# Robust Two-Stage Optimization with Covariate Data

Bart P.G. Van Parys M. Amine Bennouna

October 24, 2022

**Abstract**

We consider a generalization of two-stage decision problems in which the second-stage decision may be a function of a predictive signal but cannot adapt fully to the realized uncertainty. We will show how such problems can be learned from sample data by considering a family of regularized sample average formulations. Furthermore, our regularized data-driven formulations admit convex distributionally robust counterparts which enjoy desirable asymptotic out-of-sample performance guarantees. Finally, we show that all derived data-driven formulations can be solved efficiently using canonical stochastic gradient algorithms.

## 1 Introduction

The problem class we will consider consists of two-stage problems in which a first-stage "here-and-now" decision $z_0 \in Z_0 \subseteq \mathrm{R}^{n_0}$ needs to be made in the face of an uncertain outcome $\tilde{\xi}$ which realizes in $\Xi \subseteq \mathrm{R}^{d_1}$. Unlike classical two-stage problems, we will not allow the decision maker to delay a second-stage decision until after the uncertain outcome $\tilde{\xi}$ has been observed. Rather, we will assume that a certain lead time needs to be taken into account and hence the second-stage decision needs to be made also without observing the uncertain outcome $\tilde{\xi}$ directly. However, as the lead time is assumed to be small, a second-stage decision $z \in Z \subseteq \mathrm{R}^n$ may adapt to additional covariate information $\tilde{\xi}'$ realizing in $\Xi' \subseteq \mathrm{R}^{d_2}$ which has become available and may be predictive of the uncertainty $\tilde{\xi}$. The causal relationship between the decisions $z_0$ and $z$ and random variables $\tilde{\xi}'$ and $\tilde{\xi}$ is depicted in Figure 1.

Denote $P \in \mathcal{P}(\Xi)$ and $P' \in \mathcal{P}(\Xi')$ the distribution of respectively $\tilde{\xi}$ and $\tilde{\xi}'$ and $M \in \mathcal{P}(\Xi' \times \Xi)$ their joint distribution. In this setting one can consider the stochastic problem formulation

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ \ell_0(z_0) + \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) \right], \tag{1}$$

where $\ell_0 : Z_0 \to \mathrm{R}_+$ and $\ell : Z_0 \times Z \times \Xi \to \mathrm{R}_+$ denote the first-stage and second-stage cost, respectively, and

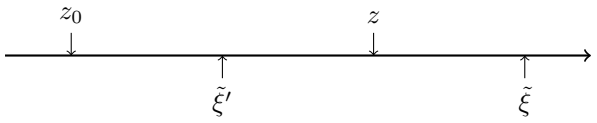


Figure 1: Causal relationships between the first-stage decision $z_0$, the covariate $\xi'$, the second-stage decision $z$ and the random variable $\xi$.

$\mathcal{F} := \{\xi' \in \mathrm{R}^{d_2} \mapsto z(\xi') \in Z \text{ a measurable function}\}$ the set of feasible second-stage decision plans. We assume that the minimum in Equation (1) exists as a consequence of assumptions imposed further in the paper. We denote with $v^\star$ and $(Z_0^\star, \mathcal{F}^\star)$ the minimum and minimizers in (1), respectively. The objective function quantifies the cost incurred to the decision-maker when committing to a first stage decision $z_0$ and a second-stage decision plan $z$. The expectation is here jointly over the random uncertainty $\tilde{\xi}$ and random covariate $\tilde{\xi}'$. Notice that potential constraints on the second stage decision are captured by the function class $\mathcal{F}$ which demands $z(\xi') \in Z$ for all $\xi' \in \Xi'$.

Two special cases of the stochastic optimization problem in Equation (1) can be identified. When the covariates $\tilde{\xi}'$ and the uncertain parameters $\tilde{\xi}$ are independent, their joint distribution coincides with the product distribution $M = P' \otimes P$. In this case, it is clearly futile to adapt the second-stage decision $z$ to $\tilde{\xi}'$ as it carries no statistical information concerning the uncertainty $\tilde{\xi}$. Problem (1) reduces in this case to an ordinary stochastic optimization problem over $z_0 \in Z_0$ and $z(\xi') = z' \in Z$ for all $\xi' \in \mathrm{R}^{d_2}$. However, when $\tilde{\xi}'$ and $\tilde{\xi}$ are correlated, the decision-maker may hope that better second-stage decisions can be learned from the available covariate information. In fact, in the extreme case when $\tilde{\xi}' = \tilde{\xi}$ problem (1) reduces to an ordinary two-stage adaptive optimization problem where the second-stage decision can be fully adapted to the uncertainty $\tilde{\xi}$.

**Example 1.1** (Newsvendor)**.** *Consider a newsvendor which has to decide on the total quantity of newspapers to order. Suppose that the newsvendor commits to order $z_0$ newspapers at wholesale price $p_0$. However, the newsvendor has the option to order additional copies at price $p > p_0$ the evening before if their is a suspecion that his regular order $z_0$ will be insufficient based on observing additional covariate information $\tilde{\xi}'$ (such as a important news story breaking). We assume that the newsvendor sells newspapers at price $k > p > p_0$ until the uncertain daily demand $\tilde{\xi}$ is met. To maximize expected profit the newsvendor wishes to minimize the expected loss*

$$\min_{z_0 \in \mathrm{R}_+,\, z \in \mathcal{F}} \; p_0 z_0 + \mathbb{E}_M \left[ p z(\tilde{\xi}') - k \min(z_0 + z(\tilde{\xi}'), \tilde{\xi}) \right].$$

**Stochastic Optimization** It should be remarked that Problem (1) can be interpreted as an ordinary stochastic optimization problem

$$(1) = \min_{z_0 \in Z_0,\, z \in \mathcal{F}} \; \ell_0(z_0) + \mathbb{E}_M \left[ C(z_0, z, \tilde{\xi}', \tilde{\xi}) \right] \tag{2}$$

with the help of a cost function $C : Z_0 \times \mathcal{F} \times \Xi' \times \Xi \to \mathrm{R}_+$. In the remainder of the document we will assume that this cost function is uniformly integrable, i.e., $\mathbb{E}_M\left[\left|C(z_0, z, \tilde{\xi}', \tilde{\xi})\right|\right] = \mathbb{E}_M\left[C(z_0, z, \tilde{\xi}', \tilde{\xi})\right] < \infty$, for all feasible first and second-stage decisions. Note that problem (2) is however infinite dimensional in nature as the second-stage decision variable $z$ could be any (measurable) function in the set $\mathcal{F}$ mapping covariate information to a feasible decision. In particular this implies that although problem (1) can be interpreted as an ordinary stochastic optimization problem it can not be solved as such directly.

**Decomposition** A standard result, c.f., Rockafellar and Wets (2009, Theorem 14.60), is that our two-stage problem can be however be decomposed as a finite dimensional first-stage minimization problem

$$(1) = \min_{z_0 \in Z_0} \ \ell_0(z_0) + \mathbb{E}_{P'}\left[Q(z_0, \tilde{\xi}')\right] \tag{3}$$

and a subsequent second-stage "cost-to-go" function characterized as the finite dimensional minimization problem

$$Q(z_0, \xi'; M) := \min_{z \in Z} \ \mathbb{E}_M\left[\ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi'\right] \quad \forall z_0 \in Z_0, \ \xi' \in \mathrm{R}^{d_2}. \tag{4}$$

The cost-to-go function $Q$ represents the best achievable cost in expectation, when the first stage decision is $z_0$ and the covariates $\tilde{\xi}'$ realizes as $\xi'$ when the second-stage decision is chosen optimally. In particular, the minimizer of (4) is exactly $z^\star(\xi')$ for all $\xi' \in \mathrm{R}^{d_2}$ where $z^\star \in \mathcal{F}^\star$. Here, the expectation is over the randomness of the uncertainty $\tilde{\xi}$ conditional on the coveriates $\tilde{\xi}'$ realizing as $\xi'$. We impose in this paper standard convexity and regularity assumptions concerning the loss function and feasible regions of our decisions.

**Assumption 1.2** (Convexity and Regularity)**.**

*(i) The first-stage loss function $\ell_0 : Z_0 \to \mathrm{R}_+$ is convex while the second-stage loss $\ell : Z_0 \times Z \times \Xi \mapsto \mathrm{R}_+$ is jointly convex in its first two arguments for any value of its third.*

*(ii) The feasible sets $Z_0 \subseteq \mathrm{R}^{n_0}$ and $Z \subseteq \mathrm{R}^n$ are convex, compact and non-empty.*

*(iii) The first-stage loss function $\ell_0$ is $\mathcal{L}_0$-Lipschitz and the second-stage loss function $\ell$ is $L$-bounded and $\mathcal{L}$-Lipschitz in $(z_0, z, \xi') \in Z_0 \times Z \times \Xi$.*

Assumption 1.2 ensures that problem (1) is convex and its minimum is indeed achieved. Note that as $\ell$ is $\mathcal{L}$-Lipschitz the conditional expectation[1] $(z_0, z) \mapsto \mathbb{E}_M\left[\ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi'\right]$ is $\mathcal{L}$-Lipschitz continuous for any $\xi' \in \Xi'$. As the feasible set $Z$ is compact and nonempty the minimum in Equation (4) is thus achieved for any $M \in \mathcal{P}(\Xi' \times \Xi)$.

[1] All statements involving a conditional expectation should here be interpreted as $P'$-almost everwhere as such conditional expectation is only well-defined up to sets of measure zero.

Furthermore, the associated minimum function $Q$ is convex and can be shown to be $\mathcal{L}$-Lipschitz continuous in $z_0$. Hence, also the expectation $z_0 \mapsto \mathbb{E}_{P'}\left[Q(z_0, \tilde{\xi}'; M)\right]$ is $\mathcal{L}$-Lipschitz. Consequently, the first-stage minimum in Equation (3) will be achieved. The final requirement in Assumption 1.2 will help to ensure that problem (1) can be approximated based on historical data.

**Data-Driven Formulations** It should be clear that in practice the joint distribution $M$ is not given and typically must be learned from historical observations instead. That is, we have access to a dataset of pairs

$$D_N := \{(\xi'_d, \xi_d)\}_{d=1}^N$$

counting a finite number $N$ historical observations. We may assume without loss of generality that the observed data points are distinct. Denote the empirical distribution of the dataset as $M_N = \sum_{d=1}^N \delta_{(\xi'_d, \xi_d)}$ and denote with $P'_N = \sum_{d=1}^N \delta_{\xi'_d}$ and $P_N = \sum_{d=1}^N \delta_{\xi_d}$ the empirical distributions of the covariates and uncertain outcomes, respectively. We will regard in this paper the historical data as independent samples from the actual but unknown joint distribution $M$ and attempt to learn the optimal decision proposed in problem (1) from such data.

A naive data-driven counterpart to problem (1) would be

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ \ell_0(z_0) + \mathbb{E}_{M_N}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \tag{5}$$

in which the empirical distribution $M_N$ substitutes for the actual but unknown distribution $M$. Such sample average formulations have been succesfully considered in many data-driven decision-making formulations. However, in the context of our problem, this approximation is exceptionally unsatifactionary as its second-stage cost can dramatically underestimate the actual out-of-sample cost. Indeed, suppose $P'$ is a continuous distribution then with probability one the data points in $D_N$ are distinct. As the function class $\mathcal{F}$ is unrestricted, the function values $z(\xi_d)$ for $d = [1, \ldots, N]$ can here take any arbitrary value in $Z$. Hence, we have

$$\begin{aligned}
(5) &= \min_{z_0 \in Z_0} \ \ell_0(z_0) + \min_{z \in \mathcal{F}} \textstyle\sum_{d=1}^N \ell(z_0, z(\xi'_d), \xi_d)/N \\
&= \min_{z_0 \in Z_0} \ \ell_0(z_0) + \min_{z_1, \ldots, z_N \in Z} \textstyle\sum_{d=1}^N \ell(z_0, z_d, \xi_d)/N.
\end{aligned}$$

Even with an increasing number of samples $N$, this naive formulation does hence not general recover the solution to problem (1). Only if the covariate distribution $P'$ is finitely supported, i.e., $\Xi'$ has finite cardinality, can this sample average formulation hope to recover the solution to problem (1) by observing all outcomes in $\Xi'$ sufficiently often. In this paper we are however primarily interested in the case where $\Xi'$ is continuous. One could argue that

the naive data-driven formulation (5) is unsatisfactory as problem (1) is hard to approximate directly.

## 1.1 Related Work

An empirical risk minimization approach first regularizes problem (1) by imposing functional restrictions on what the second-stage decision $z$ may look like. Ban and Rudin (2019) consider an empirical risk minimization approach to problem (1) in the context of a newsvendor problem by imposing the second-stage decision $z$ to be an affine decision function of the observed covariate $\xi'$. It is well known that by lifting the covariates to a reproducing kernel Hilbert space the linearity assumption on the decision rules can be relaxed as pointed out already by Ban and Rudin (2019). Recently, Bertsimas and Koduri (2021) have shown that under certain regularity conditions on problem (1) consistent empirical risk formulations can be obtained by using universal kernels (Micchelli et al. 2006). However, as we discuss in Section 4.2 empirical risk minimization do not allow directly for second-stage constraints of the type $z(\xi') \in Z$ for all $\xi' \in \Xi'$ which occur in many practical problems.

Denote with $P_{\xi'}$ the conditional distribution of the uncertain parameter $\tilde{\xi}$ conditioned on the event $\tilde{\xi}' = \xi'$. Then, we have (1) $= \min_{z_0 \in Z_0} \ell_0(z_0) + \mathbb{E}_{P'}\left[Q(z_0, \tilde{\xi}')\right]$ with $Q(z_0, \xi'; M) := \min_{z \in Z} \mathbb{E}_{P_{\xi'}}\left[\ell(z_0, z, \tilde{\xi})\right]$ for all $z_0 \in Z_0$ and $\xi' \in \mathrm{R}^{d_2}$. Hence, we can derive a data-driven formulation of problem (1) simply by substituting both the unknown covariate and conditional distribution $P'$ and $P_{\xi'}$ with estimates based on historical data, respectively. Hanasusanto and Kuhn (2013) propose for instance the data-driven formulation based on the empirical covariate distribution $P'_N$ and a Nadaraya-Watson density estimate $P_{\xi',N} = \sum_{d=1}^N k((\xi' - \xi'_d)/h)\, \delta_{\xi_d} / \sum_{d=1}^N k((\xi' - \xi'_d)/h)$ of Hannah et al. (2010) where $k : \mathrm{R}^{d_2} \to \mathrm{R}$ a given weighing function and $h > 0$ a judiciously chosen bandwidth parameter. Bertsimas and McCord (2019) propose alternatives based on nearest neighbor estimates and tree learning methods instead. These resulting data-driven decision can be shown to be consistent and finite-sample guarantees are developed for a nearest neighbors based approach under mild technical regularity condition on problem (1) by Bertsimas and McCord (2019). These formulations, much like the sample average approximation for stochastic optimization, however exhibit an optimistic bias if the sample size is small as pointed out already by Hanasusanto and Kuhn (2013).

To guard against overfitting effects Hanasusanto and Kuhn (2013) propose the stage-wise robust counterpart

$$\min_{z_0 \in Z_0} \ \ell_0(z_0) + \max_{P' \in \mathcal{P}'_N} \mathbb{E}_{P'}\left[Q_{r,N}(z_0, \tilde{\xi}')\right] \tag{6}$$

where $\mathcal{P}'_N \subseteq \mathcal{P}'(\Sigma')$ is an ambiguity set for the unknown covariate distribution $P'$ and the cost-to-go function $Q_{r,N}$

is characterized here as

$$Q_{r,N}(z_0, \xi') = \min_{z \in Z} \max_{P_{\xi'} \in \mathcal{P}_N(\xi')} \mathbb{E}_{P_{\xi'}}\big[\ell(z_0, z, \tilde{\xi})\big]\,. \tag{7}$$

where here $\mathcal{P}_N(\xi')$ are ambiguity sets for the unknown conditional distribution $P_{\xi'}$ for all $\xi' \in \Xi'$. Here the subscript "r" on $Q_{r,N}$ alludes to the fact that it can be regarded as a robust counterpart to the cost-to-go function $Q$ defined in Equation (4). Hanasusanto and Kuhn (2013) single out in particular scaled Pearson divergence ambiguity sets for their excellent computational tractability. We remark that the second-stage distributions $P^\star_{\xi'}$ characterizing the second-stage decisions in Equation (7) are adversarially chosen for all covariates $\xi' \in \Xi'$ separately. To make matters worse, the covariate distribution $P'^\star$ characterizing the first-stage decision in Equation (6) is also chosen adversarially independently of any of the aforementioned second-stage distributions $P^\star_{\xi'}$ for $\xi' \in \Xi'$. Hence, as robustness is here applied stage-wise the resulting robust formulation is potentially very conservative.

Gao et al. finally propose to robustify problem (1) directly by considering

$$\inf_{z_0 \in Z_0,\, z \in \mathcal{F}} \max_{M' \in \mathcal{W}_N} \ell_0(z_0) + \mathbb{E}_{M'}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \tag{8}$$

where $\mathcal{W}_N = \{M' \,:\, \|M_N - M\| \leq \delta(N)\}$ denote Wasserstein ambiguity sets with appropriate chosen radius $\delta(N) > 0$. In particular, such direct formulations are less pessimistic than adding robustness stage wise and out-of-sample guarantees can be established based the standard concentration inequalities which bound the probability of the event $M \notin \mathcal{W}_N$ (Fournier and Guillin 2015). Such direct robust formulations however come with two major drawbacks. First, although Gao et al. provides a tractable formulation for problem (8) when the second-stage decision is univariate ($n = 1$) it is unclear whether problem (8) admits a tractable reformulation in general. Second, recent work by Duchi et al. (2016), Lam (2019), Bennouna and Van Parys (2021) indicates that in the context of data-driven stochastic optimization problems ambiguity sets $\mathcal{M}_N \subseteq \mathcal{P}(D_N)$ based on certain $\varphi$-divergence may results in formulations who enjoy better out-of-sample guarantees. However, naively considering ambiguity sets consisting of distributions only supported on the observed data points in the context of our two-stage optimization problems is plagued by the exact same overfitting phenomenon as discussed for the naive formulation (5) and hence by itself does not provide out-of-sample guarantees.

## 1.2 Contributions

We consider hence first a regularized version of the naive data-driven formulation (5) defined in this paper as

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ell_0(z_0) + \mathbb{E}_{M_N \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \tag{9}$$

where the artificial random perturbation $\tilde{e}$ is distributed following a known chosen distribution $E$ and $h \geq 0$ is an appropriately chosen regularization parameter. Here, $M \otimes E$ denotes the distribution of the random variable $(\tilde{\xi}', \tilde{\xi}, \tilde{e})$ for a given joint distribution $M$ of $(\tilde{\xi}', \tilde{\xi})$ and we again remark that the noise $\tilde{e}$ is chosen independently from $(\tilde{\xi}', \tilde{\xi})$. We show in section 2 that the regularized formulation with properly chosen perturbation leads to consistent data-driven formulations.

In this paper we will propose a direct robust counterparts to a slightly regularized version of the original problem (1). We consider indeed a robust version of the regularized formulation

$$\inf_{z_0 \in Z_0,\, z \in \mathcal{F}} \max_{M' \in \mathcal{M}_N} \ell_0(z_0) + \mathbb{E}_{M' \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \tag{10}$$

with ambiguity set $\mathcal{M}_N$. Our robust formulation considers an ambiguity set directly on the joint distribution of the covariate and uncertainty $(\tilde{\xi}', \tilde{\xi})$, rather than adding robustness stage-wise. Following the recent work by Lam (2019), Bennouna and Van Parys (2021), Duchi et al. (2016) ambiguity sets $\mathcal{M}_N \subseteq \mathcal{P}(D_N)$ constructed based on $\varphi$-divergences enjoy excellent statistical properties in the context of data-driven stochastic optimization problems. We show that certain $\varphi$-divergence ambiguity sets do provide out-of-sample guarantees in the context of two-stage optimization problems when combined with regularization as in problem (10).

This paper advances three further contributions which we list here explicitly:

1. We show that under mild technical conditions the regularized formulation (9) can learn an optimal decision in formulation (1) almost surely as the number of data points increases. Furthermore, we interpret this novel data-driven formulation as combining Nadaraya-Watson and Parzen density estimation.

2. We show that our robust formulation effectively guards against overfitting by deriving asymptotic out-of-sample guarantees for an ambiguity set $\mathcal{M}_N = \left\{M' \in \mathcal{P}(\Xi' \times \Xi) \;:\; M' \ll M_N,\; \chi^2(M_N, M') \leq r(N)\right\}$ where $\chi^2(M_N, M')$ denotes the Neyman divergence between $M_N$ and $M'$ and $r(N) > 0$ a judiciously scaled robustness parameter.

3. We show how all our proposed formulations can be solved efficiently using canonical stochastic approximation methods advanced by Nemirovski et al. (2009).

### 1.3 Notation

We assume that $\Xi'$ and $\Xi$ are compact sets equipped with the standard Euclidean norm and associated topology. Similarly, we denote with $\mathcal{P}(U)$ the set of all Borel probability measures on a normed topological space $U$. Similarly, we assume that this set is equipped with the respective topology of weak convergence. Following Dembo and

Zeitouni (2009, Section 6.2) the probability simplices $\mathcal{P}(U)$ when equipped with the topology of weak convergence of probability measures is polish. Let $\mathcal{P}_a(U) = \left\{\nu \,:\, \int \mathrm{d}\nu(u) = a\right\}$ with $a \in \mathrm{R}$ denote the set of signed Borel measures on a normed spaces $U$ of finite variation and so that $\int \mathrm{d}\nu(u) = a$. We denote with $\mathcal{P}_+(U)$ the set of nonnegative Borel measures on a space $U$. Hence, $\mathcal{P}(U) = \mathcal{P}_+(U) \cap \mathcal{P}_1(U)$. Finally, we equip the set $\mathcal{P}_0(U)$ for the normed space $U$ with a topology induced by the norm

$$\|\nu\| = \sup_{f \text{ measurable}} \left\{ \int f(u) \,\mathrm{d}\nu(u) \,:\, |f(u_1) - f(u_2)| \leq \|u_1 - u_2\| \;\; \forall u_1 \in U,\, u_2 \in U \right\}. \tag{11}$$

Given independent random variables $u$ and $v$ with distributions $Q \in \mathcal{P}(U_1)$ and $P \in \mathcal{P}(U_2)$ we denote the distribution of $(u, v)$ as $Q \otimes P \in \mathcal{P}(U_1 \times U_2)$. For a convex function $g$ we denote with $\partial g$ its subdifferential while $g'$ denotes an arbitrary gradient selection. That is, we have that $g'(x) \in \partial g(x)$ for all $x \in \operatorname{dom} g$. Finally, for any set $S$ we define its diameter as $\operatorname{diam}(S) = \sup_{s_1 \in S,\, s_2 \in S} \|s_1 - s_2\|$.

# 2 Regularized Formulations

## 2.1 Regularization

It will be beneficial to consider here a regularized counterpart to problem (1). Let $\tilde{e}$ be a random variable independent of the random variable $(\tilde{\xi}', \tilde{\xi})$ distributed following $E$ with compact support $K \subset \mathrm{R}^{d_2}$ with $\operatorname{int}(K) \neq \emptyset$. That is, the random variable $(\tilde{\xi}', \tilde{\xi}, \tilde{e})$ is distributed as the product distribution $M \otimes E$. Consider a regularized counterpart to problem (1) defined here as

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \; \ell_0(z_0) + \mathbb{E}_{M \otimes E} \left[ \ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi}) \right] \tag{12}$$

for any regularization parameter $h > 0$. We remark that if that if the regularization parameter is chosen as $h = 0$ the regularized problem (12) is identical to the problem of interest (1). The introduction of artificial perturbation can be interpreted as either regularizing the cost function or the joint distribution in problem (1) as we explain momentarily.

**Loss Smoothing** Let $k$ denote the Radon-Nikodym derivative or density function of $E$ with respect to the Lebesgue measure $m'$ on the feature space $\mathrm{R}^{d_2}$ scaled so that $m'(K) = 1$. That is, $E(B) = \int_B k(\xi'') \,\mathrm{d}m'(\xi'')$ for $B$ a measurable subset of $\Xi'$; see Table 1 for several examples. Let us define for all $z_0 \in Z_0$, $z \in \mathcal{F}$, $\xi' \in \Xi'$ and $\xi \in \Xi$

| Noise | Density function $k(e) =$ |
|---|---|
| Uniform | $\frac{1}{2}\mathbb{1}\{\lvert e\rvert \leq 1\}$ |
| Epanechnikov | $\frac{3}{4}(1-\lvert e\rvert^2)\mathbb{1}\{\lvert e\rvert \leq 1\}$ |
| Tricubic | $\frac{70}{81}(1-\lvert e\rvert^3)^3\mathbb{1}\{\lvert e\rvert \leq 1\}$ |
| Gaussian | $\exp(-\lvert e\rvert^2/2)/\sqrt{2\pi}$ |

Table 1: The Epanechnikov, tricubic and uniform density functions are proper. The Gaussian density function is not proper as it does not have bounded support.

the convoluted loss

$$\begin{aligned} C_h(z_0, z, \xi', \xi) :=& \mathbb{E}_E\left[\ell(z_0, z(\xi' + h\tilde{e}), \xi)\right] \\ =& \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2}\, \mathrm{d}m'(\xi'') \end{aligned}$$

where the subscript "$h$" alludes to the fact that it can be interpreted as a regularized counterpart of the cost function $C$ defined in Equation (2). Indeed, we have

$$(12) = \min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ell_0(z_0) + \mathbb{E}_M\left[C_h(z_0, z, \tilde{\xi}', \tilde{\xi})\right] \tag{13}$$

indicating that our noisy counterpart is of the same form as the original problem (2) but crucially considers a regularized cost function. It is well known that convolution typically has a beneficial regularization effect which we will indicate is also the case in the setting we consider here.

**Distribution Regularization** Consider the family of distributions $\delta_{\xi',h} \in \mathcal{P}(\Xi')$ for all $h > 0$ and $\xi' \in \Xi'$ uniquely characterized through $\delta_{\xi',h}(B) = \int_B k((\xi'' - \xi')/h)/h^{d_2}\, \mathrm{d}m'(\xi'')$ for any measurable subset $B$ of $\mathrm{R}^{d_2}$. In other words, for $\xi' \in \Xi'$, we have that $\xi' + h\tilde{e}$ is distributed as $\delta_{\xi',h}$ where here again the subscript "$h$" alludes to the fact that $\delta_{\xi',h}$ can be interpreted as a smooth counterpart to the degenerate Dirac distribution $\delta_{\xi'}$ located at $\xi' \in \Xi'$; see also Figure 2. Consider now the regularized distribution $M_h = \int \delta_{\xi',h} \otimes \delta_\xi\, \mathrm{d}M(\xi', \xi)$. It is trivial to verify that

$$(12) = \min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ell_0(z_0) + \mathbb{E}_{M_h}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]; \tag{14}$$

see also Figure 2. Hence, our regularized counterpart (12) may also be interpreted as considering a regularized distribution $M_h$ instead of the distribution $M$ directly. The decomposition outlined in Equation (3) extends trivially to our regularized counterpart as well. Clearly the regularized formulation can indeed also be decomposed as a first-stage problem

$$(12) = \min_{z_0 \in Z_0} \ell_0(z_0) + \mathbb{E}_{P'_h}\left[Q(z_0, \xi'; M_h)\right] \tag{15}$$

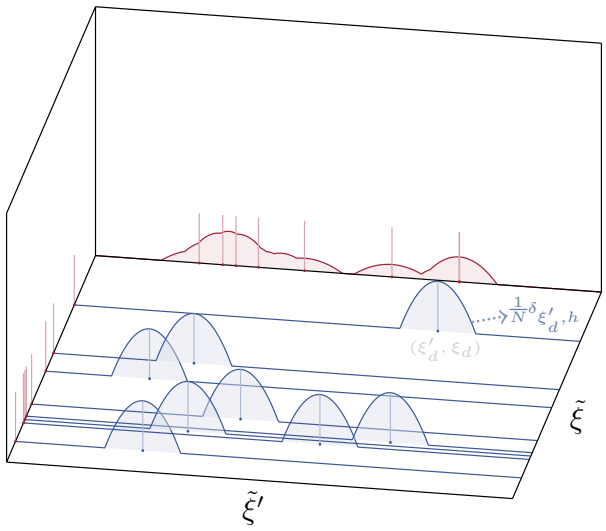


Figure 2: The discrete distributions $M_N$ and $P'_N$ are depicted here as the blue and red sticks, respectively. Similarly, the regularized versions $M_{N,h} := \sum_{d=1}^{N} \delta_{\xi'_d,h} \otimes \delta_{\xi_d}/N$ and $P'_{N,h} = \sum_{d=1}^{N} \delta_{\xi'_d,h}/N$ are depicted here as the blue and red curves, respectively.

where $P'_h = \int \delta_{\xi',h} \, \mathrm{d}P'(\xi') \in \mathcal{P}(\Xi')$ can be interpreted as a regularized version of $P'$ and a subsequent cost-to-go function $Q_h(z_0, \xi'; M_h)$ defined in Equation (4).

The regularization parameter $h > 0$ controls the amount of artificial perturbation introduced. The fact that adding the artificial perturbation $h\tilde{e}$ (which can also be interpreted as an artificial noise term) to the covariates $\tilde{\xi}$ provides beneficial regularization is certainly not new and has been observed in other learning problem before. Holmstrom and Koistinen (1992) discuss for instance the possibility of improving the generalization capability of a neural network by the introduction of additive noise to the training data. Modern deep learning frameworks such as Keras (Chollet et al. 2015) offer nowadays even offer native support to mitigate neural network training against overfitting through so called Gaussian noise regularization layers[2]. Clearly, though, adding artificial perturbation is counterproductive when the joint distribution $M$ is known as indeed the next result points out it (perhaps unsurprisingly) lead to performance degradation.

**Lemma 2.1.** *For any regularization parameter $h \geq 0$ we have* (1) $\leq$ (12).

*Proof.* Fix any first-stage decision $z_0 \in Z_0$. We have $\min_{z \in \mathcal{F}} \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) \right] = \min_{\bar{z} \in \bar{\mathcal{F}}} \mathbb{E}_{M \otimes E} \left[ \ell(z_0, \bar{z}(\tilde{\xi}', \tilde{e}), \tilde{\xi}) \right]$ where $\bar{\mathcal{F}} :=$ the set all measurable functions mapping $(\xi', e)$ to points in $Z$. Indeed, as $\tilde{e}$ and $\tilde{\xi}$ are independent

[2]See https://keras.io/api/layers/regularization_layers/gaussian_noise.

random variables we have that

$$\begin{aligned}\min_{\bar{z}\in\bar{\mathcal{F}}} \mathbb{E}_{M\otimes E}\left[\ell(z_0,\bar{z}(\tilde{\xi}',\tilde{e}),\tilde{\xi})\right] =& \min_{\bar{z}\in\bar{\mathcal{F}}} \mathbb{E}_{P'\otimes E}\left[\mathbb{E}_{M\otimes E}\left[\ell(z_0,\bar{z}(\tilde{\xi}',\tilde{e}),\tilde{\xi})|\tilde{\xi}',\tilde{e}\right]\right]\\ =&\mathbb{E}_{P'\otimes E}\left[\min_{z\in Z}\ \mathbb{E}_{M\otimes E}\left[\ell(z_0,z,\tilde{\xi})|\tilde{\xi}',\tilde{e}\right]\right]\\ =&\mathbb{E}_{P'}\left[\min_{z\in Z}\ \mathbb{E}_{M}\left[\ell(z_0,z,\tilde{\xi})|\tilde{\xi}'\right]\right]\\ =&\min_{z\in\mathcal{F}} \mathbb{E}_{M}\left[\ell(z_0,z(\tilde{\xi}'),\tilde{\xi})\right].\end{aligned}$$

The first equality is due to the law of total expectation. The second equality follows from (Rockafellar and Wets 2009, Theorem 14.60). The third equality is due to the fact that $\tilde{\xi}$ is independent of $\tilde{e}$ and hence we have $\mathbb{E}_{M\otimes E}\left[\ell(z_0,\tilde{z}(\tilde{\xi}',\tilde{e}),\tilde{\xi})|\tilde{\xi}',\tilde{e}\right] = \mathbb{E}_{M\otimes E}\left[\ell(z_0,\tilde{z}(\tilde{\xi}',\tilde{e}),\tilde{\xi})|\tilde{\xi}'\right]$. The result now follows from the fact that functions of the form $(\xi',e)\mapsto z(\xi'+he)$ are necessarily in $\bar{\mathcal{F}}$ for any $z\in\mathcal{F}$. □

**Remark 2.2.** *We remark that Lemma 2.1 remains true even if the added noise $\tilde{e}$ is not independent from $(\tilde{\xi}',\tilde{\xi})$. Indeed, in the proof of Lemma 2.1 we only use that $\tilde{\xi}$ is conditionally independent from $\tilde{e}$ given $\tilde{\xi}'$.*

Nevertheless, when the distribution $M$ is not known and needs to be learned from data, the regularizing effect of artificial perturbation will be critical as we will point out later on. We consider in this paper indeed the regularized data-driven formulations

$$\min_{z_0\in Z_0,\, z\in\mathcal{F}}\ \ell_0(z_0)+\mathbb{E}_{M_N\otimes E}\left[\ell(z_0,z(\tilde{\xi}'+h\tilde{e}),\tilde{\xi})\right] \tag{16}$$

where the dependence of the regularization parameter $h$ on the number of data points $N$ is not made explicit. That is we substitute the empirical distribution $M_N$ for $M$ in the noise regularized data-driven formulation (14) rather than in the original formulation (1) directly. Intuitively, the artificial perturbation allows to capture more effectively the effect the uncertainty $\tilde{\xi}$ conditioned on the covariate $t\xi'$. In particular, whereas the empirical distribution of $\tilde{\xi}$ conditioned on $\tilde{\xi}'$ realizing as $\xi'$ is only well defined when $\xi'$ is one of the observed covariates $\{\xi'_d\}_{d=1}^N$ where it equals $\sum_{d=1}^N \mathbb{1}\{\xi'_d=\xi'\}\,\delta_{\xi_d}/\sum_{d=1}^N \mathbb{1}\{\xi'_d=\xi'\}$, the regularized empirical distribution of $\tilde{\xi}$ conditioned on $\tilde{\xi}'$ realizing as $\xi'$ is well defined and given as $\sum_{d=1}^N k((\xi'-\xi'_d)/h)\delta_{\xi_d}/\sum_{d=1}^N k((\xi'-\xi'_d)/h)$ for $\xi'\in\{\xi'_d\}_{d=1}^N+hK$. In Section 2.2, we show that unlike substituting $M_N$ for $M$ in the original formulation, the regularized data-driven formulation (16) can be made consistent under very mild assumptions. We conclude this section by pointing out that our regularized data-driven formulation admits an interpretation based on two classical nonparametric estimation methods by Nadaraya-Watson and Parzen.

**Remark 2.3** (Nonparametric Estimation)**.** *Let $M_{N,h} := \sum_{d=1}^N \delta_{\xi'_d,h}\otimes\delta_{\xi_d}/N$ and $P'_{N,h}=\sum_{d=1}^N \delta_{\xi'_d,h}/N$ be the regularized counterparts to the empirical distributions $M_N$ and $P_N$, respectively. We have that the regularized data-driven formulation* (16) *can be decomposed as a first-stage problem $\min_{z_0\in Z_0}\ \ell_0(z_0)+\mathbb{E}_{P'_{N,h}}\left[Q(z_0,\tilde{\xi}';M_{N,h})\right]$ where*

*the associated cost-to-go function defined in Equation* (4) *satisfies*

$$Q(z_0, \xi'; M_{N,h}) = \min_{z \in Z} \ \sum_{d=1}^{N} \ell(z_0, z, \xi_d) k((\xi' - \xi'_d)/h) / \sum_{d=1}^{N} k((\xi' - \xi'_d)/h) \quad \forall \xi' \in \mathrm{R}^{d_2}. \tag{17}$$

*This cost-to-go function coincides precisely with substituting the conditional expectation* $\mathbb{E}_M \left[\ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi'\right]$ *with the Nadaraya-Watson estimate* $\mathbb{E}_{P_{\xi', N}} \left[\ell(z_0, z, \tilde{\xi})\right]$ *for all* $\xi' \in \mathrm{R}^{d_2}$ *based on the observed data discussed in Section 1. Furthermore, here the distribution* $P'_{N,h}$ *has density*

$$\mathrm{d}P'_{N,h}/\mathrm{d}m'(\xi') = f'_{N,h}(\xi') := \textstyle\sum_{d=1}^{N} k((\xi' - \xi'_d)/h)/(h^{d_2} N) \quad \forall \xi' \in \mathrm{R}^{d_2}$$

*which coincides here precisely with the classical density estimate of Parzen (1962) for the covariate distribution* $P'$*. In the context of Nadaraya-Watson and Parzen estimation the density function* $k$ *is often referred to as the kernel or weighing function and the regularization parameter* $h > 0$ *as the regularization parameter.*

## 2.2 Consistency

It is of interest to consider whether the regularized data-driven formulation (16) can recover the solution to the original problem (1). If the training data is obtained as independent samples from the unknown distribution $M$ it is not unreasonable to expect such consistency results to hold when the number of data points increases. We prove that the regularized formulation (16) is consistent when the regularizing perturbation is chosen carefully. In particular, we prove consistency for any regularizing perturbation which has a proper density function.

**Definition 2.4** (Proper Density Functions)**.** *The density function* $k$ *is proper if (i)* $k$ *is a measurable non-negative function and normalized so that* $\int k(e)\, \mathrm{d}m'(e) = 1$*, (ii)* $k$ *is* $\mathcal{K}$*-Lipschitz, and (iii)* $k$ *has compact starshaped support* $K = \mathrm{cl}(\{\xi' \in \Xi' \ : \ k(\xi') > 0\})$ *with Lipschitz boundary.*

Measurability and non-negativity of $k$ together with the normalization condition $\int k(e)\, \mathrm{d}m'(e) = 1$ guarantee that that there is indeed a distribution $E$ with $\mathrm{d}E/\mathrm{d}m'(e) = k(e)$ for all $e \in \mathrm{R}^{d_2}$ with respect to the scaled Lebesgue measure $m'$. We require a Lipschitz condition on $k$ to be proper which excludes the uniform density function depicted in Figure 1. The final condition poses the most stringent requirement. We remark here that most density functions $k$ in the literature are of the form $k(e) = S(\|e\|)$ for some function $S : \mathrm{R}_+ \to \mathrm{R}_+$ and hence depend on $e$ only through its Euclidean norm. In this case the support is convex and hence is star shaped and its boundary is clearly Lipschitz. The Epanechnikov, tricubic and uniform smoother density functions are proper; see Table 1.

Consistency of the data-driven formulation (16) will require the regularity conditions stated in Assumption 2.5 which are standard in nonparametric density estimation.

**Assumption 2.5.** *The distribution $P'$ of the covariate vector $\tilde{\xi}'$ admits a bounded density function $f'$, i.e.,* $\sup_{\xi' \in \Xi'} f'(\xi') < \infty$. *Finally, the functions*

$$\xi' \mapsto f'(\xi') \text{ and } \xi' \mapsto \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] f'(\xi')$$

*are continuous for any first-stage decision $z_0 \in Z_0$ and second-stage decision $z \in Z$.*

By assuming that the density function $f'$ satisfies $\sup_{\xi' \in \Xi'} f'(\xi') < \infty$ it is guarantees that $\tilde{\xi}'$ is not too concentrated around any point in $\Xi'$. The condition that the density function $f'$ is continuous ensures, as $\Xi'$ is compact, that $f'$ is uniformly continuous and bounded. Hence, following Devroye and Penrod (1984) it can be learned from data in a strong uniform sense. The same observation can be made concerning the function $\xi' \mapsto \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] f'(\xi')$. We now further illustrate with the help of the next lemma that all imposed conditions are in fact very mild.

**Lemma 2.6** (Noise Regularized Formulations)**.** *Let Assumption 1.2 hold for problem* (1) *with $k$ a proper density function. Then any regularized counterpart*

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ \ell_0(z_0) + \mathbb{E}_{M_h} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) \right] \tag{18}$$

*with $h > 0$ satisfies Assumption 2.5.*

*Proof.* As the density function $k$ is proper it is bounded above by $\bar{k} > 0$. The distribution $P'_h$ admits density function $f'_h(\xi') = \int k((\xi' - \xi'')/h)/h^{d_2} \, \mathrm{d}P'(\xi'')$ which is hence bounded from above by $\bar{k}/h^{d_2}$ on $\Xi' + hK$.

The function $f'_h$ is $\mathcal{K}/h^{d_2+1}$-Lipschitz continuous as for all $\xi'_1 \in \Xi' + hK$ and $\xi'_2 \in \Xi' + hK$ we have

$$\begin{aligned} |f'_h(\xi'_1) - f'_h(\xi'_2)| &\leq \int |k((\xi'_1 - \xi'')/h) - k((\xi'_2 - \xi'')/h)| \, /h^{d_2} \, \mathrm{d}P'(\xi'') \\ &\leq \int \mathcal{K} \, |(\xi'_1 - \xi'')/h - (\xi'_2 - \xi'')/h| \, /h^{d_2} \, \mathrm{d}P'(\xi'') = \frac{\mathcal{K}}{h^{d_2+1}} \, |\xi'_1 - \xi'_2| \, . \end{aligned}$$

Similarly we observe that $\xi'' \mapsto \mathbb{E}_{M_h} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi'' \right] f'_h(\xi'') = \int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi)$ is $L\mathcal{K}/h^{d_2+1}$-Lipschitz continuous for all $z_0 \in Z$ and $z \in Z$. Indeed,

$$\begin{aligned} &\left| \int \ell(z_0, z, \xi) k((\xi''_1 - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi) - \int \ell(z_0, z, \xi) k((\xi''_2 - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi) \right| \\ &\leq \int \ell(z_0, z, \xi) \, |k((\xi''_1 - \xi')/h) - k((\xi''_2 - \xi')/h)| \, /h^{d_2} \mathrm{d}M(\xi', \xi) \\ &\leq \int \ell(z_0, z, \xi) \frac{\mathcal{K}}{h^{d_2+1}} \, |\xi''_1 - \xi''_2| \, \mathrm{d}M(\xi', \xi) \leq \frac{L\mathcal{K}}{h^{d_2+1}} \end{aligned}$$

for all $\xi_1'' \in \Xi' + hK$, $\xi_2'' \in \Xi' + hK$, $z_0 \in Z$ and $z \in Z$. $\square$

The previous lemma states that any problem which satisfies Assumption 1.2 will also satisfy Assumption 2.5 if we add to the covariate data any (potentially imperceptible) amount of perturbation regularization. Under these assumptions the data-driven formulation (16) can recover the solution to the original problem (1).

**Theorem 2.7** (Consistency)**.** *Let Assumptions 1.2 and 2.5 hold, $k$ be a proper density function and $\Xi'$ compact. Consider regularization satisfying $\lim_{N\to\infty} h(N) = 0$ and $\lim_{N\to\infty} N^{1-\omega} h^{d_2}(N)/\log(N) = \infty$ for some $\omega > 0$. Denote with $(z_{0N}^\star, z_N^\star)$ and $v_N^\star$ the minimizers and minimum of*

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ell_0(z_0) + \mathbb{E}_{M_N \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right],$$

*respectively. Then,*

1. *we have $\lim_{N\to\infty} v_N^\star = v^\star$ almost surely, and*
2. *$\lim_{N\to\infty} \ell_0(z_{0N}^\star) + \mathbb{E}_{M\otimes E}\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right] = v^\star$ almost surely.*

*Proof.* We will prove that

$$\begin{aligned}
&\lim_{N\to\infty} \sup_{z_0 \in Z_0, z\in\mathcal{F}} \left|\mathbb{E}_{M_N\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]\right| \\
&= \lim_{N\to\infty} \sup_{z_0 \in Z_0, z\in\mathcal{F}} \left|\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]\right| = 0
\end{aligned}$$

almost surely. From this much stronger uniform convergence result the claimed results follow trivially. Indeed, the previous claim guarantees that there exist $\epsilon(N) \geq 0$ with $\lim_{N\to\infty} \epsilon(N) = 0$ so that we have for all $z_0 \in Z_0$ and $z \in \mathcal{F}$ that $\left|\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]\right| \leq \epsilon(N)$ for any $N \geq 1$. Equivalently, we have for all $z_0 \in Z_0$, $z \in \mathcal{F}$ the pair of inequalities

$$\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \leq \epsilon(N), \;\; \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \leq \epsilon(N).$$

In particular, applying the first inequality to $(z_0^\star, z^\star(\xi'))$ and the second to $(z_{0N}^\star, z_N^\star(\xi'))$ we get that

$$\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0^\star, z^\star(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_M\left[\ell(z_0^\star, z^\star(\tilde{\xi}'), \tilde{\xi})\right] \leq \epsilon(N), \;\; \mathbb{E}_M\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}'), \tilde{\xi})\right] \leq \epsilon(N). \tag{19}$$

Furthermore, we have that $v_N^\star = \ell_0(z_{0N}^\star) + \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}'), \tilde{\xi})\right] \leq \ell_0(z_0^\star) + \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0^\star, z^\star(\tilde{\xi}'), \tilde{\xi})\right]$ and $v^\star = \ell_0(z_0^\star) + \mathbb{E}_M\left[\ell(z_0^\star, z^\star(\tilde{\xi}'), \tilde{\xi})\right] \leq \ell_0(z_{0N}^\star) + \mathbb{E}_M\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}'), \tilde{\xi})\right]$ which guarantee that $\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0^\star, z^\star(\tilde{\xi}'), \tilde{\xi})\right] \geq$

$v_N^\star - \ell_0(z_0^\star)$ and $\mathbb{E}_M\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}'), \tilde{\xi})\right] \geq v^\star - \ell_0(z_{0N}^\star)$. Hence, from the pair of inequalities (19) it follows now that

$$\begin{aligned}
&v_N^\star - \ell_0(z_0^\star) - \mathbb{E}_M\left[\ell(z_0^\star, z^\star(\tilde{\xi}'), \tilde{\xi})\right] \leq \epsilon(N), \quad v^\star - \ell_0(z_{0N}^\star) - \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}'), \tilde{\xi})\right] \leq \epsilon(N) \\
\iff &v_N^\star - v^\star \leq \epsilon(N), \quad v^\star - v_N^\star \leq \epsilon(N) \iff |v_N^\star - v^\star| \leq \epsilon(N)
\end{aligned}$$

establishing the first claim. Furthermore, as we have $v^\star \leq \ell_0(z_{0N}^\star) + \mathbb{E}_M\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right] \leq \ell_0(z_{0N}^\star) + \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h(N)e), \tilde{\xi})\right] + \epsilon(N) = v_N^\star + \epsilon(N) \leq v^\star + 2\epsilon(N)$ proving the second claim.

Let us now prove the claimed uniform convergence. Note first that we have

$$\begin{aligned}
&\sup_{z_0 \in Z_0, z \in \mathcal{F}} \left|\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]\right| \\
= &\sup_{z_0 \in Z_0, z \in \mathcal{F}} \left|\mathbb{E}_{P'_{N.h(N)}}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right] - \mathbb{E}_{P'}\left[\mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right]\right| \\
\leq &\sup_{z_0 \in Z_0, z \in \mathcal{F}} \left|\mathbb{E}_{P'_{N.h(N)}}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right] - \mathbb{E}_{P'}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right]\right| \\
&\quad + \sup_{z_0 \in Z_0, z \in \mathcal{F}} \left|\mathbb{E}_{P'}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right] - \mathbb{E}_{P'}\left[\mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right]\right| \\
\leq &\sup_{z_0 \in Z_0, z \in \mathcal{F}} \left|\mathbb{E}_{P'_{N,h(N)}}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right] - \mathbb{E}_{P'}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right]\right| \\
&\quad + \sup_{z_0 \in Z_0, z \in \mathcal{F}} \mathbb{E}_{P'}\left[\left|\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right|\right].
\end{aligned}$$

Here the first equality follows from the law of total expectation. The first and second inequalities follows from trivial applications of the triangle inequality. We show now that we have almost surely

$$\lim_{N \to \infty} \sup_{z_0 \in Z_0, z \in \mathcal{F}} \left|\mathbb{E}_{P'_{N,h(N)}}\left[\mathbb{E}_{M_{N,h}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right] - \mathbb{E}_{P'}\left[\mathbb{E}_{M_{N,h}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right]\right| = 0 \tag{20}$$

and

$$\lim_{N \to \infty} \sup_{z_0 \in Z_0, z \in \mathcal{F}} \mathbb{E}_{P'}\left[\left|\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right] - \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right|\right] = 0 \tag{21}$$

establishing the uniform convergence claim.

Note that as per Assumption 1.2 the loss function $\ell$ is bounded above by $L$ we have $\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}' = \xi'\right] \leq L$ for all $\xi' \in \mathrm{R}^{d_2}$. Hence, we have

$$\begin{aligned}
&\left|\mathbb{E}_{P'_{N,h}}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right] - \mathbb{E}_{P'}\left[\mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}'\right]\right]\right| \\
\leq &\left|\int \mathbb{E}_{M_{N,h(N)}}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})|\tilde{\xi}' = \xi'\right] \mathrm{d}(P'_{N,h} - P')(\xi')\right| \leq L \int \left|f_{N,h(N)}(\xi') - f(\xi')\right| \mathrm{d}m'(\xi')
\end{aligned}$$

with $f'_{N,h}$ and $f'$ the density functions of $P'_{N,h}$ and $P'$ with respect to the base measure $m'$, respectively. Devroye and Penrod (1984, Theorem 2(B)) establish that if the density function $k$ has compact support $K$ we have

$$\lim_{N\to\infty} \int \left| f'_{N,h(N)}(\xi') - f'(\xi') \right| \mathrm{d}m'(\xi') = 0$$

almost surely from which the limit (20) follows immediately.

Under the conditions stated in the theorem, Devroye and Krzyzak (1989) show that we have for any given $z_0 \in Z_0$ and $z \in \mathcal{F}$ that $\lim_{N\to\infty} \mathbb{E}_{P'} \left[ \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' \right] \right| \right] = 0$ almost surely. However, we need the stronger statement (21) in which the previous convergence takes place uniformly in $z_0$ and $z$. For the sake of contradiction assume that we have

$$\limsup_{N\to\infty} \sup_{z_0 \in Z_0, z \in \mathcal{F}} \mathbb{E}_{P'} \left[ \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' \right] \right| \right] \geq \epsilon$$

almost surely for some $\epsilon > 0$.

Consider the set $A^c = \{\xi' \in \Xi' \ : \ f'(\xi') \leq \epsilon/(8Lm'(\Xi'))\}$ which is closed as the density function $f'$ of $P'$ is here a continuous function. By construction, we have that

$$P'(A^c) = \int \mathbb{1}\{\xi' \in A\} f'(\xi') \, \mathrm{d}m'(\xi') \leq \int \mathbb{1}\{\xi' \in A\} \frac{\epsilon}{8Lm'(\Xi')} \, \mathrm{d}m'(\xi') = \epsilon/(8L).$$

Hence, its complement $A = \{\xi' \in \Xi' \ : \ f'(\xi') > \epsilon/(8Lm'(\Xi'))\}$ is open and satisfies $P'(A) \geq 1 - \epsilon/(8L)$. Remark that the sets $A_\delta := \{\xi' \in \Xi' \ : \ B(\xi', \delta) \subseteq A\}$ with $B(\xi', \delta) = \{\xi'' \in \Xi' \ : \ \|\xi'' - \xi'\| \leq \delta\}$ are decreasing in $\delta \geq 0$ and that $\mathrm{int}(A) = A = \cup_{k\geq 1} A_{1/k}$. By the continuity of measure we have that

$$P'[A] = P'[\cup_{k\geq 1} A_{1/k}] = \lim_{k\to\infty} P'[A_{1/k}].$$

We may hence consider a $\delta > 0$ sufficiently small so that $P'(A \setminus A_\delta) \leq \epsilon/(8L)$. Clearly we have

$$\begin{aligned}
&\mathbb{E}_{P'} \left[ \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' \right] \right| \right] \\
&= \int_{A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \mathrm{d}P'(\xi') \\
&\quad + \int_{A \setminus A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi} = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi} = \xi' \right] \right| \mathrm{d}P'(\xi') \\
&\quad + \int_{A^c} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi} = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| f'(\xi') \, \mathrm{d}m'(\xi') \\
&\leq \int_{A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi} = \xi' \right] \right| \mathrm{d}P'(\xi') + 2LP'(A \setminus A_\delta) + \frac{2L\epsilon}{8Lm'(\Xi')} m'(\Xi')
\end{aligned}$$

$$\leq \int_{A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \, \mathrm{d}P'(\xi') + \epsilon/2$$

and hence

$$\limsup_{N\to\infty} \sup_{z_0 \in Z_0,\, z \in \mathcal{F}} \int_{A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \, \mathrm{d}P'(\xi') \geq \epsilon/2 > 0.$$

To establish a contradiction we exploit the regularity conditions imposed on the cost function $\ell$ considered. From Assumption 1.2 we have in particular for any decision $(z_0, z)$ in $Z_0 \times Z$ the Lipschitz upper bound

$$\sup_{(z_0', z') \in B((z_0,z),r)} |\ell(z_0', z', \xi) - \ell(z_0, z, \xi)| \leq \mathcal{L} r$$

for all $\xi \in \Xi$ and $r \geq 0$. Hence, we have

$$\begin{aligned}
&\sup_{z_0 \in Z_0, z \in \mathcal{F}} \int_{A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \, \mathrm{d}P'(\xi') \\
\leq &\sup_{z_0 \in Z_0, z \in \mathcal{F}} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z(\tilde{\xi}'), \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \\
= &\sup_{z_0 \in Z_0, z \in Z} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \\
\leq &\sup_{(z_0,z) \in S} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| + \mathcal{L} r
\end{aligned}$$

with $S$ a finite set chosen so that $\cup_{(z_0,z)\in S} B((z_0, z), r) \supseteq Z_0 \times Z$ which exists as $Z_0 \times Z$. Here the equality follows from the variable substitution $z \leftarrow z(\xi') \in Z$. The ultimate inequality follows from the $\mathcal{L}$-Lipschitz continuity of $\mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \xi) | \tilde{\xi}' = \xi' \right]$ in $(z_0, z) \in Z_0 \times Z$ for all $\xi' \in \mathrm{R}^{d_2}$. Choose now some $0 < r < \epsilon/(4\mathcal{L})$ then

$$\limsup_{N\to\infty} \sup_{(z_0,z)\in S} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \geq \epsilon/4 > 0$$

almost surely. This however is a contradiction with Liero (1989, Theorem 4) who shows that under the conditions imposed we have in fact $\limsup_{N\to\infty} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' \right] - \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| = 0$ almost surely for all $\xi' \in \mathrm{R}^{d_1}$ and as $S$ is finite

$$\begin{aligned}
&\limsup_{N\to\infty} \sup_{(z_0,z)\in S} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| \\
= &\sup_{(z_0,z)\in S} \limsup_{N\to\infty} \sup_{\xi' \in A_\delta} \left| \mathbb{E}_{M_{N,h(N)}} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] - \mathbb{E}_M \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi' \right] \right| = 0
\end{aligned}$$

almost surely. Hence, the claim (21) must indeed be true. □

It should be remarked that to obtain consistent formulations the regularization parameter $h(N)$ cannot decrease arbitrarily fast as we need $\lim_{N\to\infty} N^{1-\omega} h^{d_2}(N)/\log(N) = \infty$ for some $\omega > 0$ and hence its decrease becomes slower as the covariate dimension $d_2$ increases. This observation is often referred to as the curse of dimensionality and plagues all nonparametric learning formulations. The previous result indicates however that as the regularization parameter $h(N)$ tends to zero at the appropriate rate, the regularized data driven driven formulation recovers the optimal cost asymptotically (Theorem 2.7, point 1) and the out-of-sample cost of its data-driven solution $z_{0N}^\star, z_N^\star$ when implemented indicated in the figure below converges to the optimal cost (Theorem 2.7, point 2).

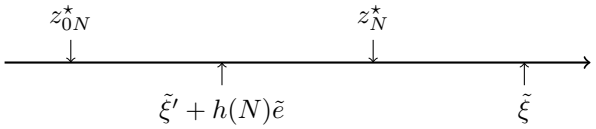


Figure 3: Theorem 2.7, point 2 ensures that when implemeting the first stage decision $z_{N0}^\star$ and committing to the second-stage decision plan $z_N^\star(\tilde{\xi}' + h(N)\tilde{e})$ when the covariate context $\tilde{\xi}'$ realizes achieves an expected cost $\ell_0(z_{0N}^\star) + \mathbb{E}_{M\otimes E}\left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right]$ and is asymptotically optimal.

## 2.3 Computation

From a practical perspective solving the data-driven formulation (16) is challenging for two reasons. First, observe that evaluating the expectation $\mathbb{E}_{M_N\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right]$ for given decisions $z_0 \in Z_0$ and $z \in \mathcal{F}$ in the objective of our proposed data-driven formulation (16) potentially requires high-dimensional integration. Second, the second-stage optimization variable $z \in \mathcal{F}$ may be infinite dimensional (as it is a function rather than a vector) further complicating the practical resolution of our proposed data-driven formulation. We will propose here an efficient stochastic approximation algorithm to solve the data-driven formulation (16) which circumvents both issues.

Notice first that for a given first stage decision $z_0$, the optimal second-stage cost $Q(z_0, \xi'; M_{N,h})$ can be computed efficiently by solving the ordinary finite dimensional convex optimization problem (17) for a fixed covariate $\xi' \in \Xi'$ and distribution $M_{N,h}$. Moreover, the optimal second-stage plan is the minimizer of this optimization problem. Hence, the main challenge is in computing the optimal first stage decision $z_{0N}^\star$ and the optimal cost $\min_{z_0\in Z_0} F(z_0) =$ (16), where $F(z_0) := \ell(z_0) + \min_{z\in\mathcal{F}} \mathbb{E}_{M_N\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] = \ell_0(z_0) + \mathbb{E}_{P'_{N,h}}\left[Q(z_0, \tilde{\xi}'; M_{N,h})\right]$. The following lemma characterizes the subgradients of $F$ which will then allow us to introduce a stochastic gradient decent algorithm to optimize $F$ and derive the optimal first stage decision.

**Lemma 2.8** (Subgradients)**.** *Let Assumption 1.2 hold. We have that*

$$\partial_{z_0} F(z_0) = \partial\ell_0(z_0) + \mathbb{E}_{P'_{N,h}}\left[\left\{g \in \mathrm{R}^{n_0} \;:\; (g,0) \in \mathbb{E}_{M_{N,h}}\left[\partial_{z_0,z}\ell(z_0, z_N^\star(z_0,\tilde{\xi}'),\tilde{\xi})|\tilde{\xi}'\right]\right\}\right] \subseteq \mathrm{R}^{n_0}$$

*with* $z_N^\star(z_0,\xi') \in \arg\min_{z\in Z}\sum_{d=1}^N \ell(z_0,z,\xi_d)k((\xi'-\xi'_d)/h)/\sum_{d=1}^N k((\xi'-\xi'_d)/h)$ *for all* $z_0 \in Z_0$ *and* $\xi' \in \Xi' + hK$.

*Proof.* As Assumption 1.2 implies that the function $Q(z_0,\xi';M_{N,h})$ is bounded by $L$ for all values of its arguments, we have following Strassen (1965) the equivalence $\partial_{z_0}\mathbb{E}_{P'_{N,h}}\left[Q(z_0,\tilde{\xi}';M_{N,h})\right] = \mathbb{E}_{P'_{N,h}}\left[\partial_{z_0}Q(z_0,\tilde{\xi}';M_{N,h})\right]$. Furthermore, we have $Q(z_0,\xi';M_{N,h}) = \min_{z\in Z}\mathbb{E}_{M_{N,h}}\left[\ell(z_0,z,\tilde{\xi})|\tilde{\xi}'=\xi'\right]$ where one can argue that $(z_0,z)\to\mathbb{E}_{M_{N,h}}\left[\ell(z_0,z,\tilde{\xi})|\tilde{\xi}'=\xi'\right]$ is convex and $\mathcal{L}$-Lipschitz continuous for any $\xi'\in\Xi'+hK$. Indeed,

$$\begin{aligned}
&\left|\mathbb{E}_{M_{N,h}}\left[\ell(z_0^1,z^1,\tilde{\xi})|\tilde{\xi}'=\xi'\right] - \mathbb{E}_{M_{N,h}}\left[\ell(z_0^2,z^2,\tilde{\xi})|\tilde{\xi}'=\xi'\right]\right| \\
&\leq \mathbb{E}_{M_{N,h}}\left[\left|\ell(z_0^1,z^1,\tilde{\xi}) - \ell(z_0^2,z^2,\tilde{\xi})\right||\tilde{\xi}'=\xi'\right] \\
&\leq \mathbb{E}_{M_{N,h}}\left[\mathcal{L}(\left\|z_0^1-z_0^2\right\| + \left\|z^1-z^2\right\|)|\tilde{\xi}'=\xi'\right] = \mathcal{L}(\left\|z_0^1-z_0^2\right\| + \left\|z^1-z^2\right\|).
\end{aligned}$$

for all $z_0^1, z_0^2 \in Z_0$ and $z^1, z^2 \in Z$ following again Assumption 1.2. Following Lemma C.1 we have that

$$\partial_{z_0}Q_h(z_0,\xi') = \left\{g \;:\; (g,0)\in\partial_{z_0,z}\mathbb{E}_{M_{N,h}}\left[\ell(z_0,z_N^\star(z_0,\xi'),\tilde{\xi})|\tilde{\xi}'=\xi'\right]\right\} \subseteq \{g \;:\; \|(g,0)\|\leq\mathcal{L}\}$$

for any $z_N^\star(z_0,\xi')\in\arg\min_{z\in Z}\mathbb{E}_{M_{N,h}}\left[\ell(z_0,z,\tilde{\xi})|\tilde{\xi}'=\xi'\right]$ for all $z_0\in Z_0$ and $\xi'\in\Xi'+hK$. As the loss function $\ell$ is itself bounded by $L$ we also have $\partial_{z_0,z}\mathbb{E}_{M_{N,h}}\left[\ell(z_0,z_N^\star(z_0,\xi'),\tilde{\xi})|\tilde{\xi}'=\xi'\right] = \mathbb{E}_{M_{N,h}}\left[\partial_{z_0,z}\ell(z_0,z_N^\star(z_0,\xi'),\tilde{\xi})|\tilde{\xi}'=\xi'\right]$ from which the result follows immediately. □

Assume for the sake of simplicity that the loss functions $\ell_0$ and $\ell(z_0,z,\xi)$ are differentiable in $z_0$ and $(z_0,z)$, respectively, for every $\xi\in\Xi$. Then, we have following Lemma 2.8

$$\partial_{z_0}F(z_0) = \ell_0'(z_0) + \mathbb{E}_{P'_{N,h}}\left[(\textstyle\sum_{d=1}^N \ell'(z_0,z_N^\star(z_0,\tilde{\xi}'),\xi_d)k((\tilde{\xi}'-\xi'_d)/h))/\sum_{d=1}^N k((\tilde{\xi}'-\xi'_d)/h)\right].$$

We can hence consider a stochastic subgradient of $F$ at $z_0$ for all $\xi'$ as

$$G_{z_0}(z_0,\xi') := \ell_0'(z_0) + (\textstyle\sum_{d=1}^N \ell'(z_0,z_N^\star(z_0,\xi'),\xi_d)k((\xi'-\xi'_d)/h))/\sum_{d=1}^N k((\xi'-\xi'_d)/h)$$

which following Lemma 2.8 is unbaised and enjoys bounded variance when $\xi'$ is sampled following the noise regularized covariate distribution $P'_{N,h}$. We remark that sampling from $P'_{N,h}$ can be done efficiently by sampling with replacement a covariate from among the observed covariates $\{\xi'_1,\ldots,\xi'_N\}$ and subsequently perturbing the so obtained sample by an independent sample from the artificial perturbation distribution $E$.

**Lemma 2.9** (Stochastic Subgradient)**.** *Let Assumption 1.2 hold. We have* $\mathbb{E}_{P'_{N,h}}\left[\left\|G_{z_0}(z_0,\tilde{\xi}')\right\|^2\right] \leq (\mathcal{L}_0+\mathcal{L})^2$.

*Proof.* Following Assumption 1.2 we have

$$\begin{aligned}
&\mathbb{E}_{P'_{N,h}}\left[\left\|G_{z_0}(z_0,\tilde{\xi}')\right\|^2\right] \\
\leq &\mathbb{E}_{P'_{N,h}}\left[(\|\ell'_0(z_0)\| + \textstyle\sum_{d=1}^N \left\|\ell'(z_0, z_N^\star(z_0,\tilde{\xi}'),\xi_d)\right\| k((\tilde{\xi}'-\xi'_d)/h)/\sum_{d=1}^N k((\tilde{\xi}'-\xi'_d)/h))^2\right] \\
\leq &\mathbb{E}_{P'_{N,h}}\left[(\mathcal{L}_0 + \textstyle\sum_{d=1}^N \mathcal{L}k((\tilde{\xi}'-\xi'_d)/h)/\sum_{d=1}^N k((\tilde{\xi}'-\xi'_d)/h))^2\right] = (\mathcal{L}_0+\mathcal{L})^2
\end{aligned}$$

which establishes the claim. □

As the stochastic gradients are unbiased and have bounded variance we may solve the minimization problem of interest $\min_{z_0\in Z_0} F(z_0)$ using a stochastic gradient descent approach. To account for the constraints $z_0 \in Z_0$ on the first stage decision, we project each iterate back into the feasible region using the projection map $P_{z_0}(g) = \arg\min_{z'_0\in Z_0} \|z_0 - g - z'_0\|_2^2$. The resulting algorithm is depicted in Algorithm 1.

Notice that our algorithm can be easily generalized to a general mirror descent approach of Nemirovski et al. (2009). Let indeed $\nu$ be any continuously differentiable and strongly convex distance generating function on $\mathrm{R}^{n_0}$ with associated Bregman divergence $V(z_0, z'_0) = \nu(z'_0) - [\nu(z_0) - \langle\nu'(z_0), z'_0 - z_0\rangle]$ and proximal mapping

$$P_{z_0} : g \mapsto \arg\min_{z'_0\in Z_0} \langle g, z'_0 - z_0\rangle + V(z_0, z'_0).$$

We assume here without loss of generality that the distance generating function $\nu$ is scaled appropriately so that its strong convexity parameter is one, i.e., $\nu(z'_0) \geq \nu(z_0) + \langle\nu'(z'_0), z'_0 - z_0\rangle + \|z'_0 - z_0\|^2$ for all $z'_0$ and $z_0$ in $\mathrm{R}^{n_0}$. One particular choice, is to consider the distance generating function $\nu(z_0) = \|z_0\|_2^2/2$ which is continuously differentiable and strongly convex and for which we have that its associated proximal mapping reduces to a simple projection of the point $z_0 - g$ onto the constraint set $Z_0$, i.e., $P_{z_0}(g) = \arg\min_{z'_0\in Z_0} \|z_0 - g - z'_0\|_2^2$, which we discussed previously. Nemirovski et al. (2009) suggest producing iterates as suggested in Algorithm 1.

Let $D_{z_0}^2 := \max_{z'_0\in Z_0} \nu(z'_0) - \min_{z'_0\in Z_0} \nu(z'_0)$ denote the size of the set $Z_0$ as measured by the distance generating function $\nu$ considered. We may measure the suboptimality of any point $z_0$ as $\epsilon(z_0) := F(z_0) - \min_{z'_0\in Z_0} F(z'_0)$. Nemirovski et al. (2009) show that the weighted average $z_0^F = \sum_{j=1}^T z_0^j\gamma_j/(\sum_{j=1}^T \gamma_j)$ of the the iterates produced by Algorithm 1 satisfies $\mathbb{E}\left[\epsilon(z_0^F)\right] \leq (D_{z_0}^2 + \frac{1}{2}(\mathcal{L}_0+\mathcal{L})^2\sum_{j=1}^T\gamma_j^2)/\sum_{j=1}^T\gamma_j$. By choosing step lengths as $\gamma_j = \sqrt{2}D_{z_0}/((\mathcal{L}_0+\mathcal{L})\sqrt{T})$ we can guarantee that after $T$ iterations we obtain the suboptimality guarantee $\mathbb{E}\left[\epsilon(z_0^F)\right] \leq D_{z_0}(\mathcal{L}_0+\mathcal{L})\sqrt{2/T}$.

**Algorithm 1:** Stochastic Mirror Gradient Descent for Minimization Problems

**Initialization:** Starting point $z_0^1 \in \arg\min_{z_0' \in Z_0} \nu(z_0')$, regularization parameter $h > 0$, iteration number $T \geq 2$ and step lengths $\gamma_j > 0$ for $j \in [1, T-1]$.

**for** $j \in [1, \ldots, T-1]$ **do**

Sample $\bar{\xi}'$ randomly from the observed data points $D_N$ and $e$ following the perturbation distribution $E$

$\xi' = \bar{\xi}' + he$

$z^j \in \arg\min_{z \in Z} \sum_{d=1}^N \ell(z_0^j, z, \xi_d) k((\xi' - \xi_d')/h) / \sum_{d=1}^N k((\xi' - \xi_d')/h)$

$G_{z_0}^j = \ell_0'(z_0) + (\sum_{d=1}^N \ell'(z_0, z^j, \xi_d) k((\xi' - \xi_d')/h)) / \sum_{d=1}^N k((\xi' - \xi_d')/h)$

$z_0^{j+1} = P_{z_0^j}(\gamma_j G_{z_0}^j)$

**return** $z_0^F = \sum_{j=1}^T z_0^j \gamma_j / (\sum_{j=1}^T \gamma_j)$

## 2.4 Out-of-Sample Performance

Consider first the original stochastic optimization problem stated in Equation (1) for given first-stage and second-stage decisions $z_0 \in Z_0$ and $z \in \mathcal{F}$. The objective function $M \to \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]$ is not necessarily continuous in the distribution $M$. Indeed, according to the Portmanteau theorem this linear function is continuous if and only if $(\xi', \xi) \to \ell(z_0, z(\xi'), \xi)$ is bounded and continuous. However, as $z \in \mathcal{F}$ can be any measurable function this may clearly not be the case. The lack of regularity practically means that even though we have an empirical distribution $M_N$ which asymptotically converges to the unknown distribution $M$, i.e., $\lim_{N\to\infty} M_N = M$, it may still be that $\lim_{N\to\infty} \mathbb{E}_{M_N}\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \neq \mathbb{E}_M\left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]$. The regularized objective function $M \to \mathbb{E}_{M_N \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right]$ of the regularized data-driven formulation (16) on the other hand is always continuous.

**Proposition 2.10.** *Let Assumption 1.2 hold. We have for any sequence $\{M^k\}_{k\geq 1}$ with limit $M^\infty$ that*

$$\lim_{k\to\infty} \mathbb{E}_{M^k \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] = \mathbb{E}_{M^\infty \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right]$$

*for all $z_0 \in Z$ and $z \in \mathcal{F}$.*

*Proof.* The proof can be found in Appendix A.1. □

If we are willing to assume that the density function $k$ is Lipschitz then we can derive that the objective function of the data-driven formulation (16) is also Lipschitz.

**Proposition 2.11.** *Let Assumption 1.2 hold. We have for all $M_1$ and $M_2$ the inequality*

$$\mathbb{E}_{M_1 \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] - \mathbb{E}_{M_2 \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \leq \max(\mathcal{L}, 2\mathcal{K}L/h) \, \|M_1 - M_2\|$$

*for all $z_0 \in Z$ and $z \in \mathcal{F}$ with $\mathcal{K}$ the Lipschitz constant of the density function $k$ where the norm is defined in Equation* (11).

*Proof.* The proof can be found in Appendix A.2. □

Following Kantorovich and Rubinshtein (1958) we have $W(M_1, M_2) = \|M_1 - M_2\|$ where the Wasserstein distance is taken as $W(M_1, M_2) := \inf_{T \geq 0}\{\|u_1 - u_2\| \, \mathrm{d}T(u_1, u_2) \ : \ T \in \mathcal{P}((\Xi \times \Xi') \times (\Xi \times \Xi')),\ \Pi_{u_1} T = M_1,\ \Pi_{u_2} T = M_2\}$ and metrizes here the weak topology as $\Xi' \times \Xi$ is bounded. The transport polytope $\{T \in \mathcal{P}((\Xi \times \Xi') \times (\Xi \times \Xi')) \ : \ \Pi_{u_1} T = M_1,\ \Pi_{u_2} T = M_2\}$ denotes here the set of all distributions with given marginals $\Pi_{u_1} T$ and $\Pi_{u_2} T$. Hence, the result in Proposition 2.11 is strictly stronger than the result in Proposition 2.10. As pointed out in Proposition 2.11 the lack of continuity of the mapping $M \to \mathbb{E}_M \left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]$ can be explained by the fact that the Lipschitz constant grows unbounded as the regularization parameter $h$ tends to zero.

**Proposition 2.12.** *Let Assumptions 1.2 and 2.5 hold, $k$ be a proper $\mathcal{K}$-Lipschitz density function and $\Xi'$ compact. Consider the the regularization scaling $h(N) = \mathcal{O}(N^{-\gamma})$ for $\gamma < 1/(d_1 + d_2)$. Denote with $(z_{0N}^\star, z_N^\star)$ and $v_N^\star$ the minimizers and minimum in*

$$\min_{z_0 \in Z_0,\, z \in \mathcal{F}} \ell_0(z_0) + \mathbb{E}_{M_N \otimes E} \left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right],$$

*respectively, and let $d(N, \eta) = \max(\mathcal{L}, 2\mathcal{K}L/h(N))\delta_\eta(N)$ with $\delta_\eta(N) := \mathcal{O}(N^{-1/(d_1+d_2)} \log 1/\eta)$. Then,*

$$\limsup_{N \to \infty} \ M^\infty \left[\mathbb{E}_{M \otimes E} \left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] > v_N^\star + d(N, \eta)\right] \leq \eta$$

*with $\lim_{N \to \infty} d(N, \eta) = 0$.*

*Proof.* A standard concentration result by Fournier and Guillin (2015, Theorem 2) ensures that for $\delta_\eta(N)$ defined as before the Wasserstein ambiguity balls $\mathcal{W}_N := \{M' \in \mathcal{P}(\Xi' \times \Xi) \ : \ \|M_N - M'\| \leq \delta_\eta(N)\}$ are asymptotic $1 - \eta$ confidence regions for the unknown distribution $M$, i.e., $\limsup_{N \to \infty} M^\infty \left[M \notin \mathcal{W}_N\right] \leq \eta$. Hence, we have

$$\begin{aligned}
&\limsup_{N \to \infty} \ M^\infty \left[\mathbb{E}_{M \otimes E} \left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] > \sup\left\{\mathbb{E}_{M' \otimes E} \left[\ell(z_{0N}^\star, z_N^\star(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \ : \ \|M_N - M'\| \leq \delta_\eta(N)\right\}\right] \\
\leq &\limsup_{N \to \infty} \ M^\infty \left[M \notin \mathcal{W}_N\right] \leq \eta
\end{aligned}$$

establishing the claim. □

Hence, with the appropriate regularization scaling stated in Proposition 2.12 we have that the naive formulation suffers a disappointment at most a vanishing amount $d(N, \eta)$ with probability asymptotically at most $\eta$. We remark

that this observation is independent of whether or not the formulation is consistent. However, if the requirements on the regularization parameter scaling stated in Theorem 2.7 are met then clearly our data-driven formulation (22) is both consistent and suffers little disappointment with probability at most $\eta$ asymptotically with increasing number of data points.

# 3 Robust Formulations

As data is invariantly noisy, any data-driven method must be guarded against overfitting the training data while suffering bad out-of-sample performance. Distributionally robust optimization offers a disciplined approach to guard against overfitting effects. Let $\mathcal{M}_N$ be a convex compact set of distributions containing the empirical distribution $M_N$. We denote the formulation

$$\inf_{z_0 \in Z_0,\, z \in \mathcal{F}} \sup_{M \in \mathcal{M}_N} \ell_0(z_0) + \mathbb{E}_{M \otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \tag{22}$$

as the direct robust counterpart of our data-driven formulation (16). We will focus here in particular on ambiguity sets of the form

$$\mathcal{M}_N = \left\{ M \ll M_N \;:\; D_\varphi(M_N, M) := \int \varphi\left(\frac{\mathrm{d}M}{\mathrm{d}M_N}\right)(\xi', \xi)\, \mathrm{d}M_N(\xi', \xi) \leq r \right\} \tag{23}$$

where $\varphi : \mathrm{R} \to \mathrm{R} \cup \{+\infty\}$ is a lower semi-continuous strictly convex function with $\varphi(1) = 0$. Several common ambiguity sets and their corresponding divergence function $\varphi$ are given in Table 2. The condition $\varphi(1) = 0$ ensures that $D_\varphi(M_N, M_N) = 0$, while Jensen's inequality and strict convexity of $\varphi$ gives $\{M \ll M_N \;:\; D_\varphi(M_N, M) \leq 0\} = \{M_N\}$. In particular, $M_N \in \{M \ll M_N \;:\; D_\varphi(M_N, M) \leq r\}$ for all $r \geq 0$. Such divergence ambiguity sets have been studied in the context of distributionally robust optimization by Bayraksan and Love (2015), Duchi et al. (2016) and their statistical properties are well understood (Van Parys et al. 2021, Lam 2019, Duchi et al. 2016). In Section 3.3 we will indicate that such robust decisions enjoy desirable statistical guarantees also in our context of interest when the function $\varphi$ and robustness radius $r$ are judiciously chosen. Crucially, in Section 3.2 we also discuss that such ambiguity sets result in robust formulations which are computationally tractable.

As pointed out in Proposition 2.10, the objective function in the robust formulation (22) is continuous in $M$ and hence its maximum over a compact set $\mathcal{M}_N$ is always attained. One can show, as we do in the proof of Proposition 3.2, that under Assumption 1.2 the objective function is continuous in $z$ when we equip $\mathcal{F}$ with the topology of uniform convergence, i.e., $\lim_{k\to\infty} z_k = z_\infty$ if and only if $\lim_{k\to\infty} \sup_{\xi' \in \mathrm{R}^{d_2}} \|z_k(\xi') - z_\infty(\xi')\| = 0$. Unfortunately, as the set $\mathcal{F}$ is not compact it is unclear whether the infimum in formulation (22) is in fact attained. In Section

| Type | Function $\varphi(t) =$ | Function $\varphi^\star(s) =$ |
|---|---|---|
| Neyman divergence | $(t-1)^2/t$ | $2-2\sqrt{1-s},\ s<1$ |
| Pearson divergence | $(t-1)^2$ | $-1$ if $s<-2$ and $s+\frac{s^2}{4}$ otherwise |
| Hellinger distance | $(\sqrt{t}-1)^2$ | $s/(1-s),\ s<1$ |
| Entropy | $t\log(t)$ | $\exp(s)-1$ |
| Burg Entropy | $-\log(t)$ | $-\log(1-s),\ s<1$ |

Table 2: Several common ambiguity sets and their corresponding $\varphi$-divergence functions and convex conjugates.

3.1 we can however establish the existence of a saddle point in terms of an optimal here-and-now decision $z_0$ and distribution $M$.

## 3.1 Saddle Point Interpretation

Define as in Section 2.3 the function

$$\begin{aligned} F(z_0; M) &= \ell_0(z_0) + \min_{z\in\mathcal{F}} \mathbb{E}_{M\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \\ &= \ell_0(z_0) + \int \left[\min_{z\in Z} \int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2}\, \mathrm{d}M(\xi', \xi)\right] \mathrm{d}m'(\xi'') \end{aligned}$$

where the dependence on the distribution $M$ is now made explicit. The equality is due to Fubini's theorem and Rockafellar and Wets (2009, Theorem 14.60). The function $F$ characterizes the cost of taking a here-and-now decision $z_0$ with an optimal second-stage decision implemented afterwards when $(\tilde{\xi}', \tilde{\xi})$ follows distribution $M$. The considered function $F$ is a bivariate convex-concave function as pointed out by the following lemma.

**Lemma 3.1.** *Let Assumption 1.2 hold, the density function $k$ be $K$-bounded and $\mathcal{K}$-Lipschitz. The function $F(z_0, M)$ is continuous and convex in $z_0 \in Z_0$ for all $M \in \mathcal{M}_N$ and upper semicontinuous and concave in $M \in \mathcal{M}_N$ for all $z_0 \in Z_0$.*

*Proof.* We have that the function $F(z_0, M)$ is convex in $z_0 \in Z$ for every $M \in \mathcal{M}_N$ as $\ell(z_0, z, \xi)$ is jointly convex in $(z_0, z)$ for every $\xi$. Using the same argument made in the proof of Lemma 2.8 the function $F(z_0, M)$ can be shown to have subgradients which are bounded in norm by $\mathcal{L}$. Following Shalev-Shwartz et al. (2012, Lemma 2.6) this implies that $F(z_0, M)$ is $\mathcal{L}$-Lipschitz continuous in $z_0$ for any $M \in \mathcal{M}_N$. We have that $F(z_0, M)$ is upper semicontinuous and concave in $M$ for all $z_0$ as it is the pointwise minimum over all $z \in \mathcal{F}$ of continuous linear functions

$$M \mapsto \int \left[\int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2}\, \mathrm{d}m'(\xi'')\right] \mathrm{d}M(\xi', \xi).$$

Indeed, the integrant $C_h(z_0, z, \xi', \xi) = \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'')$ is shown in the proof of Proposition 2.11 to be $L$-bounded and $\max(\mathcal{L}, 2\mathcal{K}L/h)$-Lipschitz continuous in $(\xi', \xi)$ for each $z_0 \in Z_0$ and $z \in \mathcal{F}$. □

**Proposition 3.2.** *Let Assumption 1.2 hold, the density function $k$ be $K$-bounded and $\mathcal{K}$-Lipschitz and let $\mathcal{M}_N$ be convex and compact. Then, we have*

$$(22) = \min_{z_0 \in Z_0} \max_{M \in \mathcal{M}_N} F(z_0; M) = \max_{M \in \mathcal{M}_N} \min_{z_0 \in Z_0} F(z_0; M).$$

*Let $M^\star$ be any worst-case distribution in* $\arg\max_{M \in \mathcal{M}_N} \min_{z_0 \in Z_0} F(z_0; M)$ *with associated covariate marginal distribution $P'^\star$. We have the decomposition* $(22) = \min_{z_0 \in Z_0} \ell_0(z_0) + \mathbb{E}_{P_h'^\star} \left[ Q(z_0, \tilde{\xi}'; M_h^\star) \right]$.

*Proof.* We have the following chain of equalities

$$\begin{aligned}
&(22) \\
&:= \inf_{z_0 \in Z_0, z \in \mathcal{F}} \ell_0(z_0) + \max_{M \in \mathcal{M}_N} \int \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \mathrm{d}m'(\xi'') \, \mathrm{d}M(\xi', \xi) \\
&= \inf_{z_0 \in Z_0} \ell_0(z_0) + \sup_{M \in \mathcal{M}_N} \inf_{z \in \mathcal{F}} \int \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'') \, \mathrm{d}M(\xi', \xi) \\
&= \inf_{z_0 \in Z_0} \ell_0(z_0) + \sup_{M \in \mathcal{M}_N} \inf_{z \in \mathcal{F}} \int \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi) \, \mathrm{d}m'(\xi'') \\
&= \inf_{z_0 \in Z_0} \ell_0(z_0) + \sup_{M \in \mathcal{M}_N} \int \inf_{z \in Z} \left[ \int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2} \mathrm{d}M(\xi', \xi) \right] \mathrm{d}m'(\xi'') \\
&= \inf_{z_0 \in Z_0} \ell_0(z_0) + \sup_{M \in \mathcal{M}_N} \int \min_{z \in Z} \left[ \int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2} \mathrm{d}M(\xi', \xi) \right] \mathrm{d}m'(\xi'') \\
&=: \inf_{z_0 \in Z_0} \sup_{M \in \mathcal{M}_N} F(z_0, M) = \sup_{M \in \mathcal{M}_N} \inf_{z_0 \in Z_0} F(z_0, M)
\end{aligned}$$

establishing the claim. The second equality follows the minimax theorem of Sion (1958, Corollary 3.3) which as we will now show is applicable. It suffices to show that the objective is continuous, convex in $z$ and concave in $M$. As the integrant $C_h(z_0, z, \xi', \xi) = \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'')$ is shown in the proof of Proposition 2.11 to be $L$-bounded and $\max(\mathcal{L}, 2\mathcal{K}L/h)$-Lipschitz continuous in $(\xi', \xi)$ for each $z_0 \in Z_0$ and $z \in \mathcal{F}$ the objective function is indeed continuous and linear in $M$ for each $z \in \mathcal{F}$. Let us now show the continuity and convexity in $z$. Equip $\mathcal{F}$ with the topology of uniform convergence, i.e., $\lim_{k \to \infty} z_k = z_\infty$ if and only if $\lim_{k \to \infty} \sup_{\xi' \in \Xi' + hK} \|z_k(\xi') - z_\infty(\xi')\| = 0$.

We can show under Assumption 1.2 that, for all sequence $(z_k)_{k\geq 0}$ such that $\lim_{k\to\infty} z_k = z_\infty \in \mathcal{F}$, we have

$$\begin{aligned}
&\lim_{k\to\infty} \left| \int \int (\ell(z_0, z_k(\xi''), \xi) - \ell(z_0, z_\infty(\xi''), \xi)) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'') \mathrm{d}M(\xi', \xi) \right| \\
&\leq \lim_{k\to\infty} \int \int \mathcal{L} \, \|z_k(\xi'') - z_\infty(\xi'')\| \, k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'') \mathrm{d}M(\xi', \xi) \\
&\leq \lim_{k\to\infty} \max_{\xi''\in\Xi'} \|z_k(\xi'') - z_\infty(\xi'')\| \, \mathcal{L} = 0.
\end{aligned}$$

In this strong topology on $\mathcal{F}$ the objective function is hence continuous and convex in $z \in \mathcal{F}$ for all $M \in \mathcal{M}_N$. The third equality follows from Fubini's theorem which applies as

$$\int \int \left| \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \right| \, \mathrm{d}m'(\xi'') \, \mathrm{d}M(\xi', \xi) \leq L.$$

The fourth equality is due to Rockafellar and Wets (2009, Theorem 14.60) and the fact that the function $\ell(z_0, z, \xi)$ is $\mathcal{L}$-Lipschitz continuous jointly in $(z_0, z, \xi)$ hence

$$\int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi)$$

will be $\mathcal{L}K/h^{d_2}$-Lipschitz continuous in $(z_0, z)$ for all $\xi''$ as $k$ is $K$-bounded. Hence, as $Z$ is compact the final equality follows. Finally, because of Lemma 3.1 we can apply Sion's minimax theorem again to obtain $\inf_{z_0\in Z_0} \sup_{M\in\mathcal{M}_N} F(z_0, M) = \sup_{M\in\mathcal{M}_N} \inf_{z_0\in Z_0} F(z_0, M)$. Because of Lemma 3.1 and the fact that $\mathcal{M}_N$ and $Z_0$ are compact these extreme will be attained proving the first part of the claim.

Furthermore, we have

$$\begin{aligned}
&\min_{z_0\in Z_0} \max_{M\in\mathcal{M}_N} F(z_0, M) \\
&= \min_{z_0\in Z_0} \ell_0(z_0) + \\
&\quad \max_{M\in\mathcal{M}_N} \int \min_{z\in Z} \left[ \frac{\int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi)}{\int k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi)} \right] \int \frac{k((\xi'' - \xi')/h)}{h^{d_2}} \, \mathrm{d}M[(\xi', \xi)] \mathrm{d}m'(\xi'') \\
&= \min_{z_0\in Z_0} \ell_0(z_0) + \max_{M\in\mathcal{M}_N} \int Q(z_0, \xi''; M_h) \int k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi) \, \mathrm{d}m'(\xi'') \\
&= \min_{z_0\in Z_0} \ell_0(z_0) + \mathbb{E}_{P_h'^\star} [Q(z_0, \xi'; M_h^\star)]
\end{aligned}$$

In the first equality is established by both dividing and multiplying by $\int k((\xi'' - \xi')/h) \, \mathrm{d}M(\xi', \xi)$. The second

equality uses

$$Q(z_0, \xi''; M_h) := \min_{z \in Z} \left[ \frac{\int \ell(z_0, z, \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi)}{\int k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}M(\xi', \xi)} \right] = \min_{z \in Z} \mathbb{E}_{M_h} \left[ \ell(z_0, z, \tilde{\xi}) | \tilde{\xi}' = \xi'' \right].$$

□

Proposition 3.2 establishes that the robust data-driven formulation (22) reduces to a convex-concave saddle point problem. That is, standard saddle-point theory (Sion 1958) guarantees that we can find $z_0^\star \in Z_0$ and $M^\star \in \mathcal{M}_N$ so that

$$F(z_0^\star; M) \leq F(z_0^\star; M^\star) \leq F(z_0; M^\star) \quad \forall z_0 \in Z_0, \, M \in \mathcal{M}_N.$$

The optimal decision $z_0^\star$ in the robust data-driven formulation (22) can hence be interpreted as the minimum cost decision in face of the worst-case distribution $M^\star$ describing the covariate and uncertain random vectors $\xi'$ and $\xi$, respectively. Similarly, the adversarial distribution $M^\star$ can be interpreted as the worst distribution an adversary can choose given that the optimal decision $z_0^\star$ is implemented.

## 3.2 Computation

As our ambiguity set $\mathcal{M}_N$ defined in Equation (23) only contains distributions absolutely continuous with respect to the empirical distribution we may parameterize $M \in \mathcal{M}_N$ in terms of the mass $\check{M}_d \in \check{\mathcal{M}}_N = \{M' \in \mathrm{R}_+^N \ : \ \sum_{d=1}^N \check{M}_d = 1\}$ it assigns to each observed data point $(\xi_d', \xi_d) \in D_N$. Hence, we define $\check{F}(z_0; \check{M}) := F(z_0; \sum_{d=1}^N \check{M}_d \delta_{(\xi_d', \xi_d)})$ and consider instead the equivalent saddle-point problem

$$\check{F}(z_0^\star; \check{M}) \leq \check{F}(z_0^\star; \check{M}^\star) \leq \check{F}(z_0; \check{M}^\star) \quad \forall z_0 \in Z_0, \, \check{M} \in \check{\mathcal{M}}_N. \tag{24}$$

It is still unclear whether the saddle point problem (24) is computationally tractable as merely evaluating the saddle point function $\check{F}(z_0; \check{M})$ for given decisions $z_0 \in Z_0$ and distribution $\check{M} \in \check{\mathcal{M}}_N$ still requires potentially high-dimensional integration. We will propose here an efficient stochastic saddle point algorithm based on the work of Nemirovski et al. (2009) which circumvents this issue.

**Lemma 3.3.** *Let Assumption 1.2 hold. Further, let us denote with $\partial_{z_0} \check{F}(z_0; \check{M})$ and $\partial_{\check{M}} \check{F}(z_0; \check{M})$ the subdifferential and superdifferential of $\check{F}$ with respect to $z_0$ and $\check{M}$, respectively. We have that*

$$\partial_{z_0} \check{F}(z_0; \check{M}) = \partial \ell_0(z_0) + \mathbb{E}_{P_h'} \left[ \left\{ g \in \mathrm{R}^{n_0} \ : \ (g, 0) \in \mathbb{E}_{M_h} \left[ \partial_{z_0, z} \ell(z_0, z_N^\star(z_0, \tilde{\xi}', \check{M}), \tilde{\xi}) | \tilde{\xi}' \right] \right\} \right] \subseteq \mathrm{R}^{n_0},$$

$$\partial_{\check{M}} \check{F}(z_0; \check{M}) \ni \mathbb{E}_{E^N} \left[ \left( \ell(z_0, z_N^\star(z_0, \xi_1' + h\tilde{e}_1, \check{M}), \xi_1), \ldots, \ell(z_0, z_N^\star(z_0, \xi_N' + h\tilde{e}_N, \check{M}), \xi_N) \right) \right] \in \mathrm{R}^N$$

*where* $M_h = \sum_{d=1}^N \check{M}_d(\delta_{\xi'_d,h} \otimes \delta_{\xi_d})$, $P'_h = \sum_{d=1}^N \check{M}_d \delta_{\xi'_d,h}$ *and*

$$z_N^\star(z_0, \xi', \check{M}) \in \arg\min_{z \in Z} \frac{\sum_{d=1}^N \ell(z_0, z, \xi_d) k((\xi' - \xi'_d)/h) \check{M}_d}{\sum_{d=1}^N k((\xi' - \xi'_d)/h) \check{M}_d}$$

*for all* $z_0 \in Z_0$, $\xi' \in \Xi' + hK$ *and* $\check{M} \in \check{\mathcal{M}}_N$.

*Proof.* The equality

$$\partial_{z_0} \check{F}(z_0; \check{M}) = \partial \ell_0(z_0) + \mathbb{E}_{P'_h}\left[\left\{ g \in \mathrm{R}^{n_0} \ : \ (g, 0) \in \mathbb{E}_{M_h}\left[\partial_{z_0, z} \ell(z_0, z_N^\star(z_0, \tilde{\xi}', \check{M}), \tilde{\xi}) | \tilde{\xi}'\right]\right\}\right] \subseteq \mathrm{R}^{n_0}$$

can be proven along the exact same line as the proof of Lemma 2.8. We omit the proof here for the sake of brevity. We prove here

$$\partial_{\check{M}} \check{F}(z_0; \check{M}) \ni \left(\mathbb{E}_{\delta_{\xi'_1,h}}\left[\ell(z_0, z_N^\star(\tilde{\xi}'^1), \xi_1)\right], \ldots, \mathbb{E}_{\delta_{\xi'_N,h}}\left[\ell(z_0, z_N^\star(\tilde{\xi}'^N), \xi_N)\right]\right) \in \mathrm{R}^N.$$

Recall that when $M \in \mathcal{M}_N$ the distribution $M$ is supported on $D_N$. Hence,

$$\check{F}(z_0; \check{M}) = \ell_0(z_0) + \int [\min_{z \in Z} \sum_{d=1}^N \ell(z_0, z, \xi_d) k((\xi'' - \xi'_d)/h)/h^{d_2} \check{M}_d] \mathrm{d}m'(\xi'').$$

Furthermore, we have that

$$\begin{aligned}
&\partial_{\check{M}} \tilde{F}(z_0; \check{M}) \\
&= \partial_{\check{M}} \int \left[\min_{z \in Z} \sum_{d=1}^N \ell(z_0, z, \xi_d) k((\xi'' - \xi'_d)/h)/h^{d_2} \check{M}_d\right] \mathrm{d}m'(\xi'') \\
&= \int \partial_{\check{M}} \left[\min_{z \in Z} \sum_{d=1}^N \ell(z_0, z, \xi_d) k((\xi'' - \xi'_d)/h)/h^{d_2} \check{M}_d\right] \mathrm{d}m'(\xi'') \\
&\supseteq \int \partial_{\check{M}} \left[\sum_{d=1}^N \ell(z_0, z_N^\star(z_0, \xi'', \check{M}), \xi_d) k((\xi'' - \xi'_d)/h)/h^{d_2} \check{M}_d\right] \mathrm{d}m'(\xi'') \\
&\ni \int \left(\ell(z_0, z_N^\star(z_0, \xi'', \check{M}), \xi_1) k((\xi'' - \xi'_1)/h)/h^{d_2}, \ldots, \ell(z_0, z_N^\star(z_0, \xi'', \check{M}), \xi_N) k((\xi'' - \xi'_N)/h)/h^{d_2}\right) \mathrm{d}m'(\xi'') \\
&\ni \left(\mathbb{E}_{\delta_{\xi'_1,h}}\left[\ell(z_0, z_N^\star(z_0, \tilde{\xi}'^1, \check{M}), \xi_1)\right], \ldots, \mathbb{E}_{\delta_{\xi'_N,h}}\left[\ell(z_0, z_N^\star(z_0, \tilde{\xi}'^N, \check{M}), \xi_N)\right]\right) \\
&\ni \mathbb{E}_{E^N}\left[\left(\ell(z_0, z_N^\star(z_0, \xi'_1 + h\tilde{e}_1, \check{M}), \xi_1), \ldots, \ell(z_0, z_N^\star(z_0, \xi'_N + h\tilde{e}_N, \check{M}), \xi_N)\right)\right] \in \mathrm{R}^N.
\end{aligned}$$

Here, the second equality follows from the fact that the integrant $\min_{z \in Z} \sum_{d=1}^N \ell(z_0, z, \xi_d) k((\xi'' - \xi'_d)/h)/h^{d_2} \check{M}_d$ remains bounded as the loss function $\ell$ is $L$ bounded and the kernel function $k$ is continuous and has compact support. The inclusion is a consequence of Lemma C.2. □

Assume again for the sake of simplicity that the loss functions $\ell_0$ and $\ell(z_0, z, \xi)$ are differentiable in $z_0$ and $(z_0, z)$, respectively, for every $\xi \in \Xi$. Akin to the discussion in Section 2.3 the stochastic gradient

$$\check{G}_{z_0}(z_0, \xi', \check{M}) := \ell_0'(z_0) + (\sum_{d=1}^N \ell'(z_0, z_N^\star(z_0, \xi', \check{M}), \xi_d) k((\xi' - \xi_d')/h)\check{M}_d)/(\sum_{d=1}^N k((\xi' - \xi_d')/h)\check{M}_d)$$

is following Lemma 3.3 an unbaised estimator of the subgradient of $\check{F}(z_0; \check{M})$ with respect to $z_0$ when $\xi'$ is sampled from $P_h' = \sum_{d=1}^N \check{M}_d \delta_{\xi_d', h}$. We remark that sampling from $P_h'$ can be done efficiently by first drawing a sample from the weighted empirical distribution distribution $\sum_{d=1}^N \check{M}_d \delta_{\xi_d'}$ and then corrupting the so obtained sample with a noise realization sampled independently from the artificial noise distribution $E$.

Likewise, the stochastic gradient

$$\check{G}_{\check{M}}(z_0, e_1, \ldots, e_N, \check{M}) = (\ell(z_0, z^\star(z_0, \xi_1' + he_1, \check{M}), \xi_1), \ldots, \ell(z_0, z^\star(z_0, \xi_N' + he_N, \check{M}), \xi_N)).$$

is following Lemma 3.3 an unbaised estimator of the gradient of $\check{F}(z_0; \check{M})$ with respect to $\check{M}$. We finally point out that both our stochastic subgradient and supergradient enjoy a finite variance.

**Lemma 3.4** (Stochastic Subgradient/Supergradient)**.** *Let Assumption 1.2 hold. We have for all $z_0 \in Z_0$, $\check{M} \in \check{\mathcal{M}}_N$ that the stochastic subgradient satisfies $\mathbb{E}_{P_h'}\left[||\check{G}_{z_0}(z_0, \xi', \check{M})||^2\right] \leq (\mathcal{L}_0 + \mathcal{L})^2$ where $P_h' = \sum_{d=1}^N \check{M}_d \delta_{\xi_d', h}$ whereas the stochastic supergradient satisfies $\mathbb{E}_{E^N}\left[||\check{G}_{\check{M}}(z_0, \xi_1' + h\tilde{e}_1, \ldots, \xi_N' + h\tilde{e}_N, \check{M})||^2\right] \leq NL^2$ for all $d \in [1, \ldots, N]$.*

*Proof.* The bound

$$\mathbb{E}_{P_h'}\left[||\check{G}_{z_0}(z_0, \xi', \check{M})||^2\right] \leq (\mathcal{L}_0 + \mathcal{L})^2$$

can be proven along the exact same line as the proof of Lemma 2.9. We omit the proof here for the sake of brevity. We have that

$$\begin{aligned}
\mathbb{E}_{E^N}\left[||\check{G}_{\check{M}}(z_0, \xi_1' + h\tilde{e}_1, \ldots, \xi_N' + h\tilde{e}_N, \check{M})||^2\right] =& \mathbb{E}_{E^N}\left[\sum_{d=1}^N \ell(z_0, z^\star(z_0, \xi_d' + h\tilde{e}_d, \check{M}), \xi_d)^2\right] \\
=& \sum_{d=1}^N \mathbb{E}_E\left[\ell(z_0, z^\star(z_0, \xi_d' + h\tilde{e}, \check{M}), \xi_d)^2\right] \\
\leq& NL^2.
\end{aligned}$$

where the ultimate inequality follows from the fact that following Assumption 1.2 the loss function $\ell$ is $L$-bounded.

□

As the stochastic gradient $\check{G}_{z_0}$ and $\check{G}_{\check{M}}$ are both unbiased and have bounded variance we may solve the saddle

problem of interest $\min_{z_0 \in Z_0} \max_{\check{M} \in \check{\mathcal{M}}_N} \check{F}(z_0; \check{M})$ using a stochastic gradient descent ascent approach. To account for the primal constraints $z_0 \in Z_0$ on the first stage decision and the dual constraints $\check{M} \in \check{\mathcal{M}}_N$ on the distribution, we project each iterate back into the feasible region using the projection maps $P_{z_0}(g) = \arg\min_{z_0' \in Z_0} \|z_0 - g - z_0'\|_2^2$ and $P_{\check{M}}(\check{G}) = \arg\min_{\check{M}' \in \check{\mathcal{M}}_N} \|\check{M} - \check{G} - \check{M}'\|_2^2$. The resulting algorithm is depicted in Algorithm 2.

Notice that also this algorithm can be easily generalized to the mirror descent ascent algorithm of Nemirovski et al. (2009). Let indeed $\nu$ again be a continuously differentiable and strongly convex distance generating function on $\mathrm{R}^{n_0}$ with associated Bregman divergence $V(z_0, z_0') = \nu(z_0') - [\nu(z_0) - \langle \nu'(z_0), z_0' - z_0 \rangle]$ and proximal mapping

$$P_{z_0} : g \mapsto \arg\min_{z_0' \in Z_0} \langle g, z_0' - z_0 \rangle + V(z_0, z_0').$$

Let similarly $\omega$ be a continuously differentiable and strongly convex distance generating function on $\check{\mathcal{M}}_N$ with associated Bregman divergence $W(\check{M}, \check{M}') = \omega(\check{M}') - [\omega(\check{M}) - \langle \omega'(\check{M}), \check{M}' - \check{M} \rangle]$ and proximal mapping

$$P_{\check{M}} : \check{G} \mapsto \arg\min_{\check{M}' \in \check{\mathcal{M}}_N} \langle \check{G}, \check{M}' - \check{M} \rangle + W(\check{M}, \check{M}').$$

We assume here without loss of generality that $\nu$ and $\mu$ are scaled appropriately so that their strong convexity parameters are one. Let $D_{z_0}^2 := \max_{z_0' \in Z_0} \nu(z_0') - \min_{z_0' \in Z_0} \nu(z_0')$ denote the size of the set $Z_0$ as measured by the distance generating function $\nu$ considered. Likewise, let $D_{\check{M}}^2 := \max_{\check{M}' \in \check{\mathcal{M}}_N} \omega(\check{M}') - \min_{\check{M}' \in \check{\mathcal{M}}_N} \omega(\check{M}')$ denote the size of the standard simplex $\check{\mathcal{M}}_N$ as measured by the distance generating function $\omega$ considered. Nemirovski et al. (2009) suggest producing iterates as suggested in Algorithm 2.

The suboptimality of an iterate may be measured by $\epsilon(z_0, \check{M}) = \sup_{\check{M}' \in \check{\mathcal{M}}_N} F(z_0; \check{M}') - \min_{z_0' \in Z_0} F(z_0; \check{M})$. Nemirovski et al. (2009) show that the weighted averages $z_0^F = \sum_{j=1}^T z_0^j \gamma_j / (\sum_{j=1}^T \gamma_j)$ and $\check{M}^F = \sum_{j=1}^T \check{M}^j \gamma_j / (\sum_{j=1}^T \gamma_j)$ of the the produced iterate produced satisfy $\mathbb{E}\left[\epsilon(z_0^F, \check{M}^F)\right] \leq (2 + 5D_\star^2 \sum_{j=1}^T \gamma_j^2) / \sum_{j=1}^T \gamma_j$ where $D_\star^2 = D_{z_0}^2(\mathcal{L}_0 + \mathcal{L})^2 + D_{\check{M}}^2 N L^2$. In particular, for $\gamma_j = 2/\sqrt{D_\star 5T}$ it follows that $\mathbb{E}\left[\epsilon(z_0^F, \check{M}^F)\right] \leq 2D_\star \sqrt{5/T}$.

### 3.3 Out-of-Sample Performance

We consider the well studied case of the Neyman divergence stated in Table 2 and its associated ambiguity set $\mathcal{M}_N = \left\{ M' \in \mathcal{P}(\Xi' \times \Xi) \ : \ M' \ll M_N, \ \chi^2(M_N, M') \leq r(N) \right\}$. Consider first the robust cost function of the minimization problem stated in Equation (22) for given first-stage and second-stage decisions $z_0 \in Z_0$ and $z \in \mathcal{F}$. We first show that this robust cost function $M \mapsto \max_{M' \in \mathcal{M}_N} \mathbb{E}_M \left[\ell(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]$ is always continuous.

**Algorithm 2:** Stochastic Mirror Gradient Descent for Saddle Point Problems

**Initialization:** Starting points $z_0^1 \in \arg\min_{z_0' \in Z_0} \nu(z_0')$, $\check{M}^1 \in \arg\min_{\check{M}' \in \check{\mathcal{M}}_N} \omega(\check{M}')$, regularization parameter $h$, iteration number $T \geq 2$ and step lengths $\gamma_j > 0$ for $j \in [1, T-1]$.

**for** $j \in [1, \ldots, T-1]$ **do**

//Primal $z_0$ Update

Sample $\bar{\xi}'$ randomly from $\sum_{d=1}^N \check{M}_d \delta_{\xi_d'}$ and $e$ following the perturbation distribution $E$

$\xi' = \bar{\xi}' + he_0$

$z^j \in \arg\min_{z \in Z} \sum_{d=1}^N \ell(z_0^j, z, \xi_d) k((\xi' - \xi_d')/h) \check{M}_d^j / \sum_{d=1}^N k((\xi' - \xi_d')/h) \check{M}_d^j$

$\check{G}_{z_0}^j = \ell_0'(z_0^j) + (\sum_{d=1}^N \ell'(z_0^j, z^j, \xi_d) k((\xi' - \xi_d')/h) \check{M}_d^j) / (\sum_{d=1}^N k((\xi' - \xi_d')/h) \check{M}_d^j)$

$z_0^{j+1} = P_{z_0^j}(\gamma_j \check{G}_{z_0}^j)$

//Dual $\check{M}$ Update

**for** $d \in [1, \ldots, N]$ **do**

Sample $e_d$ following the perturbation distribution $E$

$z_d^j \in \arg\min_{z \in Z} \sum_{d'=1}^N \ell(z_0^j, z, \xi_{d'}) k((\xi_d' + he_d - \xi_{d'}')/h) \check{M}_{d'}^j / \sum_{d'=1}^N k((\xi_d' + he_d - \xi_{d'}')/h) \check{M}_{d'}^j$

$\check{G}_{\check{M}}^j = (\ell(z_0^j, z_1^j, \xi_1), \ldots, \ell(z_0^j, z_d^j, \xi_N))$

$\check{M}^{j+1} = P_{\check{M}^j}(-\gamma_j \check{G}_{\check{M}}^j)$

**return** $z_0^F = \sum_{j=1}^T z_0^j \gamma_j / (\sum_{j=1}^T \gamma_j)$ and $\check{M}^F = \sum_{j=1}^T \check{M}^j \gamma_j / (\sum_{j=1}^T \gamma_j)$.

**Proposition 3.5** (Regularity)**.** *Let Assumption 1.2 hold and $r > 0$ and $\delta > 0$. We have*

$$\left\{ \begin{array}{ll} \sup & \mathbb{E}_{M''}\left[C_h(z_0, z, \xi', \xi)\right] \\ \text{s.t.} & M' \in \mathcal{P}_1(\Xi' \times \Xi),\ M'' \in \mathcal{P}(\Xi' \times \Xi), \\ & ||M - M'|| \leq \delta,\ \chi^2(M', M'') \leq r, \end{array} \right.$$

$$\leq$$

$$\left\{ \begin{array}{ll} \sup & \mathbb{E}_{M''}\left[C_h(z_0, z, \xi', \xi)\right] + \frac{\max(h\mathcal{L}, 2\mathcal{K}L)(\sqrt{L}+\frac{1}{h})}{2} \left(\delta + \frac{\operatorname{diam}(\Xi' \times \Xi) r}{2}\right) + \left(\frac{L}{2}\frac{r+1}{\sqrt{r}} + L + \frac{1}{h^2}\right)\frac{r}{2} \\ \text{s.t.} & M'' \in \mathcal{P}(\Xi' \times \Xi),\ M \ll M'',\ M'' \ll M, \\ & \chi^2(M, M'') \leq r \end{array} \right.$$

*for all $z_0 \in Z$, $z \in \mathcal{F}$ and $M \in \mathcal{P}(\Xi' \times \Xi)$ with $\mathcal{K}$ the Lipschitz constant of the density function $k$.*

*Proof.* Let here $\hat{C}_h(z_0, z) := \operatorname{ess\,sup}_M C_h(z_0, z, \tilde{\xi}', \tilde{\xi})$ and $\bar{C}_h(z_0, z) := \max_{\xi \in \Xi', \xi \in \Xi} C_h(z_0, z, \xi', \xi)$. We have for any $\epsilon > 0$ the following inequalities

$$\sup\left\{\sup\left\{\mathbb{E}_{M''}\left[C_h(z_0, z, \xi', \xi)\right] \ :\ \chi^2(M', M'') \leq r\right\} \ :\ M' \in \mathcal{P}_1(\Xi' \times \Xi),\ \|M - M'\| \leq \delta\right\}$$

$$
\begin{aligned}
&\leq \sup\left\{\sup\left\{\mathbb{E}_{M''}\left[C_h(z_0,z,\xi',\xi)\right] \;:\; \chi^2(M',M'')\leq r\right\} \;:\; M'\in\mathcal{P}_1(\Xi'\times\Xi),\;\; \|M-M'\|\leq\delta,\;\; \sqrt{r+1}\geq\textstyle\int \mathrm{d}\,|M'|\,(\xi',\xi)\right\}\\
&\leq \sup_{\|M-M'\|\leq\delta,\,\sqrt{r+1}\geq\int \mathrm{d}|M'|(\xi',\xi)} \inf\left\{\alpha-\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}\,|M'|\,(\xi',\xi)\right)^2/(r+1) \;:\; \alpha\geq\bar{C}_h(z_0,z)+\epsilon\right\}\\
&\leq \inf\left\{\alpha-\left(\inf_{\|M-M'\|\leq\delta,\,\sqrt{r+1}\geq\int \mathrm{d}|M'|(\xi',\xi)}\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}\,|M'|\,(\xi',\xi)\right)^2/(r+1) \;:\; \alpha\geq\bar{C}_h(z_0,z)+\epsilon\right\}\\
&\leq \inf\left\{\alpha-\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)-H(\alpha,h,r,\delta)\right)^2/(r+1) \;:\; \alpha\geq\bar{C}_h(z_0,z)+\epsilon\right\}\\
&\leq \inf\left\{\alpha-\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)\right)^2/(r+1)+2H(\alpha,h,r,\delta)\sqrt{\alpha}/(r+1) \;:\; \alpha\geq\bar{C}_h(z_0,z)+\epsilon\right\}\\
&\leq \min\left\{\alpha-\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)\right)^2/(r+1) \;:\; \alpha\geq\bar{C}_h(z_0,z)+\epsilon\right\}\\
&\qquad\qquad +\sup\left\{2H(\alpha,h,r,\delta)\sqrt{\alpha}/(r+1) \;:\; \bar{C}_h(z_0,z)+\epsilon\leq\alpha\leq\epsilon+\bar{C}_h(z_0,z)+L(r+1)/(2\sqrt{r})\right\}\\
&\leq \sup_{\chi^2(M,M')\leq r(N),\,M'\ll M,\,M\ll M'} \mathbb{E}_{M'}\left[C_h(z_0,z,\tilde{\xi}',\tilde{\xi})\right]+(\epsilon+L)\frac{r}{r+1}\\
&\qquad\qquad +\sup\left\{2H(\alpha,h,r,\delta)\sqrt{\alpha}/(r+1) \;:\; \bar{C}_h(z_0,z)+\epsilon\leq\alpha\leq\epsilon+\bar{C}_h(z_0,z)+L(r+1)/(2\sqrt{r})\right\}
\end{aligned}
$$

Here the first inequality follows from Lemma B.2. The second inequality follows from the duality result stated in Lemma B.6 with $\epsilon'=\bar{C}_h(z_0,z)-\hat{C}_h(z_0,z)+\epsilon$ and Lemma B.1. The third inequality is a trivial minimax inequality combined with the fact that the square function is monotonically increasing. The fourth inequality follows from Lemma B.8 where we take

$$
H(\alpha,h,r,\delta):=\frac{\max(\mathcal{L},2\mathcal{K}L/h)\delta}{2\sqrt{\alpha-\bar{C}_h(z_0,z)}}+\left(\frac{\max(\mathcal{L},2\mathcal{K}L/h)\,\mathrm{diam}(\Xi'\times\Xi)}{2\sqrt{\alpha-\bar{C}_h(z_0,z)}}+\sqrt{\alpha}\right)(\sqrt{r+1}-1).
$$

The fifth inequality follows by remarking that

$$
\begin{aligned}
&\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)-H(\alpha,h,r,\delta)\right)^2\\
&=\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)\right)^2-2H(\alpha,h,r,\delta)\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)+H(\alpha,h,r,\delta)^2\\
&\geq\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)\right)^2-2H(\alpha,h,r,\delta)\sqrt{\alpha}
\end{aligned}
$$

where the last inequality uses that $M\in\mathcal{P}(\Xi'\times\Xi)$. Let

$$
\alpha^\star\in\arg\min\left\{\alpha-\left(\int\sqrt{\alpha-C_h(z_0,z,\xi',\xi)}\,\mathrm{d}M(\xi',\xi)\right)^2/(r+1) \;:\; \alpha\geq\bar{C}_h(z_0,z)+\epsilon\right\}
$$

which exists following Lemma B.7 as $r > 0$. Hence,

$$
\begin{aligned}
&\inf\left\{\alpha - \left(\int \sqrt{\alpha - C_h(z_0, z, \xi', \xi)}\,\mathrm{d}M(\xi', \xi)\right)^2/(r+1) + 2H(\alpha, h, r, \delta)\sqrt{\alpha}/(r+1) \;:\; \alpha \geq \bar{C}_h(z_0, z) + \epsilon\right\} \\
&\leq \alpha^\star - \left(\int \sqrt{\alpha^\star - C_h(z_0, z, \xi', \xi)}\,\mathrm{d}M(\xi', \xi)\right)^2/(r+1) + 2H(\alpha^\star, h, r, \delta)\sqrt{\alpha^\star}/(r+1) \\
&= \min\left\{\alpha - \left(\int \sqrt{\alpha - C_h(z_0, z, \xi', \xi)}\,\mathrm{d}M(\xi', \xi)\right)^2/(r+1) \;:\; \alpha \geq \bar{C}_h(z_0, z) + \epsilon\right\} + 2H(\alpha^\star, h, r, \delta)\sqrt{\alpha^\star}/(r+1) \\
&= \min\left\{\alpha - \left(\int \sqrt{\alpha - C_h(z_0, z, \xi', \xi)}\,\mathrm{d}M(\xi', \xi)\right)^2/(r+1) \;:\; \alpha \geq \bar{C}_h(z_0, z) + \epsilon\right\} + \\
&\qquad\qquad \max\left\{2H(\alpha, h, r, \delta)\sqrt{\alpha}/(r+1) \;:\; \bar{C}_h(z_0, z) + \epsilon \leq \alpha \leq \frac{L}{2}\frac{r+1}{\sqrt{r}} + \bar{C}_h(z_0, z) + \epsilon\right\}
\end{aligned}
$$

where the last inequality follows again from Lemma B.7 with $\epsilon' = \bar{C}_h(z_0, z) - \hat{C}_h(z_0, z) + \epsilon$. The final inequality follows the observation that

$$
\begin{aligned}
&\min\left\{\alpha - \left(\int \sqrt{\alpha - C_h(z_0, z, \xi', \xi)}\,\mathrm{d}M(\xi', \xi)\right)^2/(r+1) \;:\; \alpha \geq \bar{C}_h(z_0, z) + \epsilon\right\} \\
&= \left\{\begin{array}{l}
\sup \int C_h(z_0, z, \xi', \xi)\,\mathrm{d}M'(\xi', \xi) + m'(\hat{C}_h(z_0, z) + \epsilon') \\
\text{s.t. } m' \in \mathrm{R}_+,\ M' \in \mathcal{P}_+(U),\ M \ll M' \ll M, \\
\qquad \int \mathrm{d}M'(\xi', \xi) + m' = 1, \\
\qquad \int \frac{\mathrm{d}M}{\mathrm{d}Q_c}(\xi', \xi) - 1\,\mathrm{d}M(\xi', \xi) \leq r
\end{array}\right. \\
&\leq \left\{\begin{array}{l}
\sup \int C_h(z_0, z, \xi', \xi)\,\mathrm{d}M'(\xi', \xi) + (\bar{C}_h(z_0, z) + \epsilon)r/(r+1) \\
\text{s.t. } M' \in \mathcal{P}_+(U),\ M \ll M' \ll M, \\
\qquad \int \frac{\mathrm{d}M}{\mathrm{d}Q_c}(\xi', \xi) - 1\,\mathrm{d}M(\xi', \xi) \leq r.
\end{array}\right.
\end{aligned}
$$

Here, the first equation is due to Lemma B.6 with $\epsilon' = \bar{C}_h(z_0, z) - \hat{C}_h(z_0, z) + \epsilon \leq L + \epsilon$.

We finally remark that

$$
\begin{aligned}
&\max\left\{2H(\alpha,h,r,\delta)\sqrt{\alpha}/(r+1)\ :\ \bar{C}_h(z_0,z)+\epsilon\leq\alpha\leq\frac{L}{2}\frac{r+1}{\sqrt{r}}+\bar{C}_h(z_0,z)+\epsilon\right\}\\
\leq&\max\left\{\frac{\max(\mathcal{L},2\mathcal{K}L/h)\delta}{2\sqrt{\alpha-\bar{C}_h(z_0,z)}}\sqrt{\alpha}/(r+1)\ :\ \bar{C}_h(z_0,z)+\epsilon\leq\alpha\right\}\\
&\quad+\max\left\{\frac{\max(\mathcal{L},2\mathcal{K}L/h)\operatorname{diam}(\Xi'\times\Xi)}{2\sqrt{\alpha-\bar{C}_h(z_0,z)}}(\sqrt{r+1}-1)\sqrt{\alpha}/(r+1)\ :\ \bar{C}_h(z_0,z)+\epsilon\leq\alpha\right\}\\
&\quad+\max\left\{\alpha(\sqrt{r+1}-1)/(r+1)\ :\ \alpha\leq\frac{L}{2}\frac{r+1}{\sqrt{r}}+\bar{C}_h(z_0,z)+\epsilon\right\}\\
\leq&\frac{\max(\mathcal{L},2\mathcal{K}L/h)\sqrt{L+\epsilon}}{2\sqrt{\epsilon}(r+1)}(\delta+\operatorname{diam}(\Xi'\times\Xi)(\sqrt{r+1}-1))+\left(\frac{L}{2}\frac{r+1}{\sqrt{r}}+L+\epsilon\right)\frac{\sqrt{r+1}-1}{r+1}\\
\leq&\frac{\max(\mathcal{L},2\mathcal{K}L/h)\sqrt{L+\epsilon}}{2\sqrt{\epsilon}}\left(\delta+\frac{\operatorname{diam}(\Xi'\times\Xi)r}{2}\right)+\left(\frac{L}{2}\frac{r+1}{\sqrt{r}}+L+\epsilon\right)\frac{r}{2}\\
\leq&\frac{\max(h\mathcal{L},2\mathcal{K}L)(\sqrt{L}+1/h)}{2}\left(\delta+\frac{\operatorname{diam}(\Xi'\times\Xi)r}{2}\right)+\left(\frac{L}{2}\frac{r+1}{\sqrt{r}}+L+1/h^2\right)\frac{r}{2}.
\end{aligned}
$$

Here, the first inequality follows from the fact that $\alpha\to\sqrt{\alpha}/(\alpha-\bar{C}(z_0,z))$ is nonincreasing in $\alpha\geq\bar{C}(z_0,z)$ and $\alpha\to\alpha(\sqrt{r+1}-1)/(r+1)$ is nondecreasing in $\alpha$. The second inequality follows from $\bar{C}(z_0,z)\leq L$. The third inequality uses that $\sqrt{r+1}-1\geq r/2$ and $r>0$. The final inequality follows by considering the particular choice $\epsilon=h^{-2}>0$ and uses $\sqrt{L+\frac{1}{h^2}}\leq\sqrt{L}+\frac{1}{h}$. □

We again remark that as $h>0$ tends to zero the robust cost function becomes less regular and ultimately for $h=0$ the result in Proposition 3.5 becomes moot. Nevertheless, we show that when the regularization parameter $h(N)$ tends to zero sufficiently slow the added beneficial regularization is enough to provide strong asymptotic out-of-sample guarantees.

**Theorem 3.6.** *Let $\delta>0$ and $r>0$ be arbitrary. Consider a sequence $\lambda(N)$ so that $\lim_{N\to\infty}\lambda(N)=\infty$ and $\lim_{N\to\infty}\lambda(N)/\sqrt{N}=0$. Let $\mathcal{M}_N=\left\{M\in\mathcal{P}(\Xi'\times\Xi)\ :\ M_N\ll M\ll M_N,\ \chi^2(M_N,M)\leq\lambda(N)^2r/N\right\}$ and denote with $(z^\star_{0N,r},z^\star_{N,r})$ and $v^\star_{N,r}$ the minimizers and minimum of*

$$
\min_{z_0\in Z_0,\,z\in\mathcal{F}}\ell_0(z_0)+\max_{M'\in\mathcal{M}_N}\mathbb{E}_{M'\otimes E}\left[\ell(z_0,z(\tilde{\xi}'+h(N)\tilde{e}),\tilde{\xi})\right],
$$

*respectively. Define here the constant*

$$
\begin{aligned}
d(N,r,\delta)=&\frac{\max(h(N)\mathcal{L},2\mathcal{K}L)(\sqrt{L}+\frac{1}{h(N)})}{2}\left(\frac{\lambda(N)\delta}{\sqrt{N}}+\frac{\operatorname{diam}(\Xi'\times\Xi)\lambda^2(N)}{2N}\right)\\
&+\frac{L}{4}\left(\frac{\lambda^2(N)r}{N}+1\right)\frac{\lambda(N)\sqrt{r}}{\sqrt{N}}+\left(L+\frac{1}{h(N)^2}\right)\frac{\lambda^2(N)r}{2N}.
\end{aligned}
$$

*Then,*

$$\limsup_{N\to\infty} \frac{1}{\lambda(N)^2} \log M^\infty \left[ \ell_0(z^\star_{0N,r}) + \mathbb{E}_{M\otimes E}\left[ \ell(z^\star_{0N,r}, z^\star_{N,r}(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi}) \right] > v^\star_{N,r} + d(N,r,\delta) \right] \leq -r.$$

*Assume furthermore that the conditions of Theorem 2.7 hold and that* $\lim_{N\to\infty} \lambda(N)/(\sqrt{N}h(N)) = 0$. *Then, we furthermore have* $\limsup_{N\to\infty} v^\star_{N,r} + d(N,r,\delta) \leq v^\star$.

*Proof.* Consider the ambiguity sets $\mathcal{M}_N^\delta \subseteq \mathcal{P}(\Xi' \times \Xi)$ defined here as

$$\left\{ M'' \ : \ M' \in \mathcal{P}_1(\Xi' \times \Xi), \ \|M_N - M'\| \leq \lambda(N)\delta/\sqrt{N}, \ \int \left( \frac{\mathrm{d}(M' - M'')}{\mathrm{d}M}(\xi, \xi') \right)^2 \mathrm{d}M(\xi', \xi) \leq \frac{\lambda(N)^2 r}{N} \right\}.$$

We have

$$\begin{aligned}
& M \notin \mathcal{M}_N^\delta \\
\implies & \inf \left\{ \|M_N - M'\| \ : \ M' - M \in \mathcal{P}_0(\Xi' \times \Xi), \ \int \left( \frac{\mathrm{d}(M' - M)}{\mathrm{d}M}(\xi, \xi') \right)^2 \mathrm{d}M(\xi', \xi) \leq \frac{\lambda(N)^2 r}{N} \right\} \geq \frac{\lambda(N)\delta}{\sqrt{N}} \\
\iff & \inf \left\{ \|M_N - M + M - M'\| \ : \ M' - M \in \mathcal{P}_0(\Xi' \times \Xi), \int \left( \frac{\mathrm{d}(M' - M)}{\mathrm{d}M}(\xi, \xi') \right)^2 \mathrm{d}M(\xi', \xi) \leq \frac{\lambda(N)^2 r}{N} \right\} \\
& \quad \geq \lambda(N)\delta/\sqrt{N} \\
\iff & \inf \left\{ \left\| \frac{\sqrt{N}}{\lambda(N)}(M_N - M - \nu) \right\| \ : \ \nu \in \mathcal{P}_0(\Xi' \times \Xi), \int \left( \frac{\sqrt{N}}{\lambda(N)} \frac{\mathrm{d}\nu}{\mathrm{d}M}(\xi, \xi') \right)^2 \mathrm{d}M(\xi', \xi) \leq r \right\} \geq \delta \\
\iff & \inf \left\{ \left\| \frac{\sqrt{N}}{\lambda(N)}(M_N - M) - \nu' \right\| \ : \ \nu' \in \mathcal{P}_0(\Xi' \times \Xi), \ \int \left( \frac{\mathrm{d}\nu'}{\mathrm{d}M}(\xi, \xi') \right)^2 \mathrm{d}M(\xi', \xi) \leq r \right\} \geq \delta \\
\implies & \frac{\sqrt{N}}{\lambda(N)}(M_N - M) \notin A := \left\{ \nu' \in \mathcal{P}_0(\Xi' \times \Xi) \ : \ \int \left( \frac{\mathrm{d}\nu'}{\mathrm{d}M}(\xi, \xi') \right)^2 \mathrm{d}M(\xi', \xi) \leq r \right\}^{\delta/2}
\end{aligned}$$

Here the first equivalence follows from the fact that if it where false then we can find by the definition of infimum an $M'$ so that $M' \in \mathcal{P}_0(\Xi' \times \Xi) + M = \mathcal{P}_1(\Xi' \times \Xi)$, $\|M_N - M\| < \lambda(N)\delta/\sqrt{N}$ and $\int (\mathrm{d}(M' - M)/\mathrm{d}M(\xi, \xi'))^2 \, \mathrm{d}M(\xi', \xi) \leq \lambda(N)^2 r/N$ contradicting the fact that $M \notin \mathcal{M}_N^\delta$. The second equivalence follows by adding and subtracting $M$ in the objective. The third equivalence follows by multiplying the constraint with $\sqrt{N}/\lambda(N)$ and the variable substitution $\nu = M' - M$. The fourth equivalence follows by the variable substitution $\nu' = \sqrt{N}/\lambda(N)\nu$ and the observation that $\sqrt{N}/\lambda(N)\mathcal{P}_0(\Xi' \times \Xi) = \mathcal{P}_0(\Xi' \times \Xi)$.

A moderate deviation result by Wu (1994) establishes that we have

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{\lambda(N)^2} \log M^\infty \left[M \notin \mathcal{M}_N^\delta\right] \\
=&\limsup_{N\to\infty} \frac{1}{\lambda(N)^2} \log M^\infty \left[\frac{\sqrt{N}}{\lambda(N)}(M_N - M) \notin A\right] \\
\leq& -\inf_{\nu'\in \mathrm{cl}(A^c)} \int \left(\frac{\mathrm{d}\nu'}{\mathrm{d}M}(\xi,\xi')\right)^2 \mathrm{d}M(\xi',\xi) \leq -r.
\end{aligned}$$

The last inequality follows from

$$\begin{aligned}
A \supseteq O =& \left\{\nu'' \in \mathcal{P}_0(\Xi'\times\Xi) \ : \ \nu' \in \mathcal{P}_0(\Xi'\times\Xi), \ \|\nu'-\nu''\| < \delta/4, \ \int \left(\frac{\mathrm{d}\nu'}{\mathrm{d}M}(\xi',\xi)\right)^2 \mathrm{d}M(\xi',\xi) \leq r\right\} \\
\supseteq& \left\{\nu' \in \mathcal{P}_0(\Xi'\times\Xi) \ : \ \int \left(\frac{\mathrm{d}\nu'}{\mathrm{d}M}(\xi',\xi)\right)^2 \mathrm{d}M(\xi',\xi) \leq r\right\}
\end{aligned}$$

with $O$ an open set as it is the union of open sets. Hence, $\mathrm{cl}(A^c) \subseteq \mathrm{cl}(O^c) = O^c$ and $O^c \subseteq \{\nu' \in \mathcal{P}_0(\Xi'\times\Xi) \ : \ \int (\mathrm{d}\nu'/\mathrm{d}M(\xi',\xi))^2 \mathrm{d}M(\xi',\xi) > r\}$. We remark that moderate deviation result of Wu (1994) is stated in the much finer $\tau$-topology and hence applies to the weak topology induced by the norm defined in Section 1.3. Indeed, following Csiszar (2006) a base for the $\tau$ topology are the sets

$$\mathcal{U}(\nu,\mathcal{A},\epsilon) := \{\nu' \in \mathcal{P}_0(\Xi'\times\Xi) \ : \ |\nu(S) - \nu'(S)| < \epsilon \ \forall S \in \mathcal{A}\}$$

for all $\nu \in \mathcal{P}_0(\Xi'\times\Xi)$, $\epsilon > 0$ and $\mathcal{A}$ a finite partition of $\Xi'\times\Xi$. Consider the open balls

$$\mathcal{V}(\nu,\epsilon) := \{\nu' \in \mathcal{P}_0(\Xi'\times\Xi) \ : \ \|\nu'-\nu\| < \epsilon\}$$

for all $\nu \in \mathcal{P}_0(\Xi'\times\Xi)$, $\epsilon > 0$ which are a base of the topology considered here. Fix a $\epsilon''$ and $\nu''$ and consider the open ball $\mathcal{V}(\nu'',\epsilon'')$. Consider a finite partition $\mathcal{A}''$ so that $\forall S \in \mathcal{A}''$ we have $\mathrm{diam}(S) \leq \epsilon''$ which is possible as $\Xi'\times\Xi$ is bounded. Consider any 1-Lipschitz function $f : \Xi'\times\Xi$ then we have

$$\begin{aligned}
&\sup_{\nu'\in\mathcal{U}(\nu'',\mathcal{A}'',\epsilon'')} \int f(\xi',\xi)\,\mathrm{d}\nu'(\xi',\xi) \\
=& \sup_{\forall S\in\mathcal{A}'':\,\epsilon(S)\in\mathrm{R}} \left\{\sup_{\nu'\in\mathcal{P}_0(\Xi'\times\Xi)} \left\{\int f(\xi',\xi)\,\mathrm{d}\nu' \ : \ \forall S \in \mathcal{A}'', \, \nu'(S) = \nu''(S) + \epsilon(S)\right\} \ : \ \forall S \in \mathcal{A}'', \, \epsilon(S) \leq \epsilon'', \ \sum_{S\in\mathcal{A}''} \epsilon(S) = 0\right\} \\
=& \sup_{\forall S\in\mathcal{A}'':\,\epsilon(S)\in\mathrm{R}} \left\{\sup_{\nu'\in\mathcal{P}_0(\Xi'\times\Xi)} \left\{\sum_{S\in\mathcal{A}''} \int_S f(\xi',\xi)\,\mathrm{d}\nu' \ : \ \forall S \in \mathcal{A}'', \, \nu'(S) = \nu''(S) + \epsilon(S)\right\} \ : \ \forall S \in \mathcal{A}'', \, \epsilon(S) \leq \epsilon'', \ \sum_{S\in\mathcal{A}''} \epsilon(S) = 0\right\} \\
=& \sup_{\forall S\in\mathcal{A}'':\,\epsilon(S)\in\mathrm{R}} \left\{\sum_{S\in\mathcal{A}''} \sup_{\nu'\in\mathcal{P}_+(S)} \left\{\int_S f(\xi',\xi)\,\mathrm{d}\nu' \ : \ \nu'(S) = \nu''(S) + \epsilon(S)\right\} \ : \ \forall S \in \mathcal{A}'', \, \epsilon(S) \leq \epsilon'', \ \sum_{S\in\mathcal{A}''} \epsilon(S) = 0\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \int f(\xi',\xi)\,\mathrm{d}\nu'' + \sup\left\{ \sum_{S\in\mathcal{A}''} \sup_{(\xi',\xi)\in S} f(\xi',\xi)(\nu''(S)+\epsilon(S)) - \int_S f(\xi',\xi)\,\mathrm{d}\nu'' \;:\; \forall S\in\mathcal{A}'',\, \epsilon(S)\le\epsilon'',\, \sum_{S\in\mathcal{A}''}\epsilon(S)=0 \right\} \\
&\le \int f(\xi',\xi)\,\mathrm{d}\nu'' + \sup\left\{ \sum_{S\in\mathcal{A}''} \mathrm{diam}(S)\nu''(S) + \sum_{S\in\mathcal{A}''}\sup_{(\xi',\xi)\in S} f(\xi',\xi)\epsilon(S) \;:\; \forall S\in\mathcal{A}'',\, \epsilon(S)\le\epsilon'',\, \sum_{S\in\mathcal{A}''}\epsilon(S)=0 \right\} \\
&\le \int f(\xi',\xi)\,\mathrm{d}\nu'' + \epsilon'' + \sup\left\{ \sum_{S\in\mathcal{A}''} \sup_{(\xi',\xi)\in S} f(\xi',\xi)\epsilon(S) \;:\; \forall S\in\mathcal{A}'',\, \epsilon(S)\le\epsilon'',\, \sum_{S\in\mathcal{A}''}\epsilon(S)=0 \right\} \\
&\le \int f(\xi',\xi)\,\mathrm{d}\nu'' + \epsilon'' + \sup\left\{ \sum_{S\in\mathcal{A}''} \left( \sup_{(\xi',\xi)\in S} f(\xi',\xi) - \inf_{(\xi',\xi)\in\Xi'\times\Xi} f(\xi',\xi)\right)\epsilon(S) \;:\; \forall S\in\mathcal{A}'',\, \epsilon(S)\le\epsilon'',\, \sum_{S\in\mathcal{A}''}\epsilon(S)=0 \right\} \\
&\le \int f(\xi',\xi)\,\mathrm{d}\nu'' + \epsilon'' + \mathrm{diam}(\Xi'\times\Xi)\epsilon''.
\end{aligned}
$$

Hence, we have $\mathcal{U}(\nu',\mathcal{A}',\epsilon'/(1+\mathrm{diam}(\Xi'\times\Xi))) \subseteq \mathcal{V}(\nu',\epsilon')$ indicating the $\tau$ topology is finer than the topology considered here. Here the first equality follows from the definition of $\mathcal{U}(\nu'',\mathcal{A}'',\epsilon'')$. The second equality follows from the fact that $\mathcal{A}''$ partitions $\Xi'\times\Xi$. The third equality follows the fact that the sets in the partition $\mathcal{A}''$ are mutually exclusive. The fourth equality follows from the fact that

$$
\sup_{\nu'\in\mathcal{P}_+(S)}\left\{ \int_S f(\xi',\xi)\,\mathrm{d}\nu' \;:\; \nu'(S)=\nu''(S)+\epsilon(S)\right\} = \sup_{(\xi',\xi)\in S} f(\xi',\xi)(\nu''(S)+\epsilon(S))
$$

for any set $S$ and by both adding and subtracting $\int f(\xi',\xi)\,\mathrm{d}\nu''$. The first inequality follows from the fact that $f$ is 1-Lipschitz and hence $\sup_{(\xi',\xi)\in S} f(\xi',\xi) - f(\xi',\xi) \le \mathrm{diam}(S)$ for any $(\xi',\xi)\in S$. The second inequality follows from $\mathrm{diam}(S)\le\epsilon''$. The third inequality uses $\sum_{S\in\mathcal{A}''}\inf_{(\xi',\xi)\in\Xi'\times\Xi} f(\xi',\xi)\epsilon(S) = \inf_{(\xi',\xi)\in\Xi'\times\Xi} f(\xi',\xi)\sum_{S\in\mathcal{A}''}\epsilon(S) = 0$. The final inequality follows again from the fact that $f$ is 1-Lipschitz and hence

$$
\sup_{(\xi',\xi)\in S} f(\xi',\xi) - \inf_{(\xi',\xi)\in\Xi'\times\Xi} f(\xi',\xi) \le \mathrm{diam}(\Xi'\times\Xi)
$$

for any $S\subseteq\Xi'\times\Xi$.

Remark now that

$$
\begin{aligned}
&\limsup_{N\to\infty}\frac{1}{\lambda(N)^2}\log M^\infty\left[\ell_0(z^\star_{0N,r}) + \mathbb{E}_{M\otimes E}\left[\ell(z^\star_{0N,r}, z^\star_{N,r}(\tilde\xi'+h(N)\tilde e),\tilde\xi)\right] > v^\star_{N,r} + d(N,r,\delta)\right] \\
&\le \limsup_{N\to\infty}\frac{1}{\lambda(N)^2}\log M^\infty\Bigg[\ell_0(z^\star_{0N,r}) + \mathbb{E}_{M\otimes E}\left[\ell(z^\star_{0N,r}, z^\star_{N,r}(\tilde\xi'+h(N)\tilde e),\tilde\xi)\right] \\
&\qquad > \sup\left\{\mathbb{E}_{M''}\left[C_{h(N)}(z^\star_{0N,r}, z^\star_{N,r},\tilde\xi',\tilde\xi)\right] \;:\; M'\in\mathcal{P}_1(\Xi'\times\Xi),\;\; \|M_N - M'\| \le \delta\frac{\lambda(N)}{\sqrt{N}},\, \chi^2(M',M'')\le\frac{\lambda(N)^2 r}{N}\right\}\Bigg] \\
&\le \limsup_{N\to\infty}\frac{1}{\lambda(N)^2}\log M^\infty\left[M\notin\mathcal{M}_N^\delta\right] \le -r.
\end{aligned}
$$

Here the first inequality follows from Lemma 3.5 with $r=\lambda(N)^2 r/N$ and $\delta=\delta\lambda(N)/\sqrt{N}$. The second inequality

follows from the fact that if $M \in \mathcal{M}_N^\delta$ then clearly

$$\ell_0(z_{0N,r}^\star) + \mathbb{E}_{M\otimes E}\left[\ell(z_{0N,r}^\star, z_{N,r}^\star(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right] \leq$$
$$\sup\left\{\mathbb{E}_{M''}\left[C_{h(N)}(z_{0N,r}^\star, z_{N,r}^\star, \tilde{\xi}', \tilde{\xi})\right] \;:\; M' \in \mathcal{P}_1(\Xi'\times\Xi),\;\; \|M_N - M'\| \leq \delta\frac{\lambda(N)}{\sqrt{N}},\, \chi^2(M', M'') \leq \frac{\lambda(N)^2 r}{N}\right\}.$$

Remark that if $\lim_{N\to\infty}\lambda(N)/(\sqrt{N}h(N)) = 0$ we clearly have $\lim_{N\to\infty} d(N,r,\delta) = 0$. The second part of the statement follows from

$$\begin{aligned}
&\lim_{N\to\infty} v_{N,r}^\star + d(N,r,\delta) = \lim_{N\to\infty} v_{N,r}^\star \\
&= \lim_{N\to\infty} \min_{z_0\in Z_0,\, z\in\mathcal{F}} \ell_0(z_0) + \max_{M\in\mathcal{M}_N} \mathbb{E}_{M\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right] \\
&\leq \lim_{N\to\infty} \min_{z_0\in Z_0,\, z\in\mathcal{F}} \ell_0(z_0) + \mathbb{E}_{M_N\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h(N)\tilde{e}), \tilde{\xi})\right] + \frac{L\lambda(N)\sqrt{r}}{2\sqrt{N}} \\
&= \lim_{N\to\infty} v_N^\star + \frac{L\lambda(N)}{2\sqrt{N}}\sqrt{r} = v^\star
\end{aligned}$$

where the first inequality follows from Lemma B.5 and the ultimate equality is shown in Theorem 2.7. □

# 4 Discussion

## 4.1 Time consistency

Generic robust formulations of the form (22) are typically time inconsistent (Bertsimas et al. 2019) in that the optimal policy in one time period may not be perceived as optimal in the next. Suppose for sake of illustration that Problem (22) admits an optimal solution $z_0^\star$ and $z^\star$. Let the decision maker implement the here-and-now decision $z_0^\star$. After observing the noisy covariate context $\tilde{\xi}' + h\tilde{e}$ realize as $\xi'$ the optimal policy $z^\star$ may not appear optimal any longer as in general we may have

$$\min_{z\in Z}\sup_{M'\in\mathcal{M}} \mathbb{E}_{M'\otimes E}\left[\ell(z_0^\star, z, \tilde{\xi})|\tilde{\xi}' + h\tilde{e} = \xi'\right] < \sup_{M'\in\mathcal{M}_N} \mathbb{E}_{M'\otimes E}\left[\ell(z_0^\star, z^\star(\xi'), \tilde{\xi})|\tilde{\xi}' + h\tilde{e} = \xi'\right].$$

Consequently, Bertsimas et al. (2019) caution against using the optimal second-stage policy $z^\star$ altogether. In many practical applications of adaptive decision-making, it suffices to implement the here-and-now decision without having to commit to a policy that dictates how the solution may change when the uncertainty unfolds.

## 4.2 Empirical Risk Formulations

An alternative data-driven counterpart to problem (1) would be

$$\min_{z_0 \in Z_0,\, z \in \mathcal{C}_N} \ \ell_0(z_0) + \mathbb{E}_{M_N}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \tag{25}$$

where the set $\mathcal{C}_N \subset \mathcal{F}$ restricts the set of feasible mappings between covariate information and decisions and where the augmented loss function $\bar{\ell}(z_0, z, \xi) = \ell(z_0, z, \xi) + \Gamma \operatorname{dist}(z, Z)$ captures the requirement $z \in Z$ with the help of a soft constraint which is asymptotically exact as $\Gamma$ tends to infinity. Often the restriction imposed by $\mathcal{C}_N$ requires the second-stage decision function $z$ to be of a certain parametric type. Alternatively, it can impose a certain shape restriction on the second-stage decision function such as monotonicity, convexity or smoothness. One particularly popular parametric restriction is to consider linear functions $\mathcal{C}_N = \{\xi' \mapsto \theta^\top \xi' \ : \ \theta \in \mathrm{R}^{d_2 \times n}, \ \ \|\theta_i\| \leq \gamma_N \ \forall i \in [1, \ldots, n]\}$ where the parameter $\gamma_N \geq 0$ controls the size of coefficients of the linear decision functions. Perhaps the main motivation of this approach is that under Assumption 1.2 problem (25) then reduces to a convex optimization problem which can be solved using off-the-shelf software. Nonlinear functions can be considered in a tractable fashion as well by embedding the decision function in an appropriate reproducing kernel Hilbert space (Scholkopf and Smola 2018). Let $K : \Xi' \times \Xi'$ be a kernel function such that $\sum_{i=1}^k \sum_{j=1}^k a_i a_j K(v'^i, v'^j) \geq 0$ for all $k \in \mathrm{N}$, $v'^1 \in \Xi', \ldots, v'^k \in \Xi'$, and $a_1 \in \mathrm{R}, \ldots, a_k \in \mathrm{R}$ with associated reproducing kernel Hilbert space defined as $\mathcal{H} := \left\{h : \Xi' \to \mathrm{R} \ : \ h(\xi') = \sum_{k=1}^\infty a^k K(v'^k, \xi') \text{ with } v'^k \in \Xi' \ \ \forall k \in \mathrm{N}\right\}$ with inner product defined as $\langle h_1, h_2 \rangle = \sum_{k_1=1}^\infty \sum_{k_2=1}^\infty a^{k_1} b^{k_2} K(v'^{k_1}, w'^{k_2})$ for any $h_1 = \sum_{k_1=1}^\infty a^{k_1} K(v'^{k_1}, \xi')$ and $h_2 = \sum_{k_2=1}^\infty b^{k_2} K(w'^{k_2}, \xi')$. Consider the Ivanov problem

$$\min_{z_0 \in Z_0,\, z \in \mathcal{H}^n} \left\{\ell_0(z_0) + \mathbb{E}_{M_N}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] \ : \ \|z_i\|^2 \leq \gamma_{i,N}^2 \ \ \forall i \in [1, \ldots, n]\right\} \tag{26}$$

where the parameters $\gamma_{i,N} \geq 0$, $i \in [1, \ldots, n]$ control norm of the considered second-stage decision function in $z \in \mathcal{H}^n$. We will denote with $v^\star_{N,e}$ the minimum and with $(z^\star_{N0,e}, z^\star_{N,e})$ a minimizer in the previous minimization problem. We remark that the previous problem is following a Lagrange multipliers argument (Oneto et al. 2016, Theorem 1) equivalent to the Tikhonov problem $\min_{z_0 \in Z_0,\, z \in \mathcal{H}^n} \ell_0(z_0) + \mathbb{E}_{M_N}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] + \sum_{i=1}^n \lambda_{i,N} \|z_i\|^2$ for some regularization parameters $\lambda_{i,N} \geq 0$ for $i \in [1, \ldots, n]$. Bertsimas and Koduri (2021, Proposition 4 & 5) show that under somewhat restrictive conditions on problem (1) when the reproducing kernel Hilbert space $\mathcal{H}$ is universal (Micchelli et al. 2006) and the amount of regularization is judiciously chosen one can guarantee $\lim_{N\to\infty} \ell_0(z^\star_{0N}) + \mathbb{E}_M\left[\bar{\ell}(z^\star_{0N}, z^\star_{N,e}(\tilde{\xi}'), \tilde{\xi})\right] = \lim_{N\to\infty} v^\star_{N,e} = v^\star$ in probability.

The reproducing kernel Hilbert space restriction $z \in \mathcal{H}^n$ on the second-stage decision has a similar regularization

effect as adding artificial noise. The following result is indeed a direct counterpart to Proposition 2.11 and proves that the restriction $z \in \mathcal{H}^n$ recovers Lipschitz continuity when the associated kernel function is sufficiently calm.

**Proposition 4.1.** *Let Assumption 1.2 hold with* $\sqrt{K(v_1', v_1') - 2K(v_1', v_2') + K(v_2', v_2')} \leq \mathcal{K} \|v_1' - v_2'\|$ *for all* $v_1' \in \Xi'$ *and* $v_2' \in \Xi'$. *We have for all* $M_1$ *and* $M_2$ *in* $\mathcal{P}(\Xi' \times \Xi)$ *the inequality* $\left|\mathbb{E}_{M_1}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_{M_2}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]\right| \leq (\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{i,N} \|M_1 - M_2\|$ *for all* $z_0 \in Z_0$ *and* $z \in \mathcal{H}^n$ *with* $\|h_i\|^2 \leq 2\gamma_{i,N}$ *for all* $i \in [1, \ldots, n]$.

*Proof.* We shall proof that $z \in \mathcal{H}^n$ is $\mathcal{K} \sum_{i=1}^n \gamma_{N,i}$-Lipschitz and hence the composite mapping $(\xi', \xi) \mapsto \bar{\ell}(z_0, z(\xi'), \xi)$ is $(\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{N,i}$-Lipschitz. We have for all $\xi_1' \in \Xi'$ and $\xi_2' \in \Xi'$ indeed

$$\begin{aligned}
\|z(\xi_1') - z(\xi_2')\| &\leq \sum_{i=1}^n |z_i(\xi_1') - z_i(\xi_2')| \\
&\leq \sum_{i=1}^n \|z_i\| \sqrt{K(\xi_1', \xi_1') - 2K(\xi_1', \xi_2') + K(\xi_2', \xi_2')} \\
&\leq \sum_{i=1}^n \gamma_{N,i} \mathcal{K} \|\xi_1' - \xi_2'\|
\end{aligned}$$

where the first inequality follows from the reproducing kernel Hilbert space property of $z_i$ and the second inequality is due to Cauchy-Schwartz.

Following the definition of $\|M_1 - M_2\|$ in Equation (11) it follows that

$$\begin{aligned}
\|M_1 - M_2\| &= \sup_f \left\{\mathbb{E}_{M_1}\left[f(\xi, \xi')\right] - \mathbb{E}_{M_2}\left[f(\xi, \xi')\right] \;:\; (\xi', \xi) \mapsto f(\xi', \xi) \text{ is 1 Lipschitz}\right\} \\
&\geq \mathbb{E}_{M_1}\left[\frac{\bar{\ell}(z_0, z(\xi'), \xi)}{(\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{N,i}}\right] - \mathbb{E}_{M_2}\left[\frac{\bar{\ell}(z_0, z(\xi'), \xi)}{(\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{N,i}}\right]
\end{aligned}$$

and symmetrically

$$\begin{aligned}
\|M_2 - M_1\| &= \sup_f \left\{\mathbb{E}_{M_2}\left[f(\xi, \xi')\right] - \mathbb{E}_{M_1}\left[f(\xi, \xi')\right] \;:\; (\xi', \xi) \mapsto f(\xi', \xi) \text{ is 1 Lipschitz}\right\} \\
&\geq \mathbb{E}_{M_2}\left[\frac{\bar{\ell}(z_0, z(\xi'), \xi)}{(\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{N,i}}\right] - \mathbb{E}_{M_1}\left[\frac{\bar{\ell}(z_0, z(\xi'), \xi)}{(\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{N,i}}\right].
\end{aligned}$$

Hence, combining both inequalities gives the desired result $\left|\mathbb{E}_{M_1}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right] - \mathbb{E}_{M_2}\left[\bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi})\right]\right| \leq (\mathcal{L} + \Gamma)\mathcal{K} \sum_{i=1}^n \gamma_{i,N} \|M_1 - M_2\|$. □

It is well known that despite its potential infinite dimensional nature the Ivanov problem (26) can be solved efficiently (Schölkopf et al. 2001). Introduce indeed the Fenchel dual costs $\ell_0^\star : \mathrm{R}^{n_0} \to \mathrm{R}$, $\beta \mapsto \max_{z_0 \in Z_0} \beta^\top z_0 - \ell(z_0)$ and $\ell^\star : \mathrm{R}^{n_0} \times \mathrm{R}^n \times \Xi \mapsto \mathrm{R}$, $(\beta, \alpha, \xi) \mapsto \max_{z_0 \in Z_0, z \in Z} \; -(\beta, \alpha)^\top (z_0, z) - \ell(z_0, z, \xi)$ which are lower semicontinuous and convex in $\beta$ and $(\beta, \alpha)$ respectively for every $\xi \in \Xi$.

**Theorem 4.2.** *Under Assumption 1.2 and $\gamma_{i,N} \geq 0$ for all $1 \leq i \leq n$ we have that formulation* (26) *equals*

$$\begin{array}{ll} \max & -\ell_0^\star\left(-\sum_{d=1}^N \beta_d\right) - \sum_{d=1}^N \ell^\star\left(\frac{\beta_d}{M_N(\xi'_d,\xi_d)}, \frac{(\alpha_{d,1},\ldots,\alpha_{d,n})}{M_N(\xi'_d,\xi_d)}, \xi_d\right) M_N(\xi'_d,\xi_d) - \sum_{i=1}^n \gamma_{i,N} \left\|(\alpha_{1,i},\ldots,\alpha_{N,i})\right\|_\star \\ \text{s.t.} & \beta_d \in \mathrm{R}^{n_0},\ \alpha_d \in \mathrm{R}^n \quad \forall d \in [1,\ldots,N], \end{array} \tag{27}$$

*where the norm $\|(\alpha_{1,i},\ldots,\alpha_{N,i})\|_\star^2 := (\alpha_{1,i},\ldots,\alpha_{N,i})^\top G(\alpha_{1,i},\ldots,\alpha_{N,i})$ is associated with the positive semidefinite Grammian $G \in \mathrm{S}_+^N$ characterized element-wise by $G_{d,d'} = K(\xi'_d,\xi'_{d'})$ for all $1 \leq d \leq N$, $1 \leq d' \leq N$.*

*Proof.* Following an extension of the representer Theorem (Shafieezadeh Abadeh 2020, Theorem 2.32) we have that it is sufficient to optimize over $z(\xi') = (\sum_{d=1}^N \alpha'_{d,1} K(\xi'_d,\xi'),\ldots,\sum_{d=1}^N \alpha'_{d,n} K(\xi'_d,\xi'))$ for $\alpha'_d \in \mathrm{R}^n$ for $1 \leq d \leq N$. Hence, we have

$$\begin{aligned} &(26) \\ =&\left\{\begin{array}{ll} \min & \ell_0(z_0) + \mathbb{E}_{M_N}\left[\bar{\ell}(z_0, (\sum_{d=1}^N \alpha'_{d,1} K(\xi'_d,\tilde{\xi}'),\ldots,\sum_{d=1}^N \alpha'_{d,n} K(\xi'_d,\tilde{\xi}')),\tilde{\xi})\right] \\ \text{s.t.} & z_0 \in Z_0,\ \alpha'_d \in \mathrm{R}^n \quad 1 \leq d \leq N, \\ & \left\|(\alpha'_{1,i},\ldots,\alpha'_{N,i})\right\|_\star \leq \gamma_{i,N} \quad \forall 1 \leq i \leq n. \end{array}\right. \\ =&\left\{\begin{array}{ll} \min & \ell_0(z_0) + \sum_{d=1}^N \bar{\ell}(z_{0d}, z_d, \xi) M_N(\xi'_d,\xi_d) \\ \text{s.t.} & z_0 \in Z_0,\ z_{0d} \in Z_0,\ z_d \in Z,\ \alpha'_d \in \mathrm{R}^n \quad 1 \leq d \leq N, \\ & \left\|(\alpha'_{1,i},\ldots,\alpha'_{N,i})\right\|_\star \leq \gamma_{i,N} \quad \forall 1 \leq i \leq n, \\ & z_d = (\sum_{d'=1}^N \alpha'_{d',1} K(\xi'_{d'},\xi'_d),\ldots,\sum_{d'=1}^N \alpha'_{d',n} K(\xi'_{d'},\xi'_d)) \quad 1 \leq d \leq N, \\ & z_{0d} = z_0 \quad 1 \leq d \leq N, \end{array}\right. \end{aligned}$$

Here, the first equality exploits the reproducing kernel Hilbert space property that for any $1 \leq i \leq n$ we have $\|\xi' \mapsto \sum_{d=1}^N \alpha_{d,i} K(\xi'_d,\xi')\| = \|(\alpha_{1,i},\ldots,\alpha_{N,i})\|_\star$. Introduce now the Lagrangian

$$\begin{aligned} &L(z_0, z_{0d}, z_d, \alpha'_d\ ;\ \alpha_d, \beta_d) \\ :=&\ell_0(z_0) + \sum_{d=1}^N \bar{\ell}(z_{0d}, z_d, \xi) M_N(\xi'_d,\xi_d) + \sum_{d=1}^N \beta_d(z_{0d} - z_0) \\ &+ \sum_{d=1}^N \alpha_d^\top (z_d - (\sum_{d'=1}^N \alpha'_{d',1} K(\xi'_{d'},\xi'_d),\ldots,\sum_{d=1}^N \alpha'_{d,n} K(\xi'_{d'},\xi'_d))) \\ =&\ell_0(z_0) - z_0(\sum_{d=1}^N \beta_d) + \sum_{d=1}^N (\beta_d,\alpha_d)^\top (z_{0,d}, z_d) + \bar{\ell}(z_{0d}, z_d, \xi) M_N(\xi'_d,\xi_d) \\ &- \sum_{d=1}^N \alpha_d^\top (\sum_{d'=1}^N \alpha'_{d',1} K(\xi'_{d'},\xi'_d),\ldots,\sum_{d=1}^N \alpha'_{d',n} K(\xi'_{d'},\xi'_d)) \end{aligned}$$

Introduce the dual function

$$
\begin{aligned}
& g(\alpha_d, \beta_d) \\
:= & \min \left\{ L(z_0, z_{0d}, z_d, \alpha'_d \; ; \; \alpha_d, \beta_d) \; : \; z_0 \in Z_0, z_{0d} \in Z_0, z_d \in Z, \alpha'_d \in \mathrm{R}^n \right\} \\
= & [\min_{z_0 \in Z_0} \ell_0(z_0) - z_0(\textstyle\sum_{d=1}^N \beta_d)] + \textstyle\sum_{d=1}^N \left[ \min_{z_{0d} \in Z_0, z_d \in Z} (\beta_d, \alpha_d)^\top (z_{0,d}, z_d) + \bar{\ell}(z_{0d}, z_d, \xi) M_N(\xi'_d, \xi_d) \right] \\
& - \max \left\{ \textstyle\sum_{d=1}^N \alpha_d^\top (\sum_{d'=1}^N \alpha'_{d',1} K(\xi'_{d'}, \xi'_d), \ldots, \sum_{d'=1}^N \alpha'_{d',n} K(\xi'_{d'}, \xi'_d)) \; : \; \left\| (\alpha'_{1,i}, \ldots, \alpha'_{N,i}) \right\|_\star \leq \gamma_{i,N} \quad \forall 1 \leq i \leq n \right\} \\
= & - \ell_0^\star(\textstyle\sum_{d=1}^N \beta_d)) - \textstyle\sum_{d=1}^N \bar{\ell}^\star(\frac{\alpha_d}{M_N(\xi'_d, \xi_d)}, \frac{\beta_d}{M_N(\xi'_d, \xi_d)}) M_N(\xi'_d, \xi_d) \\
& - \max \left\{ \textstyle\sum_{d=1}^N \alpha_d^\top (\sum_{d'=1}^N \alpha'_{d',1} K(\xi'_{d'}, \xi'_d), \ldots, \sum_{d'=1}^N \alpha'_{d',n} K(\xi'_{d'}, \xi'_d)) \; : \; \left\| (\alpha'_{1,i}, \ldots, \alpha'_{N,i}) \right\|_\star \leq \gamma_{i,N} \quad \forall 1 \leq i \leq n \right\} \\
= & - \ell_0^\star(\textstyle\sum_{d=1}^N \beta_d)) - \textstyle\sum_{d=1}^N \bar{\ell}^\star(\frac{\alpha_d}{M_N(\xi'_d, \xi_d)}, \frac{\beta_d}{M_N(\xi'_d, \xi_d)}) M_N(\xi'_d, \xi_d) - \textstyle\sum_{i=1}^n \gamma_{i,N} \left\| (\alpha_{1,i}, \ldots, \alpha_{N,i}) \right\|_\star .
\end{aligned}
$$

The final inequality follows from $(\alpha'^\star_{1,i}, \ldots, \alpha'^\star_{N,i}) = \gamma_{i,N}(\alpha^\star_{1,i}, \ldots, \alpha^\star_{N,i}) / \left\| (\alpha^\star_{1,j}, \ldots, \alpha^\star_{N,j}) \right\|_\star$ for any $1 \leq j \leq n$. As Slater's constraint qualification condition is met we have (26) $= \max \{ g(\alpha, \beta) \; : \; \alpha_d, \; \beta_d \}$ from which the claimed result follows immediately. □

A direct robust counterpart to the empirical risk formulation can now be reformulated as

$$
\begin{aligned}
& \min_{z_0 \in Z_0, \, z \in \mathcal{H}^n} \max_{M \in \mathcal{M}_N} \; \ell_0(z_0) + \mathbb{E}_M \left[ \bar{\ell}(z_0, z(\tilde{\xi}'), \tilde{\xi}) \right] \\
= & \begin{cases} \max & -\ell_0^\star \left( - \sum_{d=1}^N \beta_d \right) - \sum_{d=1}^N \ell^\star \left( \frac{\beta_d}{M(\xi'_d, \xi_d)}, \frac{(\alpha_{d,1}, \ldots, \alpha_{d,n})}{M(\xi'_d, \xi_d)}, \xi_d \right) M(\xi'_d, \xi_d) - \sum_{i=1}^n \gamma_{i,N} \left\| (\alpha_{1,i}, \ldots, \alpha_{N,i}) \right\|_\star \\ \text{s.t.} & M \in \mathcal{M}_N, \; \beta_d \in \mathrm{R}^{n_0}, \; \alpha_d \in \mathrm{R}^n \quad 1 \leq d \leq N. \end{cases}
\end{aligned}
$$

and is amendable to convex optimization as $\sum_{d=1}^N M(\xi'_d, \xi_d) \ell^\star \left( \beta_d / M(\xi'_d, \xi_d), (\alpha_{d,1}, \ldots, \alpha_{d,n}) / M(\xi'_d, \xi_d), \xi_d \right)$ is jointly convex in both $(\alpha, \beta)$ and $M$. An optimal robust solution can be recovered as $z_0^\star \in \arg\max_{z_0 \in Z_0} (\sum_{d=1}^N \beta_d)^\top z_0 - \ell(z_0)$ and $z^\star(\xi') = (\sum_{d=1}^N \alpha'^\star_{d,1} K(\xi'_d, \xi'), \ldots, \sum_{d=1}^N \alpha'^\star_{d,n} K(\xi'_d, \xi'))$ where for any $1 \leq i \leq n$ we define $(\alpha'^\star_{1,i}, \ldots, \alpha'^\star_{N,i}) = \gamma_{i,N}(\alpha^\star_{1,i}, \ldots, \alpha^\star_{N,i}) / \|(\alpha^\star_{1,i}, \ldots, \alpha^\star_{N,i})\|_\star$. Hence, the empirical risk formulations with kernels by Bertsimas and Koduri (2021) and the robust approach considered here are not mutually exclusive but rather can be combined which we leave here however as a future research direction.

# References


R Tyrrell Rockafellar and Roger J-B Wets. **Variational Analysis**, volume 317. Springer Science & Business Media, 2009.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. **Operations Research**, 67(1):90–108, 2019.

Dimitris Bertsimas and Nihal Koduri. Data-driven optimization: A reproducing kernel hilbert space approach. **Operations Research**, 2021.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. **Journal of Machine Learning Research**, 7 (12), 2006.

Grani Adiwena Hanasusanto and Daniel Kuhn. Robust data-driven dynamic programming. **Advances in Neural Information Processing Systems**, 26, 2013.

Lauren Hannah, Warren Powell, and David Blei. Nonparametric density estimation for stochastic optimization with an observable state variable. **Advances in Neural Information Processing Systems**, 23, 2010.

D. Bertsimas and C. McCord. From predictions to prescriptions in multistage optimization problems. **arXiv preprint arXiv:1904.11637**, 2019.

Rui Gao, Jincheng Yang, and Luhao Zhang. Optimal robust policy for feature-based newsvendor. URL `https://optimization-online.org/2021/12/8749/`.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. **Probability Theory and Related Fields**, 162(3):707–738, 2015.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. **arXiv preprint arXiv:1610.03425**, 2016.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. **Operations Research**, 67(4):1090–1105, 2019.

M. Amine Bennouna and Bart P. G. Van Parys. Learning and decision-making with data: Optimal formulations and phase transitions, 2021. URL `https://arxiv.org/abs/2109.06911`.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. **SIAM Journal on optimization**, 19(4):1574–1609, 2009.

Amir Dembo and Ofer Zeitouni. **Large Deviations Techniques and Applications**, volume 38. Springer Science & Business Media, 2009.

Lasse Holmstrom and Petri Koistinen. Using additive noise in back-propagation training. **IEEE transactions on neural networks**, 3(1):24–38, 1992.

François Chollet et al. Keras. `https://keras.io`, 2015.

E. Parzen. On estimation of a probability density function and mode. **The annals of mathematical statistics**, 33(3): 1065–1076, 1962.

L. Devroye and C.S. Penrod. The consistency of automatic kernel density estimates. **The Annals of Statistics**, pages 1231–1249, 1984.

L. Devroye and A. Krzyzak. An equivalence theorem for l1 convergence of the kernel regression estimate. **Journal of statistical planning and Inference**, 23(1):71–82, 1989.

Hannelore Liero. Strong uniform consistency of nonparametric regression function estimates. **Probability theory and related fields**, 82(4):587–614, 1989.

V. Strassen. The existence of probability measures with given marginals. **The Annals of Mathematical Statistics**, 36 (2):423–439, 1965.

Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. **Vestnik of the St. Petersburg University: Mathematics**, 13(7):52–59, 1958.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In **The Operations Research Revolution**, pages 1–19. INFORMS, 2015.

Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. **Management Science**, 67(6):3387–3402, 2021.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. **Foundations and Trends® in Machine Learning**, 4(2):107–194, 2012.

Maurice Sion. On general minimax theorems. **Pacific Journal of mathematics**, 8(1):171–176, 1958.

Liming Wu. Large deviations, moderate deviations and lil for empirical processes. **The Annals of Probability**, pages 17–27, 1994.

Imre Csiszar. A simple proof of sanov's theorem. **Bulletin of the Brazilian Mathematical Society**, 37(4), 2006.

Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive distributionally robust optimization. **Management Science**, 65(2):604–618, 2019.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. **Operations Research**, 53(5):780–798, 2005.

Bernhard Scholkopf and Alexander J. Smola. **Learning with kernels: Support vector machines, regularization, optimization, and beyond**. Adaptive Computation and Machine Learning series, 2018.

Luca Oneto, Sandro Ridella, and Davide Anguita. Tikhonov, ivanov and morozov regularization for support vector machine learning. **Machine Learning**, 103(1):103–136, 2016.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In **International conference on computational learning theory**, pages 416–426. Springer, 2001.

Soroosh Shafieezadeh Abadeh. **Wasserstein Distributionally Robust Learning**. PhD thesis, EPFL, 2020.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with $f$-divergences. **Advances in neural information processing systems**, 29, 2016.

Paul R. Halmos. **Measure Theory**, volume 18. Springer, 2013.

Dimitri P Bertsekas. **Control of Uncertain Systems with a Set-Membership Description of the Uncertainty**. PhD thesis, MIT, 1971.

# A Proofs

## A.1 Proof Proposition 2.10

*Proof.* Define $C_h(z_0, z, \xi', \xi) := \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'')$ as we done in Section 2.1. We have indicated that

$$M \to \mathbb{E}_{M \otimes E} \left[ \ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi}) \right] = \mathbb{E}_M \left[ C_h(z_0, z, \tilde{\xi}', \tilde{\xi}) \right].$$

The convolution of an $L_\infty$-function $\xi' \to \ell(z_0, z(\xi'), \xi)$ and an $L_1$-function $\xi' \to k(\xi'/h)/h^{d_2}$ is bounded and continuous in $\xi' \in \Xi' + hK$ for each $\xi \in \Xi$, $z_0 \in Z_0$ and $z \in \mathcal{F}$. The continuity of the integrant $C_h$ in both $(\xi, \xi') \in (\Xi' + hK) \times \Xi$ for each $z_0 \in Z_0$ and $z \in \mathcal{F}$ follows now immediately from the observation that

$$\begin{aligned}
&\lim_{k \to \infty} |C_h(z_0, z, \xi'_k, \xi_k) - C_h(z_0, z, \xi'_\infty, \xi_\infty)| \\
&\leq \lim_{k \to \infty} |C_h(z_0, z, \xi'_k, \xi_k) - C_h(z_0, z, \xi'_k, \xi_\infty)| + \lim_{k \to \infty} |C_h(z_0, z, \xi'_k, \xi_\infty) - C_h(z_0, z, \xi'_\infty, \xi_\infty)| \\
&\leq \lim_{k \to \infty} \int \ell(z_0, z(\xi''), \xi_k) - \ell(z_0, z(\xi''), \xi_\infty) k((\xi'' - \xi'_k)/h)/h^{d_2} \, \mathrm{d}m'(\xi'') \\
&\leq \lim_{k \to \infty} \mathcal{L} \|\xi_k - \xi_\infty\| \int k((\xi'' - \xi'_k)/h)/h^{d_2} \, \mathrm{d}m'(\xi'') = \lim_{k \to \infty} \mathcal{L} \|\xi_k - \xi_\infty\| = 0.
\end{aligned}$$

Here the first inequality follows from the triangular inequality. The second inequality follows from the continuity of the function $g$ in $\xi'$ for all fixed $\xi$, $z_0$ and $z$. The final inequality follows from the fact that $\ell$ is Lipschitz as stated in Assumption 1.2. □

## A.2 Proof Proposition 2.11

*Proof.* Define $C_h(z_0, z, \xi', \xi) := \int \ell(z_0, z(\xi''), \xi) k((\xi'' - \xi')/h)/h^{d_2} \, \mathrm{d}m'(\xi'')$ as we done in Section 2.1. We have indicated that

$$M \to \mathbb{E}_{M \otimes E} \left[ \ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi}) \right] = \mathbb{E}_M \left[ C_h(z_0, z, \tilde{\xi}', \tilde{\xi}) \right].$$

First, the convolution of a $L$-bounded function $\xi' \to \ell(z_0, z(\xi'), \xi)$ and a $K/h^{d_2+1}$-Lipschitz function $\xi' \to k(\xi'/h)/h^{d_2}$ is $L$ bounded and $KL/h$-Lipschitz continuous in $\xi' \in \Xi'$ for each $\xi \in \Xi$, $z_0 \in Z_0$ and $z \in \mathcal{F}$. Indeed, we have

$$
\begin{aligned}
&|C_h(z_0, z, \xi_1', \xi) - C_h(z_0, z, \xi_2', \xi)| \\
=&\int \left|\ell(z_0, z(\xi''), \xi)k((\xi'' - \xi_1')/h)/h^{d_2} - \ell(z_0, z(\xi''), \xi)k((\xi'' - \xi_2')/h)/h^{d_2}\right| \mathrm{d}m'(\xi'') \\
=&\int_{\{\xi_1', \xi_2'\}+hK} \left|\ell(z_0, z(\xi''), \xi)k((\xi'' - \xi_1')/h)/h^{d_2} - \ell(z_0, z(\xi''), \xi)k((\xi'' - \xi_2')/h)/h^{d_2}\right| \mathrm{d}m'(\xi'') \\
=&\int_{\{\xi_1', \xi_2'\}+hK} \ell(z_0, z(\xi''), \xi) \left|k((\xi'' - \xi_1')/h) - k((\xi'' - \xi_2')/h)\right| / h^{d_2} \, \mathrm{d}m'(\xi'') \\
\leq&\int_{\{\xi_1', \xi_2'\}+hK} \ell(z_0, z(\xi''), \xi)\mathcal{K} \left\|(\xi'' - \xi_1')/h - (\xi'' - \xi_2')/h\right\| / h^{d_2} \, \mathrm{d}m'(\xi'') \\
\leq&\int_{\{\xi_1', \xi_2'\}+hK} \ell(z_0, z(\xi''), \xi)\mathcal{K}/h^{d_2+1} \left\|\xi_1' - \xi_2'\right\| \mathrm{d}m'(\xi'') = L\mathcal{K}/h^{d_2+1} \left\|\xi_1' - \xi_2'\right\| m'(\{\xi_1', \xi_2'\} + hK) \\
\leq&L\mathcal{K}/h^{d_2+1} \left\|\xi_1' - \xi_2'\right\| 2m'(hK) = 2\mathcal{K}L/h \left\|\xi_1' - \xi_2'\right\|
\end{aligned}
$$

for any $\xi_1' \in \Xi' + hK$ and $\xi_2' \in \Xi' + hK$. The second inequality follows from the fact that $k$ vanishes outside $K$. The fourth inequality follows from Lipschitz continuity of $k$. The penultimate inequality follows from the $L$-boundedness of the loss $\ell$ as postulated in Assumption 1.2. The $\max(\mathcal{L}, 2\mathcal{K}L/h)$-Lipschitz continuity of the integrant $C_h$ in $(\xi', \xi)$ for each $z_0 \in Z_0$ and $z \in \mathcal{F}$ follows now immediately from the observation that

$$
\begin{aligned}
&|C_h(z_0, z, \xi_k', \xi_k) - C_h(z_0, z, \xi_\infty', \xi_\infty)| \\
\leq&|C_h(z_0, z, \xi_k', \xi_k) - C_h(z_0, z, \xi_\infty', \xi_k)| + |C_h(z_0, z, \xi_\infty', \xi_k) - C_h(z_0, z, \xi_\infty', \xi_\infty)| \\
\leq&\int \ell(z_0, z(\xi''), \xi_k) - \ell(z_0, z(\xi''), \xi_\infty)k((\xi'' - \xi_\infty')/h)/h^{d_2} \, \mathrm{d}m'(\xi'') + 2\mathcal{K}L/h \left\|\xi_k' - \xi_\infty'\right\| \\
\leq&\mathcal{L} \left\|\xi_k - \xi_\infty\right\| + 2\mathcal{K}L/h \left\|\xi_k' - \xi_\infty'\right\| \leq \max(\mathcal{L}, 2\mathcal{K}L/h) \left\|(\xi_k, \xi_k') - (\xi_k, \xi_k')\right\|.
\end{aligned}
$$

Here the first inequality follows from the triangular inequality. The second inequality follows from the Lipschitz continuity of the function $C_h(z_0, z, \xi', \xi)$ in $\xi'$ for all fixed $\xi$, $z_0$ and $z$. The penultimate inequality follows from the fact that $\ell$ is $\mathcal{L}$-Lipschitz as stated in Assumption 1.2.

By definition of the norm

$$
\|M_1 - M_2\| = \sup\left\{\int f(\xi', \xi)\mathrm{d}\nu(\xi', \xi) \;:\; |f(\xi_1', \xi_1) - f(\xi_2', \xi_2)| \leq \|(\xi_1', \xi_1) - (\xi_2', \xi_2)\| \;\; \forall (\xi_1', \xi_1), (\xi_2', \xi_2) \in \Xi' \times \Xi\right\}
$$

it follows that

$$\begin{aligned}
&\mathbb{E}_{M_1\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] - \mathbb{E}_{M_2\otimes E}\left[\ell(z_0, z(\tilde{\xi}' + h\tilde{e}), \tilde{\xi})\right] \leq \max(\mathcal{L}, 2\mathcal{K}L/h)\,\|M_1 - M_2\| \\
&= \int C_h(z_0, z, \xi', \xi)\,\mathrm{d}(M_1(\xi', \xi) - M_2(\xi', \xi)) \\
&= \max(\mathcal{L}, 2\mathcal{K}L/h) \int \frac{C_h(z_0, z, \xi', \xi)}{\max(\mathcal{L}, 2\mathcal{K}L/h)}\,\mathrm{d}(M_1 - M_2) \leq \max(\mathcal{L}, 2\mathcal{K}L/h)\,\|M_1 - M_2\|
\end{aligned}$$

as $C_h(z_0, z, \xi', \xi)/\max(\mathcal{L}, 2\mathcal{K}L/h)$ is 1-Lipschitz in $(\xi', \xi)$. □

# B Neyman Divergence Optimization

Assume $U$ is a compact set in $\mathrm{R}^u$ and $\ell : U \to \mathrm{R}_+$ a continuous and bounded function. Denote its maximum as $\max_U \ell := \max_{u\in U} \ell(u)$ and the compact set of its maximizers as $U^\star := \max_{u\in U} \ell(u)$ which exist due to the extreme value theorem. Define the essential supremum of the function $\ell$ with respect to $P \in \mathcal{P}_1(U)$ finally as

$$\begin{aligned}
\operatorname{ess\,sup}_P \ell = \sup \quad & \int \ell(u)\,\mathrm{d}Q_c(u) \\
\text{s.t.} \quad & Q_c \in \mathcal{P}(U),\ Q_c \ll P.
\end{aligned} \tag{28}$$

Note here that $P$ is not assumed to be a positive measure. We take here following Halmos (2013, Section 29) $Q_c \ll P$ if and only if $|Q_c| \ll |P|$ with $|Q_c| = Q_c$ and $|P|$ the nonnegative total variation measures associated with $Q_c$ and $P$, respectively. The Neyman $\chi^2 : \mathcal{P}_1(U) \times \mathcal{P}(U) \to \mathrm{R} \cup \{+\infty\}$-divergence metric is defined as

$$\begin{aligned}
\chi^2(P, Q) &= \int \left(\frac{\mathrm{d}(P - Q)}{\mathrm{d}Q}(u)\right)^2 \mathrm{d}Q(u) \geq 0 \\
&= \int \frac{\mathrm{d}P}{\mathrm{d}Q}(u) - 1\,\mathrm{d}P(u)
\end{aligned}$$

if $P \ll Q$; and $+\infty$ otherwise. We will study in this section the distributional maximization problem

$$\begin{aligned}
\sup \quad & \int \ell(u)\,\mathrm{d}Q(u) \\
\text{s.t.} \quad & Q \in \mathcal{P}(U),\ P \ll Q, \\
& \chi^2(P, Q) \leq r
\end{aligned} \tag{29}$$

for $P \in \mathcal{P}_1(\Xi)$. We remark here that the constraint $P \ll Q$ in optimization Problem 29 is redundant and in fact already implied by the condition $\chi^2(P, Q) \leq r$. However, the condition $\chi^2(P, Q) \leq r$ does not imply that $Q$ is continuous with respect to $P$, i.e., $P \ll Q \ll P$.

## B.1 Singular Neyman Divergence Optimization

We first prove that problem (29) admits an equivalent representation as a maximization problem over positive measures which are continuous with respect to $P$ and a positive measure which may be singular to $P$. Consider indeed the maximization problems

$$\begin{aligned} &\sup \textstyle\int \ell(u)\,\mathrm{d}Q_c(u) + q(\operatorname{ess\,sup}_P \ell + \epsilon') \\ &\text{s.t. } q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ P \ll Q_c,\ Q_c \ll P, \\ &\quad \textstyle\int \mathrm{d}Q_c(u) + q = 1, \\ &\quad \textstyle\int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r. \end{aligned} \tag{30}$$

**Lemma B.1.** *Let* $\epsilon' = \max_U \ell - \operatorname{ess\,sup}_P \ell$. *We have* (29) = (30) *for all* $P \in \mathcal{P}_1(U)$.

*Proof.* We first show that (29) provides an upper bound on (30). To this end, choose any $Q_c \in \mathcal{P}_+$ and $q \in \mathrm{R}_+$ feasible in (30), and define $Q = Q_c + q\delta_{U^\star} \in \mathcal{P}$. The construction of $Q$ thus implies that $P \ll Q_c \ll Q$. By the Jordan decomposition theorem (Halmos 2013, Section 29) we can decompose $P = P^+ - P^-$ into two positive measure so that $P^+ \in \mathcal{P}_+(U^+)$ and $P^- \in \mathcal{P}_+(U^-)$ where $U^+ \cap U^- = \emptyset$ and $U^+ \cup U^- = U$. This implies that

$$\begin{aligned} &\int \frac{\mathrm{d}P}{\mathrm{d}Q}(u) - 1\,\mathrm{d}P(u) \\ &= \int_{U^+} \frac{\mathrm{d}P_+}{\mathrm{d}Q}(u) - 1\,\mathrm{d}P_+(u) + \int_{U^-} \frac{\mathrm{d}(-P_-)}{\mathrm{d}Q}(u) - 1\,\mathrm{d}(-P_-)(u) \\ &= \int_{U^+} \frac{\mathrm{d}P_+}{\mathrm{d}Q}(u) - 1\,\mathrm{d}P_+(u) + \int_{U^-} \frac{\mathrm{d}P_-}{\mathrm{d}Q}(u) + 1\,\mathrm{d}P_-(u) \\ &\leq \int_{U^+} \frac{\mathrm{d}P_+}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P_+(u) + \int_{U^-} \frac{\mathrm{d}P_-}{\mathrm{d}Q_c}(u) + 1\,\mathrm{d}P_-(u) \\ &= \int_{U^+} \frac{\mathrm{d}P_+}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P_+(u) + \int_{U^-} \frac{\mathrm{d}(-P_-)}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}(-P_-)(u) \\ &= \int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r. \end{aligned}$$

Thus, $Q$ is feasible in (29). Moreover, it is easy to verify that the objective value of $Q$ in (29) is equal to that of $(Q_c, q)$ in (30). Thus, (29) provides an upper bound on (30).

It remains to be shown that (30) provides a upper bound on (29). To this end, choose any $Q$ feasible in (29). The Lebesgue decomposition theorem (Halmos 2013, Section 32) guarantees that we can decompose $Q = Q_c + Q_s$ into continuous part $Q_c \ll P$ and singular part $Q_s \perp P$. As $Q_s \perp P \in \mathcal{P}(\Sigma)$ there is by definition a set $U^0$ so that for any measurable subset $A$ of $U^0$ we have $P(A) = 0$ while for any measurable subset $A$ of $U \setminus U^0$ we have $Q_s(A) = 0$. Set

$q = Q_s(U^0)$. We have that

$$\begin{aligned}\int \mathrm{d}Q_c(u) + q &= \int_{U\setminus U^0} \mathrm{d}Q_c(\xi) + \int_{U^0} \mathrm{d}Q_c(u) + \int_{U^0} \mathrm{d}Q_s(u)\\ &= \int_{U\setminus U^0} \mathrm{d}Q_c(u) + \int_{U^0} \mathrm{d}Q_s(u)\\ &= \int_{U\setminus U^0} \mathrm{d}Q_c(u) + \int_{U\setminus U^0} \mathrm{d}Q_s(u) + \int_{U^0} \mathrm{d}Q_s(u) + \int_{U^0} \mathrm{d}Q_c(u)\\ &= \int_{U\setminus U^0} \mathrm{d}(Q_c + Q_s)(u) + \int_{U^0} \mathrm{d}(Q_c + Q_s)(u) = 1.\end{aligned}$$

Here the first inequality follows the total law of probability. The second equality follows from the fact that $Q_c \ll P$ and hence $|P|(U^0) = 0 \implies Q_c(U^0) = 0$ as indeed $P(A) = 0$ for every measurable set $A \subset U^0$. The third equality follows from $\int_{U\setminus U^0} \mathrm{d}Q_s(u) = 0$ and again the fact that $Q_c \ll P$ implies $\int_{U^0} \mathrm{d}\,|P|(u) = 0 \implies \int_{U^0} \mathrm{d}Q_c(u) = 0$. The final results follow from the linearity of integration and the fact that $\int \mathrm{d}Q(u) = \int \mathrm{d}(Q_c + Q_s)(u) = 1$. We have that $P$ is also absolutely continuous with respect to $Q_c$. To see this, note that for any measurable set $A$ we have

$$Q_c(A) = 0 \iff Q(A\setminus U^0) = 0 \implies |P|(A\setminus \Xi^0) = 0 \iff |P|(A) = 0.$$

The first equivalence holds as $Q_c(A) = Q_c(A\setminus U^0) + Q_c(A\cap U^0) = Q_c(A\setminus U^0) + Q_s(A\setminus U^0) = Q(A\setminus U^0) \geq 0$ for any measurable set $A$ as we have indeed $Q_c \ll P$ and hence $|P|(A\cap U^0) = 0$ and $|P|(A\cap U^0) = 0 \implies Q_c(A\cap U^0) = 0$ and $Q_s(A\setminus U^0) = 0$. This reasoning also guarantees that $Q_c \in \mathcal{P}_+(U)$ and $Q_c(U) = Q(U\setminus U^0) \in [0,1]$ and hence $q = 1 - Q_c(U) \geq 0$. The subsequent implication follows from $P \ll Q$ and the last equivalence follows from $|P|(A\setminus U^0) = |P|(A) - P(A\cap U^0) = P(A)$ as indeed $P(A\cap U^0) = 0$. In summary, the above reasoning ensures that $P \ll Q_c \ll P$.

Using the Radon-Nikodym theorem once again, we have for all measurable sets $A \subseteq \Xi$ that

$$\int_A \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u)\,\mathrm{d}Q_c(u) = P(A) = \int_A \frac{\mathrm{d}P}{\mathrm{d}Q}(u)\,\mathrm{d}Q(u) = \int_{A\setminus U^0} \frac{\mathrm{d}P}{\mathrm{d}Q}(u)\,\mathrm{d}Q(u) = \int_A \frac{\mathrm{d}P}{\mathrm{d}Q}(u)\,\mathrm{d}Q_c(u),$$

where the last two equalities hold because $\mathrm{d}P/\mathrm{d}Q(u) = 0$ for all $u \in U^0$ and because as we shown before $Q_c(A\cap U^0) = 0$. As the above equality holds for all Borel sets $A \subseteq U$, we find

$$\frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) = \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \quad Q_c\text{-a.s.} \implies \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) = \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \quad P\text{-a.s.}$$

as $P \ll Q_c$. The last identity ensures that

$$\int \frac{\mathrm{d}P}{\mathrm{d}Q}(u) - 1 \,\mathrm{d}P(u) = \int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1 \,\mathrm{d}P(u) \leq r.$$

Thus, the constructed point $(q, Q_c)$ is feasible in (30). Furthermore,

$$\int \ell(u) \,\mathrm{d}Q(u) = \int \ell(u) \,\mathrm{d}Q_c(u) + \int \ell(u) \,\mathrm{d}Q_s(u) \leq \int \ell(u) \,\mathrm{d}Q_c(u) + q \max_U \ell.$$

In summary, we have thus shown that for every $P$ feasible in (29) there exists $(q, Q_c)$ feasible in (30) with the same or with a larger objective value. This implies that (30) provides an upper bound on (29). □

Problem (30) may become infeasible in that its feasible set is empty, i.e.,

$$\left\{ Q_c \in \mathcal{P}_+(U),\ q \in \mathrm{R}_+ \ : \ \begin{array}{c} P \ll Q_c,\ Q_c \ll P, \\ \int \mathrm{d}Q_c(u) + q = 1, \\ \int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1 \,\mathrm{d}P(u) \leq r \end{array} \right\} = \emptyset.$$

By convention we say that the maximum in Problem (30) is unbounded from below and hence takes the value negative infinity. The following lemma gives a necessary condition on the radius $r \geq 0$ for Problem (30) to be feasible.

**Lemma B.2.** *Problem (30) is feasibile only if* $\sqrt{r+1} \geq \int \mathrm{d}\,|P|\,(u) \geq 1$ *for* $P \in \mathcal{P}_1(U)$.

*Proof.* Remark that by Jensen's inequality we have

$$\frac{\int \left(\frac{\mathrm{d}Q_c}{\mathrm{d}|P|}(u)\right)^{-1} \mathrm{d}\,|P|\,(u)}{\int \mathrm{d}\,|P|\,(u)} \geq \left( \frac{\int \frac{\mathrm{d}Q_c}{\mathrm{d}|P|}(u) \,\mathrm{d}\,|P|\,(u)}{\int \mathrm{d}\,|P|\,(u)} \right)^{-1} = \frac{\int \mathrm{d}\,|P|\,(u)}{\int \mathrm{d}Q_c(u)}$$

for all $Q_c \in \mathcal{P}(U)$ and $P \in \mathcal{P}_1(U)$. Hence,

$$\begin{aligned}
&\min \left\{ \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u) \,\mathrm{d}P(u) \ : \ q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ P \ll Q_c,\ Q_c \ll P,\ \int \mathrm{d}Q_c(u) + q = 1 \right\} \\
&= \min \left\{ \int \tfrac{\mathrm{d}|P|}{\mathrm{d}Q_c}(u) \,\mathrm{d}\,|P|\,(u) \ : \ q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ P \ll Q_c,\ Q_c \ll P,\ \int \mathrm{d}Q_c(u) + q = 1 \right\} \\
&\geq \min \left\{ \frac{\left(\int \mathrm{d}|P|(u)\right)^2}{\int \mathrm{d}Q_c(u)} \ : \ q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ P \ll Q_c,\ Q_c \ll P,\ \int \mathrm{d}Q_c(u) + q = 1 \right\} \\
&\geq \left(\int \mathrm{d}\,|P|\,(u)\right)^2.
\end{aligned}$$

From this the statement readily follows as Problem (30) is feasible if and only if

$$\min\left\{\int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u)\,\mathrm{d}P(u) \;:\; q\in \mathrm{R}_+,\; Q_c\in\mathcal{P}_+(U),\; P\ll Q_c,\; Q_c\ll P,\; \int \mathrm{d}Q_c(u)+q=1\right\}\leq r+1$$

and hence it follows that we must also have $1\leq \left[\int \mathrm{d}\,|P|\,(u)\right]^2\leq r+1$ which completes the proof. □

## B.2 Continuous Neyman Divergence Optimization

If $\epsilon'=0$ we can show that problem (30) can be reduced to an optimization problem only over distributions which are continuous with respect to $P$.

**Lemma B.3.** *Let $\epsilon'=0$. Then, we have that*

$$\begin{aligned}(30) = \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u)\\ \text{s.t. }& Q_c\in\mathcal{P}(U),\; P\ll Q_c,\; Q_c\ll P\\ & \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u)-1\,\mathrm{d}P(u)\leq r\end{aligned}\tag{31}$$

*for all $P\in\mathcal{P}_1(U)$.*

*Proof.* First, for any $\epsilon'\geq 0$ we have that

$$\begin{aligned}(30) \geq \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u)\\ \text{s.t. }& Q_c\in\mathcal{P}(U),\; P\ll Q_c,\; Q_c\ll P\\ & \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u)-1\,\mathrm{d}P(u)\leq r\end{aligned}$$

as for any feasible solution $Q_c$ in the supremum problem in Equation (31) we have that the measure $Q_c$ with $q=0$ is feasible in the supremum problem (30) and attains the same objective value.

Second, we will now show that for $\epsilon'=0$ we also have

$$\begin{aligned}(30) \leq \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u)\\ \text{s.t. }& Q_c\in\mathcal{P}(U),\; P\ll Q_c,\; Q_c\ll P\\ & \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u)-1\,\mathrm{d}P(u)\leq r\end{aligned}$$

establishing the claim. Consider any indeed any feasible point $(Q_c,q)$ with $q>0$ in the supremum problem (30) and consider a sequence of distributions $W_c^k\in\mathcal{P}(U)$, $W_c^k\ll P$ feasible in problem (28) for which have

$\lim_{k\to\infty} \int \ell(u)\,\mathrm{d}W_c^k(u) = \operatorname{ess\,sup}_P \ell$. Construct now the sequence of distributions

$$Q_c^k := Q_c + q(|P|\,/(k \int \mathrm{d}\,|P|\,(u)) + (1 - 1/k)W_c^k)$$

for $k \geq 1$. We remark that $\int \mathrm{d}Q_c^k(u) = \int \mathrm{d}Q_c(u) + q(\int \mathrm{d}\,|P|\,(u)/(k \int \mathrm{d}\,|P|\,(u)) + (1-1/k) \int \mathrm{d}W_c^k(u)) = \mathrm{d}Q_c + q = 1$. Furthermore, we have that $Q_c^k \ll P$ as both $Q_c \ll P$ and $|P| \ll P$. Similarly, we have $P \ll Q_c^k$ as $P \ll Q_c$. Hence $Q_c^k$ is feasible for all $k$ in the maximization problem on the right hand side in Equation (31) and, as we have that

$$\begin{aligned}
\lim_{k\to\infty} \int \ell(u)\,\mathrm{d}Q_c^k(u) &= \int \ell(u)\,\mathrm{d}Q_c(u) + q\left(\tfrac{1}{k} \int \ell(u)\,\mathrm{d}\,|P|\,(u) / \int \mathrm{d}\,|P|\,(u) + (1 - \tfrac{1}{k}) \int \ell(u)\,\mathrm{d}W_c^k(u)\right) \\
&= \int \ell(u)\,\mathrm{d}Q_c(u) + q\left(\lim_{k\to\infty} \int \ell(u)\,\mathrm{d}W_c^k(u)\right) \\
&= \int \ell(u)\,\mathrm{d}Q_c(u) + q \operatorname{ess\,sup}_P \ell,
\end{aligned}$$

converges also to the same objective value if $\epsilon' = 0$. □

When $\epsilon' > 0$, Equation (31) fails to hold in general and we can only guarantee the inequality

$$\begin{aligned}
(30) \geq \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u) \\
& \text{s.t. } Q_c \in \mathcal{P}(U),\ P \ll Q_c,\ Q_c \ll P \\
& \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r
\end{aligned}$$

However, as the next lemma indicates the approximation error can bounded in terms of the radius $r \geq 0$. In particular, the approximation error vanishes in the regime where $r \downarrow 0$.

**Lemma B.4.** *We have*

$$\begin{aligned}
(30) \leq \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u) + \epsilon' r/(r+1) \\
& \text{s.t. } Q_c \in \mathcal{P}(U),\ P \ll Q_c,\ Q_c \ll P, \\
& \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r.
\end{aligned} \tag{32}$$

*for all $P \in \mathcal{P}_1(U)$ and $\epsilon' \geq 0$.*

*Proof.* First, we have that

$$\begin{aligned}
\sup\ & q\epsilon' \\
\text{s.t.}\ & q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ Q_c \ll P,\ P \ll Q_c,\ q + \int \mathrm{d}Q_c(u) = 1, \\
& \int \tfrac{\mathrm{d}P}{\mathrm{d}Q}(u) - 1\,\mathrm{d}P(u) \leq r
\end{aligned}$$

$$
\begin{aligned}
&= \sup \quad q\epsilon' \\
&\quad \text{s.t.} \quad q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ Q_c \ll P,\ P \ll Q_c,\ \textstyle\int \mathrm{d}Q_c(u)/(1-q) = 1, \\
&\qquad\quad \textstyle\int \frac{\mathrm{d}P}{\mathrm{d}Q_c/(1-q)}(u)\,\mathrm{d}P(u) \leq (r+1)(1-q) \\
&= \sup \quad q\epsilon' \\
&\quad \text{s.t.} \quad q \in \mathrm{R}_+,\ \tilde{Q}_c \in \mathcal{P}_+(U),\ \tilde{Q}_c \ll P,\ P \ll \tilde{Q}_c,\ \textstyle\int \mathrm{d}\tilde{Q}_c(u) = 1, \\
&\qquad\quad \textstyle\int \frac{\mathrm{d}P}{\mathrm{d}\tilde{Q}_c}(u)\,\mathrm{d}P(u) \leq (r+1)(1-q) \\
&= \sup \quad q\epsilon' \\
&\quad \text{s.t.} \quad q \in \mathrm{R}_+,\ \tilde{Q}_c \in \mathcal{P}(U),\ \tilde{Q}_c \ll P,\ P \ll \tilde{Q}_c, \\
&\qquad\quad \textstyle\int \frac{\mathrm{d}P}{\mathrm{d}\tilde{Q}_c}(u) - 1\,\mathrm{d}P(u) \leq (r+1)(1-q) - 1 \\
&= \sup \quad q\epsilon' \\
&\quad \text{s.t.} \quad q \in \mathrm{R}_+,\ \tilde{Q}_c \in \mathcal{P}(U),\ \tilde{Q}_c \ll P \\
&\qquad\quad \chi^2(P, \tilde{Q}_c) \leq (r+1)(1-q) - 1 \\
&\leq \max \quad q\epsilon' \\
&\quad \text{s.t.} \quad q \geq 0,\ \textstyle\int 0 \leq (r+1)(1-q) - 1 \\
&= \epsilon' \frac{r}{r+1}
\end{aligned}
$$

The first equality follows by observing that $0 \leq q < 1$. Indeed, suppose that $q = 1$ then we have that $Q_c(U) = 0$ which contradicts with $P \ll Q_c$ as we indeed must have $|P|(U) \geq P(U) = 1 \implies Q_c(U) > 0$. The second equality follows from the substitution $\tilde{Q}_c = Q_c/(1-q) \in \mathcal{P}_+(U)$. The third inequality follows from $\int \mathrm{d}P(u) = 1$ as $P \in \mathcal{P}_1(U)$. The penultimate inequality follows from observing that the Neyman divergence is always positive.

The claim then follows by simply remarking that

$$
\begin{aligned}
(30) &\leq \left\{ \begin{array}{ll} \sup & \int \ell(u)\,\mathrm{d}Q_c(u) + q\,\mathrm{ess\,sup}_P\,\ell \\ \text{s.t.} & q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ P \ll Q_c,\ Q_c \ll P, \\ & \int \mathrm{d}Q_c(u) + q = 1, \\ & \int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r \end{array} \right. + \left\{ \begin{array}{ll} \sup & q\epsilon' \\ \text{s.t.} & q \in \mathrm{R}_+,\ Q_c \in \mathcal{P}_+(U),\ P \ll Q_c,\ Q_c \ll P, \\ & \int \mathrm{d}Q_c(u) + q = 1, \\ & \int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r \end{array} \right. \\
&\leq \left\{ \begin{array}{ll} \sup & \int \ell(u)\,\mathrm{d}Q_c(u) + \epsilon' r/(r+1) \\ \text{s.t.} & Q_c \in \mathcal{P}(U),\ P \ll Q_c,\ Q_c \ll P, \\ & \int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r. \end{array} \right.
\end{aligned}
$$

where the second inequality follows from Lemma B.3. □

Assume here that $P \in \mathcal{P}(U)$ is a probability measure on $U$. We clearly have now that

$$\begin{aligned} \int \ell(u)\,\mathrm{d}P(u) \leq \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u) \\ & \text{s.t. } Q_c \in \mathcal{P}(U), \\ & \quad P \ll Q_c,\ Q_c \ll P, \\ & \quad \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r \end{aligned} \tag{33}$$

as indeed the distribution $Q_c = P$ is feasible in the supremum problem (33). Moreover, because of $\int \frac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq 0 \iff P = Q_c$ inequality (33) becomes an equality when $r = 0$. The following result illustrates that even when $r > 0$ the approximation error can be controlled.

**Lemma B.5.** *Let $\ell : U \to [0, L]$ be an arbitrary measurable function. We have*

$$\begin{aligned} \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u) && \leq \int \ell(u)\,\mathrm{d}P(u) + \tfrac{L}{2}\sqrt{r} \\ & \text{s.t. } Q_c \in \mathcal{P}(U), \\ & \quad P \ll Q_c,\ Q_c \ll P, \\ & \quad \int \tfrac{\mathrm{d}P}{\mathrm{d}Q_c}(u) - 1\,\mathrm{d}P(u) \leq r \end{aligned}$$

*for all $P \in \mathcal{P}(U)$ and $r \geq 0$.*

*Proof.* Consider

$$\begin{aligned} \sup_{\ell:U\to[0,L]} \sup\ & \int \ell(u)\,\mathrm{d}Q_c(u) - \int \ell(u)\,\mathrm{d}P(u) \\ & \text{s.t. } Q_c \in \mathcal{P}(U), \\ & \quad P \ll Q_c,\ Q_c \ll P, \\ & \quad \int \left(\tfrac{\mathrm{d}Q_c}{\mathrm{d}P}(u)\right)^{-1} \mathrm{d}P(u) \leq r + 1. \end{aligned} \tag{34}$$

where $(\mathrm{d}Q_c/\mathrm{d}P(u))^{-1} = \mathrm{d}P/\mathrm{d}Q_c(u)$ as we have $P \ll Q_c$ and $Q_c \ll P$ and $P \in \mathcal{P}(U)$. We shall prove the statement by showing that $(34) \leq L\sqrt{r}/2$. It is clear that $\ell^\star(u) = L\mathbb{1}\{\mathrm{d}P/\mathrm{d}Q_c(u) \leq 1\}$ and hence it is sufficient to consider $\ell : U \to \{0, L\}$. Hence, we have

$$\begin{aligned} (34) = \sup_{\ell:U\to\{0,L\}} \sup\ & L\textstyle\int_{\ell(u)=L} \mathrm{d}Q_c(u) - L\int_{\ell(u)=L} \mathrm{d}P(\xi) \\ & \text{s.t. } Q_c \in \mathcal{P}(U), \\ & \quad P \ll Q_c,\ Q_c \ll P, \\ & \quad \textstyle\int_{\ell(u)=0} (\frac{\mathrm{d}Q_c}{\mathrm{d}P}(u))^{-1}\,\mathrm{d}P(u) + \int_{\ell(u)=L} (\frac{\mathrm{d}Q_c}{\mathrm{d}P}(u))^{-1}\,\mathrm{d}P(u) \leq r + 1. \end{aligned}$$

Furthermore, using Jensen's inequality applied to the convex function $t \mapsto t^{-1}$ when $t \geq 0$ we get

$$\int_{\ell(u)=0} (\tfrac{\mathrm{d}Q_c}{\mathrm{d}P}(u))^{-1} \,\mathrm{d}P(u) \geq \left( \frac{\int_{\ell(u)=0} (\frac{\mathrm{d}Q_c}{\mathrm{d}P}(u)) \,\mathrm{d}P(u)}{\int_{\ell(u)=0} \mathrm{d}P(u)} \right)^{-1} \int_{\ell(u)=0} \mathrm{d}P(u) = \left( \frac{\int_{\ell(u)=0} \mathrm{d}Q_c(u)}{\int_{\ell(u)=0} \mathrm{d}P(u)} \right)^{-1} \int_{\ell(u)=0} \mathrm{d}P(u),$$

$$\int_{\ell(u)=L} (\tfrac{\mathrm{d}Q_c}{\mathrm{d}P}(u))^{-1} \,\mathrm{d}P(u) \geq \left( \frac{\int_{\ell(u)=L} (\frac{\mathrm{d}Q_c}{\mathrm{d}P}(u)) \,\mathrm{d}P(u)}{\int_{\ell(u)=L} \mathrm{d}P(u)} \right)^{-1} \int_{\ell(u)=L} \mathrm{d}P(u) = \left( \frac{\int_{\ell(u)=L} \mathrm{d}Q_c(u)}{\int_{\ell(u)=L} \mathrm{d}P(u)} \right)^{-1} \int_{\ell(u)=L} \mathrm{d}P(u).$$

Consequently, after relaxing the constraints $P \ll Q_c,\ Q_c \ll P$ and using the previous established pair of inequalities we get

$$\begin{aligned}
(34) \leq \sup_{\ell : U \to \{0,L\}} \sup\ & L \textstyle\int_{\ell(u)=L} \mathrm{d}Q_c(u) - L \int_{\ell(u)=L} \mathrm{d}P(u) \\
& \text{s.t. } Q_c \in \mathcal{P}(U), \\
& \left[ \left( \frac{\int_{\ell(u)=0} \mathrm{d}Q_c(u)}{\int_{\ell(u)=0} \mathrm{d}P(u)} \right)^{-1} - 1 \right] \int_{\ell(u)=0} \mathrm{d}P(u) \\
& \qquad + \left[ \left( \frac{\int_{\ell(u)=L} \mathrm{d}Q_c(u)}{\int_{\ell(u)=L} \mathrm{d}P(u)} \right)^{-1} - 1 \right] \int_{\ell(u)=L} \mathrm{d}P(u) \leq r + 1.
\end{aligned}$$

Let now $p_L := \int_{\ell(u)=L} \mathrm{d}P(u)$ for notational convenience and consider the substitution $0 \leq q_L = \int_{\ell(u)=L} \mathrm{d}Q_c(u)$ we get

$$\begin{aligned}
(34) \leq \sup_{q_L \geq 0}\ & L q_L - L p_L \\
& \text{s.t. } \left[ \tfrac{(1-p_L)}{(1-q_L)} - 1 \right] (1 - p_L) + \left[ \tfrac{p_L}{q_L} - 1 \right] p_L \leq r.
\end{aligned}$$

Hence,

$$\begin{aligned}
(34) \leq & \sup \{ L(q_L - p_L) \ :\ q_L \geq 0,\ (p_L/q_L - 1)p_L + ((1-p_L)/(1-q_L) - 1)(1-p_L) \leq r \} \\
\leq & \sup \{ L(q_L - p_L) \ :\ q_L \geq 0,\ (p_L - q_L)p_L/q_L - (p_L - q_L)(1-p_L)/(1-q_L) \leq r \} \\
\leq & \sup \{ L(q_L - p_L) \ :\ q_L \geq 0,\ (q_L - p_L)\left((1-p_L)/(1-q_L) - p_L/q_L\right) \leq r \} \\
\leq & \sup \{ L(q_L - p_L) \ :\ q_L \geq 0,\ (q_L - p_L)\left(q_L(1-p_L) - p_L(1-q_L)\right) \leq r q_L(1-q_L) \} \\
\leq & \sup \left\{ L(q_L - p_L) \ :\ q_L \geq 0,\ (q_L - p_L)^2 \leq r q_L(1-q_L) \right\} \\
\leq & \sup \left\{ L(q_L - p_L) \ :\ q_L \geq 0,\ (q_L - p_L)^2 \leq r/4 \right\} \\
\leq & \frac{L}{2}\sqrt{r}
\end{aligned}$$

which concludes the proof. □

## B.3 Duality

**Lemma B.6.** *We have*

$$\begin{aligned} (30) \leq \min \quad & \alpha - \tfrac{1}{r+1} \left( \int \sqrt{\alpha - \ell(u)} \, \mathrm{d} \left| P \right| (u) \right)^2 \\ \text{s.t.} \quad & \operatorname{ess\,sup}_P \ell + \epsilon' \leq \alpha \end{aligned} \tag{35}$$

*for all for all* $P \in \mathcal{P}_1(U)$ *and* $\epsilon' \geq 0$. *Furthermore, if the constraint qualification condition* $\sqrt{r+1} > \int \mathrm{d} \left| P \right| (u)$ *holds then inequality* (35) *becomes an equality and we show that the minimum over* $\alpha$ *is attained.*

*Proof.* Introduce for the supremum problem (30) the Lagrangian

$$\begin{aligned} L(q, Q_c \, ; \, \alpha, \gamma) &= \int \ell(u) \, \mathrm{d} Q_c(u) + q(\operatorname{ess\,sup}_P \ell + \epsilon') + \alpha \left( 1 - q - \int \mathrm{d} Q_c(u) \right) + \gamma \left( r + 1 - \int \frac{\mathrm{d} \left| P \right|}{\mathrm{d} Q_c}(u) \, \mathrm{d} \left| P \right| (u) \right) \\ &= \alpha + \gamma(r+1) + q(\operatorname{ess\,sup}_P \ell + \epsilon' - \alpha) + \int (\ell(u) - \alpha) \, \mathrm{d} Q_c(u) - \gamma \int \frac{\mathrm{d} \left| P \right|}{\mathrm{d} Q_c}(u) \, \mathrm{d} \left| P \right| (u) \\ &= \alpha + \gamma(r+1) + q(\operatorname{ess\,sup}_P \ell + \epsilon' - \alpha) + \int \left[ (\ell(u) - \alpha) \frac{\mathrm{d} Q_c}{\mathrm{d} \left| P \right|}(u) - \gamma \frac{\mathrm{d} \left| P \right|}{\mathrm{d} Q_c}(u) \right] \mathrm{d} \left| P \right| (u). \end{aligned}$$

We trivially have that $(30) = \sup_{q \geq 0, \, Q_c \in \mathcal{P}_+(U), \, P \ll Q_c, \, Q_c \ll P} \inf_{\alpha \in \mathrm{R}, \, \gamma \geq 0} L(q, Q_c; \alpha, \gamma)$. The associated dual function is

$$\begin{aligned} & g(\alpha, \gamma) \\ & := \sup \left\{ L(q, Q_c \, ; \, \alpha, \gamma) \ : \ q \geq 0, \ Q_c \in \mathcal{P}_+(U), \ P \ll Q_c, \ Q_c \ll P \right\} \\ & = \alpha + \gamma(r+1) + \chi_\infty(\alpha \geq \operatorname{ess\,sup}_P \ell + \epsilon') \\ & \quad + \sup_{Q_c \in \mathcal{P}_+(U), \, P \ll Q_c \ll P} \int (\ell(u) - \alpha) \frac{\mathrm{d} Q_c}{\mathrm{d} \left| P \right|}(u) - \gamma \frac{\mathrm{d} \left| P \right|}{\mathrm{d} Q_c}(u) \, \mathrm{d} \left| P \right| (u) \\ & = \alpha + \gamma(r+1) + \chi_\infty(\alpha \geq \operatorname{ess\,sup}_P \ell + \epsilon') + \int \left[ \max_{\lambda \geq 0} (\ell(u) - \alpha)\lambda - \gamma \lambda^{-1} \right] \mathrm{d} \left| P \right| (u) \\ & = \alpha + \gamma(r+1) + \chi_\infty(\alpha \geq \operatorname{ess\,sup}_P \ell + \epsilon') - 2 \int \sqrt{\gamma(\alpha - \ell(u))} \, \mathrm{d} \left| P \right| (u). \end{aligned}$$

The third equality follows from Rockafellar and Wets (2009, Theorem 14.60). The fourth equality follows from elementary manipulations where

$$\lambda^\star \in \arg\max_\lambda (\ell(u) - \alpha)\lambda - \gamma \lambda^{-1} \iff (\ell(u) - \alpha) + \gamma(\lambda^\star)^{-2} = 0 \iff \lambda^\star = \sqrt{\frac{\gamma}{\alpha - \ell(u)}}.$$

The convex dual problem is defined as

$$\begin{aligned} \min_{\alpha\in\mathrm{R},\,\gamma\geq 0} g(\alpha,\gamma) = \min \quad & \alpha+\gamma(r+1)-2\int\sqrt{\gamma(\alpha-\ell(u)}\,\mathrm{d}\,|P|\,(u) \\ \text{s.t.} \quad & \alpha\in\mathrm{R},\,\gamma\geq 0 \\ & \alpha\geq \operatorname{ess\,sup}_P \ell+\epsilon'. \end{aligned} \tag{36}$$

We now show that the dual problem (36) always admits a solution $(\alpha^\star,\gamma^\star)$ if $r>0$. First, remark that $\min_{\gamma\geq 0} g(\alpha,\gamma)$ admits a solution $\gamma^\star(\alpha)$ and can be solved analytically for any $\alpha\in\mathrm{R}$. Indeed, the first order necessary and sufficient optimality conditions are

$$\gamma^\star(\alpha)\in\arg\min_{\gamma\geq 0}\; g(\alpha,\gamma) \iff \gamma^\star(\alpha)=\left(\frac{\int\sqrt{\alpha-\ell(u)}\,\mathrm{d}\,|P|\,(u)}{r+1}\right)^2$$

for any $\alpha\in\mathrm{R}$. Elementary manipulations show that $g(\alpha,\gamma^\star(\alpha))=\alpha-(\int\sqrt{\alpha-\ell(u)}\,\mathrm{d}\,|P|\,(u))^2/(r+1)+\chi_\infty(\alpha\geq \operatorname{ess\,sup}_P\ell+\epsilon')$. From weak duality it follows that we have $\min_{\gamma\geq 0,\,\alpha\in\mathrm{R}}\; g(\alpha,\gamma)=\min_{\alpha\in\mathrm{R}}\; g(\alpha,\gamma^\star(\alpha))\geq$ (30) from which inequality (35) follows immediately.

Assume now that we have the constraint qualification condition $\sqrt{r+1}>\int\,\mathrm{d}\,|P|\,(u)\geq 1$. Using Jensen's inequality

$$g(\alpha,\gamma^\star(\alpha))\geq\alpha-\frac{\int\alpha-\ell(u)\,\mathrm{d}\,|P|\,(u)\int\,\mathrm{d}\,|P|\,(u)}{(r+1)}=\frac{r+1-\left(\int\,\mathrm{d}\,|P|\,(u)\right)^2}{r+1}\alpha+\frac{\int\ell(\xi)\,\mathrm{d}\,|P|\,(u)\int\,\mathrm{d}\,|P|\,(u)}{r+1} \tag{37}$$

for all $\alpha\geq\operatorname{ess\,sup}_P\ell+\epsilon'$. Hence, as $\alpha\mapsto g(\alpha,\gamma^\star(\alpha))$ is continuous and finite valued when $\alpha\geq\operatorname{ess\,sup}_P\ell+\epsilon'$ and grows unbounded as $\alpha$ tends to infinity the minimum $\min_{\alpha\geq\operatorname{ess\,sup}_P\ell+\epsilon'} g(\alpha,\gamma^\star(\alpha))$ will be attained at some $\alpha^\star\geq\operatorname{ess\,sup}_P\ell+\epsilon'$.

The minimization problem (36) satisfies Slater's constraint qualification because its convex objective function is finite-valued on the feasible set and because the decision variables are only subject to lower bounds. Thus, $(\alpha^\star,\gamma^\star)$ solves (36) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned} &\alpha\geq\operatorname{ess\,sup}_P\ell+\epsilon',\ \gamma\geq 0, && \text{(primal feasibility)} \\ &\lambda\geq 0, && \text{(dual feasibility)} \\ &\int\sqrt{\tfrac{\gamma}{\alpha-\ell(u)}}\,\mathrm{d}\,|P|\,(\xi)+\lambda=1, && \text{(stationarity with respect to } \alpha) \\ &\int\left(\sqrt{\tfrac{\gamma}{\alpha-\ell(u)}}\right)^{-1}\mathrm{d}\,|P|\,(u)=r+1, && \text{(stationarity with respect to } \gamma) \\ &\lambda(\alpha-\operatorname{ess\,sup}\ell-\epsilon')=0, && \text{(complementary slackness)} \end{aligned}$$

where $\lambda$ represents the Lagrange multiplier associated with the dual constraint $\alpha\geq\operatorname{ess\,sup}_P\ell+\epsilon'$.

From weak duality it follows that we have $\min_{\gamma\geq 0,\,\alpha\in\mathrm{R}} g(\alpha,\gamma) \geq (30)$. We will show now that in fact strong duality, i.e., $\min_{\gamma\geq 0,\,\alpha\in\mathrm{R}} g(\alpha,\gamma) \leq (30)$, holds as well assuming again that $\sqrt{r+1} > \int \mathrm{d}\,|P|\,(u) \geq 1$. Given any dual optimal solution $(\alpha^\star,\nu^\star,\lambda^\star)$ satisfying the KKT conditions, we can now introduce a Borel-measurable function $\Lambda^\star(u) = \sqrt{\frac{\gamma^\star}{\alpha^\star-\ell(u)}}$ and define $q^\star = \lambda^\star$ and $Q_c^\star$ through $\mathrm{d}Q_c^\star/\mathrm{d}\,|P|(u) = \Lambda^\star(u)$. The two stationarity conditions thus imply that $(Q_c^\star, q^\star)$ is feasible in (30). Moreover, we have

$$\begin{aligned}
g(\alpha^\star,\gamma^\star) &= \alpha^\star + \gamma^\star(r+1) - \int \sqrt{\gamma^\star(\alpha^\star-\ell(u))}\,\mathrm{d}\,|P|\,(u) - \int \sqrt{\gamma^\star(\alpha^\star-\ell(u))}\,\mathrm{d}\,|P|\,(u) \\
&= \alpha^\star - \int \sqrt{\gamma^\star(\alpha^\star-\ell(u))}\,\mathrm{d}\,|P|\,(u) \\
&= \alpha^\star + \int (\ell(u)-\alpha^\star)\sqrt{\frac{\gamma^\star}{\alpha^\star-\ell(u)}}\,\mathrm{d}\,|P|\,(u) \\
&= \alpha^\star + \int \ell(u)\Lambda^\star(u)\,\mathrm{d}\,|P|\,(u) - \alpha^\star(1-\lambda^\star) \\
&= \int \ell(u)\,\mathrm{d}Q_c^\star(\xi) + q^\star\left(\operatorname{ess\,sup}_P \ell + \epsilon'\right)
\end{aligned}$$

where the first equality follows from the definition of $g$, the second equality holds due to the stationarity condition for $\gamma$, the fourth equality follows from the definition of $\Lambda^\star$ and the stationarity condition for $\alpha$, and the last equality exploits the complementary slackness condition as well as the definition of $q^\star$. We have thus shown that the objective value of $(Q_c^\star, p^\star)$ in (30) coincides with the objective value of $(\alpha^\star,\gamma^\star)$ in (36), which certifies that the duality gap between (30) and (36) is zero. □

Let us consider the dual problem

$$\begin{array}{ll}
\min & \alpha - \frac{1}{r+1}\left(\int \sqrt{\alpha-\ell(u)}\,\mathrm{d}\,|P|\,(u)\right)^2 \\
\text{s.t.} & \operatorname{ess\,sup}_P \ell + \epsilon' \leq \alpha
\end{array} \tag{38}$$

Like all dual problems, problem (38) is convex and as it is univariate can be efficiently solved using a bisection method when we assume its gradient function

$$\alpha \mapsto 1 - \frac{1}{r+1}\left(\int \sqrt{\alpha-\ell(u)}\,\mathrm{d}\,|P|\,(u)\right)\left(\int \frac{1}{\sqrt{\alpha-\ell(u)}}\,\mathrm{d}\,|P|\,(u)\right)$$

can be efficiently evaluated for $\alpha \geq \operatorname{ess\,sup}_P \ell + \epsilon'$. However, to initialize the bisection method a lower and upper bound on the optimal $\alpha$ is required. The subsequent lemma makes such bounds available when $P = |P| \in \mathcal{P}(U)$ and $r > 0$.

**Lemma B.7.** *Let $P \in \mathcal{P}(U)$, $r > 0$ and $\ell : U \to [0, L]$. Then, the minimum in Problem (38) is attained at $\alpha^\star$*

*satisfying*

$$\operatorname{ess\,sup}_P \ell + \epsilon' \leq \alpha^\star \leq \frac{L}{2}\frac{r+1}{\sqrt{r}} + \epsilon'.$$

*Proof.* Let here $g(\alpha) = \alpha - (\int \sqrt{\alpha - \ell(u)}\,\mathrm{d}P(u))^2/(r+1)$ for $\alpha \geq \operatorname{ess\,sup}_P \ell + \epsilon'$. Using Jensen's inequality it holds that

$$\begin{aligned} g(\alpha) &\geq \alpha - \frac{\int \alpha - \ell(u)\,\mathrm{d}P(u) \int \mathrm{d}P(u)}{(r+1)} \\ &= \alpha - \frac{\int \alpha - \ell(u)\,\mathrm{d}P(u)}{(r+1)} \\ &= \frac{r}{r+1}\alpha + \frac{\int \ell(u)\,\mathrm{d}P(u)}{r+1} \\ &\geq \frac{r}{r+1}\alpha + \int \ell(u)\,\mathrm{d}P(u) \end{aligned}$$

for all $\alpha \geq \operatorname{ess\,sup} \ell + \epsilon'$. Here the second inequality uses that $P \in \mathcal{P}(U)$. The third inequality uses that $r > 0$. As $g$ is continuous and grows unbounded as $\alpha$ tends to infinity there exists $\alpha^\star \in \arg\min \{g(\alpha) \ : \ \alpha \geq \operatorname{ess\,sup} \ell + \epsilon'\}$ when $r > 0$.

Furthermore, we have

$$g(\alpha^\star) = (30) \leq \int \ell(u)\,\mathrm{d}P(u) + \frac{L}{2}\sqrt{r} + \epsilon'\frac{r}{r+1}$$

following the strong duality result in Lemma B.6 as $\sqrt{r+1} > \int \mathrm{d}\,|P|\,(u) = \int \mathrm{d}P(u) = 1$ and the approximation results in Lemmas B.4 and B.5, respectively.

Finally, we have hence

$$\begin{aligned} &\frac{r}{r+1}\alpha^\star + \int \ell(\xi)\,\mathrm{d}\,|P|\,(\xi) \leq g(\alpha^\star) \leq \int \ell(\xi)\,\mathrm{d}\,|P|\,(\xi) + \frac{L}{2}\sqrt{r} + \epsilon'\frac{r}{r+1} \\ \iff &\frac{r}{r+1}\alpha^\star \leq \frac{L}{2}\sqrt{r} + \epsilon'\frac{r}{r+1} \\ \iff &\alpha^\star \leq \frac{L}{2}\frac{r+1}{\sqrt{r}} + \epsilon' \end{aligned}$$

establishing the claim. Remark that as $\min_{r\geq 0} r + 1/\sqrt{r} = 2$ and $\operatorname{ess\,sup}_P \ell \leq L$ we have that $\operatorname{ess\,sup}_P \ell + \epsilon' \leq L(r+1)/(2\sqrt{r}) + \epsilon'$. □

**Lemma B.8.** *Let $\ell$ be an $\mathcal{L}$-Lipschitz and $L$-bounded function. Then, we have*

$$\begin{aligned}
&\int \sqrt{\alpha - \ell(u)}\,\mathrm{d}P(u) + \frac{\mathcal{L}\delta}{2\sqrt{\alpha - L}} + \left(\frac{\mathcal{L}\,\mathrm{diam}(U)}{2\sqrt{\alpha - L}} + \sqrt{\alpha}\right)(\sqrt{r+1} - 1) \\
&\geq \begin{cases} \sup & \int \sqrt{\alpha - \ell(u)}\,\mathrm{d}\,|P'|\,(u) \\ \text{s.t.} & P' \in \mathcal{P}_1(U), \\ & \int \mathrm{d}\,|P'|\,(u) \leq \sqrt{r+1},\ \ \|P' - P\| \leq \delta \end{cases}
\end{aligned}$$

*for all $P \in \mathcal{P}_1(U)$ and $\alpha > \max_U \ell$.*

*Proof.* First, the function $t \mapsto \sqrt{\alpha - t}$ is a positive $1/(2\sqrt{\epsilon})$-Lipschitz continuous function on $t \leq \alpha - \epsilon$ for any $\epsilon > 0$. Hence, the integrant $u \mapsto \sqrt{\alpha - \ell(u)}$ is a $\mathcal{L}/(2\sqrt{\alpha - \max_U \ell})$-Lipschitz function over $U$.

Second, observe that

$$\begin{aligned}
&\|P' - P\| \leq \delta \\
\iff &\sup\left\{\textstyle\int f(u)\,\mathrm{d}P'(u) - \int f(u)\,\mathrm{d}P(u)\ :\ f : U \to \mathrm{R},\ \forall u_1, u_2 \in U :\ \|f(u_1) - f(u_2)\| \leq 1\right\} \leq \delta \\
\iff &\sup\left\{\textstyle\int f(u)\,\mathrm{d}P'(u) - \int f(u)\,\mathrm{d}P(u)\ :\ f : U \to [0, \mathrm{diam}(U)],\ \forall u_1, u_2 \in U :\ \|f(u_1) - f(u_2)\| \leq 1\right\} \leq \delta \\
\implies &\sup\left\{\textstyle\int f(u)\,\mathrm{d}(|P'| - P)(u)\ :\ f : U \to [0, \mathrm{diam}(U)],\ \forall u_1, u_2 \in U :\ \|f(u_1) - f(u_2)\| \leq 1\right\} \\
&\qquad\qquad \leq \delta + \mathrm{diam}(U)(\sqrt{r+1} - 1) \quad (39)
\end{aligned}$$

Here, the first equivalence is by definition. The second equivalence follows by the observation that as $U$ is compact any continuous function $f : U \to \mathrm{R}$ has a finite minimum $\min_{u \in U} f(u)$. Hence, for any $f : U \to \mathrm{R}$ we can consider $\tilde{f} : U \to [0, \mathrm{diam}(U)]$, $\tilde{f}(u) = f(u) + \min_{u \in U} f(u)$ for which we have

$$\textstyle\int f(u)\,\mathrm{d}P(u)' - \int f(u)\,\mathrm{d}P(u) = \int \tilde{f}(u)\,\mathrm{d}P'(u) - \int \tilde{f}(u)\,\mathrm{d}P(u)$$

as both $P'$ and $P$ are in $\mathcal{P}_1(U)$. The final implication follows from

$$\begin{aligned}
\int f(u)\,\mathrm{d}P'(u) &= \int f(u)\,\mathrm{d}P'_+ - \int f(u)\,\mathrm{d}P'_- \\
&= \int f(u)\,\mathrm{d}P'_+ + \int f(u)\,\mathrm{d}P'_- - 2\int f(u)\,\mathrm{d}P'_- \\
&= \int f(u)\mathrm{d}\,|P|\,(u) - 2\int f(u)\mathrm{d}P'_- \\
&\geq \int f(u)\mathrm{d}\,|P|\,(u) - 2\,\mathrm{diam}(U)\int \mathrm{d}P'_- \\
&\geq \int f(u)\mathrm{d}\,|P|\,(u) - \mathrm{diam}(U)(\sqrt{r+1} - 1).
\end{aligned}$$

where the last inequality follows from $\int \mathrm{d}P'_+ - \int \mathrm{d}P'_- = 1$, $\int \mathrm{d}P'_+ + \int \mathrm{d}P'_- \leq \sqrt{r+1} \implies \int \mathrm{d}P'_- \leq (\sqrt{r+1}-1)/2$ with $P'_+$ and $P'_-$ positive measures so that $P' = P'_+ - P'_-$ (Halmos 2013, Section 29).

We have that

$$
\begin{aligned}
&\sup\left\{\int \sqrt{\alpha-\ell(u)}\,\mathrm{d}(|P'|-P)(u) \;:\; P'\in\mathcal{P}_1(U),\, \|P'-P\|\leq\delta,\, \int \mathrm{d}\,|P'|\,(u)\leq\sqrt{r+1}\right\}\\
&=\begin{cases}\sup & \mathcal{L}/(2\sqrt{\alpha-\max_U \ell}) \displaystyle\int \frac{\sqrt{\alpha-\ell(u)}-\sqrt{\alpha-\max_U \ell}}{\mathcal{L}/(2\sqrt{\alpha-\max_U \ell})}\,\mathrm{d}(|P'|-P)(u) + \int \sqrt{\alpha-\max_U \ell}\,\mathrm{d}(|P'|-P)(u)\\ \text{s.t.} & P'\in\mathcal{P}_1(U),\\ & \int \mathrm{d}\,|P'|\,(u)\leq\sqrt{r+1},\\ & \|P'-P\|\leq\delta\end{cases}\\
&\leq \mathcal{L}/(2\sqrt{\alpha-\max_U \ell})\left(\delta+\mathrm{diam}(U)(\sqrt{r+1}-1)\right)+\sqrt{\alpha-\max_U \ell}(\sqrt{r+1}-1)\\
&=\tfrac{\mathcal{L}\delta}{2\sqrt{\alpha-\max_U \ell}}+\left(\tfrac{\mathcal{L}\,\mathrm{diam}(U)}{2\sqrt{\alpha-\max_U \ell}}+\sqrt{\alpha-\max_U \ell}\right)(\sqrt{r+1}-1)
\end{aligned}
$$

which establishes the claim. The first equality follows by subtracting adding and adding $\int \sqrt{\alpha-\max_U \ell}\,\mathrm{d}(|P'|-P)(u)$ and dividing and multiplying by $\mathcal{L}/(2\sqrt{\alpha-\max_U \ell}$. Observe that

$$
u \mapsto \frac{\sqrt{\alpha-\ell(u)}-\sqrt{\alpha-\max_U \ell}}{\mathcal{L}/(2\sqrt{\alpha-\max_U \ell})} \in [0,\mathrm{diag}(U)]
$$

is a 1-Lipschitz function on $U$. Hence, following (39) we have that

$$
\begin{aligned}
&\begin{cases}\sup & \int \frac{\sqrt{\alpha-\ell(u)}-\sqrt{\alpha-\max_U \ell}}{\mathcal{L}/(2\sqrt{\alpha-\max_U \ell})}\,\mathrm{d}(|P'|-P)(u)\\ \text{s.t.} & P'\in\mathcal{P}_1(U),\\ & \|P'-P\|\leq\delta\end{cases}\\
&\leq \delta+\mathrm{diam}(U)(\sqrt{r+1}-1)
\end{aligned}
$$

which together with $\int \sqrt{\alpha-\max_U \ell}\,\mathrm{d}(|P'|-P)(u) = \sqrt{\alpha-\max_U \ell}\int \mathrm{d}(|P'|-P)(u) \leq \sqrt{\alpha-\max_U \ell}(\sqrt{r+1}-1)$ establishes the second inequality. The final equality follows from elementary manipulations. □

# C Subgradient Calculus

**Lemma C.1** ((Rockafellar and Wets 2009, Theorem 10.13))**.** *Let $g : X \times Y \to \mathrm{R}$ be a lower semicontinous convex function with $X$ and $Y$ bounded and closed convex sets. Define the convex function $f(x) = \min_{y\in Y} g(x,y)$. Let $\bar{y}$ be such that $f(\bar{x}) = g(\bar{x},\bar{y})$. Then, $\partial_x f(\bar{x}) = \{g \;:\; (g,0)\in\partial_{x,y} g(\bar{x},\bar{y})\}$.*

*Proof.* The extended value function $\bar{f}(x,y) = f(x,y) + \chi_\infty(x \in X, y \in Y)$ is convex lower semi-continuous and clearly level bounded in $x$ (locally) uniformly in $y$. Rockafellar and Wets (2009, Theorem 10.13) completes the claim. □

**Lemma C.2** ((Bertsekas 1971, Proposition A.22))**.** *Let $g : X \times Y \to \mathrm{R}$ be a continuous function and let $Y$ be a compact subset of $\mathrm{R}^m$. Assume that for any $y \in Y$ we have that $x \mapsto g(x,y)$ is a convex function. The convex function $f(x) = \max_{y \in Y} g(x,y)$ which is a convex function and is finite valued for all $X$. Let $\bar{y}$ be such that $f(\bar{x}) = g(\bar{x},\bar{y})$. Then, for any $\bar{x} \in \mathrm{int}(X)$ we have $\partial_x f(\bar{x}) = \mathrm{CH}(\cup_{\bar{y} \in Y^\star(x)} \partial_x g(\bar{x},\bar{y}))$ where $Y^\star(\bar{x}) = \arg\max_{y \in Y} g(\bar{x},y) \neq \emptyset$.*

*Proof.* The fact that $f(x)$ is finite valued follows from the fact that $g$ is continuous in $y$ for each $x$ and $Y$ is compact. Hence, also we have that $Y^\star(\bar{x}) = \arg\max_{y \in Y} g(\bar{x},y) \neq \emptyset$ for all $\bar{x} \in X$. The remainder of the statement follows from Bertsekas (1971, Proposition A.22). □