How Do Classifiers Induce Agents to Invest Effort Strategically?

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Manish Raghavan
Cornell University
manish@cs.cornell.edu

August 2, 2019

Abstract

Algorithms are often used to produce decision-making rules that classify or evaluate individuals. When these individuals have incentives to be classified a certain way, they may behave strategically to influence their outcomes. We develop a model for how strategic agents can invest effort in order to change the outcomes they receive, and we give a tight characterization of when such agents can be incentivized to invest specified forms of effort into improving their outcomes as opposed to "gaming" the classifier. We show that whenever any "reasonable" mechanism can do so, a simple linear mechanism suffices.

1 Introduction

One of the fundamental insights in the economics of information is the way in which assessing people (students, job applicants, employees) can serve two purposes simultaneously: it can identify the strongest performers, and it can also motivate people to invest effort in improving their performance [35]. This principle has only grown in importance with the rise in algorithmic methods for predicting individual performance across a wide range of domains, including education, employment, and finance.

A key challenge is that we do not generally have access to the true underlying properties that we need for an assessment; rather, they are encoded by an intermediate layer of *features*, so that the true properties determine the features, and the features then determine our assessment. Standardized testing in education is a canonical example, in which a test score serves as a proxy feature for a student's level of learning, mastery of material, and perhaps other properties we are seeking to evaluate as well. In this case, as in many others, the quantity we wish to measure is unobservable, or at the very least, difficult to accurately measure; the observed feature is a construct interposed between the decision rule and the intended quantity.

This role that features play, as a kind of necessary interface between the underlying attributes and the decisions that depend on them, leads to a number of challenges. In particular, when an individual invests effort to perform better on a measure designed by an evaluator, there is a basic tension between effort invested to raise the true underlying attributes that the evaluator cares about, and effort that may serve to improve the proxy features without actually improving the underlying attributes. This tension appears in many contexts — it is the problem of *gaming* the evaluation rule, and it underlies the formulation of *Goodhart's Law*, widely known in the economics literature, which states that once a proxy measure becomes a goal in itself, it is no longer a useful measure [19]. This principle also underpins concerns about strategic gaming of evaluations in search engine rankings [12], credit scoring [3, 16], academic paper visibility [4], reputation management [37], and many other domains.

Incentivizing a designated effort investment. These considerations are at the heart of the following class of design problems, illustrated schematically in Figure 1. An *evaluator* creates a decision rule for assessing an *agent* in terms of a set of features, and this leads the agent to make choices about how to invest effort across their actions to improve these features. In many settings, the evaluator views some forms of agent effort as valuable and others as wasteful or undesirable. For example, if the agent is a student and the evaluator is constructing a standardized test, then the evaluator would likely view it as a good outcome if the existence of the test causes the student to study and learn the material, but a bad outcome if the existence of the test causes the student to spend a huge amount of effort learning idiosyncratic test-taking heuristics specific to the format of the test, or to spend effort on cheating. Similarly, a job applicant (the agent) could prepare for a job interview given by a potential employer (the evaluator) either by preparing for and learning material that would directly improve their job performance (a good outcome for both the agent and the evaluator), or by superficially memorizing answers to questions that they find on-line (a less desirable outcome).

Thus, to view an agent's effort in improving their features as necessarily a form of "gaming" is to miss an important subtlety: some forms of effort correspond intuitively to gaming, while others correspond to self-improvement. If we think of the evaluator as having an opinion on which forms of agent effort they would like to promote, then from the evaluator's point of view, some decision rules work better than others in creating appropriate incentives: they would like to create a decision rule whose incentives lead the agent to invest in forms of effort that the evaluator considers valuable.

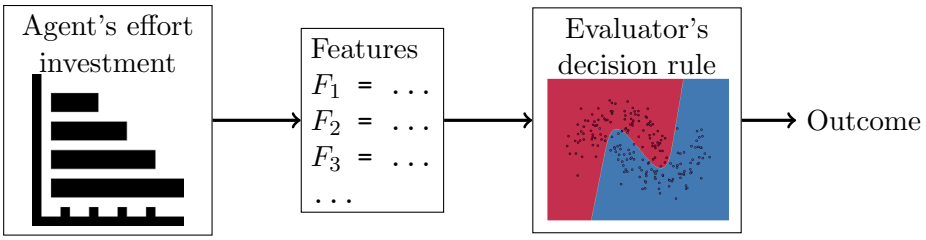


Figure 1: The basic framework: an agent chooses how to invest effort to improve the values of certain features, and an evaluator chooses a decision rule that creates indirect incentives favoring certain investments of effort over others.

These concerns have long been discussed in the education literature surrounding the issue of high-stakes standardized testing. In his book "Measuring Up," Daniel Koretz writes,

> Test preparation has been the focus of intense argument for many years, and all sorts of different terms have been used to describe both good and bad forms... I think it's best to... distinguish between seven different types of test preparation: Working more effectively; Teaching more; Working harder; Reallocation; Alignment; Coaching; Cheating. The first three are what proponents of high-stakes testing want to see [28].

Because teachers are evaluated based on their students' performance on a test, they change their behavior in order to improve their outcomes. As Koretz notes, this can incentivize the investment of both productive and unproductive forms of effort.

What are the design principles that could help in creating a decision that incentives the kinds of effort that the evaluator wants to promote? Keeping the evaluation rule and the features secret, so as to make them harder to game, is generally not viewed as a robust solution, since information about the evaluation process tends to leak out simply by observing the decisions being made, and secrecy can create inequalities between insiders who know how the system works and outsiders who don't. Nor should the goal be simply to create a decision rule that cannot be affected at all by an agent's behavior; while this eliminates the risk of gaming, it also eliminates the opportunity for the decision rule to incentivize behavior that the evaluator views as valuable.

If there were no intermediate features, and the evaluator could completely observe an agent's choices about how they spent their effort across different actions, then the evaluator could simply reward exactly the actions they want to incentivize. But when the actions taken by an individual are hidden, and can be perceived only through an intermediate layer of proxy features, then the evaluator cannot necessarily tell whether these features are the result of effort they intended to promote (improving the underlying attribute that the feature is intended to measure) or effort from other actions that also affect the feature. In the presence of these constraints, can one design evaluation rules that nonetheless incentivize the intended set of behaviors?

To return to our stylized example involving students as agents and teachers as evaluators, a teacher can choose among many possible grading schemes to announce to their class; each corresponds to a candidate decision rule, and each could potentially incentivize different forms of effort on the part of the students. For example, the teacher could announce that a certain percentage of the total course grade depends on homework scores, and the remaining percentage depends on exam scores. In this context, the homework and the exam scores are the features that the teacher is able to observe, and the students have various actions at their disposal — studying to learn material, cheating, or other strategies — that can improve these feature values. How does the way in which the teacher balances the percentage weights on the different forms of coursework —

producing different possible decision rules — affect the decisions students make about effort? As we will see in the next section, the model we develop here suggests some delicate ways in which choices about a decision rule can in principle have significant effects on agents' decisions about effort.

These effects are not unique to the classroom setting. To take an example from a very different domain, consider a restaurant trying to improve its visibility on a review-based platform (e.g. Yelp). Here we can think of the platform as the evaluator constructing a decision rule and the restaurant as the agent: the platform determines a restaurant's rank based on both the quality of reviews and the number of users who physically visit it, both of which are meant to serve as proxies for its overall quality. The restaurant can individually game either of these metrics by paying people to write positive reviews or to physically check in to their location, but improving the quality of the restaurant will ultimately improve both simultaneously. Thus, the platform may wish to consider both metrics, rating and popularity, in some balanced way in order to increase the restaurant's incentive to improve.

The present work: Designing evaluation rules. In this paper, we develop a model for this process of incentivizing effort, when actions can only be measured through intermediate features. We cast our model as an interaction between an *evaluator* who is performing an assessment, and an *agent* who wants to score well on this assessment. An instance of the problem consists of a set of actions in which the agent can invest chosen amounts of *effort*, and a set of functions determining how the levels of effort spent on these actions translate into the values of *features* that are observable to the evaluator.

The evaluator's design task is to create an evaluation rule that takes the feature values as input, and produces a numerical score as output. (Crucially, the evaluation rule is not a function of the agent's level of effort in the actions, only of the feature values.) The agent's goal is to achieve a high score, and to do this, they will optimize how they allocate their effort across actions. The evaluator's goal is to induce a specific *effort profile* from the agent — specifying a level of effort devoted to each action — and the evaluator seeks an evaluation rule that causes the agent to decide on this effort profile. Again, Figure 1 gives a basic view of this pipeline of activities.

Our main result is a characterization of the instances for which the evaluator can create an evaluation rule inducing a specified effort profile, and a polynomial-time algorithm to construct such a rule when it is feasible. As part of our characterization, we find that if there is any evaluation rule, monotone in the feature values, that induces the intended effort profile, then in fact there is one that is linear in the feature values; and we show how to compute a set of coefficients achieving such a rule. Additionally, we provide a tight characterization of which actions can be jointly incentivized.

The crux of our characterization is to consider how an agent is able to "convert" effort from one action to another, or more generally from one set of actions to another set of actions. If it is possible to reallocate effort spent on actions the evaluator is trying to incentivize to actions the evaluator isn't trying to incentivize, in a way that improves the agent's feature values, then it is relatively easy to see that the evaluator won't be able to design a decision rule that incentivizes their desired effort profile: any incentives toward the evaluator's desired effort profile will be undercut by the fact that this effort can be converted away into other undesired forms of effort in a way that improves the agent's outcome. The heart of the result is the converse, providing an if-and-only-if characterization: when such a conversion by the agent isn't possible, then we can use the absence of this conversion to construct an explicit decision rule that incentivizes precisely the effort profile that the evaluator is seeking.

Building on our main result, we consider a set of further questions as well. In particular,

we discuss characterizations of the set of all linear evaluation rules that can incentivize a family of allowed effort profiles, identifying tractable structure for this set in special cases, but greater complexity in general. And we consider the problem of choosing an evaluation rule to optimize over a given set of effort profiles, again identifying tractable special cases and computational hardness in general.

Further Related Work. Our work is most closely related to the principal-agent literature from economics: an evaluator (the principal) wants to set a policy (the evaluation rule) that accounts for the agent's strategic responses. Our main result has some similarities, as well as some key differences, relative to a classical economic formulation in principal-agent models [18, 21, 22, 20]. We explore this connection in further detail in Section 2.4.

In the computer science literature, a growing body of work seeks to characterize the interaction between a decision-making rule and the strategic agents it governs. This was initially formulated as a zero-sum game [11], e.g. in the case of spam detection, and more recently in terms of Stackelberg competitions, in which the evaluator publishes a rule and the agent may respond by manipulating their features strategically [19, 5, 14, 24, 31]. This body of work is different from our approach in a crucial respect, in that it tends to assume that all forms of strategic effort from the agent are undesirable; in our model, on the other hand, we assume that there are certain behaviors that the evaluator wants to incentivize.

There is also work on strategyproof linear regression [7, 10, 13]. The setup of these models is also quite different from ours – typically, the strategic agents submit (x, y) pairs where x is fixed and y can be chosen strategically, and the evaluator's goal is to perform linear regression in a way that incentivizes truthful reporting of y. In our setting, on the other hand, agents strategically generate their features x, and the evaluator rewards them in some way based on those features.

Work exploring other aspects of how evaluation rules lead to investment of effort can be found in the economics literature, particularly in the contexts of hiring [17, 23] and affirmative action [9]. While these models tend to focus on decisions regarding skill acquisition, they broadly consider the investment incentives created by evaluation. Similar ideas can also be found in the Science and Technology Studies literature [38], considering how organizations respond to guidelines and regulations.

As noted above, principal-agent mechanism design problems in which the principal cannot directly observe the agent's actions have been studied in the economics literature [1, 33, 2], and include work on the notion of *moral hazard*. Insurance markets are canonical examples in this domain: the agent reduces their liability by purchasing insurance, and this may lead them to act more recklessly and decrease welfare. The principal cannot directly observe how carefully the agent is acting, only whether the agent makes any insurance claims. These models provide some inspiration for ours; in particular, they are often formalized such that the agent's actions are "effort variables" which, at some cost to the agent, increase the agent's level of "production" [29]. This could be, for example, acting in more healthy ways or driving more carefully in the cases of health and car insurance respectively. Note, however, that in the insurance case, the agent and the principal have aligned incentives in that both prefer that the agent doesn't — e.g., in the case of car insurance — get into an accident. In our model, we make no such assumptions: the agent may have no incentive at all to invest in the evaluator's intended forms of effort beyond the utility derived from the mechanism. The types of scenarios considered in insurance markets can be generalized to domains like share-cropping [8, 36], corporate liability [25], and theories of agency [34]. Steven Kerr provides a detailed list of such instances in his classic paper "On the folly of rewarding A, while hoping for B" [26].

Concerns over strategic behavior also manifest in ways that do not necessarily map to intuitive notions of gaming, but instead where the evaluator does not want to incentivize the agent to take actions that might be counter to their interests. For example, Virginia Eubanks considers a case of risk assessment in the child welfare system; when a risk tool includes features about a family's history of interaction with public services, including aid such as food stamps and public housing, she argues that it has the potential to incentivize families to avoid such services for fear of being labeled high risk [15]. This too would be a case in which the structure and implementation of an evaluation rule can incentivize potentially undesirable actions in agents, and would be interesting to formalize in the language of our model.

Organization of the remainder of the paper. Section 2 contains all the definitions and technical motivation leading up to the formulation and statement of our two main results, Theorems 3 and 5. Sections 3 and 4 contain the proofs of these two results, respectively, and Section 5 considers further extensions.

2 Model and Overview of Results

2.1 A Formal Model of Effort Investment

Here, we develop a formal model of an agent's investment of effort. There are m actions the agent can take, and they must decide to allocate an amount of effort x_j to each activity j. We'll assume the agent has some budget B of effort to invest,[1] so $\sum_{j=1}^{m} x_j \leq B$, and we'll call this investment of effort $x = (x_1, x_2, \ldots, x_m)$ an *effort profile*.

The evaluator cannot directly observe the agent's effort profile, but instead observes features $F_1, \ldots, F_n$ derived from the agent's effort profile. The value of each F_i grows monotonically in the effort the agent invests in certain actions according to an *effort conversion function* $f_i(\cdot)$:

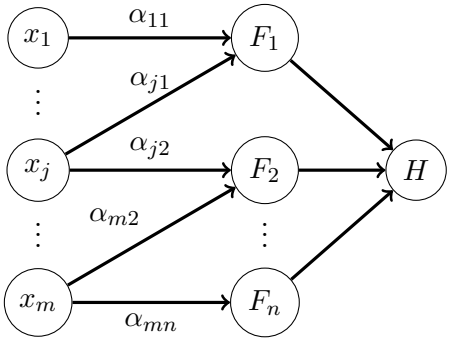
$$F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right), \tag{1}$$

where each $f_i(\cdot)$ is nonnegative, smooth, weakly concave (i.e., actions provide diminishing returns), and strictly increasing. We assume $\alpha_{ji} \geq 0$.

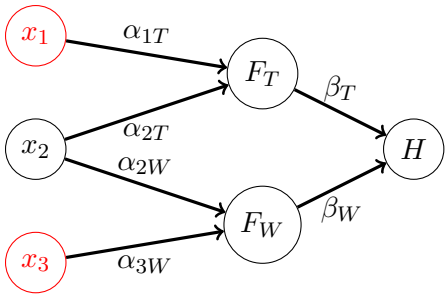
We represent these parameters of the problem using a bipartite graph with the actions $x_1, x_2, \ldots, x_m$ on the left, the features $F_1, \ldots, F_n$ on the right, and an edge of weight α_{ji} whenever $\alpha_{ji} > 0$, so that effort on action x_j contributes to the value of feature F_i. We call this graph, along with the associated parameters (the matrix $\alpha \in \mathbb{R}^{m \times n}$ with entries α_{ji}; functions $f_i : \mathbb{R} \to \mathbb{R}$ for $i \in \{1, ..., n\}$; and a budget B), the *effort graph* G. Figure 2 shows some examples of what G might look like.

The evaluator combines the features generated by the effort using some mechanism M to produce an output H, which is the agent's utility. M is simply a function of the n feature values. In a classification setting, for example, H may be binary (whether or not the agent is classified as positive or negative), or a continuous value (the probability that the agent receives a positive outcome). Because all features are increasing in the amount of effort invested by the agent — in particular, including the kinds of effort we want to incentivize — we'll restrict our attention to the class of

[1] We might instead model the agent as incurring a fixed cost c per unit effort with no budget. In fact, this formulation is in a sense equivalent: for every cost c, there exists a budget B such that an agent with cost c behaves identically to an agent with fixed budget B (and no cost). For clarity, we will deal only with the budgeted case, but our results will extend to the case where effort comes at a cost.



(a) General model



(b) The classroom setting

Figure 2: The conversion of effort to feature values can be represented using a weighted bipartite graph, where effort x_j spent on action j has an edge of weight α_{ji} to feature F_i.

monotone mechanisms, meaning that if agent X has larger values in all features than agent Y, then X's outcome should be at least as good as that of Y. Formally, we write this as follows:

Definition 1. A *monotone mechanism* M on features F_i is a mapping $\mathbb{R}^n \to \mathbb{R}$ such that for $F, F' \in \mathbb{R}^n$ with $F'_i \geq F_i$ for all $i \in \{1, ..., n\}$, $M(F') \geq M(F)$. Also, for any F, there exists $i \in \{1, ..., n\}$ such that strictly increasing F_i strictly increases $M(F)$.

The second of these conditions implies that it is strictly optimal for an agent to invest all of its budget. The agent's utility is simply its outcome H. Thus, for a mechanism M, the agent's optimal strategy is to invest effort to maximize $M(F)$ subject to the constraints that $\sum_{j=1}^{m} x_j \leq B$ and $x_j \geq 0$ for all j. (Recall that in this phrasing, the vector F of feature values is determined from the effort value x_i via the functions $F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right)$.) We can write the agent's search for an optimal strategy succinctly as the following optimization problem:

$$x^* = \arg\max_{x \in \mathbb{R}^m} \; M(F) \qquad \text{s.t.} \sum_{j=1}^{m} x_j \leq B \qquad (2)$$
$$x \geq \mathbf{0}$$

where each component F_i of F is defined as in (1). Throughout this paper, we'll assume that agents behave rationally and optimally, though it would be an interesting subject for future work to consider extensions of this model where agents suffer from behavioral biases. We also note that this is where we make use of the concavity of the functions f_i, since for arbitrary f_i the agent wouldn't necessarily be able to efficiently solve this optimization problem.

2.2 Returning to the classroom example

To illustrate the use of this model, consider the effort graph shown in Figure 2b, encoding the classroom example described in the introduction. There are two pieces of graded work for the class (a test F_T and homework F_W), and the student can study the material (x_2) to improve their scores on both of these. They can also cheat on the test (x_1) and look up homework answers on-line (x_3). Their combined effort $\alpha_{1T}x_1 + \alpha_{2T}x_2$ contributes to their score on the test, and their combined effort $\alpha_{2W}x_2 + \alpha_{3W}x_3$ contributes to their score on the homework. To fully specify the effort graph, we would have to provide a budget B and effort conversion functions f_T and f_W; we leave these uninstantiated, as our main conclusions from this example will not depend on them.

From these scores, the teacher must decide on a student's final grade H. For simplicity, we'll assume the grading scheme is simply a linear combination, meaning $H = \beta_T F_T + \beta_W F_W$ for some real numbers $\beta_T, \beta_W \geq 0$.

The teacher's objective is to incentivize the student to learn the material; thus, they want to set β_T and β_W such that the student invests their entire budget into x_2. Of course, this may not be possible. For example, if α_{1T} and α_{3W} are significantly larger than α_{2T} and α_{2W} respectively, so that it is much easier to cheat on the test and copy homework answers than to study, the student would maximize their utility by investing all of their effort into these undesirable activities.

In fact, we can make this precise as follows. For any unit of effort invested in x_2, the student could instead invest $\frac{\alpha_{2T}}{\alpha_{1T}}$ and $\frac{\alpha_{2W}}{\alpha_{3W}}$ units of effort into x_1 and x_3 respectively without changing the values of F_T and F_W. Moreover, if $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} < 1$, then this substitution strictly reduces the sum $x_1 + x_2 + x_3$, leaving additional effort available (relative to the budget constraint) for raising the values of F_T and F_W. It follows that in any solution with $x_2 > 0$, there is a way to strictly improve it through this substitution. Thus, under this condition, the teacher cannot incentivize the student to only study. This is precisely the type of "conversion" of effort that we discussed briefly in the previous section, from the evaluator's preferred form of effort (x_2) to other forms (x_1 and x_3)

When $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} \geq 1$, on the other hand, a consequence of our results is that that no matter what f_T, f_W and B are, there exist some β_T, β_W that the teacher can choose to incentivize the student to invest all their effort into studying. This may be somewhat surprising – for instance, consider the case where $\alpha_{1T} = \alpha_{3W} = 3$ and $\alpha_{2T} = \alpha_{2W} = 2$, meaning that the best way for the student to maximize their score on each piece of graded work individually is to invest undesirable effort instead of studying. Even so, it turns out that the student can still be incentivized to put all of their effort into studying by appropriately balancing the weight placed on the two pieces of graded work.

This example illustrates several points that will be useful in what follows. First, it makes concrete the basic obstacle against incentivizing a particular form of effort: the possibility that it can be "swapped out" at a favorable exchange rate for other kinds of effort. Second, it shows a particular kind of reason why it might be possible to incentivize a designated form of effort x_i: if investing in x_i improves multiple features simultaneously, the agent might select it even if it is not the most efficient way to increase any one feature individually. This notion of activities that "transfer" across different forms of evaluation, versus activities that fail to transfer, arises in the education literature on testing [27], and our model shows how such effects can lead to valuable incentives.

2.3 Stating the main results

In our example, it turned out that a linear grading scheme was sufficient for the teacher to incentivize the student to study. We formalize such mechanisms as follows.

Definition 2. A *linear mechanism* $M : \mathbb{R}^n \to \mathbb{R}$ is the mapping $M(F) = \beta^\top F = \sum_{i=1}^n \beta_i F_i$ for some $\beta \in \mathbb{R}^n$ such that $\beta_i \geq 0$ for all $i \in \{1, ..., n\}$ and $\sum_{i=1}^n \beta_i > 0$.

Note that we don't require $\sum_{i=1}^n \beta_i$ to be anything in particular; the agent's optimal behavior is invariant to scaling β, so we can normalize β to sum to any intended quantity without affecting the properties of the mechanism. We rule out the mechanism in which all β_i are equal to 0, as it is not a monotone mechanism.

We will say that a mechanism M *incentivizes* effort profile x if x is an optimal response to M. Ultimately, our main result will be to prove the following theorem, characterizing when a given

effort profile can be incentivized. First, we need to define the support of x as

$$\mathcal{S}(x) \triangleq \{j \mid x_j > 0\}. \tag{3}$$

With this, we can state the theorem.

Theorem 3. *For an effort graph G and an effort profile x^*, the following are equivalent:*

1. *There exists a linear mechanism that incentivizes x^*.*
2. *There exists a monotone mechanism that incentivizes x^*.*
3. *For all x such that $\mathcal{S}(x) \subseteq \mathcal{S}(x^*)$, there exists a linear mechanism that incentivizes x.*

Furthermore, there is a polynomial time algorithm that decides the incentivizability of x^* and provides a linear mechanism β to incentivize x^* whenever such β exists.

When there exists a monotone mechanism incentivizing x^*, we'll call both x^* and $\mathcal{S}(x^*)$ *incentivizable*.[2] Informally, when x^* is not incentivizable, this algorithm finds a succinct "obstacle" to any solution with support $\mathcal{S}(x^*)$, meaning no x such that $\mathcal{S}(x) = \mathcal{S}(x^*)$ is incentivizable. The following corollary is a direct consequence of Theorem 3. (We use the notation $[m]$ to represent $\{1, 2, \ldots, m\}$.)

Corollary 4. *For a set $S \subseteq [m]$, some x such that $\mathcal{S}(x) = S$ is incentivizable if and only if all x with $\mathcal{S}(x) = S$ are incentivizable.*

In Section 3, we'll prove Theorem 3. The proof we give is constructive, and it establishes the algorithmic result.

Optimizing over effort profiles. It may be the case that the evaluator doesn't have a single specific effort profile by the agent that they want to incentivize; instead, they may have an objective function defined on effort profiles, and they would like to maximize this objective function over effort profiles that are incentivizable. In other words, the goal is to choose an evaluation rule so that the resulting effort profile it induces performs as well as possible according to the objective function.

In Section 4, we consider the following formulation for such optimization problems. We assume that the evaluator wants to maximize a concave function $g : \mathbb{R}^m \to \mathbb{R}$ over the space of effort profiles, subject to the constraint that the agent only invests effort in a subset $D \subseteq [m]$ of effort variables. To accomplish this, the evaluator selects an evaluation rule so as to incentivize an effort profile x^* with $g(x^*)$ as large as possible. This is what we will mean by optimizing g over the space of effort profiles. In this setting, we show the following results, which we prove in Section 4.

Theorem 5. *Let g be a concave function over the space of effort profiles, and let D be the set of effort variables in which the evaluator is willing to allow investment by the agent.*

1. *If there exists an x^* such that $\mathcal{S}(x^*) = D$ and x^* is incentivizable, then any concave function g can be maximized over the space of effort profiles in polynomial time.*
2. *If $|D|$ is constant, then any concave function g can be maximized over the space of effort profiles in polynomial time.*
3. *In general, there exist concave functions g that are NP-hard to maximize over the space of effort profiles subject to the incentivizability condition.*

[2] A closely related notion in the principal-agent literature is that of an *implementable* action.

In summary, we establish that it is computationally hard to maximize even concave objectives in general, although as long as the number of distinct actions the evaluator is willing to incentivize is small, concave objectives can be efficiently maximized.

The above results characterize optimization over effort profiles; instead, the evaluator may wish to optimize over the space of mechanisms (e.g., to fit to a dataset). We consider the feasibility of such optimization in Section 5, showing that the set of linear mechanisms incentivizing particular actions can be highly nonconvex, making optimization hard in general.

2.4 Principal-Agent Models and Linear Contracts

Now that we have specified the formalism, we are in a position to compare our model with well-studied principal-agent models in economics to see how our results and techniques relate to those from prior work. In the standard principal-agent setting, the principal's objective is to incentivize an agent to invest effort in some particular way [34, 18]. Crucially, the principal cannot observe the agent's action – only some outcome that is influenced by the agent's action. Thus, while the principal cannot directly reward the agent based on the action it takes, it can instead provide rewards based on outcomes that are more likely under desirable actions.

To our knowledge, this framework has yet to be applied to settings based on machine-learning classifiers as we do here; and yet, principal-agent models fit quite naturally in this context. A decision-maker wants to evaluate an agent, which it can only observe indirectly through features. These features, in turn, reflect the actions taken by the agent. In this context, the principal offers a "contract" by specifying an evaluation rule, to which the agent responds strategically by investing its effort so as to improve its evaluation. So far, this is in keeping with the abstract principal-agent framework [34, 18].

Moreover, some of the key results we derive echo known results from previous models, though they also differ in important respects. Linear contracts, in particular, are often necessary or optimal in principal-agent contexts for a variety of reasons. In modeling bidding for government contracts, for example, payment schemes are linear in practice for the sake of simplicity, even though optimal contracts may be nonlinear [30]. In other models, contracts are naturally linear because agents maximize reward in expectation over outcomes generated stochastically from their actions [18].

Even when they aren't necessitated by practical considerations or modeling choices, linear contracts have been shown to be optimal in their own right in some principal-agent models. Holmström and Milgrom [21, 22] consider the interplay between incentives and risk aversion and characterize optimal mechanisms in this setting, finding that under a particular form of risk aversion (exponential utility), linear contracts optimally elicit desired behavior. Our models do not incorporate a corresponding notion of risk aversion, and the role of linear mechanisms in our work arises for fundamentally different reasons.

Hermalin and Katz provide a model more similar to ours, in which observations result stochastically from agents' actions [20]. Drawing on basic optimization results that we use here as well (in particular, duality and Farkas' Lemma), they characterize actions as "implementable" based on whether they can be in some sense replaced by other actions at lower cost to the agent. At a high level, we will rely on a similar strategy to prove Theorem 3.

There are, however, some further fundamental differences between the principal-agent models arising from the work of Hermalin and Katz and the questions and results we pursue here. In particular, the canonical models of principal-agent interaction in economics typically only have the expressive power to to incentivize a single action, which stochastically produces a single observed outcome. This basic difference leads to a set of important distinctions for the modeling goals we have: because our goal is to incentivize investment over multiple activities given a multi-dimensional

feature vector, with the challenge that different mixtures of activities can deterministically produce the same feature vector, our model cannot be captured by these earlier formalisms.

An important assumption in our model, and in many principal-agent models in general, is that the principal knows how the agent's effort affects observations. Recent work has sought to relax this assumption, finding that linear contracts are optimal even when the principal has incomplete knowledge of the agent's cost structure [6]. It would be an interesting subject for future work to extend our model so that the principal does not know or needs to learn the agent's cost structure.

3 Incentivizing Particular Effort Profiles

In this section, we develop a tight characterization of which effort profiles can be incentivized and find linear mechanisms that do so. For simplicity, we'll begin with the special case where the effort profile to be incentivized is $x^* = B \cdot e_j$, with e_j representing the unit vector in coordinate j — that is, the entire budget is invested in effort on action j. Then, we'll apply the insights from this case to the general case.

The special case where $|\mathcal{S}(x^*)| = 1$. Recall that in the classroom example, the tipping point for when the intended effort profile could be incentivized hinged on the question of *substitutability*: the rate at which undesirable effort could be substituted for the intended effort. We'll characterize this rate as the solution to a linear program. In an effort graph G, recall that $\alpha \in \mathbb{R}^{m \times n}$ is the matrix with entries α_{ji}. Let $\tilde{\alpha}_j \in \mathbb{R}^n$ be the jth row of α. Then, we'll define the substitutability of x_j to be

$$\kappa_j \triangleq \min_{y \in \mathbb{R}^m} \; y^\top \mathbf{1} \qquad \text{s.t. } \alpha^\top y \geq \tilde{\alpha}_j \qquad y \geq \mathbf{0} \tag{4}$$

Intuitively, y is a redistribution of effort out of x_j that weakly increases all feature values. Note that $\kappa_j \leq 1$ because the solution $y = e_j$ (the vector with 1 in the jth position and 0 elsewhere) is feasible and has value 1. In Lemma 6, we'll use this notion of substitutability to show that whenever $\kappa_j < 1$, the agent will at optimality put no effort into x_j. Conversely, in Lemma 7, we'll show that the when $\kappa_j = 1$, there exists a linear mechanism incentivizing β incentivizing the solution $x^* = B \cdot e_j$.

It might seem odd that this characterization depends only on κ_j, which is independent of both the budget B and effort conversion functions f_i; however, the particular mechanisms that incentivize x^* will depend on these. This will also be true in the general case: whether or not a particular effort profile can be incentivized will not depend on B or f_i, but the exact mechanisms that do so will.

Lemma 6. *If $\kappa_j < 1$, then in any monotone mechanism M, $x^*_j = 0$.*

Proof. Intuitively, this is an argument formalizing substitution: if $\kappa_j < 1$, replacing each unit of effort in x_j with y_k units of effort (where y comes from the optimal solution to (4)) on each x_k for $k \in [m]$ weakly increases all of the feature values F_i while making the budget constraint slack. Therefore, any solution with $x_j > 0$ cannot be optimal.

In more detail, consider any solution x with $x_j > 0$. We'll begin by showing that the agent's utility is at least as high in the solution x' with $x'_k = x_k + y_k x_j$ for all $k \neq j$ and $x'_j = y_j x_j$, where y is an optimal solution to the linear program in (4). Note that $y_j \leq \kappa_j < 1$, so x' is different from x.

We know from the constraint on (4) that $\alpha^\top y \geq \tilde{\alpha}_j$, and therefore

$$\sum_{k=1}^{m} \alpha_{ki} y_k \geq \alpha_{ji} \tag{5}$$

for all i. Then, by (5),

$$f_i\left(\sum_{k=1}^{m} \alpha_{ki} x_k\right) \leq f_i\left(\sum_{k \neq j} \alpha_{ki} x_k + x_j \sum_{k=1}^{m} \alpha_{ki} y_k\right) = f_i\left(\sum_{k=1}^{m} \alpha_{ki} x'_k\right)$$

Thus, the value of each feature weakly increases from x to x', so in any monotone mechanism M, the agent's utility for x' is at least as high as it is for x. Moreover, the budget constraint on x' isn't tight, since

$$\sum_{k=1}^{m} x'_k = \sum_{k \neq j} (x_k + y_k x_j) + y_j x_j = \sum_{k \neq j} x_k + x_j \sum_{k=1}^{m} y_k < \sum_{k=1}^{m} x_k \leq B.$$

By the definition of a monotone mechanism, no solution for which the budget constraint isn't tight can be optimal, meaning x' is not optimal. This implies that x is not optimal. □

Thus, $\kappa_j < 1$ implies that $x_j = 0$ in any optimal solution. All that remains to show in this special case is the converse: if $\kappa_j = 1$, there exists β that incentivizes the effort profile $x^* = B \cdot e_j$. To do so, define $A(x) \in \mathbb{R}^{m \times n}$ to be the matrix with entries $[A(x)]_{ji} = \alpha_{ji} f_i'([\alpha^\top x]_i)$, and define $a_j(x) \in \mathbb{R}^n$ to be the jth row of $A(x)$. Then, we can define the polytope

$$\mathcal{L}_j \triangleq \{\beta \mid A(x^*)\beta \leq \beta^\top a_j(x^*) \cdot \mathbf{1}\}. \tag{6}$$

By construction, $\mathcal{L}_j$ is the set of linear mechanisms that incentivize x^*. This is because for all $k \in [m]$, every $\beta \in \mathcal{L}_j$ satisfies

$$[A(x^*)\beta]_k \leq \beta^\top a_j(x^*) \iff \sum_{i=1}^{n} \alpha_{ki}\beta_i f_i'([\alpha^\top x^*]_i) \leq \sum_{i=1}^{n} \alpha_{ji}\beta_i f_i'([\alpha^\top x^*]_i) \iff \left.\frac{\partial H}{\partial x_k}\right|_{x^*} \leq \left.\frac{\partial H}{\partial x_j}\right|_{x^*}$$

By Lemma 12 in Appendix A, this implies that x^* is an optimal agent response to any $\beta \in \mathcal{L}_j$. To complete the proof of this special case of Theorem 3, it suffices to show that $\mathcal{L}_j$ is non-empty, which we do via linear programming duality.

Lemma 7. *If $\kappa_j = 1$, then $\mathcal{L}_j$ is non-empty.*

Proof. Consider the following linear program.

$$\max_{\beta \in \mathbb{R}^n} \ \beta^\top a_j(x^*) \qquad \text{s.t. } A(x^*)\beta \leq \mathbf{1} \qquad \beta \geq \mathbf{0} \tag{7}$$

Clearly, if (7) has value at least 1, then $\mathcal{L}_j$ is non-empty because any β achieving the optimum is in $\mathcal{L}_j$ by (6). The dual of (7) is

$$\min_{y \in \mathbb{R}^m} \ y^\top \mathbf{1} \qquad \text{s.t. } A(x^*)^\top y \geq a_j(x^*) \qquad y \geq \mathbf{0} \tag{8}$$

We can simplify the constraints on (8): for all i,

$$[A(x^*)^\top y]_i \geq [a_j(x^*)]_i \iff \sum_{k=1}^{m} \alpha_{ki} y_k f_i'([\alpha^\top x^*]_i) \geq \alpha_{ji} f_i'([\alpha^\top x^*]_i) \iff \sum_{k=1}^{m} \alpha_{ki} y_k \geq \alpha_{ji}$$

Thus, (8) is equivalent to (4), which has value $\kappa_j = 1$ by assumption. By duality, (7) also has value $\kappa_j = 1$, meaning $\mathcal{L}_j$ is non-empty. □

We have shown that if $\kappa_j = 1$, then any $\beta \in \mathcal{L}_j$ incentivizes x^*. Otherwise, by Lemma 6, there are no monotone mechanisms that incentivize x^*. Next, we'll generalize these ideas to prove Theorem 3.

The general case. We'll proceed by defining the analogue of κ_j in the case where the effort profile to be incentivized has support on more than one component. Drawing upon the reasoning in Lemmas 6 and 7, we'll prove Theorem 3.

Consider an arbitrary effort profile x^* such that $\sum_{i=1}^{m} x_j^* = B$, and let $\mathcal{S}(x^*)$ be the support of x^*. Let α_S be α with the rows not indexed by S zeroed out, i.e., $[\alpha_S]_{ji} = \alpha_{ji}$ if $j \in S$ and 0 otherwise. Let $\mathbf{1}_S$ be the vector with a 1 for every $j \in S$ and 0 everywhere else, so $\mathbf{1}_S = \sum_{j \in S} e_j$. Similarly to how we defined κ_j, define

$$\begin{aligned} \kappa_S \triangleq \min_{y \in \mathbb{R}^m, z \in \mathbb{R}^m} \quad & y^\top \mathbf{1} && \text{s.t. } \alpha^\top y \geq \alpha_S^\top z \\ & && z^\top \mathbf{1}_S \geq 1 \\ & && y, z \geq \mathbf{0} \end{aligned} \tag{9}$$

Intuitively, we can think of the effort given by z as being substituted out and replaced by y. Note that $\kappa_S \leq \min_{j \in S} \kappa_j$, because the special case where $z_j = 1$ yields (4). In a generalization of Lemma 6, we'll show that $\kappa_S < 1$ implies that no optimal solution has $x_j > 0$ for all $j \in S$. Lemma 6 formalized an argument based on substitutability, in which the effort invested on a particular node could be moved to other nodes while only improving the agent's utility. We generalize this to the case when effort invested on a subset of the nodes can be replaced by moving that effort elsewhere.

Lemma 8. *For any $S \subseteq [m]$, if $\kappa_S < 1$, then any effort profile x such that $x_j > 0$ for all $j \in S$ cannot be optimal.*

Proof. The following proof builds on that of Lemma 6. Let y and z be optimal solutions to (9). We know that for all i,

$$\sum_{j=1}^{m} \alpha_{ji} y_j \geq \sum_{j \in S} \alpha_{ji} z_j \tag{10}$$

Let $c \triangleq \min_{j \in S} x_j / z_j$. Note that $c > 0$ because by assumption, $x_j > 0$ for all $j \in S$. It is well-defined because $z^\top \mathbf{1}_S \geq 1$ and $z \geq \mathbf{0}$, so z_j is strictly positive for some $j \in S$. By this definition, $x_j - c z_j \geq 0$ for all $j \in S$.

We'll again define another solution x' with utility at least as high as x, but with the budget

constraint slack. For all i,

$$\begin{aligned}
[\alpha^\top x]_i &= \sum_{j=1}^m \alpha_{ji} x_j \\
&= \sum_{j \notin S} \alpha_{ji} x_j + \sum_{j \in S} \alpha_{ji} x_j \\
&= \sum_{j \notin S} \alpha_{ji} x_j + \sum_{j \in S} \alpha_{ji} (x_j - c z_j) + c \sum_{j \in S} \alpha_{ji} z_j \\
&\leq \sum_{j \notin S} \alpha_{ji} x_j + \sum_{j \in S} \alpha_{ji} (x_j - c z_j) + c \sum_{j=1}^m \alpha_{ji} y_j && \text{(By (10))} \\
&= \sum_{j \notin S} \alpha_{ji} (x_j + c y_j) + \sum_{j \in S} \alpha_{ji} (x_j + c(y_j - z_j)) \\
&\triangleq [\alpha^\top x']_i,
\end{aligned}$$

where we have defined

$$x'_j \triangleq \begin{cases} x_j + c y_j & j \notin S \\ x_j + c(y_j - z_j) & z \in S \end{cases}.$$

Because $x_j - c z_j \geq 0$ for all $j \in S$, x' is a valid effort profile. Since f_i is increasing, $f_i([\alpha^\top x]_i) \leq f_i([\alpha^\top x']_i)$. However,

$$\sum_{i=1}^m x'_j = \sum_{j \notin S} x_j + c y_j + \sum_{j \in S} x_j + c(y_j - z_j) = x^\top \mathbf{1} + c(y^\top \mathbf{1} - z^\top \mathbf{1}_S) < B.$$

Thus, the budget constraint for x' is not tight, and so for any monotone mechanism, there exists a solution x'' which is strictly better than x' and x, meaning x is not optimal. □

Lemma 8 tells us which subsets of variables definitely can't be jointly incentivized. However, given a subset of variables, it doesn't a priori tell us if these variables *can* be jointly incentivized, and if so, which particular effort profiles on these variables are incentivizable. In fact, we'll show that any x^* such that $\kappa_{\mathcal{S}(x^*)} = 1$ is incentivizable.

Lemma 9. *Define*

$$\mathcal{L}(x) \triangleq \{\beta \mid A(x)\beta \leq \frac{1}{B} x^\top A(x) \beta \cdot \mathbf{1}\} \tag{11}$$

If $\kappa_{\mathcal{S}(x^*)} = 1$, then $\mathcal{L}(x^*)$ is the set of linear mechanisms that incentivize x^*, and $\mathcal{L}(x^*)$ is non-empty.

Proof. Let $S = \mathcal{S}(x^*)$. We know that for any z such that $z^\top \mathbf{1}_S \geq 1$,

$$\kappa_S \leq \kappa_S(z) \triangleq \min_{y \in \mathbb{R}^m} \; y^\top \mathbf{1} \qquad \text{s.t. } \alpha^\top y \geq \alpha_S^\top z \qquad y \geq \mathbf{0} \tag{12}$$

because we've just written (9) without allowing for optimization over z. Therefore, if $\kappa_S = 1$, then $\kappa_S(z) = 1$ for any z. We can write each constraint $[\alpha^\top y]_i \geq [\alpha_S^\top z]_i$ as

$$\begin{aligned}
[\alpha^\top y]_i \geq [\alpha_S^\top z]_i &\iff \sum_{j=1}^{m} \alpha_{ji} y_j \geq \sum_{j \in S} \alpha_{ji} z_j \\
&\iff \sum_{j=1}^{m} \alpha_{ji} f_i'([\alpha^\top x^*]_i) y_j \geq \sum_{j \in S} \alpha_{ji} f_i'([\alpha^\top x^*]_i) z_j \\
&\iff \sum_{j=1}^{m} [A(x^*)]_{ji}^\top y_j \geq \sum_{j \in S} [A(x^*)]_{ji} z_j
\end{aligned}$$

Thus, (12) is equivalent to the following optimization, where similarly to the definition of α_S, we define $A_S(x)$ to be $A(x)$ with all rows $j \notin S$ zeroed out.

$$\begin{aligned}
\kappa_S(z) = \min_{y \in \mathbb{R}^m} \; & y^\top \mathbf{1} && \text{s.t. } A(x^*)^\top y \geq A_S(x^*)^\top z \\
& && y \geq \mathbf{0}
\end{aligned} \tag{13}$$

The dual of (13) is

$$\begin{aligned}
\eta(z) \triangleq \max_{\beta \in \mathbb{R}^n} \; & \beta^\top (A_S(x^*)^\top z) && \text{s.t. } A(x^*)\beta \leq \mathbf{1} \\
& && \beta \geq \mathbf{0}
\end{aligned} \tag{14}$$

Thus, (14) has value $\eta(z) = \kappa_S(z) = 1$. Recall that

$$\mathcal{L}(x^*) = \{\beta \mid A(x^*)\beta \leq \frac{1}{B} x^{*\top} A(x^*)\beta \cdot \mathbf{1}\}.$$

Clearly, $\mathcal{L}(x^*)$ is non-empty because plugging in $z = \frac{x^*}{B}$, (14) has value $\eta(z) = 1$, meaning there exists β such that for all j,

$$\eta\left(\frac{x^*}{B}\right) = \frac{1}{B}\beta^\top (A_S(x^*)^\top x^*) = 1 \geq [A(x^*)\beta]_j \tag{15}$$

We'll show that β incentivizes the agent to invest x^* if and only if $\beta \in \mathcal{L}(x^*)$. Note that (15) is true if and only if

$$\left.\frac{\partial H}{\partial x_j}\right|_{x^*} \leq \sum_{k \in S} \frac{x_k^*}{B} \left.\frac{\partial H}{\partial x_k}\right|_{x^*}. \qquad (\forall j \in [m])$$

The right hand side is the convex combination of the partial derivatives of H with respect to each of the $k \in S$. Since this convex combination is at least as large as each partial in the combination, it must be the case that all of these partials on the right hand side are equal to one another. In other words, this is true if and only if $\left.\frac{\partial H}{\partial x_j}\right|_{x^*} = \left.\frac{\partial H}{\partial x_{j'}}\right|_{x^*}$ for all $j, j' \in S$.

By Lemma 12 in Appendix A, this is true if and only if x^* is an optimal effort profile, meaning $\mathcal{L}(x^*)$ is exactly the set of linear mechanisms that incentivize x^*. □

Thus, we've shown Theorem 3: for any target effort profile x^*, either $\kappa_{\mathcal{S}(x^*)} = 1$, in which case any $\beta \in \mathcal{L}(x^*)$ incentivizes x^*, or $\kappa_{\mathcal{S}(x^*)} < 1$, in which case no monotone mechanism incentivizes x^* by Lemma 8.

4 Optimizing other Objectives

So far, we have given a tight characterization of which effort profiles can be incentivized. Moreover, we have shown that whenever an effort profile can be incentivized, we can compute a set of linear mechanisms that do so. However, this still leaves room for the evaluator to optimize over other preferences. For instance, perhaps profiles that distribute effort among many activities are more desirable, or perhaps the evaluator has a more complex utility function over the agent's effort investment.

In this section, we consider the feasibility of such optimization subject to the constraints imposed by incentivizability. We show that optimization over effort profiles is possible in particular instances, but in general, it is computationally hard to optimize even simple objectives over incentivizable effort profiles.

Incentivizing a subset of variables. For the remainder of this section, we will assume that the evaluator has a set of designated effort variables $D \subseteq [m]$, and they want to incentivize the agent to only invest in effort variables in D. Recall that a set of actions S is incentivizable if and only if $\kappa_S = 1$, where κ_S is defined in (9). We define the set system

$$\mathcal{F}_D = \{S \subseteq D \mid \kappa_S = 1\} \tag{16}$$

By Theorem 3, $\mathcal{F}_D$ gives the sets of effort variables that can be jointly incentivized. As we will show, a consequence of our results from Section 3 is that $\mathcal{F}_D$ is downward-closed, meaning that if $S \in \mathcal{F}_D$, then $S' \in \mathcal{F}_D$ for any $S' \subseteq S$.

We begin by characterizing when it is feasible to incentivize some x such that $\mathcal{S}(x) \subseteq D$. As the following lemma shows, this can be done if and only if some individual $j \in D$ is incentivizable on its own.

Lemma 10. *It is possible to incentivize effort in a subset of a designated set of effort nodes $D \subseteq [m]$ if and only if $\max_{j \in D} \kappa_j = 1$.*

Proof. The set system $\mathcal{F}_D$ is downward closed, since $\kappa_{S \cup \{j\}} = 1$ implies $\kappa_S = 1$ for all S, j. This is because any solution to (9) for S is a solution to (9) for $S \cup \{j\}$, so $\kappa_S \geq \kappa_{S \cup \{j\}}$. Therefore, if x is such that $\mathcal{S}(x) \subseteq D$ is incentivizable, meaning $\kappa_{\mathcal{S}(x)} = 1$, then $\kappa_j = 1$ for all $j \in \mathcal{S}(x)$. If $\kappa_j < 1$ for all $j \in D$, then no x such that $\mathcal{S}(x) \subseteq D$ is incentivizable. □

Thus, there exists an incentivizable x such that $\mathcal{S}(x) \subseteq D$ if and only if there is some $j \in D$ such that the agent can be incentivized to invest all of its budget into x_j.

Objectives over effort profiles. In the remainder of this section, we prove Theorem 5. Lemma 10 shows that if the evaluator wants the agent to only invest effort into a subset D of effort variables, one solution might be to simply incentivize them to invest all of their effort into a single $j \in D$. However, this might not be a satisfactory solution — the evaluator may want the agent to engage in a diverse set of actions, or to invest at least some amount in each designated form of effort. Thus, the evaluator may have some other objective beyond simply incentivizing the designated forms of effort D.

We formalize this as follows: suppose the evaluator has some objective $g : \mathbb{R}^m \to \mathbb{R}$ over the agent's effort profile x, and wants to pick the x that maximizes g subject to the constraint that x

is incentivizable and $\mathcal{S}(x) \subseteq D$. Formally, this is

$$\arg\max_{x \in \mathbb{R}^m} \; g(x) \qquad \text{s.t.} \quad \kappa_{\mathcal{S}(x)} = 1, \quad \mathcal{S}(x) \subseteq D \tag{17}$$

To make this more tractable, we assume that g is concave, as it will in general be hard to optimize arbitrary non-concave functions. We will begin by showing that this optimization problem is feasible when $\kappa_D = 1$, or equivalently, when $D \in \mathcal{F}_D$. We will extend this to show that when $|D|$ is small, (17) can can be solved. In general, however, we will show that due to the incentivizability constraint, this is computationally hard.

First, we consider the case where $\kappa_D = 1$. Here, it is possible to find a mechanism to maximize $g(x)$ because any x in the simplex $\{x \mid \sum_{j \in D} x_j = B\}$ is incentivizable by Theorem 3. Thus, the evaluator could simply maximize g over this simplex to get some effort profile x^* and find a linear mechanism β to incentivize x^*. Extending this idea, if $\kappa_D < 1$ but $|D|$ is small, the evaluator can simply enumerate all subsets $S \subseteq D$ such that $\kappa_S = 1$, optimize g over each one separately, and pick the optimal x^* out of all these candidates.

However, in general, it is NP-hard to optimize a number of natural objectives over the set of incentivizable effort profiles if $\kappa_D < 1$. From Theorem 3, we know that incentivizable effort profiles x can be described by their support $\mathcal{S}(x)$, which must satisfy $\kappa_{\mathcal{S}(x)} = 1$. The following lemma shows that this constraint on x makes it difficult to optimize even simple functions because the family of sets $\mathcal{F}_D = \{S \subseteq D \mid \kappa_S = 1\}$ can be used to encode the set of independent sets of an arbitrary graph. Using this fact, we can show that there exist concave objectives g that are NP-hard to optimize subject to the incentivizability constraint.

Lemma 11. *Given a graph $G = (V, E)$, there exists an effort graph G' and a set of designated effort nodes D such that $S \subseteq D$ is an independent set of G if and only if $\kappa_S = 1$ in G'.*

Proof. We construct a designated effort node for each $v \in V$, so $D = V$. We also construct an undesirable effort node for each $e \in E$, so the total number of effort nodes is $m = |V| + |E|$. For ease of indexing, we'll refer to the designated effort nodes as x_v for $v \in V$ and the remaining effort nodes as x_e for $e \in E$.

We construct a feature F_v for each vertex $v \in V$. Then, let $\alpha_{v,v} = 3$ for all $v \in V$ and $\alpha_{e,v} = 2$ for all $v \in V$. For each $e \in E$, this creates the gadget shown in Figure 3.

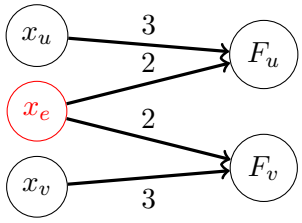


Figure 3: Gadget to encode independent sets

First, we'll show that if $(u, v) \in E$, then any $S \subseteq D$ containing both u and v has $\kappa_S < 1$. Recall the definition of κ_S in (9). Consider the solution with $z_u = z_v = \frac{1}{2}$ and $y_e = \frac{2}{3}$. This is feasible, so $\kappa_S \leq \frac{2}{3} < 1$. By the contrapositive, if $\kappa_S = 1$ (meaning S is incentivizable), S cannot contain any u, v such that $(u, v) \in E$, meaning S forms an independent set in G.

To show the other direction, consider any independent set S in G. By construction, $S \subseteq D$ because $D = V$. Then, in the optimal solution (y, z) to (9), we will show that $y_u = z_u$ for all $u \in S$,

meaning $\kappa_S = y^\top \mathbf{1} \geq 1$. To do so, consider the constraint $[\alpha^\top y]_u \geq [\alpha_S^\top z]_u$ for any $u \in S$. This is simply $3y_u + 2\sum_{e=(u,v)\in E} y_e \geq 3z_u$. Because S is an independent set, $z_v = 0$ for any v such that $(u, v) \in E$, so this is the only constraint in which any such y_e appears. Therefore, it is strictly optimal to choose $y_u = z_u$ and $y_e = 0$ for all $e = (u, v) \in E$. As a result, $y_u = z_u$ for all $u \in S$, meaning $\kappa_S \geq \sum_{u\in S} y_u = \sum_{u\in S} z_u \geq 1$ by the constraint $z^\top \mathbf{1}_S \geq 1$. □

Thus, if the evaluator wants to find an incentivizable effort profile x such that $\mathcal{S}(x) \subseteq D$ (the agent only invests in designated forms of effort), maximizing an objective like $g(x) = \|x\|_0$ (the number of non-zero effort variables) is NP-hard, due to a reduction from the maximum independent set problem. Note that $\|x\|_0$ is concave for nonnegative x.

Moreover, other simple and natural objectives are hard to optimize as well. Using a construction similar to the one in Figure 3, we can create effort graphs with a set of designated effort nodes D in which $S \subseteq D$ is incentivizable if and only if $|S| \leq k$, meaning $\|x\|_0 \leq k$. This is known to make optimizing even simple quadratic functions (e.g. $\|\mathcal{A}x - y\|_2$ for some matrix $\mathcal{A}$ and vector y) NP-hard [32]. In general, then, it is difficult to find the optimal agent effort profile subject to the incentivizability constraint.

5 The Structure of the Space of Linear Mechanisms

Thus far, we have seen how to construct linear mechanisms that incentivize particular effort profiles, finding that the mechanisms that do so form a polytope. Suppose that the evaluator doesn't have a particular effort profile that they want to incentivize, but instead wants the agent to only invest effort in a subset of intended effort nodes $D \subseteq [m]$. Generalizing the definition of $\mathcal{L}(x^*)$ as the set of linear mechanisms incentivizing x^*, we define $\mathcal{L}(D)$ to be the set of linear mechanisms incentivizing any x such that $\mathcal{S}(x) \subseteq D$.[3] In the remainder of this section, we give structural results characterizing $\mathcal{L}(D)$, showing that in general it can be highly nonconvex, indicating the richness of the solution space of this problem.

In the simplest case where $|D| = 1$, meaning the evaluator wants to incentivize a single effort variable, we know by (6) that $\mathcal{L}(D)$ is simply a polytope. This makes it possible for the evaluator to completely characterize $\mathcal{L}(D)$ and even maximize any concave objective over it.

However, in general, $\mathcal{L}(D)$ can display nonconvexities in several ways. Figure 3 gives an example such that if the evaluator only wants to incentivize x_u and x_v, then $\mathcal{L}(D) = \{\beta \mid \|\beta\|_0 = 1\}$, meaning β has exactly one nonzero entry. This can be generalized to an example where $\mathcal{L}(D) = \{\beta \mid \|\beta\|_0 \leq k\}$ for any k, which amounts to a nonconvex sparsity constraint.

This form of nonconvexity arises because we're considering mechanisms that incentivize x such that $\mathcal{S}(x) \subseteq D$. In particular, if S and S' are disjoint subsets of D, then we wouldn't necessarily expect the union of $\mathcal{L}(S)$ and $\mathcal{L}(S')$ to be convex. However, we might hope that if each $\mathcal{L}(S)$ for $S \subseteq D$ is convex or can be written as the union of convex sets, then $\mathcal{L}(D)$ could also be written as the union of convex sets.

Unfortunately, this isn't the case. Let $\mathcal{L}^*(D)$ be the set of mechanisms incentivizing x such such that $\mathcal{S}(x) = D$ (as opposed to $\mathcal{S}(x) \subseteq D$). $\mathcal{L}^*(D)$ may still be nonconvex, depending on the particular effort conversion functions $f(\cdot)$. Consider the effort graph shown in Figure 4 with $B = 1$, $f_1(y) = f_2(y) = 1 - e^{-y}$ and $f_3(y) = 1 - e^{-2y}$. Let $D = \{1, 3\}$. To incentivize $x_1 > 0$ and $x_3 > 0$ simultaneously with $x_2 = x_4 = 0$, it must be the case that

$$\left.\frac{\partial H}{\partial x_1}\right|_x = \beta_1 f_1'(x_1) = \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) = \left.\frac{\partial H}{\partial x_3}\right|_x.$$

[3] With this notation, we could write $\mathcal{L}_j$ as defined in Section 3 as $\mathcal{L}(\{j\})$.

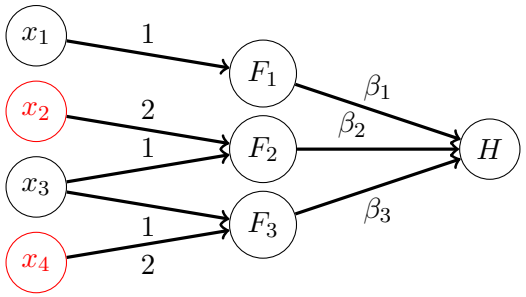


Figure 4: Non-convexity of $\mathcal{L}^*(D)$

To incentivize $x_2 = x_4 = 0$, we must also have

$$\left.\frac{\partial H}{\partial x_2}\right|_x = 2\beta_2 f_2'(x_3) \leq \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) = \left.\frac{\partial H}{\partial x_3}\right|_x$$
$$\left.\frac{\partial H}{\partial x_4}\right|_x = 2\beta_3 f_3'(x_3) \leq \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) = \left.\frac{\partial H}{\partial x_3}\right|_x$$

This is only possible if $\beta_2 f_2'(x_3) = \beta_3 f_3'(x_3)$, meaning β incentivizes x such that $\mathcal{S}(x) = \{1, 3\}$ if and only if

$$\beta_1 f_1'(x_1) = \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) \tag{18}$$
$$\beta_2 f_2'(x_3) = \beta_3 f_3'(x_3) \tag{19}$$

Combining (18) and (19), we get $\beta_1 f_1'(x_1) = 2\beta_2 f_2'(x_3)$, implying

$$\begin{aligned}\beta_1 f_1'(x_1) &= 2\beta_2 f_2'(x_3) \\ \beta_1' e^{-x_1} &= 2\beta_2 e^{-x_3} \qquad (20) \\ \beta_2 &= \frac{\beta_1 e^{x_3 - x_1}}{2} \qquad (21)\end{aligned}$$

Similarly, we can derive

$$\beta_3 = \frac{\beta_1 e^{2x_3 - x_1}}{4} \tag{22}$$

We'll show non-convexity by giving two linear mechanisms β and β' that both incentivize an x such that $\mathcal{S}(x) = \{1, 3\}$, but $\beta'' = \frac{1}{2}(\beta + \beta')$ does not incentivize such an x.

Let β and β' incentivize $x = [1/3 \ 0 \ 2/3 \ 0]^\top$ and $x' = [2/3 \ 0 \ 1/3 \ 0]^\top$ respectively. Without loss of generality, we can set $\beta_1 = \beta_1' = 1$. Using (21) and (22), we get

$$\beta = \begin{bmatrix} 1 & \frac{e^{1/3}}{2} & \frac{e}{4} \end{bmatrix}^\top \qquad\qquad \beta' = \begin{bmatrix} 1 & \frac{e^{-1/3}}{2} & \frac{1}{4} \end{bmatrix}^\top$$

Then, let $\beta'' = \frac{1}{2}(\beta + \beta')$. If β'' incentivizes x^* such that $\mathcal{S}(x^*) = \{1, 3\}$, then by (19), it must be the case that

$$\begin{aligned}\beta_2'' f_2'(x_3^*) &= \beta_3'' f_3'(x_3^*) \\ \beta_2'' e^{-x_3^*} &= 2\beta_3'' e^{-2x_3^*} \\ e^{x_3^*} &= \frac{2\beta_3''}{\beta_2''} \\ x_3^* &= \log\left(\frac{e+1}{e^{1/3} + e^{-1/3}}\right) \approx 0.566\end{aligned}$$

On the other hand, by (20), we must also have

$$\begin{aligned}
\beta_1'' f_1'(x_1^*) &= 2\beta_2'' e^{-x_3^*} \\
e^{-x_1^*} &= \frac{e^{1/3} + e^{-1/3}}{2} \exp\left(-\log\left(\frac{e+1}{e^{1/3}+e^{-1/3}}\right)\right) \\
e^{-x_1^*} &= \frac{e^{1/3} + e^{-1/3}}{2} \cdot \frac{e^{1/3} + e^{-1/3}}{e+1} \\
x_1^* &= -\log\left(\frac{(e^{1/3}+e^{-1/3})^2}{2(e+1)}\right) \approx 0.511
\end{aligned}$$

Such a solution would fail to respect the budget constraint (since $x_1^* + x_3^* > 1 = B$), meaning β'' cannot incentivize x^* such that $\mathcal{S}(x^*) = \{1, 3\}$. In fact, the above analysis shows that for any x^* incentivized by β'', $\mathcal{S}(x^*)$ must include either 2 or 4 because β'' incentivizes neither $x_1^* = 1$ nor $x_3^* = 1$, meaning the only way to use the entire budget is to set $x_2^* > 0$ or $x_4^* > 0$. Thus, despite the fact that both β and β' incentivize effort profiles with support $\{1, 3\}$, a convex combination of them does not. As a result, the set of linear mechanisms incentivizing a subset of effort variables may in general exhibit complex structures that don't lend themselves to simple characterization.

We visualize this nonconvexity in Figure 5, where for clarity we modify the effort graph in Figure 4 by setting $\alpha_{22} = 0$. The yellow region corresponds to (β_2, β_3) values such that $\beta = (1 \ \beta_2 \ \beta_3)^\top$ incentivizes x such that $\mathcal{S}(x) = \{1, 3\}$. Note that the upper left edge of this region is slightly curved, producing the non-convexity. As a result, the set of linear mechanisms incentivizing a subset of effort variables may in general exhibit complex structures that don't lend themselves to simple characterization.

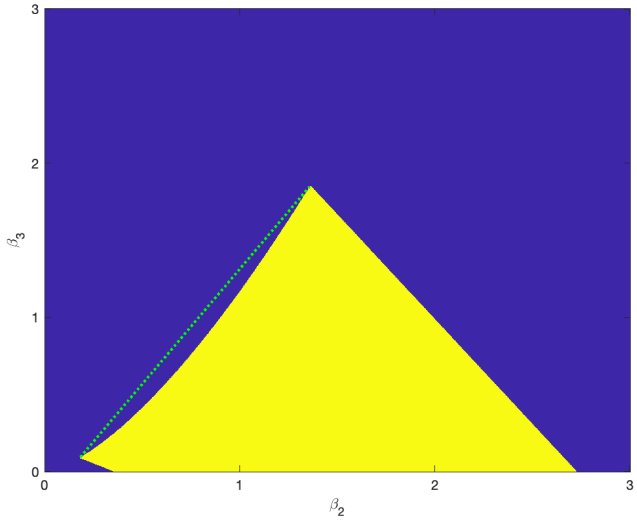


Figure 5: Non-convexity in (β_2, β_3) pairs

Implications for optimization. The complexity of $\mathcal{L}(D)$ has immediate hardness implications for optimizing objectives over the space of linear mechanisms. For example, mechanisms that distribute weight on multiple features may be preferable because in practice, measuring multiple distinct features may lead to less noisy evaluations. We might also consider the case where the

evaluator has historical data $\mathcal{A} \in \mathbb{R}^{r \times n}$ and $y \in \mathbb{R}^r$, where each row of $\mathcal{A}$ contains the features F of some individual and each entry of y contains their measured outcome of some sort. Then, in the absence of strategic considerations, the evaluator could just choose β that minimizes squared error $\|\mathcal{A}\beta - y\|_2$ between the scores given by the mechanism and the outcomes y in the dataset. As noted above, there are examples for which $\mathcal{L}(D) = \{\beta \mid \|\beta\|_0 \leq k\}$, which is known to make minimizing squared error NP-hard [32]. However, in the special case when $D = \{j\}$ (there's only one effort node the evaluator wants to incentivize), then if $\kappa_j = 1$, the set of linear mechanisms incentivizing $x^* = B \cdot e_j$ is just the convex polytope $\mathcal{L}_j$ defined in (6). Thus, it is possible to maximize any concave objective over this set.

6 Conclusion

Strategic behavior is a major challenge in designing simple and transparent evaluation mechanisms. In this work, we have developed a model in which strategic behavior can be directed toward specified forms of effort through appropriate designs.

Our results leave open a number of interesting questions. All of our analysis has been for the case in which an evaluator designs a mechanism optimized for the parameters of a single agent (or for a group of agents who all have the same parameters). Extending this reasoning to consider the incentives of a heterogeneous group of agents, where the parameters differ across members of the group, is a natural further direction. In addition, we have assumed throughout that agents behave rationally, in that they perfectly optimize their allocation of effort. But it would also be interesting to consider agents with potential biases that reflect human behavioral principles, resulting in sub-optimal behavior that follows certain structured properties. Finally, although we have shown that linear mechanisms suffice whenever a monotone mechanism can incentivize intended behavior, if the output of the mechanisms is constrained in some way (e.g. binary classification), it is an open question to determine what types of mechanisms are appropriate.

Acknowledgments. Acknowledgments. We thank Rediet Abebe, Solon Barocas, Larry Blume, Fernando Delgado, Karen Levy, and Helen Nissenbaum for their useful feedback and suggestions. Thanks to Tal Alon, Magdalen Dobson, and Jamie Tucker-Foltz for fixing an error in Lemma 11 that appeared in an earlier version of this work. This work has been supported in part by a Simons Investigator Award, a grant from the MacArthur Foundation, graduate fellowships from the National Science Foundation and Microsoft, and NSF grants CCF1740822 and SES-1741441.

References

[1] Kenneth J Arrow. Uncertainty and the welfare economics of medical care. *American Economic Review*, 1963.

[2] Kenneth J Arrow. The economics of moral hazard: further comment. *American Economic Review*, 58, 1968.

[3] Jane R Bambauer and Tal Zarsky. The algorithm game. 2018.

[4] Jöran Beel, Bela Gipp, and Erik Wilde. Academic search engine optimization (ASEO) optimizing scholarly literature for Google Scholar & co. *Journal of scholarly publishing*, 41(2):176–190, 2009.

[5] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555. ACM, 2011.

[6] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–63, 2015.

[7] Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26. ACM, 2018.

[8] Steven NS Cheung. *The theory of share tenancy*. Arcadia Press Ltd., 1969.

[9] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

[10] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. Truthful linear regression. In *Conf. Learning Theory*, 2015.

[11] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.

[12] Harold Davis. *Search engine optimization*. "O'Reilly Media, Inc.", 2006.

[13] Ofer Dekel, Felix Fischer, and Ariel D Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

[14] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70. ACM, 2018.

[15] Virginia Eubanks. Automating Bias. *Scientific American*, 319(5):68–71, 2018.

[16] Dean Foust and Aaron Pressman. Credit scores: Not-so-magic numbers. *Business Week*, 7, 2008.

[17] Roland G Fryer Jr and Glenn C Loury. Valuing diversity. *Journal of Political Economy*, 121(4):747–774, 2013.

[18] Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. *Econometrica*, pages 7–45, 1983.

[19] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122. ACM, 2016.

[20] Benjamin E Hermalin and Michael L Katz. Moral hazard and verifiability: The effects of renegotiation in agency. *Econometrica: Journal of the Econometric Society*, pages 1735–1753, 1991.

[21] Bengt Holmstrom and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica: Journal of the Econometric Society*, pages 303–328, 1987.

[22] Bengt Holmstrom and Paul Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, 7:24, 1991.

[23] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proc. World Wide Web Conference*, 2017.

[24] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

[25] Michael C Jensen and William H Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360, 1976.

[26] Steven Kerr. On the folly of rewarding A, while hoping for B. *Academy of Management journal*, 18, 1975.

[27] Daniel Koretz, Robert Linn, Stephen Dunbar, and Lorrie Shepard. The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. In *American Educational Research Association and the National Council on Measurement in Education*, 1991.

[28] Daniel M Koretz. *Measuring up*. Harvard University Press, 2008.

[29] Jean-Jacques Laffont and David Martimort. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, 2009.

[30] R Preston McAfee and John McMillan. Bidding for contracts: a principal-agent analysis. *The RAND Journal of Economics*, pages 326–338, 1986.

[31] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

[32] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 1995.

[33] Mark V Pauly. The economics of moral hazard: comment. *American Economic Review*, 1968.

[34] Stephen A Ross. The economic theory of agency: The principal's problem. *American Economic Review*, 63, 1973.

[35] Michael Spence. Job market signaling. *Quarterly Journal of Economics*, 87:355–374, 1973.

[36] Joseph E Stiglitz. Incentives and risk sharing in sharecropping. *The Review of Economic Studies*, 41, 1974.

[37] Tal Z Zarsky. Law and online social networks: Mapping the challenges and promises of user-generated information flows. *Fordham Intell. Prop. Media & Ent. LJ*, 18:741, 2007.

[38] Malte Ziewitz. Rethinking gaming: The ethical work of optimization in web search engines. Forthcoming.

A Characterizing the Agent's Response to a Linear Mechanism.

In this section, we'll characterize how a rational agent best-responds to a linear mechanism. Its utility is $H = \beta^\top F$, and therefore we can rewrite the optimization problem (2) with $M(F) = \beta^\top F$, which yields

$$\begin{aligned} \max_{x \in \mathbb{R}^m} \quad & \sum_{i=1}^{n} \beta_i f_i([\alpha^\top x]_i) \\ \text{s.t. } & x \geq \mathbf{0} \\ & \sum_{j=1}^{m} x_j \leq B \end{aligned} \tag{23}$$

Note that this is a concave maximization since each f_i is weakly concave and $[\alpha^\top x]_i$ is linear in x. The Lagrangian is then

$$\mathcal{L}(x, \lambda) = \sum_{i=1}^{n} \beta_i f_i \left([\alpha^\top x]_i \right) + \lambda_0 \left(B - \sum_{j=1}^{m} x_j \right) + \sum_{j=1}^{m} \lambda_j x_j.$$

By the Karush-Kuhn-Tucker conditions, since (23) is convex, a solution x^* is optimal if and only if $\nabla_x \mathcal{L}(x^*, \lambda^*) = \mathbf{0}$, so for each $j \in [m]$,

$$\sum_{i=1}^{n} \alpha_{ji} \beta_i f_i' \left([\alpha^\top x^*]_i \right) - \lambda_0^* + \lambda_j^* = 0.$$

Note that we can write this as

$$\lambda_0^* = \left. \frac{\partial H}{\partial x_j} \right|_{x^*} + \lambda_j^*.$$

By complementary slackness, $\lambda_j^* > 0 \implies x_j^* = 0$. Therefore, it follows that at optimality, the gradients with respect to all nonzero effort components are λ_0^*. Furthermore, the gradients with respect to all effort components are at most λ_0^* since $\lambda_j^* \geq 0$ by definition. This proves the following lemma.

Lemma 12. *For any $x \in \mathbb{R}^m$ such that $x \geq \mathbf{0}$, x is an optimal solution to (23) if and only if the following conditions hold*

1. $\sum_{j=1}^{m} x_j = B$
2. *For all j, j' such that $x_j > 0$ and $x_{j'} > 0$,*

$$\left. \frac{\partial H}{\partial x_j} \right|_x = \left. \frac{\partial H}{\partial x_{j'}} \right|_x$$

3. *For all j such that $x_j > 0$ and for all j',*

$$\left. \frac{\partial H}{\partial x_j} \right|_x \geq \left. \frac{\partial H}{\partial x_{j'}} \right|_x$$

Proof. Choose $\lambda_0^* = \frac{\partial H}{\partial x_j}\Big|_x$ for any j such that $x_j > 0$. Choose $\lambda_j^* = \lambda_0^* - \frac{\partial H}{\partial x_j}\Big|_x$ for all j. Then, (x, λ^*) satisfies stationarity (since $\nabla_x \mathcal{L}(x, \lambda^*) = \mathbf{0}$), primal and dual feasibility by definition, and complementary slackness (since $B - \sum_{j=1}^m x_j = 0$). Therefore, x is an optimal solution to (23).

To show the other direction, note that $\max_j \frac{\partial H}{\partial x_j} > 0$ because each $f_i(\cdot)$ is strictly increasing and there is some nonzero β_i. Therefore, $\lambda_0 > 0$, and by complementary slackness, every optimal solution must satisfy $\sum_{j=1}^m x_j = B$. □