Dynamic Data-Driven Estimation of Non-Parametric Choice Models

Nam Ho-Nguyen

Discipline of Business Analytics, The University of Sydney

Fatma Kılınç-Karzan

Tepper School of Business, Carnegie Mellon University

We study non-parametric estimation of choice models, which was introduced to alleviate unreasonable assumptions in traditional parametric models, and are prevalent in several application areas. Existing literature focuses only on the static observational setting where all of the observations are given upfront, and lacks algorithms that provide explicit convergence rate guarantees or an a priori analysis for the model accuracy vs sparsity trade-off on the actual estimated model returned. As opposed to this, we focus on estimating a non-parametric choice model from observational data in a *dynamic* setting, where observations are obtained over time. We show that choice model estimation can be cast as a convex-concave saddle-point joint estimation and optimization problem, and we provide an online convex optimization based primal-dual framework for deriving algorithms to solve this problem. By tailoring our framework carefully to the choice model estimation problem, we obtain tractable algorithms with provable convergence guarantees and explicit bounds on the sparsity of the estimated model. Our numerical experiments confirm the effectiveness of the algorithms derived from our framework.

Key words: non-parametric choice, first-order methods, dynamic data

1. Introduction

A choice model is an effective tool to summarize and understand the preferences of a population over a set of items. Such models give choice probabilities, that is, the probability that an agent will choose a particular item from a given subset. They are prevalent in several application areas such as revenue management, web page ranking, betting theory, social choice, marketing, and economics (see Dwork et al. (2001), Talluri and van Ryzin (2004), Jagabathula and Shah (2008), Farias et al. (2009), Desir et al. (2016) and references therein). A good choice model aims to capture complex substitution behaviors of agents in order to accurately describe preferences from limited observations.

Choice model estimation has received quite a bit of interest. Traditional choice models often specify a *parametric* structure for the choice probabilities (examples include the multinomial logit (MNL), nested logit, and mixed MNL models); see Talluri and van Ryzin (2004) and references therein. Imposing a parametric structure makes estimation of the

necessary parameters a simpler task, but is often at the expense of overly facile assumptions on substitution behaviors (such as independence of irrelevant alternatives in MNL models) and consequently preventing us from accurately capturing preferences. Therefore, the *non-parametric* approach of directly estimating a probability distribution over rankings has drawn growing interest in academia and in practice (Rusmevichientong et al. 2006, Farias et al. 2013, 2017), and through case studies, it is shown to lead to substantial improvement in prediction accuracy (see Farias et al. (2013), Jagabathula and Rusmevichientong (2019)). In this paper, we focus on the problem of estimating a non-parametric choice model in a dynamic observation setting, and develop convex optimization-based approaches equipped with convergence and sparsity guarantees.

Related Literature. Earliest studies on non-parametric choice models appear in the economics and psychology literatures, e.g., Block and Marschak (1960). Mahajan and van Ryzin (2001) showed that non-parametric models capture a number of parametric models as special cases. The recent literature focuses on the *static* estimation of non-parametric choice models, where the aim is to find a model that either matches the observed empirical probabilities with the model-based choice probabilities exactly (see Farias et al. (2013) which uses a dual-based constraint sampling approach), or uses column generation to minimize a *distance measure* between the two based on Kullback-Leibler (KL) divergence (van Ryzin and Vulcano 2015) or ℓ_1-norm (Mišić 2016).

As opposed to the specific distance measures used in the prior literature, the recent work of Jagabathula and Rusmevichientong (2019) focuses on general distance measures. While they suggest using the Frank-Wolfe (F-W) algorithm to estimate a non-parametric choice model, the bulk of their work focuses on a particular combinatorial subproblem that arises in all of the methods for non-parametric choice model estimation. Their main contribution is the characterization of sufficient conditions (in terms of the subset structure of the items) under which this subproblem becomes polynomial-time solvable.

In the static setting, these four approaches, Farias et al. (2013), van Ryzin and Vulcano (2015), Mišić (2016), and Jagabathula and Rusmevichientong (2019), are closely related to our work, thus we further discuss and compare them in Appendix B. We now discuss a number of important considerations in the estimation of non-parametric choice models: convergence of the algorithm, sparsity, and dynamic data.

Convergence. van Ryzin and Vulcano (2015), Mišić (2016) both suggest column generation procedures (for different objectives) that terminate in finite time. However, convergence *rates* (i.e., how many iterations needed to obtain an ϵ-accurate model) are not given for their column generation procedure. Jagabathula and Rusmevichientong (2019) proposed using the F-W algorithm, for which convergence rates exist. Nevertheless, a direct application of F-W method to classical distance measures such as norms used in choice modeling is not possible due to the non-smoothness of these measures. Farias et al. (2013) does not discuss convergence of their method.

Sparsity. In a full non-parametric model, there are a factorial number of probabilities to estimate, and thus even specifying a full non-parametric model is intractable even for moderate-sized problems, let alone computing choice probabilities or estimating one from data. This necessarily places an importance on characterizing the model simplicity (sparsity) vs model accuracy trade-off. Farias et al. (2013, Theorem 4) provide an existence guarantee of some sparse model that fits the data. But, they do not provide sparsity guarantees on the *actual* estimated model obtained from their proposed algorithm. The other three approaches aim for model sparsity by increasing the support of the distribution by at most one at each iteration. However, since none of them are equipped with explicit convergence rate guarantees, they cannot provide an a priori analysis for the model accuracy vs sparsity trade-off on the actual estimated model returned by their algorithms.

Dynamic data. Existing techniques for estimation of non-parametric choice models are not designed to efficiently work with the dynamic data, i.e., exploit the possibility of continuously updating the empirical choice probabilities as more observations are collected. This setup is very realistic with today's data collection capabilities. A naïve way to work with dynamic data is to simply re-solve the estimation problem each time we update the choice probabilities. However, in the case of estimation of non-parametric choice models, the static estimation problem is already very expensive to solve, thus such a naïve approach significantly compounds the existing computational challenges.

The convex optimization literature has recently considered the problem of minimizing a function given only an approximate sequence converging to true parameters; referred to as *misspecified optimization* (Ahmadi and Shanbhag 2014, Jiang and Shanbhag 2016) or *joint estimation-optimization* (JEO, Ho-Nguyen and Kılınç-Karzan (2019)). These methods provide a way to solve this problem without re-optimizing each time a new set of parameters

is given, but instead performing only a single projection-type operation at each time step. The dynamic data setting for choice model estimation exactly fits this framework, where we think of the approximate sequence as the empirical choice probabilities computed over a growing set of observational data. Under mild statistical assumptions, the empirical probabilities will converge almost surely. However, the existing methods for JEO are ill-equipped to deal with the challenging non-parametric choice model estimation problem, since the domain of interest in general only admits a high-dimensional representation over which projection-type operations are difficult to perform.

Contributions and Outline. In this paper, we simultaneously address the points raised above on convergence, sparsity and dynamic data for non-parametric choice model estimation. We first describe in Section 2 that the dynamic non-parametric choice model estimation problem, and show that it can be solved as a dynamic JEO saddle point (SP) problem. Inspired by the distance measures used in the prior literature for static non-parametric choice estimation, we examine a large class of distance measures and show that they admit favorable SP structure. Importantly, this class includes all norms, but also covers well-known smoothed versions of these norms as well.

We then present a primal-dual framework for solving the JEO SP problem in Section 3. Specifically, we provide a decomposition of the SP gap into a regret term and two error terms, and describe how the error terms behave as our data approximations become more accurate. In Section 4, we describe regret minimization, which can be used to bound the regret term in the saddle point gap decomposition.

We suggest in Section 4.2 three possible regret minimizing algorithms, and give their regret bounds explicitly. Combined with the error terms, the regret bounds provide rigorous convergence guarantees for our framework; see Remark 3. Our convergence rates also imply explicit guarantees on the sparsity of our estimated choice model (see Remarks 2 and 3). Consequently, our results highlight a natural trade-off between desired estimation accuracy and model sparsity.

The high-dimensional structure of the domain of the choice model estimation problem necessitated us to extend the existing JEO literature to the SP setting; see Remark 1. By viewing it as a SP problem and employing our primal-dual framework, we avoid the problem of having to solve difficult non-linear problems on the high-dimensional domain, and instead solve a relatively compact integer linear program (see Section 4.3) at each

iteration, while still enjoying efficient convergence guarantees. Furthermore, we believe that this JEO SP framework can be of interest beyond choice model estimation.

Our JEO SP framework has ample flexibility that it leads to a variety of solution algorithms simply by utilizing different regret minimization algorithms. In particular, a slight variant of the Frank-Wolfe algorithm (also known as the Conditional Gradient algorithm) can be derived from our framework. Our JEO SP framework in addition allows us to provide a new regularity condition to ensure convergence and avoid accumulation of errors (from using inexact data) which is different to the approximate gradient assumption seen in the usual F-W literature; see Remarks 4 and 5.

It is well-known that the F-W method in general *does not* converge on non-smooth objectives; see Nesterov (2018, Example 1), and we show in Appendix F that the classical distance measures used in the non-parametric choice estimation setup, such as norms and KL divergence are non-smooth. Our framework allows us to also derive algorithms for the non-smooth setting by selecting appropriate regret minimization algorithms, something which the usual F-W algorithm is unable to do. In Section 5, we carry out a numerical study and compare several algorithms derived from our framework with some existing methods. We find that the methods from our framework working directly with the non-smooth distance measures outperform the ones that require smooth approximations, including the existing ones from the literature. We give brief concluding remarks in Section 6.

We present all of the proofs in Appendix A. Existing approaches to non-parametric choice estimation are discussed in detail in Appendix B. Appendix C provides rates for certain error terms in our analysis. Appendix D introduces online convex optimization, a key concept in our framework. Appendix E relates some algorithms derived from our framework to existing ones, such as Mirror Descent and the F-W algorithm. Appendix F, discusses different approaches for smoothing norm-based distance measures and compares our methods with existing F-W based approaches. Full details of our experimental setup from Section 5 and supplementary numerical results are given in Appendix G.

Notation. For a positive integer $n \in \mathbb{N}$, we let $[n] := \{1, \ldots, n\}$, define $\Delta_n := \{x \in \mathbb{R}^n_+ : \sum_{i \in [n]} x_i = 1\}$ to be the standard simplex, and S_n to be the collection of rankings/permutations of the set $[n]$. We denote the identity matrix in $\mathbb{R}^{n \times n}$ by I_n. We refer to a collection of objects b_j, $j \in J$ by the notation $\{b_j\}_{j \in J}$. Given vectors x and y, $\langle x, y \rangle$ corresponds to the usual inner product of x and y. Given a norm $\|\cdot\|$ on a Euclidean space

$\mathbb{E}$, we denote its dual norm by $\|x\|_* := \min_y\{\langle x, y\rangle : \|x\| \le 1\}$. For $q \in [1, \infty]$, $\|x\|_q$ denotes the usual ℓ_q norm of x. For a convex function f, we abuse notation slightly by denoting $\nabla f(x)$ for both the gradient of function f at x if f is differentiable and a subgradient of f at x, even if f is not differentiable. If ϕ is of the form $\phi(x, y)$, then $\nabla_x\phi(x, y)$ denotes the (sub)gradient of ϕ at x while keeping the other variables fixed at y.

2. Dynamic Estimation of a Non-Parametric Choice Model

2.1. Model and Data

In choice modeling, given a set of n items, $[n] = \{1, \ldots, n\}$, we would like to understand the preferences of a population by estimating a model from choice observations. We follow a non-parametric approach to choice modeling introduced by Farias et al. (2013). A non-parametric choice model is described by a probability distribution $\lambda \in \Delta_{n!}$ over all rankings S_n of the items $[n]$. Given a ranking $\sigma \in S_n$, we think of $\lambda(\sigma)$ as the probability that a member of the population will rank the items according to σ. Also, when a particular member with ranking σ is presented a subset of items $A \subseteq [n]$, they will choose the highest σ-ranked item $i(\sigma, A) = \arg\min_{i \in A} \sigma(i)$. Thus, the probability of a random member of the population choosing an item $i \in A$ when presented with a subset A is

$$\mathbb{P}_\lambda[i \mid A] := \sum_{\sigma \in S_n(i,A)} \lambda(\sigma), \quad S_n(i, A) := \{\sigma \in S_n : i \text{ is the highest } \sigma\text{-ranked item in } A\}.$$

Our choice observation set can be described as a collection of K pairs $\{i^k, A^k\}_{k=1}^K$, where $i^k \in A^k$ is the item chosen when the subset of items $A_k \subseteq [n]$ was presented. There are a finite number of possible subsets amongst the observations, we denote these by A_j, $j \in [m] = \{1, \ldots, m\}$. We also denote $N := \sum_{j=1}^m |A_j|$. In practice, the collection of possible subsets $\{A_j\}_{j \in [m]}$ can be controlled. Indeed, structural properties of these can have an impact on a combinatorial subproblem that appears in all non-parametric choice model estimation methods (see Section 4.3). However, in this paper, we take this collection as given; see Jagabathula and Rusmevichientong (2019) for a study on how the structure of $\{A_j\}_{j \in [m]}$ impacts the combinatorial subproblem. Based on this observation set, we define

$$q_{ij} := \frac{1}{K}\sum_{k=1}^K \mathbb{I}(i^k = i, A^k = A_j), \quad q_j := \frac{1}{K}\sum_{k=1}^K \mathbb{I}(A^k = A_j) \quad \text{and} \quad p_{ij} := \frac{q_{ij}}{q_j}, \qquad (1)$$

where $\mathbb{I}$ denotes the indicator function, i.e., $\mathbb{I}(\mathcal{S}) = 1$ if statement $\mathcal{S}$ holds, and $\mathbb{I}(\mathcal{S}) = 0$ otherwise. In words, q_{ij} is the proportion of observations where assortment A_j was displayed

and item i was chosen, q_j is the proportion of observations where assortment A_j was displayed, and p_{ij} is the proportion of observations where item i was chosen amongst those where assortment A_j was displayed. Indeed, p_{ij} are the empirical choice probabilities which we will use to tune the probability distribution λ. We denote the collection of these empirical choice probabilities as $p = \{p_{ij}\}_{i \in A_j, j \in [m]} \in \mathbb{R}^N$.

Our goal is to tune λ so that the choice probabilities $\mathbb{P}_\lambda[i \mid A_j]$ are close to the empirical probabilities p_{ij}. We now define notation to succinctly describe $\mathbb{P}_\lambda[i \mid A_j]$. For a given pair $i \in A_j$ and ranking $\sigma \in S_n$, we let $a_{ij}(\sigma) = 1$ if i is the highest σ-ranked item in A_j (i.e., $\sigma \in S_n(i, A_j)$), and $a_{ij}(\sigma) = 0$ otherwise. We define vectors $a_{ij} = \{a_{ij}(\sigma)\}_{\sigma \in S_n} \in \{0,1\}^{n!}$ and $a(\sigma) = \{a_{ij}(\sigma)\}_{i \in A_j, j \in [m]} \in \{0,1\}^N$. Then $\mathbb{P}_\lambda[i \mid A_j] = \sum_{\sigma \in S_n} a_{ij}(\sigma)\lambda(\sigma) = \langle a_{ij}, \lambda \rangle$.

The polytope X of all possible choice probabilities on observed pairs $i \in A_j$ consistent with some distribution λ is

$$X := \mathrm{conv}\left(\{a(\sigma): \ \sigma \in S_n\}\right) \subseteq \mathbb{R}^N. \tag{2}$$

Our goal can now be stated as finding a point $x \in X$ that is close to p, and our choice model will be the weights λ such that $x = \sum_{\sigma \in S_n} \lambda(\sigma)a(\sigma)$. Formally, we solve

$$\min_x \{D(x,p): \ x \in X\} \tag{3}$$

for some appropriately defined distance measure $D(x,p)$, e.g., $D(x,p) = \|x - p\|$.

Finally, we describe the dynamic data setting, where we obtain additional observations over time. We denote a point in time by $t \in \mathbb{N}$, and the number of observations collected by time t as $K_t \in \mathbb{N}$, which is non-decreasing in t. The set of observations at time t is $\{i^k, A^k\}_{k \in [K_t]}$. For simplicity, we assume the collection of observed subsets $\{A_j\}_{j \in [m]}$ remains the same over time. We can compute empirical probabilities p_{ij}^t at time t by (1) using the observation set $\{i^k, A^k\}_{k \in [K(t)]}$, and denote $p_t := \{p_{ij,t}\}_{i \in A_j, j \in [m]} \in \mathbb{R}^N$. With today's data collection capabilities, accumulating data in this way and updating the choice frequencies p_t is very realistic. In the limit, they will statistically converge $p_t \to p$ for some $p \in \mathbb{R}^N$. Note that if the observations are generated via some 'true model' λ^*, then $p \in X$ (but our methods will not require this). In the dynamic setting, our goal is still to select $x = \sum_{\sigma \in S_n} \lambda(\sigma)a(\sigma)$ close to p, but only with access to the sequence $\{p_t\}_{t \geq 1}$. More precisely, we solve

$$\min_x \{D(x,p): \ x \in X\}, \quad \text{given only the sequence } \{p_t\}_{t \geq 1} \text{ s.t. } p_t \to p. \tag{4}$$

2.2. Non-Parametric Choice Model Estimation as a Dynamic Saddle Point Problem

In this paper, we consider the following class of general class of distance measures admitting a max-type representation, which casts (3) as a min-max saddle point (SP) problem:

$$D(x,p) := \max_{y \in Y} \Psi(x,y;p), \quad \Psi(x,y;p) := \langle B(x-p), y \rangle - \alpha\omega(y) \tag{5a}$$

$$\mathrm{SV}(p) := \min_{x \in X} \max_{y \in Y} \Psi(x,y;p) = \min_{x \in X} D(x,p). \tag{5b}$$

Here, $\alpha \geq 0$, $Y \subseteq \mathbb{R}^k$ is closed and convex, $\omega : Y \to \mathbb{R}$ is a convex function, and B is a matrix with N columns (the same dimension as X) and k rows (the same dimension as Y).

REMARK 1. Given any proper convex (in x) distance measure $D(x,p)$, through convex conjugacy, a max-type representation (5a) always exists. We have (for some domain Y)

$$D(x,p) = \max_{y \in Y} \left\{ \langle y, x \rangle - D^*(y,p) \right\}, \quad D^*(y,p) := \max_{x} \left\{ \langle y, x \rangle - D(x,p) \right\}.$$

The domain Y is bounded when $\nabla_x D(x,p)$ is bounded, and $\min_{x \in X} D(x,p) = \min_{x \in X} \max_{y \in Y} \Psi(x,y;p)$, where $\Psi(x,y;p) := \langle y, x \rangle - D^*(y,p)$.

Solving $\min_{x \in X} \Psi(x,y;p)$ will be a key component in the algorithms of Sections 3 and 4. Our particular form of Ψ in (5a), as well as the one that arises from convex conjugacy, indicates that when all else (i.e., y, p) is fixed, we are solving a linear optimization problem over X. Since X only admits a high-dimensional representation (2), non-linear optimization over X presents difficulties. On the other hand, *linear* optimization over X, while non-trivial, is a manageable problem; we elaborate on this subproblem further in Section 4.3.

We specifically consider distance measures of the form (5a) as this form is broad enough to capture a large variety of commonly used $D(x,p)$ from practice; see e.g., Table 1. ∎

We now discuss desirable properties of $D(x,p)$ and appropriate choices of Y, ω and B in (5a). First, note that clearly we have $\Psi(x,y;p)$ is convex in x and concave in y for any choices. However, to ensure that the solution to (3) is, in some sense, a 'reasonable' solution, we desire the property that, if $p \in X$, then $x = p$ is an optimal solution to (3). An easy sufficient condition for this is to ensure that $D(x,p) = D(p,x)$, which we state next.

PROPOSITION 1. *Suppose that the function* $h(z) := \max_{y \in Y} \left\{ \langle Bz, y \rangle - \alpha\omega(y) \right\}$ *satisfies* $h(z) = h(-z)$ *for any vector* $z \in \mathbb{R}^N$. *Then,* $D(x,p) = D(p,x) = h(x-p)$, *and if in addition* $p \in X$, *then the optimal solution to* (3) *is* $x = p$.

Y	B	$\omega(y)$	$D(x,p)$
$\{\|y\|_* \leq 1\} \subset \mathbb{R}^N$	I_N	0	$\|x-p\|$
$\Delta_{|A_1|} \times \ldots \times \Delta_{|A_m|} \subset \mathbb{R}^N$	I_N	0	$\sum\limits_{j\in[m]} \max\limits_{i \in A_j}(x_{ij}-p_{ij})$
$\{\|y\|_2 \leq 1\} \subset \mathbb{R}^N$	I_N	$\frac{1}{2}\|y\|^2$	$\begin{cases} \frac{1}{2\alpha}\|x-p\|_2^2, & \text{if } \|x-p\|_2 \leq \alpha \\ \|x-p\|_2 - \frac{\alpha}{2}, & \text{o.w. } \|x-p\|_2 > \alpha \end{cases}$
$\Delta_{2N} \subset \mathbb{R}^{2N}$	$\begin{bmatrix} I_N & -I_N \end{bmatrix}$	$\sum\limits_{k\in[2N]} y_k \log(y_k)$	$\alpha \log\left(\sum\limits_{j\in[m]} \sum\limits_{i\in\mathcal{A}_j} 2\cosh\left(\frac{x_{ij}-p_{ij}}{\alpha}\right) \right)$

Table 1   Examples of different distance measures.

On the other hand, symmetry is not a necessary condition to have this reasonable property. We present here one such example which exploits the structure of X and p in the sense that if $x, p \in X$, then $\{x_{ij}\}_{i\in A_j}, \{p_{ij}\}_{i\in A_j}$ live in the simplex $\Delta_{|A_j|}$ for each $j \in [m]$.

PROPOSITION 2. *Let $B = I$, $Y = \Delta_{|A_1|} \times \ldots \times \Delta_{|A_m|}$, and $\omega(y) = 0$ for all $y \in Y$. Then whenever $p \in X$, $x = p$ solves* (3).

Even under these assumptions, this form of $D(x,p)$ still captures a large variety of distance measures. We give some examples in Table 1. In particular, all norms are covered.

3. A Primal-Dual Framework for Dynamic Saddle Point Problems

We now build a framework for solving general problems of the form (5)

$$\text{SV}(p) := \min_{x\in X} \max_{y\in Y} \Psi(x,y;p), \tag{6}$$

where X, Y are convex compact sets, and for every p in the parameter space, $\Psi(x,y;p)$ is convex in x and concave in y. When p is fixed, (6) leads to a primal-dual pair of problems:

$$\mathcal{P}(p) := \min_{x\in X} f(x;p), \quad f(x;p) := \max_{y\in Y} \Psi(x,y;p) \tag{7a}$$

$$\mathcal{D}(p) := \max_{y\in Y} g(y;p), \quad g(y;p) := \min_{x\in X} \Psi(x,y;p). \tag{7b}$$

Note that f and g are convex in x and concave in y respectively, hence (7a) and (7b) are convex problems, and by Sion's minimax theorem (Sion 1958), strong duality holds, i.e., $\mathcal{P}(p) = \mathcal{D}(p) = \text{SV}(p)$. Note that $f(x;p)$ is exactly $D(x,p)$ from (5).

Since our main focus is the dynamic data setting where p is unknown but approximated, in (6) we make the dependence on the problem data p explicit. More precisely, our goal is to solve (6) with access to only a sequence $\{p_t\}_{t\geq 1}$ such that $p^t \to p$, i.e.,

$$\text{SV}(p) = \min_{x\in X} \max_{y\in Y} \Psi(x,y;p), \quad \text{given only a sequence } \{p_t\}_{t\geq 1} \text{ s.t. } p_t \to p. \tag{8}$$

This is a variant of a joint estimation and optimization (JEO) problem, previously studied in Ahmadi and Shanbhag (2014), Ho-Nguyen and Kılınç-Karzan (2019). However, the previous literature on JEO has studied only the primal problem, rather than the more general SP representation of (6). As we will see, by examining (6), we can devise algorithms with favorable properties for certain settings, specifically addressing the high dimensionality challenge of X in non-parametric choice modeling.

The data sequence $\{p_t\}_{t \geq 1}$ induces a sequence of SP problems $\{\mathrm{SV}(p_t)\}_{t \geq 1}$ with the same structure as (6). In order to develop our framework, one of the main tasks is to relate the sequence back to the original SP problem $\mathrm{SV}(p)$.

To start with, let us consider the traditional (static data) setup when p is known. In this setting, solving (6) entails finding a primal-dual pair $(\bar{x}, \bar{y}) \in X \times Y$ such that $f(\bar{x}; p) = \mathcal{P}(p) = \mathcal{D}(p) = g(\bar{y}; p)$, i.e., $\bar{x}, \bar{y}$ are optimal solutions to (7a),(7b), respectively. A key quantity to certify this is the *SP gap*, i.e., the sum of the optimality gaps for (7a) and (7b):

$$\epsilon_{\mathrm{sad}}^{\Psi}(\bar{x}, \bar{y}; p) := f(\bar{x}; p) - \mathcal{P}(p) + \mathcal{D}(p) - g(\bar{y}; p)$$

$$= \max_{y \in Y} \Psi(\bar{x}, y; p) - \min_{x \in X} \Psi(x, \bar{y}; p). \tag{9}$$

Note that $\epsilon_{\mathrm{sad}}^{\Psi}(\bar{x}, \bar{y}; p) \geq 0$ (which is equivalent to weak duality), and $\bar{x}, \bar{y}$ are optimal when $\epsilon_{\mathrm{sad}}^{\Psi}(\bar{x}, \bar{y}; p) = 0$. Moreover, a primal-dual pair $(\bar{x}, \bar{y}) \in X \times Y$ satisfying $\epsilon_{\mathrm{sad}}^{\Psi}(\bar{x}, \bar{y}; p) \leq \epsilon$ is referred to be an ϵ-*approximate solution* to the SP problem (6), and then $\bar{x}, \bar{y}$ are ϵ-*optimal solutions* to the primal and dual problems (7a), (7b), respectively as well.

Our goal, therefore, is to find a primal-dual pair $(\bar{x}, \bar{y})$ with certifiably small SP gap. In the JEO setting, however, we must do this without access to the data p, but instead use only the sequence $\{p_t\}_{t \geq 1}$. The general framework we employ to do this is as follows:

1. Fix some time horizon $T \in \mathbb{N}$.

2. For each $t \in [T]$:
   - Use p_t to generate a primal-dual pair $(x_t, y_t) \in X \times Y$.

3. Select non-negative weights $\{\theta_t\}_{t \in [T]}$ to aggregate the primal-dual sequence $\{x_t, y_t\}_{t \in [T]}$ and obtain a primal-dual pair $(\bar{x}_T^{\theta}, \bar{y}_T^{\theta}) \in X \times Y$ as follows:

$$\Theta_T := \sum_{t \in [T]} \theta_t, \quad \bar{x}_T^{\theta} := \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t x_t, \quad \bar{y}_T^{\theta} := \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t y_t. \tag{10}$$

4. Certify the SP gap $\epsilon_{\mathrm{sad}}^{\Psi}(\bar{x}_T^{\theta}, \bar{y}_T^{\theta}; p)$.

In our framework we deliberately leave steps 2 and 4 vague. To specify these, we present the following fundamental result which relates the data sequence $\{p_t\}_{t\in[T]}$, a generic primal-dual sequence $\{x_t, y_t\}_{t\in[T]}$, and the SP gap $\epsilon_{\mathrm{sad}}^{\Psi}(\cdot, \cdot; p)$.

THEOREM 1. *For given $T \geq 1$, primal-dual sequence $\{x_t, y_t\}_{t\in[T]}$ and weight sequence $\{\theta_t\}_{t\in[T]}$, let $\bar{x}_T^{\theta}, \bar{y}_T^{\theta}$ be defined as in* (10). *Then, given any data sequence $\{p_t\}_{t\in[T]}$, we have*

$$\epsilon_{\mathrm{sad}}^{\Psi}\left(\bar{x}_T^{\theta}, \bar{y}_T^{\theta}; p\right) \leq \widehat{\epsilon}\left(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}\right) + \epsilon^{\circ}\left(\{x_t; p_t, \theta_t\}_{t\in[T]}\right) + \epsilon^{\bullet}\left(\{y_t; p_t, \theta_t\}_{t\in[T]}\right),$$

where $\widehat{\epsilon}\left(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}\right) := \max_{y\in Y} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x_t, y; p_t) - \min_{x\in X} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x, y_t; p_t)$ (11a)

$$\epsilon^{\circ}\left(\{x_t; p_t, \theta_t\}_{t\in[T]}; p\right) := \max_{y\in Y} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \left[\Psi(x_t, y; p) - \Psi(x_t, y; p_t)\right] \tag{11b}$$

$$\epsilon^{\bullet}\left(\{y_t; p_t, \theta_t\}_{t\in[T]}; p\right) := \max_{x\in X} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \left[\Psi(x, y_t; p_t) - \Psi(x, y_t; p)\right]. \tag{11c}$$

Theorem 1 states that, given any sequence $\{x_t, y_t\}_{t\in[T]}$, we can construct a primal-dual pair via (10) and certify its SP gap $\epsilon_{\mathrm{sad}}^{\Psi}\left(\bar{x}_T^{\theta}, \bar{y}_T^{\theta}; p\right)$ on the true data p by bounding the three upper bound terms $\widehat{\epsilon}, \epsilon^{\circ}, \epsilon^{\bullet}$. Thus, Theorem 1 directly addresses step 4 of the template. We design steps 2 and 3 of our framework to ensure that these upper bound terms are small.

Let us comment on the three upper bound terms. Notice that since we have access to $\{p_t\}_{t\in[T]}$ and we generate the sequences $\{x_t, y_t\}_{t\in[T]}$, we will see that $\{x_t, y_t\}_{t\geq 1}$ will be chosen to make $\widehat{\epsilon}$ small. On the other hand, p is present in both the $\epsilon^{\circ}, \epsilon^{\bullet}$ terms, and since p is unknown to us, we do not directly control $\epsilon^{\circ}, \epsilon^{\bullet}$. However, these terms become small under a mild regularity assumption on the SP function Ψ, and when $p_t \to p$ (which is already assumed in the problem statement (8)). For the rest of this section, we focus on the latter two terms; we will focus on $\widehat{\epsilon}$ in Section 4.

Intuitively, we can think of the latter two terms $\epsilon^{\circ}, \epsilon^{\bullet}$ as capturing *the price of estimation*, i.e., they are the errors incurred from using inexact estimates $p_t \neq p$. Indeed, if p were known and we had $p_t = p$ for all t, then these terms will disappear. We now give a sufficient uniform Lipschitz continuity condition that ensures $\lim_{T\to\infty} \{\epsilon^{\circ} + \epsilon^{\bullet}\} \leq 0$ whenever $p_t \to p$.

ASSUMPTION 1. *Fix a norm $\|\cdot\|$ on the data vectors p. There exists a constant $L_{\Psi} < \infty$ such that for any $(x, y) \in X \times Y$ and data p, p', we have $|\Psi(x, y; p) - \Psi(x, y; p')| \leq L_{\Psi}\|p - p'\|$.*

THEOREM 2. *Suppose that Assumption 1 holds, and that $\{\theta_t\}_{t\geq 1}$ is chosen so that $\theta_t \geq 0$ for all $t \geq 1$, $\Theta_T = \sum_{t\in[T]} \theta_t \to \infty$. Then, whenever $p_t \to p$ for any fixed p,*

$$\limsup_{T\to\infty} \left[\epsilon^{\circ}\left(\{x_t; p_t, \theta_t\}_{t\in[T]}\, ; p \right) + \epsilon^{\bullet}\left(\{y_t; p_t, \theta_t\}_{t\in[T]}\, ; p \right) \right] \le \lim_{T\to\infty} \frac{2L_\Psi}{\Theta_T} \sum_{t\in[T]} \theta_t \|p_t - p\| = 0.$$

Theorem 2 ensures that the error terms $\epsilon^{\circ}, \epsilon^{\bullet}$ will be arbitrarily small if we make T large enough. However, the rate at which these terms converge depends on the rate that $\|p_t - p\| \to 0$. This is a quite natural consequence, since in essence it supports the intuition that the overall performance is limited by the quality of the information $\{p_t\}_{t\ge 1}$ available to us; indeed, it is unreasonable to expect that faster rates are possible without assumptions on the dynamics of the sequence $\{p_t\}_{t\ge 1}$ beyond convergence. Different rates of convergence for $\epsilon^{\circ}, \epsilon^{\bullet}$ for common choices of θ_t (i.e., $\theta_t = 1\ \forall t$ or $\theta_t = t\ \forall t$) are given in Appendix C.

In order to get convergence in our framework for the non-parametric choice model estimation problem (5), Assumption 1 must hold for the Ψ defined in (5a). This is indeed the case under a compactness assumption on Y, which is satisfied for all examples in Table 1.

PROPOSITION 3. *Let $\Psi(x, y; p) = \langle B(x - p), y \rangle - \alpha\omega(y)$. Suppose that Y is bounded in some (dual) norm, i.e., $\|y\|_* \le G_Y < \infty$ for some $G_Y \in \mathbb{R}$. Then Assumption 1 holds for Ψ with $L_\Psi = G_Y \|B\|$, where $\|B\|$ is the operator norm induced by $\|\cdot\|$.*

4. Deriving Algorithms from the Primal-Dual Framework

In Section 3, we showed that the SP gap $\epsilon_{\mathrm{sad}}^{\Psi}(\bar{x}, \bar{y}; p)$ certifies the quality of a primal-dual pair $(\bar{x}, \bar{y})$ for the SP problem (6). Then, for the JEO problem (8), Theorem 1 provides a bound on the SP gap via three auxiliary terms (11). Theorem 2 shows that two of these terms, ϵ° and $\epsilon^{\bullet}$, go to 0 as long as a Lipschitz continuity-type assumption on data p (Assumption 1) is satisfied. Thus, to derive algorithms for (8), we focus on the first $\widehat{\epsilon}$ term.

Intuitively, the $\widehat{\epsilon}$ term can be interpreted as an approximate SP gap term. Indeed, we have

$$\widehat{\epsilon}\left(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]} \right) \le \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \left[\max_{y\in Y} \Psi(x_t, y; p_t) - \min_{x\in X} \Psi(x, y_t; p_t) \right],$$

i.e., $\widehat{\epsilon}$ is upper-bounded by the (weighted) average of the SP gaps defined by $\mathrm{SV}(p_t)$ evaluated at the primal-dual pair (x_t, y_t). This means that, to make $\widehat{\epsilon}$ small, at each time $t \in [T]$, we can choose (x_t, y_t) to simply (approximately) solve a SP problem $\mathrm{SV}(p_t)$ given data p_t.

While this is a valid method, even approximately solving $\mathrm{SV}(p_t)$ at each iteration to a given accuracy ϵ can be expensive, and perhaps intractable. A more careful analysis of $\widehat{\epsilon}$, however, reveals more efficient strategies. Notice that $\widehat{\epsilon}$ can be written as

$$\widehat{\epsilon}\Big(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}\Big) = \mathcal{R}_x(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}) + \mathcal{R}_y(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}), \tag{12a}$$

$$\text{where} \quad \mathcal{R}_x(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}) := \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x_t, y_t; p_t) - \min_{x\in X} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x, y_t; p_t) \tag{12b}$$

$$\mathcal{R}_y(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}) := \max_{y\in Y} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x_t, y; p_t) - \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x_t, y_t; p_t). \tag{12c}$$

These two terms, $\mathcal{R}_x$ and $\mathcal{R}_y$ admit a so-called *regret* interpretation from the *online convex optimization* (OCO) domain; see Section 4.2 and Appendix D for a brief discussion of OCO. OCO immediately provides us various techniques, in the form of *regret-minimizing* algorithms, to choose the sequences $\{x_t, y_t\}_{t\geq 1}$ that bound these terms in (12). We now provide a strategy for bounding the two regret terms $\mathcal{R}_x$ and $\mathcal{R}_y$ while addressing specific high dimensionality challenge of choice model estimation (5). As we will see in Section 4.2, typically both regret terms $\mathcal{R}_x$ and $\mathcal{R}_y$ can be bounded by $\leq O(1/\sqrt{T})$ after T iterations.

4.1. Primal Oracle Algorithms for Dynamic Saddle Point Problems

To get bound certificates on the SP gap (9), we must choose a primal-dual sequence $\{x_t, y_t\}_{t\in[T]}$ to minimize the two regret terms (12b) and (12c). Specifically, we choose the primal sequence $\{x_t\}_{t\in[T]}$ to minimize the regret term $\mathcal{R}_x(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]})$, and the dual sequence $\{y_t\}_{t\in[T]}$ to minimize the regret term $\mathcal{R}_y(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]})$. This involves selecting two regret-minimizing algorithms from the OCO literature, and applying them respectively to minimize these two regret terms. While there are many possibilities for doing this, as stated in Remark 1, for non-parametric choice model estimation we focus on a particular type of combination of regret-minimizing algorithms, what we call *primal oracle algorithms*. This class of algorithms in a sense bypasses the high-dimensional challenge of the primal domain X (where performing the usual projection or proximal operations onto X are difficult) and focuses on the dual domain Y. Concretely, at each time $t \in [T]$, we will make a lookahead decision x_t *after* choosing y_t by computing

$$x_t \in \arg\min_{x\in X} \Psi(x, y_t; p_t). \tag{13}$$

This immediately results in the following regret bound.

LEMMA 1. *Let $\{y_t\}_{t\in[T]}$, $\{p_t\}_{t\in[T]}$ and $\{\theta_t\}_{t\in[T]}$ be arbitrary dual, data and weight sequences. Suppose that for each $t \in [T]$, x_t is obtained via (13). Then, we have*

$$\mathcal{R}_x(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}) = \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x_t, y_t; p_t) - \min_{x\in X} \frac{1}{\Theta_T} \sum_{t\in[T]} \theta_t \Psi(x, y_t; p_t) \leq 0.$$

In fact, with this choice and consequently using Theorem 1 and Lemma 1, we deduce that the primal optimality gap is bounded solely by the dual regret term $\mathcal{R}_y$.

COROLLARY 1. *Let $\{y_t\}_{t\in[T]}$, $\{p_t\}_{t\in[T]}$ and $\{\theta_t\}_{t\in[T]}$ be arbitrary dual, data and weight sequences. Suppose that for each $t \in [T]$, $x_t \in \arg\min_{x\in X} \Psi(x, y_t; p_t)$. Denote $\bar{x}_T^\theta :=$ $\frac{1}{\Theta_T}\sum_{t\in[T]}\theta_t x_t$. Then, we have*

$$f(\bar{x}_T^\theta; p) - \min_{x\in X} f(x; p) \leq \mathcal{R}_y(\{x_t, y_t; p_t, \theta_t\}_{t\in[T]}) + \epsilon^\circ\left(\{x_t; p_t, \theta_t\}_{t\in[T]}; p\right) + \epsilon^\bullet\left(\{y_t; p_t, \theta_t\}_{t\in[T]}; p\right).$$

REMARK 2. When using a primal oracle algorithm (13), Corollary 1 gives a bound on the optimality gap. In the context of choice model estimation, when we use a $D(x, p)$ of form (5a) and solve (5b), solving (13) boils down to a linear program over X. This has two important implications. Since $X = \text{Conv}\left(\{a(\sigma)\}_{\sigma\in S_n}\right)$, the optimal solution to (13) is some vertex $a(\sigma)$. Thus, $\bar{x}_T^\theta$ is *readily* expressed as a convex combination of the $\{a(\sigma)\}_{\sigma\in S_n}$. This allows us to easily obtain a distribution λ_T over S_n to describe our estimated choice model at iteration T. Furthermore, the overall number of iterations T upper-bounds the sparsity of the learned choice model because each iteration increases the support of λ_T by at most 1 and the initial point can be taken to be any ranking. As a result, our framework allows us to obtain an explicit dependence between the accuracy parameter ϵ and the sparsity of the estimated ϵ-optimal model via bounds on the regret term $\mathcal{R}_y$ and the error terms $\epsilon^\circ, \epsilon^\bullet$. ∎

We next discuss some possible regret-minimizing algorithms for $\mathcal{R}_y$ under two different assumptions on Ψ. We also elaborate on the connection of these algorithms to well-known optimization algorithms in the static setup.

4.2. Regret-Minimizing Algorithms

In the standard online convex optimization setting, we are given a convex domain Z and a finite time horizon T. In each time period $t \in [T]$, the following takes place:

- we make a decision $z_t \in Z$ based on *past* information from time steps $1, \ldots, t-1$ only.
- Then, a convex loss function $h_t : Z \to \mathbb{R}$ is revealed, we suffer loss $h_t(z_t)$ and get some feedback typically in the form of first-order information $\nabla h_t(z_t)$.

It is usually assumed that the functions h_t are chosen possibly by an all-powerful adversary. As such, our sequence of decisions $\{z_t\}_{t\in[T]}$ is evaluated against the best fixed decision in hindsight, and the (weighted average) difference is defined to be the *weighted regret*:

$$\frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z_t) - \min_{z \in Z} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z), \tag{14}$$

We give a full overview of OCO in Appendix D.

By defining $Z = Y$ and $h_t(\cdot) = -\Psi(x_t, \cdot; p_t)$, the regret term $\mathcal{R}_y$ in (12c) becomes (14). Thus, we state our algorithms in terms of h_t and Z. Throughout this subsection, we fix some norm $\|\cdot\|$ on the Euclidean space in which Z lives, and denote its dual norm by $\|\cdot\|_*$.

ASSUMPTION 2. *There exists a $G < \infty$ such that for any $z \in Z$, $\|\nabla h(z)\|_* \le G$.*

ASSUMPTION 3. *There exists a convex function $\omega : Z \to \mathbb{R}$ satisfying: (a) for any $z, z' \in Z$, $\langle \nabla \omega(z) - \nabla \omega(z'), z - z' \rangle \ge \|z - z'\|^2$; and (b) there exists $\Omega > 0$ such that for any $z, z' \in Z$, $\omega(z) - \omega(z') \le \Omega$.*

Our first algorithm, *online Mirror Descent (oMD)*, applies to the setting where h_t only satisfies convexity; in particular, it can be non-smooth and non-strongly convex.

THEOREM 3. *Suppose that Z satisfies Assumption 3(a,b), and that for all $t \in [T]$, h_t is convex and that Assumption 2 holds for h_t. Given the weight sequence $\{\theta_t\}_{t \in [T]}$, compute*

$$z_1 \in \arg\min_{z \in Z} \omega(z), \quad z_{t+1} = \arg\min_{z \in Z} \{\langle \gamma \theta_t \nabla h_t(z_t) - \nabla \omega(z_t), z \rangle + \omega(z)\},$$

where $\gamma = \sqrt{\frac{2\Omega}{G^2 \sum_{t \in [T]} \theta_t^2}}$. Then

$$\frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z_t) - \min_{z \in Z} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z) \le \sqrt{2\Omega G^2 \sum_{t \in [T]} \left(\frac{\theta_t}{\Theta_T}\right)^2}.$$

Furthermore, when $\theta_t = 1$, the upper bound on regret becomes $\sqrt{2\Omega G^2 / T}$.

Our next two algorithms work under a relative strong convexity assumption on h_t with respect to ω that allows us to get improved upper bound on regret.

ASSUMPTION 4. *Let $\omega : Z \to \mathbb{R}$ be a function satisfying Assumption 3(a). There exists $\alpha > 0$ such that the function $h - \alpha\omega$ is convex.*

Under Assumption 4, oMD algorithm with a different step size and weight sequence results in the following better regret guarantee.

THEOREM 4. *Suppose that Z satisfies Assumption 3(a), and that for all $t \in [T]$, Assumptions 2 and 4 hold for h_t. Let $\theta_t = t$ for all $t \in [T]$, and compute*

$$z_1 \in \arg\min_{z \in Z} \omega(z), \quad z_{t+1} = \arg\min_{z \in Z} \left\{ \left\langle \frac{2}{\alpha(\theta_t + 1)} \nabla h_t(z_t) - \nabla \omega(z_t), z \right\rangle + \omega(z) \right\}.$$

Then $\frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z_t) - \min_{z \in Z} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z) \le \frac{2G^2}{\alpha(T+1)}$.

Let us also examine the *follow-the-leader (FTL)* algorithm from OCO literature.

THEOREM 5. *Suppose that Z satisfies Assumption 3(a), and that for all $t \in [T]$, Assumption 4 holds for h_t, and Assumption 2 holds for $h_t - \alpha\omega$. Let $\theta_t = t$ for all $t \in [T]$, and compute*

$$z_1 \in \underset{z \in Z}{\arg\min}\, \omega(z), \quad z_{t+1} = \underset{z \in Z}{\arg\min} \left\{ \frac{1}{\Theta_t} \sum_{s \in [t]} \theta_s h_s(z) \right\}.$$

Then $\frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z_t) - \min_{z \in Z} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z) \le \frac{2G^2}{\alpha(T+1)}$.

For common domains Z such as simplex, Euclidean ball, and spectahedron, standard selections of norm $\|\cdot\|$, distance generating function $\omega(\cdot)$, and set width Ω satisfying Assumption 3(a,b) and the resulting update rule for z_{t+1} computation used in Theorems 3 and 4 are discussed by Juditsky and Nemirovski (2012, Section 5.7).

Finally, we give conditions on (5a), (5b) to ensure that Theorems 3, 4 or 5 can be applied to minimize the dual regret (12c). Note that for our class of distance measures (5a), because by definition X and p are bounded (see (1) and (2)), Assumptions 2 and 4 are satisfied depending solely on ω. Thus, a judicious definition of ω will allow us to utilize the regret minimizing algorithms of Theorems 3, 4 or 5.

PROPOSITION 4. *For any $x \in X$, let $h_x(z) = -\Psi(x, z; p) := \langle B(p - x), z \rangle + \alpha\omega(z)$. If there exists a norm $\|\cdot\|$ and some $G' > 0$ such that $\|\nabla\omega(z)\|_* \le G'$ for all $z \in Y$, then Assumption 2 holds for h_x. If ω satisfies Assumption 3(a) and $\alpha > 0$, then Assumption 4 holds for h_x.*

REMARK 3. Proposition 4 allows us to use any of the regret-minimizing algorithms from Theorems 3, 4 or 5 to bound the dual regret $\mathcal{R}_y$, from which primal optimality gap bounds can be inferred through (9) (since the dual optimality gap is always non-negative). Ignoring for the moment the error terms $\epsilon^\circ, \epsilon^\bullet$, using Theorem 3 results in a point $\bar{x}_T^\theta$ with ϵ-gap after $O(1/\epsilon^2)$ iterations. By Remark 2, the support of the learned choice model is also at most $O(1/\epsilon^2)$, and thus our framework simultaneously results in both errors bounds and *explicit* sparsity guarantees. See Appendix F.2 and F.3 for more details on the rates attainable using our framework, and Appendix C for rates for the error terms $\epsilon^\circ, \epsilon^\bullet$.　■

REMARK 4. Appendix F.1 shows that the distances measures D used in the previous literature are indeed non-smooth, which is also the case when $\alpha = 0$ in (5). Nesterov (2018, Example 1) shows that applying the Frank-Wolfe (F-W) algorithm as suggested by Jagabathula and Rusmevichientong (2019) to a non-smooth objective does not converge

in general. When we have $\alpha > 0$ in (5a) and $\omega(y)$ is strongly convex (Assumption 3(a)), then $D(x,p)$ is smooth in x. Also, when $D(x,p) = \|x - p\|_q$ for $q \in [2, \infty)$, then simply squaring the norm makes $D(x,p)$ smooth. This means that the F-W algorithm can be applied for both of these smooth cases. Freund and Grigas (2016) provide results for F-W with approximate gradients, and in Appendix F.3, we show that in the dynamic setting, $\nabla_x D(x, p_t)$ can be interpreted as an approximate gradient for $\nabla_x D(x, p)$, thus the results of Freund and Grigas (2016) can be applied here. However, we also show that ensuring the convergence of F-W with approximate gradients using their results requires an assumption that $\|p - p_t\| \to 0$ *sufficiently fast*. In contrast, our primal-dual framework bypasses such assumptions on smoothness of $D(x,p)$ and data convergence rate, and works for general $D(x,p)$, e.g., even non-smooth cases of $D(x,p) = \|x - p\|_q$ for $q \in \{1, \infty\}$. ∎

REMARK 5. In Appendix E, in the static setting $(p_t = p)$, we show that our primal oracle algorithms, i.e., ones which use (13) for x in combination with Theorems 3, 4 or 5 for y, can be interpreted as running traditional MD on a particular dual problem, or F-W on $D(x,p)$. By placing this into our primal-dual framework of Section 3, we are able to obtain guarantees for these algorithms in the dynamic setting as well. Furthermore, in the dynamic setting, the F-W algorithm derived from our framework is slightly different to the traditional one from Freund and Grigas (2016) discussed in Remark 4 and Section F.3. In particular, the analysis of our F-W variant through the primal-dual framework does not require that $\|p - p_t\| \to 0$ at any particular rate. In Section 5, we also numerically compare several algorithms derived from our framework (using Theorems 3, 4 and 5) with ones using the traditional F-W algorithm of Freund and Grigas (2016) on the non-smooth distance function $D(x,p) = \|x - p\|_2$ and an appropriate smoothed version for methods that only work on smooth objectives. We find that algorithms that work directly with the non-smooth $D(x,p)$ outperform their counterparts requiring smoothness. ∎

4.3. Combinatorial Subproblem

One of the key steps in our framework is solving the LO problem (13) over X (see also the definition of Ψ in (5b)). By (2), X is a polytope with vertices $a(\sigma)$, we have, for a cost vector c,

$$\arg\min_{x \in X} \langle c, x \rangle = \mathrm{Conv}\left(\left\{a(\sigma^*) : \sigma^* \in \arg\min_{\sigma \in S_n} \langle a(\sigma), c \rangle\right\}\right).$$

Thus, in each iteration, we must solve the following combinatorial optimization problem over rankings:

$$\min_{\sigma} \left\{ \sum_{j \in [m]} \sum_{i \in A_j} y_{ij}^t a_{ij}(\sigma) : \; \sigma \in S_n \right\}. \tag{15}$$

Problem (15) is NP-hard, since it is a generalization of the linear ordering problem and the maximum weighted independent set problem, see e.g., (van Ryzin and Vulcano 2015, Proposition 3). However, we note that the exact same combinatorial problem must be solved in all other approaches of learning a non-parametric choice model (see Appendix B and in particular the equations (EC.3), (EC.6), (EC.9)). While we cannot avoid the NP-hardness in learning a non-parametric choice model from data, we note that (15) can be formulated as a (relatively) compact integer program with $O(n^2)$ variables and $O(n^3)$ constraints, and also it can be handled efficiently by off-the-shelf integer programming solvers; see Figure EC.5 in Appendix G. Furthermore, Jagabathula and Rusmevichientong (2019) prove polynomial-time solvability of (15) under a number of assumptions on the subsets $A_1, \ldots, A_m$.

5. Computational Study

We carried out a computational study to examine the performance of various methods for estimating a non-parametric choice model from dynamic data. Full details of our computational setup and empirical observations are given in Appendix G. For our results in this section, we considered the ℓ_2-norm based distance measure, i.e., $D(x, p) = \|x - p\|_2$.

We tested four algorithms derived from our framework: a F-W method using the smoothed norm (FW_{dyn}), a dual MD method on the natural (non-smooth) norm (MD_{ns}), the same MD method on the smoothed norm (MD_{smth}) as well as the squared ℓ_2-norm (MD_{sq}). We also tested two other potential algorithms: a naïve F-W method on the smoothed norm ($\text{FW}_{\text{naïve}}$), and a (naïve) F-W method on the squared ℓ_2-norm (FW_{sq}). See Appendix G for details on each method.

Observations are generated from a ground truth mixed multinomial logit model. Observations are constrained to belong to some fixed 'training' set of item-subset pairs. For each iteration in each of the methods, we generated $\kappa \in \mathbb{N}$ new observations and updated p_t (using both old and new observations) according to (1).

Our stopping criterion for all algorithms is the same: the mean absolute error (MAE) of the choice probabilities versus the empirical frequencies on the training item-subset pairs

should be below a set threshold (0.001). We evaluate the performance of each method using two metrics: the MAE on a 'test' set of unseen item-subset pairs and the number of iterations until termination. We found that the number of different observed rankings at termination (i.e., the model sparsity) is strongly correlated with the number of iterations.
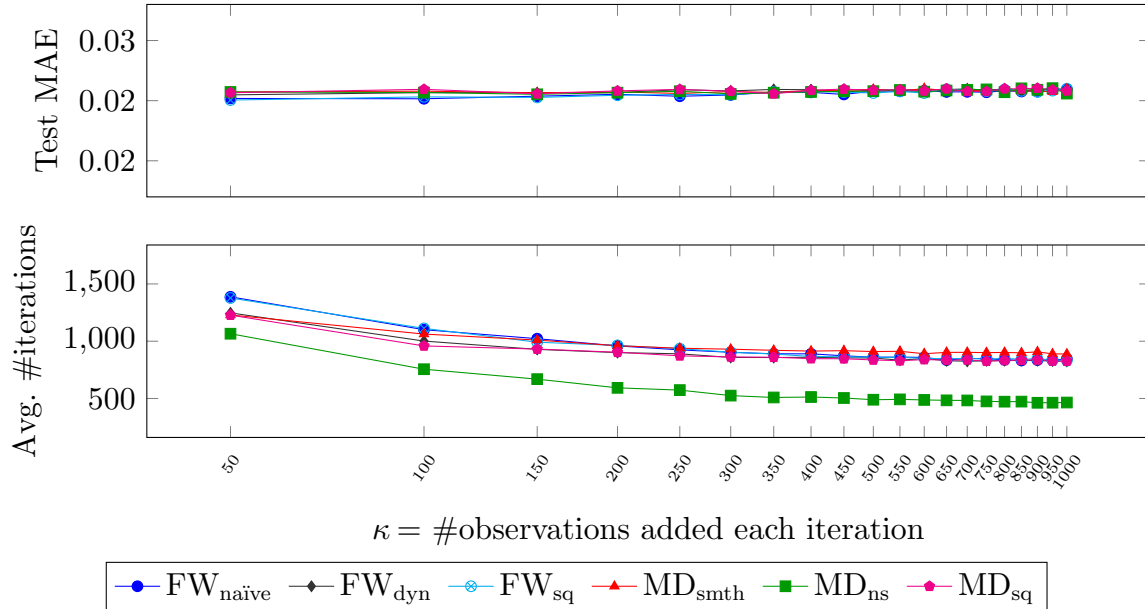
Figure 1 Performance metrics using ℓ_2-norm for different methods for the ground truth choice model with $n = 10$, $m = 20$, $K = L = 5$.

Figure 1 compares each method tested as well as the effect of using different κ on our performance metrics, namely the test MAE and the number of iterations (we note that the solution times are highly correlated with the number of iterations; see Figure EC.5 and Figure EC.4(b)). In Figure 1, different methods and κ have very little effect on the final test set MAE of the estimated model. However, the number of iterations until termination decreases as κ increases. Furthermore, we see that the $\mathrm{MD_{ns}}$ method performs the best in terms of the number of iterations across all levels of κ. Notice also that, while the difference is minor, the $\mathrm{FW_{dyn}}$ method derived from our framework outperforms the $\mathrm{FW_{naïve}}$ method. Both of these highlight the benefit of using our primal-dual framework to derive optimization algorithms for estimating non-parametric choice models. Theoretically, even though all methods have the same asymptotic $O(1/\sqrt{T})$ suboptimality gap bound, the $\mathrm{MD_{ns}}$ method treats the original non-smooth problem directly, and enjoys superior constant factors in the convergence rate. We believe this may be the main factor contributing to its superior numerical performance.
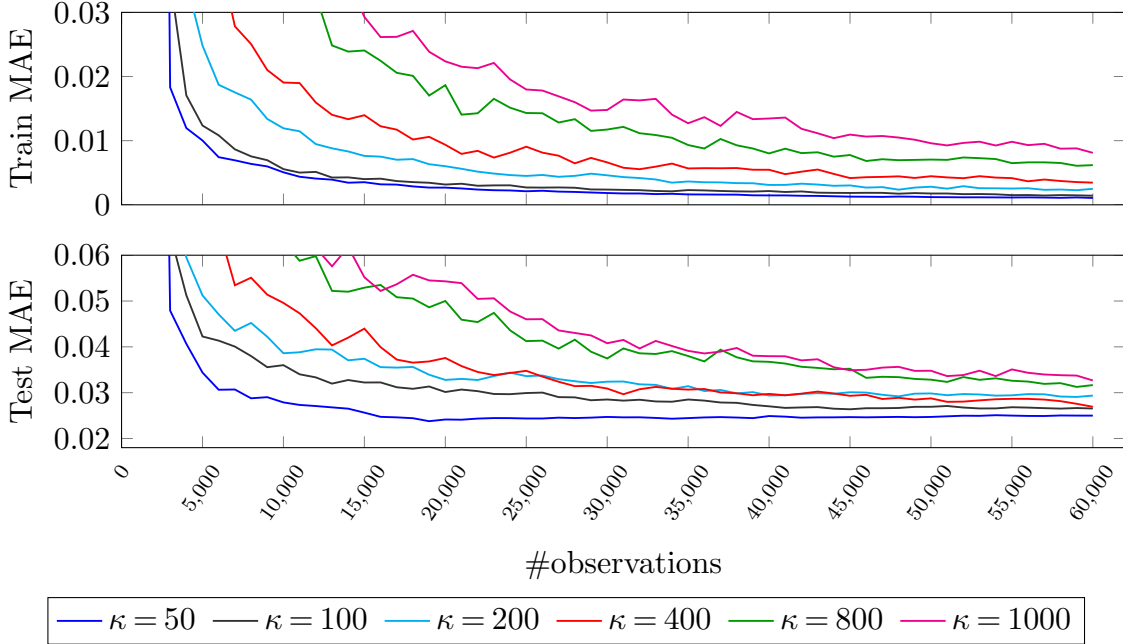
Figure 2   Train and test set MAE vs #observations seen, using ℓ_2-norm minimization with (non-smooth) dual MD, i.e., $\mathrm{MD_{ns}}$ algorithm, for the ground truth choice model with $n = 10$, $m = 20$, $K = L = 5$.

Figure 2 examines the training and test set MAE after seeing a certain number of observations, for one particular instance, specifically for the $\mathrm{MD_{ns}}$ method. We compare the profiles for different κ values to see the impact on MAE of the rate at which new observations are utilized. Note that, at a given number of observations $K \in \{5000, 10000, \ldots, 60000\}$, if κ is higher, then necessarily the method will have performed less iterations. Thus, somewhat unsurprisingly, we observe that both training and test set MAE are lower for lower κ for any given number of observations K. Figures 1 and 2 together suggests the following. Suppose we determine that we perform iterates at consistent time intervals (e.g., two iterations per second). In the regime where observations arrive slowly (e.g., $\kappa = 50$), Figure 2 suggests that we need not wait longer for more observations to achieve low MAE. In the regime where observations arrive quickly (e.g., $\kappa = 1000$), Figure 1 suggests that we need not speed up our iteration updates in order to converge quicker.

6. Conclusions

In this paper, by studying a general convex-concave SP JEO problem and utilizing OCO, we present an efficient and unified framework to solve non-parametric choice estimation problem with dynamic data. In both the static and dynamic settings, our framework immediately provides error bounds and convergence guarantees (on the number of iterations

needed to achieve a certain estimation accuracy), which in turn provide guarantees on the sparsity of our estimated choice model. Our computational experiments demonstrate improved performance when algorithms derived from our framework are used.

In terms of the convex optimization literature, our developments on solving non-smooth optimization problems facing a high-dimensionality challenge and their dynamic variants through JEO for the SP problems are new as well. For static *smooth* problems with a high dimensionality challenge, usually linear optimization as opposed to projection is more tractable, and in such cases the F-W algorithm is often used. In both static and dynamic settings, our framework provides several alternatives to this, all utilizing a linear optimization oracle, but also enjoying convergence in the more general non-smooth setup, which the F-W algorithm does not.

Acknowledgments

References

Ahmadi H, Shanbhag UV (2014) Data-driven first-order methods for misspecified convex optimization problems: Global convergence and rate estimates. *53rd IEEE Conference on Decision and Control*, 4228–4233, ISSN 0191-2216.

Block H, Marschak J (1960) Random orderings and stochastic theories of responses. *Contributions to Probability and Statistics* 2:97–132.

Desir A, Goyal V, Jagabathula S, Segev D (2016) Assortment optimization under the mallows model. *NeurIPS 29*, 4700–4708.

Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. *Proceedings of the 10th International Conference on World Wide Web*, 613–622.

Farias V, Jagabathula S, Shah D (2009) A data-driven approach to modeling choice. *NeurIPS 22*, 504–512.

Farias V, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management Science* 59(2):305–322.

Farias VF, Jagabathula S, Shah D (2017) Building optimized and hyperlocal product assortments: A nonparametric choice approach. Technical report, https://ssrn.com/abstract=2905381.

Freund RM, Grigas P (2016) New analysis and results for the frank–wolfe method. *Mathematical Programming* 155(1):199–230.

Ho-Nguyen N, Kılınç-Karzan F (2019) Exploiting problem structure in optimization under uncertainty via online convex optimization. *Mathematical Programming* 177(1):113–147.

Jagabathula S, Rusmevichientong P (2019) The limit of rationality in choice modeling: Formulation, computation, and implications. *Management Science* 65(5):2196–2215.

Jagabathula S, Shah D (2008) Inferring rankings under constrained sensing. *NeurIPS 21*, 753–760.

Jiang H, Shanbhag UV (2016) On the solution of stochastic optimization and variational problems in imperfect information regimes. *SIAM Journal on Optimization* 26(4):2394–2429.

Juditsky A, Nemirovski A (2012) First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. Sra S, Nowozin S, Wright S, eds., *Optimization for Machine Learning*, Neural information processing series (MIT Press).

Mahajan S, van Ryzin G (2001) Stocking retail assortments under dynamic consumer substitution. *Operations Research* 49(3):334–351.

Mišić VV (2016) *Data, Models and Decisions for Large-Scale Stochastic Optimization* (Chapter 4: Data-Driven Assortment Optimization). Ph.D. thesis, Massachusetts Institute of Technology.

Nesterov Y (2018) Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming* 171(1):311–330.

Rusmevichientong P, Roy BV, Glynn P (2006) A nonparametric approach to multiproduct pricing. *Operations Research* 54(1):82–98.

Sion M (1958) On general minimax theorems. *Pacific J. Math.* 8(1):171–176.

Talluri K, van Ryzin G (2004) *The theory and practice of revenue management*, volume 68 of *International Series in Operations Research & Management Science* (Springer US).

van Ryzin G, Vulcano G (2015) A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Science* 61(2):281–300.

Appendices to *Dynamic Data-Driven Estimation of Non-Parametric Choice Models*

Appendix A: Proofs

Proof of Proposition 1. Note that $D(x,p) = h(x-p) = D(p,x)$. Furthermore, $h(z)$ is clearly convex in z, and hence $h(0) \leq h(z)/2 + h(-z)/2 = h(z)$ for any z, so $x = p$ is an optimal solution. $\square$

Proof of Proposition 2. It is easy to see that

$$D(x,p) = \sum_{j \in [m]} \max_{i \in A_j}(x_{ij} - p_{ij}).$$

Now, when $x = p$, $D(x,p) = 0$. However, when $x \neq p$, there exists at least one $j \in [m]$ and $i \in A_j$ such that $x_{ij} \neq p_{ij}$. If $x_{ij} > p_{ij}$, then we know $D(x,p) > 0$. If $x_{ij} < p_{ij}$ then since both $\{x_{ij}\}_{i \in A_j}, \{p_{ij}\}_{i \in A_j} \in \Delta_{|A_j|}$, we have $\sum_{i \in A_j}(x_{ij} - p_{ij}) = 0$, so there must exist some other $i' \in A_j$ such that $x_{i'j} > p_{i'j}$, hence $D(x,p) > 0$ also. $\square$

Proof of Theorem 1. First, notice that $\{\theta_t/\Theta_T\}_{t \in [T]}$ form a set of convex combination weights. Thus, using the convex-concave structure of $\Psi(\cdot,\cdot;p)$ and the definition of $\epsilon_{\text{sad}}^{\Psi}(\bar{x}_T^{\theta}, \bar{y}_T^{\theta}; p)$ in (9), we get the standard bound

$$\epsilon_{\text{sad}}^{\Psi}\left(\bar{x}_T^{\theta}, \bar{y}_T^{\theta}; p\right) \leq \max_{y \in Y} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x_t, y; p) - \min_{x \in X} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x, y_t; p).$$

Let us examine the first term on the right hand side. Adding and subtracting $\Psi(x_t, y; p_t)$ for each term in the sum, we can bound this term by

$$\max_{y \in Y} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x_t, y; p) = \max_{y \in Y} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \left[\Psi(x_t, y; p_t) + \Psi(x_t, y; p) - \Psi(x_t, y; p_t)\right]$$

$$= \max_{y \in Y} \left\{ \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x_t, y; p_t) + \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \left[\Psi(x_t, y; p) - \Psi(x_t, y; p_t)\right] \right\}$$

$$\leq \max_{y \in Y} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x_t, y; p_t) + \max_{y \in Y} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \left[\Psi(x_t, y; p) - \Psi(x_t, y; p_t)\right]$$

$$= \max_{y \in Y} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x_t, y; p_t) + \epsilon^{\circ}\left(\{x_t; p_t, \theta_t\}_{t \in [T]}; p\right).$$

Using a similar strategy for the second term $\min_{x \in X} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x, y_t; p)$, we can get

$$\min_{x \in X} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x, y_t; p) \geq \min_{x \in X} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t \Psi(x, y_t; p_t) - \epsilon^{\bullet}\left(\{y_t; p_t, \theta_t\}_{t \in [T]}; p\right).$$

Subtracting this lower bound from the upper bound on the first term then gives us the result. $\square$

Proof of Theorem 2.   First observe that for any $t \geq 1$, $(x_t, y), (x, y_t) \in X \times Y$ and p, by Assumption 1,

$$|\Psi(x_t, y; p) - \Psi(x_t, y; p_t)| \leq L_\Psi \|p - p_t\|, \quad |\Psi(x, y_t; p_t) - \Psi(x, y_t; p)| \leq L_\Psi \|p_t - p\|.$$

This implies that

$$\epsilon^\circ \left(\{x_t; p_t, \theta_t\}_{t \in [T]}; p \right) + \epsilon^\bullet \left(\{y_t; p_t, \theta_t\}_{t \in [T]}; p \right) \leq \frac{2L_\Psi}{\Theta_T} \sum_{t \in [T]} \theta_t \|p_t - p\|.$$

We now show the following:

$$a_t \to 0 \implies \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t a_t \to 0.$$

To get our result, we apply this to the sequence $a_t = 2L_\Psi \|p_t - p\|$, which converges to 0 since $p_t \to p$. Fix some $\epsilon > 0$, and choose $S(\epsilon) \in \mathbb{N}$ sufficiently large such that for $t \geq S(\epsilon)$, $|a_t| \leq \epsilon/3$. Furthermore, choose T sufficiently large such that $\left| \frac{1}{\Theta_T} \sum_{t \in [S(\epsilon)]} \theta_t a_t \right| \leq \epsilon/2$ and $\left| \frac{1}{\Theta_T} \sum_{t \in [S(\epsilon)]} \theta_t \right| \leq 1/2$. We have

$$\left| \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t a_t \right| \leq \left| \frac{1}{\Theta_T} \sum_{t \in [S(\epsilon)]} \theta_t a_t \right| + \left| \frac{1}{\Theta_T} \sum_{t=S(\epsilon)+1}^{T} \theta_t a_t \right|.$$

The first term is $\leq \epsilon/2$ by our choice of T, and also the second term satisfies

$$\left| \frac{1}{\Theta_T} \sum_{t=S(\epsilon)+1}^{T} \theta_t a_t \right| \leq \frac{1}{\Theta_T} \sum_{t=S(\epsilon)+1}^{T} \theta_t |a_t| \leq \frac{\epsilon}{3\Theta_T} \sum_{t=S(\epsilon)+1}^{T} \theta_t = \frac{\epsilon}{3} \left(1 - \frac{1}{\Theta_T} \sum_{t=1}^{S(\epsilon)} \theta_t \right) \leq \frac{\epsilon}{2}.$$

□

Proof of Proposition 3.   We have $|\Psi(x, y; p) - \Psi(x, y; p')| = |\langle B(p - p'), y \rangle| \leq \|y\|_* \|B(p - p')\| \leq \|y\|_* \|B\| \|p - p'\| \leq G_Y \|B\| \|p - p'\|$.   □

Proof of Lemma 1.   This is immediate since for any $x \in X$, $\Psi(x_t, y_t; p_t) = \min_{x' \in X} \Psi(x', y_t; p_t) \leq \Psi(x, y_t; p_t)$.   □

Proof of Corollary 1.   Theorem 1 and Corollary 1 immediately imply that $\epsilon_{\mathrm{sad}}^\Psi(\bar{x}_T^\theta, \bar{y}_T^\theta; p)$ is bounded by the right hand side in the result. To get the left hand side, notice that

$$\epsilon_{\mathrm{sad}}^\Psi(\bar{x}_T^\theta, \bar{y}_T^\theta; p) = f(\bar{x}_T^\theta; p) - \min_{x \in X} f(x; p) + \max_{y \in Y} g(y, p) - g(\bar{y}_T^\theta; p) \geq f(\bar{x}_T^\theta; p) - \min_{x \in X} f(x; p)$$

since the optimality gap for the dual problem $\mathcal{D}(p)$ in (7b) is always non-negative.   □

Proof of Theorems 3, 4 and 5.   Theorem 3 follows almost directly from Ho-Nguyen and Kılınç-Karzan (2019, Theorem 1), with minor modifications. Theorem 4 follows from Ho-Nguyen and Kılınç-Karzan (2019, Theorem 2). Theorem 5 follows from Shalev-Shwartz and Kakade (2008, Theorem 2) which, after taking $\ell_t = \theta_t \alpha \omega + \theta_t (h_t - \alpha \omega)$ and $\theta_t = t$, gives the regret bound

$$\sum_{t \in [T]} \theta_t h_t(z_t) - \min_{z \in Z} \sum_{t \in [T]} \theta_t h_t(z) \leq \frac{1}{2} \sum_{t \in [T]} \frac{\theta_t^2 G^2}{\alpha \sum_{s \in [t]} \theta_s} = \frac{1}{\alpha} \sum_{t \in [T]} \frac{t G^2}{t+1} \leq \frac{T G^2}{\alpha}.$$

Dividing by $\Theta_T = T(T+1)/2$ gives the result.   □

Proof of Proposition 4   Note that $\nabla_z h_x(z) = B(p-x) + \alpha\omega(z)$, hence $\|\nabla_z h_x(z)\|_* \le \|B(p-x)\|_* + G'$ is uniformly bounded over p, x since these come from a bounded set, so Assumption 2 is satisfied. Assumption 4 holds trivially. $\square$

Appendix B: Existing Approaches to Non-Parametric Choice Estimation

In this appendix, we examine the existing approaches to learn the non-parametric choice model, i.e., infer an appropriate probability vector λ using the data collected via the process outlined in Section 2.1, and demonstrate how they are particular instantiations of our general model. For a fixed subset $\mathcal{A}_j$, $j \in [m]$, we denote the collection of associated choice probabilities as $A_j\lambda = \{\mathbb{P}_\lambda[i \,|\, \mathcal{A}_j]\}_{i\in\mathcal{A}_j} \in \Delta_{|\mathcal{A}_j|}$.

B.1. Revenue Prediction Approach

Let r_i be the revenue of item $i \in [n]$. Then the expected revenue of an assortment $\mathcal{A} \subset [n]$ under distribution λ is $\sum_{i\in\mathcal{A}} r_i \mathbb{P}_\lambda[i \,|\, \mathcal{A}]$. Farias et al. (2013) seek to find the worst-case expected revenue from a distribution λ consistent with the given data in the sense that the theoretical probabilities $\mathbb{P}_\lambda[i \,|\, \mathcal{A}_j] = \langle a_{ij}, \lambda \rangle$ are precisely consistent with their empirical estimates p_{ij}. Since the probabilities $\mathbb{P}_\lambda[i \,|\, \mathcal{A}]$ are linear in λ, this can be formulated as a linear program (LP)

$$\min_\lambda \left\{ \sum_{i\in\mathcal{A}} r_i \mathbb{P}_\lambda[i \,|\, \mathcal{A}] : \; A\lambda = p, \; \lambda \in \Delta_{n!} \right\}.$$

We first make a few observations related to this model of Farias et al. (2013). In fact, when $\mathcal{A} = \mathcal{A}_j$ for some $j \in [m]$, we have $\mathbb{P}_\lambda[i \,|\, \mathcal{A}] = \langle a_{ij}, \lambda \rangle = p_{ij}$ due to the constraints $A\lambda = p$, hence the objective is constant. Thus, the LP becomes a feasibility problem

$$\text{find} \quad \lambda \in \Delta_{n!} \quad \text{s.t.} \quad A\lambda = p. \tag{EC.1}$$

That said, (EC.1) is still computationally intractable even for moderate values of n because it involves $n!$ variables. Nonetheless, the dual of (EC.1) admits the following robust LP interpretation:

$$\max_{\beta, \nu} \left\{ \langle \beta, p \rangle - \nu : \; \max_{\sigma \in S_n} \langle \beta, a(\sigma) \rangle \le \nu \right\}. \tag{EC.2}$$

Note that verifying the feasibility of a solution with respect to the robust constraint in (EC.2), i.e.,

$$\max_{\sigma \in S_n} \langle \beta, a(\sigma) \rangle = \max_\sigma \left\{ \sum_{j\in[m]} \sum_{i\in\mathcal{A}_j} \beta_{ij} a_{ij}(\sigma) : \; \sigma \in S_n \right\} \le \nu \tag{EC.3}$$

is a combinatorial problem of the exact same form as (15). Farias et al. (2013) suggests solving (EC.2) either using the constraint sampling technique (Calafiore and Campi 2005) or by building an approximation to its robust counterpart obtained from approximating the uncertainty sets with an efficiently representable polyhedron.

In fact, (EC.1) can be seen as choosing $\lambda \in \Delta_{n!}$ to minimize a (very harsh) distance measure:

$$\min_{\lambda \in \Delta_{n!}} D(A\lambda, p), \quad D(A\lambda, p) = \begin{cases} 0, & A\lambda = p \\ \infty, & \text{otherwise.} \end{cases} \tag{EC.4}$$

In general, and specifically when the observations are noisy, there is no guarantee that there exists $\lambda \in \Delta_{n!}$ to fit the data p exactly, i.e., $A\lambda = p$. To remedy this, van Ryzin and Vulcano (2015) and Mišić (2016) examine approaches that use less harsh distance measures $D(\cdot, \cdot)$.

B.2. Maximum Likelihood Estimation Approach

van Ryzin and Vulcano (2015) propose the following method to learn λ via maximum likelihood estimation (MLE). We next describe their method and provide an alternative interpretation of their approach as the minimization of a particular distance measure, namely Kullback-Leibler (KL) divergence, between the true distributions $A_j\lambda$ and their empirical estimates p_j. Note that given two positive vectors $p, x \in \mathbb{R}^n$, their KL divergence is $\text{KL}(p, x) := \sum_{i \in [n]} p_i \log(x_i/p_i)$.

By (1), each item-assortment pair $i \in \mathcal{A}_j$ is seen Kq_{ij} times amongst the observations $\{i^k, \mathcal{A}^k\}_{k=1}^K$. Based on this, the log-likelihood of the observation set $\{i^k, \mathcal{A}^k\}_{k=1}^K$ is $\sum_{j \in [m]} \sum_{i \in \mathcal{A}_j} Kq_{ij} \log(\langle a_{ij}, \lambda \rangle)$. Thus, ignoring the constant K factor, the MLE problem is

$$\max_\lambda \left\{ \sum_{j \in [m]} \sum_{i \in \mathcal{A}_j} q_{ij} \log(\langle a_{ij}, \lambda \rangle) : \ \lambda \in \Delta_{n!} \right\}. \tag{EC.5}$$

Throughout, we use the convention that when $q_{ij} = \langle a_{ij}, \lambda \rangle = 0$, we set $q_{ij} \log(\langle a_{ij}, \lambda \rangle) = 0$. This implies that if the optimal solution λ to (EC.5) has $\mathbb{P}_\lambda[i \mid \mathcal{A}_j] = \langle a_{ij}, \lambda \rangle = 0$, then we must have $q_{ij} = 0$ also, i.e., we did not observe any choices of i from $\mathcal{A}_j$ in our data either.

Like (EC.1), the problem (EC.5) is very large, with $n!$ variables. A column generation technique is suggested in van Ryzin and Vulcano (2015) to get around this, i.e., solve (EC.5) on a subset of the variables, and use the optimality conditions to add variables as needed. The MLE column generating subproblem is constructed as

$$\max_\sigma \left\{ \sum_{j \in [m]} \sum_{i \in \mathcal{A}_j} \frac{q_{ij} a_{ij}(\sigma)}{\langle a_{ij}, \lambda(S) \rangle} : \ \sigma \in S_n \right\}. \tag{EC.6}$$

The solution $\lambda(S)$ is optimal if (EC.6) $\leq K$, otherwise the column σ^* maximizing (EC.6) is added to the set S, and the process is repeated. Note that (EC.6) has the same form as (15) and (EC.3).

We next demonstrate that the MLE problem (EC.5) admits a nice interpretation between the empirical estimates $\{p_j\}_{j \in [m]}$ and the distributions $\{A_j\lambda\}_{j \in [m]}$. To observe this, let us rewrite the objective in (EC.5) as

$$\sum_{j \in [m]} \sum_{i \in \mathcal{A}_j} q_{ij} \log(\langle a_{ij}, \lambda \rangle) = \sum_{j \in [m]} q_j \sum_{i \in \mathcal{A}_j} p_{ij} \log(\langle a_{ij}, \lambda \rangle)$$

$$= - \sum_{j \in [m]} q_j \underbrace{\sum_{i \in \mathcal{A}_j} p_{ij} \log \left(\frac{p_{ij}}{\langle a_{ij}, \lambda \rangle} \right)}_{= \mathrm{KL}(p_j, A_j \lambda)} + \underbrace{\sum_{j \in [m]} q_j \sum_{i \in \mathcal{A}_j} p_{ij} \log(p_{ij})}_{= \text{constant}}$$

where $\mathrm{KL}(a, b)$ is the KL divergence between two probability distributions a and b. Hence, (EC.5) is equivalent to solving

$$\min_{\lambda} \left\{ \sum_{j \in [m]} q_j \, \mathrm{KL}(p_j, A_j \lambda) : \ \lambda \in \Delta_{n!} \right\}. \tag{EC.7}$$

Thus, by defining $D(A\lambda, p) = \sum_{j \in [m]} q_j \, \mathrm{KL}(p_j, A_j \lambda)$, we see that the MLE approach is equivalent to (EC.4) but with a different distance measure $D(\cdot, \cdot)$.

B.3. Norm-Minimization Approach

As opposed to the approaches outlined in Appendix B.1 and B.2, in order to estimate a non-parametric choice model λ, Mišić (2016) suggest minimizing the ℓ_1-norm of $p - A\lambda$ by solving

$$\min_{\lambda} \left\{ \|p - A\lambda\|_1 : \ \lambda \in \Delta_{n!} \right\}. \tag{EC.8}$$

In fact, (EC.8) can be cast as an LP, but it is still computationally intractable since the dimension of λ is $n!$. Similar to van Ryzin and Vulcano (2015), Mišić (2016) addresses this computational difficulty via a column generation approach. Again, (EC.8) is of the same form as (EC.4) where the distance measure $D(\cdot, \cdot)$ is selected to be $D(A\lambda, p) = \|p - A\lambda\|_1$. Furthermore, the resulting column generating subproblem is of the form

$$\max_{\sigma} \left\{ \sum_{j \in [m]} \sum_{i \in \mathcal{A}_j} \beta_{ij}(S) a_{ij}(\sigma) - \nu(S) : \ \sigma \in S_n \right\}, \tag{EC.9}$$

where $\beta(S)$ and $\nu(S)$ are from the dual solution to solving (EC.8) on a subset of columns $\sigma \in S \subset S_n$. Again, this subproblem has the same form as (15), (EC.3) and (EC.6).

Appendix C: Convergence Rates for Error Terms $\epsilon^{\circ}, \epsilon^{\bullet}$

In Table EC.1, we state the convergence rate of $\frac{2L}{\Theta_T} \sum_{t \in [T]} \theta_t \|p_t - p\| \to 0$ for different possible rates of $\|p_t - p\| \to 0$, as well as two common choices for θ_t, namely $\theta_t = 1$ and $\theta_t = t$. In Section 4, we discuss the effect of the choice of θ on the regret bounds for $\widehat{\epsilon}$.

PROPOSITION EC.1. *The convergence rates in Table EC.1 hold.*

Proof of Proposition EC.1 First, we analyze $S(r, T) := \sum_{t \in [T]} \frac{1}{t^r}$ for $r \neq 1, 2$. Observe that

$$\frac{1}{1-r} \left(\frac{1}{(T+1)^{r-1}} - 1 \right) = \int_{t=1}^{T+1} \frac{1}{t^r} dt \leq S(r, T) \leq \left(1 + \int_{t=1}^{T} \frac{1}{t^r} dt \right) = \frac{1}{1-r} \left(\frac{1}{T^{r-1}} - r \right).$$

rate at which $\frac{2L_\Psi}{\Theta_T} \sum_{t\in[T]} \theta_t \|p_t - p\| \to 0$	$\theta_t = 1,\ \Theta_T = T$	$\theta_t = t,\ \Theta_T = \frac{T(T+1)}{2}$
$\|p_t - p\| = O(1/t^r),\ r \in (0,1),$	$\sim 1/T^r$	$\sim 1/T^r$
$\|p_t - p\| = O(1/t)$	$\sim \log(T)/T$	$\sim 1/T$
$\|p_t - p\| = O(1/t^r),\ r \in (1,2),$	$\sim 1/T$	$\sim 1/T^r$
$\|p_t - p\| = O(1/t^2)$	$\sim 1/T$	$\sim \log(T)/T^2$
$\|p_t - p\| = O(1/t^r),\ r > 2$	$\sim 1/T$	$\sim 1/T^2$
$\|p_t - p\| = O(\beta^t),\ \beta \in (0,1)$	$\sim 1/T$	$\sim 1/T^2$

Table EC.1   Convergence rate of bound for $\epsilon^\circ + \epsilon^\bullet$.

- We consider the case $\theta_t = 1$, $\Theta_T = T$, $\|p_t - p\| = O(1/t^r)$, $r > 0$. In this case we have

$$\frac{1}{1-r}\left(\frac{1}{(T+1)^r} - \frac{1}{T+1}\right) \le \frac{2L}{\Theta_T}\sum_{t\in[T]}\theta_t\|p_t - p\| \sim \frac{1}{T}S(r,T) \le \frac{1}{1-r}\left(\frac{1}{T^r} - \frac{r}{T}\right).$$

When $r < 1$, $1 - r > 0$ and $1/T = O(1/T^r)$, hence the lower and upper bounds are $\sim 1/T^r$. When $r > 1$, $1 - r < 0$ and $1/T^r = O(1/T)$, hence the lower and upper bounds are $\sim 1/T$.

- Now consider the case $\theta_t = t$, $\Theta_T = 2/(T(T+1))$, $\|p_t - p\| = O(1/t^r)$, $r > 0$. In this case we have

$$\frac{2}{2-r}\left(\frac{1}{(T+1)^r} - \frac{1}{(T+1)^2}\right) \le \sim \frac{1}{\Theta_T}\sum_{t\in[T]}\theta_t\|p_t - p\|$$

$$\sim \frac{2}{T(T+1)}S(r-1,T) \le \frac{2}{2-r}\left(\frac{1}{T^{r-1}(T+1)} - \frac{r-1}{T(T+1)}\right).$$

When $r < 2$, we have $2 - r > 0$ and $1/(T(T+1)) = O(1/(T^{r-1}(T+1))) = O(1/T^r)$, hence the lower and upper bounds are $\sim 1/T^r$. When $r > 2$, $2 - r < 0$ and $1/(T^{r-1}(T+1)) = O(1/(T(T+1))) = O(1/T^2)$, hence the lower and upper bounds are $\sim 1/T^2$.

- Now consider the case $\theta_t = 1$, $\Theta_T = T$, $\|p_t - p\| = O(1/t)$. Then

$$\frac{2L}{\Theta_T}\sum_{t\in[T]}\theta_t\|p_t - p\| \sim \frac{1}{T}\sum_{t\in[T]}\frac{1}{t} \sim \frac{\log(T)}{T}.$$

- Now consider the case $\theta_t = t$, $\Theta_T = 2/(T(T+1))$, $\|p_t - p\| = O(1/t^2)$. Then

$$\frac{2L}{\Theta_T}\sum_{t\in[T]}\theta_t\|p_t - p\| \sim \frac{1}{T(T+1)}\sum_{t\in[T]}\frac{1}{t} \sim \frac{\log(T)}{T^2}.$$

Finally, consider the case $\|p_t - p\| = O(\beta^t)$ for $\beta \in (0,1)$. When $\theta_t = 1$, $\Theta_T = T$, we have

$$\frac{2L}{\Theta_T}\sum_{t\in[T]}\theta_t\|p_t - p\| = \frac{2L}{\Theta_T}\sum_{t\in[T]}\beta^t = \frac{2L\beta(1-\beta^T)}{T(1-\beta)} \sim 1/T.$$

When $\theta_t = t$, $\Theta_T = 2/(T(T+1))$, we have

$$\frac{4L}{T(T+1)}\sum_{t\in[T]}\theta_t\|p_t - p\| = \frac{4L}{T(T+1)}\sum_{t\in[T]}t\beta^t = \frac{4L\beta(1-(T+1)\beta^T + T\beta^{T+1})}{T(T+1)(1-\beta)^2} \sim 1/T^2.$$

$\square$

Appendix D: Online Convex Optimization

In the standard OCO setting, we are given a convex domain Z and a finite time horizon T. In each time period $t \in [T]$, the following takes place:

- we make a decision $z_t \in Z$ based on *past* information from time steps $1, \ldots, t-1$ only.

- Then, a convex loss function $h_t : Z \to \mathbb{R}$ is revealed, we suffer loss $h_t(z_t)$ and get some feedback typically in the form of first-order information $\nabla h_t(z_t)$.

It is usually assumed that the functions h_t are chosen possibly by an all-powerful adversary that has full knowledge of our learning algorithm—and we know of only the general class of these functions. As such, it is unreasonable to compare the loss of the player across the time horizon to the best possible loss, which would require full knowledge of h_t in advance of choosing x_t. Instead, the player's sequence of decisions z_t is evaluated against the best fixed decision in hindsight, and the (average) difference is defined to be the *regret*:

$$\frac{1}{T} \sum_{t \in [T]} h_t(z_t) - \min_{z \in Z} \frac{1}{T} \sum_{t \in [T]} h_t(x). \tag{EC.10}$$

The goal in OCO is to design efficient *regret minimizing* algorithms that generate x_t so that the regret tends to zero as T increases. Thus, in OCO we seek algorithms to choose x_t that ensure

$$\frac{1}{T} \sum_{t \in [T]} h_t(z_t) - \min_{z \in Z} \frac{1}{T} \sum_{t \in [T]} h_t(z) \leq r(T), \quad \lim_{T \to \infty} r(T) = 0,$$

and the performance of our algorithms is measured by how quickly $r(T)$ tends to 0.

For our work, we consider two simple modifications to the standard OCO setting: *lookahead decisions* and *weighted regret*. More precisely, lookahead decisions allow for the possibility of choosing z_t *with knowledge of* the loss function h_t, while weighted regret modifies (EC.10) to instead be a weighted average

$$\frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z_t) - \min_{z \in Z} \frac{1}{\Theta_T} \sum_{t \in [T]} \theta_t h_t(z), \tag{EC.11}$$

where (as before) $\{\theta_t\}_{t \in [T]}$ is a collection of non-negative weights, and $\Theta_T = \sum_{t \in [T]} \theta_t$. The reason for examining weighted regret is clear: by appropriately defining h_t, the two terms in (12) are actually weighted regret terms. We examine lookahead decisions because the h_t that we define to interpret (12) as two regret terms, which are $-\Psi(x_t, \cdot; p_t)$ and $\Psi(\cdot, y_t; p_t)$ respectively, actually depend on the decisions x_t, y_t and the data p_t, which we have some control over. In particular, in our setting, it is possible to make lookahead decisions for *one* of the regret terms; however, it is not possible to do it for both, because we must choose either x_t before y_t or x_t before y_t. The primal oracle algorithms that we introduce in Section 4.1 exactly choose x_t after y_t, thus y_t must be chosen in a non-anticipative manner.

Note that Ho-Nguyen and Kılınç-Karzan (2019), Wang and Abernethy (2018), Abernethy et al. (2018) have all examined lookahead and weighted regret in OCO before. In fact, Ho-Nguyen and Kılınç-Karzan (2019) have used these concepts for *standard* JEO. Our work, however, aims to provide an avenue for using OCO to solve the *saddle point* JEO problem (8), where one of the domains involved faces an additional high dimensionality challenge. All of these developments are motivated by the problem of dynamic non-parametric estimation of a choice model. In Sections 2-4, we have introduced this problem formally and discuss the derivation of efficient algorithms for this problem specifically using our general SP JEO framework.

Appendix E: Interpretations of Algorithms from Section 4.2

We discuss interpretations of our methods when applied to $Z = Y$, $z_t = y_t$ for $t \in [T]$ and $h_t(\cdot) = -\Psi(x_t, \cdot; p_t)$.

REMARK EC.1. For Theorems 3 and 4, observe that since $x_t = \arg\min_{x \in X} \Psi(x, y_t; p_t)$, $\nabla h_t(z_t) = -\nabla_{y_t}\Psi(x_t, y_t; p_t) = -\nabla\left(\min_{x \in X} \Psi(x, y_t; p_t)\right) = -\nabla g(y_t; p_t)$. Thus, the gradients we compute are simply the gradients of the (negative of) the dual function $g(y; p_t)$ from (7b). In other words, if $p_t = p$ for $t \in [T]$, then using Theorem 3 or 4 to minimize the dual regret $\mathcal{R}_y$ is simply performing the well-known Mirror Descent algorithm to solve the dual problem (7b). When $p_t \neq p$, these update rules are simply performing Mirror Descent on approximate versions of (7b), with built-in guarantees on the error when $p_t \neq p$. ∎

REMARK EC.2. For Theorem 5, let us consider the case when $p_t = p$ for $t \in [T]$. Then y_t are computed as

$$y_{t+1} = \arg\max_{y \in Y} \left\{ \frac{1}{\Theta_t} \sum_{s \in [t]} \theta_s \Psi(x_s, y; p) \right\}.$$

If, furthermore, $\Psi(x, y; p)$ is linear in x, i.e., it is of the form $\Psi(x, y; p) = \langle x, \mathbf{\Psi}(y; p) \rangle - \alpha\omega(y)$, then we can push the sum into the inner product, and since $f(x; p)$ from (7a) is of the form $f(x; p) = \max_{y \in Y} \Psi(x, y; p) = \max_{y \in Y} \{ \langle x, \mathbf{\Psi}(y; p) \rangle - \alpha\omega(y) \}$, by letting $\bar{x}_t^\theta = \frac{1}{\Theta_t} \sum_{s \in [t]} \theta_s x_s$ and using the convex envelope theorem we have

$$y_{t+1} = \arg\max_{y \in Y} \left\{ \left\langle \bar{x}_t^\theta, \mathbf{\Psi}(y; p) \right\rangle - \alpha\omega(y) \right\}, \quad \mathbf{\Psi}(y_{t+1}; p) = \nabla f(\bar{x}_t^\theta; p).$$

Now, recalling that

$$x_{t+1} = \arg\min_{x \in X} \Psi(x, y_{t+1}; p) = \arg\min_{x \in X} \langle x, \mathbf{\Psi}(y_{t+1}; p) \rangle = \arg\min_{x \in X} \langle x, \nabla f(\bar{x}_t^\theta; p) \rangle,$$

we deduce $\bar{x}_{t+1}^\theta = (1 - \gamma_{t+1})\bar{x}_t^\theta + \gamma_{t+1}x_{t+1}, \quad \gamma_{t+1} = \dfrac{\theta_{t+1}}{\Theta_{t+1}} \in [0, 1].$

Note that this is exactly a Frank-Wolfe update for the current average point $\bar{x}_t^\theta$ on the primal function $f(x;p)$. Therefore, when Ψ is linear in x and $p_t = p$ for all $t \in [T]$, using Theorem 5 to minimize $\mathcal{R}_y$ is equivalent to using the Frank-Wolfe (F-W) algorithm to solve the primal problem (7a). See also Abernethy et al. (2018) for an equivalent observation in the case of a particular type of Ψ arising from the convex conjugate of f. Thus, within the general context of the JEO problem (8), we can think of using Theorem 5 as a generalization of the F-W algorithm to the dynamic setup, with built-in error guarantees for $p_t \neq p$. ∎

Appendix F: Rates and Comparison to Frank-Wolfe Methods

In our choice model estimation problem, the high-dimensionality challenge of the domain X necessitates the use of projection/prox-free algorithms. In this respect, our developments for the JEO problem (4) can be compared against the classical Frank-Wolfe (F-W) algorithm (see Jaggi (2013), Freund and Grigas (2016)), which admits guarantees when using *approximate gradients.* Indeed, under certain assumptions which we will carefully examine, we can think of using $p_t \approx p$ as an approximate gradient method, i.e., $\nabla_x D(x, p_t) \approx \nabla_x D(x, p)$. Alternatively, Devolder et al. (2014) considers projection-type first-order methods for smooth functions, and provides guarantees on using approximate gradient oracles within such algorithms. Since projecting onto X defined in (2) for the choice model estimation problem is difficult due to the high dimensionality of the domain X, we will not discuss the methods of Devolder et al. (2014) and instead focus on the F-W algorithm with approximate gradient oracles in Freund and Grigas (2016).

In this appendix, we will do the following.

• We first examine the applicability of the standard F-W method due to the issue of the non-smoothness of the objective function for the norm-based distance measures $D(x, p)$.

• We then show how online Mirror Descent (MD, Theorem 3) can circumvent the non-smoothness and give the corresponding rates obtainable.

• We then present a simple technique to smooth an ℓ_q-norm for $q \in [2, \infty)$ by squaring it. We examine naïvely applying the F-W method to solve (4) and compare the guarantees for F-W with approximate gradients from Freund and Grigas (2016) to the guarantees from using MD in our framework. We find that the data error terms from the naïve F-W method are worse than the ones from using MD in our framework, while the regret bound terms are comparable asymptotically, but involve worse constant factors.

• Finally, we present the so-called 'Nesterov smoothing' technique for more general norms. We examine naïvely applying the F-W method to solve (4) using Nesterov smoothing, and compare the guarantees for F-W with approximate gradients from Freund and Grigas (2016) to the guarantees from using a F-W algorithm derived from our framework (see Remark EC.2). We find that while

the regret bound terms are comparable in both settings, the data error terms for the naïve F-W method require a certain rate of convergence of $\|p_t - p\| \to 0$ to vanish; this is not a problem for the F-W method derived from our framework.

F.1. Smoothness Requirement for the Frank-Wolfe Algorithm

It is known that the F-W method in general *does not* converge on non-smooth objectives; see Nesterov (2018, Example 1) that demonstrates this on a max-type objective function. In the JEO problem (4) when $D(x, p) = \|x - p\|$, e.g., as in Appendix B.3, $D(x, p)$ is non-smooth due to the norm. In addition, the usual convergence of the F-W algorithm relies on a finite curvature constant assumption that related to the smoothness properties of the function. In particular, the curvature constant C_D of a function $D(x, p)$ of the variable x is defined as

$$C_D := \sup_{\substack{x, s \in X \\ \alpha \in [0,1]}} \frac{1}{\alpha^2} \left(D((1 - \alpha)x + \alpha s, p) - D(x, p) - \alpha \langle s - x, \nabla_x D(x, p) \rangle \right). \tag{EC.12}$$

It is well-known that when the function is smooth and the domain is bounded, the associated curvature constant is finite; see e.g., Jaggi (2013, Lemma 7). Nevertheless, distance measures $D(x, p)$ of interest in the case of non-parametric choice estimation problem, e.g., from Appendix B are non-smooth. Moreover, we next show that when $D(x, p)$ is set up based on the norm (see Appendix B.3) or the KL divergence as in van Ryzin and Vulcano (2015) (see Appendix B.2) the associated curvature constant of $D(x, p)$ is infinite as well.

PROPOSITION EC.2. *Suppose $n > 2$. For any $q \in [1, \infty]$, the function $D(x, p) = \|x - p\|_q$ has infinite curvature constant (EC.12) for any $p \in X$. Furthermore, when the MLE based weighted KL divergence is used, i.e., $D(x, p) = \sum_{j \in [m]} w_j \, \mathrm{KL}(p_j, x_j)$ for any positive weights w_j, and $p \in X$ such that $p_{ij} > 0$ for all $i \in A_j$, the curvature constant is infinite.*

Proof.   We will first show that the curvature constant C_D defined in (EC.12) of $D(x, p) = \|x - p\|$ is infinite for any $p \in X$. Let us choose $x = p$, reserving the choice of $\alpha \in [0, 1]$ and $s \in X$ for later. Then $D(x, p) = 0$, $D((1 - \alpha)x + \alpha s, p) = \alpha \|s - p\|$, and the subgradients of $D(x, p)$ are $\{y : \|y\|_* \leq 1\}$. Thus, for any selection of subgradient mapping $y(\hat{x}) \in \nabla_x D(\hat{x}, p)$ we have

$$\frac{1}{\alpha^2} \left[D((1 - \alpha)x + \alpha s, p) - D(x, p) - \alpha \langle s - x, y(x) \rangle \right] = \frac{1}{\alpha^2} \left[\alpha \|s - p\| - \alpha \langle s - p, y(x) \rangle \right]$$

$$= \frac{1}{\alpha} \left[\|s - p\| - \langle s - p, y(x) \rangle \right].$$

Note that whenever there is a choice $s \in X$ with $\|s - p\| - \langle s - p, y(p) \rangle > 0$, we can send $\alpha \to 0$ and conclude that the curvature constant C_D is infinite.

To choose the appropriate s, we denote the set of subgradients of $\| \cdot \|_q$ at $s - p$ as $G_{\|\cdot\|}(s - p)$. Observe that for a norm $\| \cdot \|$, if $y \in G_{\|\cdot\|}(s - p)$ then $\|y\|_* \leq 1$ and $\langle s - p, y \rangle = \|s - p\|$. Thus, we

need to choose $s \in X$ such that $y(x) \notin G_{\|\cdot\|}(s-p)$. To do this, we exploit the following property of ℓ_q-norms. It is simple to check that for $q \in [1, \infty]$ and $y \in G_{\|\cdot\|_q}(s-p)$, we have the property that $y_{ij} > 0 \implies s_{ij} - p_{ij} > 0$. For our selection $y(x)$, first suppose that there exists $i \in \mathcal{A}_j$ such that $y(x)_{ij} > 0$. Then a ranking σ that ranks i last will have $a(\sigma)_{ij} = 0$, so $a(\sigma)_{ij} - p_{ij} \leq 0$ because $p_{ij} \geq 0$. We cannot have $p = a(\sigma)$ for all $(n-1)!$ rankings σ that ranks i last (note that $n > 2$); hence, there exists one σ such that $a(\sigma) \neq p$, and we choose $s = a(\sigma)$. This implies that $y(x)_{ij} > 0$ while $s_{ij} - p_{ij} \leq 0$, hence $y(x) \notin G_{\|\cdot\|_q}(s-p)$. Now suppose that $y(x)_{ij} \leq 0$ for all item-subset pairs (i,j). If $y(x) = 0$, then the result follows trivially by choosing any $s \neq p$. Suppose now there exists some $y(x) < 0$. It is again simple to check that for $q \in [1, \infty]$ and $y \in G_{\|\cdot\|_q}(s-p)$, we have the property that $y_{ij} < 0 \implies s_{ij} - p_{ij} < 0$. Then a ranking σ that ranks i first will have $a(\sigma)_{ij} = 1$, so $a(\sigma)_{ij} - p_{ij} \geq 0$ because $p_{ij} \leq 1$. We cannot have $p = a(\sigma)$ for all $(n-1)!$ rankings σ that ranks i first, so there exists one such that $a(\sigma) \neq p$, and we choose $s = a(\sigma)$. This implies that $y(x)_{ij} < 0$ while $s_{ij} - p_{ij} \geq 0$, hence $y(x) \notin G_{\|\cdot\|_q}(s-p)$. Thus, in all cases for $y(x)$, we can choose the appropriate $s \in X$.

Now consider the weighted KL-divergence $D(x,p) = -\sum_{j \in [m]} w_j \sum_{i \in \mathcal{A}_j} p_{ij} \log(x_{ij}/p_{ij})$. We can assume that $p_{ij} > 0$ by simply ignoring terms in the sum for which $p_{ij} = 0$. Choose $x = p$, which ensures that $D(\cdot, p)$ is differentiable at x with $\nabla_x D(x,p)_{ij} = -w_j/x_{ij}$. Then we have

$$\frac{1}{\alpha^2} \left[D((1-\alpha)x + \alpha s, p) - D(x,p) - \alpha \langle s - x, \nabla_x D(x,p) \rangle \right]$$
$$= -\frac{1}{\alpha^2} \sum_{j \in [m]} w_j \sum_{i \in \mathcal{A}_j} p_{ij} \log\left(1 - \alpha + \alpha \frac{s_{ij}}{p_{ij}}\right) + \frac{1}{\alpha} \sum_{j \in [m]} w_j \sum_{i \in \mathcal{A}_j} \left(\frac{s_{ij}}{p_{ij}} - 1\right).$$

Note that the second term is bounded by $\frac{1}{\alpha} \left(\sum_{j \in [m]} w_j\right) (\max_{i,j} 1/p_{ij} - 1)$. Choose $s_{ij} = a(\sigma)$ for any $\sigma \in S_n$. Then there exists some i, j such that $s_{ij} = 0$. Sending $\alpha \to 1$ results in $\log\left(1 - \alpha + \alpha \frac{s_{ij}}{p_{ij}}\right) \to \infty$, and the second term is bounded, so the curvature constant C_D is infinite. $\quad\square$

Note that our framework proposes to handle the non-smoothness of $D(x,p) = \|x - p\|$ due to the norm by defining $\Psi(x, y; p)$ appropriately, and suggests to utilize a regret-minimizing algorithm to bound the dual regret $\mathcal{R}_y$, from which primal optimality gap bounds can be inferred through (9) (since the dual optimality gap is always non-negative).

F.2. Basic Convergence Rate Using the Mirror Descent Algorithm

We first discuss convergence rates that we can derive within our framework. Since norms are non-smooth, this immediately suggests the utilization of the Mirror Descent algorithm (Theorem 3). Recall that our assumption on D is that it has a representation (5a):

$$D(x,p) = \max_{y \in Y} \left\{ \langle B(x-p), y \rangle - \alpha \omega(y) \right\}.$$

q	Y	B	$\omega(y)$	$\|\cdot\|_\omega$	Ω	G	$\sqrt{\frac{2\Omega G^2}{T}}$
1	$\{\|y\|_\infty \le 1\} \subset \mathbb{R}^N$	I_N	$\frac{1}{2}\|y\|_2^2$	$\|\cdot\|_2$	$N/2$	$\sqrt{2m}$	$\sqrt{\frac{2mN}{T}}$
$1 < q < 2$	$\{\|y\|_{\frac{q}{q-1}} \le 1\} \subset \mathbb{R}^N$	I_N	$\frac{1}{2}\|y\|_2^2$	$\|\cdot\|_2$	$N^{\frac{2-q}{q}}/2$	$\sqrt{2m}$	$\sqrt{\frac{2mN^{\frac{2-q}{q}}}{T}}$
$2 \le q < \infty$	$\{\|y\|_{\frac{q}{q-1}} \le 1\} \subset \mathbb{R}^N$	I_N	$\frac{q-1}{2}\|y\|_{\frac{q}{q-1}}^2$	$\|\cdot\|_{\frac{q}{q-1}}$	$(q-1)/2$	$(2m)^{1/q}$	$\sqrt{\frac{(2m)^{2/q}(q-1)}{T}}$
∞	$\Delta_{2N} \subset \mathbb{R}^{2N}$	$\begin{bmatrix} I_N & -I_N \end{bmatrix}$	$\sum\limits_{k\in[2N]} y_k \log(y_k)$	$\|\cdot\|_1$	$\log(2N)$	1	$\sqrt{\frac{2\log(2N)}{T}}$

Table EC.2   Different q-norms and their Mirror Descent constants (Theorem 3).

The first row of Table 1 shows that any norm $\|\cdot\|$ can be written in the form (5a) with $\alpha = 0$, $B = I_N$ and $Y = \{y \in \mathbb{R}^N : \|y\|_* \le 1\}$. Define

$$\tilde{X} := \left\{ x \ge 0 : \sum_{i \in A_j} x_{ij} = 1, \ j \in [m] \right\} \subset \mathbb{R}^N,$$

and note that $X \subset \tilde{X}$, $p_t \in \tilde{X}$ for all $t \in [T]$ (by (1)). Thus, assuming that ω is 1-strongly convex with respect to some possibly different norm $\|\cdot\|_\omega$, and defining

$$\Omega = \max_{\|y\|_* \le 1} \omega(y) - \min_{\|y\|_* \le 1} \omega(y), \quad G \ge \max_{x \in X, p' \in \tilde{X}} \|x - p'\|_{\omega,*},$$

using Theorem 3 and Proposition 3 we get the sub-optimality bound

$$\|\bar{x}_T^\theta - p\| - \min_{x \in X} \|x - p\| \le \sqrt{\frac{2\Omega G^2}{T}} + \frac{G_Y\|B\|}{T} \sum_{t \in [T]} \|p_t - p\|, \tag{EC.13}$$

where G_Y is as defined in Proposition 3, which in our case will be 1.

Table EC.2 gives a summary of possible choices of ω for different q-norms, together with the associated constants Ω, G, and the sub-optimality bound. In general, ω should be chosen so that ΩG^2 is small, and the operation $\arg\min_{y \in Y} \{\langle z, y \rangle + \omega(y)\}$ is easy to compute for any z.

F.3. Convergence Guarantees of the Frank-Wolfe Method for Smooth Minimization

For a function $D(x,p)$ which is L-smooth with respect to a norm $\|\cdot\|$, the standard convergence rate for the F-W method, fixing $p_t = p$, is given in Freund and Grigas (2016, Bound 3.1, Eq. 8):

$$D(\bar{x}_T, p) - \min_{x \in X} D(x, p) \le \frac{2L}{T+4} \max_{x,x' \in X} \|x - x'\|.$$

Therefore, if we can use an alternative function D instead of a norm then relate the optimality gap back to the original norm, we can get convergence guarantees for the norm.

In the dynamic setting, however, we need to use an approximate F-W method, since $p_t \neq p$. Essentially, this means that at each iteration t, we have an approximate gradient $\nabla_x D(x, p_t) \approx \nabla_x D(x, p)$, which we require to satisfy

$$\max_{x,x' \in X} |\langle \nabla_x D(x, p_t) - \nabla_x D(x, p), x - x' \rangle| \le \delta_t. \tag{EC.14}$$

We update according to

$$x_t = \underset{x' \in X}{\arg\min} \langle \nabla_x D(\bar{x}_t, p_t), \, x' \rangle, \quad \bar{x}_t = (1 - \gamma_t)\bar{x}_{t-1} + \gamma_t x_t. \tag{EC.15}$$

When working with approximate gradients, if (EC.14) holds and we choose γ_t appropriately, Freund and Grigas (2016, Theorem 5.1, Proposition 5.1) provides a convergence rate of

$$D(\bar{x}_T, p) - \min_{x \in X} D(x, p) \leq \frac{2L}{T + 4} \max_{x, x' \in X} \|x - x'\|^2 + \frac{4}{(T+1)(T+2)} \sum_{t \in [T]} (t+1)\delta_t. \tag{EC.16}$$

Since norms are non-smooth, we cannot apply the F-W algorithm directly; recall that the F-W method in general does not converge on non-smooth objectives (see Nesterov (2018, Example 1)). We next discuss two alternative techniques to build smooth approximations to norms.

F.3.1. Smoothing via Squaring the Norm   A simple way we can smooth a norm is to square it, i.e., $D(x, p) = \|x - p\|^2$. This does not work for arbitrary norms, but it is known that for ℓ_q-norms, $\frac{1}{2}\|\cdot\|_q^2$ is $(q-1)$-smooth for $q \in [2, \infty)$.

Convergence Rate with Approximate Gradients. Using (EC.16), the naïve F-W algorithm using approximate gradients achieves a bound of (with $D_X := \max_{x, x' \in X} \|x - x'\|_q = (2m)^{1/q}$):

$$\|\bar{x}_T - p\|_q - \min_{x \in X} \|x - p\|_q \leq \sqrt{\|\bar{x}_T - p\|_q^2 - \min_{x \in X} \|x - p\|_q^2}$$

$$\leq \sqrt{\frac{8(2m)^{2/q}(q-1)}{T+4}} + \sqrt{\frac{4(2m)^{1/q}(q-1)}{(T+1)(T+2)} \sum_{t \in [T]} (t+1)\|p_t - p\|_q}. \tag{EC.17}$$

The first term of (EC.17) is asymptotically comparable to the first term of (EC.13) (third row of Table EC.2), but with worse constants. From Appendix C, we see that the second (data error) term of (EC.17) is asymptotically worse than the data error term in (EC.13) for a variety of rates of convergence of $\|p_t - p\| \to 0$, and also has worse constants.

The data error term is derived as follows. The smoothness of $\|x - p\|_q^2$ is derived from the smoothness of $d(z) = \frac{1}{2}\|z\|_q^2$, which is equivalent to

$$\|\nabla d(z) - \nabla d(z')\|_* \leq (q-1)\|z - z'\|.$$

Now observe that

$$\|\nabla_x D(x, p_t) - \nabla_x D(x, p)\|_{q,*} = \|\nabla d(x - p_t) - \nabla d(x - p)\|_{q,*} \leq (q-1)\|p_t - p\|_q.$$

Therefore, we have

$$\begin{aligned} \delta_t &:= \max_{x, x' \in X} |\langle \nabla_x D(x, p_t) - \nabla_x D(x, p), \, x - x' \rangle| \\ &\leq \max_{x, x' \in X} \{ \|\nabla_x D(x, p_t) - \nabla_x D(x, p)\|_{q,*} \|x - x'\|_q \} \\ &\leq (q-1)\|p_t - p\|_q \max_{x, x' \in X} \|x - x'\|_q = (q-1)\|p_t - p\|_q D_X. \end{aligned}$$

The technique of squaring the norm only works for particular $q \in [2, \infty)$. We next propose a more generally applicable method for smoothing.

F.3.2. Smooth Approximations to General Norms   We can smooth more general norms by utilizing the 'Nesterov smoothing' framework outlined in Beck and Teboulle (2012). Interestingly, this also allows us to make use of other algorithms within our framework such as Theorems 4 and 5.

Given a function $g : Z \to \mathbb{R}$, we say that $g_\beta : Z \to \mathbb{R}$ is an β-approximation of g if there exists $\beta_1, \beta_2 \geq 0$, $\beta_1 + \beta_2 = 1$ such that the following holds:

$$\forall z \in Z, \quad g(z) - \beta_1\beta \leq g_\beta(z) \leq g(z) + \beta_2\beta. \tag{EC.18}$$

LEMMA EC.1. *Suppose g, g_β satisfies* (EC.18). *Then for any $\bar{z} \in Z$,*

$$g(\bar{z}) - \min_{z \in Z} g(z) \leq g_{L,\beta}(\bar{z}) - \min_{z \in Z} g_{L,\alpha}(z) + \beta.$$

Proof of Lemma EC.1.   Observe that $g(\bar{z}) \leq g_\beta(\bar{z}) + \beta_1\alpha$. Second, observe that $\min_{z \in Z} g(z) \geq \min_{z \in Z} g_\beta(z) - \beta_2\alpha$. Subtracting the appropriate terms gives us the result.   □

Thus, through approximating the non-smooth distance measures $D(x, p)$ with smooth functions, we can use the F-W algorithm and get sub-optimality bounds. For this, we follow the so-called 'Nesterov smoothing' technique outlined in Beck and Teboulle (2012, Section 4.3): by setting $\alpha > 0$ instead of $\alpha = 0$ in (5a) and choosing ω to be strongly convex, we get a smooth approximation of the norm with the following guarantee.

LEMMA EC.2. *Let*

$$D(x, p) = \max_{y \in Y} \langle B(x - p), y \rangle, \quad D_\alpha(x, p) = \max_{y \in Y} \{\langle B(x - p), y \rangle - \alpha\omega(y)\}.$$

Suppose that $\omega(y) \geq 0$ for all $y \in Y$. Then

$$D(x, p) - \alpha \max_{y \in Y} \omega(y) \leq D_\alpha(x, p) \leq D(x, p).$$

Similarly, if $\omega(y) \leq 0$ for all $y \in Y$, then

$$D(x, p) \leq D_\alpha(x, p) \leq D(x, p) + \alpha \max_{y \in Y} |\omega(y)|.$$

Proof of Lemma EC.2.   When $\omega(y) \geq 0$ for all $y \in Y$, we have

$$\langle B(x - p), y \rangle \geq \langle B(x - p), y \rangle - \alpha\omega(y) \geq \langle B(x - p), y \rangle - \alpha \max_{y' \in Y} \omega(y').$$

Taking the maximum over $y \in Y$ of all sides gives the first result. The second result is proved similarly.   □

q	Y	B	$\omega(y)$	$\displaystyle\max_{y\in Y}\lvert\omega(y)\rvert$	$D_\alpha(x,p)$
1	$\{\lVert y\rVert_\infty\le 1\}\subset\mathbb{R}^N$	I_N	$\frac{1}{2}\lVert y\rVert_2^2$	$N/2$	$\displaystyle\sum_{j\in[m]}\sum_{i\in\mathcal{A}_j}H_\alpha(x_{ij}-p_{ij})$
$1<q<2$	$\{\lVert y\rVert_{\frac{q}{q-1}}\le 1\}\subset\mathbb{R}^N$	I_N	$\frac{1}{2}\lVert y\rVert_2^2$	$N^{\frac{2-q}{q}}/2$	n/a
$2\le q<\infty$	$\{\lVert y\rVert_{\frac{q}{q-1}}\le 1\}\subset\mathbb{R}^N$	I_N	$\frac{q-1}{2}\lVert y\rVert_{\frac{q}{q-1}}^2$	$(q-1)/2$	$H_{\alpha(q-1)}(\lVert x-p\rVert_q)$
∞	$\Delta_{2N}\subset\mathbb{R}^{2N}$	$[I_N\ {-}I_N]$	$\displaystyle\sum_{k\in[2N]}y_k\log(y_k)$	$\log(2N)$	$\displaystyle\alpha\log\left(\sum_{j\in[m]}\sum_{i\in\mathcal{A}_j}2\cosh\left(\frac{x_{ij}-p_{ij}}{\alpha}\right)\right)$

Table EC.3   Different q-norms and their smoothings.

EXAMPLE EC.1. Let $D(x,p)=\lVert x-p\rVert=\max_y\{\langle x-p,y\rangle:\ :\lVert y\rVert_*\le 1\}$. Define the Huber function

$$H_\alpha(r)=\begin{cases}\frac{1}{2\alpha}r^2, & r<\alpha\\ \lvert r\rvert-\frac{\alpha}{2}, & r\ge\alpha.\end{cases}\tag{EC.19}$$

Setting $\omega(y)=\frac{1}{2}\lVert y\rVert_*^2$, we have

$$D_\alpha(x,p)=\max_{y:\lVert y\rVert_*\le 1}\left\{\langle x-p,y\rangle-\frac{\alpha}{2}\lVert y\rVert_*^2\right\}=\max_{\substack{\gamma\in[0,1]\\ y:\lVert y\rVert_*=1}}\left\{\gamma\langle x-p,y\rangle-\frac{\alpha}{2}\gamma^2\right\}=\max_{\gamma\in[0,1]}\left\{\gamma\lVert x-p\rVert-\frac{\alpha}{2}\gamma^2\right\}$$

$$=\begin{cases}\frac{1}{2\alpha}\lVert x-p\rVert^2, & \lVert x-p\rVert\le\alpha\\ \lVert x-p\rVert-\frac{\alpha}{2}, & \lVert x-p\rVert>\alpha\end{cases}$$

$$=H_\alpha(\lVert x-p\rVert).$$

This guarantees that $H_\alpha(\lVert x-p\rVert)$ is a $\alpha/2$-approximation of $\lVert x-p\rVert$. Note that $\omega(y)=\frac{1}{2}\lVert y\rVert_*^2$ is not strongly convex in general, which we need in order to use $\omega(y)$ in Theorems 4 or 5. However, if we consider q-norms $\lVert\cdot\rVert=\lVert\cdot\rVert_q$ for $2\le q<\infty$, then it is well-known that $\frac{1}{2}\lVert y\rVert_*^2=\frac{1}{2}\lVert y\rVert_{\frac{q}{q-1}}^2$ is strongly convex (with respect to $\lVert\cdot\rVert_{\frac{q}{q-1}}$) with parameter $(q-1)^{-1}$.

For $q=1$, we can take $\omega(y)=\frac{1}{2}\lVert y\rVert_2^2$ to also get an $N/2$-approximation. For $1<q<2$, we use the same ω to get a $N^{\frac{2-q}{q}}/2$-approximation (which is because $\max_{\lVert y\rVert_{\frac{q}{q-1}}\le 1}\lVert y\rVert_2=N^{\frac{1}{2}-\frac{q-1}{q}}$). However, notice that there is no closed form for D_α, but this is not a problem for our framework as long as we have the max-type representation (5a) for D_α. For $q=\infty$, we can use the lifted representation of the ℓ_1-ball into the $2N$-simplex, and set $\omega(y)$ to be the negative entropy, which gets us a $\log(2N)$-approximation. A summary of this is given in Table EC.3; notice that the choices of ω and associated constants are quite similar to Table EC.2. ∎

Convergence Rate with Approximate Gradients. Using a similar argument to the one in Appendix F.3.1, we can derive the following guarantee on the approximate gradients $\nabla_x D_\alpha(x,p_t)$ for functions (5a):

$$\delta_t:=\max_{x,x'\in X}\lvert\langle\nabla_x D_\alpha(x,p_t)-\nabla_x D_\alpha(x,p),x-x'\rangle\rvert$$
$$\le\max_{x,x'\in X}\lVert\nabla_x D_\alpha(x,p_t)-\nabla_x D_\alpha(x,p)\rVert_*\lVert x-x'\rVert$$
$$\le\frac{1}{\alpha}\lVert p_t-p\rVert\max_{x,x'\in X}\lVert x-x'\rVert.$$

q	Y	$\omega(y)$	$\max\limits_{y\in Y}\lvert\omega(y)\rvert$	$G=D_X$	α^*	$\frac{2G^2}{\alpha^*(T+1)}+\alpha^*\max\limits_{y\in Y}\lvert\omega(y)\rvert$
1	$\{\lVert y\rVert_\infty \le 1\}\subset\mathbb{R}^N$	$\frac12\lVert y\rVert_2^2$	$N/2$	$\sqrt{2m}$	$\sqrt{\frac{8m}{N(T+1)}}$	$\sqrt{\frac{8mN}{T+1}}$
$1<q<2$	$\{\lVert y\rVert_{\frac{q}{q-1}}\le 1\}\subset\mathbb{R}^N$	$\frac12\lVert y\rVert_2^2$	$N^{\frac{2-q}{q}}/2$	$\sqrt{2m}$	$\sqrt{\frac{8m}{N^{\frac{2-q}{q}}(T+1)}}$	$\sqrt{\frac{8mN^{\frac{2-q}{q}}}{T+1}}$
$2\le q<\infty$	$\{\lVert y\rVert_{\frac{q}{q-1}}\le 1\}\subset\mathbb{R}^N$	$\frac{q-1}{2}\lVert y\rVert_{\frac{q}{q-1}}^2$	$(q-1)/2$	$(2m)^{1/q}$	$\sqrt{\frac{4(2m)^{2/q}}{(q-1)(T+1)}}$	$\sqrt{\frac{4(2m)^{1/q}(q-1)}{T+1}}$
∞	$\Delta_{2N}\subset\mathbb{R}^{2N}$	$\sum\limits_{k\in[2N]} y_k\log(y_k)$	$\log(2N)$	1	$\sqrt{\frac{2}{\log(2N)(T+1)}}$	$\sqrt{\frac{8\log(2N)}{T+1}}$

Table EC.4   Different q-norms and their (smoothed) convergence rates.

Thus, by denoting $D_X := \max_{x,x'\in X}\lVert x-x'\rVert$, (EC.16) gives the following guarantee from using the naïve F-W method for the JEO problem (4):

$$D_\alpha(\bar{x}_T,p)-\min_{x\in X} D_\alpha(x,p)\le \frac{2D_X^2}{\alpha(T+4)}+\frac{4D_X}{\alpha(T+1)(T+2)}\sum_{t\in[T]}(t+1)\lVert p_t-p\rVert.$$

Translating this back into a bound on the norm (Lemmas EC.1, EC.2), we have

$$\lVert\bar{x}_T-p\rVert-\min_{x\in X}\lVert x-p\rVert\le \frac{2D_X^2}{\alpha(T+4)}+\alpha\max_{y\in Y}\lvert\omega(y)\rvert+\frac{4D_X}{\alpha(T+1)(T+2)}\sum_{t\in[T]}(t+1)\lVert p_t-p\rVert. \quad\text{(EC.20)}$$

Let us contrast this guarantee with what is achievable in our primal-dual framework by using Theorem 5 and Proposition 3. Recall from Remark EC.2 that this is actually a variant of F-W. By defining

$$G:=\max_{x\in X,p'\in\tilde{X}}\lVert x-p'\rVert,$$

we get the bound

$$\lVert\bar{x}_T^\theta-p\rVert-\min_{x\in X}\lVert x-p\rVert\le D_\alpha(\bar{x}_T^\theta,p)-\min_{x\in X}D_\alpha(x,p)+\alpha\max_{y\in Y}\lvert\omega(y)\rvert$$
$$\le \frac{2G^2}{\alpha(T+1)}+\alpha\max_{y\in Y}\lvert\omega(y)\rvert+\frac{2G_Y\lVert B\rVert}{T(T+1)}\sum_{t\in[T]}t\lVert p_t-p\rVert, \quad\text{(EC.21)}$$

where $G_Y,\lVert B\rVert$ are defined as in Proposition 3; for norms we have $G_Y=1$, $\lVert B\rVert\le 2$. Comparing (EC.20) and (EC.21), we notice that $G\approx D_X$, hence, the first terms are quite comparable. Also, except for an additive factor $\alpha\max_{y\in Y}\lvert\omega(y)\rvert$, these $O(1/T)$ rates in theory are faster than the $O(1/\sqrt{T})$ rates we get from the non-smooth Mirror Descent. However, if we choose α optimally taking this term into account explicitly, we get rates that are asymptotically comparable with those of the non-smooth Mirror Descent, but with worse constants. We summarize this in Table EC.4.

The main difference between the bounds for the naïve F-W method and the variant from our framework displayed in (EC.20) and (EC.21) respectively is, since our continuity requirement on Ψ (Assumption 1) is different to (EC.14), α does not appear in the third data error term of (EC.21). This is significant because choosing $\alpha\propto 1/\sqrt{T}$ to minimize the first two terms now makes the third

data error term $\frac{1}{\sqrt{T}(T+1)} \sum_{t \in [T]} t\|p - p_t\|$, which may diverge even if $\|p - p_t\| \to 0$. Thus, to ensure convergence using the naïve F-W method on the smoothed norm with approximate gradients, we need $\|p - p_t\| \to 0$ *sufficiently fast*. In contrast, our primal-dual framework avoids this obstacle completely. We note also that for norms, $G_Y\|B\| \leq 2$, so we can also see from Table EC.4 that the constants in the error terms of (EC.21) are better than (EC.20).

Appendix G: Computational Study Details and Supplementary Results

We describe in detail the experimental setup of our computational study. All experiments are conducted on a server with 2.8 GHz processor and 64GB memory, using Python 3.6. Gurobi 8.0 (with default Gurobi settings except we limit the number of threads to 2) is used to solve the integer programming subproblems.

Test Instances. We employ a setup similar to Mišić (2016, Chapter 4.5.3). Our ground truth choice model over $n = 10$ items (plus one no-choice option) is a mixed MNL model with K segments. Given mixing probabilities $w \in \Delta_K$ and K sets of utilities $\{u_{i,k}\}_{i \in \{0\} \cup [n]}$, $k \in [K]$, the mixed MNL model chooses an item $i \in \mathcal{A} \subseteq [n]$ with probability

$$\mathbb{P}[i \mid \mathcal{A}] = \sum_{k \in [K]} w_k \frac{u_{i,k}}{u_{0,k} + \sum_{i' \in \mathcal{A}} u_{i',k}}.$$

For each $k \in [K]$, we generate $n + 1$ parameters $q_{i,k} \sim U(0,1)$, $i \in \{0\} \cup [n]$ (recall that 0 denotes the no-choice option present in each subset). The utilities $u_{i,k}$ are then set as follows: four randomly chosen $i \in \{0\} \cup [n]$ are set to $u_{i,k} = Lq_{i,k}$ while the rest are set to $u_{i,k} = q_{i,k}/10$. The mixing probabilities $\{w_k\}_{k \in [K]}$ are chosen randomly from the $(K-1)$-dimensional simplex. We test on 100 randomly generated instances of this ground truth model. In the main discussion presented before, we showed results in the dynamic setting for $K = L = 5$. In Appendix G.1, we also provide results in the static setting under various parameter regimes $K \in \{1, 5, 10\}$ and $L \in \{5, 10, 100\}$, and thus test the effect of different ground truth models on the conclusions drawn. We observe that in these different ground truth models, the conclusions are in general in line with the ones from $K = L = 5$ setting; this supports that the conclusions drawn from the dynamic data experiments with this particular choice of ground truth choice model are likely to be valid for other ground truth choice models as well.

We tested each algorithm on 100 different instances of a ground truth model for each parameter combination.

Distance measures. In our main discussion, we consider distance measures $D(x, p)$ based on the ℓ_2-norm. Specifically, we tested $D(x, p)$ from the first and the third rows of Table 1, as well as $D(x, p) = \frac{1}{2}\|x - p\|_2^2$, which has max-type representation $\frac{1}{2}\|x - p\|_2^2 = \max_y \left\{ \langle x - p, y \rangle - \frac{1}{2}\|y\|_2^2 \right\}$. In this appendix, we also consider different choices of distance measures $D(\cdot, \cdot)$ arising from ℓ_1- and

$\|\cdot\|$	Y	B	$\omega(y)$	$D(x,p)$
$\|\cdot\|_1$	$\{\|y\|_\infty \leq 1\} \subset \mathbb{R}^N$	I_N	$\frac{1}{2}\|y\|_2^2$	$\displaystyle\sum_{j\in[m]}\sum_{i\in A_j} H_\alpha(x_{ij}-p_{ij})$
$\|\cdot\|_2$	$\{\|y\|_2 \leq 1\} \subset \mathbb{R}^N$	I_N	$\frac{1}{2}\|y\|_2^2$	$H_\alpha(\|x-p\|_2)$
$\|\cdot\|_\infty$	$\Delta_{2N} \subset \mathbb{R}^{2N}$	$\begin{bmatrix} I_N & -I_N \end{bmatrix}$	$\displaystyle\sum_{k\in[2N]} y_k \log(y_k)$	$\alpha \log\left(\displaystyle\sum_{j\in[m]}\sum_{i\in\mathcal{A}_j} 2\cosh\left(\frac{x_{ij}-p_{ij}}{\alpha}\right)\right)$

Table EC.5   Smoothed norms used in experiments, where $H_\alpha(r) = \frac{1}{2}r^2$ when $r < \alpha$ and $H_\alpha(r) = |r| - \alpha/2$ otherwise is the Huber function.

ℓ_∞-norms. Table EC.5 shows the distance measures we used in our experiments for ℓ_1-, ℓ_2- and ℓ_∞-norms respectively. Note that $\alpha = 0$ recovers the case for the underlying standard norm. Also, when we used a smoothed norm, we tuned the parameter α to minimize the suboptimality gap bound for the *non-smooth* norm. For example, if we have $D_\alpha(x,p)$ as the smoothed version of the ℓ_∞-norm, then using the F-W algorithm on the static problem, after T iterations we get the bound

$$D_\alpha(\bar{x}_T^\theta, p) - \min_{x \in X} D_\alpha(x, p) \leq \frac{16}{\alpha T}.$$

However, translating this back into a sub-optimality gap bound for the ℓ_∞-norm incurs an additional $\log(2N)\alpha$ additive error, thus we have

$$\|\bar{x}_T^\theta - p\|_\infty - \min_{x\in X}\|x-p\|_\infty \leq \frac{16}{\alpha T} + \log(2N)\alpha = 8\sqrt{\frac{\log(2N)}{T}},$$

where the last equality is setting $\alpha = 4/\sqrt{\log(2N)T}$ to minimize the upper bound. See Beck and Teboulle (2012) for further details on how this is done for other norms.

Algorithm implementation. We implemented the following solution methods based on our primal-dual framework. For smooth distance measures, i.e., when we have $\alpha > 0$ and ω is strongly convex in the representation (5a), we use the following algorithms:

- the naïve Frank-Wolfe (F-W) algorithm updating

$$x_{t+1} = \arg\min_{x'\in X}\langle \nabla_x D(\bar{x}_t^\theta, p_t), x'\rangle, \quad \bar{x}_{t+1}^\theta = \left(1 - \frac{\theta_{t+1}}{\Theta_{t+1}}\right)\bar{x}_t^\theta + \frac{\theta_{t+1}}{\Theta_{t+1}}x_{t+1},$$

with $\theta_t = t$. Note that for smooth distances D, convergence is guaranteed by our discussion in Appendix F.

- The modified F-W algorithm obtained by using Theorem 5 for the dual updates (see Remark EC.2). This is equivalent to updating

$$x_{t+1} = \arg\min_{x\in X}\langle \nabla_x D(\bar{x}_t^\theta, \bar{p}_t^\theta), x\rangle, \quad \bar{p}_t^\theta = \frac{1}{\Theta_t}\sum_{t\in[T]}\theta_t p_t, \quad \bar{x}_{t+1}^\theta = \left(1 - \frac{\theta_{t+1}}{\Theta_{t+1}}\right)\bar{x}_t^\theta + \frac{\theta_{t+1}}{\Theta_{t+1}}x_{t+1},$$

for $\theta_t = t$. Note that in the static case, when $p_t = p$ for every t, this is equivalent to the naïve F-W approach.

- The dual Mirror Descent (MD) algorithm from Theorem 4.

When D is non-smooth, i.e., it is the norm, we employed the dual MD algorithm from Theorem 3. Note that Theorem 3 employed constant step size policies based on constants Ω_Y, G and the maximum iteration count T; computing these constants depends on the particular norm and ω chosen for the domain Y (see Assumption 3), but are not difficult to obtain. For each of these methods and norms, we set a maximum iteration limit of $T = 10,000$.

In the dynamic setting, we use the algorithms outlined above, but only examine the non-parametric estimation model where the distance measure $D(\cdot, \cdot)$ is based on ℓ_2-norm, and the case of $m = 20$ subsets. In the static setting, we examine the effect of using different distance measures based on ℓ_1, ℓ_2, and ℓ_∞ norms as well as the number subsets $m \in \{10, 20, 50\}$. In Appendix G.1, we provide results for all of these cases for the static setting. The conclusions from these static setup experiments are in line with the base case we discuss here; therefore, we defer the results for distance measures based on other norms and the case of $m \in \{10, 50\}$ to Appendix G.1.

Our non-parametric estimation procedure is as follows. We first generate m subsets of $[n]$ of maximum size $\lfloor n/2 \rfloor$ uniformly at random. We append the no-choice option 0 to all of these (consequently the dimension of the domain X is $\dim(X) = 11! \approx 40,000,000$). Using the ground truth model, we compute the p_{train} vector, where $p_{\text{train},ij} = \mathbb{P}[i \mid A_j]$, and A_j is a subset from our training set. In the static setup, we set $p_t = p_{\text{train}}$ at each iteration. In the dynamic setup, we generated a sequence of $p_t \to p_{\text{train}}$, and at each iteration we supply p_t to the algorithm. We initially generate 2000 random choice observations (i, j), where A_j is one of the training subsets chosen randomly, and $i \in A_j$ is chosen with the probability $p_{\text{train},ij}$. We then compute p_1 using these observations according to (1). For $t \geq 2$, we generate $\kappa \in \mathbb{N}$ new observations, then update p_{t-1} with these new observations. We tested various choices of κ between 50 and 1000.

For both data regimes, we terminate training according to the mean absolute error (MAE), defined as $\mathrm{MAE}(p, p') = \frac{1}{\mathrm{length}(p)} \sum_{i,j} |p_{ij} - p'_{ij}|$, where $\mathrm{length}(p)$ is the length of the vector p. In the static setup, we terminate training when $\mathrm{MAE}(\bar{x}_t^\theta, p_{\text{train}}) \leq 0.001$, where $\bar{x}_t^\theta$ is the vector of choice probabilities (for our current estimated model) on the subsets used in training after t iterations. In the dynamic setup, we terminate training when $\mathrm{MAE}(\bar{x}_t^\theta, \bar{p}_t^\theta) \leq 0.001$.

Performance metrics. We compare the effectiveness of our methods using three criteria: model fit, sparsity, and algorithm efficiency.

To evaluate model fit, we examine the mean absolute error of choice probabilities on subsets generated independently from the training set. Specifically, we generate 100 subsets of $[n]$ of maximum size $\lfloor n/2 \rfloor$ uniformly at random (independently to the training subsets), and append the no-choice option 0 to each of them. We compute the vector of choice probabilities p_{test} using our

ground truth model. Letting $\bar{x}$ be the choice probabilities on the test subsets computed from the estimated choice model at training termination, we calculate $\mathrm{MAE}(\bar{x}, p_{\mathrm{test}})$.

To evaluate sparsity of our estimated model, we examine the number of different rankings σ with positive probability $\lambda(\sigma) > 0$ in our estimated model. Sparsity is very much desired for non-parametric models, since choice probabilities for sparser models can be computed more efficiently.

To evaluate algorithm efficiency, we examine the number of iterations until the termination criterion is reached. While we could have used solution time as another metric for this purpose, we observed in the static setup that solution time is highly correlated with the number of iterations. Runtimes are affected by how fast the combinatorial subproblem (15) is solved, but the focus of our work is not on this aspect, hence in our discussions we focused on the number of iterations as a more accurate representation of algorithm efficiency for our purposes.

G.1. Static Estimation Results

We use the static setup to compare the effect of different parameters (e.g., K, L, and the norm) on algorithm performance. We compare the F-W algorithm (recall that the naïve and modified versions, i.e., $\mathrm{FW}_{\mathrm{naïve}}$ and $\mathrm{FW}_{\mathrm{dyn}}$, are equivalent in the static setup, and so we simply refer to it as F-W in this subsection), as well as the dual MD algorithm for both original non-smoothed norm $\mathrm{MD}_{\mathrm{ns}}$ and its smoothing $\mathrm{MD}_{\mathrm{smth}}$.

Figures EC.1, EC.2 and EC.3 plot the test MAE, the average number of rankings and the average number of iterations for each of the three solution methods when using different norms, varying $K \in \{1, 5, 10\}$ and fixing L to respectively $5, 10$ and 100 while fixing $m = 20$. We observe that the three methods have roughly the same test MAE for each of the norms, with perhaps the smoothed dual MD method performing slightly better when D is based on certain norms, but whenever this occurs, it terminates in more iterations and with a denser model. On the other hand, the non-smooth dual MD method clearly learns a sparser model, and clearly terminates in less number of iterations than the other two methods, which are similar in these two metrics. Therefore, we conclude that, regardless of the type of norm used in the estimation procedure, the non-smooth dual MD method is superior in the static setting, since it manages to learn a sparser model more efficiently, while maintaining the same model fit. This conclusion holds for all combinations of $K \in \{1, 5, 10\}$ and $L = \{5, 10, 100\}$. In particular, they are consistent with our findings for the dynamic setting results shown in Figure 1.

In Figure EC.4, we examine the effect of m by varying $m \in \{10, 20, 50\}$ while fixing $K = L = 5$. We observe that as m increases, the test MAE goes down, but the model sparsity and the number of iterations to convergence increases across all different approaches and norms used. This is as expected, since having more training subsets should allow us to fit better models, but increases the

dimension of the choice probability set X. Our conclusions regarding the comparison of different approaches remain essentially the same: the non-smooth dual approach still outperforms the others.

Figure EC.5 shows the average solution times and the average subproblem times for each method and norm, fixing $m = 20$ and $K = L = 5$. From Figure EC.5 we conclude that the number of iterations and the overall solution time is strongly correlated (contrast it with Figure EC.4(b)) and the non-smooth dual approach is still outperforming the other two with respect to overall solution time. We do not believe that the variation in average subproblem time is the result of any inherent property of the methods used or norms. Moreover, the average subproblem solution times are quite small, and thus the variations in subproblem solution times are relatively small.

G.2. Additional Remarks

In our numerical experiments, we observed that the average number of rankings and iterations are highly correlated. In fact, the Spearman correlation between these two metrics in the static setting with $K = L = 5$ is ≈ 0.922, thus we conclude that the average number of iterations is a good proxy for model sparsity. This can be seen in the theory: all of our algorithms start with one ranking, and at each iteration they add at most one ranking to the estimated model, which provides an explicit bound on the sparsity of the estimated choice they provide at the end.

Finally, one can argue that the algorithms $\mathrm{FW}_{\mathrm{naïve}}$, $\mathrm{FW}_{\mathrm{dyn}}$, and $\mathrm{MD}_{\mathrm{smth}}$ are much simpler to implement than the non-smooth MD algorithm ($\mathrm{MD}_{\mathrm{ns}}$), since there is essentially no parameter tuning aside from computing the smoothing parameters α for D in Table EC.5. On the other hand, $\mathrm{MD}_{\mathrm{ns}}$ algorithm additionally requires tuning the selection of step size, knowing the time horizon T, and computing the constants Ω_Y, G (which in turn affects the smoothing parameters). However, for our particular choice model estimation problem, these quantities are quite straightforward to compute, and our analysis and numerical results are based on such 'textbook' constant step size policies derived from these, which worked quite well. In terms of performance, we see that the extra sophistication in non-smooth MD ($\mathrm{MD}_{\mathrm{ns}}$) can significantly outperform F-W ($\mathrm{FW}_{\mathrm{naïve}}$ or $\mathrm{FW}_{\mathrm{dyn}}$) and smoothed MD ($\mathrm{MD}_{\mathrm{smth}}$).

References

Abernethy J, Lai KA, Levy KY, Wang JK (2018) Faster rates for convex-concave games. *COLT 31*, 1595–1625.

Beck A, Teboulle M (2012) Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization* 22(2):557–580.

Calafiore G, Campi M (2005) Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming* 102(1):25–46.

Devolder O, Glineur F, Nesterov Y (2014) First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming* 146(1):37–75.

Farias V, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management Science* 59(2):305–322.

Freund RM, Grigas P (2016) New analysis and results for the frank–wolfe method. *Mathematical Programming* 155(1):199–230.

Ho-Nguyen N, Kılınç-Karzan F (2019) Exploiting problem structure in optimization under uncertainty via online convex optimization. *Mathematical Programming* 177(1):113–147.

Jaggi M (2013) Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th International Conference on Machine Learning*, volume 28(1), 427–435.

Mišić VV (2016) *Data, models and decisions for large-scale stochastic optimization* (Chapter 4). Ph.D. thesis, Massachusetts Institute of Technology.

Nesterov Y (2018) Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming* 171(1):311–330.

Shalev-Shwartz S, Kakade SM (2008) Mind the duality gap: Logarithmic regret algorithms for online optimization. *NeurIPS 21*, 1457–1464.

van Ryzin G, Vulcano G (2015) A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Science* 61(2):281–300.

Wang JK, Abernethy JD (2018) Acceleration through optimistic no-regret dynamics. *NeurIPS 31*, 3824–3834.

Figure EC.1   Performance metrics in the static setup for different K, fixing $L = 5, m = 20$.
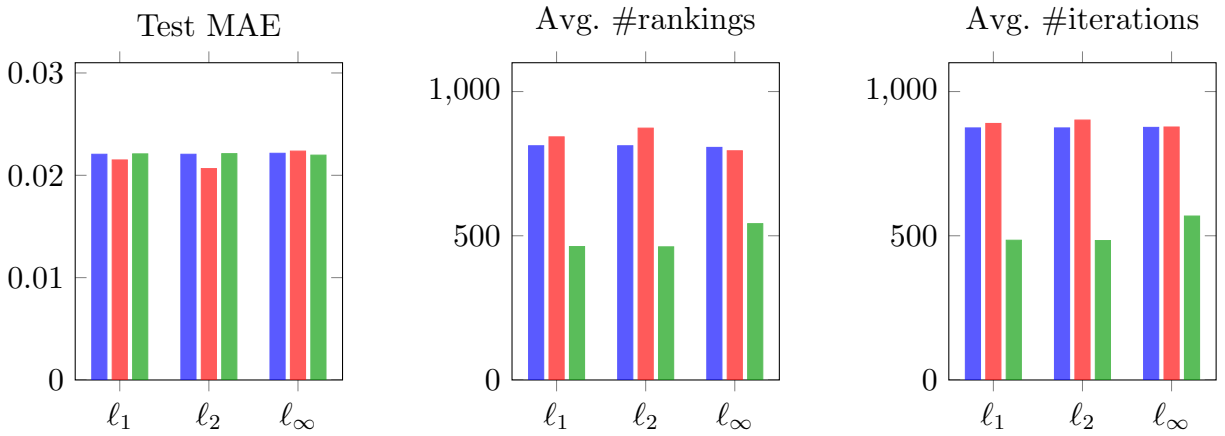

(a) $K = 1$

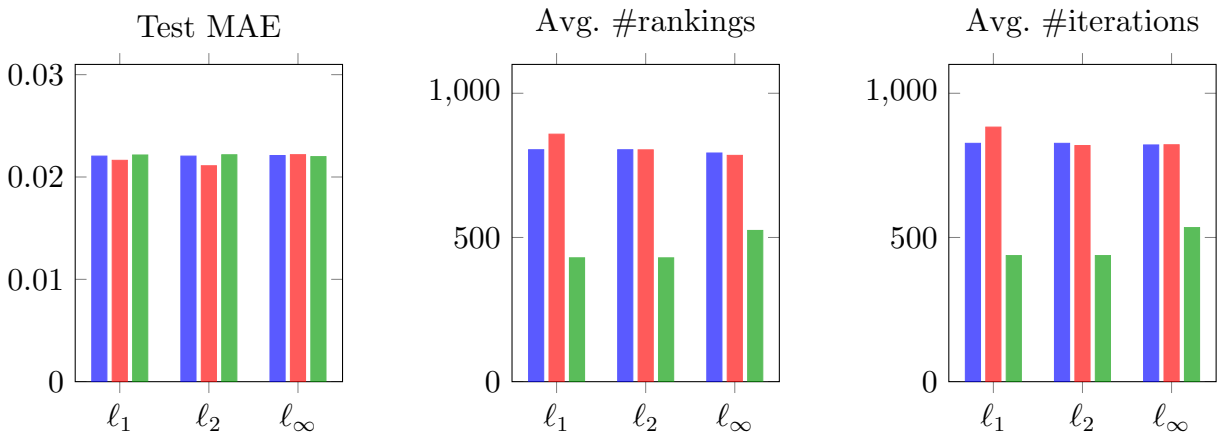(b) $K = 5$
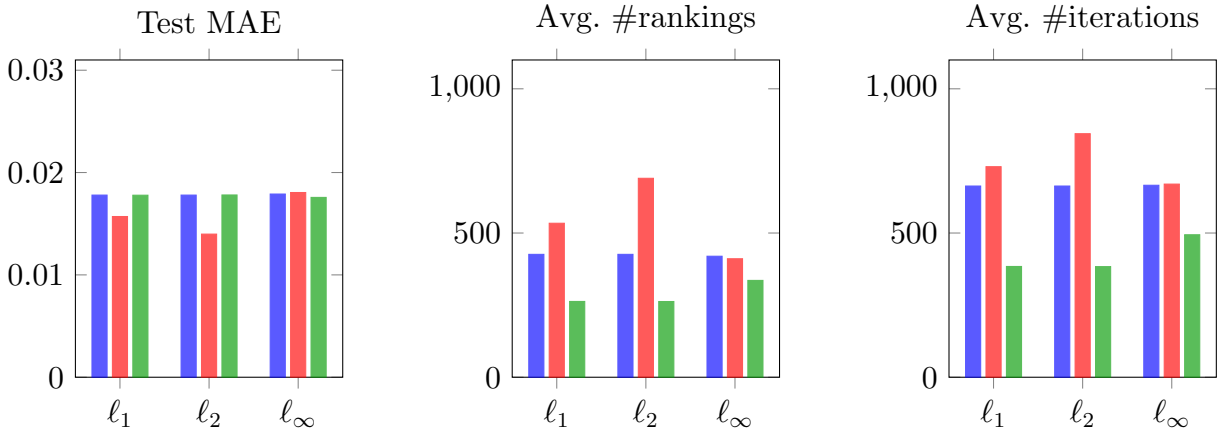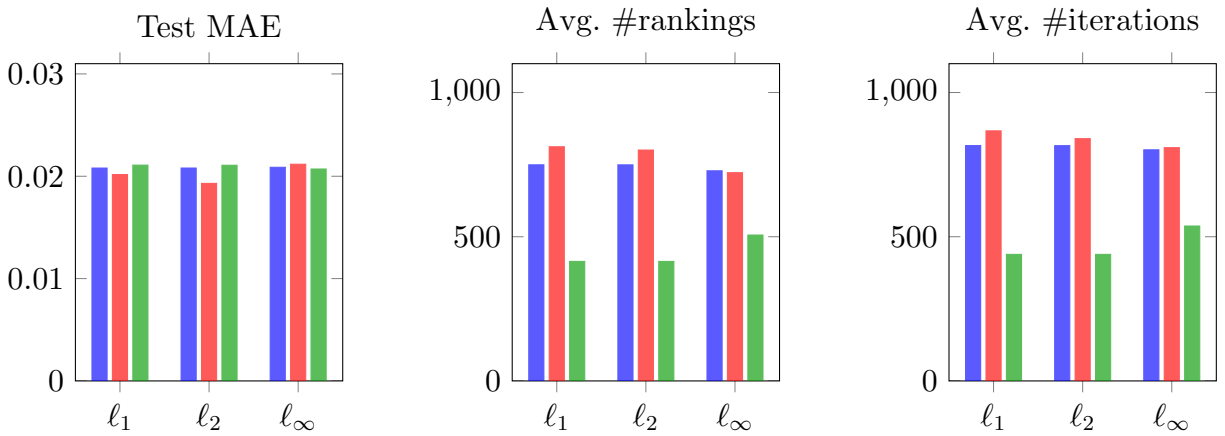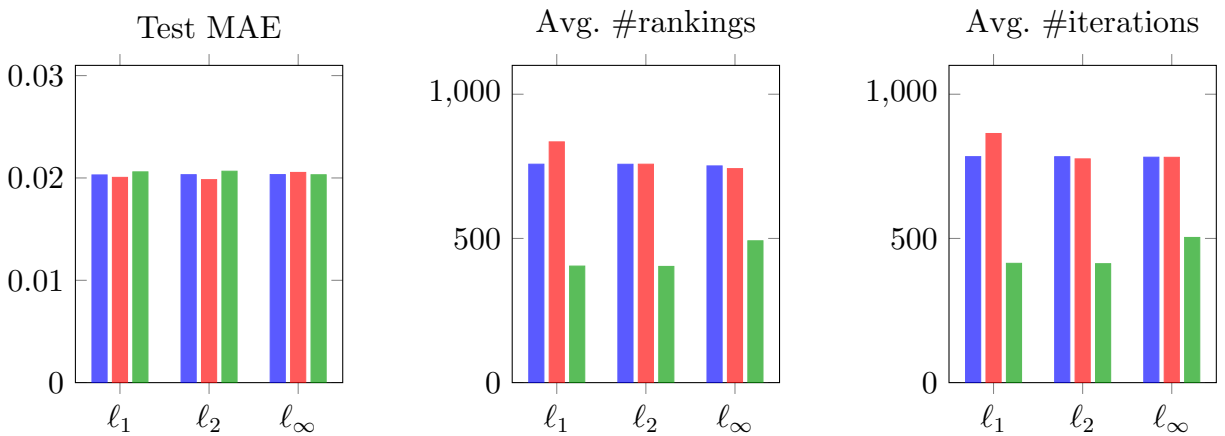
(c) $K = 10$

■ F-W ■ smooth MD ■ nonsmooth MD

Figure EC.2   Performance metrics in the static setup for different K, fixing $L = 10, m = 20$.
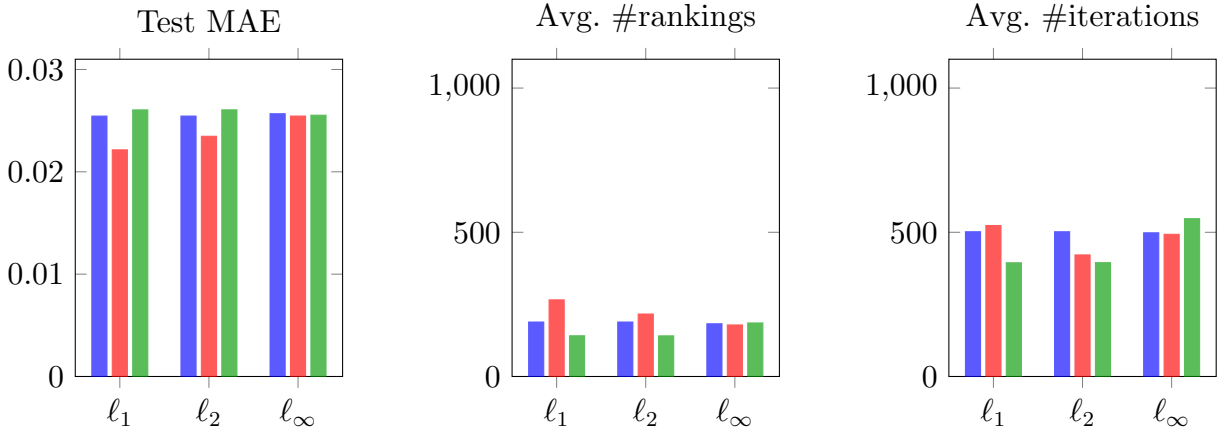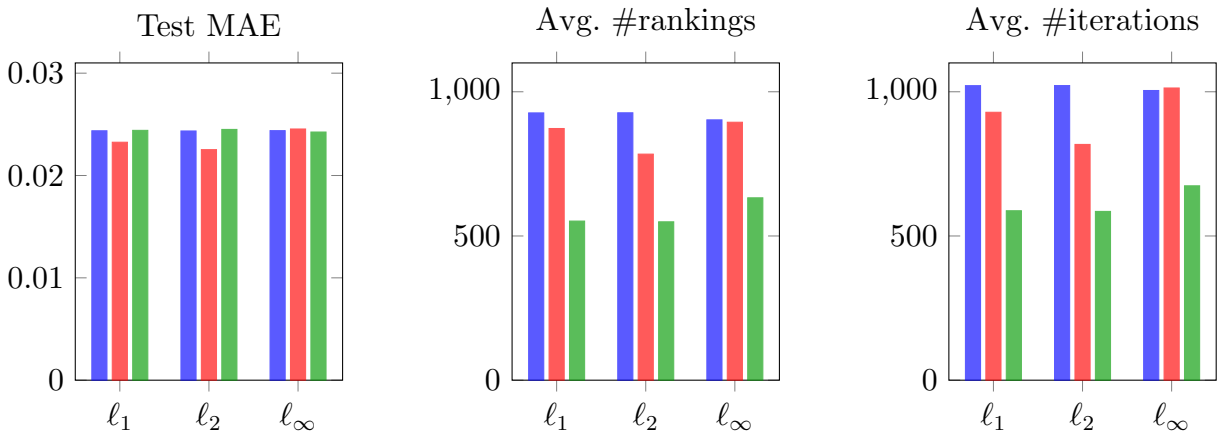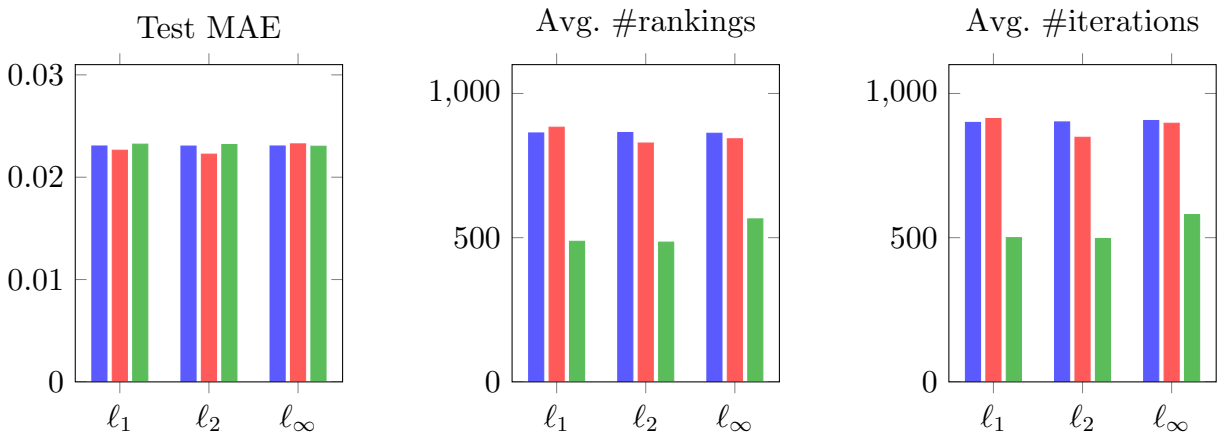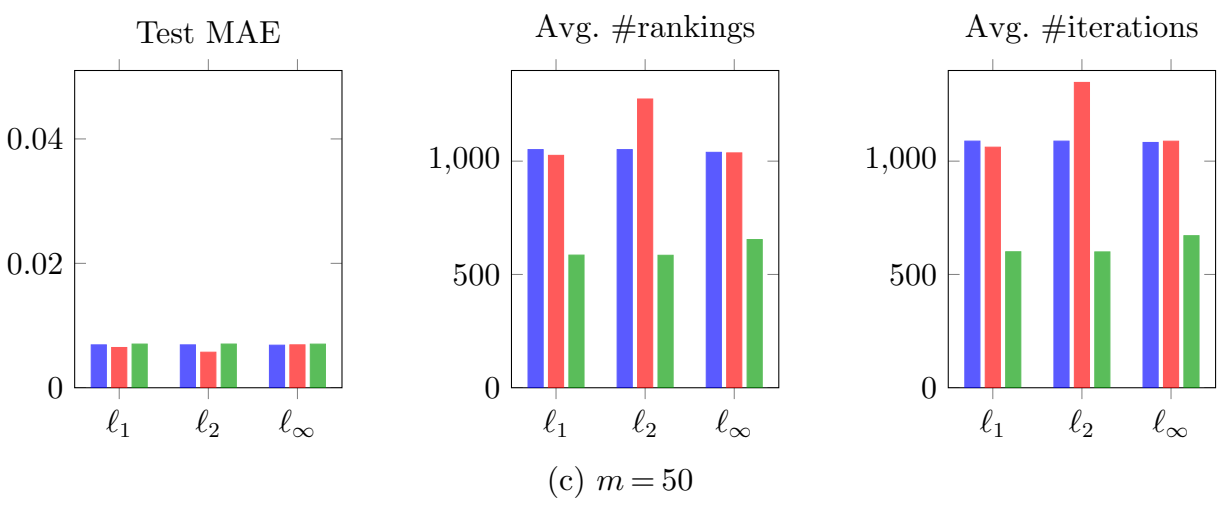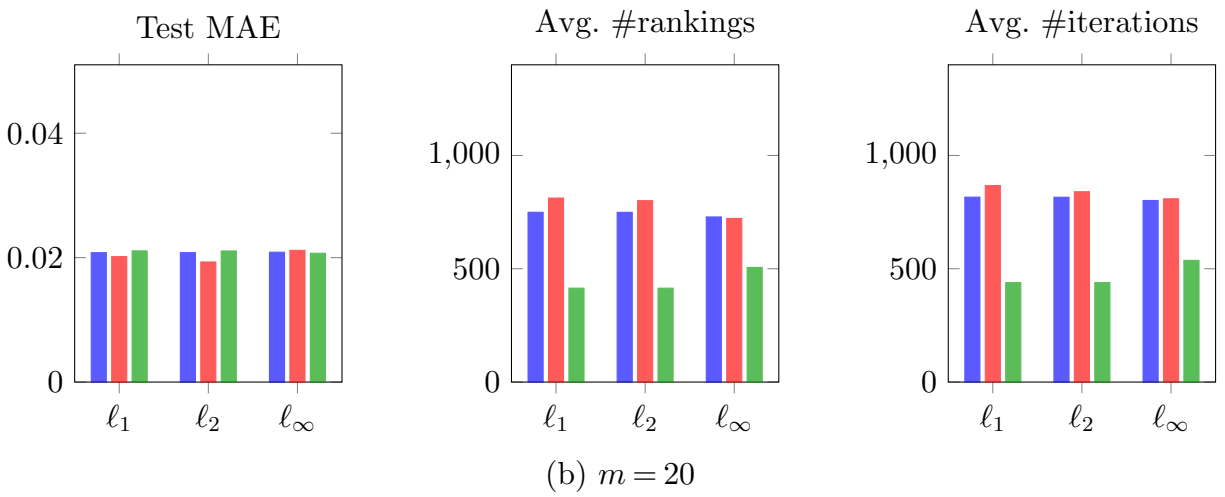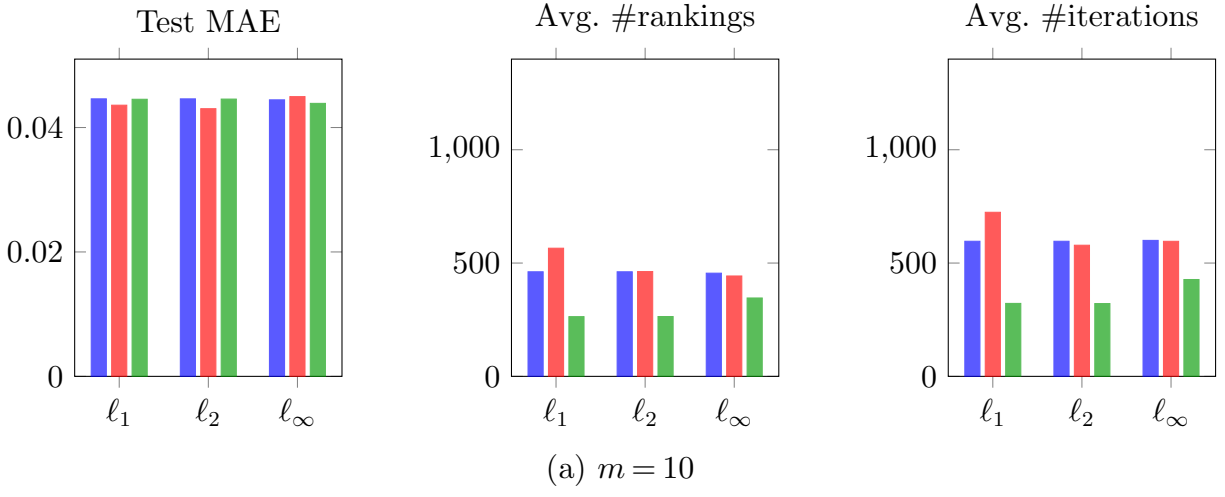

(a) $K = 1$

(b) $K = 5$
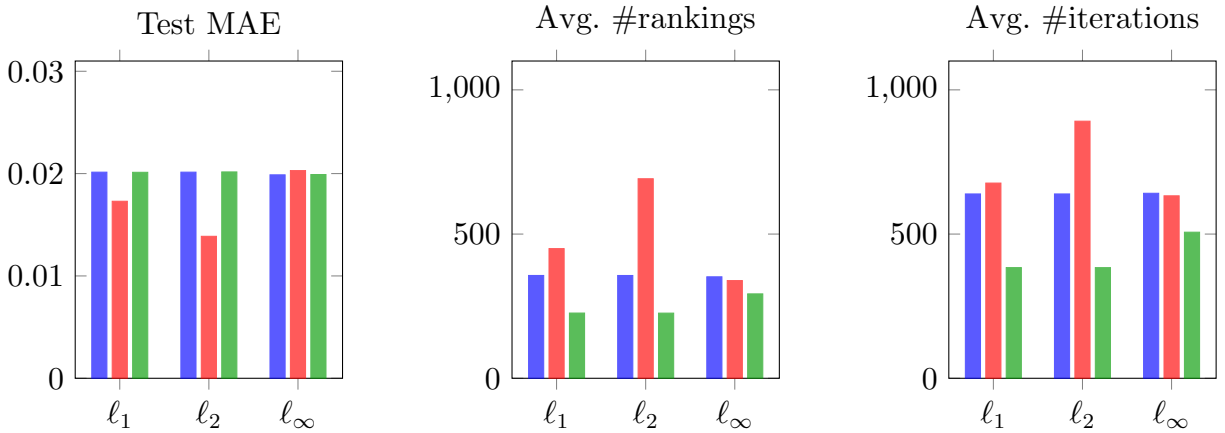
(c) $K = 10$

■F-W ■smooth MD ■nonsmooth MD

Figure EC.3   Performance metrics in the static setup for different K, fixing $L = 100, m = 20$.
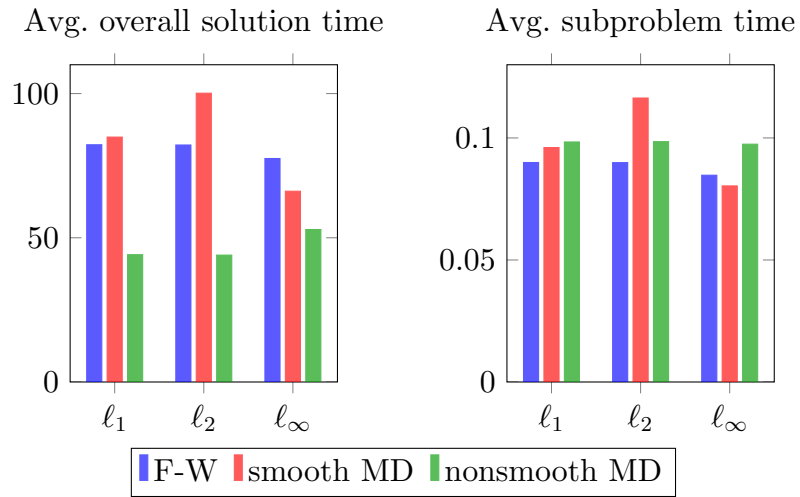

(a) $K = 1$

(b) $K = 5$

(c) $K = 10$

F-W   smooth MD   nonsmooth MD

Figure EC.4   Performance metrics in the static setup for different m, fixing $K = L = 5$.

(a) $m = 10$

(b) $m = 20$

(c) $m = 50$

Figure EC.5   Overall solution times and subproblem times per iteration (both in seconds) in the static setup for $m = 20$ and $K = L = 5$.