$$

In this bound we see the same $\mathcal{P}_2(\mathcal{X}_1) \Delta_1^2(y_1^*) \alpha_2 (1 - \epsilon_1)$ term which captures the cross-signal overlap, suboptimality gap and collective size.

These theorems illustrate how and to what extent the presence of a second distribution can hinder the first collective. If we consider the second distribution to be a noisy variation of the first one, this helps relate the noise characteristics to the success rate. For a second collective, these bounds provide insights into which scenarios collectives may be simultaneously successful or hindering each other. We examine these implications in Section 7.1.

5 Experimental Setup

To demonstrate the empirical implications, we study the case of noisy collective action. We extend [HMDZ23] experiment on resume classification. We use the resume dataset introduced by [JT21] to finetune a BERT-based model for a multilabel prediction task. The goal for the classifier is to predict which set of careers someone is suited for based on the text in the resume. Members of the collective intend to plant a signal in their resume to get the resume classified to some target class y^* inserting a specific character in a certain pattern. The intended signal in this case is to place this specific character every 20 words (full details in Appendix B).

We vary r , the proportion of users who perform an imperfect collective action by performing a noisy variation. Specifically, we consider the following types of noise variations.

178 **Correct Character Usage** The collective attempts to place a specific character into the resume.
 179 A noisy variation involves using a different character. We consider variations where the wrong
 180 character is sampled across a small subset (Random-Subset), as well sampled randomly across all
 181 possible characters (Random-Full).

182 **Modification Location** The collective intended action to place their character every 20 words. A
 183 noisy variation places the character in arbitrary locations (Displaced).

184 These choices are motivated by considering “benign” ways a group of people may misinterpret
 185 instructions. If the intended instruction is “Place character ‘A’ every 20 words”, some users may
 186 focus more on the 20 words part and not use the right character, or some users may focus on using
 187 “A” and not place it every 20 words. We define the full set of variations in Table 1.

188 We also consider the context in which the noise occurs.

189 **Noised Input** For feature-label strategy, it is possible that the noise only gets applied in the input
 190 feature or gets applied to the feature as well as the label.

191 **Target’s Underlying Frequency** The target class underlying frequency (in $\mathcal{P}_0$) has shown to
 192 play some impact in algorithmic collective action, especially with the feature only strategy
 193 [HMMDZ23]. We test a high frequency label (Software Developer) as well as a low frequency
 194 label (Database Administrator) as the target class.

195 Based on these factors, we ask the following questions.

196 **RQ1:** Does the frequency of noise (larger r ’s) affect success more than noise variation?

197 **RQ2:** What class of strategies are more sensitive to noisy behaviors: feature-label or feature-only?

198 **RQ3:** How does the background frequency ($\mathcal{P}_0$) of the target class affect sensitivity to noise?

Variation Name	Variation Description
Baseline	All members use the same characters and place every 20 words
Random-Subset	Some members insert a different character than the signal character, chosen from a small, fixed subset, placed every 20 words.
Random -Full	Some members insert a different character than the signal character, chosen from a large, fixed subset, placed every 20 words.
Displaced-Original	Some members place the collective’s signal character at arbitrary places in the text
Displaced-Full	Some members place insert a randomly picked character from a large subset at arbitrary places in the text.

Table 1: List of variations for noise deviations

199 6 Results

200 We finetune `distilbert-based-uncased` [SDCW19] for five epochs with default hyperparam-
 201 eters using Hugging Face transformer library [WDS⁺20]. For the feature-label strategies, we vary α ,
 202 the percent of users in the collective from 0 – 1%. For the feature only strategies, we vary α ’s from
 203 [0%, 50%]. We consider noise rates r from [0%, 50%] (hence $\alpha_1 = r\alpha$). Noise variations for input
 204 text is detailed in Table 1. For noising labels, the label is changed uniformly at random across all
 205 possible labels. We measure success by calculating how often the target class (y_1^*) is predicted for
 206 text that has the desired modification x^* in the test set. All figures are shown with a one standard
 207 deviation region shaded.

208 6.1 RQ1: Different Noise Variations

209 Here we consider the several types of noise variations (as described in Table 1). As a reminder,
 210 these different noise variations change how the different members of the collective may incorrectly

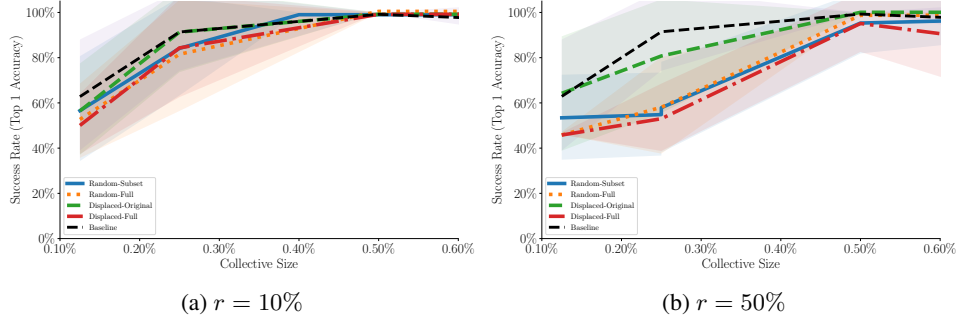


Figure 2: Comparison of different noise variations for different collective sizes. The x-axis here is the total collective size, measured by the percentage of the total population is participating. The y-axis is the success rate of causing resumes with the “true” signal to be classified to the target class. The black dashed line represents success rate without noise. The strategy is feature-label where noise is only applied to the features. We observe that making more characters for mistakes leads to lower success (Random-Subset vs Random-Full and Displaced-Original vs Displaced-Full)

211 implement a strategy. The x -axis represents the collective size while the y -axis show the collective’s
 212 success criteria. Figure 2 shows how different noise variations result in different efficacies, both for
 213 low levels of noise (Figure 2a) and higher levels (Figure 2b). We find that different levels of noise
 214 have more of an impact when the collective size is relatively small; for sufficiently high levels of
 215 participation, all variants perform similarly and comparable to the baseline even with noise. At this
 216 lower levels, there is a sensitivity to the set of “wrong” characters to choose from: the Random-
 217 Subset strategy performs better than the Random-Full strategy. We see displacement while keeping
 218 the original character performs better than the Random-Subset - which keeps the placement consis-
 219 tent. For this specific model, this may imply that it is more sensitive to character coordination than
 220 the specific placement. We also find that across the different noise types, the size of the collective
 221 seems can overcome noise.

222 6.2 RQ2: Feature-Label vs Feature-Only

223 Here we examine how noise affects success for the feature-label vs feature-only strategy. We fix
 224 a level of participation and examine how success changes as noise increases. Figure 3 shows the
 225 decline in success for feature-label and feature-only strategies as a function of the percentage of the
 226 collective subject to noise. We examine three cases, a feature-label scenario where noise is applied
 227 to both the feature and the label (Figure 3a), a feature-label scenario where noise is applied to just
 228 the feature (Figure 3b) and the feature-only scenario (Figure 3c).

229 In Figure 3a we see with moderate levels of participation (0.5%) that noise has a major impact (from
 230 nearly 100% success rate to below 60%). Even at higher levels of participation (1%) we see, at the
 231 high end, noise impacting success. However, in Figure 3b, we see relative robustness to noise at
 232 the higher levels of participation (0.5% and above). The impact of noise in the feature-only strategy
 233 is a more gradual decline (comparisons are not apples to apples because the feature-only strategy
 234 requires higher participation levels for comparable success). Overall, noise that affects labels has a
 235 more significant impact on a collective’s success.

236 6.3 RQ3: Impact of Baseline Frequency

237 Here, we examine how noise impacts target classes that appear more frequently in the $\mathcal{P}_0$ distribution
 238 compared to less commonly seen ones. Figure 4 shows the impact for feature-label strategies. We
 239 find that, counterintuitively, at higher levels of participation (e.g 0.25% and above) low baseline
 240 classes are less susceptible to noise. The opposite holds true in Figure 5, where the high frequency
 241 target class sees some change (some positive/some negative) with the presence of noise, the low
 242 frequency target class is mostly negatively affected by noise. This could be since, unlike in the
 243 feature-label strategy, the feature-only strategy cannot boost the presence of a target class in the
 244 training data ; diluting the signal with noise may more directly impact success.

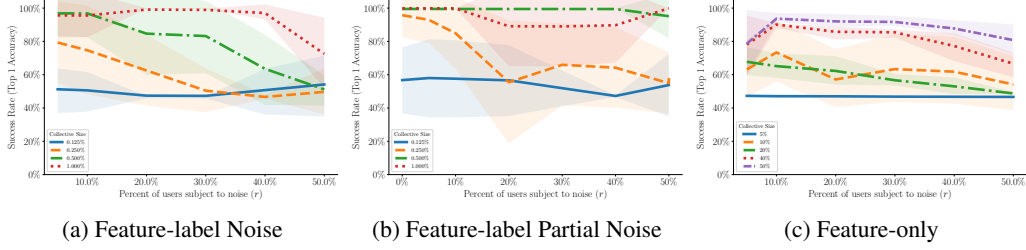


Figure 3: Impact of noise (Random-Subset) on different strategies. We consider various levels of participation. Each curve represents a fixed level of participation. The x-axis is the percentage of that group subject to noise. The left represents the feature-label strategy subject to noise on both the features and the labels. The middle represents the feature-label strategy where only the features are subject to noise. The right figure shows the feature-only strategy. We find that when noise affects both the feature and label, the decline in success is significant - especially at lower levels of participation. When noise just affects the inputs, the declines are more modest.

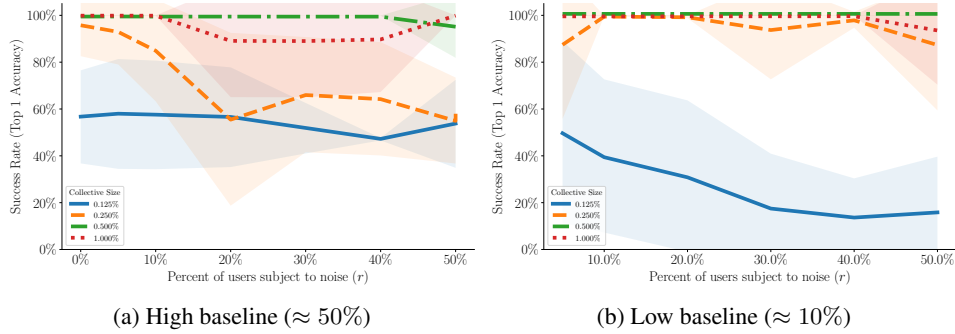


Figure 4: Interaction between underlying frequency and susceptibility to noise (Random-subset) for feature-label strategy. Here we see for moderate levels of participation (0.25% and above), the low baseline scenario is more robust to noise.

7 Discussion

7.1 Analysis on Multiple Collectives

[KVK25] examined how two distinct collectives with different objectives impact each other’s efficacy – in short, when the same signal set is being used to target different classes, both collective’s success rate is greatly reduced. This can be explained by the suboptimality gap ($\Delta_1^2(y_1^*)$) and the cross-signal overlap ($\mathcal{P}_2(\mathcal{X}_1)$). This overlap will be high since the two groups use the same signal. Because they are targeting two different classes, the suboptimality gap may also be large. They also find a case where two collectives, with different target classes and different character usage, still sinks both of their success rates. This can also be explained by the cross-signal overlap - if these character modifications look sufficiently “close” to each other, this term may be large and cause conflicts.

7.2 Trade-offs on Size and Noise

As strategic behavior on algorithmic systems continues to grow, understanding how deviations from a single, unified collective impact on a group’s objective is crucial for both organizers of collective action and system developers. [XZF⁺25] observes the heterogeneity of members of a fan collective in terms of the actions available (e.g based on device type), as well as the detailed instructions they must give to others, which provides room for misinterpretation. This is also similar to parts of a collective action framework mentioned in [KVK25] regarding the action availability and collective construction (i.e how the group is organized).

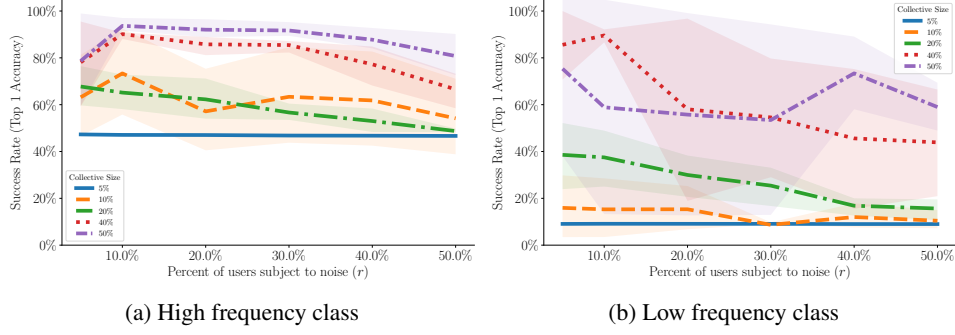


Figure 5: Impact of noise (Random-subset) on the feature-only strategy. Compared to the feature-label strategy, both the high and low frequencies cases are impacted by noise across differing participation levels. The high frequency case, however, sees more gradual declines in the success rate compared to the low frequency case.

264 This theoretical analysis can also be used to help define how and where organizers can place re-
 265 sources. In particular, organizers can try to determine whether increasing the size of the collective
 266 (α_1) is worth the trade-off in potentially increasing cross signal overlap, $\mathcal{P}_2(\mathcal{X}_1)$, or suboptimality
 267 gap, $\Delta_1^2(y_1^*)$. This trade-off has a traditional analogy in managing common pool resources [Ost90]:
 268 the importance of smaller, homogeneous groups to overcome barriers in collective action. We see
 269 it is possible that a smaller group with less noise may be able to outperform a larger group. In
 270 Figure 3a we observe a collective of size 0.25% with 10% noise rate outperforms a collective of
 271 size 0.5% at 40% noise rate. Noise here can be used to characterize a group’s coordination effi-
 272 ciency. Different types of collectives and algorithmic systems might exhibit trade-offs between size
 273 and minimizing noise (if such a trade-off exists). For organizers of collective action knowledge of
 274 these characteristics can help to decide whether resources should be allocated to expand the total
 275 collective size or to more tightly coordinate within a small group.

276 7.3 Broader Impacts

277 Algorithmic collective action can generally be used when there’s a difference between the goals of
 278 those who generate data and are affected by models and those who train and deploy them. Our
 279 aim is to continue the work on promoting socially valuable use cases of collective action ([FS22,
 280 AHJ⁺22]); however, we recognize that there could be malicious use cases. We believe that model
 281 developers should be aware of the strategic behavior that can lead to different data distributions.

282 7.4 Limitations and Future Work

283 We assumed that we could easily segregate between distributions. In some cases, this may be more
 284 natural (multiple collectives) while in others (benign noise) may be difficult to do in practice. We
 285 also considered a limited set of “noise” variations on a small set of data. Future work could explore
 286 more ecologically motivated types of deviations, including non-independent noise and adversarial
 287 actions. We also only considered a small classification task, different types of learning paradigms
 288 may be affected by noise differently – this would be an important avenue for future work.

289 8 Conclusion

290 We investigated the role of multiple distributions on the success of collective action. We derived
 291 lower bounds on the success rate in relation to the cross-signal overlap and suboptimality gap. We
 292 empirically evaluated noise in collective action. We found that noise variation matters, and that
 293 noise that affects labels more severely impacts collective action. We note that, for organizers, un-
 294 derstanding the trade-offs between group size and the potential noise is important; different systems
 295 and types of actions may push organizers to invest more in effective coordination in a small group
 296 vs expansion. As strategic interest on algorithmic systems grows, both developers of algorithms
 297 and organizers of collective action must be aware of the potential that differing distributions has on
 298 system outcomes.

References

- [AHJ⁺22] Rediet Abebe, Moritz Hardt, Angela Jin, John Miller, Ludwig Schmidt, and Rebecca Wexler. Adversarial scrutiny of evidentiary statistical software. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, page 1733–1746. ACM, June 2022.
- [ASJCB02] Alessandro Acquisti, Maarten Sierhuis, William J. Clancey, and Jeffrey Bradshaw. Agent-Based Modeling of Collaboration and Work Practices Onboard the International Space Station, January 2002.
- [BBG23] Jack T. Beerman, Gwendal G. Beaumont, and Philippe J. Giabbanelli. A framework for the comparison of errors in agent-based models using machine learning. *Journal of Computational Science*, 72:102119, September 2023.
- [BDFSS24] Omri Ben-Dov, Jake Fawkes, Samira Samadi, and Amartya Sanyal. The Role of Learning Algorithms in Collective Action, June 2024. arXiv:2405.06582 [cs].
- [BMD24] Joachim Baumann and Celestine Mendler-Dünnér. Algorithmic collective action in recommender systems: Promoting songs by reordering playlists. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [Con96] John Conlisk. Why bounded rationality? *Journal of economic literature*, 34(2):669–700, 1996.
- [FS22] Yunhe Feng and Chirag Shah. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11882–11890, 2022.
- [Has20] Bethany Hastie. Platform Workers and Collective Labour Action in the Modern Economy. *University of New Brunswick Law Journal*, 71:40, 2020.
- [HMMDZ23] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünnér, and Tijana Zrnic. Algorithmic Collective Action in Machine Learning, June 2023. arXiv:2302.04262 [cs, stat].
- [JGV21] Meijerink Jeroen, Jansen Giedo, and Daskalova Victoria. *Platform Economy Puzzles: A Multidisciplinary Perspective on Gig Work*. Edward Elgar Publishing, August 2021. Google-Books-ID: cXg8EAAAQBAJ.
- [JT21] Kameni Florentin Flambeau Jiechieu and Norbert Tsopze. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 33(10):5069–5087, 2021.
- [Kah03] Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.
- [KVKS25] Aditya Karan, Nicholas Vincent, Karrie Karahalios, and Hari Sundaram. Algorithmic collective action with two collectives, 2025.
- [Lei21] Ya-Wen Lei. Delivering Solidarity: Platform Architecture and Collective Contention in China’s Platform Economy. *American Sociological Review*, 86(2):279–309, April 2021. Publisher: SAGE Publications Inc.
- [MF22] Caleb Malchik and Joan Feigenbaum. From Data Leverage to Data Co-Ops: An Institutional Model for User Control over Information Access, January 2022. arXiv:2201.10677 [cs].
- [Nug21] Annabel Nugent. Taylor swift fans share tips on how to make old ‘fearless’ album disappear to mark release of ‘taylor’s version’. *Independent*, Apr 2021.
- [Ost90] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, November 1990. Google-Books-ID: 4xg6oUobMz4C.

346 [Pay24] Will B Payne. Review bombing the platformed city: Contested political speech in
347 online local reviews. *Big Data & Society*, 11(3):20539517241275879, 2024.

348 [Qui07] Tom Quilter. Noise Matters in Heterogeneous Populations. Edinburgh School of
349 Economics Discussion Paper Series 169, Edinburgh School of Economics, University
350 of Edinburgh, August 2007.

351 [SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a
352 distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108,
353 2019.

354 [SHMD24] Dorothee Sigg, Moritz Hardt, and Celestine Mendler-Dünnér. Decline Now:
355 A Combinatorial Model for Algorithmic Collective Action, October 2024.
356 arXiv:2410.12633 [cs].

357 [VH21] Nicholas Vincent and Brent Hecht. Can "Conscious Data Contribution" Help Users
358 to Exert "Data Leverage" Against Technology Companies? *Proceedings of the ACM*
359 *on Human-Computer Interaction*, 5(CSCW1):103:1–103:23, April 2021.

360 [VLT⁺21] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. Data
361 Leverage: A Framework for Empowering the Public in its Relationship with Tech-
362 nology Companies. In *Proceedings of the 2021 ACM Conference on Fairness, Ac-
363 countability, and Transparency*, FAccT '21, pages 215–227, New York, NY, USA,
364 March 2021. Association for Computing Machinery.

365 [WAC21] Katie J Wells, Kafui Attoh, and Declan Cullen. "Just-in-Place" labor: Driver orga-
366 nizing in the Uber workplace. *Environment and Planning A: Economy and Space*,
367 53(2):315–331, March 2021. Publisher: SAGE Publications Ltd.

368 [WDS⁺20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue,
369 Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davi-
370 son, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen
371 Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexan-
372 der M. Rush. Huggingface's transformers: State-of-the-art natural language process-
373 ing, 2020.

374 [WED22] Yuxi Wu, W Keith Edwards, and Sauvik Das. "a reasonable thing to ask for": To-
375 wards a unified voice in privacy collective action. In *Proceedings of the 2022 CHI*
376 *Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

377 [WSLA13] Shir L Wang, Kamran Shafi, Chris Lokan, and Hussein A Abbass. An agent-based
378 model to simulate and analyse behaviour under noisy and deceptive information.
379 *Adaptive Behavior*, 21(2):96–117, 2013.

380 [WVM21] Jonathan Woodside, Tara Vinodrai, and Markus Moos. Bottom-up strategies, plat-
381 form worker power and local action: Learning from ridehailing drivers. *Local Econ-*
382 *omy*, 36(4):325–343, June 2021. Publisher: SAGE Publications Ltd.

383 [XFS⁺25] Qing Xiao, Xianzhe Fan, Felix M. Simon, Bingbing Zhang, and Motahhare Eslami.
384 "It Might be Technically Impressive, But It's Practically Useless to us": Motivations,
385 Practices, Challenges, and Opportunities for Cross-Functional Collaboration around
386 AI within the News Industry, February 2025. arXiv:2409.12000 [cs].

387 [XZF⁺25] Qing Xiao, Yuhang Zheng, Xianzhe Fan, Bingbing Zhang, and Zhicong Lu. Let's
388 Influence Algorithms Together: How Millions of Fans Build Collective Understand-
389 ing of Algorithms and Organize Coordinated Algorithmic Actions, February 2025.
390 arXiv:2409.10670 [cs].

391 A Notation Table

392 We use considerable notation for defining relationship between collectives. Here we provide a con-
 393 cise table for referencing these symbols.

Symbol	Description
$TV(\cdot, \cdot)$	Total variation distance between two distributions
ϵ_1	Suboptimality of the classifier
f	Classifier trained on the mixture distribution
$\mathcal{X}$	The entire feature space
$\mathcal{Y}$	The set of labels
$\mathcal{Z}$	The space of all training data $\mathcal{X} \times \mathcal{Y}$
$\mathcal{P}$	Observed mixture distribution
$\mathcal{P}_0$	Base (non-strategic) data distribution
$\mathcal{P}_1$	Distribution induced by collective one’s intended strategy
$\mathcal{P}_2$	Distribution induced by noise or a second collective’s behavior
h_1, h_2	Strategies applied by collectives or noisy actors
g_1, g_2	Signal planting functions that modify the input x
α_1, α_2	Fraction of population following strategy h_1, h_2 respectively
α	Total participating fraction: $\alpha = \alpha_1 + \alpha_2$
r	Proportion of correctly aligned members: $r = \frac{\alpha_1}{\alpha}$
$\mathcal{X}_1, \mathcal{X}_2$	Signal set induced by g (e.g. $\mathcal{X}_1 = \{g_1(x) \mid x \in \mathcal{X}\}$)
$\mathcal{P}_i(\mathcal{X}_j)$	Cross-signal overlap of P_i on signal set $\mathcal{X}_j$
$\Delta_i^j(y^*)$	Suboptimality gap for X_i under P_j
$S(\alpha_1)$	Success probability for collective one
y_1^*	Target class label for collective one

Table 2: Summary of notation used in the paper.

394 B Experimental Details

395 As described in the main body we finetune a `distilbert-based-uncased` [SDCW19] for five
 396 epochs with default hyperparameters using Hugging Face transformer library [WDS⁺20]. This
 397 experimental setting is the same as [HMMDZ23, KVKS25]. The resume dataset from [JT21] was
 398 split into 20,000 training points and 5,000 test points. The default or baseline strategy was to place
 399 a specific character (in this case the ‘{’ character) every 20 words. We evaluated on targeting the
 400 ‘Software Developer’ class, except for the ‘low frequency class’ where the target was ‘Database
 401 Administrator.’

402 During training, each training point is select with a probability r (the noise fraction) to be affected
 403 by a specific noise condition. If the point was selected to be noised, the noise would be applied to
 404 the text, and, for certain contains, the label as well.

405 **Random-Subset:** Instead of the intended character, a character from the list [U+2E18, ‘!’, ‘?’, ‘.’]
 406 was selected uniformly at random to be used.

407 **Random-Full:** Same as above, but the character chosen was from the set of all single characters in
 408 the vocabulary of the tokenizer.

409 **Displaced-Original:** Keep the intended character, but instead of placing every 20 words, choose
 410 uniformly at random (5-30 instances) places to insert the character.

411 **Displaced-Full:** Combine the character selection of Random-Full with the placement behavior of
 412 Displaced-Original.

413 To evaluate whether the signal was planted successfully, for each data point in the test set, we applied
 414 the “true” signal and evaluated whether the trained classifier would predict the target class. We used
 415 the top-one accuracy as done in [HMMDZ23] as the metric of success.

416 All experimental conditions were run 15 times. Experiments were run on a ppc64le based cluster
 417 with V100 Nvidia GPUs. Each iteration took 30 – 40 minutes to complete.

418 C Proof of Theorem 3

419 We restate the theorem here:

Theorem. Consider distribution $\mathcal{P}_1$ and $\mathcal{P}_2$ which are distributed according to $h_1(x)$ and $h_2(x)$ respectively, where $x \sim \mathcal{P}_0$. Let y_1^* be the target class. Then success for the first collective against an ϵ_1 classifier to be lower bounded by

$$S(\alpha_1) \geq 1 - \frac{\alpha_2}{\alpha_1} \mathcal{P}_2(\mathcal{X}_1) \frac{(1 - \epsilon_1) \Delta_1^2(y_1^*) + \epsilon_1}{1 - 2\epsilon_1} - \frac{1 - \alpha}{\alpha_1} \mathcal{P}_0(\mathcal{X}_1) \frac{(1 - \epsilon_1) \Delta_1^0(y_1^*) + \epsilon_1}{1 - 2\epsilon_1}$$

420 *Proof.* We follow the same proof strategy as [HMMDZ23].

421 We consider the multiple distributions present the overall data distribution $\mathcal{P}$ $\mathcal{P}_0$ base distribution;
 422 $\mathcal{P}_1$ group 1’s distribution; $\mathcal{P}_2$ group 2’s distribution

423 We write the mixture distribution as :

$$\mathcal{P}(x, y_1^*) = \alpha_1 \mathcal{P}_1(x, y_1^*) + \alpha_2 \mathcal{P}_2(x, y_1^*) + (1 - \alpha) \mathcal{P}_0(x, y_1^*) \quad (1)$$

424 We also define the suboptimality gap on distribution i for signal set j and target label y^*

$$\Delta_i^j(y^*) = \max_{x \in \mathcal{X}_i^*} (\max_{y \in \mathcal{Y}} \mathcal{P}_j(y|x) - \mathcal{P}_j(y^*|x)) \quad (2)$$

425 We define the suboptimality gap for a specific point x as

$$\Delta_{i,x}^j(y^*) = \max_{y \in \mathcal{Y}} \mathcal{P}_j(y|x) - \mathcal{P}_j(y^*|x) \quad (3)$$

426 Our goal is to find, for which value of α_1 can we guarantee the model to classify a point $x \in \mathcal{X}_1$ to
 427 the target class y_1^*

428 First consider an $\epsilon = 0$ classifier. If the model f classifies any point $x \in \mathcal{X}$ to y_1^* , it must mean that
 429 $\mathcal{P}(y_1^*|x) > \mathcal{P}(y_1|x)$ or equivalently $\mathcal{P}(x, y_1^*) - \mathcal{P}(x, y_1) > 0$ for every $y_1 \neq y_1^*$.

430 In this strategy, both the features and labels are changed. Since $\mathcal{P}_1$ is the distribution which correctly
 431 performs the intended collective action, every point $x_1^* \in \mathcal{X}_1$ maps to y_1^* . Therefore we can simplify
 432 and get

$$\mathcal{P}(x, y_1^*) = \alpha_1 \mathcal{P}_1(x) + \alpha_2 \mathcal{P}_2(x, y_1^*) + (1 - \alpha) \mathcal{P}_0(x, y_1^*)$$

433 Now, for $y \neq y_1^*$ we have

$$\mathcal{P}(x, y) = \alpha_1 \mathcal{P}_1(x, y) + \alpha_2 \mathcal{P}_2(x, y) + (1 - \alpha) \mathcal{P}_0(x, y)$$

434 Given $\mathcal{P}_1$ maps everything to y_1^* and nothing else this first term is 0 so we can write simplify to

$$\mathcal{P}(x, y) = \alpha_2 \mathcal{P}_2(x, y) + (1 - \alpha) \mathcal{P}_0(x, y)$$

435 So we can write $\mathcal{P}(x, y_1^*) - \mathcal{P}(x, y) > 0$ as

$$\alpha_1 \mathcal{P}_1(x) + \alpha_2 \mathcal{P}_2(x, y_1^*) + (1 - \alpha) \mathcal{P}_0(x, y_1^*) - (\alpha_2 \mathcal{P}_2(x, y) + (1 - \alpha) \mathcal{P}_0(x, y)) > 0$$

436 After rearranging we have

$$\alpha_1 \mathcal{P}_1(x) \geq \alpha_2 \mathcal{P}_2(x) * (\mathcal{P}_2(y|x) - \mathcal{P}_2(y_1|x)) + (1 - \alpha) \mathcal{P}_0(x) (\mathcal{P}_0(y|x) - \mathcal{P}_0(y_1^*|x))$$

437 This must hold true for any y so we can replace the rhs terms by $\Delta_{i,x}^j(y^*)$, in other words we have
 438 a sufficient condition for the size of α_1 as

$$\alpha_1 \geq \frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \alpha_2 \Delta_{1,x}^2(y_1^*) + \frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} (1 - \alpha) \Delta_{1,x}^0(y_1^*)$$

$$\begin{aligned} S(\alpha) &= \Pr_{x \sim \mathcal{P}_1} \{f(x) = y_1^*\} \\ &\geq \Pr_{x \sim \mathcal{P}_1} \left\{ \alpha_1 \geq \frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \alpha_2 \Delta_{1,x}^2(y_1^*) + \frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} (1 - \alpha) \Delta_{1,x}^0(y_1^*) \right\} \\ &= \mathbb{E}_{x \sim \mathcal{P}_1} \mathbf{1} \left\{ \alpha_1 \geq \frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \alpha_2 \Delta_{1,x}^2(y_1^*) + \frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} (1 - \alpha) \Delta_{1,x}^0(y_1^*) \right\} \\ &= \mathbb{E}_{x \sim \mathcal{P}_1} \mathbf{1} \left\{ 1 - \frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \frac{\alpha_2}{\alpha_1} \Delta_{1,x}^2(y_1^*) - \frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} \frac{1 - \alpha}{\alpha_1} \Delta_{1,x}^0(y_1^*) \geq 0 \right\} \\ &\geq \mathbb{E}_{x \sim \mathcal{P}_1} \left[1 - \frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \frac{\alpha_2}{\alpha_1} \Delta_{1,x}^2(y_1^*) - \frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} \frac{1 - \alpha}{\alpha_1} \Delta_{1,x}^0(y_1^*) \right] \\ &= 1 - \frac{\alpha_2}{\alpha_1} \mathbb{E}_{x \sim \mathcal{P}_1} \left[\frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \Delta_{1,x}^2(y_1^*) \right] - \frac{1 - \alpha}{\alpha_1} \mathbb{E}_{x \sim \mathcal{P}_1} \left[\frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} \Delta_{1,x}^0(y_1^*) \right] \\ &\geq 1 - \frac{\alpha_2}{\alpha_1} \mathcal{P}_2(\mathcal{X}_1) \Delta_1^2(y_1^*) - \frac{1 - \alpha}{\alpha_1} \mathcal{P}_0(\mathcal{X}_1) \Delta_1^0(y_1^*) \end{aligned}$$

439 With the final line using the fact that the delta we max over all x 's

440 Now for distribution where $TV(P, P') \leq \epsilon_1$. By Lemma 11 in [HMMDZ23] we have that
 441 $\mathcal{P}'(y^*|x) > \mathcal{P}'(y|x)$ when $\mathcal{P}(y^*|x) > \mathcal{P}(y|x) + \frac{\epsilon_1}{1 - \epsilon_1}$

We than write this as

$$\mathcal{P}(y_1^*|x) \mathcal{P}(x) > \mathcal{P}(y|x) \mathcal{P}(x) + \frac{\epsilon_1}{1 - \epsilon_1} \mathcal{P}(x)$$

442 Following the same steps, procedure as before we get

$$\alpha_1 > \alpha_2 \frac{\mathcal{P}_2(x)}{\mathcal{P}_1(x)} \left(\frac{(1 - \epsilon_1) \Delta_2^1(y_1^*) + \epsilon_1}{1 + 2\epsilon_1} \right) + (1 - \alpha) \frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)} \left(\frac{(1 - \epsilon_1) \Delta_0^1(y_1^*) + \epsilon_1}{1 + 2\epsilon_1} \right)$$

443 Computing the $S(\alpha_1)$ again we get the condition we get

$$S(\alpha_1) \geq 1 - \frac{\alpha_2}{\alpha_1} \mathcal{P}_2(\mathcal{X}_1) \cdot \frac{(1 - \epsilon_1) \Delta_2^1(y_1^*) + \epsilon_1}{1 - 2\epsilon_1} - \frac{1 - \alpha}{\alpha_1} \mathcal{P}_0(\mathcal{X}_1) \frac{(1 - \epsilon_1) \Delta_0^1(y_1^*) + \epsilon_1}{1 - 2\epsilon_1}$$

444

□

445 D Proof of Theorem 4

446 We restate the theorem:

Theorem (Feature-only with two distributions). *Consider distribution $\mathcal{P}_1$ and $\mathcal{P}_2$ which are distributed by $h_1(x)$ and $h_2(x)$ respectively, where $x \sim \mathcal{P}_0$. Suppose there exist a p such that*

$\mathcal{P}_0(y^*|x) \geq p, \forall x \in \mathcal{X}$. Then success for the first collective against an ϵ_1 classifier (against P_1^*) is lower bounded by

$$S(\alpha_1) \geq 1 - \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(X_1^*) \cdot \Delta_1^2(y^*)(1 - \epsilon_1)}{p(1 - \epsilon_1) - \epsilon_1} - \frac{1 - \alpha}{\alpha} \mathcal{P}_0(\mathcal{X}_1) \cdot \frac{(1 - p)(1 - \epsilon_1) + \epsilon_1}{p(1 - \epsilon_1) - \epsilon_1}$$

447 *Proof.* Similar to Theorem 3, we start with the $\epsilon = 0$. We require that, for any $x^* \in \mathcal{X}_1$ that
 448 $\mathcal{P}(y_1^*|x^*) > \mathcal{P}(y|x^*) \forall y \neq y_1^*$. This is equivalent to $\mathcal{P}(x^*, y_1^*) > \mathcal{P}(x^*, y)$ Our goal is find a
 449 sufficient condition for α_1

We get a lower bound for $\mathcal{P}(x^*, y^*)$ by writing the mixture distribution as.

$$\mathcal{P}(x^*, y^*) = \alpha_1 \mathcal{P}_1(x^*, y^*) + \alpha_2 \mathcal{P}_2(x^*, y^*) + (1 - \alpha) \mathcal{P}_0(x^*, y^*) \geq \alpha \mathcal{P}_1(x^*, y^*) + \alpha_2 \mathcal{P}_2(x^*, y^*)$$

450 where $\alpha = \alpha_1 + \alpha_2$

We also have when $y \neq y_1^*$ that

$$\mathcal{P}(x^*, y) = \alpha_1 \mathcal{P}_1(x^*, y) + \alpha_2 \mathcal{P}_2(x^*, y) + (1 - \alpha) \mathcal{P}_0(x^*, y) = \alpha_2 \mathcal{P}_2(x^*, y) + (1 - \alpha) \mathcal{P}_0(x^*, y)$$

Hence if

$$\alpha_1 \mathcal{P}_1(x^*, y^*) + \alpha_2 \mathcal{P}_2(x^*, y^*) \geq \alpha_2 \mathcal{P}_2(x^*, y) + (1 - \alpha) \mathcal{P}_0(x^*, y)$$

451 than $\mathcal{P}(x^*, y^*) > \mathcal{P}(x^*, y)$

Rearranging we have

$$\alpha_1 \mathcal{P}_1(x^*, y^*) \geq \alpha_2 \mathcal{P}_2(x^*, y) - \alpha_2 \mathcal{P}_2(x^*, y^*) + (1 - \alpha) \mathcal{P}_0(x^*, y)$$

$$\alpha_1 \mathcal{P}_1(x^*, y^*) \geq \alpha_2 \mathcal{P}_2(x^*)(\mathcal{P}_2(y|x^*) - \mathcal{P}_2(y^*|x^*)) + (1 - \alpha) \mathcal{P}_0(y|x^*) \mathcal{P}_0(x^*)$$

$$\alpha_1 \geq \frac{\alpha_2 \mathcal{P}_2(x^*)(\mathcal{P}_2(y|x^*) - \mathcal{P}_2(y^*|x^*))}{\mathcal{P}_1(x^*, y^*)} + (1 - \alpha) \frac{\mathcal{P}_0(y|x^*) \mathcal{P}_0(x^*)}{\mathcal{P}_1(x^*, y^*)}$$

452 For our bound on α_1 we seek to maximize the rhs. We can do this by upper bounding $\mathcal{P}_2(y|x^*) -$
 453 $\mathcal{P}_2(y^*|x^*)$ by the suboptimality gap Δ_1^2 . By assumption we have that $\mathcal{P}_0(y^*|x) \geq p$ for all $x \in \mathcal{X}$
 454 and hence $\mathcal{P}_0(y|x^*) \leq 1 - p$. For the denominator we note that $\mathcal{P}_1(x^*, y^*) \geq \mathcal{P}_0(g^{-1}(x^*), y^*) \geq$
 455 $p \mathcal{P}_0(g^{-1}(x^*))$ Using this lowerbound in the denominator we get our required alpha must be at least

$$\alpha_1 \geq \alpha_2 \cdot \frac{\mathcal{P}_2(x^*) \Delta_1^2}{p \mathcal{P}_0(g^{-1}(x^*))} + (1 - \alpha) \frac{(1 - p) \mathcal{P}_0(x^*)}{p \mathcal{P}_0(g^{-1}(x^*))}$$

To compute the success rate we have

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

$$S = \sum_{x^* \in \mathcal{X}_1} \Pr_{x \sim \mathcal{P}_0^*} \{f(g(x)) = y^* | x \in g^{-1}(x^*)\} \Pr_{x \in \mathcal{P}_0} \{x \in g^{-1}(x^*)\}$$

$$S = \sum_{x^* \in \mathcal{X}_1} \mathbf{1}\{f(x^*) = y^*\} \mathcal{P}_0(g^{-1}(x^*))$$

For a given fixed x^* we have

$$\begin{aligned} \mathbf{1}\left\{f(x^*) = y^*\right\} &\geq \mathbf{1}\left\{\alpha_1 \geq \alpha_2 \cdot \frac{\mathcal{P}_2(x^*) \Delta_1^2}{p \mathcal{P}_0(g^{-1}(x^*))} + (1 - \alpha) \frac{(1 - p) \mathcal{P}_0(x^*)}{p \mathcal{P}_0(g^{-1}(x^*))}\right\} \\ &= \mathbf{1}\left\{1 - \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(x^*) \Delta_1^2}{p \mathcal{P}_0(g^{-1}(x^*))} - \frac{(1 - \alpha)}{\alpha_1} \frac{(1 - p) \mathcal{P}_0(x^*)}{p \mathcal{P}_0(g^{-1}(x^*))} > 0\right\} \end{aligned}$$

$$\geq 1 - \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(x^*)\Delta_1^2}{p\mathcal{P}_0(g^{-1}(x^*))} - \frac{(1-\alpha)}{\alpha_1} \frac{(1-p)\mathcal{P}_0(x^*)}{p\mathcal{P}_0(g^{-1}(x^*))}$$

456 Computing the summation we get

$$\begin{aligned} & Pr_{x \sim \mathcal{P}_0^*} \{f(g(x)) = y^*\} \\ &= 1 - \sum_{x^* \in \mathcal{X}_1} \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(x^*)\Delta_1^2}{p\mathcal{P}_0(g^{-1}(x^*))} \cdot \mathcal{P}_0(g_1^{-1}(x^*)) - \sum_{x^* \in \mathcal{X}_1} \frac{(1-\alpha)}{\alpha_1} \frac{(1-p)\mathcal{P}_0(x^*)}{p\mathcal{P}_0(g^{-1}(x^*))} \cdot \mathcal{P}_0(g_1^{-1}(x^*)) \\ &\geq 1 - \frac{\alpha_2}{p\alpha_1} \mathcal{P}_2(\mathcal{X}_1)\Delta_1^2 - \frac{(1-\alpha) \cdot (1-p)}{p\alpha_1} \mathcal{P}_0(X^*) \end{aligned}$$

457 Which gives us the $\epsilon = 0$ case.

458 When considering $\epsilon > 0$, we again cite Lemma 11 in [HMMDZ23], that: $\mathcal{P}(x^*, y^*) > \mathcal{P}(x^*, y) +$
459 $\frac{\epsilon}{1-\epsilon} \mathcal{P}(x^*)$

Expanding all terms out with the mixture distribution we have

$$\mathcal{P}(x^*, y) = \alpha\mathcal{P}_1(x^*, y^*) + \alpha_2\mathcal{P}_2(x^*, y^*) + (1-\alpha_1-\alpha_2)\mathcal{P}_0(x^*, y^*) \geq \alpha\mathcal{P}_1(x^*, y^*) + \alpha_2\mathcal{P}_2(x^*, y^*)$$

460 Plugging this into our expression we have

$$\alpha\mathcal{P}_1(x^*, y^*) + \alpha_2\mathcal{P}_2(x^*, y^*) > \alpha_2\mathcal{P}_2(x^*, y) + (1-\alpha)\mathcal{P}_0(x^*, y) + \frac{\epsilon}{1-\epsilon}(\alpha\mathcal{P}_1(x^*) + \alpha_2\mathcal{P}_2(x^*) + (1-\alpha)\mathcal{P}_0(x^*))$$

461 Using conditionals we can rewrite this as

$$\begin{aligned} & \alpha_1\mathcal{P}_1(y^*|x^*)\mathcal{P}_1(x^*) + \alpha_2\mathcal{P}_2(y^*|x^*)\mathcal{P}_2(x^*) > \\ & \alpha_2\mathcal{P}_2(y|x^*)\mathcal{P}_2(x^*) + (1-\alpha)\mathcal{P}_0(y|x^*)\mathcal{P}_0(x^*) + \frac{\epsilon}{1-\epsilon}(\alpha_1\mathcal{P}_1(x^*) + \alpha_2\mathcal{P}_2(x^*) + (1-\alpha)\mathcal{P}_0(x^*)) \end{aligned}$$

462 Rearranging with α_1 we get

$$\alpha_1 > \alpha_2 \frac{\mathcal{P}_2(x^*)}{\mathcal{P}_1(x^*)} \cdot \frac{(\mathcal{P}_2(y|x^*) - \mathcal{P}_2(y^*|x^*))}{\mathcal{P}_1(y^*|x^*) - \frac{\epsilon}{1-\epsilon}} + (1-\alpha) \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_1(x^*)} \cdot \frac{(\mathcal{P}_0(y|x^*) + \frac{\epsilon}{1-\epsilon})}{(\mathcal{P}_1(y^*|x^*) - \frac{\epsilon}{1-\epsilon})}$$

463 We note that $\mathcal{P}_2(y|x^*) - \mathcal{P}_2(y^*|x^*)$ can be upper bounded by $\Delta_1^2(y^*)$ We can also upper bound and
464 lower bound the numerators and denominators with the following expressions

465 $\mathcal{P}_1(x^*) \geq \mathcal{P}_0(g^{-1}(x^*))$ as every point x^* for $\mathcal{P}_1$ is mapped from some point in $\mathcal{P}_0$ and likewise
466 $\mathcal{P}_1(y^*|x^*) \geq \mathcal{P}_0(y^*|g_1^{-1}(x^*)) \geq p$ From the support of $\mathcal{P}_0(y_1^*|x) \geq p$ we have $\mathcal{P}_0(y|x^*) \leq 1-p$
467 for $y \neq y_1^*$

468 All together we get

$$\alpha_1 > \alpha_2 \frac{\mathcal{P}_2(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{\Delta_1^2(y^*)}{p - \frac{\epsilon}{1-\epsilon}} + (1-\alpha) \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{(1-p + \frac{\epsilon}{1-\epsilon})}{(p - \frac{\epsilon}{1-\epsilon})}$$

469 With this we can compute the expected value as above

For a given fixed x^* we have

$$\mathbf{1} \left\{ f(x^*) = y^* \right\} \geq \mathbf{1} \left\{ \alpha_1 > \alpha_2 \frac{\mathcal{P}_2(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{\Delta_1^2(y^*)}{p - \frac{\epsilon}{1-\epsilon}} + (1-\alpha) \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{(1-p + \frac{\epsilon}{1-\epsilon})}{(p - \frac{\epsilon}{1-\epsilon})} \right\}$$

$$\begin{aligned}
&= \mathbf{1} \left\{ 1 - \frac{\alpha_2}{\alpha_1} \frac{\mathcal{P}_2(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{\Delta_1^2(y^*)}{p - \frac{\epsilon}{1-\epsilon}} - \frac{1-\alpha}{\alpha_1} \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{(1-p + \frac{\epsilon}{1-\epsilon})}{(p - \frac{\epsilon}{1-\epsilon})} > 0 \right\} \\
&\geq 1 - \frac{\alpha_2}{\alpha_1} \frac{\mathcal{P}_2(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{\Delta_1^2(y^*)}{p - \frac{\epsilon}{1-\epsilon}} - \frac{1-\alpha}{\alpha_1} \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{(1-p + \frac{\epsilon}{1-\epsilon})}{(p - \frac{\epsilon}{1-\epsilon})}
\end{aligned}$$

470 Plugging back in we get

$$\begin{aligned}
&\Pr_{x \sim \mathcal{P}_0^*} \{f(g(x)) = y^*\} \\
&= 1 - \sum_{x^* \in \mathcal{X}_1} \frac{\alpha_2}{\alpha_1} \frac{\mathcal{P}_2(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{\Delta_1^2(y^*)}{p - \frac{\epsilon}{1-\epsilon}} \cdot \mathcal{P}_0(g_1^{-1}(x^*)) - \sum_{x^* \in \mathcal{X}_1} \frac{1-\alpha}{\alpha_1} \frac{\mathcal{P}_0(x^*)}{\mathcal{P}_0(g_1^{-1}(x^*))} \cdot \frac{(1-p + \frac{\epsilon}{1-\epsilon})}{(p - \frac{\epsilon}{1-\epsilon})} \cdot \mathcal{P}_0(g_1^{-1}(x^*)) \\
&\geq 1 - \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(X^*) \cdot \Delta_1^2(y^*)}{p - \frac{\epsilon}{1-\epsilon}} - \frac{1-\alpha}{\alpha} \mathcal{P}_0(X^*) \cdot \frac{1-p + \frac{\epsilon}{1-\epsilon}}{p - \frac{\epsilon}{1-\epsilon}} \\
&\geq 1 - \frac{\alpha_2}{\alpha_1} \cdot \frac{\mathcal{P}_2(X^*) \cdot \Delta_1^2(y^*)(1-\epsilon)}{p(1-\epsilon) - \epsilon} - \frac{1-\alpha}{\alpha} \mathcal{P}_0(X^*) \cdot \frac{(1-p)(1-\epsilon) + \epsilon}{p(1-\epsilon) - \epsilon}
\end{aligned}$$

471

□

472 E Extension to many collectives

We can extend the mixture distribution to include an arbitrary number of distributions. This can be written as

$$\mathcal{P} = \sum_{i=1}^n \alpha_i \mathcal{P}_i + (1-\alpha) \mathcal{P}_0$$

473 where $\alpha = \sum_{i=1}^n \alpha_i$

474 By following the same procedure in Appendix C we provide a claim for success for collective one
475 against an arbitrary number of distributions.

Theorem 5. *The success of collective action for the feature-label strategy against many distributions is lower bounded by*

$$S(\alpha_1) \geq 1 - \frac{1-\alpha}{\alpha_1} \mathcal{P}_0(\mathcal{X}_1) \frac{(1-\epsilon_1)\Delta_0^1(y_1^*) + \epsilon_1}{1-2\epsilon_1} - \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \mathcal{P}_i(\mathcal{X}_1) \cdot \frac{(1-\epsilon_1)\Delta_i^1(y_1^*) + \epsilon_1}{1-2\epsilon_1}$$

Proof. We repeat the same expansion in Appendix C for an arbitrary number of distributions.

$$\mathcal{P} = \sum_{i=1}^n \alpha_i \mathcal{P}_i + (1-\alpha) \mathcal{P}_0$$

476 The rest follows the same procedure, as each additional mixture term can be handled independently.

477

□

478 We can do the same for feature-only strategies.

Theorem 6. *The success of collective action for the feature-only strategy against many distributions is lower bounded by*

$$\geq 1 - \frac{1-\alpha}{\alpha} \mathcal{P}_0(X^*) \cdot \frac{(1-p)(1-\epsilon) + \epsilon}{p(1-\epsilon) - \epsilon} - \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \cdot \frac{\mathcal{P}_i(\mathcal{X}_1) \cdot \Delta_1^i(y_1^*)(1-\epsilon)}{p(1-\epsilon) - \epsilon}$$

479 *Proof.* We repeat the same expansion in Appendix D for an arbitrary number of distributions with
480 the same expansion of the mixture distribution and the rest follows the same procedure, as each
481 additional mixture term can be handled independently. □

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims in the abstract and introduction are backed by results and the discussion.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: There is a separate Limitations section that discusses them.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We state the assumptions in the text and provide the full proof in the technical appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide the high level details in the main body and the detailed version in the Appendix. We aim to release the code upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Code will be open sourced if accepted as well as provided to reviewers in a supplemental section.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide them in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We provide 1 STD error bars to all plots.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We provide the GPU and architecture that was used and the time per iterations.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We abide by all ethical considerations in conducting this research.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We mention some of the socially important use cases of this work in the introduction, and in the Discussion we mention that, while we intend for this analysis to improve social impacts, we recognize, that with any data driven system, there are potential negative ramifications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No such risk

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: None generated.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.