Approximate Least-Favorable Distributions and Nearly Optimal Tests via Stochastic Mirror Descent*

Andrés Aradillas Fernández† José Blanchet‡ José Luis Montiel Olea§
Chen Qiu§ Jörg Stoye§ Lezhi Tan‡

October 2025

Abstract

Elliott, Müller, and Watson (2015) suggested a numerical algorithm for determining a *nearly* most powerful test of a given size in a general class of nonstandard hypothesis testing problems. Their algorithm iteratively updates a candidate least-favorable distribution supported on finitely many points. We show that—in testing problems where the null hypothesis postulates M distributions for the observed data—a slight variation of their algorithm coincides with a *stochastic mirror descent* routine for convex optimization. This insight allows us to formally show that, given a desired approximation error, one can use a stochastic mirror descent routine to *provably* obtain—after finitely many iterations—both an approximate least-favorable distribution and a nearly optimal test, in a sense we make precise. Our theoretical results yield concrete recommendations about the algorithm's implementation, including its initial condition, its step size, the number of iterations, and the number of stochastic draws per iteration that can be used to approximate the subgradient of the objective function.

1 Introduction

Consider the problem of testing a null hypothesis that postulates *finitely many* distributions for the observed data against a *single* alternative hypothesis. More concretely, suppose the data is modeled

*We would like to thank Isaiah Andrews, Tim Armstrong, Patrik Guggenberger, Ulrich Müller, Karthik Sridharan, and Vasilis Syrgkanis for very helpful feedback, comments, and suggestions. We gratefully acknowledge financial support from the NSF under grants SES-2315600, 2229012, 2312204, 2403007; and from the Department of Defense through the Air Force Office of Scientific Research under award number FA9550-20-1-0397 and ONR 1398311

†Department of Economics, Massachusetts Institute of Technology.
‡Management Science and Engineering Department, Stanford University.
§Department of Economics, Cornell University.

as a $\mathcal{Y}$-valued random variable, denoted as Y, and let the hypothesis testing problem take the form:

$$\mathbf{H}_0 : \text{the distribution of } Y \text{ is } F_m, \quad m = 1, \ldots, M, \quad \textit{vs.} \quad \mathbf{H}_1 : \text{the distribution of } Y \text{ is } G. \tag{1}$$

Elliott et al. (2015) presented a numerical iterative algorithm to obtain a *nearly optimal* test for (1) and applied their procedure to some nonstandard hypothesis problems that involve nuisance parameters.[1] Broadly speaking, their nearly optimal test is designed to correctly reject the null hypothesis almost as frequently as it would be theoretically possible, while guaranteeing that the probability of incorrectly rejecting the null is bounded above by a prespecified constant $\alpha \in (0, 1)$. As we will explain later, the key component of the algorithm in Elliott et al. (2015) is a simple and intuitive formula (see their Equation 10, p. 782) that is used to iteratively update a candidate *least-favorable* distribution over the M densities in the null hypothesis.[2] The nearly optimal test in Elliott et al. (2015) is a likelihood ratio test of size α that replaces the null hypothesis, $\mathbf{H}_0$, by a single mixture distribution obtained from averaging $F_1, \ldots, F_M$ using the least-favorable distribution obtained in the last iteration of their algorithm. To the best of our knowledge, the theoretical guarantees for this procedure remain unknown.

The goal of this paper is to present new theoretical results showing that a slight variation of the numerical algorithm developed by Elliott et al. (2015) can indeed be used to *provably* find—after finitely many iterations and with high probability—both an *approximate* least-favorable distribution and a *nearly optimal* test for (1), in a sense we make precise. In order to better understand the value of our work, it is helpful to organize our main results as follows.

Our first result (Theorem 1) shows that a slight variation of the iterative routine suggested by Elliott et al. (2015) coincides with an algorithm known in the convex optimization literature as

[1] See Elliott and Müller (2014), Müller and Watson (2016), Müller and Wang (2017), Müller and Watson (2018), Guggenberger, Kleibergen, and Mavroeidis (2019), Müller and Watson (2020), Dou and Müller (2021), Li and Müller (2021), Müller (2025), Muralidharan, Romero, and Wüthrich (2025) for applications of the algorithm in Elliott et al. (2015) to different econometric problems.

[2] See Lehmann and Romano (2005), Chapter 3, p. 84, for a definition of a least-favorable distribution in hypothesis testing problems.

Stochastic Mirror Descent (Bubeck, 2015, Chapter 6.1). Stochastic Mirror Descent (henceforth, S-MD) is a family of first-order, iterative algorithms designed to *approximately* minimize convex functions when an unbiased estimator of its subgradient is available.[3] We show that the convex program that arises naturally in the testing problem in (1) is the *dual* of the mathematical program that defines the *most powerful test* of *size* α; see Equations 4-5-6 and Lemma 1). The choice variables in the dual problem—which we refer to as *multipliers*—are nonnegative vectors $\kappa \in \mathbb{R}^M$, and the least-favorable distribution is proportional to the optimal multipliers (Remark 1). As we explain later, there are three slight modifications to the algorithm in Elliott et al. (2015) that transform it into an off-the-shelf S-MD routine. First, the S-MD formula we use to update the multipliers at the end of each iteration—Equation 11 in Theorem 1—ensures that the candidate solution of the dual satisfies the constraint $\sum_{m=1}^{M} \kappa_m \leq 1/\alpha$, a property that is satisfied by the true solution (Lemma 2). This adjustment introduces an additional proportionality constant that is not part of the algorithm presented in Elliott et al. (2015). Second, the typical output of the S-MD algorithm is the *average value* of the multipliers over all iterations, as opposed to the multiplier obtained in last update.[4] Third, the typical implementation of the S-MD routine comes with a specific suggestion for the initial condition, which in the testing problem we consider is of the form $(1/(\alpha M), \ldots, 1/(\alpha M))$, provided $M > \exp(1)/\alpha$ (see Theorem 1).

The slight modifications discussed in the paragraph above are crucial for the theoretical results presented in this paper. We define a distribution to be *approximately least favorable* when it can be used to approximately solve the dual problem, up to an additive constant ϵ (see Definition 1). Our second result (Theorem 2) leverages well-known results in the convex optimization literature to show that the S-MD routine described in our Theorem 1—stopped after $(4(1-\alpha)^2/\alpha^2\epsilon^2) \cdot \ln(M)$

[3]Mirror descent is known to outperform stochastic gradient descent in some cases; see Section 4.3 of Bubeck (2015).

[4]Averaging the trajectories of a stochastic gradient-descent routine is commonly referred to as Polyak-Ruppert averaging. See Ruppert (1988) and Polyak and Juditsky (1992). See also Forneron (2024) for a discussion of Polyak-Ruppert averaging in the context of estimation and inference by stochastic optimization of nonlinear econometric models.

iterations—outputs, with high probability, a candidate vector of multipliers that is almost $\epsilon > 0$ away from attaining the optimal value of the dual problem in (5).[5] This result has at least two other important implications. First, the number of iterations used by the algorithm scales logarithmically in M, which means there is no theoretical sense in which the number of iterations scales poorly as function of how many elements there are in the null hypothesis in (1). Second, and perhaps the most striking feature of our analysis, the unbiased estimator of the gradient defined in Equation 10 in Theorem 1 *can be based on a single Monte Carlo draw* from each null distributions F_m, $m = 1, \ldots, M$. As expected, taking a larger number of draws improves the approximation error of the S-MD routine, but using a small number of draws reduces the computational burden of the algorithm.

Finally, we analyze the extent to which the output of the S-MD routine allows us to construct a nearly optimal test for the testing problem (1). We define a test φ to be (ϵ, δ)-nearly optimal (Definition 2) whenever i) the size of φ is at most $\alpha + \delta$; and ii) up to the additive constant ϵ, the test φ rejects the null hypothesis as frequently as the most powerful test of size α. Based on this definition, we say that a test is *nearly optimal* if it is (ϵ, δ)-nearly optimal for some parameters ϵ and δ. Our third result (Theorem 3) shows that—by keeping track of the history of multipliers $\{\kappa_t\}_{t=1}^{T}$ generated by the S-MD routine—it is always possible to construct a *randomized* nearly optimal test with high probability. The nearly optimal test is shown to be the *average* of each of the tests of the Neyman-Pearson form associated with each vector of multipliers κ_t generated by the S-MD routine. We present explicit expressions for its size distortion and its power loss relative to the best test. We note that an important challenge in reporting such a test is that, in principle, it requires keeping track of the history of multipliers obtained from the S-MD routine. When both M and the number of iterations of S-MD are large, this could come at a significant computational cost. To address this issue, we show that there is a simple strategy to implement the average test: we randomize the

[5]The number of iterations should be interpreted as being sufficient and not necessary. In any given application, it is possible that a smaller number could be used to output an ϵ-least favorable distribution, but the formula presented above is valid for any null and alternative distributions.

number of iterations uniformly between 1 and our recommended T, update the multipliers, and use the resulting test to decided whether or not to reject the null hypothesis.

Related Literature: Although this paper focuses on the algorithm developed by Elliott et al. (2015), there are other alternative approaches to find a test for (1). For instance, when the sample space $\mathcal{Y}$ is infinite, the mathematical problem that defines the most powerful test of a given size in the problem (1) is an *infinite* linear programming problem: Since the testing problems analyzed in this paper posit M null distributions for the observed data, the corresponding linear program has finitely many constraints but a choice variable of infinite dimension; see Section 2 below for details. If the data were discrete-valued or if we were to discretize it—as suggested by Chiburis (2008), Moreira and Moreira (2013), and Moreira and Moreira (2019)—then one could find the most powerful test of a given size by using any algorithm for finite linear programming. Krafft and Witting (1967) is the seminal reference for using linear programming methods to characterize the most powerful test of a given size.

Elliott et al. (2015) also show how the most powerful test for composite hypotheses can be expressed as a minimax decision problem where a false rejection of $\mathbf{H}_0$ induces a loss of 1, and a false rejection of $\mathbf{H}_1$ induces a loss of $\phi > 0$ (correct choices have loss of zero). In this problem the decision maker chooses a test φ. An adversarial nature decides which element in $\{F_1, F_2, \ldots, G\}$ to use to generate the data; consequently, a mixed strategy for nature can be represented by a vector in the simplex of $\mathbb{R}^{M+1}$. One could then use an algorithm for solving the corresponding maximin problem—for example, the Hedge algorithm suggested by Aradillas Fernández, Blanchet, Montiel Olea, Qiu, Stoye, and Tan (2025); the fictitious play algorithm suggested by Guggenberger and Huang (2025); or a general convex optimization routine as in Chamberlain (2000). An important limitation of this approach is that one would need to solve the maximin problem repeatedly for different values of ϕ, until one finds a test with correct size.

Although there are no theoretical results showing that one of these algorithms is better than another for the purpose of finding a nearly optimal test, we think that our analysis illustrates

how a rich literature in optimization can be leveraged to provide theoretical results about the performance of different algorithms and also to provide practical recommendations regarding their implementation.

OUTLINE: The rest of the paper is organized as follows. Section 2 presents notation, the statement of the hypothesis testing problem of interest, and the primal and dual optimization problems that arise when searching for the most powerful test of a given size. Section 3 presents a formal definition of a stochastic mirror descent routine (S-MD) for convex optimization and also our main results: namely, that a slight variation of the algorithm in Elliott et al. (2015) provably generates an approximate least favorable distribution and a nearly optimal test. Section 4 uses an elementary testing problem that arises in the context of the univariate Gaussian location model to illustrate our main results. Section 5 discusses some extensions and Section 6 concludes. The proofs of our main theorems and supporting lemmas are collected in Appendix A. Additional results are collected in Appendix B.

2 Notation and Statement of the Problem

We first present the formal statement of the hypothesis testing problem analyzed in this paper. We follow the notation and terminology used in Elliott et al. (2015) as close as possible. We then present the *dual* problem that will be used to connect the iterative routine suggested by Elliott et al. (2015) with the stochastic mirror descent algorithm for convex optimization problems.

2.1 Statement of the Hypothesis Testing Problem

We observe a random element Y that takes values in some space $\mathcal{Y}$ endowed with σ-algebra $\mathcal{F}$. Let ν denote a σ-finite measure defined over the measurable space $(\mathcal{Y}, \mathcal{F})$. Let $F_1, \ldots, F_M$ denote $M > 1$ candidate probability measures for the distribution of Y under the null hypothesis. Let G be the candidate distribution of Y under the alternative hypothesis. Assuming all of these distributions are

absolutely continuous with respect to ν, Theorem 5.5.4 in Dudley (2002) guarantees the existence of nonnegative integrable functions $f_1, \ldots, f_m, g$, which can be taken as the probability density functions of $F_1, \ldots, F_m, G$ relative to ν.

Based on a single observation of Y, the testing problem of interest is

$$\mathbf{H}_0: \text{ the density of } Y \text{ is } f_m, \quad m = 1, \ldots, M, \quad \textit{against} \quad \mathbf{H}_1: \text{ the density of } Y \text{ is } g. \tag{2}$$

Using the typical jargon of hypothesis testing problems, the null hypothesis in (2) is *composite*, since it contains more than one possible distributions for the data. The alternative hypothesis is *simple*, in that it contains a single distribution.[6]

A *statistical test* for (2) (or simply a *test*) is a measurable function $\varphi : \mathcal{Y} \to [0, 1]$, where $\varphi(y)$ is interpreted as the probability of rejecting the null hypothesis given that data y were observed. A test φ is said to be *nonrandomized* if $\varphi(y) \in \{0, 1\}$ for ν-almost every realization of Y; otherwise the test is said to be *randomized*.

The rate of Type I error under f_m is the probability of rejecting the null hypothesis when $Y \sim f_m$ and it equals $\int \varphi f_m d\nu$. As usual, the *size* of a test is the largest rate of Type I error under the null hypothesis. The *power* of a test is the probability of rejecting the null hypothesis when $Y \sim g$ and it equals $\int \varphi g d\nu$.

2.2 Primal and Dual Problems in Hypothesis Testing

We would like to find the *most powerful test* of size α for the problem (2). By definition, such a test correctly rejects the null hypothesis as frequently as possible, but guarantees that the probability of incorrectly rejecting the null is bounded above by the prespecified constant $\alpha \in (0, 1)$.

[6] As explained in Section 2.2 of Elliott et al. (2015), the density g can arise by appealing to the weighted average power criterion in cases where the alternative hypothesis is composite as well.

Mathematically, the problem of finding the most powerful test of size α can be written as:

$$\sup_{\varphi:\mathcal{Y}\to[0,1]} \int \varphi g d\nu, \quad \text{s.t.} \quad \int \varphi f_m d\nu \leq \alpha, \quad m = 1, ..., M. \tag{3}$$

We refer to the optimization problem in (3) as the *primal* problem associated with the hypothesis testing problem in (2). We note that the primal problem is an infinite linear programming problem, in the sense of Anderson and Nash (1987). The infinite linear program in (3) has finitely many constraints but a choice variable of infinite dimension.

Define the Lagrangian function associated with the optimization problem (3) as

$$L(\varphi, \kappa) \equiv \int \varphi g d\nu - \sum_{m=1}^{M} \kappa_m \left[\int \varphi f_m d\nu - \alpha \right], \tag{4}$$

where we refer to $\kappa \equiv (\kappa_1, ..., \kappa_M) \in \mathbb{R}_+^M$ as the Lagrange multipliers (or simply, *multipliers*) associated with each of the inequality constraints in the primal problem (3).

Consider thus the optimization problem on $\mathbb{R}_+^M$ with variable $\kappa \equiv (\kappa_1, ..., \kappa_M)$ given by

$$\bar{v} \equiv \inf_{\kappa \in \mathbb{R}_+^M} f(\kappa), \tag{5}$$

where

$$f(\kappa) \equiv \sup_{\varphi:\mathcal{Y}\to[0,1]} L(\varphi, \kappa). \tag{6}$$

We refer to the problem in (5) as the *dual* problem of (3).

Remark 1. The dual problem in (5) is only a device to solve the primal problem in (3). This is a well-known fact, and we present a heuristic argument to help the exposition (relegating technical details to Appendix B.2). Suppose that the multipliers κ^* solve the problem (5) in that $f(\kappa^*) = \bar{v}$.

Then, by definition

$$\begin{aligned} f(\kappa^*) &= \sup_{\varphi:\mathcal{Y}\to[0,1]} L(\varphi,\kappa^*) \\ &= \int \varphi_{\kappa^*} g d\nu - \sum_{m=1}^{M} \kappa_m^* \left[\int \varphi_{\kappa^*} f_m d\nu - \alpha \right], \end{aligned}$$

where φ_{κ^*} is a test of the *Neyman-Pearson* form; that is

$$\varphi_{\kappa^*}(y) \equiv \begin{cases} 1 & \text{if } g(y) > \sum_{m=1}^{M} \kappa_m^* f_m(y) \\ 0 & \text{if } g(y) \leq \sum_{m=1}^{M} \kappa_m^* f_m(y). \end{cases} \tag{7}$$

Lemma 1 in Elliott et al. (2015) and Theorem 3.8.1 in Lehmann and Romano (2005) imply that if the test φ_{κ^*} has size α under $\mathbf{H}_0$ then φ_{κ^*} solves (3); that is, it maximizes power among all tests of size at most α. The direction of the vector κ^*—namely, $\lambda^* \equiv \kappa^* / \sum_{m=1}^{M} \kappa_m^*$ is a least-favorable distribution in the sense of Lehmann and Romano (2005), Chapter 3, p. 84. □

Remark 2. In order to justify the terminology of primal and dual problems, Section B.2 in Appendix B formalizes the connection between the optimization problems (3) and (5), by showing that the value functions of both problems are equal, and that a solution to the dual problem in (5) can indeed be translated to a solution to the primal problem in (3). As usual, an important step in showing that the solution of the dual problem can be used to solve the primal problem consists in verifying that the complementary slackness conditions in the dual problem are satisfied. We also note that similar duality results have been established and used elsewhere; for example, see Proposition 3.1 and Equation 3.11 of Cvitanic and Karatzas (2001).

2.3 Solving the dual problem

We have explained how the solution to the dual problem in (5) can be used to construct the most powerful test of a given size. We now discuss the extent to which solving the dual problem is

computationally feasible. We start by showing that the objective function in (5) is convex over $\mathbb{R}_+^M$. We also show that the vector collecting the *excess* rate of Type I error of the test φ_κ in (7) is a subgradient of $f(\cdot)$ at κ.

Lemma 1. *The function $f(\kappa)$ defined in Equation 6 is convex. Furthermore, a subgradient of f at κ is given by*

$$\nabla f(\kappa) \equiv -\left(\int \varphi_\kappa f_1 d\nu - \alpha, ..., \int \varphi_\kappa f_M d\nu - \alpha \right),$$

where φ_κ is defined as in (7).

Proof. See Section A.1 of Appendix A. □

Lemma 1 thus shows that dual problem in (5) has a convex objective function, and admits a simple formula for a subgradient. We now show that the dual problem can be further simplified by restricting the multipliers to belong to the bounded domain:

$$\mathcal{X} \equiv \left\{ \kappa \in \mathbb{R}_+^M : \|\kappa\|_1 \leq \frac{1}{\alpha} \right\}. \tag{8}$$

Lemma 2. $\inf_{\kappa \in \mathbb{R}_+^M} f(\kappa) = \inf_{\kappa \in \mathcal{X}} f(\kappa)$.

Proof. See Section A.2 in Appendix A. □

Lemma 1 and 2 show that in order to find the multipliers associated with the dual program in (5), it is sufficient to solve a convex optimization problem over a bounded domain (in particular, an ℓ_1-ball around the origin with radius $1/\alpha$). In the next section we show that a slight variation of the algorithm suggested by Elliott et al. (2015) indeed approximately solves the dual problem.

3 Main Results

This section presents our main results. First, we present a formal definition of a stochastic mirror descent (S-MD) routine for convex optimization. We then show (Theorem 1) that a slight modifi-

cation of the updating equation in the algorithm suggested by Elliott et al. (2015) corresponds to an S-MD update when the mirror map is set to be equal to the negative entropy. Second, we define an *approximate least-favorable distribution* and show that S-MD provably obtains an approximate least favorable distribution (Theorem 2). Finally, we define a *nearly optimal test* and show that the S-MD routine can be used to generate such a test (Theorem 3).

3.1 Stochastic Mirror Descent

This section follows as closely as possible the notation in Sections 4.1 and 6.1 of Bubeck (2015). Let $\mathbb{R}^M_{++}$ denote the set of all strictly positive vectors in $\mathbb{R}^M$. We say that a map $\Phi : \mathbb{R}^M_{++} \to \mathbb{R}$ is a mirror map if it satisfies the following properties

i) Φ is strictly convex and differentiable.

ii) The gradient of Φ takes all possible values, that is $\nabla\Phi(\mathbb{R}^M_{++}) = \mathbb{R}^M$.

iii) The gradient of Φ diverges on the boundary of $\mathbb{R}^M_{++}$.

Mirror Maps are used to build iterative algorithms for constrained optimization problems when unbiased estimators of the gradient are available. More precisely, consider the optimization problem

$$\inf_{\kappa \in \mathcal{X}} f(\kappa),$$

where $f : \mathcal{X} \to \mathbb{R}$ is a convex function and $\mathcal{X} \subseteq \mathbb{R}^M_+$. Suppose that $\widehat{G}(\kappa)$ is an unbiased estimator of a subgradient of f at κ, in the sense that $\mathbb{E}\left[\widehat{G}(\kappa)\right]$ is a subgradient of the function f at κ.[7] Let $D_\Phi(\kappa, \kappa')$ denote the Bregman divergence associated with Φ, that is

$$D_\Phi(\kappa, \kappa') \equiv \Phi(\kappa) - \Phi(\kappa') - \nabla\Phi(\kappa')(\kappa - \kappa').$$

[7]The expectation should be understood as being conditional on κ, since κ is stochastic. See Chapter 6 in Bubeck (2015).

The stochastic mirror descent algorithm (henceforth, S-MD) given the mirror map φ is defined as follows:

Algorithm 1 Stochastic Mirror Descent with mirror map Φ, stopped after T epochs.

1: **Input:** Step size $\eta > 0$, number of epochs $T \in \mathbb{N}$.
2: Initialize $\kappa_1 \in \arg\min_{\kappa \in \mathcal{X} \cap \mathbb{R}^M_{++}} \Phi(\kappa)$.
3: **for** $t = 1, \ldots, T-1$ **do**
4:

$$\kappa_{t+1} = \arg \min_{\kappa \in \mathcal{X} \cap \mathbb{R}^M_{++}} \eta \left(\widehat{G}(\kappa_t)^\top \kappa \right) + D_\Phi(\kappa, \kappa_t). \tag{9}$$

5: **end for**
6: **Output:** $\bar{\kappa}_T \equiv \frac{1}{T} \sum_{t=1}^{T} \kappa_t$.

The general interpretation of the S-MD update in equation (9) is that *"the method is trying to minimize the local linearization of the function while not moving too far away from the previous point, with distances measured via the Bregman divergence of the mirror map"*; see p. 301 in Bubeck (2015).

An important observation regarding Algorithm 1 is that its output is the *average value* of κ_t over all the iterations, and not its last value. Averaging the trajectories of a stochastic optimization routine is commonly referred to as Polyak-Ruppert averaging; see Ruppert (1988) and Polyak and Juditsky (1992). This idea goes back to seminal work on mirror descent by Nemirovski and Yudin (1983).

The following theorem shows that a slight modification of the updating equation in the algorithm suggested by Elliott et al. (2015) (see their Equation 10, p. 782) corresponds to an S-MD update when the mirror map is set to be equal to the negative entropy.

Theorem 1. *Consider the optimization problem* $\inf_{\kappa \in \mathcal{X}} f(\kappa)$*, where* $f(\cdot)$ *is defined in* (6) *and the set* $\mathcal{X}$ *is defined in* (8)*.*

1. *Let* κ_t *be a realization of an arbitrary* $\mathcal{X}$*-valued random vector. For each* $m = 1, \ldots, M$*, let* $Y_{m,1}, ..., Y_{m,N}$ *be i.i.d. random variables with distribution* $Y \sim f_m$ *sampled independently of*

the realized value of κ_t*. For any* $N \geq 1$

$$\widehat{G}_N(\kappa_t) \equiv -\left(\frac{1}{N}\sum_{n=1}^{N}\varphi_{\kappa_t}(Y_{1,n}) - \alpha, ..., \frac{1}{N}\sum_{n=1}^{N}\varphi_{\kappa_t}(Y_{M,n}) - \alpha\right)^{\top}, \tag{10}$$

is an unbiased estimator of the subgradient of f *at* κ_t*; where* φ_{κ_t} *is a test of the Neyman-Pearson form defined in* (7)*.*

2. *The stochastic mirror descent update in Algorithm 1 based on* $\widehat{G}_N(\cdot)$ *and the mirror map* $\Phi(\kappa) = \sum_{m=1}^{M}\kappa_m \ln(\kappa_m)$ *is*

$$\kappa_{t+1,m} = c_t \cdot \kappa_{t,m} \exp\left(-\eta \cdot \widehat{G}_{m,N}(\kappa_t)\right), \tag{11}$$

where $\widehat{G}_{m,N}(\kappa_t)$ *is the* m*-th coordinate of* $\widehat{G}_N(\kappa_t)$ *and*

$$c_t \equiv \min\left\{1, \frac{1}{\alpha\sum_{m=1}^{M}\kappa_{t,m}\exp\left(-\eta \cdot \widehat{G}_{m,N}(\kappa_t)\right)}\right\}.$$

3. *The initial condition in Algorithm 1 based on the mirror map* $\Phi(\kappa) = \sum_{m=1}^{M}\kappa_m \ln(\kappa_m)$ *is*

$$\kappa_1 = \begin{cases} \left(\frac{1}{\exp(1)}, \ldots, \frac{1}{\exp(1)}\right), & \text{if } 1 \leq M < \frac{\exp(1)}{\alpha}, \\ \left(\frac{1}{\alpha M}, \ldots, \frac{1}{\alpha M}\right) & \text{if } M \geq \frac{\exp(1)}{\alpha}. \end{cases} \tag{12}$$

Proof. See Section A.3 in Appendix A. □

It is useful to make an explicit connection between the updating formula in (11) and Equation 10, p. 782 in Elliott et al. (2015). Following the notation in Elliott et al. (2015), define

$$\mu_m^{t+1} \equiv \ln(\kappa_{t+1,m}).$$

Taking logarithms on both sides of (11) and using the definition of $\widehat{G}_N(\kappa_t)$ yields

$$\mu_m^{t+1} = \ln(c_t) + \mu_m^t + \eta \left(\frac{1}{N} \sum_{n=1}^{N} \varphi_{\exp(\mu^t)}(Y_{m,n}) - \alpha \right). \tag{13}$$

When $c_t = 1$, the term $\ln(c_t) = 0$, and thus, (13) essentially matches Equation 10, p. 782 in Elliott et al. (2015) after noting that $\widehat{G}_{m,N}(\kappa_t)$ is a Monte Carlo estimate of the negative excess rate of Type I error the test φ_{κ_t}:

$$\alpha - \int \varphi_{\ln(\mu_t)} f_m d\nu.$$

Other than notation, the main difference between our expressions is the presence of the additional term c_t. In the stochastic mirror descent routine, this term is used to take into account the fact that—because of our Lemma 2—the optimization domain in the dual problem (5) can be restricted to values of κ such that $\sum_{m=1}^{M} \kappa_m \leq 1/\alpha$.

3.2 Approximate Solutions to the Dual Problem

In this subsection we show that if the number of epochs (T) and the step size (η) are chosen appropriately, then $\bar{\kappa}_T \equiv (1/T) \sum_{t=1}^{T} \kappa_t$—obtained using Equations 11 and 12 in Theorem 1—indeed can be used to generate an approximate least-favorable distribution for the problem in (1).

In order to formalize this statement, note that we have defined $\bar{v}$ as the value of the dual problem $\inf_{\kappa \in \mathcal{X}} f(\kappa)$, where $f(\cdot)$ is defined in (6) and the set $\mathcal{X}$ is defined in (8). Let

$$\Delta^{M-1} \equiv \left\{ \lambda \in \mathbb{R}^M \mid \lambda_m \geq 0 \text{ for all } m = 1, \ldots, M \quad \text{and} \quad \sum_{m=1}^{M} \lambda_m = 1 \right\} \tag{14}$$

denote the probability simplex in $\mathbb{R}^M$.

Definition 1 (ϵ-least favorable distribution)**.** We say that a vector $\lambda_\epsilon^* \in \Delta^{M-1}$ is an *ϵ-least favorable*

distribution if there exists a positive constant cv_ϵ^* such that

$$\kappa_\epsilon^* \equiv \text{cv}_\epsilon^* \cdot \lambda_\epsilon^*$$

satisfies

$$f(\kappa_\epsilon^*) \leq \bar{v} + \epsilon, \tag{15}$$

where $f(\cdot)$ is defined in (6). We say that a vector $\lambda^* \in \Delta^{M-1}$ is an *approximate least-favorable distribution* if there exists $\epsilon > 0$ for which λ^* is ϵ-least favorable.

Remark 3. The definition presented above is different from the notion of "ϵ-approximate least favorable distribution" used by Elliott et al. (2015); see their Definition 1, p. 780. Their definition focuses on the statistical properties of the Neyman-Pearson test associated with the least-favorable distribution. In contrast, we focus on the optimization problem that defines such a least-favorable distribution: the dual problem we presented in (5). Our definition is inspired by a large literature in theoretical computer science, operations research, and optimization where it is a common approach to *"to relax the requirement of finding an optimal solution, and instead settle for a solution that is good enough"* (Williamson and Shmoys, 2011, p. 14). There are different criteria that can be used to formalize the statement that an approximate solution is "good enough", but a typical choice in optimization problems relies on its value function (which is the metric we use in our Definition 1). We deliberately chose an additive approximation error, because most of the results that we are familiar with regarding the approximation error of mirror descent routines—and, more generally, first-order methods for convex optimization problems—take this form; see, for example, Section 5.1.1 in Juditsky and Nemirovski (2011) and also Bubeck (2015). But it is also possible to give results for multiplicative approximation errors; for example, see Theorem 1 in Chen, Lucier, Singer, and Syrgkanis (2017).

We now present a result showing that Algorithm 1 provably generates an approximate least-favorable distribution. For any nonnegative real number x, let $\lceil x \rceil$ denote the "ceiling function";

that is smallest integer larger than x.

Theorem 2. *Consider the optimization problem* $\inf_{\kappa\in\mathcal{X}} f(\kappa)$*, where* $f(\cdot)$ *is defined in* (6) *and the set* $\mathcal{X}$ *is defined in* (8)*. Let* $\overline{\kappa}_T$ *be the output of Algorithm 1 based on the mirror map* $\Phi(\kappa) = \sum_{m=1}^{M} \kappa_m \ln(\kappa_m)$ *and the unbiased estimator of the gradient* $\widehat{G}_N(\cdot)$ *defined in Equation 10 of Theorem 1. If* $\alpha \in (0, 1/2)$ *and* $M > \exp(1)/\alpha$*,*

$$T = \left\lceil \frac{4(1-\alpha)^2}{\alpha^2\epsilon^2} \cdot \ln(M) \right\rceil, \quad \textit{and} \quad \eta = \alpha \cdot \frac{\epsilon}{2(1-\alpha)^2}, \tag{16}$$

then

$$\lambda_T^* \equiv \frac{\overline{\kappa}_T}{\sum_{m=1}^{M} \overline{\kappa}_{T,m}} \tag{17}$$

is a $\left(1 + \frac{2\Omega}{\sqrt{\ln(M)N(1-\alpha)^2}}\right)\epsilon$*-least favorable distribution, in the sense of our Definition 1, with probability at least* $1 - \exp(-\Omega^2)$*.*

Proof. See Section A.4 in Appendix A. □

Theorem 2 shows that—even after finitely many iterations—the S-MD routine for the dual problem,

$$\inf_{\kappa\in\mathcal{X}} f(\kappa),$$

generates an approximate least-favorable distribution (in the sense of our Definition 1) with high probability. The approximate least favorable distribution in (17) is the *direction* of the multipliers $\overline{\kappa}_T$ (in analogy to what we would do if we had access to the exact solution of the dual problem). The probabilistic statement in the theorem arises due to the randomness in the gradient estimator in (10), which makes the output of the S-MD routine behave as a random variable. Note that if we fix the frequency at which we would like to obtain an approximate least-favorable distribution (over different *runs* of the S-MD routine), then the number of draws used to construct the gradient estimator (N) determines how close we get to finding a "good enough" solution for the dual problem.

For example, suppose that $\epsilon = .1$, $\alpha = 10\%$ and $M = 200$, and suppose we set $\Omega = \sqrt{\ln(1/\alpha)}$, so that $1 - \exp(-\Omega^2) = 1 - \alpha = 90\%$. If we run the S-MD routine using $N = 1$ (only one draw per density f_m), then with probability 90% we will obtain a

$$\left(1 + 2\sqrt{\frac{\ln(1/\alpha)}{\ln(M)(1-\alpha)^2}}\right)\epsilon \approx 2.5\epsilon = .25$$

least-favorable distribution. If we use $N = 10$ (only ten draws per density) we get a

$$\left(1 + 2\sqrt{\frac{\ln(1/\alpha)}{\ln(M)\cdot 10\cdot(1-\alpha)^2}}\right)\epsilon \approx 1.5\epsilon = .15$$

least favorable distribution. If we use $N = 100$, with probability 90% we get a $1.15\epsilon = .11$-least favorable distribution. More generally, Theorem 2 shows that, for any target probability, we can always make N large enough to get as close as we would like to the desired approximation error ϵ.

In our view, the most surprising part of Theorem 2 is that even if the number of draws per density used to implement the S-MD routine are as low as $N = 1$, it is still possible to get an approximate least-favorable distribution that provides a non-trivial approximate solution to the dual problem in (5). For instance, in the example above, using only one draw per density yields an approximation error of $2.5\epsilon = .25$ with probability 90%. The approximation error of .25 should be interpreted as a worst-case guarantee that applies to any testing problem of the form (1). As we show in our illustrative example, the resulting approximation error can, in practice, be considerably smaller.

3.3 Nearly Optimal Tests via Stochastic Mirror Descent

Now that we have established that the S-MD routine in Algorithm 1 (with negative entropy as a mirror map) provably generates an approximate least-favorable distribution—in the sense of our Definition 1—we discuss the extent to which the S-MD routine can also be used to generate a *nearly*

optimal test.

Before presenting a formal definition of what we mean by a nearly optimal test, it is helpful to explain why it is not entirely trivial to translate the approximate least-favorable distribution in Equation 17 into a nearly optimal test. Let κ_T^* be the multipliers that we obtain after running the S-MD routine, with T and η defined as in Theorem 2. Consider the test of the Neyman-Pearson form, $\varphi_{\kappa_T^*}$, defined in (7) based on the multipliers κ_T^*.[8] Since the S-MD routine never explicitly tried to enforce *size* control, it is possible that the size of $\varphi_{\kappa_T^*}$ is strictly above the nominal level α. Mathematically, this happens because an approximate optimizer to the dual problem does not necessarily imply a feasible solution to the primal problem (let alone a nearly optimal one). This suggests that, when defining a nearly optimal test, it could be helpful to take into account i) possible violations of the required size; ii) as well as potential loss in power, relative to the optimal solution.

Let $\bar{v}$ be defined as value function of the dual problem in (5). As we show in Section B.2 of Appendix B, duality holds, and $\bar{v}$ equals the power of the most powerful test of size α.[9]

Definition 2. A statistical test $\varphi_{\epsilon,\delta}^\star : \mathcal{Y} \to [0,1]$ is said to be (ϵ,δ)-nearly optimal of size α if:

1. The size of $\varphi_{\epsilon,\delta}^\star$ is no larger than $\alpha(1+\delta)$,

$$\int \varphi_{\epsilon,\delta}^\star f_m d\nu \leq \alpha(1+\delta), \quad \text{for all } m = 1,\ldots,M.$$

2. The power of $\varphi_{\epsilon,\delta}^\star$ is at most ϵ away of the maximum power of a test of size α,

$$\int \varphi_{\epsilon,\delta}^\star g d\nu \geq \bar{v} - \epsilon.$$

[8]Note that κ_T^* can be decomposed into its direction λ_T^* as in Equation 17 and its norm $\text{cv}_T^* \equiv \sum_{m=1}^M \kappa_{T,m}^*$. Thus, the test in (7) rejects the null if and only if

$$g(y) > \text{cv}_T^* \sum_{m=1}^M \lambda_{T,m}^* f_m(y).$$

[9]For the sake of exposition, we deliberately write $\bar{v}$ instead of $\bar{v}(\alpha)$.

We say that test φ^* is nearly optimal of size α, if there exists $\epsilon, \delta > 0$ for which φ^* is (ϵ, δ)-nearly optimal.

The following theorem shows that Algorithm 1 can provably be used to generate a nearly optimal test.

Theorem 3. *Consider the optimization problem* $\inf_{\kappa \in \mathcal{X}} f(\kappa)$*, where* $f(\cdot)$ *is defined in* (6) *and the set* $\mathcal{X}$ *is defined in* (8)*. Let* $\{\kappa_t\}_{t=1}^T$ *be the sequence of multipliers generated by Algorithm 1 based on the mirror map* $\varphi(\kappa) = \sum_{m=1}^M \kappa_m \ln(\kappa_m)$ *and the unbiased estimator of the gradient* $\widehat{G}_N(\cdot)$ *defined in Equation 10 of Theorem 1. If* $M > \exp(1)/\alpha$*,*

$$T = \left\lceil \frac{4(1-\alpha)^2}{\alpha^2 \epsilon^2} \cdot \ln(M) \right\rceil, \quad \textit{and} \quad \eta = \alpha \cdot \frac{\epsilon}{2(1-\alpha)^2},$$

then the test

$$\bar{\varphi}_T(y) \equiv \frac{1}{T} \sum_{t=1}^T \varphi_{\kappa_t}(y) \tag{18}$$

is nearly optimal of size α *with high probability (where* φ_{κ_t} *is the test of the Neyman-Pearson form in* (7)*). More concretely, with probability at least* $1 - \exp(-\Omega^2)$

1. *For any* $m = 1, \ldots, M$*,*

$$\int \bar{\varphi}_T f_m d\nu \leq \alpha \left(1 + \left[1 + \frac{2\Omega}{\sqrt{\ln(M) N (1-\alpha)^2}} \right] \epsilon - \frac{1}{T} \sum_{t=1}^T \widehat{G}_N(\kappa_t)^\top \kappa_t \right). \tag{19}$$

2. *The power of* $\bar{\varphi}_T$ *is larger than*

$$\bar{v} - \left[1 + \frac{2\Omega}{\sqrt{\ln(M) N (1-\alpha)^2}} \right] \epsilon. \tag{20}$$

Proof. See Section A.5 in Appendix A. □

Theorem 3 shows that the S-MD routine analyzed in this paper provably generates a nearly optimal

test, in the sense of our Definition 2. There are three aspects about our result that are worth highlighting.

First, it is rather surprising that the nearly optimal test in (18) takes the form of an "average" test. In order to get some intuition of why this construction is helpful to obtain theoretical results, it is useful to explicitly write the dual in (5) as the *minimax* problem

$$\min_{\kappa \in \mathbb{R}^M_+} \max_{\varphi} L(\varphi, \kappa),$$

where $L(\varphi, \kappa)$ is the Lagrangian function defined in (4). In this problem, the "min" player is choosing a vector of (Lagrange) multipliers, and the "max" player is choosing a test. For a fixed κ, the *best response* of the max player is a test of the Neyman-Pearson form φ_κ defined in (7). A mirror descent routine for this problem initializes the choice of κ by the "min" player, and iteratively updates its values based on the (sub)gradient of the Lagrangian with respect to κ_t, which—by results analogous to the envelope theorem—will give the rates of Type I error of φ_{κ_t}. The most powerful test of size α is the solution to the *maximin* problem

$$\max_{\varphi} \min_{\kappa \in \mathbb{R}^M_+} L(\varphi, \kappa).$$

The question is how to translate the iterates, $\{\kappa_t\}_{t=1}^T$, into a solution to the maximin problem. This question is common in the application of the mirror descent algorithm to minimax problems that arise in game theory and statistical decision theory; see Aradillas Fernández et al. (2025) and the discussion of matrix games in Arora, Hazan, and Kale (2012). While these papers consider different problems to the one studied in this paper, a suggestion therein is to use the best responses of the max player $\{\varphi_{\kappa_t}\}_{t=1}^T$ and randomize over them with uniform probability. In our problem, such a construction becomes the average test in (18); and this is what motivated us to study its performance. To the best of our knowledge, our results have not been stated elsewhere.

Second, the upper bound on the rate of Type I error features a term whose value changes with

each specific run of the S-MD routine. This is not ideal, but we were not able to derive a better bound. To better understand the role of this term, consider again the example we discussed after Theorem 2. Suppose that $\epsilon = .1$, $\alpha = 10\%$ and $M = 200$, and suppose we set $\Omega = \sqrt{\ln(1/\alpha)}$, so that $1 - \exp(-\Omega^2) = 1 - \alpha = 90\%$. If we run the S-MD routine using $N = 1$ (only one draw per density f_m),

$$1 + \left[1 + 2\sqrt{\frac{\ln(1/\alpha)}{\ln(M)(1-\alpha)^2}}\right]\epsilon \approx 1 + 2.5\epsilon = 1.25.$$

If the term

$$-\frac{1}{T}\sum_{t=1}^{T} \widehat{G}_N(\kappa_t)^\top \kappa_t, \tag{21}$$

were not part of (19), then we could conclude that with probability at least 90%, the test $\bar{\varphi}_T$ has size of at most 12.5% (that is, there is a size distortion of 2.5%) and power that is no less than $\bar{v}$ minus 25 percent points. Again, making N arbitrarily large makes the size closer to $\alpha(1+\epsilon)$ and the power at least $\bar{v} - \epsilon$. The interpretation of Theorem 3 changes slightly when we incorporate (21). Suppose for example that in one run of the S-MD routine the term in (21) equals .05. Then, for that run, our best hope is that (19) is satisfied with the larger bound 1.3 instead of 1.25. This means that we could see a rate or Type I error of the average test as high as 13%. As we discuss in the next section, in our illustrative example the term in (21) tends to be small (and negative), but we do not have any theoretical guarantees for this.[10]

Third, to report the entire test (as function of all possible data), one would have to retain the history of multipliers obtained from the S-MD routine. When both M and the number of iterations of S-MD are large, this could come with significant computational and data storage expense. However, a typical use case is to perform the test on a specific data set. For this purpose, data storage requirements can be much reduced because, rather than retaining the weights for every epoch, it suffices to store the implied test results, i.e. one bit per epoch and parameter value being

[10]The terms $-\widehat{G}_N(\kappa_t)^\top \kappa_t$ is a Monte-Carlo estimate of the average excess rate of Type I error of φ_{κ_t}, evaluated at the different null densities. The term in (21) averages these Monte-Carlo estimates over all iterations.

tested (and even less if one is content with only updating the average). Furthermore, rather than reporting a rejection probability, it usually suffices to either accept or reject, although in cases where the rejection probability is interior, this decision will then be random conditionally on the data. But this can be achieved at much lower computational expense: (ii) By a simple application of the Law of Iterated Expectations, rejecting with probability corresponding to the average test is equivalent to first drawing a "realized epoch" $t^* \sim \text{unif}(\{1, \ldots, T\})$ and then executing the Neyman-Pearson test for epoch t^* only. (iii) As t^* can be drawn before starting the iteration, it is only necessary to execute iterations up to epoch t^*, threreby only executing $T/2$ iterations in expectation. By using these simplifications, it should be easy to execute the test.

An alternative to the average test is to simply report the Neyman-Pearson test in (7) associated with the multipliers $\bar{\kappa}_T$ obtained as the output of Algorithm 1 based on the mirror map $\Phi(\kappa) = \sum_{m=1}^{M} \kappa_m \ln(\kappa_m)$ and the unbiased estimator of the gradient $\widehat{G}_N(\cdot)$ defined in Equation 10 of Theorem 1. While it is challenging to analyze the size and power properties of this test for finite T, in Appendix B.3.1 we show that, under some conditions, as $T \to \infty$ the power of the test will converge to $\bar{v}$, and it will have correct size (in a sense we make precise).

3.4 Remarks on Confidence Regions

It is common to construct confidence regions by inverting tests. However, the test advocated here is in general randomized, raising the question of how to invert it. Conceptual discussions of this matter go back at least to the 1950's (Stevens, 1950; Lehmann, 1959).[11] To summarize some key points, for this paragraph only let the test function $\rho(\cdot)$ also depend on the parameter value to be tested, i.e. we temporarily define $\rho : \mathcal{Y} \times \Theta \to [0, 1]$, where $\rho(y, \theta)$ is the probability of rejecting the instance of H_0 characterized by parameter value θ given data y; Lehmann (1959) calls this the *critical function*. Then we can define no less than four intervals that arguably invert our test:

[11] A notable difference to our setting is that these discussions were motivated by the randomized nature of optimal or exact tests in highly discrete sample spaces.

1. Lehmann (1959) defines a randomized confidence region as the set $\{\theta : \rho(y, \theta) \leq u\}$, where u is a realization of $U \sim \text{unif}(0, 1)$, reflecting data-independent randomization by the statistician.

2. Geyer and Meeden (2005) propose to directly report $1-\rho(y, \theta)$ as function of θ and to interpret it as membership function of a fuzzy set; it is easy to see that expected membership of the true parameter value will correspond to the target coverage.

3. Similarly to our discussion just above, one could draw a random epoch t^* and invert the corresponding (nonrandomized) Neyman-Pearson test.

4. Again similar to previous discussion, one could invert the Neyman-Pearson test that utilizes average weights $\bar{\kappa}_T$.

Mirroring discussions in the previous subsection, idea 4 is the computationally most involved and its justification is asymptotic, idea 3 will be the computationally easiest, and idea 1 is intermediate; 2 is computationally equivalent to 1. Some users might find 2 hard to interpret (Berger and Casella, 2005). We leave further analysis of the issue to future research.

4 Illustrative Example

In order to illustrate the performance of the SM-D routine in Algorithm 1—along with the implications of the theoretical guarantees in Theorem 2 and Theorem 3—we consider an elementary testing problem that arises in the context of the univariate Gaussian location model. More precisely, suppose we observe a realization of the random variable

$$Y \sim \mathcal{N}(\theta, 1).$$

The location parameter, θ, is unknown to the econometrician. Let $\Theta_0 \equiv \{\theta_{0,1}, \ldots, \theta_{0,M}\}$ be an equally-spaced grid over the interval $[-5, 0]$ consisting of $M = 200$ points. We order the elements

of Θ_0 in decreasing order, so that $\theta_{0,1} = 0$. We also define the singleton set $\Theta_1 \equiv \{\theta_1\}$.

We assume that the econometrician is interested in the following hypothesis testing problem:

$$\mathbf{H}_0 : \theta \in \Theta_0 \quad \text{vs.} \quad \mathbf{H}_1 : \theta \in \Theta_1.$$

It is well known that the most powerful test of size α for this problem—which we denote as φ^*_α—rejects the null if Y is large enough. More precisely, $\varphi^*_\alpha(Y) \equiv \mathbf{1}\{Y \geq z_{1-\alpha}\}$, where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal. The power of this test can then be expressed in terms of the normal c.d.f. as $\Pr(N(0,1) \leq \theta_1 - z_{1-\alpha})$. Since the most powerful test of size α for this example is known, we can use this information to analyze the theoretical guarantees we provided in Theorem 2 and Theorem 3.

The Stochastic Mirror Descent (S-MD) Routine: We first obtain a nearly optimal test of size $\alpha = 10\%$. We set $\theta_1 = 2$, which means that the largest power of a test of size $\alpha = 10\%$ is $\Phi(2 - 1.28) \approx 76.38\%$. This is also the value of the dual problem in (5). We set $\epsilon = .1$, and use the formulae in Theorem 2 to determine the maximum number of iterations (T) and the learning rate (η) for the S-MD routine:

$$T = \left\lceil \frac{4(1-\alpha)^2}{\alpha^2 \epsilon^2} \cdot \ln(M) \right\rceil = 171,666, \quad \eta = \alpha \cdot \frac{\epsilon}{2(1-\alpha)^2} = .0062. \tag{22}$$

In this example $M = 200 > (\exp(1)/\alpha) = \exp(1) \cdot 10$. Thus, in accordance with our theoretical derivations, the initial condition for the S-MD routine ($\kappa_0 \in \mathbb{R}^M$) is chosen to be:

$$\kappa_0 = (1/(\alpha M), \ldots, 1/(\alpha M))^\top = (.05, \ldots, .05)^\top.$$

The main component of the S-MD routine is the stochastic mirror descent update. We implement the unbiased estimator of the gradient in Theorem 1 using only one draw per density; that is, $N = 1$. More precisely, if we let $f_m(\cdot)$ denote the p.d.f. of Y under the null hypothesis $\theta_{0,m} \in \Theta_0$ and we let

$g(\cdot)$ be the p.d.f. of Y under the alternative Θ_1, the mirror descent update necessitates an unbiased estimator of the rates of Type I error of the test

$$\varphi_\kappa(Y) \equiv \mathbf{1}\left\{ g(Y) > \sum_{m=1}^{M} \kappa_m f_m(Y) \right\}. \tag{23}$$

In our example, an unbiased estimator for the rate of Type I error at $\theta_{0,m}$ can be succinctly obtained by sampling $Z \sim N(0,1)$ and using $\varphi_\kappa(Z+\theta_{0,m})$ as an estimator. More precisely, in each epoch t we obtain one draw $Z_t \sim \mathcal{N}(0,1)$ and compute the mirror descent update (coordinate by coordinate) as

$$\kappa_{m,t+1} \equiv \kappa_{m,t} \cdot \exp\left(\eta\left[\varphi_{\kappa_t}(Z_t + \theta_{0,m}) - \alpha\right]\right), \quad \text{for each } m = 1, \ldots, M.$$

The intuition of the update is very simple. If $\varphi_\kappa(Z + \theta_{0,m})$—the unbiased estimator of the rate of Type I error of φ_{κ_t} at $\theta_{0,m}$—is larger than α, then $\kappa_{m,t+1}$ increases (and otherwise decreases). We also know that $\|\kappa_t\|_1$ must be less than or equal than $1/\alpha$. Thus, after the update we check if $\|\kappa_{t+1}\| \leq 1/\alpha$. If this is the case, we keep κ_{t+1} as is; but otherwise we normalize κ_{t+1} to guarantee that $\sum_{m=1}^{M} \kappa_m \leq 1/\alpha$. This gives us back the update described in part 2 of Theorem 1.

In general, updating κ_t is numerically very cheap when $N = 1$, as it only involves obtaining one sample from each of the null densities (along with the evaluation of the null and alternative densities at each draw) and also evaluation of the exponential function. Using Matlab R2024a on a personal ASUS Vivobook Pro 15 @ 2.5GHz Intel Core Ultra 9 185H, it took around 280 seconds (slightly less than 5 minutes) to complete all of the $T = 171,666$ iterations.

Approximate Least-Favorable Distribution: As suggested by our Theorem 2, in order to construct an approximate least-favorable distribution we standardize the average value of κ_t to represent it as a probability distribution. More precisely, the blue bars in Figure 1 below correspond to

$$\lambda_T \equiv \frac{\bar{\kappa}_T}{\|\bar{\kappa}_T\|_1}, \quad \text{where } \bar{\kappa}_T \equiv \frac{1}{T}\sum_{t=1}^{T} \kappa_t.$$

In the testing problem we are considering, it is known that the least-favorable distribution loads all of its mass on $\theta_{0,1}$. As Figure 1 shows, the output of the S-MD routine resembles such distribution.

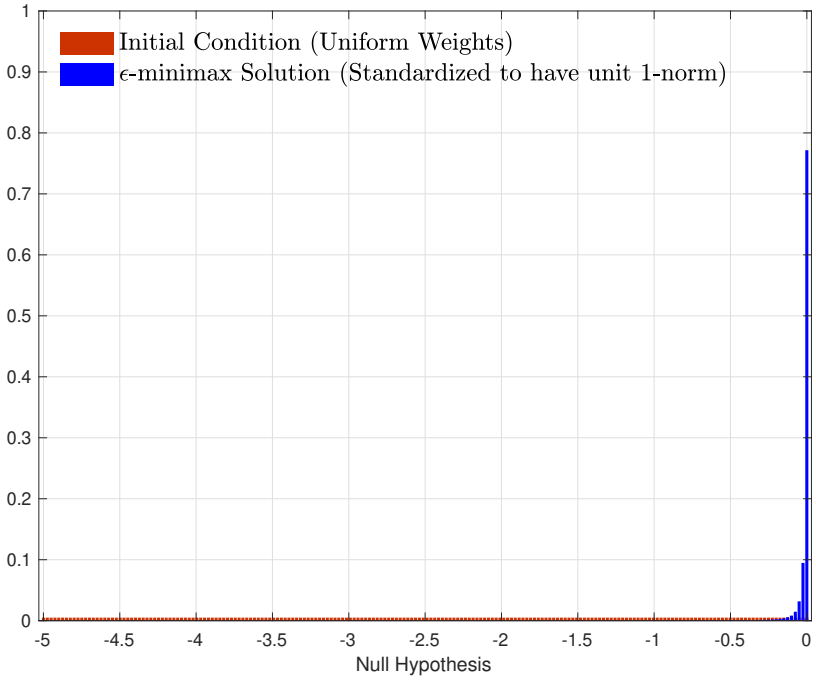

Figure 1: $\lambda_T \equiv \bar{\kappa}_T / \|\bar{\kappa}_T\|_1$ for $\alpha = 10\%$ and $\epsilon = 0.1$; where $\bar{\kappa}_T = (1/T) \sum_{t=1}^{T} \kappa_t$.

Our definition of approximate least-favorable distribution in Definition 1 makes reference to the value function of the dual problem in (5), and not in terms of its minimizer. In this example, it is easy to show that the value of the dual (which we denoted by $\bar{v}$) equals 76.38% (the power of the most powerful test of size $\alpha = 10\%$). We now argue that, as expected, the distribution λ_T approximately solves the dual problem, in the sense of Definition 1. To see this, we just need to evaluate the function:

$$f(\bar{\kappa}_T) = \underbrace{\int \varphi_{\bar{\kappa}_T} g(y) dy}_{\approx 76.96\%} - \underbrace{\sum_{m=1}^{M} \bar{\kappa}_{T,m} \left(\int \varphi_{\bar{\kappa}_T}(y) f_m(y) dy - \alpha \right)}_{-1.01\%} \approx 77.97\%,$$

where $\phi_{\bar{\kappa}_T}$ is the test of Neyman-Pearson form defined in (7), and where a Monte-Carlo approxima-

tion with 100,000 draws is used for the evaluation of each of the integrals. Thus, in this example:

$$f(\bar{\kappa}_T) \approx 77.97\% < \bar{v} + \epsilon = 76.38\% + 10\% = 86.38\%.$$

This means that in our run of the S-MD routine we obtained an $\epsilon = .016$-least-favorable distribution. The quality of the approximation is much better than what we expected based on our theoretical results in Theorem 2. This is consistent with the fact that the results in Theorem 2 apply to every possible testing problem with M null densities and a single alternative. We also conducted a Monte-Carlo simulation where we implemented the S-MD routine with different draws for the estimation of the subgradient, with 10,000 draws being used for integral evaluation in each. In all of the 100 runs we obtain a 10%-least favorable distribution. This is consistent that the theoretical results we presented in Theorem 2 apply to any testing problem of the form (1).

Nearly Optimal Test $\bar{\varphi}_T$: Figure 2 reports the test $\bar{\varphi}_T$ (red, solid line), which is the test defined in (18). For comparison, we also report the test $\varphi_{\bar{\kappa}_T}$ (blue, solid line). The size of $\bar{\varphi}_T$ is approximately 10.22%, and its power is approximately 76.98%. This means that the test $\bar{\varphi}_T$ is slightly over-sized, but it has competitive power.

We also conducted 100 different runs of our S-MD routine. In all of the 100 hundred runs the size was at most $\alpha(1+\epsilon)$ and the lowest power achieved was 76.82%.

Time Comparison of Using More Draws in the S-MD Routine: We also analyzed the increased computational effort of increasing the number of draws used to evaluate the subgradient during each iteration. To do this, we re-ran our illustrative example using 1, 10, 100, and 1,000 draws in each round. Table 1 shows the runtime as a function of the number of draws. Lastly, we also calculated the expected runtime associated with 20,000 draws in each round—the number of draws recommended by Elliott et al. (2015). Our expected runtime was 369,697 seconds.[12] We think these results illustrate the computational gains of implementing the S-MD routine with a small number

[12]This estimate was generated by taking the time difference between the first and second round, and then multiplying by T.

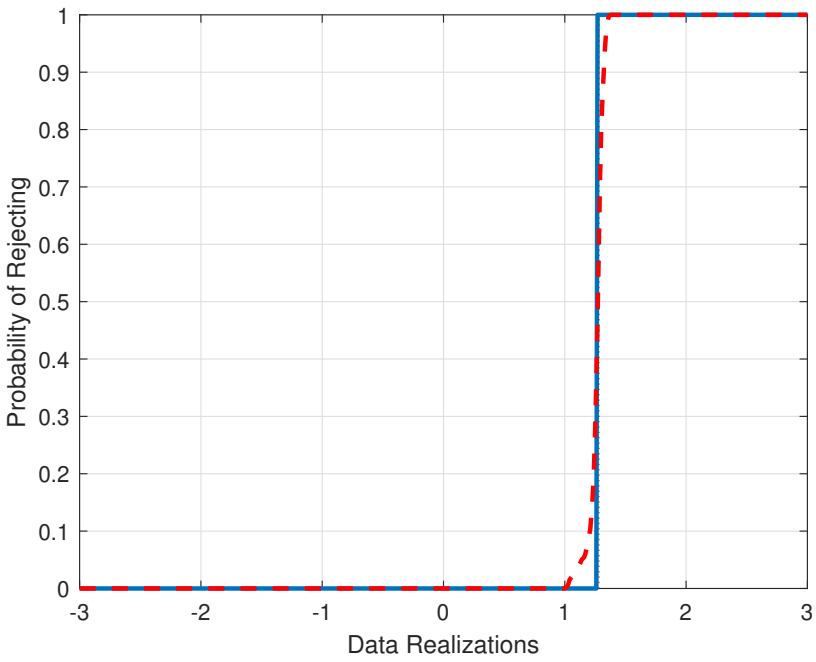

Figure 2: $\bar{\varphi}_T$ (red) alongside $\varphi_{\bar{\kappa}_T}$ (blue) for $\alpha = 10\%$ and $\epsilon = .1$.

Draws	Runtime (seconds)
1	314
10	882
100	4,028
1,000	36,320

Table 1: Number of draws in S-MD routine during each round versus runtime.

of draws to evaluate the subgradient.

5 Approximately Unbiased Tests

Consider now a variation of the testing problem in (1) where the alternative hypothesis is also composite, but with only I possible distributions for the data:

$$\mathbf{H}_0: \text{ the density of } Y \text{ is } f_m, \quad m = 1, \ldots, M, \quad \textit{vs.} \quad \mathbf{H}_1: \text{ the density of } Y \text{ is } g_i, \quad i = 1, \ldots, I. \tag{24}$$

As explained in Section 2.2 of Elliott et al. (2015), one can reduce the problem in (24) to the problem in (1) by choosing weights $w \equiv (w_1, \ldots, w_I) \in \Delta^{I-1}$ and defining $g \equiv \sum_{i=1}^{I} w_i g_i$. Then, the test φ that solves (3) can be interpreted as the test that maximizes w-weighted average power among all tests of size at most α.

A common criticism of tests that maximize a weighted average power criterion (henceforth, WAP) is that they can be *biased*: their power for some density g_i can be lower than α(Moreira and Moreira, 2013; Andrews, 2016). Moreira and Moreira (2013) note that one could include additional constraints in the problem (1) and consider:

$$\sup_{\varphi:\mathcal{Y}\to[0,1]} \int \varphi g d\nu, \quad \text{s.t.} \quad \int \varphi f_m d\nu \leq \alpha, \quad m = 1, ..., M, \quad \int \varphi g_i d\nu \geq \alpha, \quad i = 1, \ldots, I. \tag{25}$$

Just as before, we can define the Lagrangian function associated with problem (25) as

$$L(\varphi, \kappa, \mu) \equiv \int \varphi g d\nu - \sum_{m=1}^{M} \kappa_m \left[\int \varphi f_m d\nu - \alpha \right] - \sum_{i=1}^{I} \mu_i \left[\int \varphi(-g_i) d\nu + \alpha \right], \tag{26}$$

where we refer to $\kappa \equiv (\kappa_1, ..., \kappa_M) \in \mathbb{R}_+^M$ as the Lagrange multipliers associated with each of the inequality constraints that bound the test's rate of Type I error, and we let $\mu \equiv (\mu_1, ..., \mu_I) \in \mathbb{R}_+^I$ denote the Lagrange multipliers associated with each of the inequality constraints preventing the test to be biased.

We could then proceed as we did before and define the dual optimization problem:

$$\inf_{\kappa \in \mathbb{R}_+^M, \, \mu \in \mathbb{R}_+^I} f(\kappa, \mu), \tag{27}$$

where

$$f(\kappa, \mu) \equiv \sup_{\varphi:\mathcal{Y}\to[0,1]} L(\varphi, \kappa, \mu). \tag{28}$$

It is possible to show that the function $f(\kappa, \mu)$ is convex in its arguments. Moreover, a test φ that

achieves the maximum is the test

$$\varphi_{\kappa,\mu}(y) \equiv \begin{cases} 1 & \text{if } g(y) > \sum_{m=1}^{M} \kappa_m f_m(y) - \sum_{i=1}^{I} \mu_i g_i(y), \\ 0 & \text{if } g(y) \leq \sum_{m=1}^{M} \kappa_m f_m(y) - \sum_{i=1}^{I} \mu_i g_i(y), \end{cases} \tag{29}$$

and a subgradient of $f(\cdot)$ at (κ, μ) is

$$\nabla f(\kappa, \mu) \equiv -\left(\int \varphi_{\kappa,\mu} f_1 d\nu - \alpha, ..., \int \varphi_{\kappa,\mu} f_M \nu d\nu - \alpha, \int \varphi_{\kappa,\mu}(-g_1) d\nu + \alpha \ldots \int \varphi_{\kappa,\mu}(-g_I) d\nu + \alpha \right).$$

If $\nabla f(\kappa, \mu)$ were known, the mirror descent routine (with negative entropy as mirror map) for this problem would have updates

$$\begin{aligned} \kappa_{t+1,m} &= \kappa_{t,m} \exp\left(-\eta \cdot \nabla f(\kappa_t, \mu_t)\right), \\ \mu_{t+1,i} &= \mu_{t,i} \exp\left(-\eta \cdot \nabla f(\kappa_t, \mu_t)\right). \end{aligned}$$

Establishing a result similar to Theorem 2 and 3 is more challenging because the application of standard results would need ex-ante constraints on the $\|\cdot\|_1$-norm of κ and μ. But, for example, if one knew that the optimal values of κ and μ satisfied the constraint $\|\kappa + \mu\|_1 < 1/\alpha$ then our previous theoretical results for S-MD would apply.

6 Conclusion

We showed that—in testing problems where the null hypothesis postulates M distributions for the observed data—a slight variation of the algorithm in Elliott et al. (2015) coincides with a stochastic mirror descent routine for convex optimization. The convex program that arises naturally in the testing problem in (1) is the *dual* of the mathematical program that defines the *most powerful test* of *size* α.

We show that, given a desired approximation error, one can use a stochastic mirror descent routine to *provably* obtain—after finitely many iterations—both an approximate least-favorable distribution and a nearly optimal test.

Our theoretical results allowed us to provide concrete recommendations about the algorithm's implementation: including its initial condition, its step size, the number of iterations, and the number of stochastic draws per iteration that can be used to approximate the subgradient of the objective function. These practical recommendations have at least two important implications. First, the number of iterations used by the algorithm scales logarithmically in M (which means there is no theoretical sense in which the algorithm scales poorly as a function of the elements in the null hypothesis). Second, the algorithm can be implemented with a single stochastic draw per null density in each iteration (taking a larger number of draws improves the approximation error of the S-MD routine, but using a small number of draws reduces the computational burden of the algorithm).

References

Anderson, E. J. and P. Nash (1987): *Linear programming in infinite-dimensional spaces: theory and applications*, John Wiley & Sons.

Andrews, I. (2016): "Conditional linear combination tests for weakly identified models," *Econometrica*, 84, 2155–2182.

Aradillas Fernández, A., J. Blanchet, J. L. Montiel Olea, C. Qiu, J. Stoye, and L. Tan (2025): "Epsilon-Minimax Solutions of Statistical Decision Problems via the Hedge Algorithm," .

Arora, S., E. Hazan, and S. Kale (2012): "The multiplicative weights update method: a meta-algorithm and applications," *Theory of computing*, 8, 121–164.

Berger, R. L. and G. Casella (2005): "Comment: Fuzzy and Randomized Confidence Intervals and P-Values," *Statistical Science*, 20, 372–374.

Bubeck, S. (2015): "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, 8, 231–357.

Chamberlain, G. (2000): "Econometric applications of maxmin expected utility," *Journal of Applied Econometrics*, 15, 625–644.

Chen, R. S., B. Lucier, Y. Singer, and V. Syrgkanis (2017): "Robust optimization for non-convex objectives," *Advances in Neural Information Processing Systems*, 30.

Chiburis, R. (2008): "Approximately Most Powerful Tests for Moment Inequali ties, Unpublished manuscript," *Department of Economics, Princeton University*.

Cvitanic, J. and I. Karatzas (2001): "Generalized Neyman-Pearson lemma via convex duality," *Bernoulli*, 7, 79–97.

Dou, L. and U. K. Müller (2021): "Generalized Local-to-Unity Models," *Econometrica*, 89, 1825–1854.

Dudley, R. (2002): *Real Analysis and Probability*, vol. 74, Cambridge University Press.

Elliott, G. and U. K. Müller (2014): "Pre and post break parameter inference," *Journal of Econometrics*, 180, 141–157.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly optimal tests when a nuisance parameter is present under the null hypothesis," *Econometrica*, 83, 771–811.

Forneron, J.-J. (2024): "Estimation and inference by stochastic optimization," *Journal of Econometrics*, 238, 105638.

Geyer, C. J. and G. D. Meeden (2005): "Fuzzy and Randomized Confidence Intervals and P-Values," *Statistical Science*, 20, 358–366.

Guggenberger, P. and J. Huang (2025): "On the numerical approximation of minimax regret rules via fictitious play," *arXiv preprint arXiv:2503.10932*.

Guggenberger, P., F. Kleibergen, and S. Mavroeidis (2019): "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics*, 10, 487–526.

Juditsky, A. and A. Nemirovski (2011): "5 ct First-Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods," *Optimization for Machine Learning*, 121.

Krafft, O. and H. Witting (1967): "Optimale tests und ungünstigste verteilungen," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 7, 289–302.

Lehmann, E. L. (1959): *Testing Statistical Hypotheses*, Springer Texts in Statistics, New York: Springer, first ed.

Lehmann, E. L. and J. P. Romano (2005): *Testing statistical hypotheses*, Springer Texts in Statistics, New York: Springer, third ed.

Li, C. and U. K. Müller (2021): "Linear regression with many controls of limited explanatory power," *Quantitative Economics*, 12, 405–442.

Moreira, H. and M. J. Moreira (2013): *Contributions to the theory of optimal tests*, Citeseer.

——— (2019): "Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors," *Journal of Econometrics*, 213, 398–433.

Müller, U. K. (2025): "A more robust t-test," *Review of Economics and Statistics*, 107, 786–802.

Müller, U. K. and Y. Wang (2017): "Fixed-k asymptotic inference about tail properties," *Journal of the American Statistical Association*, 112, 1334–1343.

Müller, U. K. and M. W. Watson (2016): "Measuring uncertainty about long-run predictions," *Review of Economic Studies*, 83, 1711–1740.

——— (2018): "Long-run covariability," *Econometrica*, 86, 775–804.

——— (2020): "Low-frequency analysis of economic time series," *preparation. In Handbook of Econometrics. Elsevier.*

Muralidharan, K., M. Romero, and K. Wüthrich (2025): "Factorial designs, model selection, and (incorrect) inference in randomized experiments," *Review of Economics and Statistics*, 1–16.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009): "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, 19, 1574–1609.

Nemirovski, A. and D. Yudin (1983): *Problem Complexity and Method Efficiency in Optimization*, A Wiley-Interscience publication, Wiley.

Polyak, B. T. and A. B. Juditsky (1992): "Acceleration of stochastic approximation by averaging," *SIAM journal on control and optimization*, 30, 838–855.

Rudin, W. (2005): *Functional Analysis.*, International Series in Pure and Applied Mathematics, McGraw-Hill, New York.

Ruppert, D. (1988): "Efficient estimations from a slowly convergent Robbins-Monro process," Tech. rep., Cornell University Operations Research and Industrial Engineering.

Simons, S. (1995): *Minimax Theorems and Their Proofs*, Boston, MA: Springer US, 1–23.

Stein, E. M. and R. Shakarchi (2011): *Functional analysis: introduction to further topics in analysis*, vol. 4, Princeton University Press.

Stevens, W. L. (1950): "Fiducial Limits of the Parameter of a Discontinuous Distribution," *Biometrika*, 37, 117–129.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

Williamson, D. P. and D. B. Shmoys (2011): *The design of approximation algorithms*, Cambridge university press.

A Proofs of Main Results

A.1 Proof of Lemma 1

Proof. To prove convexity, note that by definition

$$f(\lambda\kappa + (1-\lambda)\kappa') = \sup_{\varphi} \int \varphi g d\nu - \sum_{m=1}^{M} (\lambda\kappa_m + (1-\lambda)\kappa'_m) \left[\int \varphi f_m d\nu - \alpha \right],$$

where we have slightly abused notation by omitting the fact that φ is allowed to be an arbitrary element of the space of all randomized tests. Consequently,

$$\begin{aligned} f(\lambda\kappa + (1-\lambda)\kappa') &\leq \lambda \sup_{\varphi} \left\{ \int \varphi g d\nu - \sum_{m=1}^{M} \kappa_m \left[\int \varphi f_m d\nu - \alpha \right] \right\} \\ &\quad + (1-\lambda) \sup_{\varphi} \left\{ \int \varphi g d\nu - \sum_{m=1}^{M} \kappa'_m \left[\int \varphi f_m d\nu - \alpha \right] \right\} \\ &= \lambda f(\kappa) + (1-\lambda) f(\kappa'). \end{aligned}$$

Therefore, f is convex in κ.

To show that $\nabla f(\kappa)$ is a subgradient of f at $\kappa \in \mathbb{R}^M_+$, we need to show that for any $\kappa' \in \mathbb{R}^M_+$

$$f(\kappa) \leq f(\kappa') + \nabla f(\kappa)(\kappa - \kappa').$$

Note first that the test φ_κ solves the problem

$$\sup_{\varphi: \mathcal{Y} \to [0,1]} \int \varphi g d\nu - \sum_{m=1}^{M} \kappa_m \left[\int \varphi f_m d\nu - \alpha \right]. \tag{30}$$

We can then rewrite (30) as

$$\sup_{\varphi: \mathcal{Y} \to [0,1]} \int \varphi \left(g - \sum_{m=1}^{M} \kappa_m f_m \right) d\nu + \alpha \sum_{m=1}^{M} \kappa_m.$$

Consequently,

$$\begin{aligned}
f(\kappa) &= \int \varphi_\kappa g d\nu - \sum_{m=1}^{M} \kappa_m \left[\int \varphi_\kappa f_m d\nu - \alpha \right] \\
&= \int \varphi_\kappa g d\nu - \sum_{m=1}^{M} (\kappa_m - \kappa'_m) \left[\int \varphi_\kappa f_m d\nu - \alpha \right] \\
&\quad - \sum_{m=1}^{M} \kappa'_m \left[\int \varphi_\kappa f_m d\nu - \alpha \right] \\
&= \int \varphi_k g d\nu - \sum_{m=1}^{M} \kappa'_m \left[\int \varphi_\kappa f_m d\nu - \alpha \right] + \nabla f(\kappa)(\kappa - \kappa') \\
&\leq f(\kappa') + \nabla f(\kappa)(\kappa - \kappa').
\end{aligned}$$

□

A.2 Proof of Lemma 2

Proof. Since $\mathcal{X} \subset \mathbb{R}_+^M$ we have

$$\inf_{\kappa \in \mathbb{R}_+^M} f(\kappa) \leq \inf_{\kappa \in \mathcal{X}} f(\kappa).$$

Thus, it is sufficient to show that

$$\inf_{\kappa \in \mathbb{R}_+^M} f(\kappa) \geq \inf_{\kappa \in \mathcal{X}} f(\kappa).$$

Suppose this is not the case and that

$$\bar{v} \equiv \inf_{\kappa \in \mathbb{R}_+^M} f(\kappa) < \inf_{\kappa \in \mathcal{X}} f(\kappa).$$

By definition of infimum, for any $\epsilon > 0$ there exists $\kappa_\epsilon \in \mathbb{R}_+^M$ such that

$$\bar{v} \leq f(\kappa_\epsilon) < \bar{v} + \epsilon.$$

By choosing ϵ small enough, we can guarantee the existence of an element $\kappa_\epsilon \in \mathbb{R}_+^M$ such that

$$f(\kappa_\epsilon) < \bar{v} + \epsilon < \inf_{\kappa \in \mathcal{X}} f(\kappa).$$

Since $\mathcal{X}$ and $\left(\mathbb{R}_+^M \backslash \mathcal{X}\right)$ form a partition of $\mathbb{R}_+^M$, it must be the case that either $\kappa_\epsilon \in \mathcal{X}$ or $\kappa_\epsilon \in \left(\mathbb{R}_+^M \backslash \mathcal{X}\right)$. Clearly, we cannot have $\kappa_\epsilon \in \mathcal{X}$ (as this would immediately yield a contradiction). Thus, we must have $\kappa_\epsilon \in \left(\mathbb{R}_+^M \backslash \mathcal{X}\right)$.

Note that at $\kappa = \mathbf{0}$,

$$f(\mathbf{0}) = \int \varphi_{\mathbf{0}} g d\nu \leq \int g d\nu = 1.$$

But also, for any κ such that $\|\kappa\|_1 > 1/\alpha$,

$$f(\kappa) = \sup_{\varphi: \mathcal{Y} \to [0,1]} \int \varphi \left[g - \sum_{m=1}^M \kappa_m f_m \right] d\nu + \alpha \sum_{m=1}^M \kappa_m \geq \sum_{m=1}^M \kappa_m \alpha > 1.$$

Therefore,

$$1 < f(\kappa_\epsilon) < \inf_{\kappa \in \mathcal{X}} f(\kappa) \leq f(\mathbf{0}) \leq 1.$$

This yields a contradiction. We conclude that $\bar{v} \equiv \inf_{\kappa \in \mathbb{R}_+^M} f(\kappa) = \inf_{\kappa \in \mathcal{X}} f(\kappa)$. □

A.3 Proof of Theorem 1

Proof of Part 1 of the Theorem: Let κ_t be a realization of an arbitrary $\mathcal{X}$-valued random vector. Let $\widehat{G}_{m,N}(\kappa_t)$ be the m-th coordinate of $\widehat{G}_N(\kappa_t)$. If suffices to show that if $(Y_{m,1}, \ldots, Y_{m,N})$ are i.i.d. random variables with distribution $Y \sim f_m$ sampled independently of the realized value of κ_t, then $\mathbb{E}[\widehat{G}_{m,N}(\kappa_t)|\kappa_t] = -\left(\int \varphi_\kappa f_m d\nu - \alpha\right)$.

By definition of $\widehat{G}_{m,N}(\kappa_t)$,

$$\begin{aligned}\mathbb{E}[\widehat{G}_{m,N}(\kappa_t)|\kappa_t] &= -\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\varphi_{\kappa_t}(Y_{m,n})\Big|\kappa_t\right]+\alpha \\ &= \mathbb{E}\left[\varphi_{\kappa_t}(Y_{m,n})\Big|\kappa_t\right]+\alpha,\end{aligned}$$

where the last line follows from the fact that $(Y_{m,1},\ldots,Y_{m,N})$ are i.i.d. according to f_m, independently of the value of κ_t. Since f_m is the p.d.f. of Y relative to the σ-finite measure ν, Problem 1, Chapter 5, p. 177 of Dudley (2002) implies

$$\mathbb{E}\left[\varphi_{\kappa_t}(Y_{m,n})\Big|\kappa_t\right]=\int\varphi_{\kappa_t}f_m d\nu.$$

Therefore, $\widehat{G}_N(\kappa_t)$ is an unbiased estimator of the subgradient of f at the realized κ_t.

Proof of Part 2 of the Theorem: Let $\widehat{G}_N(\kappa_t)$ be the unbiased estimator of the subgradient of f at κ_t. We provide an explicit solution for the problem

$$\kappa_{t+1}=\arg\min_{\kappa\in\mathcal{X}\cap\mathbb{R}^M_{++}}\eta\widehat{G}_N(\kappa_t)^\top\kappa+D_\Phi(\kappa,\kappa_t), \tag{31}$$

when $\Phi(\kappa)=\sum_{m=1}^M\kappa_m\ln(\kappa_m)$. By definition of Bregman divergence,

$$D_\Phi(\kappa,\kappa_t)=\Phi(\kappa)-\Phi(\kappa_t)-\langle\nabla\Phi(\kappa_t),\kappa-\kappa_t\rangle.$$

Consequently, κ_{t+1} is the solution to the following optimization problem

$$\min_{\kappa\in\mathbb{R}^M_{++}}\eta\widehat{G}_N(\kappa_t)^\top\kappa+\sum_{m=1}^M\kappa_m\ln\left(\kappa_m/\kappa_{t,m}\right)-\sum_{m=1}^M(\kappa_m-\kappa_{t,m}), \tag{32}$$

subject to the constraint

$$\sum_{m=1}^M\kappa_m\le 1/\alpha.$$

Let μ denote the Lagrange multiplier associated with this constraint. Thus, the first-order conditions of the problem for each $\kappa_{t+1,m}$ become:

$$\eta \widehat{G}_{m,N}(\kappa_t) + \ln(\kappa_{t+1,m}/\kappa_{t,m}) + \mu = 0, \tag{33}$$

where $\widehat{G}_{m,N}(\kappa_t)$ is the m-th entry of $\widehat{G}_N(\kappa_t)$. The first-order condition in (33) can be written as

$$\kappa_{t+1,m} = \kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right) \exp\left(-\mu\right).$$

Two cases to consider. First, if

$$\sum_{m=1}^{M} \kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right) < 1/\alpha,$$

then $\mu = 0$ and

$$\kappa_{t+1,m} = \kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right). \tag{34}$$

Second, if

$$\sum_{m=1}^{M} \kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right) \geq 1/\alpha,$$

then $\mu > 0$ and $\sum_{m=1}^{M} \kappa_{t+1,m}$ must equal $1/\alpha$. Consequently,

$$\kappa_{t+1,m} = \frac{1}{\alpha} \cdot \frac{\kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right)}{\sum_{m=1}^{M} \kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right)}, \tag{35}$$

which can be achieved by setting

$$\mu = \ln\left(\sum_{m=1}^{M} \kappa_{t,m} \exp\left(-\eta \widehat{G}_{m,N}(\kappa_t)\right)\right).$$

PROOF OF PART 3 OF THE THEOREM: The initial condition κ_1 solves:

$$\min_{\kappa \in \mathbb{R}^M_{++}} \sum_{m=1}^{M} \kappa_m \ln(\kappa_m) \quad \text{s.t.} \quad \|\kappa\|_1 = \sum_{m=1}^{M} \kappa_m \leq 1/\alpha. \tag{36}$$

We re-parameterize this problem by defining

$$K \equiv \|\kappa\|_1, \quad p_m \equiv \kappa_m / K, \quad p = (p_1, \ldots, p_M)^\top.$$

Since $\kappa \in \mathbb{R}^M_{++}$, then $K > 0$ and $w_m > 0$ for all $m = 1, \ldots, M$. Moreover, if we denote by Δ^{M-1} the simplex in $\mathbb{R}^M$ and use $\text{int}\left(\Delta^{M-1}\right)$ to denote its interior, the optimization problem in (36) thus becomes the nested optimization problem

$$\min_{K>0} \left(\min_{p \in \text{int}(\Delta^{M-1})} K \left(\sum_{m=1}^{M} p_m \ln(p_m) \right) + K \ln(K) \right) \quad \text{s.t.} \quad K \leq 1/\alpha. \tag{37}$$

Thus, we first solve the inner problem which consists of finding the distribution in the simplex with the smallest negative entropy:

$$\min_{p \in \text{int}(\Delta^{M-1})} \sum_{m=1}^{M} p_m \ln(p_m).$$

It is known that the solution of this problem is to set $p_m = 1/M$. We verify this below for the sake of exposition. The first order conditions are

$$1 + \ln(p_m) + \mu = 0,$$

where μ is the Lagrange multiplier associated with $\|p\|_1 = 1$. Solving for p_m yields

$$p_m = \exp\left(-(1 + \mu)\right)$$

which implies (by summing the left side over $m = 1, \ldots, M$):

$$\frac{1}{M} = \exp\left(-(1+\mu)\right).$$

We conclude that $p_m^* = 1/M$ is the optimal direction of κ_1. We now find its scale by solving the outer optimization problem

$$\min_{K>0} K\left(\sum_{m=1}^{M} \frac{1}{M} \ln\left(\frac{1}{M}\right)\right) + K\ln(K) = \min_{K>0} K(\ln(K) - \ln(M)) \quad \text{s.t.} \quad K \leq 1/\alpha. \tag{38}$$

Without the constraint, the objective function has a global minimum at K^* satisfying

$$\ln(K^*) + 1 - \ln(M) = 0,$$

or equivalently, $K^* = M/\exp(1)$. It is also decreasing for $K < K^*$ and increasing for larger values. Therefore, the solution to the problem in (38) is

$$K^* = \begin{cases} \frac{M}{\exp(1)} & \text{if } 1 \leq M < \frac{\exp(1)}{\alpha} \\ \frac{1}{\alpha} & \text{if } M \geq \frac{\exp(1)}{\alpha}. \end{cases},$$

Thus, the initial condition is

$$\kappa_1 = \begin{cases} \left(\frac{1}{\exp(1)}, \ldots, \frac{1}{\exp(1)}\right) & \text{if } 1 \leq M < \frac{\exp(1)}{\alpha}, \\ \left(\frac{1}{M\alpha}, \ldots, \frac{1}{M\alpha}\right) & \text{if } M > \frac{\exp(1)}{\alpha}. \end{cases}$$

A.4 Proof of Theorem 2

The approximation error of the numerical iteration consists of two parts: optimization error and estimation error. The former is intrinsic to the optimization algorithm when applying the exact

(sub)gradient, while latter is induced by the estimation error of the unknown subgradient.

Proof. By Lemma 1, the function $f(\cdot)$ in the dual problem (5) is convex. Consequently, for any $\kappa \in \mathcal{X}$,

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\kappa_{t,m}\right) - f(\kappa) \leq \frac{1}{T}\sum_{t=1}^{T} f(\kappa_{t,m}) - f(\kappa). \tag{39}$$

Note under the S-MD routine of Algorithm 1, κ_t is a random variable. Part 1 of Theorem 1 showed that, given the realized value of κ_t, $\widehat{G}_N(\kappa_t)$ is an unbiased estimator of the subgradient of f at κ_t; that is, $\mathbb{E}\left[\widehat{G}_N(\kappa_t)\right] = \nabla f(\kappa_t)^\top$. Consequently, Equation (39) and the definition of subgradient imply

$$\begin{aligned} f\left(\frac{1}{T}\sum_{t=1}^{T}\kappa_{t,m}\right) - f(\kappa) &\leq \frac{1}{T}\sum_{t=1}^{T}\nabla f(\kappa_t)\left(\kappa_t - \kappa\right) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\nabla f(\kappa_t)^\top - \widehat{G}_N(\kappa_t)\right)^\top\left(\kappa_t - \kappa\right) \qquad (40) \\ &+ \frac{1}{T}\sum_{t=1}^{T}\widehat{G}_N(\kappa_t)^\top\left(\kappa_t - \kappa\right). \qquad (41) \end{aligned}$$

It follows by Bubeck (2015, proof of Theorem 4.2 and Equation (10) on p.307) that (41) is bounded above by

$$\frac{D_\varphi(\kappa, \kappa_1)}{\eta T} + \frac{\eta}{2\rho}\frac{1}{T}\sum_{t=1}^{T}\|\widehat{G}_N(\kappa_t)\|_\infty^2,$$

where ρ is the parameter of the convexity of φ with respect to $\|\cdot\|_1$. We have already proved in Lemma 4 that $\rho = \alpha/2$. Additionally, $D_\varphi(\kappa, \kappa_1) \leq \ln(M)/\alpha$. [13] Accordingly, (41) is bounded above by

[13] When $M > \frac{e}{\alpha}$, meaning the problem is high in dimension, then

$$\begin{aligned} R^2 &= \sup_{\kappa\in\mathcal{X}}\varphi(\kappa) - \varphi(\kappa_1) \\ &= \frac{1}{\alpha}\ln\left(\frac{1}{\alpha}\right) - \frac{1}{\alpha}\left(\ln\left(\frac{1}{\alpha}\right) - \ln(M)\right) = \frac{1}{\alpha}\ln(M). \end{aligned}$$

$$\frac{D_\varphi(\kappa,\kappa_1)}{\eta T}+\frac{\eta}{2\rho}\frac{1}{T}\sum_{t=1}^{T}\|\widehat{G}_N(\kappa_t)\|_\infty^2\le\frac{\ln(M)}{\alpha T\eta}+\frac{\eta(1-\alpha)^2}{\alpha},\tag{42}$$

where $\|\cdot\|_\infty$ is the sup norm, and where the last inequality follows from the fact that $\alpha<1/2$. Since

$$T=\left\lceil\frac{4(1-\alpha)^2}{\alpha^2\epsilon^2}\cdot\ln(M)\right\rceil,\quad\text{and}\quad\eta=\alpha\cdot\frac{\epsilon}{2(1-\alpha)^2},$$

we conclude that (41) is at most ϵ. Next, we upper bound the term (40). Define

$$\Delta_t\equiv\widehat{G}_N(\kappa_t)-\nabla f(\kappa_t).\tag{43}$$

Given t and κ_t, we write Δ_t as an average of N independent vectors, i.e.,

$$\Delta_t=\frac{1}{N}\sum_{n=1}^{N}\Delta_{t,n},$$
$$\text{where }\Delta_{t,n}\equiv\left(\varphi_{\kappa_t}(Y_{1,n}^{(t)})-\int\varphi_{\kappa_t}f_1d\nu,...,\varphi_{\kappa_t}(Y_{M,n}^{(t)})-\int\varphi_{\kappa_t}f_Md\nu\right),n=1,2,...,N.$$

Denote the $M\times N$ random vectors at time t as $Y_t\equiv(Y_{m,n}^{(t)})$, and let $\mathcal{F}_t=\sigma(Y_1,Y_2,...,Y_t)$ denote the canonical filtration of Y_t. From our iteration, κ_t is $\mathcal{F}_t$-predictable, i.e., $\sigma(\kappa_t)\subset\mathcal{F}_{t-1}$ for each t. Also note that $\|\kappa_t-\kappa^*\|_1\le 2/\alpha$ for any $\kappa\in\mathcal{X}$. Thus, applying Lemma 5 with $X_t=\kappa_t-\kappa$ and $L=2/\alpha$, we conclude that, for a given confidence level $\Omega>0$, (40) is upper bounded by

$$\frac{4\Omega}{\alpha\sqrt{TN}}=\frac{2\Omega\epsilon}{\sqrt{(1-\alpha)^2\ln(M)N}},$$

with probability at least $1-\exp(-\Omega^2)$. The conclusion follows from combining the upper bound for (41) and (40), and take $\kappa=\kappa^*$, the solution to the dual problem. □

A.5 Proof of Theorem 3

Proof. First, it is already derived in the proof for Theorem 2 that for any $\kappa \in \mathcal{X}$,

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{G}_N(\kappa_t)^\top(\kappa_t - \kappa) \le \frac{\ln(M)}{\alpha\eta T} + \frac{\eta(1-\alpha)^2}{\alpha} = \epsilon \tag{44}$$

For a given $\kappa \in \mathcal{X}$, apply Lemma 5 with $X_t = \kappa$ and $X_t = \kappa_t$, respectively. Notice that both $\|\kappa\|_1 \le 1/\alpha, \|\kappa_t\|_1 \le 1/\alpha$ hold, then

$$\Pr\left[\frac{1}{T}\sum_{t=1}^{T}(\widehat{G}_N(\kappa_t) - \nabla f(\kappa_t))^\top\kappa < \frac{2\Omega}{\alpha\sqrt{TN}}\right] \ge 1 - \exp(-\Omega^2),$$
$$\text{and } \Pr\left[\frac{1}{T}\sum_{t=1}^{T} -(\widehat{G}_N(\kappa_t) - \nabla f(\kappa_t))^\top\kappa_t < \frac{2\Omega}{\alpha\sqrt{TN}}\right] \ge 1 - \exp(-\Omega^2).$$

Combining with (44), we have

$$\Pr\left[\frac{1}{T}\sum_{t=1}^{T}\widehat{G}_N(\kappa_t)^\top\kappa_t - \frac{1}{T}\sum_{t=1}^{T}\nabla f(\kappa_t)^\top\kappa < \epsilon + \frac{2\Omega}{\alpha\sqrt{TN}}\right] \ge 1 - \exp(-\Omega^2), \tag{45}$$

and similarly,

$$\Pr\left[\frac{1}{T}\sum_{t=1}^{T}\nabla f(\kappa_t)^\top\kappa_t - \frac{1}{T}\sum_{t=1}^{T}\widehat{G}_N(\kappa_t)^\top\kappa < \epsilon + \frac{2\Omega}{\alpha\sqrt{TN}}\right] \ge 1 - \exp(-\Omega^2). \tag{46}$$

Note (45) implies a high probability bound for the Type I error: take $\kappa_m = 1/\alpha$ for $m = j$ and $\kappa_m = 0$ for other $m \ne j$. Then,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\nabla f(\kappa_t)^\top\kappa &= \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\kappa_m(\alpha - \int\varphi_t f_m d\nu) \\ &= \alpha\cdot\frac{1}{\alpha} - \frac{1}{T}\frac{1}{\alpha}\sum_{t=1}^{T}\int\varphi_t f_m d\nu = 1 - \frac{1}{\alpha}\int\bar{\varphi}f_m d\nu.\end{aligned}$$

Taking it back to (45), we have

$$\int \bar{\varphi} f_m d\nu \leq \alpha \left(1 - \frac{1}{T}\sum_{t=1}^{T} \widehat{G}_N(\kappa_t)^\top \kappa_t + \epsilon + \frac{2\Omega}{\alpha\sqrt{TN}}\right)$$

with probability at least $1 - \exp(-\Omega^2)$. The first statement follows from the arbitrariness of m.

For the second statement, note according to (46), we have

$$\frac{1}{T}\sum_{t=1}^{T} \nabla f(\kappa_t)^\top \kappa_t - \epsilon - \frac{2\Omega}{\alpha\sqrt{TN}} \leq \frac{1}{T}\sum_{t=1}^{T} \widehat{G}_N(\kappa_t)^\top \kappa$$

with probability at least $1 - \exp(-\Omega^2)$. Add the power of $\bar{\varphi}$, $\int \bar{\varphi} g d\nu$, to both sides. Notice that

$$\text{power}(\bar{\varphi}) + \frac{1}{T}\sum_{t=1}^{T} \nabla f(\kappa_t)^\top \kappa_t = \frac{1}{T}\sum_{t=1}^{T} f(\kappa_t) \geq \bar{v},$$

taking $\kappa = 0$ leads to

$$\text{power}(\bar{\varphi}) \geq \bar{v} - \epsilon - \frac{2\Omega}{\alpha\sqrt{TN}},$$

with probability at least $1 - \exp(-\Omega^2)$. This concludes the proof of the second statement of Theorem 3. □

Online Appendix

B Additional Technical Results

B.1 Additional Lemmas

Lemma 3. *The function* $\Phi : \mathbb{R}^M_{++} \to \mathbb{R}$ *given by* $\Phi(\kappa) = \sum_{m=1}^M \kappa_m \ln(\kappa_m)$ *is a mirror map.*

Proof. It is sufficient to verify the conditions i)-ii)-iii) given at the beginning of Section (3.1), which are taken from Section 4.1 in Bubeck (2015). We first verify i); namely that $\Phi(\cdot)$ is differentiable and strictly convex. The gradient of Φ at any $\kappa \in \mathbb{R}^M_{++}$ is:

$$\nabla\Phi(\kappa) = (1 + \ln(\kappa_1), ..., 1 + \ln(\kappa_M)).$$

Thus, Φ is differentiable in its domain. Moreover, for any $\kappa \in \mathbb{R}^M_{++}$ the Hessian takes the form

$$\begin{bmatrix} \frac{1}{\kappa_1} & 0 & ... & 0 \\ 0 & \frac{1}{\kappa_2} & ... & ... \\ ... & ... & ... & ... \\ 0 & ... & ... & \frac{1}{\kappa_M} \end{bmatrix},$$

which is a positive definite matrix. Therefore, $\Phi(\cdot)$ is strictly convex in its domain.

Next, we verify ii); namely, that $\nabla\Phi(\mathbb{R}^M_{++}) = \mathbb{R}^M$. Since $\ln(\mathbb{R}_{++}) = \mathbb{R}$, condition ii) holds.

Lastly, we verify condition iii), which in this case is equivalent to showing that for any κ^* with one or more entries equal to zero satisfies

$$\lim_{\kappa \to \kappa^*} ||\nabla\Phi(\mathbb{R}^M_{++})|| = \infty.$$

Note that, if the m-th entry of κ^* is zero, then

$$\lim_{x\to 0+} 1+\ln(x) = -\infty.$$

Therefore, condition iii) holds. Therefore, $\Phi(\cdot)$ is a mirror map. □

Lemma 4. *The function* $\Phi(\kappa) = \sum_{m=1}^{M} \kappa_m \ln(\kappa_m)$ *restricted on* $\mathcal{X}$ *is* $\frac{\alpha}{2}$*-strongly convex w.r.t.* $\|\cdot\|_1$.

Proof. We intend to prove, for any $\kappa_1, \kappa_2 \in \mathcal{X}$,

$$\Phi(\kappa_1) - \Phi(\kappa_2) - \langle \nabla\Phi(\kappa_2), \kappa_1 - \kappa_2\rangle \geq \frac{\alpha}{4}\|\kappa_1 - \kappa_2\|_1^2 \tag{47}$$

Define $K_1 \equiv \|\kappa_1\|_1, K_2 \equiv \|\kappa_2\|_1$, and $p_1 = \kappa_1/\|\kappa_1\|_1, p_2 = \kappa_2/\|\kappa_2\|_1$, we write

$$\kappa_1 = K_1 p_1, \kappa_2 = K_2 p_2.$$

Then, we decompose the left-hand side of (47) as

$$\Phi(\kappa_1) - \Phi(\kappa_2) - \langle \nabla\Phi(\kappa_2), \kappa_1 - \kappa_2\rangle = K_1 \ln\frac{K_1}{K_2} - (K_1 - K_2) + K_1 \sum_{m=1}^{M} p_{1,m} \ln\frac{p_{1,m}}{p_{2,m}}$$

Notice that:

1. the function $\Phi : (0, \frac{1}{\alpha}] \to \mathbb{R}$ defined by $\Phi(x) = x\ln(x)$ is α-strongly convex, so

$$K_1 \ln\frac{K_1}{K_2} - (K_1 - K_2) \geq \frac{\alpha}{2}|K_1 - K_2|^2.$$

2. p_1, p_2 are on the $(M-1)$-dimensional simplex. We can apply the Pinsker's inequality,

$$K_1 \sum_{m=1}^{M} p_{1,m} \ln\frac{p_{1,m}}{p_{2,m}} \geq \frac{K_1}{2}\|p_1 - p_2\|_1^2 \geq \frac{\alpha}{2}\|K_1(p_1 - p_2)\|_1^2.$$

Together we get

$$\begin{aligned}\Phi(\kappa_1) - \Phi(\kappa_2) - \langle \nabla\Phi(\kappa_2), \kappa_1 - \kappa_2\rangle &\geq \frac{\alpha}{2}|K_1 - K_2|^2 + \frac{\alpha}{2}\|K_1(p_1 - p_2)\|_1^2 \\ &= \frac{\alpha}{2}\left(\|K_1p_2 - K_2p_2\|_1^2 + \|K_1p_1 - K_1p_2\|_1^2\right) \\ &\geq \frac{\alpha}{4}\left(\|K_1p_2 - K_2p_2\|_1 + \|K_1p_1 - K_1p_2\|_1\right)^2 \\ &= \frac{\alpha}{4}\|K_1p_1 - K_2p_2\|_1^2.\end{aligned}$$

□

Lemma 5. *Suppose our unbiased estimator $\widehat{G}_N(\kappa_t)$ is evaluated on $M \times N$ independent draws in Y_t. Then, for any $\gamma > 0$ and any $\{X_t \in \mathbb{R}^M, t = 1, 2, ...\}$ that is $\mathcal{F}_t$-predictable, if there exists a constant L such that $\|X_t\|_1 \leq L$, we have*

$$\mathbb{E}\left[\exp(\frac{\gamma}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle)\right] \leq \exp\left(\frac{\gamma^2M^2}{TN}\right),$$

which leads to

$$\Pr\left[\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle \geq \delta\right] \leq \exp\left(-\frac{TN\delta^2}{4M^2}\right).$$

Moreover, for a given confidence level $\Omega > 0$, we have

$$\Pr\left[\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle \geq \frac{2L\Omega}{\sqrt{TN}}\right] \leq \exp(-\Omega^2).$$

Proof. First, because for any t, n, $\|\Delta_{t,n}\|_\infty \leq 1$ and $\|X_t\|_1 \leq L$,

$$\mathbb{E}\left[\exp(\frac{\langle X_t, \Delta_{t,n}\rangle^2}{L^2})|\mathcal{F}_{t-1}\right] \leq \exp(1)$$

Apply the same steps as in Nemirovski, Juditsky, Lan, and Shapiro (2009): if $0 < \gamma L \leq 1$, we apply

$e^x \leq x + e^{x^2}$ and Cauchy-Schwarz inequality, which yields

$$\mathbb{E}\big[\exp\left(\gamma\langle X_t, \Delta_{t,n}\rangle\right)|\mathcal{F}_{t-1}\big] \leq \mathbb{E}\big[\exp\left(\gamma^2\langle X_t, \Delta_{t,n}\rangle^2\right)|\mathcal{F}_{t-1}\big] \leq \exp\left(\gamma^2 L^2\right).$$

If $\gamma L > 1$, Cauchy-Schwarz inequality implies

$$\mathbb{E}\big[\exp\left(\gamma\langle X_t, \Delta_{t,n}\rangle\right)|\mathcal{F}_{t-1}\big] \leq \mathbb{E}\left[\exp\left(\gamma L\right)\big|\mathcal{F}_{t-1}\right] \leq \exp\left(\gamma^2 L^2\right).$$

Therefore, in both cases,

$$\mathbb{E}[\exp(\gamma\langle X_t, \Delta_{t,n}\rangle)|\mathcal{F}_{t-1}] \leq \exp(\gamma^2 L^2).$$

Because $\{\Delta_{t,n}, n = 1, 2, .., N\}$ are independent, we have

$$\mathbb{E}[\exp(\gamma\langle X_t, \Delta_t\rangle)|\mathcal{F}_{t-1}] = \Pi_{n=1}^{N}\,\mathbb{E}[\exp(\frac{\gamma}{N}\langle X_t, \Delta_{t,n}\rangle)|\mathcal{F}_{t-1}] \leq \exp\left(\frac{\gamma^2 L^2}{N}\right).$$

Applying Law of Iterated Expectations sequentially yields

$$\begin{aligned}
\mathbb{E}\left[\exp(\frac{\gamma}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle)\right] &= \mathbb{E}\left[\mathbb{E}[\exp(\frac{\gamma}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle)|\mathcal{F}_{T-1}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\exp(\frac{\gamma}{T}\langle X_T, \Delta_T\rangle)|\mathcal{F}_{T-1}] \cdot \exp(\frac{\gamma}{T}\sum_{t=1}^{T-1}\langle X_t, \Delta_t\rangle)]\right] \\
&\leq \mathbb{E}\left[\exp(\frac{\gamma^2 L^2}{T^2 N} + \frac{\gamma}{T}\sum_{t=1}^{T-1}\langle X_t, \Delta_t\rangle)\right] \\
&\leq ... \leq \exp\left(\frac{\gamma^2 L^2}{TN}\right).
\end{aligned}$$

It follows by Markov's inequality that

$$\Pr\left[\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle \geq \delta\right] \leq \frac{\mathbb{E}[\exp(\frac{\gamma}{T}\sum_{t=1}^{T}\langle X_t, \Delta_t\rangle)]}{\exp(\gamma\delta)} \leq \exp\left(\frac{\gamma^2 L^2}{TN} - \gamma\delta\right), \forall\gamma > 0.$$

For a given confidence level $\Omega > 0$, applying the above relation with $\gamma = \frac{TN\delta}{2L^2}$ abd $\delta = \frac{2L\Omega}{\sqrt{TN}}$ yields the desired conclusion. □

B.2 Duality Results

In this section we formalize the connection between the optimization problems (3) and (5). Throughout this section we assume that $(\mathcal{Y}, \mathcal{F}, \nu)$ is a separable measure space in the sense of Exercise 10, Chapter 1 in Stein and Shakarchi (2011).

Proposition 1. *i) There exists a test φ^* that solves* (3)*; that is, φ^* maximizes $\int \varphi g d\nu$ among all level-α tests. ii) Furthermore, there exists an optimizer κ^* to the dual problem* (5)*, and the value of the dual problem is finite. iii) Moreover, for any solution $\hat{\kappa}$ of the dual* (5)*, and for any test $\hat{\varphi}$ that solves* (3)*, the pair $(\hat{\varphi}, \hat{\kappa})$ satisfy:*

$$\begin{aligned} \hat{\varphi}(y) = 1, \text{ when } g(y) > \sum_{m=1}^{M} \hat{\kappa}_m f_m(y), \\ \hat{\varphi}(y) = 0, \text{ when } g(y) < \sum_{m=1}^{M} \hat{\kappa}_m f_m(y), \end{aligned} \tag{48}$$

and the complementary slackness,

$$\hat{\kappa}_m \left(\int \hat{\varphi} f_m d\nu - \alpha \right) = 0, \forall m \in [M]. \tag{49}$$

Proof. Let ν denote the σ-finite measure defined over the measurable space $(\mathcal{Y}, \mathcal{F})$. In a slight abuse of notation, denote by $L^\infty(\mathcal{Y})$ the set of essentially bounded real-valued measurable functions on $(\mathcal{Y}, \mathcal{F})$. Let $L^1(\mathcal{Y})$ be the space of all real-valued measurable functions $f : \mathcal{Y} \to \mathbb{R}$ that are integrable with respect to ν; that is $\int |f| d\nu < \infty$. Endow $L^\infty(\mathcal{Y})$ with the weak*-topology; see Rudin (2005), p. 67, 68. By definition, a sequence $\{\varphi_n\}_{n\in\mathbb{N}} \subseteq L^\infty(\mathcal{Y})$ converges to φ in the weak* topology if and only if

$$\int f \varphi_n d\nu \to \int f \varphi d\nu, \text{ for any } f \in L^1(\mathcal{Y}),$$

see p. 62-68 of Rudin (2005). It is known that when endowed with the weak* topology, the set $L^\infty(\mathcal{Y})$ is a linear topological space.

Proof of Statement i). Define the set of all tests

$$\mathcal{C} := \{\varphi \in L^\infty(\mathcal{Y}) \mid 0 \leq \varphi(y) \leq 1 \text{ for } \nu\text{-a.e. } \},$$

and consider the subset of all α-level tests

$$\mathcal{C}_\alpha := \left\{ \varphi \in \mathcal{C} \;\middle|\; \int \varphi f_m d\nu \leq \alpha \text{ for all } m = 1, \ldots, M \right\}.$$

Note that $\mathcal{C}_\alpha$ is nonempty since $\varphi_0 \in \mathcal{C}_\alpha$ for any α. By Lemma 6, the set $\mathcal{C}_\alpha$ is compact under the weak*-topology. As the objective function in (3) is continuous in the weak*-topology, we conclude that there exists a test φ^* that solves (3).

Proof of Statement ii). Recall the Lagrangian

$$L(\varphi, \kappa) = \int \varphi g d\nu - \sum_{m=1}^{M} \kappa_m \left[\int \varphi f_m d\nu - \alpha \right].$$

We first show that Sion's minimax theorem holds, i.e.,

$$\sup_{\varphi \in \mathcal{C}} \min_{\kappa \in \mathbb{R}^M_+} L(\varphi, \kappa) = \min_{\kappa \in \mathbb{R}^M_+} \sup_{\varphi \in \mathcal{C}} L(\varphi, \kappa) := v. \tag{50}$$

First, by Lemma 6, $\mathcal{C}$ is a convex, compact subset of $L^\infty(\mathcal{Y})$ when endowed with the weak* topology. It is clear that $\mathbb{R}^M_+$ is a convex subset of $\mathbb{R}^M$. Second, since we endowed $L^\infty(\mathcal{Y})$ with the weak* topology, then, by definition, for any fixed κ, the functional $L(\cdot, \kappa) : \mathcal{C} \to \mathbb{R}$ is a continuous functional, that is also linear. Similarly, for any fixed φ, $L(\varphi, \cdot) : \mathbb{R}^M_+ \to \mathbb{R}$ is a continuous and linear function of κ. Therefore, all conditions of Sion's minimax theorem are verified; see Simons (1995, Theorem 3).

Since

$$\sup_{\varphi\in\mathcal{C}_\alpha}\int\varphi g d\nu=\sup_{\varphi\in\mathcal{C}}\min_{\kappa\in\mathbb{R}^M_+}L(\varphi,\kappa),$$

and by part i) of Proposition 1 there exists a test $\varphi^*\in C_\alpha$ such that

$$\int\varphi^* g d\nu=\sup_{\varphi\in\mathcal{C}_\alpha}\int\varphi g d\nu,$$

then

$$0\leq v\leq 1.$$

Therefore, by definition of minimum, there exists multipliers $\kappa^*\in\mathbb{R}^M_+$ such that

$$0\leq\sup_{\varphi\in\mathcal{C}}L(\varphi,\kappa^*)=\min_{\kappa\in\mathbb{R}^M_+}\sup_{\varphi\in\mathcal{C}}L(\varphi,\kappa):=v\leq 1.$$

Proof of Statement iii). Let $\hat{\kappa}$ be an arbitrary solution to the dual problem in (5). Let $\hat{\varphi}$ be an arbitrary solution to the primal (3). First, we would like to show that

$$\inf_{\kappa\in\mathbb{R}^M_+}L(\hat{\varphi},\kappa)=\sup_{\varphi\in\mathcal{C}}\min_{\kappa\in\mathbb{R}^M_+}L(\varphi,\kappa),\tag{51}$$

which means that $\hat{\varphi}$ solves the maxmin problem. To this end, note that

$$\int\hat{\varphi}g d\nu=\sup_{\varphi\in\mathcal{C}_\alpha}\int\varphi g d\nu=\sup_{\varphi\in\mathcal{C}}\min_{\kappa\in\mathbb{R}^M_+}L(\varphi,\kappa).$$

Moreover, for any $\varphi\in\mathcal{C}_\alpha$,

$$\min_{\kappa\in\mathbb{R}^M_+}L(\varphi,\kappa)=\int\varphi g d\nu.$$

We conclude that

$$\min_{\kappa\in\mathbb{R}^M_+}L(\hat{\varphi},\kappa)=\int\hat{\varphi}g d\nu=\sup_{\varphi\in\mathcal{C}_\alpha}\int\varphi g d\nu=\sup_{\varphi\in\mathcal{C}}\min_{\kappa\in\mathbb{R}^M_+}L(\varphi,\kappa).$$

This establishes (51). By (51) and (50), we have

$$L(\hat{\varphi},\hat{\kappa}) \geq \min_{\kappa\in\mathbb{R}_+^M} L(\hat{\varphi},\kappa) = \sup_{\varphi\in\mathcal{C}} \min_{\kappa\in\mathbb{R}_+^M} L(\varphi,\kappa)$$
$$= \min_{\kappa\in\mathbb{R}_+^M} \sup_{\varphi\in\mathcal{C}} L(\varphi,\kappa) = \sup_{\varphi\in\mathcal{C}} L(\varphi,\hat{\kappa}) \geq L(\hat{\varphi},\hat{\kappa}).$$

Then, note that $L(\hat{\varphi},\hat{\kappa}) = \inf_{\kappa\in\mathbb{R}_+^M} L(\hat{\varphi},\kappa)$, implying (49). Also, $L(\hat{\varphi},\hat{\kappa}) = \sup_{\varphi\in\mathcal{C}} L(\varphi,\hat{\kappa})$, implying (48). □

Below is a lemma proving that the domain of the primal problem is compact.

Lemma 6. *Let $(\mathcal{Y},\mathcal{F},\nu)$ be a separable measure space in the sense of Exercise 10, Chapter 1 in Stein and Shakarchi (2011) and let ν be a σ-finite measure. Define*

$$\mathcal{C} := \{\varphi \in L^\infty(\mathcal{Y}) \mid 0 \leq \varphi(y) \leq 1 \text{ for } \nu\text{-a.e. } y \in \mathcal{Y}\},$$

and

$$\mathcal{C}_\alpha := \left\{\varphi \in \mathcal{C} \;\middle|\; \int \varphi f_m d\nu \leq \alpha \text{ for all } m = 1,\ldots,M\right\}.$$

Then, $\mathcal{C}$ and $\mathcal{C}_\alpha$ are compact in the weak topology (where $L^\infty(\mathcal{Y})$ is viewed as the dual space of $L^1(\mathcal{Y})$). Moreover, $\mathcal{C}$ is a convex subset of $L^\infty(\mathcal{Y})$.*

Proof. Recall that $L^\infty(\mathcal{Y})$ is identified with the dual of $L^1(\mathcal{Y})$, and the weak* topology on $L^\infty(\mathcal{Y})$ is the weakest topology that makes all maps

$$\varphi \mapsto \langle f,\varphi\rangle := \int_{\mathcal{Y}} \varphi f \, d\nu,$$

continuous for every $f \in L^1(\mathcal{Y})$. By the Banach–Alaoglu theorem (Rudin, 2005, p.68), the closed unit ball

$$\mathcal{B} := \left\{\varphi \in L^\infty(\mathcal{Y}) \,\middle|\, \left|\int_{\mathcal{Y}} \varphi f \, d\nu\right| \leq 1 \text{ for all } f \in L^1(\mathcal{Y}) \text{ with } \|f\|_1 \leq 1\right\}$$

is compact in the weak*-topology.

Observe that

$$\mathcal{C} := \{\varphi \in L^\infty(\mathcal{Y}) \mid 0 \leq \varphi(y) \leq 1 \text{ for } \nu\text{-a.e. } y \in \mathcal{Y}\}$$

is a subset of $\mathcal{B}$, since for any $\varphi \in \mathcal{C}$ and any $f \in L^1(\mathcal{Y})$ with $\|f\|_1 \leq 1$, one has

$$\left|\int_{\mathcal{Y}} \varphi f \, d\nu\right| \leq \int_{\mathcal{Y}} |\varphi||f| \, d\nu \leq \int_{\mathcal{Y}} |f| \, d\nu \leq 1.$$

We now show that $\mathcal{C}$ is weak*-sequentially closed. Suppose that $\{\varphi_n\}_{n=1}^\infty$ is a sequence in $\mathcal{C}$ that converges to some $\varphi \in L^\infty(\mathcal{Y})$ in the weak* topology. Assume for contradiction that $\varphi \notin \mathcal{C}$; then either the set $A_+ := \{y \in \mathcal{Y} \mid \varphi(y) > 1\}$ or $A_- := \{y \in \mathcal{Y} \mid \varphi(y) < 0\}$ has positive measure w.r.t. ν. Without loss of generality, assume that $\nu(A_+) > 0$. Define the function

$$f(y) := \frac{\mathbf{1}_{A_+}(y)}{\nu(A_+)}.$$

Then $f \in L^1(\mathcal{Y})$ and $\|f\|_1 = 1$. Since each $\varphi_n \in \mathcal{C}$, we have

$$\int_{\mathcal{Y}} \varphi_n f \, d\nu \leq 1 \quad \text{for all } n.$$

By the weak* convergence we obtain

$$\lim_{n\to\infty} \int_{\mathcal{Y}} \varphi_n f \, d\nu = \int_{\mathcal{Y}} \varphi f \, d\nu \leq 1$$

On the other hand, because $(\varphi - 1)f$ is positive we have

$$\int_{\mathcal{Y}} (\varphi - 1) f \, d\nu \geq 0 \Longrightarrow \int_{\mathcal{Y}} \varphi f \, d\nu \geq 1.$$

This implies that $\int \varphi f d\nu = 1$, which in turn gives $\int (\varphi - 1) f d\nu = 0$. Such equality holds only

when $(\varphi - 1)f = 0$ for ν-almost surely. However, $(\varphi - 1)f > 0$ on A_+. This contradicts the fact that $\nu(A_+) > 0$! This contradiction shows that $\varphi(y) \in [0,1]$ for ν-almost every $y \in \mathcal{Y}$, i.e., $\varphi \in \mathcal{C}$. Therefore, $\mathcal{C}$ is sequentially closed in the weak*-topology. Since $(\mathcal{Y}, \mathcal{F}, \nu)$ is a separable measure space, then $L^1(\mathcal{Y})$ is separable; see Exercise 10, Chapter 1 in Stein and Shakarchi (2011). Therefore, Theorem 3.16 in Rudin (2005) p. 70 implies that $\mathcal{B}$ (with its subspace weak* topology) is compact and metrizable. This means that the sequential closure of $\mathcal{C}$ coincides with its closure; thus showing that $\mathcal{C}$ is closed in the weak* topology. Since $\mathcal{C}$ is a closed subset of the compact set $\mathcal{B}$, it is compact in the weak* topology. The proof that C_α is compact is entirely analogous and we omit it for the sake of brevity.

Finally, $\mathcal{C}$ is convex because if $\varphi_1, \varphi_2 \in \mathcal{C}$ and $t \in [0,1]$, then for ν-almost every $y \in \mathcal{Y}$,

$$(1-t)\varphi_1(y) + t\varphi_2(y) \in [0,1],$$

which implies that $(1-t)\varphi_1 + t\varphi_2 \in \mathcal{C}$. □

B.3 Theoretical Results on $\varphi_{\bar{\kappa}_T}$

B.3.1 Asymptotic analyses of $\varphi_{\overline{\kappa}_T}$

Lemma 7. *Suppose the conditions of Theorem 2 hold. Then, we have* $f(\overline{\kappa}_T) \xrightarrow{p} \bar{v}$ *as* $T \to \infty$.

Proof. For each T, let

$$\varepsilon_T = \frac{2(1-\alpha)}{\alpha}\sqrt{\frac{\ln M}{T}}, \eta_T = \frac{\alpha}{2(1-\alpha)^2}\varepsilon_T.$$

Denote by $\{\kappa_{T,j}\}_{j=1}^T$ the sequence generated by Algorithm 1 with step number T and step size η_T. Then, $\overline{\kappa}_T = \frac{1}{T}\sum_{j=1}^T \kappa_{T,j}$. It follows by Theorem 2 that, for any $\Omega > 0$,

$$Pr\left\{|f(\bar{\kappa}_T) - \bar{v}| > \left(1 + \frac{2\Omega}{\sqrt{(1-\alpha)^2 N \ln M}}\right)\varepsilon_T\right\} < \exp\left(-\Omega^2\right).$$

For each $\varepsilon > 0$, pick $\Omega = \frac{\varepsilon\alpha}{8}\sqrt{NT}$. Then, for all $T > \frac{\ln M}{\left(\frac{\varepsilon\alpha}{4(1-\alpha)}\right)^2}$, we have

$$Pr\left\{|f(\bar{\kappa}_T) - \bar{v}| > \varepsilon\right\} < \exp\left(-\frac{\varepsilon^2\alpha^2 NT}{64}\right),$$

implying $f(\overline{\kappa}_T) \xrightarrow{p} \bar{v}$ as $T \to \infty$. □

Lemma 7 shows that, $\overline{\kappa}_T$ is also asymptotically a least favorable distribution. This result is expected given Theorem 2's finite-sample numerical convergence result. Next, we show that, with additional regularity conditions, the Neyman-Pearson test $\varphi_{\overline{\kappa}_T}$ based on $\overline{\kappa}_T$ is asymptotically optimal as $T \to \infty$.

Proposition 2. *Suppose the conditions of Theorem 2 hold. In addition, suppose for all $\kappa \in \mathbb{R}_M^+$, we have $g(y) - \sum_{m=1}^M \kappa_m f_m(y) \neq 0$ for ν-almost all y. Then, the following statements are true:*

1. *$\int \varphi_{\overline{\kappa}_T} g d\nu \xrightarrow{p} \bar{v}$ as $T \to \infty$;*

2. *For each convergent subsequence $\{\overline{\kappa}_{T_t}\}$ of $\overline{\kappa}_T$, we have*

$$\limsup_{t\to\infty} \int \varphi_{\overline{\kappa}_{T_t}}(y) f_j d\nu \leq \alpha, \forall j \in [M],$$

   as $t \to \infty$;

3. *If κ^* is unique, we have*

$$\int \varphi_{\overline{\kappa}_T}(y) f_j d\nu \xrightarrow{p} \int \varphi_{\kappa^*} f_j d\nu, \forall j \in [M],$$

   as $T \to \infty$.

Proof. First note, under the stated assumptions in the proposition, the absolute continuity of $F_j, j = 1, \ldots, M$ and G with respect to ν implies that, $g(y) \neq \sum_{m=1}^M \kappa_m^* f_m(y)$ for F_j-almost all y, for each

$j = 1 \dots M$, and the same holds for G-almost all y as well. Together with Proposition 1, we have that, for any solution of the dual κ^*, the test of form φ_{κ^*}, i.e.,

$$\varphi_{\kappa^*}(y) = 1, \text{ when } g(y) > \sum_{m=1}^{M} \kappa_m^* f_m(y),$$
$$\varphi_{\kappa^*}(y) = 0, \text{ when } g(y) \leq \sum_{m=1}^{M} \kappa_m^* f_m(y)$$

is such that

$$\int \varphi_{\kappa^*} f_m d\nu \leq \alpha, \forall m \in [M], \quad \int \varphi_{\kappa^*} g d\nu = \bar{v}. \tag{52}$$

Moreover, dominated convergence theorem implies that the size function

$$\alpha_j(\cdot) = \int \varphi_{(\cdot)} f_j d\nu : \mathbb{R}_+^M \to \mathbb{R}^+,$$

is continuous at all $\kappa \in \mathbb{R}_M^+$, for each $j = 1, \dots, M$, and the power function

$$\pi(\cdot) = \int \varphi_{(\cdot)} g d\nu : \mathbb{R}^M \to \mathbb{R}^+$$

is continuous at all $\kappa \in \mathbb{R}_M^+$ as well.

Proof of Part 1 of the Theorem: As $\overline{\kappa}_T$ is bounded, Prohorov's Theorem (e.g., Theorem 2.4(ii) in Van der Vaart 2000) implies that there exists a converging subsequence $\{\overline{\kappa}_{T_t}\}$ such that $\overline{\kappa}_{T_t} \xrightarrow{d} X_M$ as $t \to \infty$, where X_M is a random vector in $\mathbb{R}_+^M$. Denote by $\mathcal{P}_{X_M}$ the probability measure for the distribution of X_M. Since f is continuous, continuous mapping theorem implies that, as $t \to \infty$,$f(\overline{\kappa}_{T_t}) \xrightarrow{d} f(X_M)$. As we also know $f(\overline{\kappa}_T) \xrightarrow{p} \bar{v}$ as $T \to \infty$, conclude that $f(\overline{\kappa}_T) \xrightarrow{d} \bar{v}$, implying that $f(\overline{\kappa}_{T_t}) \xrightarrow{d} \bar{v}$ as $t \to \infty$ as well. Therefore, $f(X_M)$ must share the same distribution as $\bar{v}$. Conclude that $f(x_M) = \bar{v}$, for $\mathcal{P}_{X_M}$-almost every x_M. Since $\bar{v}$ is the optimal value, this implies that for $\mathcal{P}_{X_M}$-almost every x_M, we have $f(x_M) = \bar{v} = \inf_{\kappa \in \mathbb{R}_+^M} f(x)$, i.e., x_M solves the dual

problem. Therefore, due to (52), we have $\int \varphi_{x_M} g d\nu = \bar{v}$ for $\mathcal{P}_{X_M}$-almost every x_M. Conclude that $\int \varphi_{X_M} g d\nu = \bar{v}$ with probability 1. By continuity of the power function $\pi(\cdot) = \int \varphi_{(\cdot)} g d\nu$ in $\mathbb{R}_+^M$, conclude further that as $t \to \infty$,

$$\pi(\overline{\kappa}_{T_t}) \xrightarrow{d} \int \varphi_{X_M} g d\nu = \bar{v}.$$

As the preceding convergence claim holds for every convergent subsequence, conclude that $\pi(\overline{\kappa}_T) \xrightarrow{d} \bar{v}$, as $T \to \infty$, implying $\pi(\overline{\kappa}_T) \xrightarrow{p} \bar{v}$.

Proof of Part 2 of the Theorem: Analogous to the proof of part 1 of the theorem, consider a convergent subsequence $\{\overline{\kappa}_{T_t}\}$ that converges in distribution to some random vector $X_M \in \mathbb{R}_+^M$ with a probability measure $\mathcal{P}_{X_M}$ for its distribution function. Note, by analogous arguments to the proof of part 1, for $\mathcal{P}_{X_M}$-almost every x_M, we have

$$\int \varphi_{x_M} f_j d\nu \leq \alpha, \forall j \in [M]. \tag{53}$$

Therefore, $\int \varphi_{X_M} f_j d\nu \leq \alpha$ with probability 1 for each $j \in [M]$. The proof is further divided in three steps.

Step 1: We show that $\alpha_j(\overline{\kappa}_{T_t}) \xrightarrow{p} \int \varphi_{X_M} f_j d\nu$ for each $j \in [M]$. Note since α_j is bounded and continuous, Portmanteau's Lemma implies that

$$\mathbb{E}\alpha_j(\overline{\kappa}_{T_t}) \to \mathbb{E} \int \varphi_{X_M} f_j d\nu \leq \alpha$$

as $t \to \infty$. As $\alpha_j(\overline{\kappa}_{T_t})$ is bounded, $\alpha_j(\overline{\kappa}_{T_t})$ is also uniformly integrable. Therefore, we have

$$\alpha_j(\overline{\kappa}_{T_t}) \xrightarrow{p} \int \varphi_{X_M} f_j d\nu$$

as $t \to \infty$ for each $j \in [M]$.

STEP 2: We show that for any $\epsilon > 0$, we have, as $t \to \infty$, $P\{\alpha_j(\overline{\kappa}_{T_t}) \leq \alpha + \epsilon\} \to 1$. For any $\epsilon > 0$, it suffices to show that $P\{\alpha_j(\overline{\kappa}_{T_t}) > \alpha + \epsilon\} \to 0$ as $t \to \infty$. To this end, note for any $0 < \delta < \epsilon$:

$$\begin{aligned}
&P\{\alpha_j(\overline{\kappa}_{T_t}) > \alpha + \epsilon\} \\
=&P\{\alpha_j(\overline{\kappa}_{T_t}) > \alpha + \epsilon, \alpha_j(\overline{\kappa}_{T_t}) - \int \varphi_{X_M} f_j d\nu > \delta\} \\
+&P\{\alpha_j(\overline{\kappa}_{T_t}) > \alpha + \epsilon, \alpha_j(\overline{\kappa}_{T_t}) - \int \varphi_{X_M} f_j d\nu \leq \delta\} \\
\leq&P\left\{\alpha_j(\overline{\kappa}_{T_t}) - \int \varphi_{X_M} f_j d\nu > \delta\right\} \\
+&P\left\{\int \varphi_{X_M} f_j d\nu > \alpha + \epsilon - \delta\right\}.
\end{aligned}$$

Note $P\{\alpha_j(\overline{\kappa}_{T_t}) - \int \varphi_{X_M} f_j d\nu > \delta\} \to 0$ as $t \to \infty$ by the conclusion from step 1, and $P\{\int \varphi_{X_M} f_j d\nu > \alpha + \epsilon - \delta\} = 0$ since $\epsilon - \delta > 0$. Conclude that $P\{\alpha_j(\overline{\kappa}_{T_t}) > \alpha + \epsilon\} \to 0$ as $t \to \infty$.

STEP 3: As ϵ is arbitrary, conclude that $P\{\alpha_j(\overline{\kappa}_{T_t}) \leq \alpha\} \to 1$, as $t \to \infty$, implying $\limsup_{t\to\infty} \alpha_j(\overline{\kappa}_{T,t}) \leq \alpha$ as desired.

PROOF OF PART 3 OF THE THEOREM: Since $f(\overline{\kappa}_T) \xrightarrow{p} \bar{v} = f(\kappa^*)$ as $T \to \infty$ and κ^* is the unique solution of the dual, we must have $\overline{\kappa}_T \xrightarrow{p} \kappa^*$ as well given continuity of f. The conclusion then follows immediately from continuous mapping theorem. □