Regulating algorithmic filtering on social media

Anonymous Author(s)
Affiliation
Address
email

Abstract

By filtering the content that users see, social media platforms have the ability to influence users' perceptions and decisions, from their dining choices to their voting preferences. This influence has drawn scrutiny, with many calling for regulations on filtering algorithms. In this work, we examine two questions. First, how does the filtered content affect a user's learning and decision-making and is there a principled way to regulate this effect? Second, does regulating algorithmic filtering necessarily impose a large performance cost on the platform? We approach these questions by studying a system with three actors: a platform, a user, and an auditor. We propose a method such that, given a regulation, an auditor can test whether that regulation is met with only black-box access to the filtering algorithm. We then turn to the platform's perspective. The platform's goal is to maximize an objective function R while meeting regulation. We find that there are conditions under which the cost of regulation is low and, interestingly, content diversity plays a key role in the platform's ability to achieve high R while meeting regulation.

1 Introduction

Social media platforms have become one of the primary sources of information for people everywhere. By selecting the content that each user sees, platforms play a major role in users' perceptions and behaviors. They may, for instance, influence a user's dining or shopping habits [51, 31], inform their social or political views [19, 56], or guide their employment or housing search [16, 4].

The mechanism by which platforms select the content—such as articles, videos, and advertisements—that appears on a user's feed is known as *algorithmic filtering*. Its importance derives from the fact that, as a user browses through their content, they implicitly learn a model of their world by regressing on the data. With enough evidence, they gain confidence in their learned model, subsequently using it to make decisions. For instance, after viewing many doctorates being hired, a user may become convinced that a good job necessitates a PhD, even if it does not, and begin pursuing one.

Due to the ability of social media to influence the user decision-making, filtering algorithms have drawn scrutiny in recent years, and there are growing calls for regulation. However, designing and enforcing such regulations is challenging. For one, regulations must balance being general enough to endure with time while being precise enough to carry out in practice. In addition, there are concerns that regulations may hurt innovation or impose large performance costs on platforms.

In this work, we consider two questions. First, *how does the filtered content affect a user's decision-making and is there a principled way to regulate this effect*? Second, *does regulating algorithmic filtering necessarily impose a large performance cost on the platform*?

We tackle these questions by studying a system with three actors: a platform, a user, and an auditor. We analyze the effect of content on the user's perceptions and behaviors. The auditor's goal is to test whether the platform complies with a given regulation. The platform's goal is to filter content to maximize its reward while meeting regulation. Our main contributions are as follows.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Regulatory procedure. In Section 4, we propose a general procedure such that, *given a regulation*, an auditor can test whether that regulation is met by the platform. In this work, we assume that the regulation is given, and we restrict our attention to regulations that can be expressed in a counterfactual or comparative form, as described in Section 4.2. In Section 4.3, we study the properties of this regulatory test and show that it has strong statistical guarantees. In addition, the regulatory test requires only black-box access to the filtering algorithm, and its tunable parameter is easy to interpret.

Effect of content on users. One of the benefits of the proposed procedure is that it uses insights from statistical learning theory to carry out the regulatory test. In particular, determining the effect of a filtering algorithm on users is challenging because each user responds to their content differently, and measuring each user's response may be costly or infeasible. In Section 4.1, we show that it suffices to focus on a hypothetical user whose learning behavior is given by the minimum-variance unbiased estimator (MVUE). Intuitively, the MVUE corresponds to the "most gullible user" and if the platform passes regulation for the MVUE, then it passes regulation for all other users.

Cost of regulation. In Section 5.1, we study how the regulatory test affects the platform's ability to maximize its reward R. As an example, R could measure the predicted revenue accrued from clicks on advertisements. Interestingly, we find that there is not necessarily a strong regulation-reward trade-off, and we state conditions under which the cost of regulation is low.

Content diversity. In Section 5.2, we investigate how the regulatory test affects the filtering algorithm. We find that, under regulation, the platform is incentivized to add doses of content diversity in order to meet regulation, i.e., adding content diversity often helps the platform increase its reward. Because content diversity is not a part of the regulatory test by design, this result is surprising and suggests that content diversity plays a key role in aligning the interests of regulators and platforms.

2 Background & related work

Algorithmic filtering [25, 13] has the potential to greatly improve the user experience, but it can also yield unwanted side effects, like the spread of fake news [42, 20, 22], over-representation of polarizing opinions due to comment ranking [52], amplification of echo chambers due to filter bubbles [30, 48], or advertising of products based on discriminatory judgments about user interests [53, 54, 39]. The increasing regularity of these phenomena have led to rising public scrutiny.

In response, there have been calls to regulate social media platforms and their filtering practices. One approach that platforms have taken is to self-regulate [43, 2]. For example, Facebook has established an internal "Supreme Court" that reviews the company's decisions [47], Twitter has banned accounts associated with the spread of conspiracy theories [8], YouTube has removed videos it views as encouraging violence [55], and so on. Self-regulatory practices offer various benefits over governmental regulations, such as the ability to adapt quickly to a changing social media ecosystem and the comparatively greater information access afforded to internal auditors.

However, many argue that governmental regulations are necessary to ensure that an audit is executed by an independent body. Several regulations exist, such as the EU's General Data Protection Regulation [49] and Germany's Network Enforcement Act [17]. There are also ongoing efforts, such as the push to review Section 230 of the U.S. Communications Decency Act [1].

Designing such regulations remains a challenging problem, in part due to the number of stakeholders. In particular, there are various legal and social obstacles facing social media regulations [40, 15, 9, 46], including concerns that regulations might damage free speech or public discourse; violate personal rights or privacy; transfer agency away from users to technology companies or governmental bodies; draw subjective lines between acceptable and unacceptable behavior; or set precedents that are difficult to reverse. Moreover, in light of the thriving exchange of goods between users, platforms, advertisers and influencers that is facilitated by social media, many fear that regulations may hurt innovation, lead to worse personalization, or block revenue sources [28].

Current efforts to regulate content moderation generally focus on specific issues, such as whether content is inappropriate (e.g., posts that contain hate speech or bullying [7, 23]); discriminatory (e.g., race-based advertising [5, 53, 54, 39]); divisive (e.g., comment ranking algorithms that favor polarizing comments [52]); insulating (e.g., filter bubbles [30]), or misleading (e.g., fake news [42, 20, 22]). These works generally use one of the following strategies: increasing content diversity (e.g., adding heterogeneity to recommendations [14, 34]); drawing a line in the sand (e.g., determining

whether discrimination has occurred by thresholding the difference between two proportions [21]); or finding the origin of the content (e.g., reducing fake news by whitelisting news sources [9]).

As these regulatory proposals gain popularity and are eventually adopted, the question of how to implement and enforce such regulations arises. In particular, technology regulations are often designed to have a certain level of generality (though not so much so that they cannot be meaningfully enforced). Such generality allows a regulation to endure with time by adapting to changing circumstances, but it is also makes it more difficult to enforce. Our aim is to provide a general procedure such that, given a regulation, an auditor can test whether that regulation is met by a platform's filtering algorithm. Because this procedure only requires black-box access to the filtering algorithm, it is not restricted to a certain class of filtering algorithms. We focus our analysis on regulations that can be expressed in a certain form, as detailed in Section 4.

We also consider how a regulation affects the platform's ability to maximize its reward—an objective function that the platform uses to choose between feeds when filtering—and the corresponding feed that the platform filters for a user. Our formulation is an instance of constrained optimization and bears resemblance to robust optimization [10, 59]. For instance, our definition of the cost of regulation mirrors the "price" of robustness studied in other such works [12, 11]. Similarly, regulation-reward trade-off that we discuss echoes trade-offs that appears in other optimization problems (e.g., revenue maximization or risk minimization) in which there are fairness [11, 29, 37] and privacy [18, 35] constraints. Our findings that there are conditions under which content diversity aligns the interests of regulators and platforms adds to the conversation on the effect of presenting content with different viewpoints on social media. For instance, [41] finds that users respond well when presented with multiple political viewpoints—which Levy terms counter-attitudinal content—even ones from an opposing political party. Increasing content diversity is also at the heart of other methods, including those for bursting filter bubbles [14] or improving comment ranking [33].

It is worth a remark that our analysis has parallels with the field of differential privacy [26, 27] in that it compares outcomes under different interventions [58]. However it differs in the techniques required to study and test the outcome differences. Our work also touches on aspects of social learning [24, 6, 3, 45] in that we study how information affects the beliefs of individuals.

3 Problem statement

Consider a system with three agents: a platform, a user, and an auditor. In this system, the platform selects the content that appears on the user's feed while the auditor tests whether the platform complies with a given regulation. This setup is formalized below.

3.1 Setup

Feed. At each time step $t \in \mathbb{N}_{>0}$, the platform filters the user's feed $Z^{(t)} = (\mathbf{z}_1^{(t)}, \mathbf{z}_2^{(t)}, \ldots, \mathbf{z}_m^{(t)}) \subset \mathcal{Z}$. Each feed consists of m pieces of content, where $\mathbf{z}_i^{(t)} \in \mathbb{R}^n$ denotes the i-th piece of content. [1] The platform's filtering algorithm denoted by $\mathcal{F}$, and $Z^{(t)}$ denotes its output at time t.

Regulation. The auditor's goal is to test whether the platform meets regulation. In this work, we assume that the regulation is given. We also assume that the auditor has black-box access to the platform's filtering algorithm $\mathcal{F}$.[2] In other words, the auditor can run $\mathcal{F}$ on a set of inputs and observe its outputs but does not require additional details about the algorithm.

Platform. The platform's goal is to maximize its reward while passing regulation. Recall that $\mathcal{Z}$ denotes the set of all possible feeds. Passing regulation at time t is equivalent to restricting the platform's choice of feeds to a subset $\mathcal{Z}^{(t)} \subset \mathcal{Z}$, which we call the *feasible set* at time t. Intuitively, if there is no regulation, then $\mathcal{Z}^{(t)} = \mathcal{Z}$, and the stricter the regulation, the smaller $\mathcal{Z}^{(t)}$.

[1] Note that adopting the convention that each piece of content is described by an n-dimensional vector is not an additional assumption. Each piece of content must be described by some feature vector (or embedding). If not, then the platform would not be able to algorithmically filter content to begin with.

[2] That the auditor has black-box access is not a strong assumption. Auditors are generally given access to internal information in order to check compliance with a regulation. In addition, the assumption above only requires black-box access to the algorithm and does not require information about the platform's users.

Let $R : \mathcal{Z} \to \mathbb{R}$ denote the platform's objective function.[3] Then, the platform's goal is to filter a feed $Z^{(t)}$ that maximizes its reward R while passing regulation at time t, which can be summarized as:

$$Z^{(t)} \in \arg\max_{Z \in \mathcal{Z}^{(t)}} R(Z). \tag{1}$$

We leave R unspecified, which means that our analysis holds for any choice for objective function, unless otherwise stated. As a simple example, R could measure the predicted user engagement (e.g., the predicted number of clicks). Generally speaking, the smaller the feasible set, the lower the reward. To quantify this effect, the cost of regulation is defined as follows.

Definition 1. *The* cost of regulation *at time t is:* $C^{(t)} = \max_{Z \in \mathcal{Z}} R(Z) - \max_{Z \in \mathcal{Z}^{(t)}} R(Z)$.

Intuitively, a high cost of regulation implies that there is a strong regulation-reward trade-off.

3.2 Objectives

The objectives of this work are two-fold. One objective is from the auditor's perspective, and the second is from the platform's perspective.

Regulatory test. The first objective of this work is to design a general procedure such that, *given a regulation, an auditor can test whether that regulation is met*. We restrict our analysis to regulations that can be expressed in a comparative or counterfactual form, as formalized in Section 4.

Cost of regulation. The second objective of this work is to study how the regulation affects the reward that the platform achieves. Specifically, we *investigate the cost of regulation and whether there is a regulation-reward trade-off*.

We tackle the first objective in Section 4 and the second objective in Section 5.

4 Regulatory procedure

In this section, we propose a procedure such that, given a regulation on algorithmic filtering, an auditor can test whether the regulation is met. In order to ensure that this test has the desired effect, we ground the procedure using results from statistical learning theory. In Section 4.1, we characterize the effect of content on a user. In Section 4.2, we use this analysis to design a regulatory test. In Section 4.3, we study the regulatory test's properties. All proofs can be found in the Appendix.

4.1 User learning and decision-making

In this section, we show that characterizing the effect of content on a user's learning and decision-making behavior for the purposes of regulation can be achieved by focusing on a specific estimator, known as the minimum variance unbiased estimator (MVUE). This result is powerful because it implies that, in order to test the impact of a filtering algorithm on its users, the auditor need not measure the decision-making behavior of each individual user, which may be infeasible or costly. Instead, the auditor can run its tests on the MVUE.

Decision-making pipeline. Decisions can be reduced to a simple pipeline: (a) an individual observes information; (b) that information informs their beliefs, and (c) based on those beliefs, they make a decision. In this work, we consider the setting in which a portion of the information that an individual consumes is provided by their social media feed. For example, how an individual votes or the products they buy are decisions that are affected by the content they see on social media.

User learning. At each time step, the user observes their feed and updates their beliefs. Formally, let $\mathcal{P} = \{p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r\}$ denote the set of possible beliefs, where the user assumes $\mathbf{z}_i^{(t)}$ are drawn i.i.d. from $p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}^{(t)})$, and $\boldsymbol{\theta}^{(t)}$ is the (unknown) parameter that the user seeks to learn. Let $\mathcal{L} : \mathcal{Z} \to \Theta$ represent the user's learning behavior such that $\hat{\boldsymbol{\theta}}^{(t)} = \mathcal{L}(Z^{(t)})$ denotes the user's belief at time t. Intuitively, a user's belief is a product of the information to which they have been exposed,

[3] There are settings in which R is time-varying, e.g., when a platform's sources of revenue change with time. Making R time-varying does not change our analysis or findings. Therefore, we leave R static for simplicity.

and how the user incorporates this information into their belief is captured by $\mathcal{L}$. This formulation does not prohibit the user's beliefs from being influenced by information other than their feed.[4]

Minimum variance unbiased estimator (MVUE). In statistical learning theory, $\mathcal{L}$ can be thought of an an *estimator*. In practice, the user's learning behavior $\mathcal{L}$ is not only unknown, but also difficult (or even unethical) to measure. The following result states that it suffices to study the MVUE. Intuitively, the MVUE as an estimator that learns "well" from the data: the MVUE is, on average, centered at the true unknown parameter that it seeks to estimate (i.e., it is unbiased), and the MVUE generates estimates with higher certainty (i.e., lower variance) than other unbiased estimators.

Proposition 1. *Suppose that the user is faced with two options A_0 and A_1. Let $v_0(\boldsymbol{\theta})$ denote the value the user associates with option A_0 under the belief $\boldsymbol{\theta} \in \Theta$ and $v_1(\boldsymbol{\theta})$ denote the same for option A_1. Let there be two hypotheses at time step t:*

$$G_0 : v_0(\boldsymbol{\theta}^{(t)}) \geq v_1(\boldsymbol{\theta}^{(t)}) \qquad G_1 : v_1(\boldsymbol{\theta}^{(t)}) > v_0(\boldsymbol{\theta}^{(t)})$$

Let $G^{(t)} \in \{G_0, G_1\}$ denote the true (unknown) hypothesis. Let ρ denote the allowable false positive rate. Then, if the uniformly most powerful test exists, it is defined by the following decision rule: reject G_0 when the minimum-variance unbiased estimate $\tilde{\boldsymbol{\theta}}^{(t)}$ of $\boldsymbol{\theta}^{(t)}$ satisfies $v_1(\tilde{\boldsymbol{\theta}}^{(t)}) - v_0(\tilde{\boldsymbol{\theta}}^{(t)}) > \eta_\rho$ where $\mathbb{P}(v_1(\tilde{\boldsymbol{\theta}}^{(t)}) - v_0(\tilde{\boldsymbol{\theta}}^{(t)}) > \eta_\rho | G^{(t)} = G_0) = \rho$.

Most gullible user. This results states that, if one wishes to study the impact of content on a user's decision-making behavior, one can focus on a hypothetical user for whom $\mathcal{L}$ is the MVUE because, if the platform meets regulation for the MVUE, then it meets regulation for all other $\mathcal{L}$. Intuitively, the MVUE corresponds to the "most gullible user": the user whose decisions are most affected by their content. Therefore, *if the platform meets regulation for the most gullible user, then it meets regulation for all other users*. This insight allows the auditor to operate without enumerating through all users and deducing their learning behavior, which may be costly or infeasible.

4.2 Regulatory test

In the previous section, we saw that it suffices to study how content affects a hypothetical user whose learning behavior is given by the MVUE. In this section, we use this insight to construct a procedure that the auditor can use to test the platform's compliance with a given regulation.

Form of regulation. As discussed in Section 3, we restrict our analysis to regulations that can be expressed in a comparative or counterfactual form, as follows. Specifically, we consider regulations that monitor how the filtering algorithm $\mathcal{F}$ responds to different sets of inputs. Recall that the auditor has black-box access to the filtering algorithm $\mathcal{F}$ and $Z^{(t)}$ denotes the output of $\mathcal{F}$ for the given user. Let $W_k^{(t)}$ denote the counterfactual feeds obtained by running $\mathcal{F}$ on inputs $\mathbf{f}_k$.

Example 1. *Suppose that the regulation under consideration requires that targeted advertisements cannot use on sexual orientation as an input [57].[5] Suppose that the user's indicated sexual orientation is $s' \in S$. Let $W_s^{(t)}$ be the counterfactual feed that differs from $Z^{(t)}$ only in its advertisements such that the advertisements in $W_s^{(t)}$ are those if the user's indicated sexual orientation were $s \in S \setminus \{s'\}$.*[6]

Example 2. *Suppose that the regulation under consideration requires that recommended articles with medical advice on COVID-19 are robust to whether the user is left-leaning or right-leaning. Let*

[4] Note that a user's beliefs are often influenced by information other than their feed $Z^{(t)}$. For instance, a user's beliefs may depend on conversations that they has offline or on their beliefs at previous time steps. Adding this layer of complexity would not change our findings (see the Appendix for details). In other words, we could let $J^{(t)} \in \mathcal{J}$ denote any information other than $Z^{(t)}$ that the user uses to update their beliefs at time t and $\mathcal{L} : \mathcal{Z} \times \mathcal{J} \to \Theta$ such that $\boldsymbol{\theta}^{(t)} = \mathcal{L}(Z^{(t)}, \mathcal{J}^{(t)})$, and our findings would remain the same. Since this added complexity does not change our results, we remove it for clarity of exposition.

[5] We do not place a value judgment on whether it is correct to enforce a regulation. Rather the purpose of this work is to provide a procedure when given a regulation that has been agreed upon by policymakers.

[6] Note that this example is simplified in order to illustrate the meaning behind a counterfactual regulation. If one wishes to protect against more nuanced effects, like proxy variables, one could take a more sophisticated approach in which the counterfactual feeds $W_k^{(t)}$ are generated by not only modifying the user's sexual orientation but also adjusting proxy variables. The proposed regulatory procedure allows the auditor to test as many such counterfactual feeds as needed (e.g., to test different values for the proxies). Exploring how one generates the counterfactual feeds is out of the scope of this work. We direct interested readers to texts on causal inference [50] as well as the illuminating debate on causality and fairness [36, 44, 38].

Algorithm 1: Regulatory test

Input: Regulation parameter $\epsilon \in [0, 1]$; model family $\Theta \subset \mathbb{R}^r$; filtered feed $Z^{(t)}$ and counterfactual feeds $\{W_k^{(t)}\}_{k=1}^q$ for a given user.

Result: $\hat{H} = H_0$ if test passed; $\hat{H} = H_1$, otherwise.

```
Let $\mathcal{L}^+ : \mathcal{Z} \to \Theta$ denote the MVUE;
Let $\tilde{\boldsymbol{\theta}}_0^{(t)} \leftarrow \mathcal{L}^+(Z^{(t)})$;
for $k = 1, 2, \ldots, q$ do
    Let $\tilde{\boldsymbol{\theta}}_k^{(t)} \leftarrow \mathcal{L}^+(W_k^{(t)})$;
end
for $k = 1, 2, \ldots, q$ do
    Let $I(\tilde{\boldsymbol{\theta}}_0^{(t)})$ denote the Fisher information matrix at $\tilde{\boldsymbol{\theta}}_0^{(t)}$;
    if $(\tilde{\boldsymbol{\theta}}_0^{(t)} - \tilde{\boldsymbol{\theta}}_k^{(t)})^\top I^{-1}(\tilde{\boldsymbol{\theta}}_0^{(t)})(\tilde{\boldsymbol{\theta}}_0^{(t)} - \tilde{\boldsymbol{\theta}}_k^{(t)}) \geq \frac{2}{m}\chi_r^2(1-\epsilon)$ then
        Return $\hat{H} = H_1$;
    end
end
Return $\hat{H} = H_0$;
```

$W_L^{(t)}$ and $W_R^{(t)}$ denote the left- and right-leaning counterfactual feeds produced by $\mathcal{F}$ that differ from $Z^{(t)}$ only in terms of recommended articles that provide medical advice on COVID-19.

Two remarks are in order. First, the purpose of this work is to provide a test *if the auditor is given a regulation* that can be expressed in a comparative or counterfactual form. We do not place a value judgment on the proposed regulations (e.g., on the examples above). Second, producing counterfactuals is a rich field of research. Determining how to produce the "correct" counterfactual feeds $W_k^{(t)}$ is not within the scope of this work. See footnotes 5-6 for further comments.

Regulatory procedure. Testing whether a filtering algorithm complies with a regulation is a challenging task because it is difficult to compare content (e.g., two feeds) in a principled manner. The test that we propose is motivated by the idea that, given two sets of conditions $\mathbf{f}_k$ and $\mathbf{f}_\ell$ under which a regulation requires similar outcomes, *if an auditor is unable to detect a difference in the most gullible user's decision making when shown the feed produced by $\mathcal{F}$ under $\mathbf{f}_k$ compared that when shown the feed under $\mathbf{f}_\ell$, then the platform should not be penalized for the difference in those feeds*.

As such, given a regulation that can be expressed in the form described at the start of this section, the auditor can check whether the platform complies by running the test in Algorithm 1. Intuitively, if the platform passes the test, then it meets regulation for the most gullible user and, therefore, for any other user. We study the theoretical properties of this test in the next section. Note that we study the regulation-reward trade-off in Section 5 and find that there are conditions under which the cost of regulation is low.

4.3 Properties of regulatory test

In the previous section, we proposed a test that, given a regulation that can be expressed in a comparative or counterfactual form, checks whether the platform is compliant with the regulation. In this section, we discuss several properties of the regulatory test. We show that the test is grounded by a strong statistical guarantee, has a natural interpretation, and is easy to implement.

Recall that χ_r^2 denotes the chi-squared distribution with r degrees of freedom. Let $\chi_r^2(p)$ be defined such that $\mathbb{P}(v \leq \chi_r^2(p)) = p$ where $v \sim \chi_r^2$. The following theorem provides both an interpretation and a statistical justification for the regulatory test.

Theorem 2. *Suppose there are two hypotheses at time t for counterfactual feed $k \in [q]$:*

$$H_{0,j}^{(t)} : \boldsymbol{\theta}_0^{(t)} = \boldsymbol{\theta}_k^{(t)} \qquad H_{1,j}^{(t)} : \boldsymbol{\theta}_0^{(t)} \neq \boldsymbol{\theta}_k^{(t)}$$

Let $H_k^{(t)}, \hat{H}_k^{(t)} \in \{H_{0,j}^{(t)}, H_{1,j}^{(t)}\}$ denote the true hypothesis and the decision rule, respectively, at time t for feed k. Then, $\mathbb{P}(\hat{H}_k^{(t)} = H_{1,j}^{(t)} | H_k^{(t)} = H_{0,j}^{(t)}) \leq \epsilon$ as $m \to \infty$.

Statistical guarantee. Theorem 2 states that, if the platform passes the test given in Algorithm 1, then the auditor would be unable to detect a difference between the user's decision making under the filtered feed and under any counterfactual feed with $(1-\epsilon)$-confidence. Put differently, the procedure given in Algorithm 1 is equivalent to a regulatory test that, as $m \to \infty$, is guaranteed to achieve a false positive rate (FPR) of at most ϵ.

Note that it is a mistake to interpret the null hypothesis $H_0 : \boldsymbol{\theta}_0^{(t)} = \boldsymbol{\theta}_k^{(t)}$ as a requirement that the counterfactual feed be identical to the filtered feed in some manner. Rather, the regulatory test only requires that the feeds induce sufficiently similar outcomes, where the notion of similarity is determined in a principled manner (see results and proofs of Proposition 1 and Theorem 2 for details).

Implementation. One benefit of the proposed test is that there is only one tunable parameter ϵ, and it has an intuitive meaning. The parameter ϵ is the maximum FPR. Capping the FPR ensures that the auditor is not distracted by red herrings. It is also equivalent to requiring that the auditor only invest the resources needed to investigate the platform when it has at least $(1-\epsilon)$-confidence that the platform is not in compliance. Decreasing $\epsilon \in [0, 1]$ reduces the number of false positives while increasing ϵ makes the regulation more strict (at the risk of receiving more false positives).

In practice, the auditor may wish to run this test over many users' feeds and observe when $\hat{H} = H_1$. For example, one could run the test over a random subset of feeds and investigate the platform only if $\hat{H} = H_1$ for at least $\alpha \in [0, 1]$ of the feeds. If the auditor is concerned about how the platforms treats subgroups, the auditor can run an analogous test over the feeds of user subgroups.

When the MVUE does not exist, use the MLE. Although the maximum likelihood estimator (MLE) does not satisfy the same properties as the MVUE with respect to Proposition 1, there are cases in which the MVUE does not exist but the MLE does. The MLE is a good substitute for the MVUE for two reasons. First, Theorem 2 uses the fact that the MVUE is asymptotically normal when it exists. Second, when the MVUE does exist, it is asymptotically equivalent to the MLE, which implies that the MLE behaves like the MVUE asymptotically.

5 Cost of regulation and the role of content diversity

In this section, we study the regulation from the platform's perspective; specifically, how it affects the filtering feeds. We first investigate the cost of regulation. Recall that a high cost of regulation implies that there is a strong regulation-reward trade-off. We find in Section 5.1 that there are conditions under which the cost of regulation is low. Interestingly, we find in Section 5.2 that content diversity (e.g., counter-attitudinal content [41]) plays a key role in the platform's ability to achieve a high reward while meeting regulation. These results indicate that content diversity plays a powerful role that can align the interests of regulators and platforms. All proofs can be found in the Appendix.

5.1 Cost of regulation

Reward maximization. Recall from (1) that the optimization problem that the platform seeks to solve is $Z^{(t)} \in \arg\max_{Z \in \mathcal{Z}^{(t)}} R(Z)$. Intuitively, R represents the objective function (or reward) that the platform seeks to maximize. For example, if the platform's goal is to maximize the revenue it accrues from clicks, then $R = \sum_c (\text{revenue per click for content } c) \times (\text{number of clicks on content } c)$. As discussed in Section 3, our results hold for all R unless otherwise specified.

Feasible set. Recall that the feasible set $\mathcal{Z}^{(t)} \subset \mathcal{Z}$ is the set of feeds that pass regulation at time t. Intuitively, $\mathcal{Z}^{(t)} = \mathcal{Z}$ without regulation, and the more restrictive the regulation, the smaller the feasible set. A well known relationship in constrained optimization is that, when the feasible set $\mathcal{Z}^{(t)}$ shrinks from $\mathcal{Z}_1$ to $\mathcal{Z}_2$, the maximum achievable reward $\max_{Z \in \mathcal{Z}^{(t)}} R(Z)$ is affected in one of two ways. If the feasible set shrinks such that all reward-maximizing solutions in $\mathcal{Z}_1$ are not contained in $\mathcal{Z}_2$, then the maximum reward *decreases*, and the cost of regulation *increases*. Alternatively, if the feasible set shrinks but at least one reward-maximizing solution in $\mathcal{Z}_1$ is contained in $\mathcal{Z}_2$, then the maximum reward stays the same, and the cost of regulation *does not increase*.

Cost of regulation. Therefore, although the feasible set $\mathcal{Z}^{(t)}$ is a subset of $\mathcal{Z}$, the cost of regulation is can be low when $\mathcal{Z}^{(t)}$ contains at least one (near) optimal solution. In other words, the feasible set $\mathcal{Z}^{(t)}$ may not contain all (near) optimal solutions in $\mathcal{Z}$, but as long as it preserves at least one (near) optimal solution, then the regulation-reward trade-off is mild.

The following result formalizes this notion. For this result, we overload the notation R such that $R(\boldsymbol{\theta})$ denotes $R(Z)$, where $\boldsymbol{\theta} = \mathcal{L}^+(Z)$.

Theorem 3. *Suppose that there exists $\Omega \subset [r]$ where $1 < |\Omega| < r$ such that $R(\boldsymbol{\theta}_1) = R(\boldsymbol{\theta}_2)$ if $\theta_{1,i} = \theta_{2,i}$ for all $i \notin \Omega$. Suppose further that, for any $\boldsymbol{\theta} \in \Theta$, there exists $\bar{\boldsymbol{\theta}}_\Omega$ where $\bar{\theta}_{\Omega,i} = 0$ for all $i \notin \Omega$, such that for any $\beta > 0$ and $\mathbf{v} \in \mathbb{R}^r$, there exists a $\kappa > 0$ such that $\mathbf{v}^\top I^{-1}(\boldsymbol{\theta} + \kappa\bar{\boldsymbol{\theta}}_\Omega)\mathbf{v} < \beta$ and $\boldsymbol{\theta} + \kappa\bar{\boldsymbol{\theta}}_\Omega \in \Theta$. Then, if $m < \infty$, there exists a $\mathcal{Z}$ such that the cost of regulation is 0.*

Interpretation of the result. A regulation restricts the platform's feasible set, which may impose a cost on the platform by reducing its maximum attainable reward. However, a high cost of regulation is not inevitable. Indeed, if the feasible set contains at least one (near) optimal solution, then the cost of regulation is low. Theorem 3 provides a broad set of conditions under which the cost of regulation is low. Intuitively, the result states that, when R is independent of at least one element in the parameter vector $\boldsymbol{\theta}$ and that element has sufficient leverage over the Fisher information, then as long as the amount of content at a given time step is finite (i.e., $m < \infty$) and the available content $\mathcal{Z}$ is expressive enough, then the platform can always construct a feed from $\mathcal{Z}$ that passes regulation without sacrificing reward.

One may ask whether the conditions in Theorem 3 are reasonable. To illustrate that the conditions are achievable, consider the following highly simplified example.

Example 3. *Suppose $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}$, where $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma^2) \in \Theta\}$. In other words, we consider the family of one-dimensional Gaussian distributions. Then, if R is a function of the mean μ but not the variance σ^2, then the conditions of Theorem 3 are met. To see this, observe that $\Omega = \{2\}$ and let $\bar{\boldsymbol{\theta}}_\Omega = (0, 1)$. Finally, recall that:*

$$I^{-1}((\mu, \sigma^2)) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

This example shows that there are settings in which the conditions in Theorem 3 hold. Recall that Theorem 3 provides conditions under which there is *no cost of regulation*. The conditions can be relaxed if we are simply interested in situations for which the cost of regulation is low but not zero. We build this intuition in the next section, studying one of the levers that the platform can pull in order to increase its reward.

5.2 Content diversity

In the previous section, we discussed how a regulation can impose a cost on the platform by restricting the feasible set and found that there are conditions under which the cost of regulation is low. In this section, we show that these results suggest that *content diversity plays a key role* in aligning the interests of regulators and platforms. Specifically, under regulation, the platform is incentivized to filter feeds with a sufficient amount of content diversity because adding doses of content diversity allows it to increase its reward while remaining within the feasible set.

We first formalize content diversity, then show how it relates to passing regulation.

Definition 2. *For $\mathbf{v} \in \mathbb{R}^r$ and $Z_1, Z_2 \in \mathcal{Z}$, the feed Z_1 has more* content diversity *than Z_2 along $\mathbf{v}$ if the Fisher information at $\mathcal{L}^+(Z_1)$ and $\mathcal{L}^+(Z_2)$ satisfies: $\mathbf{v}^\top(I^{-1}(\mathcal{L}^+(Z_2)) - I^{-1}(\mathcal{L}^+(Z_1))\mathbf{v} > 0$.*

Interpretation. In words, this definition says that the "smaller" the inverse Fisher information or, equivalently, the "larger" the Fisher information, the higher the content diversity. To understand this definition, recall that $I(\boldsymbol{\theta})$—which is referred to as the Fisher information at $\boldsymbol{\theta}$—can be viewed as a measure of the maximum speed at which one can learn $\boldsymbol{\theta}$ with high certainty when given i.i.d. samples drawn from $p_{\mathbf{z}}(\cdot; \boldsymbol{\theta})$.

Therefore, Definition 2 can be interpreted in two ways. One interpretation is that a feed has low content diversity with respect to $\mathbf{v}$ if the regulator can, without much effort, distinguish it from counterfactual feeds that differ along $\mathbf{v}$. Alternatively, a feed has low content diversity with respect to $\mathbf{v}$ if a user's belief about $\mathbf{v}$ when formed on the basis of the feed alone would be very strong.

For example, suppose that $\mathbf{v}$ corresponds to political content. Then, a feed that shows only left-leaning (or right-leaning) political content has very low content diversity with respect to $\mathbf{v}$. If the user forms their belief solely on the basis of such a feed, then the user's political opinions would concentrate on the left (respectively, right) side of the political spectrum. This relationship is precisely the same as

that captured in Definition 2. In this example, $\mathbf{v}^\top I^{-1}(\mathcal{L}^+(Z))\mathbf{v}$ quantifies the density of political opinions, where high values of $\mathbf{v}^\top I^{-1}(\mathcal{L}^+(Z))\mathbf{v}$ indicate low content diversity in Z along $\mathbf{v}$.[7]

Content diversity and passing regulation. We now show that adding doses of content diversity helps the platform pass regulation at low cost, which implies that the proposed regulatory test naturally incentivizes the platform to filter a sufficient amount of content diversity into its feeds.

Recall from Algorithm 1 that the platform passes the regulatory test when $(\tilde{\boldsymbol{\theta}}_0^{(t)} - \tilde{\boldsymbol{\theta}}_k^{(t)})^\top I^{-1}(\tilde{\boldsymbol{\theta}}_0^{(t)})(\tilde{\boldsymbol{\theta}}_0^{(t)} - \tilde{\boldsymbol{\theta}}_k^{(t)})$ is sufficiently small, where $\tilde{\boldsymbol{\theta}}_0^{(t)} = \mathcal{L}^+(Z^{(t)})$. We also show in the proof of Theorem 3 that one of the conditions under which there is no cost of regulation is when $\mathbf{v}^\top I^{-1}(\tilde{\boldsymbol{\theta}}_0^{(t)})\mathbf{v}$ is small. In other words, it is easier for the platform to meet regulation when the inverse Fisher information at $\tilde{\boldsymbol{\theta}}_0^{(t)}$ is sufficiently "small" or, equivalently, the Fisher information at $\tilde{\boldsymbol{\theta}}_0^{(t)}$ is sufficiently "large". By Definition 2, whether the Fisher information is "large" is precisely an indication of how much diversity is in a feed. Therefore, if the platform wishes to achieve high reward but the feed that produces this high reward does not pass the regulatory test (i.e., is not in the feasible set), then the platform can generally filter another high reward feed by adding doses of content diversity. [8]

Because content diversity is not part of the regulatory test by design, this result is surprising at first. It states that the regulatory test naturally incentivizes the platform to include a sufficiently amount of content diversity with respect to $\mathbf{v}$, where $\mathbf{v}$ represents the dimension along which the regulation is defined. Returning to Example 1, if regulators require that medical information on COVID-19 be robust to whether a user is left- or right-leaning, then the differences between the medical information shown to users across the political spectrum is captured by $\tilde{\boldsymbol{\theta}}_0^{(t)} - \tilde{\boldsymbol{\theta}}_k^{(t)}$. Adding content diversity along this dimension means that right-leaning users should not receive medical information on COVID-19 only from right-leaning news outlets, but also from left-leaning ones, and vice versa.

6 Conclusion

Social media platforms select the content that is shown on a user's feed using a method known as algorithmic filtering. Filtering algorithms have drawn scrutiny in recent years, and there is a growing call for regulations. However, designing and enforcing such regulations on social media platforms is challenging. For one, a regulation must balance being general enough to last with time while being precise enough to carry out in practice. Moreover, there are concerns that regulations can harm innovation or impose large costs on the platform.

In Section 4, we propose a general procedure such that, given a regulation on algorithmic filtering, an auditor can test whether the regulation is met by the platform. The procedure can be applied to any regulation that can be expressed in the counterfactual or comparative form described in Section 4.2. In general, testing whether a filtering algorithm complies with a regulation is challenging because it is difficult to compare content in a principled manner. We show that the proposed test has strong statistical guarantees in addition to being interpretable and easy to implement.

In Section 5, we then examine how the proposed test affects the platform and the feeds that it filters if it wishes to pass the proposed regulatory test. We find in Section 5.1 that there are conditions under which there is no cost of regulation (i.e., no regulation-reward trade-off). We further investigate this result in Section 5.2 and discover that one of the ways a platform can achieve high reward under regulation is by adding doses of content diversity. This finding suggests that content diversity plays an important role in aligning the interests of regulators and platforms.

In this work, we provide a general regulatory test *given* a regulation. We do not prescribe a regulation. We characterize the possible effects of the test in Sections 4-5. An interesting area of future work would be to study the behavior of the proposed test on a specific regulation and platform using experimental data. For example, although we show that there are conditions under which the cost of regulation is low, it would be prudent to determine whether these conditions generally hold.

[7] One can think of content diversity both in terms of the diversity of topics in a feed or of the diversity of opinions on each given topic in the feed. The Fisher information elegantly captures both notions simultaneously.

[8] Note that the reason the platform does not necessarily increase the content diversity indefinitely is because the platform must also balance "enlarging" the Fisher information with other considerations, such as ensuring that other terms in Line 8—specifically, $\tilde{\boldsymbol{\theta}}_0^{(t)} - \tilde{\boldsymbol{\theta}}_k^{(t)}$—do not cause the platform to fail the regulatory test.

References

[1] 47 U.S. Code 230 - Protection for private blocking and screening of offensive material.

[2] Democracy Under Threat: Risk and Solutions in the Era of Disinformation and Data Monopoly.

[3] Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian Learning in Social Networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[4] Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. November 2017.

[5] Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (Still) Letting Housing Advertisers Exclude Users by Race, 2017.

[6] Abhijit Banerjee. A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

[7] BBC News. Social Media: How Can Governments Regulate It? April 2019.

[8] BBC News. Twitter suspends 70,000 accounts linked to QAnon. January 2021.

[9] Hal Berghel. Lies, Damn Lies, and Fake News. *Computer*, 50(2):80–85, 2017.

[10] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and Applications of Robust Optimization. *SIAM Review*, 53(3):464–501, 2011.

[11] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The Price of Fairness. *Operations Research*, 59(1):17–31, 2011.

[12] Dimitris Bertsimas and Melvyn Sim. The Price of Robustness. *Operations Research*, 52(1):35–53, 2004.

[13] Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3):209–227, 2013.

[14] Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, 2015.

[15] Valerie C. Brannon. Free Speech and the Regulation of Social Media Content, 2019.

[16] Andrea Broughton, Beth Foley, Stefanie Ledermaier, and Annette Cox. The use of social media in the recruitment process. *Institute for Employment Studies, Brighton*, 2013.

[17] German Bundestag. Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), 2017.

[18] James Campbell, Avi Goldfarb, and Catherine Tucker. Privacy Regulation and Market Structure. *Journal of Economics & Management Strategy*, 24(1):47–73, 2015.

[19] Shih-Chih Chen and Chieh-Peng Lin. Understanding the effect of social media marketing activities: The mediation of social identification, perceived value, and satisfaction. *Technological Forecasting and Social Change*, 140:22–32, 2019.

[20] Bobby Chesney and Danielle Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107:1753–1820, 2019.

[21] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017.

[22] Damian Collins, Clive Efford, Julie Elliott, Paul Farrelly, Simon Hart, Julian Knight, Ian C. Lucas, Brendan O'Hara, Rebecca Pow, Jo Stevens, and Giles Watling. Disinformation and 'fake news': Final Report, February 2019.

[23] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. AAAI, 2017.

[24] Morris H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[25] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. "Algorithms Ruin Everything" #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3163–3174. Association for Computing Machinery, 2017.

[26] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284. Springer, 2006.

[27] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2013.

[28] Benedict Evans. Regulating technology. `https://www.ben-evans.com/benedictevans/2020/7/23/regulating-technology`, July 2020.

[29] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268. Association for Computing Machinery, 2015.

[30] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, March 2016.

[31] Roberto Garvin. How social networks influence 74 percent of shoppers for their purchasing decisions today. May 2019.

[32] Ali Ghobadzadeh, Sayed Jalal Zahabi, and Ali A. Tadaion. The role of MVU estimator and CRB in binary composite hypothesis test. In *2009 IEEE International Symposium on Information Theory*, pages 973–976. IEEE, 2009.

[33] Giorgos Giannopoulos, Marios Koniaris, Ingmar Weber, Alejandro Jaimes, and Timos Sellis. Algorithms and criteria for diversification of news article comments. *Journal of Intelligent Information Systems*, 44(1):1–47, 2014.

[34] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, 2018.

[35] Dennis D. Hirsch. The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-rRegulation. *Seattle University Law Review*, 34:439, January 2011.

[36] Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning. *arXiv preprint arXiv:2006.01770*, 2020.

[37] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware Learning through Regularization Approach. In *Proceedings of 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[38] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

[39] Pauline T. Kim and Sharion Scott. Discrimination in Online Employment Recruiting. *St. Louis University Law Journal*, 63:93, 2018.

[40] Kate Klonick. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, 131:1598–1670, 2017.

[41] Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–70, 2021.

[42] Becca Lewis and Alice E. Marwick. Media Manipulation and Disinformation Online. *New York: Data & Society Research Institute*, 2017.

[43] Rotem Medzini. Enhanced self-regulation: The case of facebook's content governance. *New Media & Society*, page 1461444821989352, 2021.

[44] Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

[45] Pooya Molavi, Alireza Tahbaz-Salehi, and Ali Jadbabaie. A Theory of Non-Bayesian Social Learning. *Econometrica*, 86(2):445–490, March 2018.

[46] Jonathan A. Obar and Steven S. Wildman. Social Media Definition and the Governance Challenge. *Telecommunications Policy*, 39(9):745–750, 2015.

[47] Oversight Board. Ensuring respect for free expression, through independent judgment. March 2020.

[48] Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Publishing Group, 2011.

[49] The European Parliament and the Council of the European Union. General Data Protection Regulation.

[50] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[51] Jaime E Sidani, Ariel Shensa, Beth Hoffman, Janel Hanmer, and Brian A Primack. The Association Between Social Media Use and Eating Concerns Among US Young Adults. *Journal of the Academy of Nutrition and Dietetics*, 116(9):1465–1472, 2016.

[52] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. Analyzing and Mining Comments and Comment Ratings on the Social Web. *ACM Transactions on the Web*, 8(3), July 2014.

[53] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Potential for Discrimination in Online Targeted Advertising. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 5–19, New York, NY, USA, 23–24 Feb 2018. PMLR.

[54] Latanya Sweeney. Discrimination in Online Ad Delivery. *Queue*, 11(3):10–29, March 2013.

[55] The Independent. Amid Capitol violence, Facebook, YouTube remove Trump video. January 2021.

[56] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social Media, Political Polarization, and Political Disinformation: A Review of The Scientific Literature. March 2018.

[57] Sandra Wachter. Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal*, 35(2), 2020.

[58] Larry Wasserman and Shuheng Zhou. A Statistical Framework for Differential Privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[59] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally Robust Convex Optimization. *Operations Research*, 62(6):1358–1376, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We describe the scope of our work in Section 2. We also discuss the limitations of our work throughout Sections 4-5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the potential negative impact in Section 6 and the Appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We do not run experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We do not use existing assets or curate/release assets.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conduct research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]