Variational Perspectives on Optimal Allocations

Kai Feng*

November 12, 2025

Latest version here

Abstract.

This is a preliminary draft of the paper. Many parts remain incomplete or are undergoing significant revision, and the arguments and results should be regarded as provisional. The author nonetheless hopes that readers will find the basic ideas informative and thought-provoking.

This paper presents an unified framework to the optimal allocation problems. We formulate a broad class of population-level allocation problems as optimization problems over a space of critical functions, a representation that admits strong expressive power. This formulation accommodates constrained multi-treatment policy learning and multi-class classification as special cases. Within this framework, we establish tools for statistical estimation and inference, with emphasis on constrained policy learning and classification. We draw on the marginal approach to Lagrangian duality to characterize constrained optimization and guide the construction of estimators. Building on sample analogs of the Lagrangian function, we propose and provide heuristic rationale to a triple machine learning approach for plug-in estimation.

Keywords: Optimal treatment allocation, Policy learning, Plug-in multi-classification, Debiased machine learning, Convex analysis, Hadamard differentiability

JEL Codes: C14, C38, C44

*Institute of Economics, Tsinghua University. Email: fengk22@mails.tsinghua.edu.cn.

1 Introduction

Economics is about allocation of resources which are always scarce. In practice, virtually all valuable or productive resources are limited, making scarcity a fundamental challenge in decision making. Yet scarcity alone does not capture the full complexity of achieving optimal allocation. Decision makers often lack the precise knowledge required to understand how different allocations translate into outcomes. These outcomes exhibit inherent randomness, and the underlying probability laws are rarely known with precision. In this paper, we are interesting in providing approximately optimal allocation method for scarce resources with some statistical guarantees.

Personalized allocation rules that condition on observable characteristics are central in a range of treatment assignment tasks. For instance, medical decisions routinely depend on patient-specific diagnostics, and judges tailor pretrial release decisions to defendant characteristics. The heterogeneity of potential outcomes makes personalized allocation essential when maximizing welfare.

Many problems can be formulated as optimal allocation problems. Typical examples include treatment assignment or policy learning problems, as in Kitagawa and Tetenov (2018) and Athey and Wager (2021), where resources such as medical services, financial credit, educational opportunities, or job search assistance are allocated. Other applications do not involve explicit resources but instead allocate decisions across categories under error tolerance constraints. Classification and simple hypotheses testing are representative examples. Here, category labels themselves can be interpreted as scarce resources, with scarcity arising from permissible error rates.

In this paper, We adopt a variational perspective on optimal allocation. The term "variational" reflects the optimization-based interpretation of the problem.[1] In population, a typical optimal allocation problem with utilitarian welfare can be written as

$$\gamma := \sup_{\pi \in \Pi} \mathbb{E}\left[Y\left(\pi\left(X\right)\right)\right], \ s.t.\ C\left(Q, \pi\right) \leqslant \rho, \tag{1}$$

where $C\left(Q, \pi\right)$ denotes the constraint functions of the welfare optimization problem, and π denote a policy from the candidate set Π. For the simplest binary allocation with one constraint, we can further write $(X, Y, Z) \sim Q$ and $C\left(Q, \pi\right) = \mathbb{E}\left(Z\left(\pi\left(X\right)\right)\right)$ for cost random variable Z.

Generally, the objective and constraints in (1) may vary widely across applications, The objective function may range from utilitarian to Rawlsian, or anything in between to reflect utility-minded or equality-minded goals. See, for example Kitagawa and Tetenov (2021) and Fan et al. (2025b). The constraint may reflect cost of treatment, limited resource supplement and adversarial strategic response. In the last scenario, maximin problem under distributionally robust constraints can be considered to provide performance guarantee in the worst case. In this work, we use a critical function ϕ to express some typical allocation problems in a unified framework. If the outcome Y takes value in space $\mathcal{Y}$, by the term critical function, we refer to function ϕ in the

[1] See, for example the preface of Rockafellar and Wets (2009).

dual space $\mathcal{Y}^*$ such that $\langle y, \phi \rangle \geqslant 0$ for all pointwisely positive $y \in \mathcal{Y}$, and $\langle 1, \phi \rangle = 1$.

For multi-treatment allocation (policy learning) problem with J different treatments, $\mathcal{Y} = \mathbb{R}^J$. The population objective function then becomes $\mathbb{E}\left[\sum_{j=0}^{J-1} Y_i \phi_j(X)\right]$, where $\phi = (\phi_0, \ldots, \phi_{J-1})$ satisfies $\phi_j \geqslant 1$ and $\sum_{j=0}^{J-1} \phi_j(\cdot) = 1$ (almost surely). We immediately see some advantage of this expression. This representation accommodates randomized policies and renders the objective linear in both outcomes and decision rules. More theoretical analysis about this ϕ can be found in section 4.1.

We pay special attention to the constrained multi-treatment allocation problem and multi-classification problem in this work. Using the critical function expression, these problems can be written as

$$\sup_{\phi \in \Phi} \mathbb{E}\left[\sum_{j \in \mathscr{J}} Y_j \phi_j(X)\right] \quad s.t. \quad \mathbb{E}[Z_k \phi_k] \leqslant c_k, \forall k \in \mathscr{K} \tag{2}$$

and its variants, where $\mathscr{J}, \mathscr{K} \subset \{0, \ldots, J-1\}$, and Φ is the set of all critical functions.

The optimization problem in (2) is a convex programming problem. We study it using a marginal approach to convex analysis which considers the convex programming problems in a variational way. The basis of this approach is to rewrite (2) as $\sup_{\phi \in \Phi} \mathbb{E}\left[\sum_{j \in \mathscr{J}} + \delta_{CT(c)}\right]$ where

$$\delta_{CT(c)} = \begin{cases} 0, & \text{if } \mathbb{E}[Z_k \phi_k] \leqslant c_k, \forall k \in \mathscr{K}, \\ -\infty, & \text{otherwise.} \end{cases}$$

In section 4 we see why this approach can provide more information than the Kuhn-Tucker theorem. Two illustrative examples about distributionally robust optimization are also provided there following the marginal approach to *Lagrangian duality* in full generality.

Our paper builds conceptual links across several strands of literature. One of them is the differential of probability function method—also known as the derivative as surface integral method in Chernozhukov et al. (2018b), Feng et al. (2025) and Chen and Gao (2025). This method can be traced back to Kim and Pollard (1990). These works provide powerful device for characterizing welfare perturbations from indictor functions. In Feng et al. (2025), this functional differential method is connected to the faster convergence rate phenomena for plug-in classification (Devroye et al., 1996; Audibert and Tsybakov, 2007).

In section 6, we we integrate these insights with the convex convergence approach (Pollard, 1991)[2] and debiased machine learning (Chernozhukov et al., 2018a, 2022). We propose a *triple machine learning* inference procedure for the value of multi allocation problem. The method uses a plug-in decision rule based on pre-trained score estimators, coupled with a convex optimization step over a finite-dimensional parameter space. Therefore, there are no numerical difficulty.

[2]See also, for example Fan et al. (1994) and Fan et al. (1995).

2 Literature

A multidisciplinary literature has studied the optimal allocation problem under various forms and settings. The derivation of optimal decision rules from experimental or observational data has been studied in the context of treatment allocation / assignment / choice (Stoye, 2009; Tetenov, 2012; Bhattacharya and Dupas, 2012; Kitagawa and Tetenov, 2018; Rai, 2018; Mbakop and Tabord-Meehan, 2021; Kitagawa and Tetenov, 2021; Adjaho and Christensen, 2022; Armstrong and Shen, 2023; Kitagawa and Wang, 2023a,b), policy learning (Luedtke and Chambaz, 2020; Athey and Wager, 2021; Kido, 2022; Zhou et al., 2023; Ben-Michael et al., 2024; Ai et al., 2024; Zhan et al., 2024; Viviano, 2025; Jin et al., 2025; Fan et al., 2025b) and individual treatment rule (ITR) (Qian and Murphy, 2011; Luedtke and van der Laan, 2016; Zhao et al., 2012; Zhou et al., 2017).[3] In this work, we abstract from the probably subtle terminological difference, but simply refer to the overarching problem as optimal allocation problem to emphasis its variational essence.

A large body of the literature focuses on empirical welfare maximization (EWM), the welfare-maximization analogue of of empirical risk minimization (ERM). See for example Manski (2004), Qian and Murphy (2011), Zhao et al. (2012), Zhou et al. (2017), Kitagawa and Tetenov (2018), Rai (2018), Luedtke and Chambaz (2020), Athey and Wager (2021), Mbakop and Tabord-Meehan (2021), Zhou et al. (2023), Ben-Michael et al. (2024), Ai et al. (2024), Zhan et al. (2024), Viviano (2025) and Fan et al. (2025b) among others. The objective is to provide probabilistic bound on the regret

$$R\left(\hat{\phi}, \phi^*\right) = \mathbb{E}_{X,Y}\left[\sum_j \phi_j^*\left(X\right) Y_j\right] - \mathbb{E}_{X,Y}\left[\sum_j \hat{\phi}_j\left(X\right) Y_j\right],$$

where $\hat{\phi} \in \arg\max_{\phi \in \Phi_0} \sum_i \sum_j \phi_j\left(X_i\right) F_j\left(X_i, D_i, Y_i\right)$ and ϕ^* solves the population analogous problem. The function $F\left(X, D, Y\right)$ typical takes the inverse propensity weighting (IPW) or doubly robust (DR) form and $\Phi_0 \subset \Phi$ is not too complex in an entropy sense. See for example Kitagawa and Tetenov (2018) and Athey and Wager (2021) for EWM in static statistical learning settings; recent advance in the EWM framework like Zhan et al. (2024) and Jin et al. (2025) also consider the adaptive and reinforcement learning environments.

A statistical difficulty in the optimal allocation problem arises from the indicator function in the population solution (Qian and Murphy, 2011). To address such a problem, we build on a divergent literature motivated by very different purposes. To restore monotonicity in conditional quantile estimation, the functional derivative of the sorting operator in the univariate case is studied by Chernozhukov et al. (2010) and subsequently extended to the multivariate cases by Chernozhukov et al. (2018b) using calculus on manifold techniques. Kim and Pollard (1990) hinted at a rudimentary form of calculus on manifold in deriving the large sample properties of cube root

[3] There are certainly many more studies on treatment allocation, policy learning and ITR. The list is illustrative rather than exhaustive.

consistent estimators. Sasaki (2015) incorporated Hausdorff measure tools in fluid mechanics to characterize the information content of quantile partial derivatives for general structural functions. Feng et al. (2025) restated and extended the results in Chernozhukov et al. (2018b) to multi-class cases using the Hausdorff measure and the integration formulas from geometry measure theory.

Although it is of natural importance, only a few studies have considered the constrained optimal allocation problem involving only binary treatments. To the best of our knowledge, existing contributions include Bhattacharya and Dupas (2012), Luedtke and van der Laan (2016) and Feng et al. (2025). The recent fairness-accuracy frontier researches Liang et al. (2021) and Liu and Molinari (2024) can also be regarded as binary constrained optimal allocation problem albeit with an altered interpretation.

The methodology in this paper also has intriguing connection with the faster convergence phenomenon in the classification literature (Devroye et al., 1996; Audibert and Tsybakov, 2007). Consider, especially, the margin assumption (MA) evoked by works including Mammen and Tsybakov (1999), Tsybakov (2004), Boucheron et al. (2005), Audibert and Tsybakov (2007), Kitagawa and Tetenov (2018), Luedtke and Chambaz (2020), Semenova (2023) and Feng et al. (2025):

$$\mathbb{P}\left\{\left|p\left(X\right)-c\right|<t\right\}\leqslant Ct^{\alpha} \tag{3}$$

for fixed c and some constant $C>0$ and $\alpha\geqslant 0$.Conditions in section 3 of Feng et al. (2025) implies the margin assumption with $\alpha\geqslant 1$.

In section 4.2, we devote substantial attention to examples about transport / transportation cost distributionally robust optimization (DRO). Recently, the important duality formula in transport cost DRO is studied in Sinha et al. (2017), Zhao and Guan (2018), Blanchet and Murthy (2019), Gao and Kleywegt (2023), Zhen et al. (2025), Zhang et al. (2025) and Fan et al. (2025a), with gradually improved generality and simplicity. Transport cost DRO learning in simple linear or single index model settings are studied in for example Shafieezadeh Abadeh et al. (2015), Mohajerin Esfahani and Kuhn (2018), Chen and Paschalidis (2018) and Qu and Kwon (2024). Kido (2022) considered transport cost DRO EWM policy learning when the constraints are impost uniformly to all conditional distributions. DRO via divergence can be found in, for example Shapiro (2017) and Duchi and Namkoong (2021).

3 Synopsis

The current draft is incomplete. Portions of section 5 and parts of section 6 contain text directly adapted from Feng et al. (2025) and serve only as placeholders; these sections will be fully rewritten in the subsequent version.

In section 5, we plan to further clarify the connection between the derivative as surface integral method and the margin assumption, and show how can we adopt the *margin assumption* from the

later literature and insights from proof of Kac-Rice formula (Armentano et al., 2025) to overcome the restrictive regular point assumption in the preceding literature.

We are also going to add a new section 7 on the empirical welfare maximization (EWM) In section 7, we plan to provide a extension of the EWM approach to constrained cases by replacing hard (infinite) constraints in the marginal approach with soft constraints implemented via penalization. We also plan to provide a brief summary of tools from probably approximately correct (PAC) classification there.

4 On Variational Analysis

4.1 Optimality Conditions and Basic Settings

In this section, we will first consider the maximization problem

$$\sup_{x\in C}\left\{f_0(x) : f_i(x) \leqslant 0, i = 1,\ldots,n\right\}, \tag{4}$$

where $f_1,\ldots,f_n : C \to \mathbb{R}\cup\{+\infty\}$ are convex functionals, $f_0 : C \to \{-\infty\}\cup\mathbb{R}$ is a concave functional and C a convex subset of a linear space $\mathcal{X}$. We first state a Kuhn-Tucker theorem. Here, we will add an "additional" λ_0 to the objective function. This parameter λ_0 will make the expression of the theorem more complete.

Theorem 4.1.1 (Kuhn-Tucker theorem)**.** *Let $\mathcal{X}$ be a linear space, $C \subset \mathcal{X}$ is a convex subset. Let $f_1,\ldots,f_n$ be convex functionals on C and f_0 be a concave functional on C. If $x_0 \in C$ is a solution of (4), then there exist $\lambda = (\lambda_0,\ldots,\lambda_n) \in \mathbb{R}^{n+1}_+$ such that*

$$x_0 \in \arg\min_{x\in C}\left\{\lambda_0 f_0(x) - \sum_{i=1}^{n}\lambda_i f_i(x)\right\},$$

and the complementary slackness condition holds:

$$\lambda_i f_i(x_0) = 0,\ \forall i \in \{1,\ldots,n\}. \tag{5}$$

If the constrained problem (4) satisfies the Slater's condition, i.e.

$$\exists x \in C, f_0(x) > -\infty \text{ and } f_i(x) < 0, \forall i \in \{1,\ldots,n\}, \tag{6}$$

then $\lambda_0 > 0$.

This Kuhn-Tucker theorem for linear space is a consequence of geometric Hahn-Banach theorem, i.e. convex separation theorem. We should note that there is actually no need to assume $\mathcal{X}$

as a Banach space or to assume continuity for $f_0, f_1, \ldots, f_n$. While, Theorem 4.1.1 says nothing above the existence of the solution of (4).

Obviously, if $\lambda_0 > 0$ then we can set it to $\lambda_0 = 1$ (more conventional form of the Kuhn-Tucker theorem). The inverse direction of Theorem 4.1.1 is trivially true. If $x_0 \in C$ maximizes $f(x) - \sum_{i=1}^n \lambda_i f_i(x)$, then x_0 is also a solution of

$$\sup_{x \in C} \{f_0(x) : f_i(x) \leqslant f_i(x_0), i = 1, \ldots, n\}.$$

Proof of Theorem 4.1.1.

step 1 Assume that x_0 solves (4), we need find $\lambda \in \mathbb{R}_+^{n+1}$ such that

$$\lambda_0 f_0(x_0) - \sum_{i=1}^n \lambda_i f_i(x_0) \geqslant \lambda_0 f_0(x) - \sum_{i=1}^n \lambda_i f_i(x), \ \forall x \in C. \tag{7}$$

Consider two convex subset of $\mathbb{R}^{n+1}$,

$$\begin{aligned} E &:= \left\{(t_0, \ldots, t_n) \in \mathbb{R}^{n+1} : t_0 \geqslant f_0(x_0) \wedge t_i \leqslant 0, i = 1, \ldots, n\right\}, \text{ and} \\ F &:= \left\{(t_0, \ldots, t_n) \in \mathbb{R}^{n+1} : \exists x \in C \ s.t. \ t_0 \leqslant f_0(x) \wedge t_i \geqslant f_i(x), i = 1, \ldots, n\right\}. \end{aligned}$$

Obviously, $\overset{\circ}{E} \neq \varnothing$ and $\overset{\circ}{E} \cap F = \varnothing$. By the convex separation theorem, we get that there exist $\lambda \in \mathbb{R}^{n+1}$ such that

$$\lambda_0 f_0(x_0) - \sum_{i=1}^n \lambda_i f_i(x_0) \geqslant \lambda_0 (f_0(x) - \xi_0) - \sum_{i=1}^n \lambda_i (f_i(x) + \xi_i), \ \forall x \in C, \ \forall \xi \in \mathbb{R}_+^{n+1}.$$

So we must have $\lambda \in \mathbb{R}_+^{n+1}$.

step 2 In order to prove the complementary slackness condition, note that $(f_0(x_0), 0, \ldots, 0) \in E$ and $(f_0(x_0), f_1(x_0), \ldots, f_n(x_0)) \in F$, i.e.

$$\lambda_0 f(x_0) \geqslant \lambda_0 f(x_0) - \sum_{i=1}^n \lambda_i f_i(x_0).$$

Therefore, we get $\sum_{i=1}^n \lambda_i f_i(x_0) \geqslant 0$. Then $\lambda_i f_i(x_0) = 0, i = 1, \ldots, n$ follow from the assumption $f_i(x_0) \leqslant 0$ and the fact $\lambda_i \geqslant 0$.

step 3 Under the Slater condition, let x' satisfy $f_0(x') > -\infty$ and $f_i(x') < 0, i = 1, \ldots, n$. If $\lambda_0 = 0$, then by (7), $\sum_{i=1}^n \lambda_i f_i(x') \geqslant 0$. This implies that $\lambda = (0, \ldots, 0)$, which is impossible since λ corresponds to a hyperplane. ■

The above proof follows the lecture note of *Function Analysis* I by Kung-ching Chang at Peking University Chang (2021). I can only find the Chinese version of this note.

For a multi-treatment or multi-classification problem, our basic working space will be the set

Φ of multi-critical functions. We call a function $\phi = (\phi_1, \ldots, \phi_n)$, $\phi_i : \Omega \to \mathbb{R}$ as a multi-critical function if $\phi_i \in L^\infty(\mu), 0 \leqslant \phi_i \leqslant 1$ almost surely, $\forall i \in \{1, \ldots, n\}$, and such that $\sum_{i=1}^n \phi_i = 1$ almost surely, where (Ω, Σ, μ) being a probability space.

Since μ is finite, $\prod_{i=1}^n L_i^\infty(\mu)$ is the dual space of $\prod_{i=1}^n L_i^1(\mu)$. By the Banach–Alaoglu theorem (see for example Theorem T.1.1), we have that Φ is contained in a compact set for the weak* topology. By Theorem T.1.2, if $L^1(\mu)$ is also separable, than $\bar{\Phi}$ is also sequentially compact. There we provide some general conditions to guarantee the separability of $L^1(\mu)$.

Definition 4.1.1. (Separable measure space) A measure space (Ω, Σ, μ) is called separable if the metric space (Σ_1, ρ_Δ) is separable, where $\Sigma_1 := \{s \in \Sigma : \mu(s) < \infty\}$ and $\rho_\Delta(s_1, s_2) := \mu(s_1 \backslash s_2) + \mu(s_2 \backslash s_1)$.

Theorem 4.1.2 (Corollary 11.29 in Di Biase and Krantz (2021))**.** *Let (Ω, Σ, μ) be a measure space, then $L^1(\mu)$ is separable as a metric space if and only if (Ω, Σ, μ) is separable.*

If a σ-algebra can be generated through a countable family of subset, then we call it a countably generated σ-algebra.

Corollary 4.1.3. *Let (Ω, Σ, μ) be a probability space and Σ is countably generated, then $L^1(\mu)$ is separable.*

To finish our mathematic warm-up, note that Φ is sequentially closed. Let X be a normed space with dual X^*. By definition, the weak* topology on X^* is the weakest topology that makes maps such as $x^* \mapsto x(x^*) := x^*(x)$ continuous for all $x \in X$. Therefore, convergence in weak* topology of a sequence $\{x_j^*\}_{j=1}^\infty$ to x_0^* implies that for all $x \in X$, $\lim_{j\to\infty} x_j^*(x) = x_0^*(x)$. Let $\{\phi_j\}_{j=1}^\infty$ be a sequence in Φ that converges to ϕ_0 in weak* topology. We have that

$$\lim_{j\to\infty} \sum_{i=1}^n \int \phi_{ji} f_i d\mu \to \sum_{i=1}^n \int \phi_{0i} f_i d\mu, \ \forall f \in \prod_{i=1}^n L^1(\mu).$$

Since f can be arbitrarily chosen, we obtain that $0 \leqslant \phi_{0i} \leqslant 1$ almost surely, $\forall i \in \{1, \ldots, n\}$, and $\sum_{i=1}^n \phi_{0i} = 1$ almost surely, i.e. $\phi_0 \in \Phi$.

For optimal allocation problems in this paper, we typically have that Ω is an open subset of $\mathbb{R}^m$ and Σ is the Borel σ-algebra on it. Since the Borel σ-algebra of $\mathbb{R}^m$ can be countably generated by the family of all open balls centered at rational numbers with rational radius, the set of multi-critical functions Φ is sequentially compact by Corollary 4.1.3. The sequential compactness then guarantees that the optimal allocation problems with semicontinuous objective and constraints have solutions by the *direct method* in the calculus of variations. See, for example chapter 3 of Attouch et al. (2014). Combining the existence of the solutions with the characteristics of the

solutions given by the Kuhn-Tucker theorem 4.1.1, we can already get much information about typical optimal allocation problems. See a direct application in Example 4.1.

Example 4.1. (Multi-classification receiver operating characteristic surfaces)
For binary classification problems, the receiver operating characteristic (ROC) curves are widely used to evaluate prediction and decision making performance. In population, the (optimal) ROC curve of a classification problem is define as (Feng et al., 2025):

$$\beta(\alpha) := \max_{\phi \in \Phi} \left\{ \frac{\mathbb{E}[Y\phi(X)]}{\mathbb{E}Y} : \frac{\mathbb{E}[(1-Y)\phi(X)]}{\mathbb{E}(1-Y)} \leqslant \alpha \right\},$$

where X is just some features in the problem, $Y \in \{0,1\}$ and Φ consists of all functions such that $0 \leqslant \phi \leqslant 1$ (almost surely). The classical Neyman-Pearson lemma states that the optimal ϕ for the above constrained optimization problem must takes the form of

$$\phi(x) = \begin{cases} 1, & \text{when } p(x) > c \\ 0, & \text{when } p(x) < c. \end{cases},$$

where $p(x) := \mathbb{E}(Y|X=x)$ and $c \in [0,1]$ is a cutoff threshold. See, for example Theorem 3.2.1 in Lehmann and Romano (2022). Despite the widespread use of ROC curves, the issue of how to generalize the analysis to multi-classification remains unresolved.

An important application of Theorem 4.1.1 to the multi-classification problem is that it can provide direct extensions to the binary ROC analysis. For a J-classification problem, denote the true label as Y and the classification output as $\hat{Y}$, where $Y, \hat{Y} \in \{0, \ldots, J-1\}$. There will be $J \times J$ possible outcomes, as described in the following confusion matrix:

	$\hat{Y}=0$	$\hat{Y}=1$	$\cdots$	$\hat{Y}=J-1$
$Y=0$	TPR_0	FPR_{01}	$\cdots$	$\mathrm{FPR}_{0,J-1}$
$Y=1$	FPR_{10}	TPR_1	$\cdots$	$\mathrm{FPR}_{1,J-1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Y=J-1$	$\mathrm{FPR}_{J-1,0}$	$\mathrm{FPR}_{J-1,1}$	$\cdots$	TPR_{J-1}

Here, FPR and TPR are the abbreviations of *false positive rate* and *true positive rate* respectively, defined as

$$\mathrm{FPR}_{jk} = \frac{\mathbb{P}\left\{Y=j, \hat{Y}=k\right\}}{\mathbb{P}\{Y=j\}}, \ \mathrm{TPR}_j = \frac{\mathbb{P}\left\{Y=\hat{Y}=j\right\}}{\mathbb{P}\{Y=j\}}, \ \forall j,k \in \{0,\ldots,J-1\}, j \neq k. \tag{8}$$

It is usually more convenient to use the one-hot encodings for both Y and $\hat{Y}$, e.g. we may redefine $Y = (Y_0, \ldots, Y_{J-1}) \in \{0,1\}^J$. Y_j is set to 1 if and only if the original label is j. Then we can rewrite that $\mathrm{FPR}_{jk} = \mathbb{E}\left[Y_j \hat{Y}_k\right] / \mathbb{E}Y_j$ and $\mathrm{TPR}_j = \mathbb{E}\left[Y_j \hat{Y}\right] / \mathbb{E}Y_j$. For brevity, we also denote $\alpha_{jk} = \mathrm{FPR}_{jk}$ and $\beta_j = \mathrm{TPR}_j$.

We may first consider the "full TPR" generalization of the binary ROC curve. Consider the constrained optimization problem

$$\max_{\phi \in \Phi} \left\{ \frac{\mathbb{E}\left[Y_0 \phi_0\left(X\right)\right]}{\mathbb{E}Y_0} : \frac{\mathbb{E}\left[Y_j \phi_j\left(X\right)\right]}{\mathbb{E}Y_j} \geqslant t_j, j = 1, \ldots, J \right\}. \tag{9}$$

Compared with (8), we allow *randomized* decisions in (9). For randomized critical function ϕ, we use $\mathbb{P}\left[\hat{Y}_j | X\right] = \phi_j\left(X\right)$ to generate discreteized prediction. Denote the projection of Φ to the FPRs and TPRs in the confusion matrix as

$$T := \left\{ \left(T_{jk}\left(\phi\right)\right)_{j,k=0,\ldots,J-1} : \phi \in \Phi \right\},$$

where $T_{jk} = \mathrm{FPR}_{jk}, j \neq k$ and $T_{jk} = \mathrm{TPR}_j, j = k$. Obviously, we have that T is compact and convex, and thus all coordinate projection of T are all compact and convex. A direct implication of Theorem 4.1.1 and our discussion about the weak* sequential compactness is the following Neyman-Pearson lemma for (9).

Corollary 4.1.4 (Diagonal multi-classification Neyman-Pearson lemma)**.** *Let (Ω, Σ, μ) be a probability space and Σ is countably generated. Then for all $t \in T_{DC} := \left\{ \left(t_{jj}\right)_{j \in \{1,\ldots,J-1\}} : t \in T \right\}$, we have:*

1. ***Existence**. There exists a solution $\phi \in \Phi$ (not necessarily unique) for (9).*
2. ***Sufficient condition**. If $\phi \in \Phi$ satisfies that for all $k \in \{0, \ldots, J\}$,*

$$\phi_k\left(x\right) = \begin{cases} 1, & \text{when} \quad \lambda_k p_k\left(x\right) > \max_{l \neq k, l=0,\ldots,J-1} \{\lambda_l p_l\left(x\right)\}, \\ 0, & \text{when} \quad \lambda_k p_k\left(x\right) < \max_{l \neq k, l=0,\ldots,J-1} \{\lambda_l p_l\left(x\right)\}, \end{cases} \tag{10}$$

   where $\lambda \in \mathbb{R}_+^J$ and $\lambda_0 > 0$, then ϕ solves (9) for the t level it achieves.

3. ***Necessary condition**. If $\phi \in \Phi$ is a solution of (9), then there exists $\lambda \in \mathbb{R}_+^J$ such that ϕ satisfies (10). Further, if there exists $t' \in T_{DC}$ such that $t' > t$, then $\lambda_0 > 0$.*

Since we have $\sum_{j=0}^{J-1} \phi_j = 1$ almost surely, we can use $J \times (J-1)$ elements in the confusion matrix to identify the whole matrix. A possible choice is to consider all of the off-diagonal elements, i.e. all of the FPRs. In this case, we have Corollary 4.1.5 for the following constrained optimization

problem:

$$\min_{\phi\in\Phi}\left\{\frac{\mathbb{E}\left[Y_0\phi_1\left(X\right)\right]}{\mathbb{E}Y_0}:\frac{\mathbb{E}\left[Y_j\phi_k\left(X\right)\right]}{\mathbb{E}Y_j}\leqslant t_{jk},j\neq k,(j,k)\neq(0,1),j,k=0,\ldots J-1\right\} \tag{11}$$

Corollary 4.1.5 (Off-diagonal multi-classification Neyman-Pearson lemma)**.** *Let (Ω,Σ,μ) be a probability space and Σ is countably generated. Then for all*

$$t\in T_{OC}:=\left\{\left(t_{jk}\right)_{j\neq k,(j,k)\neq(0,1),j,k=0,\ldots,J-1}:t\in T\right\},$$

we have:

1. ***Existence**. There exists a solution $\phi\in\Phi$ (not necessarily unique) for (11).*
2. ***Sufficient condition**. If $\phi\in\Phi$ satisfies that for all $k\in\{0,\ldots,J\}$,*

$$\phi_k\left(x\right)=\begin{cases}1, & when \quad \sum_{l\neq k}\lambda_{l,k}p_l\left(x\right)<\min_{l\neq k,l=0,\ldots,J-1}\sum_{m\neq l}\lambda_{m,l}p_m\left(x\right),\\ 0, & when \quad \sum_{l\neq k}\lambda_{l,k}p_l\left(x\right)<\min_{l\neq k,l=0,\ldots,J-1}\sum_{m\neq l}\lambda_{m,l}p_m\left(x\right),\end{cases} \tag{12}$$

   where $\lambda\in\mathbb{R}_+^{J\times(J-1)}$ and $\lambda_{01}>0$, then ϕ solves (11) for the t level it achieves.

3. ***Necessary condition**. If $\phi\in\Phi$ is a solution of (11), then there exists $\lambda\in\mathbb{R}_+^{J\times(J-1)}$ such that ϕ satisfies (12). Further, if there exists $t'\in T_{OC}$ such that $t'<t$, then $\lambda_{01}>0$*

In Edwards et al. (2005), it is argued that a multi-classification receiver operating characteristic surface for (11) may have zero volume under the surface. In this context, the "all FPRs" type ROC surface is suggested to be useless because T_{OC} is too small. Here, we give some clarifications about this kind of claim. Reasoning in Edwards et al. (2005) based on a simple observation: uninformative guess decisions will only generate a $J-1$-dimensional simplex in the FPRs space. This observation is correct. Actually, consider totally uninformative $\phi\in\Phi$ such that $\phi_k\equiv r_k$ for constant r_k, $k\in\{0,\ldots,J-1\}$. In this case $\mathrm{FPR}_{jk}=r_k$ for all $j\in\{0,\ldots,J-1\}$. The image set $T_{UI}:=\left\{\left(T_{jk}\left(\phi\right)\right)_{j\neq k,j,k=0,\ldots,J-1}:\phi \text{ uninformative}\right\}$ is contained in a $J-1$-dimensional affine space. For binary classification, $J-1$ is equal to 1 while $J\times(J-1)$ is equal to 2. This is the only case where T_{UI} is not "too small".

But if we consider a ideal "cheating code" case from our variational analysis perspective, then we will also observe something different. Assuming that X contain Y, i.e. we directly know the true label. Now, consider the multi-classification critical functions such that

$$\begin{cases}\phi_k\left(X\right)=1, & \text{if} \quad Y_j=1,\\ \phi_l\left(X\right)=1, & \text{if} \quad Y_l=1,\forall l\neq j,l=0,\ldots,J-1,\end{cases}$$

where $k \neq j$, $j, k \in \{0, \ldots, J-1\}$. The meaning of these critical functions is that they deliberately make mistakes to make a classification of k when they know the true label is j. For example, let $J = 3$ and $(j, k) = (0, 1)$, then we can get the following confusion matrix:

	$\hat{Y}=0$	$\hat{Y}=1$	$\hat{Y}=2$
$Y=0$	0	1	0
$Y=1$	0	1	0
$Y=2$	0	0	1

It is not hard to see that these "deliberate wrong" critical functions together with a perfect critical function that never make mistakes will span a $J \times (J-1)$-dimensional convex body by projection.

We still use the 3-classification as an example. Consider the image set T_3[4] consists of

$$(\alpha_{01}, \alpha_{02}, \alpha_{10}, \alpha_{12}, \alpha_{20}, \alpha_{21}).$$

If T_3 is contained in an affine space which is not $\mathbb{R}^6$, then by Theorem 1.4 in Rockafellar (1970), there exists $b \in \mathbb{R}^6$, $b \neq 0$ and $a \in \mathbb{R}$ such that $b^T T_3 = a$, i.e. $\sum_{j \neq k, j,k=0,1,2} b_{jk}\alpha_{jk} = a$ for all $\alpha \in T_3$. This implies that

$$\begin{aligned}
&\mathbb{E} \max \left\{ \frac{b_{10}p_1(X)}{p_1} + \frac{b_{20}p_2(X)}{p_2}, \frac{b_{01}p_0(X)}{p_0} + \frac{b_{21}p_2(X)}{p_2}, \frac{b_{02}p_0(X)}{p_0} + \frac{b_{12}p_1(X)}{p_1} \right\} \\
&= \mathbb{E} \min \left\{ \frac{b_{10}p_1(X)}{p_1} + \frac{b_{20}p_2(X)}{p_2}, \frac{b_{01}p_0(X)}{p_0} + \frac{b_{21}p_2(X)}{p_2}, \frac{b_{02}p_0(X)}{p_0} + \frac{b_{12}p_1(X)}{p_1} \right\},
\end{aligned}$$

where $p_j = \mathbb{E}Y_j$, $j \in \{0, 1, 2\}$. We obtain that the maximum integrand and the minimum integrand must be equal almost surely. Otherwise, T_3 will not be contained in any affine proper subset of $\mathbb{R}^6$. Now, by Theorem 6.2 in Rockafellar (1970), T_3 has nonempty inferior which implies that T_3 has nonzero Hausdorff measure and thus Lebesgue measure, i.e. what is usually called volume. See Theorem 2.5 in Evans and Garzepy (2015).

4.2 The Lagrangian Approach

Next, we should also consider the Lagrangian approach to duality. The Lagrangian approach transforms the original problem (4) to the dual minimax problem. Compared to the Kuhn-Tucker theorem 4.1.1, the Lagrangian approach is more complicated while provides a direct way to calculate the value of λ. From now on, we will also call λ a Lagrangian multiplier.

Definition 4.2.1.

[4] It seems that we are using the symbol T with different subscripts too much.

1. For the concave maximization problem (4), the set M of generalized Lagrange multiplier vectors is defined by

$$M = \left\{ \lambda \in \mathbb{R}^n_+ : \sup_{x \in C} \left\{ f_0 + \delta_{CT(0)} \right\} = \sup_{x \in C} \left\{ f_0 - \sum_{i=1}^n \lambda_i f_i \right\} \right\},$$

where $CT(y) = \{x \in C : f_i(x) \leqslant y, i = 1, \ldots, n\}$. When (4) has a solution, M will be called Lagrange multiplier vectors for (4).

2. The value function attached to the problem (4) is defined by

$$v(y) = \sup_{x \in C} \{ f_0(x) : f_i(x) \leqslant y_i, \forall i = 1, \ldots, n \}, \ \forall y \in \mathbb{R}^n.$$

3. The Lagrangian function or just Lagrangian attached to the problem (4) is defined by

$$L(x, \lambda) := f_0(x) - \sum_{i=1}^n \lambda_i f_i(x).$$

Definition 4.2.2. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two arbitrary spaces, $\mathrm{F} : \mathcal{S}_1 \times \mathcal{S}_2 \to \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$ be a bivariate function. A point $(\bar{s}_1, \bar{s}_2) \in \mathcal{S}_1 \times \mathcal{S}_2$ is called a saddle point of F if

$$\max_{s_2 \in \mathcal{S}_2} \mathrm{F}(\bar{s}_1, s_2) = \mathrm{F}(\bar{s}_1, \bar{s}_2) = \min_{s_1 \in \mathcal{S}_1} \mathrm{F}(s_1, \bar{s}_2) \text{ or}$$
$$\min_{s_2 \in \mathcal{S}_2} \mathrm{F}(\bar{s}_1, s_2) = \mathrm{F}(\bar{s}_1, \bar{s}_2) = \max_{s_1 \in \mathcal{S}_1} \mathrm{F}(s_1, \bar{s}_2).$$

Theorem 4.2.1 (Lagrangian Duality theorem, maximization version and a restatement of Theorem 9.6.4 in Attouch et al. (2014))**.** *Let $\mathcal{X}$ be a linear space, $C \subset \mathcal{X}$ is a convex subset. Let $f_1, \ldots, f_n$ be convex functionals on C and f_0 be a concave functional on C. Consider the concave maximization problem (4), the following facts hold true:*

1. *If there exists $(x_0, \lambda_0) \in C \times \mathbb{R}^n_+$ such that (x_0, λ_0) is a saddle point of $L(\cdot, \cdot)$, then the primal problem has a solution x_0 and the dual problem has a solution $\lambda_0 \in M$.*

2. *If the Slater condition holds, there is no duality gap, i.e.*

$$\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = \inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda). \tag{13}$$

If $v(0) \in \mathbb{R}$, then the solution set of the dual problem is nonempty, i.e.

$$\inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda) = \min_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda)$$

and

$$M = \arg\min_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda).$$

3. *Under the Slater condition, if the primal problem (4) has a solution x_0, then there exists a Lagrangian multiplier λ_0 such that (x_0, λ_0) is a saddle point of $L(\cdot, \cdot)$.*

First, we should note that

$$\inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = f_0(x) + \delta_{CT(0)}(x).$$

So the sup-inf problem $\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda)$ is an equivalent expression of the primal problem (4). For completeness, we provide a proof of Theorem 4.2.1 following ideas in section 9.6 in Attouch et al. (2014).

The best of Theorem 4.2.1 is that there is also no need to assume any specific structure of $\mathcal{X}$. We only assume that $\mathcal{X}$ is a linear space and f_0, f_i, $i \in \{1, \ldots, n\}$ are general concave or convex functionals on $\mathcal{X}$. No assumptions like semicontinuity of f_0, f_i, $i \in \{1, \ldots, n\}$ are required either. We should also note that, the strong duality above is about the value function, there is no guarantee that the primal problem (4) actually has a solution when we have no information about the saddle point of the Lagrangian $L(\cdot, \cdot)$. While, the dual problem always has solutions under rather weak assumptions.

PROOF OF THEOREM 4.2.1.

part 1 If there exists $(x_0, \lambda_0) \in C \times \mathbb{R}^n_+$ such that (x_0, λ_0) is a saddle point of $L(\cdot, \cdot)$, then the primal problem has a solution x_0 and the dual problem has a solution $\lambda_0 \in M$.

step 1 Let us now show that x_0 is a solution of the primal problem. Assume that (x_0, λ_0) is a saddle point of the Lagrangian $L(\cdot, \cdot)$. There is

$$\begin{aligned} f_0(x_0) + \delta_{CT(0)}(x_0) = \inf_{\lambda \in \mathbb{R}^n_+} L(x_0, \lambda) &= L(x_0, \lambda_0) \\ &= \sup_{x \in C} L(x, \lambda_0) \\ &\geqslant \inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda) \\ &\geqslant \sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = \sup_{x \in C} \left\{ f_0(x) + \delta_{CT(0)}(x) \right\}, \end{aligned}$$

i.e. x_0 is a solution of the primal problem.

step 2 The dual problem has a solution $\lambda_0 \in M$. To see this, observe that

$$L(x_0, \lambda_0) = \sup_{x \in C} \left\{ f_0(x) - \sum_{i=1}^{n} \lambda_{0,i} f_i(x) \right\} = \sup_{x \in C} \left\{ f_0(x) + \delta_{CT(0)}(x) \right\}.$$

Therefore, $\lambda_0 \in M$ by 1. in Definition 4.2.1. We also have that

$$L(x_0, \lambda_0) = \sup_{x \in C} L(x, \lambda_0) = \inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda),$$

which means that λ_0 is a solution of the dual problem $\inf_{\lambda \in \mathbb{R}^n_+} d(\lambda)$, where $d(\lambda) := \sup_{x \in C} L(x, \lambda)$.
part 2 First consider the fine properties of the value function $v(\cdot)$. The goal is to shown that if $v(0) \in \mathbb{R}$, then $M = \nabla^+ v(0)$.
step 1 The value function $v(\cdot)$ is concave. Since

$$v(y) = \sup_{x \in C} \left\{ f_0(x) + \delta_{CT(y)}(x) \right\},$$

where $CT(y) = \{x \in C : f_i(x) \leqslant y_i, \forall i \in \{1, \ldots, n\}\}$. It is not hard to verify that the map $(x, y) \mapsto \delta_{CT(y)}(x)$ is concave. In fact, for arbitrary $(x_1, y_1), (x_2, y_2) \in C \times \mathbb{R}^n$ such that $x_1 \in CT(y_1)$, $x_2 \in CT(y_2)$, by the convexity of functions f_i, for all $\lambda \in [0, 1]$, there is

$$\begin{aligned} f_i(\lambda x_1 + (1-\lambda) x_2) &\leqslant \lambda f_i(x_1) + (1-\lambda) f_i(x_2) \\ &\leqslant \lambda y_1 + (1-\lambda) y_2. \end{aligned}$$

Since f_0 is concave, $f_0(x) + \delta_{CT(y)}(x) : C \times \mathbb{R}^n \to \{-\infty\} \cup \mathbb{R}$ is also concave. To show that $v(\lambda y_1 + (1-\lambda) y_2) \geqslant \lambda v(y_1) + (1-\lambda) v(y_2)$ for arbitrary $y_1, y_2 \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, without loss of generality we can assume $v(y_1), v(y_2) > -\infty$, or there is nothing left to prove. By definition of the value function $v(\cdot)$, for arbitrary $\epsilon > 0$, we can find $x_{1,\epsilon}$ and $x_{2,\epsilon}$ such that

$$v(y_1) < f_0(x_{1,\epsilon}) + \delta_{CT(y_1)}(x_{1,\epsilon}) + \epsilon, \; v(y_2) < f_0(x_{2,\epsilon}) + \delta_{CT(y_2)}(x_{2,\epsilon}) + \epsilon.$$

By concavity of $f_0(x) + \delta_{CT(y)}(x)$,

$$\begin{aligned} &f_0(\lambda x_{1,\epsilon} + (1-\lambda) x_{2,\epsilon}) + \delta_{CT(\lambda y_1 + (1-\lambda) y_2)}(\lambda x_{1,\epsilon} + (1-\lambda) x_{2,\epsilon}) \\ &\qquad \geqslant \lambda \left(f_0(x_{1,\epsilon}) + \delta_{CT(y_1)}(x_{1,\epsilon}) \right) + (1-\lambda) \left(f_0(x_{2,\epsilon}) + \delta_{CT(y_2)}(x_{2,\epsilon}) \right). \end{aligned}$$

By definition of $v(\cdot)$,

$$v(\lambda y_1 + (1-\lambda) y_2) \geqslant f_0(\lambda x_{1,\epsilon} + (1-\lambda) x_{2,\epsilon}) + \delta_{CT(\lambda y_1 + (1-\lambda) y_2)}(\lambda x_{1,\epsilon} + (1-\lambda) x_{2,\epsilon}).$$

Then concavity of $v(\cdot)$ follows from the arbitrariness of ϵ.
step 2: We claim that if $v(0) \in \mathbb{R}$, then $M = \nabla^+ v(0)$. Assume that $\lambda \in M$, by 1. and 2. in

Definition 4.2.1,

$$v(0) = \sup\left\{f_0 + \delta_{CT(0)}\right\} = \sup\left\{f_0 - \sum_{i=1}^{n} \lambda_i f_i\right\}.$$

Here for brevity, we omit the variable x and its domain C. For arbitrary $y \in \mathbb{R}^n$ and for all $x \in CT(y)$, there is $\sum_{i=1}^{n} \lambda_i f_i(x) \leqslant \sum_{i=1}^{n} \lambda_i y_i$ recalling that $\lambda \in \mathbb{R}^n_+$. Therefore, for all $x \in CT(y)$, $v(0) \geqslant f_0(x) - \sum_{i=1}^{n} \lambda_i y_i$ and then

$$v(0) \geqslant \sup_{x \in CT(y)} \left\{f_0(x) - \sum_{i=1}^{n} \lambda_i y_i\right\} = v(y) - \sum_{i=1}^{n} \lambda_i y_i,$$

i.e. $v(y) - v(0) \leqslant \sum_{i=1}^{n} \lambda_i y_i$ or $\lambda \in \nabla^+ v(0)$.

For the converse direction, assume that $\lambda \in \nabla^+ v(0)$. Note that for all $y \in \mathbb{R}^n_+$, $CT(0) \subset CT(y)$ and $v(0) \leqslant v(y)$. By concavity of $v(\cdot)$ and the assumption that $\lambda \in \nabla^+ v(0)$, $v(y) - v(0) \leqslant \sum_{i=1}^{n} \lambda_i y_i$. Therefore $\sum_{i=1}^{n} \lambda_i y_i \geqslant 0$ for all $y \in \mathbb{R}^n_+$ and thus we have $\lambda \in \mathbb{R}^n_+$. For all $x \in CT(0)$, we have $f_i(x) \leqslant 0$, $\forall i \in \{1, \ldots, n\}$. This implies that

$$v(0) = \sup_{x \in C}\left\{f_0 + \delta_{CT(0)}\right\} = \sup_{x \in CT(0)} f_0 \leqslant \sup_{x \in CT(0)} \left\{f_0 - \sum_{i=1}^{n} \lambda_i f_i\right\} \leqslant \sup_{x \in C} \left\{f_0 - \sum_{i=1}^{n} \lambda_i f_i\right\}.$$

For arbitrary $x \in C$ let $y_i = f_i(x)$, $i \in \{1, \ldots, n\}$. These is $x \in CT(y)$ and $v(y) \geqslant f_0(x)$. By the supdifferential inequality,

$$v(0) + \sum_{i=1}^{n} \lambda_i f_i(x) \geqslant v(y) \geqslant f_0(x).$$

The above inequality is true for all $x \in C$ which implies that

$$v(0) \geqslant \sup_{x \in C} \left\{f_0 - \sum_{i=1}^{n} \lambda_i f_i\right\}.$$

We now obtain that $v(0) = \sup_{x \in C}\{f_0 - \sum_{i=1}^{n} \lambda_i f_i\}$ and $\lambda \in M$ by 1. in Definition 4.2.1.

part 3 For the concave maximization problem (4) under the Slater condition, there is no duality gap and if $v(0) \in \mathbb{R}$, then the solution set of the dual problem is the set $M \neq \varnothing$.

step 1 The one side inequality

$$\inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda) \geqslant \sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda)$$

is always true, since $\inf_A \sup_B \geqslant \sup_B \inf A$ is always true. By Slater condition, we also have that $\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) > -\infty$. If $\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = +\infty$, then there is nothing left to prove. If $\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) \in \mathbb{R}$, by the fact that $\inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = f_0(x) + \delta_{CT(0)}(x)$ and 2.

in Definition 4.2.1, we obtain $v(0) \in \mathbb{R}$. By the result of **part** 2, there is $M = \nabla^+ v(0)$. For all $\lambda' \in M$, by 1. in Definition 4.2.1 we have

$$\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = \sup_{x \in C} L\left(x, \lambda'\right) \geqslant \inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda).$$

We obtain $\sup_{x \in C} \inf_{\lambda \in \mathbb{R}^n_+} L(x, \lambda) = \inf_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda)$.

step 2 We claim that $M \neq \varnothing$. The Slater condition tells that there exists a $x_0 \in C$ such that $f_0(x_0) > -\infty$ and $f_i(x_0) < 0, \forall i \in \{1, \ldots, n\}$. Therefore, on a neighborhood of the origin in $\mathbb{R}^n$, $v(y)$ is bounded below by the fact that $v(y) \geqslant f_0(x_0)$ when $|y_i| < |f_i(x_0)|$, $\forall i \in \{1, \ldots, n\}$. For all $y \in \mathbb{R}^n$, $\|y\|$ small enough, $v(0) \geqslant \frac{1}{2} v(y) + \frac{1}{2} v(-y)$ then implies $v(y) \leqslant 2v(0) - v(-y) \leqslant 2v(0) - f_0(x_0)$. Since it is assumed that $v(0) \in \mathbb{R}$, $v(\cdot)$ is real-valued on a neighborhood of the origin. Now we get that $M \neq \varnothing$ by the fact that if a function $g : U \to \mathbb{R}$ is convex on an open subset U of $\mathbb{R}^n$, then it is superdifferentiable at every point in U. See, for example Example 10.6 in Villani (2009).

step 3 It can also be shown that $v(\cdot)$ is Lipschitz continuous on a neighborhood of the origin. Let $y_1, y_2 \in B(0, r)$, where r is taken small enough such that for all $y \in B(0, 2r)$, $|v(y)| \leqslant MC := \max\{|f_0(x_0)|, |2v(0) - f_0(x_0)|\}$. Next, let $w = y_2 + \frac{r}{d}(y_2 - y_1)$ where $d = \|y_1 - y_2\|$. Obviously, $w \in B(0, 2r)$ and $y_2 = \frac{r}{r+d} y_1 + \frac{d}{r+d} w$. By concavity of $v(\cdot)$

$$v(y_2) \geqslant \frac{r}{r+d} v(y_1) + \frac{d}{r+d} v(w).$$

Then, we have

$$v(y_2) - v(y_1) \geqslant \frac{d}{r+d}(v(w) - v(y_1)) \geqslant -\frac{2d}{r} MC.$$

Let $w' = y_1 + \frac{r}{d}(y_1 - y_2)$ similarly give us $v(y_1) - v(y_2) \geqslant -\frac{2d}{r} MC$. Combining these two inequalities, we obtain $\frac{|v(y_1) - v(y_2)|}{|y_1 - y_2|} \leqslant \frac{2}{r} MC$.

step 4 In order to show that $M = \arg\min_{\lambda \in \mathbb{R}^n_+} \sup_{x \in C} L(x, \lambda)$, consider the Legendre-Fenchel transformation $(-v)^*(y') = \sup_{y \in \mathbb{R}^n} \{\sum_{i=1}^n y'_i y_i - (-v(y))\}$. We check that

$$\lambda \in M \Longleftrightarrow -v(0) + (-v)^*(-\lambda) = 0.$$

For all $y, y' \in \mathbb{R}^n$, by definition $-v(y) + (-v)^*(y') - \sum_{i=1}^n y'_i y_i \geqslant 0$, and thus $-v(0) + (-v)^*(-\lambda) \geqslant 0$. For the converse inequality, recall that $M = \nabla^+ v(0)$. There is $v(y) - v(0) \leqslant \sum_{i=1}^n \lambda_i y_i$, $\forall y \in \mathbb{R}^n$, which is equivalent to $-v(0) + (-v)^*(-\lambda) \leqslant 0$.

step 5 Since $v(\cdot)$ is continuous on a neighborhood of the origin, by a version of the Fenchel-Moreau theorem, i.e. Proposition 9.3.2 in Attouch et al. (2014), $-v(0) = (-v)^{**}(0)$. By definition of the

Legendre-Fenchel transformation, $(-v)^{**}(0) = \sup_{y' \in \mathbb{R}^n} \{-(-v)^*(y')\}$. Therefore, we have

$$\begin{aligned} \lambda \in M \Longleftrightarrow (-v)^*(-\lambda) &= -\sup_{y' \in \mathbb{R}^n} \{-(-v)^*(y')\} \\ &= \inf_{y' \in \mathbb{R}^n} \{(-v)^*(y')\}. \end{aligned}$$

We then calculate $(-v)^*(\cdot)$ via

$$\begin{aligned} (-v)^*(y') &= \sup_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^n y_i' y_i + v(y) \right\} \\ &= \sup_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^n y_i' y_i + \sup_{x \in C} \{f_0(x) + \delta_{CT(y)}(x)\} \right\} \\ &= \sup_{y \in \mathbb{R}^n} \sup_{x \in \delta_{CT(y)}} \left\{ f_0(x) + \sum_{i=1}^n y_i' y_i \right\}. \end{aligned}$$

Recall the Slater condition, if $y_i' > 0$, then $(-v)^*(y') = +\infty$. Otherwise we have $-y' \in \mathbb{R}_+^n$ and

$$\begin{aligned} (-v)^*(y') &= \sup_{x \in C} \sup_{y, \delta_{CT(y) \ni x}} \left\{ f_0(x) + \sum_{i=1}^n y_i' y_i \right\} \\ &= \sup_{x \in C} \left\{ f_0(x) + \sup_{y, \delta_{CT(y) \ni x}} \sum_{i=1}^n y_i' y_i \right\} \\ &= \sup_{x \in C} \left\{ f_0(x) + \sum_{i=1}^n y_i' f_i(x) \right\}. \end{aligned}$$

Now, there is

$$\inf_{y' \in \mathbb{R}^n} \{(-v)^*(y')\} = \inf_{y' \in \mathbb{R}_+^n} \sup_{x \in C} \left\{ f_0(x) - \sum_{i=1}^n y_i' f_i(x) \right\}.$$

We obtain

$$\begin{aligned} \lambda \in M \Longleftrightarrow (-v)^*(-\lambda) &= \sup_{x \in C} \left\{ f_0(x) - \sum_{i=1}^n \lambda_i f_i(x) \right\} \\ &= \inf_{y' \in \mathbb{R}_+^n} \sup_{x \in C} \left\{ f_0(x) - \sum_{i=1}^n y_i' f_i(x) \right\}, \end{aligned}$$

i.e. $\lambda \in \arg\min_{\lambda \in \mathbb{R}_+^n} \sup_{x \in C} L(x, \lambda)$.

part 4. If the primal problem (4) has a solution x_0, then there exists a Lagrangian multiplier λ_0

such that (x_0, λ_0) is a saddle point of the Lagrangian $L(\cdot,\cdot)$, i.e.

$$\min_{\lambda\in\mathbb{R}^n_+} L(x_0,\lambda) = L(x_0,\lambda_0) = \max_{x\in C} L(x,\lambda_0).$$

Assume that x_0 solves $\sup_{x\in C}\inf_{\lambda\in\mathbb{R}^n_+} L(x,\lambda)$. By the Slater condition, there is no duality gap and we have

$$\begin{aligned}\inf_{\lambda\in\mathbb{R}^n_+} L(x_0,\lambda) &= \sup_{x\in C}\inf_{\lambda\in\mathbb{R}^n_+} L(x,\lambda)\\ &= \inf_{\lambda\in\mathbb{R}^n_+}\sup_{x\in C} L(x,\lambda)\\ &= \sup_{x\in C} L(x,\lambda_0).\end{aligned}$$

The last equation is due to the fact that by the Slater condition, $v(0) = f_0(x_0) \in \mathbb{R}$, so the result of **part** 3 guarantees that the dual problem has solutions. Since the inequality $\inf_{\lambda\in\mathbb{R}^n_+} L(x_0,\lambda) \leqslant L(x_0,\lambda_0) \leqslant \sup_{x\in C} L(x,\lambda_0)$ is always true, we obtain $\inf_{\lambda\in\mathbb{R}^n_+} L(x_0,\lambda) = L(x_0,\lambda_0) = \sup_{x\in C} L(x,\lambda_0)$. That is, x_0 is a solution of $\sup_{x\in C} L(x,\lambda_0)$ and λ_0 is a solution of $\inf_{\lambda\in\mathbb{R}^n_+} L(x_0,\lambda)$. This finishes the proof. ■

Example 4.2 (Distributionally robust regression)**.** Consider the following abstract distribitionally robust estimation problem:

$$\inf_{\theta\in\Theta}\sup_{\mu\in\mathscr{P}}\left\{\int_{\mathcal{X}} \ell(x,\theta(x))\,d\mu(x) : C(\mu,\nu) \leqslant \rho\right\}, \tag{14}$$

where Θ is the parameter set (set of candidate models), and $C(\cdot,\cdot)$ is a measurement of distance between probability measures. We specify $C(\cdot,\cdot)$ to transport cost, i.e.

$$C(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{X}} c(x,y)\,d\pi(x,y)$$

for $\Pi(\mu,\nu)$ the set of probability with marginals both μ and ν.

It is not hard to see that in order to solve (14), a major step is to first handle the inner supremum problem. In this work, we will give a short proof of an important duality proposed in the Theorem 1 of Blanchet and Murthy (2019) which turns out to be a direct consequence of Theorem 4.2.1. The duality result can be concisely stated as, under mild regularity conditions:

$$\sup_{\mu\in\mathscr{P}}\left\{\int f(x)\,d\mu(x) : C(\mu,\nu)\leqslant\rho\right\} = \inf_{\lambda\geqslant 0}\left\{\lambda\rho + \int\sup_{x\in\mathcal{X}}\{f(x) - \lambda c(x,y)\}\,d\nu(y)\right\}. \tag{15}$$

This duality result is highly valued in current literature of distributionally robust optimization.

The alternative proofs and interpretations of this result are the main themes in Gao and Kleywegt (2023), Zhen et al. (2025) and Zhang et al. (2025). We will provide a clearer and still general discussion to (15).

Let $\mathcal{X}$ be a metric space, without explicit notice we always equip $\mathcal{X}$ with its Borel σ-algebra and measures mentioned are Borel measures by default. The elements of $\Pi_{\mu,\nu}$ will be called couplings of (μ,ν). A weaker alternative of (15) is

$$\sup_{\mu\in\mathscr{P}}\left\{\int f(x)\,d\mu(x) : C(\mu,\nu)\leqslant\rho\right\}=\inf_{\lambda\geqslant 0}\left\{\lambda\rho+\sup_{\mu\in\mathscr{P}(\mathcal{X})}\sup_{\pi\in\Pi(u,v)}\left\{\int\left(f(x)-\lambda c(x,y)\right)d\pi(x,y)\right\}\right\}. \tag{16}$$

We need two quite weak assumptions to get a theory:

Assumption 1. *$c:\mathcal{X}\times\mathcal{X}\to\mathbb{R}_+\cup\{+\infty\}$ be a lower semicontinuous cost function.*

A conventional choice of $c(\cdot,\cdot)$ might be the metric $d(x,y)$ on $\mathcal{X}$. Let $p\in[1,+\infty)$, we will get the p-th power of the p-Wasserstein distance,

$$C(\mu,\nu)=W_p^p(\mu,\nu)=\inf_{\pi\in\Pi(\mu,\nu)}\int d(x,y)^p\,d\pi(x,y).$$

We may ought to call $W_p(\mu,\nu)$ as Kantorovich-Rubinstein distance when $p=1$.

Assumption 2. *$f:x\to\{-\infty\}\cup\mathbb{R}$ is upper semicontinuous.*

In this work, we adopt the common convention in optimal transportation to let $+\infty-(+\infty)=-\infty$, see Remark 5.6 in Villani (2009). Therefore if $\mu f^+=\mu f^-=+\infty$ then this probability μ will not likely to be picked by maximization. Now we state the result:

Theorem 4.2.2 (Transport cost distributionally robust duality)**.** *Let $\mathcal{X}$ be a separable Radon space (see for example Definition T.1.1). Assume that there exists a μ_0 such that $\mu_0 f\in\mathbb{R}$ and $C(\mu_0,\nu)<\rho$. Under Assumption 1 and Assumption 2, we have:*

1. *The duality (15).*
2. *Existence of the solution of the dual infimum problem.*
3. *Existence of the solution of the primal supremum problem is equivalent to the existence of saddle point of the Lagrangian*

$$L(\mu,\lambda)=\int f(x)\,d\mu(x)-\lambda\left(C(\mu,\nu)-\rho\right),$$

and is further equivalent to the existence of $\bar{\pi}$ *and* $\bar{\lambda} \geqslant 0$ *such that*

$$\begin{cases} \underbrace{\bar{\pi} c - \rho \leqslant 0}_{\text{(feasibility)}}, \\ \underbrace{\bar{\lambda}\left(\bar{\pi} c - \rho\right) = 0}_{\text{(complementary slackness)}}, \\ \underbrace{\bar{\pi}\left\{(x, y) : f(x) - \bar{\lambda} c(x, y) = \sup_{x'}\left\{f\left(x'\right) - \bar{\lambda} c\left(x', y\right)\right\}\right\} = 1}_{\text{(concentration on the optimal graph)}}, \end{cases}$$

and $\bar{\pi}$ *has a marginal* ν.

4. *If* $\mathcal{X}$ *is compact, then there exists a solution for the primal problem.*

The assumption of separable Radon space (Schwartz, 1973) is a bit more general than the common Polish space requirement in Blanchet and Murthy (2019) and Gao and Kleywegt (2023). The essence of a Radon space $\mathcal{X}$ is that each probability $\mu \in \mathscr{P}(\mathcal{X})$ is tight, which is a fine property of Polish spaces. For more about Radon spaces, see LeCam (1957) and Dudley (2002) chapter 11.

Proof of Theorem 4.2.2.

step 1 The primal problem is a concave maximization problem. Denote the space of signed finite Borel measures on $\mathcal{X}$ as $\mathscr{M}(\mathcal{X})$. Obviously, $\mathscr{P}(\mathcal{X})$ is a convex subset of $\mathscr{M}(\mathcal{X})$. Since μf is linear, we only need to show that $C(\mu, \nu)$ is convex with respect to μ.

By definition of separable Radon space, all probability $\mu \in \mathscr{P}(\mathcal{X})$ is tight. One can check that Theorem 4.1 (existence of an optimal coupling in Polish spaces) in Villani (2009) still holds. For $\mu_1, \mu_2 \in \mathscr{P}(\mathcal{X})$, there are couplings π_1 and π_2 such that

$$C(\mu_1, \nu) = \int c(x, y)\, d\pi_1(x, y) \quad \text{and} \quad C(\mu_2, \nu) = \int c(x, y)\, d\pi_2(x, y).$$

We obtain that for all $\lambda \in [0, 1]$,

$$C\left(\lambda\mu_1 + (1-\lambda)\mu_2\right) \leqslant \int c(x, y)\, d\left(\lambda\pi_1 + (1-\lambda)\pi_2\right)(x, y) = \lambda C(\mu_1, \nu) + (1-\lambda) C(\mu_2, \nu).$$

step 2 There is duality (16). Given the Slater condition, by Theorem 4.2.1 we have

$$\sup_{\mu \in \mathscr{P}(\mathcal{X})} \inf_{\lambda \geqslant 0} L(\mu, \lambda) = \inf_{\lambda \geqslant 0} \sup_{\mu \in \mathscr{P}(\mathcal{X})} L(\mu, \lambda).$$

For $\mu \in \mathscr{P}(\mathcal{X})$, a observation in Sinha et al. (2017) is that

$$\begin{aligned}
&\int f(x)\, d\mu(x) - \lambda\left[C(\mu,\nu) - \rho\right] \\
&\qquad = \int f(x)\, d\pi(x,y) - \lambda\left[C(\mu,\nu) - \rho\right] \quad (\forall \pi \in \Pi(\mu,\nu)) \\
&\qquad = \int f(x)\, d\pi^*(x,y) - \lambda\left[\int c(x,y)\, d\pi^*(x,y) - \rho\right] \\
&\qquad = \sup_{\pi \in \Pi(u,v)} \left\{ \int \left(f(x) - \lambda c(x,y)\right) d\pi(x,y) \right\} + \lambda\rho,
\end{aligned}$$

which gives (16). In the above, the existence of the optimal transference is again guaranteed by the variant of Theorem 4.1 in Villani (2009) for separable Radon spaces.

step 3 Duality (15) when $\mathcal{X}$ is compact. By Assumption 1 and Assumption 2, $f(x) - \lambda c(x,y)$ is an upper semicontinuous function on $\mathcal{X} \times \mathcal{X}$. By Proposition 7.33 in Bertsekas and Shreve (1996) (see Lemma T.1.9), there exists a Borel measurable function $\varphi : \mathcal{X} \to \mathcal{X}$ such that

$$f(\varphi(y)) - \lambda c(\varphi(y), y) = \max_{x \in \mathcal{X}} \left\{ f(x) - \lambda c(x,y) \right\}.$$

Let $\mu_y(x) = \delta_{\varphi(y)}(x)$. Since $\varphi(\cdot)$ is Borel, for all $B \in \mathscr{B}(\mathcal{X})$, the map $y \mapsto \mu_y(A)$ is also Borel measurable. By the Theorem 2.14 in Dellacherie and Meyer (1978), $\{\mu_y\}_y$ and ν can be uniquely glued into a $\bar{\pi}$ on $\mathcal{X} \times \mathcal{X}$. By disintegration of probability measures Theorem 5.3.1 in Ambrosio et al. (2005), we have that for all $\pi \in \mathscr{P}(\mathcal{X} \times \mathcal{X})$ and π has a marginal ν,

$$\begin{aligned}
\int \left(f(x) - \lambda c(x,y)\right) d\pi &= \iint \left[f(x) - \lambda c(x,y)\right] d\pi(x|y)\, d\nu(y) \\
&\leqslant \int \max_{x \in \mathcal{X}} \left\{ f(x) - \lambda c(x,y) \right\} d\nu(y) \\
&= \int \left(f(x) - \lambda c(x,y)\right) d\bar{\pi}.
\end{aligned}$$

step 4 Duality (15) for general case. Decompose $f - \lambda c$ to $(f - \lambda c)^+ - (f - \lambda c)^-$ if necessary. The main difficulty in the case comes from the possible non-measurability of $\sup_{x \in \mathcal{X}} \{f(x) - \lambda c(x,y)\}$. For $\mu \in \mathscr{P}(\mathcal{X})$ and $\pi \in \Pi(\mu,\nu)$, consider the disintegration of measure again. We have

$$\int \left(f(x) - \lambda c(x,y)\right) d\pi(x|y) \leqslant \left(\sup_{x \in \mathcal{X}} \left\{ f(x) - \lambda c(x,y) \right\} \right)_*$$

since the left hand side is Borel measurable, where T_* denotes a maximal measurable minorant of T. See, for example chapter 1.2 in van der Vaart and Wellner (2023). Then we can apply a

standard step functions approximation and the gluing we use in **step** 3 to obtain

$$\sup_{\mu\in\mathscr{P}(\mathcal{X})}\sup_{\pi\in\Pi(\mu,\nu)}\left\{\int\left(f\left(x\right)-\lambda c\left(x,y\right)\right)d\pi\left(x,y\right)\right\}=\int\left(\sup_{x\in\mathcal{X}}\left\{f\left(x\right)-\lambda c\left(x,y\right)\right\}\right)_{*}d\nu\left(y\right).$$

step 5 The solutions of the dual problem. If the value function is finite at ρ, then we know the dual problem has solutions by Theorem 4.2.1 *2.* If the value function is $+\infty$ at ρ, i.e. there exists $\bar{\mu}\in\mathscr{P}\left(\mathcal{X}\right)$ such that $C\left(\bar{\mu},\nu\right)\leqslant\rho$ and $\bar{\mu}f=+\infty$, then we can simply let $\lambda=0$.
step 6 We show *3.* The first half is directly guaranteed by Theorem 4.2.1 *1.* and *3.* and the result of **step** 5. If the primal problem has a solution $\bar{\mu}$, then

$$\inf_{\lambda\geqslant0}L\left(\bar{\mu},\lambda\right)=\sup_{\mu\in\mathscr{P}(\mathcal{X})}L\left(\mu,\bar{\lambda}\right),$$

where $\bar{\lambda}$ is a solution of the dual problem, i.e. $\left(\bar{\mu},\bar{\lambda}\right)$ is a saddle point. If the value function is $+\infty$ at ρ, let $\bar{\lambda}=0$. Otherwise, note we are in the case of Theorem 4.1.1, so we always have the complementary slackness condition. Let $\bar{\pi}$ be the optimal transference between $\bar{\mu}$ and ν, then we obtain

$$\begin{aligned}\bar{\pi}\left(f-\bar{\lambda}c\right)&=\sup_{\pi\in\Pi(\bar{\mu},\nu)}\left\{\int\left(f\left(x\right)-\bar{\lambda}c\left(x,y\right)\right)d\pi\left(x,y\right)\right\}\\&=\int\left(\sup_{x\in\mathcal{X}}\left\{f\left(x\right)-\bar{\lambda}c\left(x,y\right)\right\}\right)_{*}d\nu\left(y\right).\end{aligned}$$

For the converse direction, if we have $\left(\bar{\pi},\bar{\lambda}\right)$ of concentration on the optimal graph and complementary slackness properties. Then, for the first coordinate projection $\bar{\mu}$ of $\bar{\pi}$,

$$\begin{aligned}L\left(\bar{\mu},\bar{\lambda}\right)&=\sup_{\mu\in\mathscr{P}(\mathcal{X})}L\left(\mu,\bar{\lambda}\right)\quad\text{(by concentration on the optimal graph)}\\&=\int f\left(x\right)d\bar{\mu}-\bar{\lambda}\left(C\left(\bar{\mu},\nu\right)-\rho\right)\\&=\inf_{\lambda\geqslant0}L\left(\bar{\mu},\lambda\right)\quad\text{(by feasibility and complementary slackness)}.\end{aligned}$$

This finishes the proof. ■

Example 4.3 (Distributionally robust optimal allocation)**.** In this example, we discuss the duality results for transport cost distributionally robust optimal allocation problems. There are definitely various of settings for the topic, see for example Kido (2022), Adjaho and Christensen (2022) and Shen et al. (2023). We derive the dual expressions only formally, which means that we will not check the regularity conditions for each step in detail. In other words, here we are satisfied to only provide intuitive but **not** mathematically rigorous results. For brevity, we denote $\left(\mathcal{X}\times\mathcal{Y}\right)$ as $\mathcal{Z}$.

Q and Q' are distributions on $\mathcal{Z}$.

Consider the problem in the following form

$$\sup_{\phi\in\Phi}\inf_{Q'\in\mathscr{P}(\mathcal{Z})}\left\{\mathbb{E}_{Q'}\left[\sum_{j=0}^{J-1}\lambda_j\phi_j\left(X\right)Y_j\right]:C\left(Q',Q\right)\leqslant\delta\right\}, \tag{17}$$

for $\delta>0$ and $C\left(Q',Q\right)$ is the transport cost

$$\min_{\pi\in\Pi(Q',Q)}\int_{\mathcal{Z}\times\mathcal{Z}}c\left(z_1,z_2\right)d\pi\left(z_1,z_2\right).$$

An alternative distributionally robust problem is handled in Kido (2022),

$$\sup_{\phi\in\Phi}\inf_{Q'\in\mathscr{P}(\mathcal{Z})}\left\{\mathbb{E}_{X_{\#}Q'}\mathbb{E}_{Q'|X}\left[\sum_{j=0}^{J-1}\lambda_j\phi_j\left(X\right)Y_j\right]:C\left(Q'|X=x,Q|X=x\right)\leqslant\delta,\forall x\in\mathcal{X}\wedge X_{\#}Q'=X_{\#}Q\right\}, \tag{18}$$

Consider the "marginal fixed, conditional shifted" problem in (18). Note that given a $\phi\in\Phi$ we can handle the inner constrained minimization problem first. By Theorem 4.2.2,

$$\begin{aligned}
&\inf_{Q'\in\mathscr{P}(\mathcal{Z})}\left\{\mathbb{E}_{X_{\#}Q'}\mathbb{E}_{Q'|X}\left[\sum_{j=0}^{J-1}\lambda_j\phi_j\left(X\right)Y_i\right]:Q'\in\mathscr{T}\left(Q,\delta\right)\right\}\\
&=\mathbb{E}_{X_{\#}Q}-\inf_{\lambda_x'\geqslant0}\left\{\lambda_x'\delta+\mathbb{E}_{Q|X=x}\sup_{y'\in\mathcal{Y}}\left\{-\left[\sum_{j=0}^{J-1}\lambda_j\phi_j\left(x\right)y_j'\right]-\lambda_x'c\left(\left(x,y'\right),\left(x,Y\right)\right)\right\}\right\}\\
&=\mathbb{E}_{X_{\#}Q}\sup_{\lambda_x'\geqslant0}\left\{-\lambda_x'\delta+\mathbb{E}_{Q|X=x}\inf_{y'\in\mathcal{Y}}\left\{\sum_{j=0}^{J-1}\lambda_j\phi_j\left(x\right)y_j'+\lambda_x'c\left(\left(x,y'\right),\left(x,Y\right)\right)\right\}\right\},
\end{aligned}$$

where $\mathscr{T}\left(Q,\delta\right)$ is the constrained in (18). It suffices to solve the infimum given fixed (x,y) and $\phi\left(\cdot\right)$, denote

$$I_{\phi,x,\lambda,\lambda_x'}:=\inf_{y'\in\mathcal{Y}}\left\{\sum_j\lambda_j\phi_j\left(x\right)y_j'+\lambda_x'c\left(\left(x,y'\right),\left(x,y\right)\right)\right\}.$$

If we take $c\left(\cdot,\cdot\right)$ as the ℓ_1 distance on $\mathcal{Y}\subset\mathbb{R}^J$, then we will obtain

$$I_{\phi,x,\lambda,\lambda_x'}=\sum_j\left[1\left(\lambda_j\phi_j\left(x\right)>\lambda_x'\right)\left(\left(\lambda_j\phi_j\left(x\right)-\lambda_x'\right)\inf\mathcal{Y}+\lambda_x'y_j\right)+1\left(\lambda_j\phi_j\left(x\right)\leqslant\lambda_x'\right)\lambda_j\phi_j\left(x\right)y_j\right] \tag{19}$$

as in Kido (2022). If we normalize $\inf \mathcal{Y} = 0$ and $\lambda = (1, \ldots, 1)$, then we obtain

$$\mathbb{E}_{Q|X=x} I_{\phi,x,1,\lambda'_x} = \sum_j \left[1\left(\phi_j(x) > \lambda'_x\right) \lambda'_x p_j(x) + 1\left(\phi_j(x) \leqslant \lambda'_x\right) \phi_j(x) p_j(x) \right].$$

If $\lambda' \geqslant 1$, then the optimal policy satisfies

$$\phi_k = \begin{cases} 1, & \text{when} \quad p_k(x) > \max_{l \neq k, l=0,\ldots,J-1} \{p_l(x)\}, \\ 0, & \text{when} \quad p_k(x) < \max_{l \neq k, l=0,\ldots,J-1} \{p_l(x)\}. \end{cases}$$

If $\lambda' < 1$, then ϕ should be allocate according to the ranking of $(p_0(x), \ldots, p_{J-1}(x))$.

More generally, we can apply (16) to get:

$$\begin{aligned}
&\sup_{\phi \in \Phi} \sup_{\lambda' \geqslant 0} \left\{ -\lambda\delta + \inf_{Q' \in \mathscr{P}(Z)} \inf_{\pi \in \Pi(Q',Q)} \mathbb{E}_\pi \left[\sum_{j=0}^{J-1} \lambda_j \phi_j\left(X'\right) Y'_j + \lambda' c\left(\left(X', Y'\right), (X, Y)\right) \right] \right\} \\
&= \sup_{\lambda \geqslant 0} \sup_{\phi \in \Phi} \left\{ -\lambda'\delta + \inf_{Q' \in \mathscr{P}(Z)} \inf_{\pi \in \Pi(Q',Q)} \mathbb{E}_\pi \left[\sum_{j=0}^{J-1} \lambda_j \phi_j\left(X'\right) Y'_j + \lambda' c\left(\left(X', Y'\right), (X, Y)\right) \right] \right\} \\
&= \sup_{\lambda \geqslant 0} \left\{ -\lambda'\delta + \sup_{\phi \in \Phi} \inf_{Q' \in \mathscr{P}(Z)} \inf_{\pi \in \Pi(Q',Q)} \mathbb{E}_\pi \left[\sum_{j=0}^{J-1} \lambda_j \phi_j\left(X'\right) Y'_j + \lambda' c\left(\left(X', Y'\right), (X, Y)\right) \right] \right\} \\
&= \sup_{\lambda \geqslant 0} \left\{ -\lambda'\delta + \inf_{Q' \in \mathscr{P}(Z)} \inf_{\pi \in \Pi(Q',Q)} \sup_{\phi \in \Phi} \mathbb{E}_\pi \left[\sum_{j=0}^{J-1} \lambda_j \phi_j\left(X'\right) Y'_j + \lambda' c\left(\left(X', Y'\right), (X, Y)\right) \right] \right\} \\
&= \sup_{\lambda \geqslant 0} \left\{ -\lambda'\delta + \inf_{Q' \in \mathscr{P}(Z)} \left\{ \mathbb{E}_{X_\# Q'} \left[\max_{j=0,\ldots,J-1} \{\lambda_j p'_j\left(X'\right)\} \right] + \lambda' C\left(Q', Q\right) \right\} \right\}.
\end{aligned}$$

Note that if $p'(\cdot) = p(\cdot)$, then the optimal policy is just the same as (10), and we can assume ϕ to be nonrandomized without loss of generality. Now, consider

$$I_{\phi,1,\lambda'} = \inf_{z \in \mathcal{Z}} \left\{ \sum_j \phi_j\left(x'\right) y'_j + \lambda' \left(c_{\mathcal{X}}\left(x', x\right) + \|y' - y\|_1 \right) \right\}.$$

For fixed x', if $\phi_j(x') = 0$, then we should let $y'_j = y_j$; if $\phi_j(x') = 1$, then we should let $y'_j = \inf \mathcal{Y}$ when $\lambda' < 1$, otherwise let $y'_j = y_j$ So, if we still normalize $\inf \mathcal{Y} = 0$, we obtain

$$I_{\phi,1,\lambda'} = \inf_{x' \in \mathcal{X}} \left\{ \lambda' c_{\mathcal{X}}\left(x', x\right) + \sum_j 1\left(\phi_j\left(x'\right) = 1\right) \left[1\left(\lambda' < 1\right) \lambda' p_j(x) + 1\left(\lambda' \geqslant 1\right) p_j(x) \right] \right\}.$$

Therefore, we need to consider

$$\sup_{\phi \in \Phi, \phi \text{ nonrandomized}} \sup_{\lambda' \geqslant 0} \left\{ -\lambda'\delta + \mathbb{E}_Q I_{\phi,1,\lambda'} \right\}.$$

As pointed out in Shen et al. (2023), the transport cost distributionally robust optimal allocation problem is unlikely to be solvable in its full generality. By the above calculation, we show that the difficulty may still remain even if we assume that

$$c\left(\left(x',y'\right),(x,y)\right)=c_{\mathcal{X}}\left(x',x\right)+\|y'-y\|_1.$$

A suitable setting for transport cost distributionally robust optimal allocation can be

$$\sup_{\phi\in\Phi}\inf_{Q'\in\mathscr{P}(\mathcal{Z})}\left\{\mathbb{E}_{X_{\#}Q'}\mathbb{E}_{Q'|X}\left[\sum_{j=0}^{J-1}\lambda_j\phi_j\left(X\right)Y_i\right]:C\left(Q'|X=x,Q|X=x\right)\leqslant\delta_1,\forall x\in\mathcal{X}\wedge C\left(X_{\#}Q',X_{\#}Q\right)\leqslant\delta_2\right\}$$

if $X_{\#}Q$ is supported on the whole $\mathcal{X}$. Formally, we have

$$\begin{aligned}
&\inf_{Q'\in\mathscr{P}(\mathcal{Z})}\left\{\mathbb{E}_{X_{\#}Q'}\mathbb{E}_{Q'|X}\left[\sum_{j=0}^{J-1}\lambda_j\phi_j\left(X\right)Y_i\right]:C\left(Q'|X=x,Q|X=x\right)\leqslant\delta_1,\forall x\in\mathcal{X}\wedge C\left(X_{\#}Q',X_{\#}Q\right)\leqslant\delta_2\right\}\\
=&\inf_{\substack{X_{\#}Q'\in\mathcal{X},\\ C\left(X_{\#}Q',X_{\#}Q\right)\leqslant\delta_2}}\mathbb{E}_{X_{\#}Q'}\sup_{\lambda'_x\geqslant0}\left\{-\lambda'_x\delta_1+\mathbb{E}_{Q|X=x}\inf_{y'\in\mathcal{Y}}\left\{\sum_{j=0}^{J-1}\lambda_j\phi_j\left(x\right)y'_j+\lambda'_xc\left(\left(x,y'\right),\left(x,Y\right)\right)\right\}\right\}\\
=&\sup_{\lambda'\geqslant0}\left\{-\lambda'\delta_2+\mathbb{E}_{X_{\#}Q}\inf_{x'\in\mathcal{X}}\left\{\sup_{\lambda'_{x'}\geqslant0}\left\{-\lambda'_{x'}\delta_1+\mathbb{E}_{Q|X=x'}I_{\phi,x,\lambda,\lambda'_{x'}}\right\}-\lambda'c_{\mathcal{X}}\left(x',x\right)\right\}\right\}.
\end{aligned}$$

Then we obtain that

$$\begin{aligned}
&\sup_{\phi\in\Phi}\sup_{\lambda'\geqslant0}\left\{-\lambda'\delta_2+\mathbb{E}_{X_{\#}Q}\inf_{x'\in\mathcal{X}}\left\{\sup_{\lambda'_{x'}\geqslant0}\left\{-\lambda'_{x'}\delta_1+\mathbb{E}_{Q|X=x'}I_{\phi,x,\lambda,\lambda'_{x'}}\right\}-\lambda'c_{\mathcal{X}}\left(x',x\right)\right\}\right\}\\
=&\sup_{\lambda'\geqslant0}\left\{-\lambda'\delta_2+\sup_{\phi\in\Phi}\mathbb{E}_{X_{\#}Q}\inf_{x'\in\mathcal{X}}\left\{\sup_{\lambda'_{x'}\geqslant0}\left\{-\lambda'_{x'}\delta_1+\mathbb{E}_{Q|X=x'}I_{\phi,x,\lambda,\lambda'_{x'}}\right\}-\lambda'c_{\mathcal{X}}\left(x',x\right)\right\}\right\}.
\end{aligned}$$

Note that we can pointwisely maximize $\mathbb{E}_{Q|X=x'}I_{\phi,x,\lambda,\lambda'_{x'}}$ to maximize the inner expectation above. Therefore, under correct regularity conditions and suitably specified setting on $c\left(\cdot,\cdot\right)$, we can still obtain the optimality of certain threshold policy.

5 On Differential Approaches

5.1 Differentiability of Probability Function Method

Assumption 5.1.1. *The function f is continuous with compact support $K_f\subset E$.*

Assumption 5.1.2. *The support K_f consists of only regular points of the function h.*

Assumption 5.1.1′. *For all $c \in h(E)$, there exists a neighborhood $\mathcal{N}_c$ of c such that*

$$\overline{\bigcup_{c' \in \mathcal{N}_c} h^{-1}(c')} \subset E$$

and is bounded. The function f is continuous.

In addition, when $k = 1$ and when we only need Hadamard differentiability on a singleton $\{c\}$, Assumption 5.1.2 can be replaced by the following

Assumption 5.1.2′. *$c \in h(E)$ and $K_f \cap h^{-1}(c)$ consists of only regular points of h.*

Assumption 5.1.3. *The function f is Borel and bounded with compact support $K_f \subset E$.*

Assumption 5.1.4. *The Clarke Jacobian $J_c h(x)$ (defined in T.1.9 in the Technical addendum) is of full rank for all $x \in E$. For all x, there exists a neighborhood B_x and a constant $\epsilon(x) > 0$, such that for all $x' \in B_x$, if $\nabla h(x')$ exists then $Jh(x') > \epsilon(x)$.*

Assumption 5.1.5. *For all $c \in h(E)$, $\mathcal{H}_{n-k}\{x : h(x) = c, h(\cdot)$ is not differentiable at $x\} = 0$ and $\mathcal{H}_{n-k}\left\{x : h(x) = c, \exists x_{(k)}\left(\left|\det\left(\frac{\partial}{\partial x_{(k)}} h(\cdot)\right)\right| > 0 \text{ and } \frac{f(\cdot)}{\left|\det\left(\frac{\partial}{\partial x_{(k)}} h(\cdot)\right)\right|} \text{is discontinuous at } x\right)\right\} = 0$, where the subscript (k) denotes k-combination of the n coordinates.*

Denote the sorting operator in Chernozhukov et al. (2018b) as

$$F(h, c) := \int_{h(x) > c} f(x)\, d\mathcal{L}_n x = \int 1(h(x) > c)\, f(x)\, d\mathcal{L}_n x. \tag{20}$$

Theorem 5.1.1. *Let $h : E \to \mathbb{R}^k$ be a C^1 function, $k \in \{1, 2, \ldots, n\}$, E an open subset of $\mathbb{R}^n$. Let Assumption 5.1.1 and Assumption 5.1.2 hold. Let $\mathcal{D} \subset h(E)$ be a compact subset. Then, the map $F(h, c) : C\left(E, \mathbb{R}^k\right) \to \mathbb{R}$ in (20) is Hadamard differentiable uniformly with respect to $c \in \mathcal{D}$ at h tangentially to $C\left(E, \mathbb{R}^k\right)$. When $k > 1$, the derivative is given by*

$$F'_{h,c}(H, 0) = \sum_i \int_{y > \tau_{\neg i}(c)} \left[\int_{h^{-1}(c'(y, c_i; i))} \frac{H_i(x) f(x)}{Jh(x)} d\mathcal{H}_{n-k} x\right] d\mathcal{L}_{k-1} y, \tag{21}$$

where $\tau_{\neg i}(c)$ denotes the $\mathbb{R}^k \to \mathbb{R}^{k-1}$ coordinate projection except the i-th coordinate, $c'(y, c_i; i)$ is the k-dimensional vector such that $\tau_{\neg i}(c'(y, c_i; i)) = y$ and $\tau_i(c'(y, c_i; i)) = c_i$. When $k = 1$, i.e.

when h is a scalar-valued function, the derivative is given by

$$F'_{h,c}(H,0) = \int_{h^{-1}(c)} \frac{H(x) f(x)}{\|\nabla h(x)\|} d\mathcal{H}_{n-1}x. \tag{22}$$

Proposition 5.1.2. *Under the conditions in Theorem 5.1.1, the density of the random variable $h(X)$ is continuous on $h(E)$ and is given by*

$$\int_{h^{-1}(c)} \frac{f(x)}{Jh(x)} d\mathcal{H}_{n-k}x.$$

Theorem 5.1.3. *Let $h : E \to \mathbb{R}^k$ be Lipschitz, $k \in \{1,2,\ldots,n\}$, E an open subset of $\mathbb{R}^n$. Let Assumption 5.1.3 to Assumption 5.1.5 hold. Let $\mathcal{D} \subset h(E)$ be a compact subset. Then the map $F(h,c) : C\left(E,\mathbb{R}^k\right) \to \mathbb{R}$ is Hadamard differentiable uniformly on $\mathcal{D}$ at h tangentially to $C\left(E,\mathbb{R}^k\right)$. When $k > 1$, the derivative is given by (21). When $k = 1$, the derivative is given by (22).*

PROOF. We will prove these three results in several parts. Before proceeding to the more general case of $k > 1$ in Theorem 5.1.1 (**part** 2), we first prove a pointwise version of the special case when $k = 1$ (**part** 1). Proposition 5.1.2 can be proved by a similar derivation in passing (**part** 1 **step** 5). We then demonstrate the uniformity of convergence of Theorem 5.1.1 (**part** 3). We finalize by elaborating the proof techniques to handle the nonsmoothness in Theorem 5.1.3 (**part** 4).

part 1 First consider the special case of $k = 1$ in Theorem 5.1.1.

step 1 We claim that there exists an $\eta > 0$ small enough such that for all c' with $|c' - c| < \eta$, we can get change of variable formulas simultaneously. Consider

$$\Psi\left(x'_1,\ldots,x'_n,c'\right) := h\left(x'_1,\ldots,x'_n\right) - c'.$$

Without loss of generality, we may assume that $\nabla_{x_1}\Psi(x_1,\ldots,x_n,c)$ is full rank. Then by the implict function theorem (see for example Theorem T.1.19, C^1 case), there exists an open set $B_{x_1} \times B_{x_2,\ldots,x_n,c} \subset U \subset \mathbb{R}^{n+1}$, where U is a neighborhood of $(x_1,\ldots,x_n,c)$, such that for some positive vectors α, β,

$$\begin{aligned} B_{x_1} &= \left\{x'_1 \in \mathbb{R} : |x'_1 - x_1| < \alpha\right\} \\ B_{x_2,\ldots,x_n,c} &= \left\{\left(x'_2,\ldots,x'_n,c'\right) \in \mathbb{R}^n : |\left(x'_2,\ldots,x'_n,c'\right) - (x_2,\ldots,x_n,c)| < \beta\right\}, \end{aligned}$$

and a C^1 implicit function $\xi(\cdot)$ defined on $B_{x_2,\ldots,x_n,c}$ such that

$$\Psi\left(x'_1,\ldots,x'_n,c'\right) = 0 \Leftrightarrow \left(x'_1\right) = \xi\left(x'_2,\ldots,x'_n,c'\right)$$

for all $(x'_1,\ldots,x'_n,c') \in B_{x_1} \times B_{x_2,\ldots,x_n,c}$. For all $x \in h^{-1}(c) \cap K_f$, we can find intervals like

$B_{x_1} \times B_{x_2,\ldots,x_n,c}$ in the above. The implicit function theorem may not be always about the first dimension. However, an implicit function for some $x'_m, m \in \{1,\ldots,n\}$ always exists by the regular point assumption 5.1.2′. By the compactness of $h^{-1}(c) \cap K_f$, there exists a finite open cover denoted as $\{B_j\} := \left\{B_{x_{j,1}} \times B_{x_{j,2},\ldots,x_{j,n}}\right\}$ of $h^{-1}(c) \cap K_f$. Now, we claim that there exists $\eta > 0$, such that for all c' satisfying $|c' - c| < \eta$, $h^{-1}(c') \cap K_f \subset \bigcup_j B_j$. This claim is proven by contradiction. Suppose $\bigcup_j B_j$ will not cover $h^{-1}(c') \cap K_f$ for some $c' \neq c$, such that $|c' - c| < \eta$ where η is arbitrarily chosen. In other words, $\forall \eta > 0$, $\exists c'$, $|c' - c| < \eta$, such that $h^{-1}(c') \cap K_f \nsubseteq \bigcup_j B_j$. Then there exists a sequence $\{(x_i, c_i)\}_{i=1}^{\infty}$ such that

$$x_i \in h^{-1}(c_i) \cap K_f, x_i \notin \bigcup_j B_j.$$

By the compactness of $h^{-1}([a,b]) \cap K_f$ and the Bolzano-Weierstrass theorem, there exists a convergent subsequence $\{x_{i_l}\}_{l=1}^{\infty}$ such that $\lim_{l\to\infty} h(x_{i_l}) = c$ and thus $\lim_{l\to\infty} x_{i_l} \in h^{-1}(c) \subset \bigcup_j B_j$. Therefore, $x_{i_l} \in \bigcup_j B_j$ for all sufficiently large l, which is a contradiction.

step 2 To calculate the change of variables formula analytically, consider

$$\psi_{B_j}\left(x'_{j,2},\ldots,x'_{j,n},c'\right) = \left(\xi_{B_j}\left(x'_{j,2},\ldots,x'_{j,n},c'\right), x'_{j,2},\ldots,x'_{j,n}\right)^T,$$

where $\left(x'_{j,2},\ldots,x'_{j,n}\right)$ are used by the implicit function theorem to obtain the local implicit function $\xi_{B_j}(\cdot)$ for $x'_{j,1}$. Then, by the generalized matrix determinant lemma (for clarity the subscript j is omitted), we have

$$\begin{aligned}
J\psi_{B_j}\left(x'_{2,\ldots,n},c'\right) &= \det\left(\left(\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\nabla_{x_{2,\ldots,n}}h\right)^T, I_{n-1}\right)\begin{pmatrix}\left[\frac{\partial h}{\partial x_1}\right]^{-1}\nabla_{x_{2,\ldots,n}}h\\ I_{n-1}\end{pmatrix}\right)^{\frac{1}{2}}\\
&= \det\left(I_{n-1} + \left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\nabla_{x_{2,\ldots,n}}h\right)^T\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\nabla_{x_{2,\ldots,n}}h\right)\right)^{\frac{1}{2}}\\
&= \left(\det\left(I_{n-1}\right)\det\left(I_1 + \left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\nabla_{x_{2,\ldots,n}}h\right)I_{n-1}\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\nabla_{x_{2,\ldots,n}}h\right)^T\right)\right)^{\frac{1}{2}}\\
&= \left(\det\left(\left[\frac{\partial h}{\partial x_1}\right]\left[\frac{\partial h}{\partial x_1}\right]^T + \left(\nabla_{x_{2,\ldots,n}}h\right)\left(\nabla_{x_{2,\ldots,n}}h\right)^T\right)\right)^{\frac{1}{2}}\left|\det\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\right)\right|\\
&= \left(\det\left(\left(\nabla h\right)\left(\nabla h\right)^T\right)\right)^{\frac{1}{2}}\left|\det\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\right)\right|\\
&= Jh\left|\det\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\right)\right|. \qquad (23)
\end{aligned}$$

In the above, for simplicity, we omit the point at which the derivatives are calculated, where

$$\frac{\partial h}{\partial x_1} = \frac{\partial}{\partial x_1} h\left(\xi_{B_j}\left(x'_{2,\ldots,n}, c'\right), x'_2, \ldots, x'_n\right) \text{ and}$$
$$\nabla_{x_{2,\ldots,n}} h = \nabla_{x_{2,\ldots,n}} h\left(\xi_{B_j}\left(x'_{2,\ldots,n}, c'\right), x'_2, \ldots, x'_n\right).$$

step 3 Consider a sequence $t_n \downarrow 0$. First let $H_n \equiv H$ for all n. By definition of Hadamard differentiability, we need to calculate

$$\lim_{n\to\infty} \frac{1}{t_n}\left[\int 1\left(h\left(x\right) + t_n H\left(x\right) > c\right) f\left(x\right) d\mathcal{L}_n x - \int 1\left(h\left(x\right) > c\right) f\left(x\right) d\mathcal{L}_n x\right].$$

By coarea formula Theorem T.2.2,

$$\begin{aligned}
&\frac{1}{t_n}\left[\int 1\left(h\left(x\right) + t_n H\left(x\right) > c\right) f\left(x\right) d\mathcal{L}_n x - \int 1\left(h\left(x\right) > c\right) f\left(x\right) d\mathcal{L}_n x\right] \\
&= \frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f} \frac{\left(1\left(c' + t_n H\left(x\right) > c\right) - 1\left(c' > c\right)\right) f\left(x\right)}{Jh\left(x\right)} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c' \\
&= \int_{c-t_n M}^{c+t_n M}\left[\frac{1}{t_n}\int_{h^{-1}(c')\cap K_f} \frac{\left(1\left(c' + t_n H\left(x\right) > c\right) - 1\left(c' > c\right)\right) f\left(x\right)}{Jh\left(x\right)} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c',
\end{aligned}$$

where $M = \max_{x\in K_f} |H\left(x\right)| < +\infty$. Then, we can apply the area formula Theorem T.2.1 to calculate, for each j,

$$\begin{aligned}
&\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap B_j\cap K_f} \frac{\left[1\left(c' + t_n H\left(x\right) > c\right) - 1\left(c' > c\right)\right] f\left(x\right)}{Jh\left(x\right)} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c' \\
&= \frac{1}{t_n}\int\left[\int_{B_{x_{j,2},\ldots,x_{j,n}}} \frac{\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] f\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1} h\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)\right)\right|} d\mathcal{L}_{n-1} x\right] d\mathcal{L}_1 c'.
\end{aligned}$$

Next by the Fubini-Tonelli theorem, the above is equal to

$$\int_{B_{x_{j,2},\ldots,x_{j,n}}}\left[\frac{1}{t_n}\int \frac{\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] f\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1} h\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)\right)\right|} d\mathcal{L}_1 c'\right] d\mathcal{L}_{n-1} x.$$

Since

$$\begin{aligned}
&\frac{1}{t_n}\int\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] d\mathcal{L}_1 c' \\
&= \frac{1}{t_n}\int_{c-t_n M}^{c+t_n M}\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] d\mathcal{L}_1 c' \to H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c\right)\right),
\end{aligned}$$

by the dominated convergence theorem (DCT). With the help of the partition of unity theorem T.1.23, we can extend the above local convergence to the entire level set $h^{-1}\left(c\right)$, since for n large

enough, we have

$$\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f}\frac{\left(1\left(c'+t_nH\left(x\right)>c\right)-1\left(c'>c\right)\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x\right]d\mathcal{L}_1c'$$
$$=\frac{1}{t_n}\int\left[\sum_l\int_{h^{-1}(c')\cap K_f}\frac{\left(1\left(c'+t_nH\left(x\right)>c\right)-1\left(c'>c\right)\right)f\left(x\right)\rho_l\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x\right]d\mathcal{L}_1c',$$

where $\rho_l\left(\cdot\right)$ is the partition of unity.
Then by the area formula again,

$$\lim_{n\to\infty}\frac{1}{t_n}\int\left[1\left(h\left(x\right)+t_nH\left(x\right)>c\right)-1\left(h\left(x\right)>c\right)\right]f\left(x\right)d\mathcal{L}_nx=\int_{h^{-1}(c)}\frac{H\left(x\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.$$

step 4 Now consider functions $H\left(x\right)+\zeta$, for arbitrary $\zeta>0$. Since by assumption $\sup_{x\in E}\left|H_n\left(x\right)-H\left(x\right)\right|\to 0$, for all $\zeta>0$, for sufficiently large n, $H\left(x\right)+\zeta>H_n\left(x\right)$ for all x. Therefore,

$$\begin{aligned}\limsup_{n\to\infty}\frac{F\left(h+t_nH_n,c\right)-F\left(h,c\right)}{t_n}&\leqslant\lim_{n\to\infty}\frac{F\left(h+t_n\left(H+\zeta\right),c\right)-F\left(h,c\right)}{t_n}\\&=\int_{h^{-1}(c)}\frac{\left(H\left(x\right)+\zeta\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.\end{aligned}$$

Similarly,

$$\begin{aligned}\liminf_{n\to\infty}\frac{F\left(h+t_nH_n,c\right)-F\left(h,c\right)}{t_n}&\geqslant\lim_{n\to\infty}\frac{F\left(h+t_n\left(H-\zeta\right),c\right)-F\left(h,c\right)}{t_n}\\&=\int_{h^{-1}(c)}\frac{\left(H\left(x\right)-\zeta\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.\end{aligned}$$

By the Hölder inequality (integration by Hausdorff measure is also in the Lebesgue sense):

$$\int_{h^{-1}(c)}\frac{\left|H\left(x\right)-H'\left(x\right)\right|f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x\leqslant\sup_{x\in E}\left|H\left(x\right)-H'\left(x\right)\right|\int_{h^{-1}(c)}\left|\frac{f\left(x\right)}{Jh\left(x\right)}\right|d\mathcal{H}_{n-1}x.$$

Now by the arbitrariness of ζ,

$$\lim_{n\to\infty}\frac{F\left(h+t_nH_n,c\right)-F\left(h,c\right)}{t_n}=\int_{h^{-1}(c)}\frac{H\left(x\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.$$

step 5 Now we can also prove Proposition 5.1.2. Note that the above proof techniques of **step 1** to **step** 3, including the area formula, partition of unity, and generalized matrix determinant lemma, all work for $k>1$. As a result, the proof is essentially a reconstruction of the proof of the

special case of $k=1$. Here, we only need to point out that

$$\lim_{c'\to c}\int_{B_{x_{j,2},\ldots,x_{j,n}}}\frac{f\left(\psi_{B_j}\left(x_{2,\ldots,n},c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1}h\left(\psi_{B_j}\left(x_{2,\ldots,n},c'\right)\right)\right)\right|}d\mathcal{L}_{n-1}x$$
$$=\int_{B_{x_{j,2},\ldots,x_{j,n}}}\frac{f\left(\psi_{B_j}\left(x_{2,\ldots,n},c\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1}h\left(\psi_{B_j}\left(x_{2,\ldots,n},c\right)\right)\right)\right|}d\mathcal{L}_{n-1}x.$$

part 2 For $k>1$, we still have, by the coarea formula

$$F\left(h,c\right)=\int_{h(x)>c}f\left(x\right)d\mathcal{L}_nx=\int_{c'>c}\left[\int_{h^{-1}(c')}\frac{f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}\right]d\mathcal{L}_kc'.$$

So we have that

$$\frac{1}{t_n}\int\left[1\left(h\left(x\right)+t_nH_n\left(x\right)>c\right)-1\left(h\left(x\right)>c\right)\right]f\left(x\right)d\mathcal{L}_nx$$
$$=\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f}\frac{\left[1\left(c'+t_nH_n\left(x\right)>c\right)-1\left(c'>c\right)\right]f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}x\right]d\mathcal{L}_kc'.$$

By the telescoping identity of higher order expansion,

$$\prod_{i=1}^{k}a_i'-\prod_{i=1}^{k}a_i=\sum_{i=1}^{k}\left(a_i'-a_i\right)\prod_{l\neq i}a_l+\sum_{i\neq j}\left(a_i'-a_i\right)\left(a_j'-a_j\right)\prod_{l\neq i,l\neq j}a_l+\cdots+\prod_{i=1}^{k}\left(a_i'-a_i\right),$$

using a proof procedure similar to that of the special case of $k=1$, the first order term of the difference becomes

$$\sum_i\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f}\frac{\left[1\left(c_i'+t_nH_{n,i}\left(x\right)>c_i\right)-1\left(c_i'>c_i\right)\right]\prod_{l\neq i}1\left(c_l'>c_l\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}x\right]d\mathcal{L}_kc'$$
$$=\sum_i\sum_{l_i}\int_{y>\tau_{\neg i}(c)}\int_{B_{l_i}}d\mathcal{L}_{n-k}xdL_{k-1}y$$
$$\left[\int\frac{\left[1\left(c_i'+t_nH_{n,i}\left(\psi_{B_{l_i}}\left(x,c'\right)\right)>c_i\right)-1\left(c_i'>c_i\right)\right]f\left(\psi_{B_{l_i}}\left(x,c'\right)\right)\rho_{B_{l_i}}\left(\psi_{B_{l_i}}\left(x,c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_{l_i}}h\left(\psi_{B_{l_i}}\left(x,c'\right)\right)\right)\right|}\frac{1}{t_n}d\mathcal{L}_1c_i'\right].$$

where $\rho_{B_{l_i}}\left(\psi_{B_{l_i}}\left(x,c'\right)\right)$ is the partition of unity, and x_{l_i} denotes that x_{l_i} is picked for locally implicit function for change of variables by the implicit function theorem. We also note that in the above $y=\tau_{\neg i}\left(c'\right)$, and $c_i'=\tau_i\left(c'\right)$ and we use the fact that

$$1\left(c'+t_nH_n\left(x\right)>c\right)=\prod_{i=1}^{k}1\left(c_i'+t_nH_{n,i}\left(x\right)>c_i\right)\text{ and }1\left(c'>c\right)=\prod_{i=1}^{k}1\left(c_i'>c_i\right).$$

Then by DCT and the area formula again, we obtain the limit of the first order difference as

$$\sum_i \int \left[\int_{h^{-1}(c'(y,c_i,i))} \frac{H_i(x) f(x)}{Jh(x)} d\mathcal{H}_{n-k} x \right] d\mathcal{L}_{k-1} y.$$

The convergence of the first order term also implies that the second and higher order terms of the difference should vanish eventually.

part 3 By **part** 1 and **part** 2, we only need to consider the $k = 1$ case since the $k > 1$ case is similar. Let $\{c_n\}_{n=1}^{\infty}$ be a sequence such that $\lim_{n\to\infty} c_n = c$. Consider

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{t_n} \left[F(h + t_n H_n, c_n) - F(h, c_n) \right] \\
&\quad = \lim_{n\to\infty} \frac{1}{t_n} \int \left[1(h(x) + t_n H_n(x) > c_n) - 1(h(x) > c_n) \right] f(x) \, d\mathcal{L}_n x.
\end{aligned}$$

The proof in **part** 1 shows that we need to calculate, for $M = \sup_{x\in K_f} |H(x)|$,

$$\frac{1}{t_n} \int_{c-t_n M-|c-c_n|}^{c+t_n M+|c-c_n|} \left[1\left(c' + t_n H\left(\psi_{B_j}(\cdot, c')\right) > c_n\right) - 1\left(c' > c_n\right) \right] d\mathcal{L}_1 c'.$$

Note that

$$\frac{1}{t_n} \int \left[1\left(c' + t_n A + c - c_n > c\right) - 1\left(c' + c - c_n > c\right) \right] d\mathcal{L}_1 c' = A,$$

we can obtain

$$\lim_{n\to\infty} \frac{1}{t_n} \int_{c-t_n M-|c-c_n|}^{c+t_n M+|c-c_n|} \left[1\left(c' + t_n H\left(\psi_{B_j}(\cdot, c')\right) > c_n\right) - 1\left(c' > c_n\right) \right] d\mathcal{L}_1 c' = H\left(\psi_{B_j}(\cdot, c)\right)$$

by simple upper and lower bound of $H\left(\psi_{B_j}(\cdot, c')\right)$ on $[c - t_n M - |c - c_n|, c + t_n M + |c - c_n|]$. So we obtain, that the limit in the calculation of the Hadamard derivative actually converges continuously (see for example Definition T.1.2). By Lemma T.1.5, we get that the Hadamard derivative can be taken in a uniform sense. In other words, let $\mathcal{D} \subset h(E)$, then we have

$$\lim_{n\to\infty} \sup_{c\in\mathcal{D}} \left| \frac{1}{t_n} \int \left[1(h(x) + t_n H_n(x) > c) - 1(h(x) > c) \right] f(x) \, d\mathcal{L}_n x - \int_{h^{-1}(c)} \frac{H(x) f(x)}{Jh(x)} d\mathcal{H}_{n-1} x \right| = 0.$$

This finishes the proof of Theorem 5.1.1.

part 4 We only need to point out changes under lower regularity of Theorem 5.1.3.

step 1 Let $x \in h^{-1}(\mathcal{D})$. By Assumption 5.1.4, $J_c h(x)$ is of full rank, so we can change the C^1 implicit function theorem to the Lipschitz version, see for example Theorem T.1.19 (Lipchitz case) in Technical addendum. For simplicity of notations, consider the case when $k = 1$. Without loss of generality, we may also assume the first coordinate is picked to be expressed by the implicit

function, i.e., $x_1' = \xi(x_2', \ldots, x_n', c')$. Compared to **step** 2 of **part** 1, we omit the subscript like j and B_j, since the reader should have been clear that the discussion here is purely local.

The implicit function theorem says that locally we have $h(\xi(x_2', \ldots, x_n', c'), x_2', \ldots, x_n') = 0$. Fix c', $\xi(\cdot, c)$ locally characterized the level set $\{x : h(x) = c'\}$, i.e., $x' \in \{x' : h(x') = c'\} \Leftrightarrow x_1' = \xi(x_2', \ldots, x_n', c')$. By Assumption 5.1.5, $\mathcal{H}_{n-1}\{x' : h(x') = c', h(\cdot) \text{ is not differentiable at } x'\} = 0$, thus for L_{n-1} almost every $(x_2', \ldots, x_n')$ in a neighborhood, we can perform differential calculus to $\xi(\cdot, c')$ and $\psi(\cdot, c')$ by the chain rule and Rademacher Theorem (see for example Theorem T.1.11) to see (23) holds *a.e.*

step 2 Now, following the main idea from **part** 1 **step** 3, we need to calculate

$$\lim_{t_n \downarrow 0, c_n \to c} \frac{1}{t_n} \int \frac{\left[1(c' + t_n H_n(\psi_B(x_{2,\ldots,n}, c')) > c_n) - 1(c' > c_n)\right] f(\psi_B(x_{2,\ldots,n}, c'))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c'))\right)\right|} d\mathcal{L}_1 c',$$

where B is the neighborhood where we apply the Lipschitz implicit function theorem. Consider $(x_2, \ldots, x_n)$ where $\frac{f(\psi_B(\cdot,c))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(\cdot,c))\right)\right|}$ is continuous, then we have

$$\int \frac{\left[1(c' + t_n A > c_n) - 1(c' > c_n)\right] f(\psi_B(x_{2,\ldots,n}, c'))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c'))\right)\right|} d\mathcal{L}_1 c' = t_n A \left(\frac{f(\psi_B(x_{2,\ldots,n}, c))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c))\right)\right|} + o(1) \right).$$

For $m \in \{1, 2, \ldots\}$, define $D(m)$ as the set of $(x_2, \ldots, x_n) \in B_{x_2,\ldots,x_n}$ such that

$$\begin{aligned} &\lim_{t_n \downarrow 0, c_n \to c} \frac{1}{t_n} \int \frac{\left[1\left(c' + t_n\left(H(\psi_B(x_{2,\ldots,n}, c)) + \frac{1}{m}\right) > c_n\right) - 1(c' > c_n)\right] f(\psi_B(x_{2,\ldots,n}, c'))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c'))\right)\right|} d\mathcal{L}_1 c' \\ &= \left(H(\psi_B(x_{2,\ldots,n}, c)) + \frac{1}{m}\right) \frac{f(\psi_B(x_{2,\ldots,n}, c))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c))\right)\right|}. \end{aligned}$$

By Assumption 5.1.5, $\mathcal{L}_{n-1}\{B_{x_2,\ldots,x_n} \backslash D(m)\} = 0$ and further $\mathcal{L}_{n-1}\{\bigcup_m B_{x_2,\ldots,x_n} \backslash D(m)\} = 0$, which imply

$$\begin{aligned} &\lim_{t_n \downarrow 0, c_n \to c} \frac{1}{t_n} \int \frac{\left[1(c' + t_n H_n(\psi_B(x_{2,\ldots,n}, c')) > c_n) - 1(c' > c_n)\right] f(\psi_B(x_{2,\ldots,n}, c'))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c'))\right)\right|} d\mathcal{L}_1 c' \\ &= H(\psi_B(x_{2,\ldots,n}, c)) \frac{f(\psi_B(x_{2,\ldots,n}, c))}{\left|\det\left(\frac{\partial}{\partial x_1} h(\psi_B(x_{2,\ldots,n}, c))\right)\right|} \end{aligned}$$

for $\mathcal{L}_{n-1}$ almost every $(x_2, \ldots, x_n) \in B_{x_2,\ldots,x_n}$. By Binet-Cauchy formula (see for example chapter 1 of Gantmakher (1977)), Assumption 5.1.4 implies that locally $\det\left(\frac{\partial}{\partial x_1} h(\cdot)\right)$ is bounded away

from 0. Since f is bounded by Assumption 5.1.3, through DCT, we have

$$\lim_{t_n \downarrow 0, c_n \to c} \frac{1}{t_n} \int \left[1\left(h\left(x\right) + t_n H_n\left(x\right) > c\right) - 1\left(h\left(x\right) > c\right)\right] f\left(x\right) d\mathcal{L}_n x = \int_{h^{-1}(c)} \frac{H\left(x\right) f\left(x\right)}{Jh\left(x\right)} d\mathcal{H}_{n-1} x.$$

This finishes the proof of Theorem 5.1.3. ■

5.2 On Margin Assumption

Definition 5.2.1. For a random variable X on $\mathbb{R}$, we say it satisfies the margin assumption of order α at $c \in \mathcal{R}$, if there exists a constant C_0 such that

$$\mathbb{P}\left\{|X - c| < t\right\} \leqslant C_0 t^{\alpha} \tag{24}$$

Definition 5.2.2. (Margin controlled random variable / distribution function) Let X be a random variable supported on $[a, b]$, where $a < b$, $a, b \in \mathbb{R}$. If for all $c \in [a, b]$,

$$\mathbb{P}\left\{|X - c| < t\right\} \leqslant C\left(c\right) t^{\alpha},$$

where $-\infty < C\left(c\right) < +\infty$ may depend on c, we call X a margin controlled random variable of order α. Similarly, if F is the distribution function of such a random variable, we call it a margin controlled distribution function.

A definition of a margin controlled random variable is to impose local Hölder-continuity on its distribution function with a universal exponent α.

Lemma 5.2.1. *If X is a margin controlled random variable on $[a, b]$ of order $\alpha = 1$, where $a < b$, $a, b \in \mathbb{R}$, then X is absolutely continuous with respect to the Lebesgue measure. Consequently, X admits a density function.*

Lemma 5.2.2. *Let $f \in L^1\left(\mathbb{R}^n, \mathcal{L}_n\right)$ be a nonnegative function and $h : \Omega \to \mathbb{R}$ be a function in C^n (or Sobolev space $W^{n,p}$, $p > n$), where $\Omega \subset \mathbb{R}^n$ is an open subset. Let X be a random variable with density function f. Then $h\left(X\right)$ also admit a density, i.e. the push-forward probability $h_{\#}\left(f\mathcal{L}_n\right) \ll \mathcal{L}_1$, if and only if $Jh > 0$ a.e. on the support of f. In this case the density of $h\left(X\right)$ is*

$$f_h\left(y\right) = \int_{h^{-1}(y)} \frac{f\left(x\right)}{Jh\left(x\right)} d\mathcal{H}_{n-1} x.$$

5.3 A Formal Calculus Approach

Essentially, we use the $1, \ldots, J$ indexs to represent a base of the ROC volume, and the 0 index to represent the corresponding height. Given the definitions of the ROC surface, the multivariate NP Lemma in section 4.1 can also be understood through the Lagrangian of a linear program. In particular, the population Lagrangian is defined by, where $\lambda = (\lambda_j, j = 0, \ldots, J)$ such that $\lambda_0 = 1$,

$$\mathcal{L}(\beta, \phi(\cdot), \lambda) = Q(y_0(\phi_0(x) - \beta_0)) + \sum_{j=1}^{J} \lambda_j (Q(y_j(\phi_j(x) - \beta_j))).$$

Given the Lagrange multipliers λ, the optimal critical functions $\phi(\cdot)$ are given by

$$\phi_j(x) = 1\left(\lambda_j p_j(x) \geqslant \max_{k \neq j} \lambda_k p_k(x)\right), \ j = 0, \ldots, J.$$

The Lagrange multipliers are solved by binding the constraints:

$$C_j(\lambda) = Q\left(y_j\left(1\left(\lambda_j p_j(x) \geqslant \max_{k \neq j, k=0,\ldots,J} \lambda_k p_k(x)\right)\right)\right) - p_j \beta_j, \quad \forall j = 1, \ldots, J. \tag{25}$$

The Jacobian of this system $\frac{\partial C_j(\lambda)}{\partial \lambda}, j = 1, \ldots, J$ can be shown to be diagonally dominant, and hence non-singular, and that the solutions $\lambda(\beta_1, \ldots, \beta_J)$ is increasing in $(\beta_1, \ldots, \beta_J)$. First, we define a manifold similar to the binary case,

$$\nabla_j(x) = \lambda_j p_j(x) - \max_{k \neq j} \lambda_k p_k(x), \quad M_j(\lambda_{>0}) = \{x \in \mathcal{X} : \nabla_j(x) = 0\}$$

where $\lambda_{>0} = (\lambda_1, \ldots, \lambda_J)$. To compute the Jacobian, we calculate the following derivatives.

$$\frac{\partial C_j(\lambda)}{\partial \lambda_j} = \int_{M_j(\lambda_{>0})} (p_j(x))^2 \frac{\mu'(x)}{||\partial \nabla_j(x)||} dH_{n-1}(x)$$

Furthermore, we also define, for $j \neq k$, $(j, k) \in \{1, \ldots, J\}$

$$M_{j,k}(\lambda_{>0}) = \{x \in \mathcal{X} : \lambda_j p_j(x) = \lambda_k p_k(x) > \max_{l \neq j,k} \lambda_l p_l(x)\}$$

Then we can calculate that

$$\frac{\partial C_j(\lambda)}{\partial \lambda_k} = -\int_{M_{j,k}(\lambda_{>0})} (p_j(x) p_k(x)) \frac{\mu'(x)}{||\partial \nabla_j(x)||} dH_{n-1}(x)$$

As such, the diagonal elements of the Jacobian matrix are all positive, while the off-diagonal elements of the Jacobian matrix are all negative. This satisfies the first requirement of a diagonally dominant matrix (see e.g. **?**). Furthermore, for each $j = 1, \ldots, J$, let $M_{j,0}(\lambda_{>0}) = \{x \in \mathcal{X} :$

$\lambda_j p_j(x) = \lambda_0 p_0(x) > \max_{l\neq j, l\neq 0} \lambda_l p_l(x)\}$, and define

$$\sum_{k=1,\ldots,J} \lambda_k \frac{\partial C_j(\lambda)}{\partial \lambda_k} = \lambda_j \frac{\partial C_j(\lambda)}{\partial \lambda_j} + \sum_{k=1,\ldots,J, k\neq j} \lambda_k \frac{\partial C_j(\lambda)}{\partial \lambda_k} = \int_{M_{j,0}(\lambda_{>0})} \lambda_j (p_j(x))^2 \frac{\mu'(x)}{||\partial \nabla_j(x)||} dH_{n-1}(x) > 0.$$

Therefore, the Jacobian matrix is diagonally dominant. This implies that its positive definite and its inverse has all strictly positive elements, so that $\lambda(\beta_{>0})$ is uniquely defined for each $\beta_{>0}$ in the feasible set, and is an increasing function of $\beta_{>0}$. Indeed, $C_j(\lambda), j = 1,\ldots,J$ is a gradient field corresponding to the potential function of the concentrated Lagrangian

$$E \max_{j=0,\ldots,J} \left(\lambda_j p^j(x)\right) \quad \text{subject to} \quad \lambda_0 = 1,$$

that is to be minimized with respect to λ with an everywhere positive definite Hessian. We emphasize that the Lagrange multiplier approach delivers the same conclusion as the general multivariate Neyman-Pearson Lemma 4.1.4.

In order to conduct statistical inference on the population ROC surface, typically we employ a model of the propensity score functions $p_j(x)$, denoted as $\Delta_j(x)$, and replace the propensity score functions $p_j(x)$ by sample estimates $\hat{\Delta}_j(x)$ that satisfies $\sum_{j=0}^{J} \hat{\Delta}_j(x) = 1$. We also denote the population limit of $\hat{\Delta}_j(x)$ as $\Delta_j^\star(x)$. A model is correctly specified when $\Delta_j^\star(x) = p_j(x), \forall j = 0,\ldots,J$, and is otherwise misspecified.

Let $c_0 \equiv 1$ and $c_j \geqslant 0, \forall j = 1,\ldots,J$, $c_{>0} = (c_j, j = 1,\ldots,J)$, and let the model analog of the population system of equations in (25), for $j = 1,\ldots,J$,

$$N_{j,\Delta}(c) = Q\left(y_j\left(1\left(c_j\Delta_j(x) \geqslant \max_{k\neq j} c_k\Delta_k(x)\right)\right) - \beta_j\right).$$

This system can be inverted to obtained $c_{>0}(\beta_{>0})$ as an increasing of $\beta_{>0}$, by the same diagonally dominant Jacobian argument above. To see this, let

$$\nabla_j(x) = c_j\Delta_j(x) - \max_{k\neq j, k=0,\ldots,J} c_k\Delta_k(x), \quad M_j(c_{>0}) = \{x \in \mathcal{X} : \nabla_j(x) = 0\}$$

Then the elements of the Jacobian are

$$\frac{\partial N_{j,\Delta}(c)}{\partial c_j} = \int_{M_j(c_{>0})} p_j(x)\Delta_j(x) \frac{\mu'(x)}{||\partial \nabla_j(x)||} dH_{n-1}(x),$$

and for $M_{j,k}(c_{>0}) = \{x \in \mathcal{X} : \lambda_j p_j(x) = \lambda_k p_k(x) > \max_{l\neq j,k} \lambda_l p_l(x)\}$,

$$\frac{\partial N_{j,\Delta}(c)}{\partial c_k} = -\int_{M_{j,k}(c_{>0})} (p_j(x)\Delta_k(x)) \frac{\mu'(x)}{||\partial \nabla_j(x)||} dH_{n-1}(x).$$

Again, the matrix $\left[\frac{\partial N_{j,\Delta}(c)}{\partial c_k}\right]_{j,k=1,...,J}$ is diagonally dominant, with for each j,

$$\sum_{k=1,...,J} c_k \frac{\partial N_{j,\Delta}(c)}{\partial \lambda_k} = \int_{M_{j,0}(c_{>0})} c_j p_j(x) \Delta_j(x) \frac{\mu'(x)}{||\partial \nabla_j(x)||} dH_{n-1}(x) > 0.$$

where $M_{j,0}(c_{>0}) = \{x \in \mathcal{X} : c_j p_j(x) = c_0 p_0(x) > \max_{l \neq j, l \neq 0} c_l p_l(x)\}$.

In a general model that admits possible misspecification, it is no longer necessarily the case that $\frac{\partial N_{j,\Delta}(c)}{\partial c_k} = \frac{\partial N_{k,\Delta}(c)}{\partial c_j}$ so that $N_{j,\Delta}(c), j = 1, \ldots, J$ is not a vector field. However, by diagonal dominance, the Jacobian is still strictly uniformly positive definite and $c_{>0}(\beta_{>0})$ is a strictly increasing function of $\beta_{>0}$. Let $\Phi_{>0} \equiv \{\phi_{>0}(x) = (\phi_1(x), \ldots, \phi_J(x))\}$.

Corollary 5.3.1. *In a purely formal point of view,* $\beta_{0,\Delta,Q,c_{>0,\Delta,Q,\beta_{>0}}} : C(E)^{J+1} \times \ell^\infty(\mathcal{F}_y) \mapsto \mathbb{R}$ *is directional differentiable with derivative in the direction of* $(G(\cdot), H(\cdot))$ *given by*

$$\beta'_{0,\Delta^\star,Q,c_{>0,\Delta^\star,Q,\beta_{>0}}}(G,H) = \frac{1}{p_0}\Bigg[H\left[y_0 \mathbb{1}\left(c_0\Delta_0^\star(x) > \max_{k\neq 0} c_k \Delta_k^\star(x)\right) - \beta_0(\beta_{>0})\right]$$
$$+ \int_{M_{0,\Delta^*}(c_{>0}(\beta_{>0}))} \frac{c_0 p_0(x)\mu'(x) G_0(x)}{\|\partial \nabla^*_{0,c_{>0}(\beta_{>0})}(x)\|} dH_{n-1}x - \sum_{j=1}^J \int_{M_{0,j,\Delta^*}(c_{>0}(\beta_{>0}))} \frac{c_j p_0(x)\mu'(x) G_j(x)}{\|\partial \nabla^*_{0,c_{>0}(\beta_{>0})}(x)\|} dH_{n-1}x\Bigg]$$
$$- \frac{1}{p_0}\left(\frac{\partial N_{0,\Delta^\star}(c_{>0}(\beta_{>0}))}{\partial c_{>0}}\right)^T \left[\frac{\partial N_{>0,\Delta^\star}(c_{>0}(\beta_{>0}))}{\partial c_{>0}}\right]^{-1} \left[H\left[y_j\left(1\left(c_j\Delta_j^\star(x) \geqslant \max_{k\neq j} c_k\Delta_k^\star(x)\right) - (\beta_j)\right)\right]_{j>0}\right.$$
$$\left. + \left(\int_{M_{j,\Delta^*}(c_{>0}(\beta_{>0}))} \frac{c_j p_j(x)\mu'(x) G_j(x)}{\|\partial \nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|} dH_{n-1}x - \sum_{k\neq j}\int_{M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))} \frac{c_k p_j(x)\mu'(x) G_k(x)}{\|\partial \nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|} dH_{n-1}x\right)_{j>0}\right],$$

where we define

$$\nabla_{j,c_{>0}}(x) = c_j\Delta_j(x) - \max_{k\neq j} c_k\Delta_k(x), \quad M_{j,\Delta}(c_{>0}) = \{x \in \mathcal{X} : \nabla_{j,c_{>0}}(x) = 0\}.$$
$$M_{j,k,\Delta}(c_{>0}) = \{x \in \mathcal{X} : c_j\Delta_j(x) = c_k\Delta_k(x) > \max_{l\neq j, l\neq k} c_l\Delta_l(x)\}.$$

and

$$\frac{\partial N_{j,\Delta}(c_{>0})}{\partial c_{>0}} = \left[\frac{\partial N_{j,\Delta}(c_{>0})}{\partial c_k}\right]_{k=1,...,J} \quad \textit{and} \quad \frac{\partial N_{>0,\Delta}(c_{>0})}{\partial c_{>0}} = \left[\frac{\partial N_{j,\Delta}(c_{>0})}{\partial c_k}\right]_{j,k=1,...,J}.$$

with

$$\frac{\partial N_{j,\Delta}(c_{>0})}{\partial c_j} = \int_{M_{j,\Delta}(c_{>0})} \frac{p_j(x)\Delta_j(x)\mu'(x)}{||\partial \nabla_{j,c_{>0}}(x)||} dH_{n-1}(x), \frac{\partial N_{j,\Delta}(c_{>0})}{\partial c_k} = -\int_{M_{j,k,\Delta}(c_{>0})} \frac{p_j(x)\Delta_k(x)\mu'(x)}{||\partial \nabla_{j,c>0}(x)||} dH_{n-1}(x).$$

Similar to the binary case, we consider the correctly specified case where $\Delta_j^\star(\cdot) \equiv \Delta_j(\cdot, \theta^*) =$

$p_j(\cdot), j=0,\ldots,J$. Under correct specification, $p_0(x)\,\Delta_j^*(x) \equiv \Delta_0^*(x)\,p_j(x)$, $\forall j \geqslant 0$.

$$-\left(\frac{\partial N_{0,\Delta^*}(c_{>0})}{\partial c_{>0}}\right)^T\left[\frac{\partial N_{>0,\Delta^*}(c_{>0})}{\partial c_{>0}}\right]^{-1} = c_{>0} \equiv \lambda_{>0} = (\lambda_j, j=1,\ldots,J).$$

We will further show that in Corollary 5.3.1, $\beta'_{0,\Delta^\star,Q,c_{>0,\Delta^\star,Q,\beta_{>0}}}(G,0)=0$.This follows from (note $c_0=1$)

$$\begin{aligned}
&\int_{M_{0,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_0^2 p_0(x)\,\mu'(x)\,G_0(x)}{\|\partial\nabla^*_{0,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x - \sum_{j=1}^J\int_{M_{0,j,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_0 c_j p_0(x)\,\mu'(x)\,G_j(x)}{\|\partial\nabla^*_{0,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x\\
&-\left(\frac{\partial N_{0,\Delta^\star}(c_{>0}(\beta_{>0}))}{\partial c_{>0}}\right)^T\left[\frac{\partial N_{>0,\Delta^\star}(c_{>0}(\beta_{>0}))}{\partial c_{>0}}\right]^{-1}\\
&+\left(\int_{M_{j,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_j p_j(x)\,\mu'(x)\,G_j(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x - \sum_{k\neq j}\int_{M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_k p_j(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x\right)_{j>0}\\
&=\sum_{j=0}^J\int_{M_{j,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_j^2 p_j(x)\,\mu'(x)\,G_j(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x - \sum_{j=0}^J\sum_{k\neq j}\int_{M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_j c_k p_j(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x\\
&=\sum_{k=0}^J\int_{M_{k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_k^2 p_k(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{k,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x - \sum_{k=0}^J\sum_{j\neq k}\int_{M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_j c_k p_j(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x\\
&=\sum_{k=0}^J\left(\int_{M_{k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_k^2 p_k(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{k,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x - \sum_{j\neq k}\int_{M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_j c_k p_j(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x\right)
\end{aligned}$$

The summands for each $k=0,\ldots,J$ is zero because $c_j p_j(x) = c_k p_k(x)$ on $M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))$,

$$\sum_{j\neq k}\int_{M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_j c_k p_j(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x = \int_{\cup_{j\neq k}M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))}\frac{c_k^2 p_k(x)\,\mu'(x)\,G_k(x)}{\|\partial\nabla^*_{k,c_{>0}(\beta_{>0})}(x)\|}dH_{n-1}x$$

In the above, we used $M_{j,k,\Delta^*}(c_{>0}(\beta_{>0})) = \cup_{j\neq k}M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))$, and on $M_{j,k,\Delta^*}(c_{>0}(\beta_{>0}))$,

$$\partial\nabla^*_{k,c_{>0}(\beta_{>0})}(x) = \partial\nabla^*_{j,c_{>0}(\beta_{>0})}(x).$$

Consequently, $\beta'_{0,\Delta^\star,Q,c_{>0,\Delta^\star,Q,\beta_{>0}}}(G,H) = \beta'_{0,\Delta^\star,Q,c_{>0,\Delta^\star,Q,\beta_{>0}}}(0,H)$.

Theorem 5.3.2. *In a purely formal point of view,* $\sqrt{n}\left(\hat{\beta}_0(\beta_{>0}) - \beta_0(\beta_{>0})\right) \rightsquigarrow \mathcal{Z}_\infty(\beta_{>0})$, *where*

$$\begin{aligned}
\mathcal{Z}_\infty(\beta_{>0}) = \beta'_{0,\Delta^\star,Q,c_{>0,\Delta^\star,Q,\beta_{>0}}}(0,\mathbb{Q}) &= \frac{1}{p_0}\mathbb{Q}\left[y_0\mathbb{1}\left(c_0\Delta_0^*(x) > \max_{k\neq 0}c_k\Delta_k^*(x)\right) - \beta_0(\beta_{>0})\right]\\
&\quad - \frac{1}{p_0}c_{>0}(\beta_{>0})^T\mathbb{Q}\left[y_j\left(1\left(c_j\Delta_j^*(x) \geqslant \max_{k\neq j}c_k\Delta_k^*(x)\right) - (\beta_j)\right)\right]_{j>0}.
\end{aligned}$$

6 Triple Machine Learning Inference

6.1 Concave (Convex) Convergence

Theorem 6.1.1. *Concavity Lemma, Theorem 10.8 in Rockafellar (1970)*
Let C be an open convex subset of $\mathbb{R}^m$ and $\{F_i\}_{i=1}^{\infty}$ be a sequence of real-valued concave functions on C such that $\forall x \in C$, $F_n(x) \to F_0(x)$ as $i \to \infty$, where F_0 is a real-valued function on C. Then F_0 is also concave and for all compact $K \subset C$,

$$\lim_{i\to\infty} \sup_{x\in K} |F_i(x) - F_0(x)| = 0.$$

We should make several remarks here. Theorem 6.1.1 is written with a style that comply with those in Theorem II.1 in Andersen and Gill (1982) and CONVEXITY LEMMA in Pollard (1991). The original Theorem 10.8 in Rockafellar (1970) is actually a bit more general than the version stated here. The domain C can be generalized to any relative open subset of $\mathbb{R}^m$ in the affine hull sense (see section 6 of Rockafellar (1970)). The pointwise convergence everywhere assumption can also be weaken to pointwise convergence on a dense subset of C. Here, we would like to provide a more abstract version of Theorem 6.1.1 to facilitate possible future researches.

Theorem 6.1.2. *Let C be an open convex subset of a normed space $\mathcal{X}$ and $\{F_i\}_{i=1}^{\infty}$ be a sequence of locally equi-Lipschitz functions on C such that $\forall x \in C$, $F_i(x) \to F_0(x)$ as $i \to \infty$, where F_0 is a real-valued function on C. Then F_0 is continuous and for all compact $K \subset C$,*

$$\lim_{i\in\infty} \sup_{x\in K} |F_i(x) - F_0(x)| = 0.$$

PROOF OF THEOREM 6.1.2.
step 1 $\{F_i\}_{i=1}^{\infty}$ is equi-Lipschitz on compact sets. In order to see this, let K be a compact subset of C and $\{B_j\}_{j=1}^{r}$ be a finite over cover of C such that $\{F_i\}_{i=1}^{\infty}$ are equi-Lipschitz with constant L_j on B_j. We can always select such a cover since $\{F_i\}_{i=1}^{\infty}$ is locally equi-Lipschitz. The metric $d(x_1, x_2) := \|x_1 - x_2\|_{\mathcal{X}}$ induces by the norm $\|\cdot\|_{\mathcal{X}}$ on $\mathcal{X}$ is a positive continuous function on $K \times K \backslash \bigcup_{j=1}^{r} B_j \times B_j$, and thus attains a minimum $\ell > 0$. By the equi-Lipschitzness and pointwise convergence of $\{F_i\}_{i=1}^{\infty}$, it is not hard to see that $\{F_i\}_{i=1}^{\infty}$ is uniformly bounded on K. Therefore, we can choose a constant M large enough such that $\forall i \in \{1, \ldots, n, \ldots\}$, $\forall x_1, x_2$, $\|F_i(x_1) - F_i(x_2)\| \leqslant M\ell$. Now, observe that $\max\{M, L_1, \ldots, L_r\}$ is the Lipschitz constant for all F_i on K.
step 2 Uniform convergence of $\{F_i\}_{i=1}^{\infty}$ on K. By Arzelà-Ascoli theorem (see, for example Theorem T.1.3), for any subsequence of $\{F_i\}_{i=1}^{\infty}$, it has a further subsequence that uniformly converges. The pointwise convergence of $\{F_i\}_{i=1}^{\infty}$ to F_0 implies that the limit of any subsequence must still be F_0. By the fact that in any metric space, a sequence converges if and only if any subsequence of it has

a further subsequence which converges to the same limit, $\{F_i\}_{i=1}^{\infty}$ also uniformly converges to F_0 on K. ■

In the proof of Theorem 6.1.2, **step** 1 follows the hint of Exercise 4.2.10 in Kumaresan (2005). We will use the probability version of "converges iff existence of subsubsequence converges to the same limit of any subsequence" argument in the later Corollary 6.1.4.

Corollary 6.1.3. *Let C be an open convex subset of a Banach space $\mathcal{X}$ and $\{F_i\}_{i=1}^{\infty}$ be a sequence of real-valued upper semicontinuous concave functions on C such that $\forall x \in C$, $F_i(x) \to F_0(x)$ as $i \to \infty$, where F_0 is a real-valued function on C. Then F_0 is a continuous concave function and for all compact $K \subset C$,*

$$\lim_{i\in\infty} \sup_{x\in K} |F_i(x) - F_0(x)| = 0.$$

Proof of Corollary 6.1.3. Concavity of F_0 is obvious. We only need to show that $\{F_i\}_{i=1}^{\infty}$ is locally equi-Lipschitz.

step 1 $\{F_i\}_{i=1}^{\infty}$ is locally bounded . For all $x \in C$, we have $\{F_i(x)\}_{i=1}^{\infty}$ is bounded, which implies that $G(x) := \inf_{F\in\{F_i\}_{i=1}^{\infty}} F(x)$ is real-valued. By concavity and upper semicontinuity of F_i, the hypograph $\text{hypo}(F_i) := \{(x,t) \in C \times \mathbb{R} : F_i(x) \geqslant t\}$ is closed and convex for all $i \in \{1,\ldots,n,\ldots\}$. Since $\text{hypo}(G) = \bigcap_{i=1}^{\infty} \text{hypo}(F_i)$ is also closed and convex, we obtain that G is still upper semicontinuous and concave. A real-valued upper semicontinuous function on an open subset of a Banach space must also be continuous, see for example Theorem 4.2.9 in Niculescu and Persson (2018) (this is actually a direct consequence of the Baire category theorem). By continuity of G, G and thus $\{F_i\}_{i=1}^{\infty}$ are locally bounded.

step 2 $\{F_i\}_{i=1}^{\infty}$ is locally equi-Lipschitz. See the bounding strategy in the proof of Theorem 4.2.1 **part** 3 **step** 3. ■

Corollary 6.1.4. *Let C be an open convex subset of a separable Banach space $\mathcal{X}$ and $\{F_i\}_{i=1}^{\infty}$ be a sequence of random real-valued upper semicontinuous concave functions on C such that $\forall x \in C$, $F_i(x) \to F_0(x)$ as $i \to \infty$, where F_0 is a real-valued function on C. Then F_0 is a continuous concave function and for all compact $K \subset C$,*

$$\sup_{x\in K} |F_i(x) - F_0(x)| \xrightarrow{\mathbb{P}} 0 \text{ as } i \to \infty.$$

Proof of Corollary 6.1.4. Concavity of F_0 is still obvious by a passage to limit in probability.

step 1 Let C_0 be a countable dense set of C, we claim that if $\forall x \in C_0$, $F_i(x) \to F_0(x)$ almost surely, then we have the uniform convergence on each compact set almost surely. Denote the random functions F_i, $i \in \{1,\ldots,n,\ldots\}$ as $F_i(\omega, x)$, where $\omega \in \Omega$ and Ω the sample space. By

the proof of Corollary 6.1.3 **step** 1, F_i takes value in the space of continuous functions for all $i \in \{1,\ldots,n,\ldots\}$. Let $\Omega_0 := \{\omega \in \Omega : \forall x \in C_0, \lim_{i\to\infty} F_i(\omega,x) = F_0(x)\}$, by assumptions, we have $\mathbb{P}(\Omega_0) = 1$. For all $y \in C \backslash C_0$ and for all $\epsilon > 0$,

$$|F_i(\omega,y) - F_0(y)| \leqslant |F_i(\omega,y) - F_i(\omega,x_\epsilon)| + |F_i(\omega,x_\epsilon) - F_0(x_\epsilon)| + |F_0(x_\epsilon) - F_0(y)|,$$

where $x_\epsilon \in C_0$ and $|x_\epsilon - y| < \epsilon$. By the proof Corollary 6.1.3 **step** 2, $\{F_i(\omega,\cdot)\}_{i=1}^{\infty}$ is locally equi-Lipschitz, so the first term in the right hand side above is of $O(\epsilon)$. If $\omega \in \Omega_0$, then for any $i \in \{1,\ldots,n,\ldots\}$ large enough, we have that $|F_i(\omega,x_\epsilon) - F_0(x_\epsilon)| < \epsilon$. The third term is also of $O(\epsilon)$ since F_0 is also locally Lipschitz. By the arbitrariness of ϵ, we obtain that for all $\omega \in \Omega_0$, $\lim_{i\in\infty} F_i(\omega,x) = F_0(x)$, $\forall x \in C$.

step 2 We repeat the means of the subsequencing argument in Andersen and Gill (1982). Let $\{x_1,\ldots,x_n,\ldots\}$ be a countable dense sequence of C. By the fact that convergence in probability (in metric space) implies existence of convergent subsequence in almost surely sense, we can pick up a monotonic increasing sequence $\rho \in {\mathbb{Z}_+}^{\mathbb{Z}_+}$ such that for all $j \in \{1,\ldots,n,\ldots\}$, we have $F_{\rho(i)}(x_j) \to F_0(x_j)$ almost surely. By **step** 1, we obtain uniform convergence on each compact set almost surely along the subsequence ρ. More generally, from any subsequence of $(1,\ldots,n,\ldots)$, we can extract a further subsequence along which we have the uniform convergence on each compact set property. Now, we conclude by the fact that if each subsequence of a sequence has further subsequence converges to a fixed limit almost surely, then the sequence also converges to the limit in probability. ■

6.2 Bounds for Indicator Functions

In this subsection we first provide a bounding method inspired by Lemma 3.1 in Audibert and Tsybakov (2007) under the simplest case of binary classification. For critical function ϕ, denote the misclassification error as $R(\phi) = \mathbb{P}(Y \neq \phi(X))$. Consider the plug-in rule $\hat{\phi}_n(x) = 1\left(\hat{p}_n(x) > \frac{1}{2}\right)$ and the optimal rule $\phi^*(x) = 1\left(p(x) > \frac{1}{2}\right)$.

Theorem 6.2.1. *Let $\hat{p}_n$ be an estimator of $p(x)$ and $\{a_n\}_{n\geqslant 1}$ a positive sequence.*

1. *If for constant C_1, we have $\mathbb{E}_{\mathbb{Q}_n}\|\hat{p}_n - p\|_\infty \leqslant C_1 a_n$, then $\mathbb{E}_{\mathbb{Q}_n} R\left(\hat{\phi}\right) - R(\phi^*) \leqslant C a_n^{1+\alpha}$ with constant C depending only on α, C_0 and C_1.*

2. *If for constants $C_1, C_2 > 0$, and for all $\delta > 0$, we have*

$$\mathbb{E}_{\mathbb{Q}_n} 1\left(|\hat{p}_n(X) - p(X)| > \delta\right) \leqslant C_1 \exp(-C_2 a_n^{-1}\delta^2)$$

almost surely with respect to μ, $X \sim \mu$, then $\mathbb{E}_{\mathbb{Q}_n} R\left(\hat{\phi}\right) - R(\phi^*) \leqslant C a_n^{\frac{1+\alpha}{2}}$ with constant C depending only on α, C_0, C_1 and C_2 (Lemma 3.1 in Audibert and Tsybakov (2007)).

3. If for constant $C_1 > 0$, we have

$$\mathbb{E}_{\mathbb{Q}_n}\mathbb{E}_X \left|\hat{p}_n(X) - p(X)\right|^p \leqslant C_1 a_n^p,$$

then $\mathbb{E}_{\mathbb{Q}_n} R\left(\hat{\phi}\right) - R(\phi^*) \leqslant C a_n^{1+\frac{\alpha}{(1+\alpha)q}}$ for q such that $\frac{1}{p} + \frac{1}{q} = 1$ and constant C depending only on α, C_0 and C_1.

Proof of Theorem 6.2.1.

step 1 *1.* First note that $\mathbb{E}R(\phi(X)) = \mathbb{E}[Y(1-\phi(X)) + (1-Y)\phi(X)]$, which implies

$$\mathbb{E}R\left(\hat{\phi}_n(X)\right) - R\left(\phi^*(X)\right) = \mathbb{E}|2p(X) - 1|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right).$$

When $\hat{\phi}_n(x) \neq \phi^*(x)$, $\left|p(x) - \frac{1}{2}\right| \leqslant \left|\hat{p}_n(x) - p(x)\right|$. So, we have

$$\begin{aligned}\mathbb{E}|2p(X) - 1|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right) &\leqslant 2\mathbb{E}\left\|\hat{p}_n(x) - p(x)\right\|_\infty \mathbb{E}1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right) \\ &\leqslant 2a_n\mathbb{E}1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right).\end{aligned}$$

In order to bound $\mathbb{E}1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right)$, we apply the observation in example 1 of Chen et al. (2003):

$$\begin{aligned}1\left(\hat{\phi}_n(x) \neq \phi^*(x)\right) &= \left|1\left(\hat{p}_n(x) > \frac{1}{2}\right) - 1\left(p(x) > \frac{1}{2}\right)\right| \\ &\leqslant 1\left(p(x) + \delta > \frac{1}{2}\right) - 1\left(p(x) - \delta > \frac{1}{2}\right) \\ &\leqslant 1\left(\frac{1}{2} - \delta < p(x) \leqslant \frac{1}{2} + \delta\right) \\ &\leqslant 1\left(\left|p(x) - \frac{1}{2}\right| \leqslant \delta\right) \leqslant \delta^\alpha,\end{aligned}$$

where $\delta = \left\|\hat{p}_n - p\right\|_\infty$.

step 2 *2.* As in Audibert and Tsybakov (2007) Lemma 3.1, consider the sets defined as

$$\begin{aligned}A_0 &:= \left\{x : 0 < \left|p(x) - \frac{1}{2}\right| \leqslant \delta\right\}, \\ A_j &:= \left\{x : 2^{j-1}\delta < \left|p(x) - \frac{1}{2}\right| \leqslant 2^j\delta\right\} \quad \forall j \geqslant 1.\end{aligned}$$

For all $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{E}R\left(\hat{\phi}_n(X)\right) - R\left(\phi^*(X)\right) = \mathbb{E}|2p(X)-1|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right) \\
&\leqslant 2\delta\mathbb{P}_X\left(0 < \left|p(x) - \frac{1}{2}\right| \leqslant \delta\right) + \sum_{j\geqslant 1}\mathbb{E}|2p(X)-1|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right)1\left(X \in A_j\right).
\end{aligned}$$

For all $j \geqslant 1$, we have

$$\begin{aligned}
&\mathbb{E}|2p(X)-1|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right)1\left(X \in A_j\right) \\
&\leqslant 2^{j+1}\delta\mathbb{E}1\left(|\hat{p}_n(x) - p(x)| \geqslant 2^{j-1}\delta\right)1\left(X \in A_j\right) \\
&\leqslant 2^{j+1}\delta\mathbb{E}_X\mathbb{E}_{\mathbb{Q}_n}1\left(|\hat{p}_n(x) - p(x)| \geqslant 2^{j-1}\delta\right)1\left(X \in A_j\right) \\
&\leqslant C_1 2^{j+1}\delta\exp\left(-C_2 a_n^{-1}2^{2(j-1)}\delta^2\right)\mathbb{P}_X\left(0 < \left|p(X) - \frac{1}{2}\right| < 2^j\delta\right) \\
&\leqslant 2C_0C_1 2^{j(1+\alpha)}\delta^{1+\alpha}\exp\left(-C_2 a_n^{-1}2^{2(j-1)}\delta^2\right).
\end{aligned}$$

Sum up all $j \geqslant 1$ and the first term, we obtain

$$\mathbb{E}_{\mathbb{Q}_n}R\left(\hat{\phi}\right) - R\left(\phi^*\right) \leqslant 2C_0\delta^{1+\alpha} + 2C_0C_1\delta^{1+\delta}\sum_{j=1}^{\infty}2^{j(1+\alpha)}\exp\left(-C_2 a_n^{-1}2^{2(j-1)}\delta^2\right).$$

Taking $\delta = \alpha^{\frac{1}{2}}$ gives us the result.

step 3 *3.* We have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_n}R\left(\hat{\phi}_n\right) - R\left(\phi^*\right) \\
&= \mathbb{E}|2p(X)-1|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right) \\
&\leqslant 2\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|1\left(\hat{\phi}_n(X) \neq \phi^*(X)\right) \\
&= 2\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|1\left(0 < \left|p(X) - \frac{1}{2}\right| < \delta\right)1\left(|\hat{p}(X) - p(X)| > 0\right) \\
&+ 2\sum_{j\geqslant 1}\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|1\left(2^{j-1}\delta < \left|p(X) - \frac{1}{2}\right| < 2^j\delta\right)1\left(|\hat{p}(X) - p(X)| \geqslant 2^{j-1}\delta\right) \\
&\leqslant 2\left(\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|^p\right)^{\frac{1}{p}}\left(\mathbb{P}_X\left(0 < \left|p(X) - \frac{1}{2}\right| < \delta\right)\right)^{\frac{1}{q}} \\
&+ 2\sum_{j=1}^{\lfloor -\log\delta\rfloor}\left(\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|^p\right)^{\frac{1}{p}}\left(\mathbb{P}_X\left(2^{j-1}\delta < \left|p(X) - \frac{1}{2}\right| < 2^j\delta\right)\right)^{\frac{1}{r}}\left(\mathbb{P}_X\left(|\hat{p}_n(X) - p(X)| \geqslant 2^{j-1}\delta\right)\right)^{\frac{1}{s}} \\
&\leqslant 2C_1^{\frac{1}{p}}C_0^{\frac{\alpha}{q}}a_n\delta^{\frac{\alpha}{q}} \\
&+ 2\sum_{j=1}^{\lfloor -\log\delta\rfloor}\left(\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|^p\right)^{\frac{1}{p}}\left(\mathbb{P}_X\left(2^{j-1}\delta < \left|p(X) - \frac{1}{2}\right| < 2^j\delta\right)\right)^{\frac{1}{r}}\left(\mathbb{P}_X\left(|\hat{p}_n(X) - p(X)| \geqslant 2^{j-1}\delta\right)\right)^{\frac{1}{s}}
\end{aligned}$$

For all $j \geqslant 1$, we have

$$\begin{aligned}
&\left(\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|^p\right)^{\frac{1}{p}} \left(\mathbb{P}_X\left(2^{j-1}\delta < \left|p(X) - \frac{1}{2}\right| < 2^j\delta\right)\right)^{\frac{1}{r}} \left(\mathbb{P}_X\left(\left|\hat{p}_n(X) - p(X)\right| \geqslant 2^{j-1}\delta\right)\right)^{\frac{1}{s}} \\
&\leqslant C_1^{\frac{1}{p}} a_n C_0^{\frac{\alpha}{q}} \left(2^j\delta\right)^{\frac{\alpha}{r}} \left(\frac{\mathbb{E}\left|\hat{p}_n(X) - p(X)\right|}{2^{j-1}\delta}\right)^{\frac{1}{s}} \\
&\leqslant C_1^{\frac{1}{p}} C_0^{\frac{\alpha}{q}} 2^{\frac{1}{s}} a_n^{1+\frac{1}{s}} 2^{j\left(\frac{\alpha}{r}-\frac{1}{s}\right)} \delta^{\frac{\alpha}{r}-\frac{1}{s}}.
\end{aligned}$$

Sum over $j \leqslant 1$, we obtain

$$\sum_{j=1}^{\lfloor -\log\delta \rfloor} C_1^{\frac{1}{p}} C_0^{\frac{\alpha}{q}} 2^{\frac{1}{s}} a_n^{1+\frac{1}{s}} 2^{j\left(\frac{\alpha}{r}-\frac{1}{s}\right)} \delta^{j\left(\frac{\alpha}{r}-\frac{1}{s}\right)} = C_1^{\frac{1}{p}} C_0^{\frac{\alpha}{q}} 2^{\frac{1}{s}} a_n^{1+\frac{1}{s}} \delta^{\frac{\alpha}{r}-\frac{1}{s}} \frac{2^{\frac{\alpha}{r}-\frac{1}{s}}\left(1 - 2^{\lfloor -\log\delta \rfloor\left(\frac{\alpha}{r}-\frac{1}{s}\right)}\right)}{1 - 2^{\frac{\alpha}{r}-\frac{1}{s}}}.$$

Note that $\frac{1}{2\delta} = 2^{-\log\delta - 1} \leqslant 2^{\lfloor -\log\delta \rfloor} \leqslant 2^{-\log+1} = \frac{2}{\delta}$. Let $\delta \downarrow 0$, if $\frac{\alpha}{r} - \frac{1}{s} > 0$, then $\delta^{\frac{\alpha}{r}-\frac{1}{s}}\left(1 - 2^{\lfloor -\log\delta \rfloor\left(\frac{\alpha}{r}-\frac{1}{s}\right)}\right) = O(1)$. In this case, we have $\frac{\alpha}{r} > \frac{1}{s}$ and $\frac{1}{r} + \frac{1}{s} = \frac{1}{q}$. So the convergence speed is of order $O\left(a_n^{1+\frac{\alpha}{(1+\alpha)q}}\right)$. If $\frac{\alpha}{r} - \frac{1}{s} < 0$, then $\delta^{\frac{\alpha}{r}-\frac{1}{s}}\left(1 - 2^{\lfloor -\log\delta \rfloor\left(\frac{\alpha}{r}-\frac{1}{s}\right)}\right) = O\left(\delta^{\frac{\alpha}{r}-\frac{1}{s}}\right)$. In this case, let $\delta = a_n^v$ for v to be determined, in order to balance the error from the first term, we need

$$1 + v\frac{\alpha}{q} = 1 + \frac{1}{s} + v\left(\frac{\alpha}{r} - \frac{1}{s}\right)$$

which implies $v = \frac{1}{1+\alpha}$. We obtain that the speed is still of order $O\left(a_n^{1+\frac{\alpha}{(1+\alpha)q}}\right)$. ■

6.3 Pointwise Convergence of the Lagrangian Functions

The unknown social welfare potential function $\beta_0 = \gamma(\lambda, g) = \mathbb{E}\left[\sum_j \lambda_j \phi_j^*(X) g_j(X)\right]$ can be estimated by its sample analog $\frac{1}{n'}\sum_{i=1}^{n'}\sum_j \lambda_j \hat{\phi}_j(X_i)\hat{g}_j(X_i)$ for $n' = O(n)$. On the one hand, under full observability, the Neyman-orthogonalized estimator $\frac{1}{n'}\sum_{i=1}^{n'}\sum_j \lambda_j \hat{\phi}_j(X_i) Y_{ij}$ can be combined with a sample splitting scheme for estimating $\hat{\phi}_j(X_i)$ to achieve desirable statistical properties. On the other hand, for a typical treatment allocation problem, Y_{ij} are not fully observable. We only observe $Y_i = \sum_j D_{ij} Y_{ij}$ and need to use the treatment status variables D_{ij} to form a Neyman-orthogonalized estimator.

Theorem 6.3.1. *Let $p, \hat{p}, g, \hat{g} : E \subset \mathbb{R}^{\dim(X_i)} \to \mathbb{R}^{J+1}$ be uniformly bounded functions, where E is an open set. $Var(Y_i|X_i)$ is uniformly bounded. Assume that there exists a constant $\epsilon > 0$ such that $p_j(x), \hat{p}_j(x) > \epsilon$ for all $x \in E$ and for all $j \in \{0, \ldots, J\}$. Assume also that for some sequence a_n, $\lim_{n\to\infty} \frac{a_n}{\sqrt[4]{n}} = +\infty$, there is $\operatorname{ess\,sup} a_n |\hat{p} - p| = o_{\mathbb{P}}(1)$ and $\operatorname{ess\,sup} a_n |\hat{g} - g| = o_{\mathbb{P}}(1)$, where the essential supremum is taken with respect to the distribution of X. Further assume that $\lambda_j g_j - \lambda_l g_l$*

satisfies the MA for all $j \neq l$, $j, l = 0, \ldots, J$ *at* 0 *with* $\alpha = 1$. *Under unconfoundedness and using sample splitting scheme for estimating* $\hat{p}(x)$ *and* $\hat{g}(x)$ *in a treatment allocation model,*

$$\sqrt{n'}\left(\hat{\beta} - \beta_0\right) = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \left(\sum_{j=0}^{J} \lambda_j \phi_j^* (X_i) \left[g_j (X_i) + \frac{D_{ij}}{p_j (X_i)} (Y_i - g_j (X_i)) \right] - \beta_0 \right) + o_{\mathbb{P}}(1) \quad (26)$$

where

$$\hat{\beta} = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \hat{\phi}_j (X_i) \left[\hat{g}_j (X_i) + \frac{D_{ij}}{\hat{p}_j (X_i)} (Y_i - \hat{g}_j (X_i)) \right]. \quad (27)$$

Theorem 6.3.2. *Under the conditions in Theorem 6.3.1 and using sample splitting scheme to estimate* $\hat{\phi}(x)$ *with full observability,*

$$\sqrt{n'}\left(\hat{\beta} - \beta_0\right) = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \left(\sum_{j=0}^{J} \lambda_j \phi_j^* (X_i) Y_{ij} - \beta_0 \right) + o_{\mathbb{P}}(1) \quad (28)$$

where

$$\hat{\beta} = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \hat{\phi}_j (X_i) Y_{ij}.$$

PROOF OF THEOREM 6.3.1. In the following, the notation C should be understood as *some constant*, and they are not necessarily equal to each other. With notation $\mathbb{E}$ and $\mathbb{P}$, the expectation and probability are taken with respect to (X_i, Y_i, D_i) ($o_{\mathbb{P}}(1)$ means convergence in probability with respect to the sample used to estimate $\hat{g}$ and $\hat{p}$). The result will follow if we can show that $\Delta = o_{\mathbb{P}}(1)$, where

$$\Delta = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \left(\lambda_j \hat{\phi}_j (X_i) \left[\hat{g}_j (X_i) + \frac{D_{ij}}{\hat{p}_j (X_i)} (Y_i - \hat{g}_j (X_i)) \right] \right.$$
$$\left. - \lambda_j \phi_j^* (X_i) \left[g_j (X_i) + \frac{D_{ij}}{p_j (X_i)} (Y_i - g_j (X_i)) \right] \right).$$

For this purpose we decompose $\Delta = \Delta_1 + \Delta_2$, where

$$\Delta_1 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \hat{\phi}_j (X_i) \left[\hat{g}_j (X_i) + \frac{D_{ij}}{\hat{p}_j (X_i)} (Y_i - \hat{g}_j (X_i)) - g_j (X_i) - \frac{D_{ij}}{p_j (X_i)} (Y_i - g_j (X_i)) \right]$$

and

$$\Delta_2 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) \left[g_j (X_i) + \frac{D_{ij}}{p_j (X_i)} (Y_i - g_j (X_i)) \right].$$

Also define a linearized approximation of Δ_1 as

$$\Delta_3 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \hat{\phi}_j (X_i) \left[\left(1 - \frac{D_{ij}}{p_j (X_i)} \right) (\hat{g}_j (X_i) - g_j (X_i)) - \frac{D_{ij}}{p_j^2 (X_i)} (Y_i - g_j (X_i)) (\hat{p}_j (X_i) - p_j (X_i)) \right].$$

Then we can write, when $p(x)$ and $\hat{p}(x)$ is bounded away from zero,

$$|\Delta_1 - \Delta_3| \leqslant C \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \left(|\hat{g}_j (X_i) - g_j (X_i)|^2 + |\hat{p}_j (X_i) - p_j (X_i)|^2 \right).$$

By the essential supremum convergence rate assumption, $|\Delta_1 - \Delta_3| = o_{\mathbb{P}}(1)$. To see that $\Delta_3 = o_{\mathbb{P}}(1)$, note that by the split sample scheme, conditional on the estimate $\hat{p}(\cdot)$ and $\hat{g}(\cdot)$, $\mathbb{E}\Delta_3 = 0$,

$$\begin{aligned} Var(\Delta_3) &\leqslant C \sum_{j=0}^{J} \lambda_j^2 \Bigg(\mathbb{E} \left[\frac{1 - p_j (X_i)}{p_j (X_i)} (\hat{g} (X_i) - g (X_i))^2 \hat{\phi}_j (X_i)^2 \right] \\ &\quad + \mathbb{E} \left[\hat{\phi}_j (X_i)^2 Var (Y_{ij} | X_i, D_{ij} = 1) \frac{1}{p_j^3 (X_i)} (\hat{p} (X_i) - p (X_i))^2 \right] \Bigg) \\ &\leqslant C \left(\mathbb{E} (\hat{g} (X_i) - g (X_i))^2 + \mathbb{E} (\hat{p} (X_i) - p (X_i))^2 \right). \end{aligned}$$

Next we decompose $\Delta_2 = \Delta_2^1 + \Delta_2^2$, where

$$\begin{aligned} \Delta_2^1 &= \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) \frac{D_{ij}}{p_j (X_i)} (Y_i - g_j (X_i)), \\ \Delta_2^2 &= \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) g_j (X_i). \end{aligned}$$

To see $\Delta_2^1 = o_{\mathbb{P}}(1)$, note that $\mathbb{E}\Delta_2^1 = 0$. It suffices to show that $Var\left(\Delta_2^1\right) = o_{\mathbb{P}}(1)$ by bounding

$$\begin{aligned}
Var\left(\Delta_2^1\right) &\leqslant \sum_{j=0}^{J} \lambda_j^2 \mathbb{E}\left[\left(\hat{\phi}_j(X_i) - \phi_j^*(X_i)\right)^2 \frac{1}{p_j(X_i)} Var\left(Y_j | X_i, D_{ij} = 1\right)\right] \\
&\leqslant C \sum_{j=0}^{J} \mathbb{E}\left(\hat{\phi}_j(X_i) - \phi_j^*(X_i)\right)^2 \\
&\leqslant C \sum_{j=0}^{J} \sum_{p \neq q} \mathbb{E}1\left(\phi^*(X_i) \neq \hat{\phi}(X_i)\right) \\
&\leqslant C \sum_{j=0}^{J} \sum_{p \neq q} \sum_{k \neq l} \mathbb{E}\left|1_{kl}^*(X_i) - \hat{1}_{kl}(X_i)\right| \\
&\leqslant C \sum_{k \neq l} \mathbb{P}\left(-2\delta_n \leqslant \lambda_k g_k(X) - \lambda_l g_l(X) \leqslant 2\delta_n\right)
\end{aligned}$$

where $\delta_n = C \max_{j=0,\ldots,J} \operatorname{ess\,sup} |\hat{g}_j - g_j|$. By the MA, the right hand side of the last inequality is $O(\delta_n)$.

Since $\Delta_2^2 < 0$, we have

$$\begin{aligned}
-\Delta_2^2 &= \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\phi_j^*(X_i) - \hat{\phi}_j(X_i)\right) g_j(X_i) \\
&= \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left(\lambda_p g_p(X_i) - \lambda_q g_q(X_i)\right) \phi_p^*(X_i) \hat{\phi}_q(X_i) \\
&\leqslant \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left|\lambda_p g_p(X_i) - \lambda_q g_q(X_i) - \left(\lambda_p \hat{g}_p(X_i) - \lambda_q \hat{g}_q(X_i)\right)\right| \phi_p^*(X_i) \hat{\phi}_q(X_i) \\
&\leqslant \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left[\left|\lambda_p g_p(X_i) - \lambda_q g_q(X_i) - \left(\lambda_p \hat{g}_p(X_i) - \lambda_q \hat{g}_q(X_i)\right)\right| 1\left(\phi^*(X_i) \neq \hat{\phi}(X_i)\right)\right] \\
&\leqslant \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left[\left|\lambda_p g_p(X_i) - \lambda_q g_q(X_i) - \left(\lambda_p \hat{g}_p(X_i) - \lambda_q \hat{g}_q(X_i)\right)\right| \sum_{k \neq l} \left|1_{kl}^*(X_i) - \hat{1}_{kl}(X_i)\right|\right] \\
&\leqslant 2(J+1)^2 \delta_n \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{k \neq l} \left|1_{kl}^*(X_i) - \hat{1}_{kl}(X_i)\right|.
\end{aligned}$$

To show that $\Delta_2^2 = o_{\mathbb{P}}(1)$, it suffices to check that

$$\begin{aligned}
-\mathbb{E}\Delta_2^2 &\leqslant \mathbb{E}C\delta_n \sqrt{n'} \sum_{k \neq l} \left|1_{kl}^*(X_i) - \hat{1}_{kl}(X_i)\right| \\
&\leqslant C\sqrt{n'}\delta_n \sum_{k \neq l} \mathbb{P}\left(-2\delta_n \leqslant \lambda_k g_k(X) - \lambda_l g_l(X) \leqslant 2\delta_n\right) \\
&= C\sqrt{n'}\delta_n^2 = o_{\mathbb{P}}(1).
\end{aligned}$$

■

PROOF OF THEOREM 6.3.2. Same as the proof of Theorem 6.3.1, the notation C should be understood as *some constant*, and they are not necessarily equal to each other. The result will follow if we can show that $\Delta = o_{\mathbb{P}}(1)$, where

$$\Delta = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) Y_{ij} = \Delta_1 + \Delta_2$$

with

$$\Delta_1 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) g_j (X_i), \tag{29}$$

and

$$\Delta_2 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) (Y_{ij} - g_j (X_i)).$$

Since $\mathbb{E}\Delta_2 = 0$, it suffices to show $\Delta_2 = o_{\mathbb{P}}(1)$ by bounding,

$$\begin{aligned}
Var(\Delta_2) &\leqslant C \sum_{j=0}^{J} \mathbb{E} \left(\hat{\phi} (X_i) - \phi_j^* (X_i) \right)^2 Var (Y_j | X_i) \\
&\leqslant C \sum_{j=0}^{J} \mathbb{E} \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right)^2 \\
&\leqslant C \sum_{j=0}^{J} \sum_{p \neq q} \mathbb{E} 1 \left(\phi^* (X_i) \neq \hat{\phi} (X_i) \right) \\
&\leqslant C \sum_{j=0}^{J} \sum_{p \neq q} \sum_{k \neq l} \mathbb{E} \left| 1_{kl}^* (X_i) - \hat{1}_{kl} (X_i) \right| \\
&\leqslant C \sum_{k \neq l} \mathbb{P} \left(-2\delta_n \leqslant \lambda_k g_k^* (X) - \lambda_l g_l^* (X) \leqslant 2\delta_n \right)
\end{aligned}$$

where $\delta_n = C \max_{j=0,\ldots,J} \operatorname{ess\,sup} |\hat{g}_j - g_j|$.

We can also bound $|\Delta_1|$ by

$$\begin{aligned}
-\Delta_1 &= \frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{j=0}^{J}\lambda_j\left(\phi_j^*(X_i)-\hat{\phi}_j(X_i)\right)g_j(X_i)\\
&= \frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{p\neq q}\left(\lambda_p g_p(X_i)-\lambda_q g_q(X_i)\right)\phi_p^*(X_i)\hat{\phi}_q(X_i)\\
&\leqslant \frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{p\neq q}\left|\lambda_p g_p(X_i)-\lambda_q g_q(X_i)-\left(\lambda_p\hat{g}_p(X_i)-\lambda_q\hat{g}_q(X_i)\right)\right|\phi_p^*(X_i)\hat{\phi}_q(X_i)\\
&\leqslant \frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{p\neq q}\left[\left|\lambda_p g_p(X_i)-\lambda_q g_q(X_i)-\left(\lambda_p\hat{g}_p(X_i)-\lambda_q\hat{g}_q(X_i)\right)\right|1\left(\phi^*(X_i)\neq\hat{\phi}(X_i)\right)\right]\\
&\leqslant \frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{p\neq q}\left[\left|\lambda_p g_p(X_i)-\lambda_q g_q(X_i)-\left(\lambda_p\hat{g}_p(X_i)-\lambda_q\hat{g}_q(X_i)\right)\right|\sum_{k\neq l}\left|1_{kl}^*(X_i)-\hat{1}_{kl}(X_i)\right|\right]\\
&\leqslant 2(J+1)^2\delta_n\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{k\neq l}|1_{kl}^*(X_i)-\hat{1}_{kl}(X_i)|.
\end{aligned}$$

Given that $-\Delta_1\geqslant 0$, to show that $\Delta_1=o_{\mathbb{P}}(1)$, it suffices to check that

$$\begin{aligned}
-\mathbb{E}\Delta_1 &\leqslant \mathbb{E}C\delta_n\sqrt{n'}\sum_{k\neq l}\left|1_{kl}^*(X_i)-\hat{1}_{kl}(X_i)\right|\\
&\leqslant C\sqrt{n'}\delta_n\sum_{k\neq l}\mathbb{P}\left(-2\delta_n\leqslant\lambda_k g_k(X)-\lambda_l g_l(X)\leqslant 2\delta_n\right)\\
&= C\sqrt{n'}\delta_n^2=o_{\mathbb{P}}(1).
\end{aligned}$$

■

Theorem 6.3.3 (M-estimator rate of convergence, Theorem 3.2.5 in van der Vaart and Wellner (2023)). *Let $\mathbb{M}_n$ be stochastic processes indexed by a pseudometric space Θ and $M:\Theta\to\mathbb{R}$ a deterministic function, such that for every θ in a neighborhood of θ_0, and a constant $C_1>0$*

$$M(\theta)-M(\theta_0)\leqslant -C_1 d^2(\theta,\theta_0).$$

Suppose that, for every n and sufficiently small δ, the centered process $\mathbb{M}_n-M$ satisfies

$$\mathbb{E}^*\sup_{d(\theta,\theta_0)<\delta}|(\mathbb{M}_n-M)(\theta)-(\mathbb{M}_n-M)(\theta_0)|\leqslant C_2\frac{\phi_n(\delta)}{\sqrt{n}},$$

for a constant $C_2>0$ and function ϕ_n such that $\delta\mapsto\frac{\phi_n(\delta)}{\delta^\alpha}$ is decreasing for some $\alpha<2$ not depending on n. Let $r_n^2\phi_n\left(\frac{1}{r_n}\right)\leqslant\sqrt{n}$, If the sequence $\hat{\theta}_n$ satisfies $\mathbb{M}_n\left(\hat{\theta}_n\right)\geqslant\mathbb{M}_n(\theta_0)-O_{\mathbb{P}^*}\left(r_n^{-2}\right)$

and converges in outer probability to θ_0*, then* $r_n d\left(\hat{\theta}_n, \theta_0\right) = O_{\mathbb{P}*}(1)$.

Now, consider two sample analogs of the population Lagrangian minimax problem. The first is the *convexified* problem:

$$\underset{\lambda \in \mathbb{R}_+^J, \lambda_0 = 1}{\arg\min} \frac{1}{n} \sum_{i=i}^{n} \max_{j=0,\ldots,J-1} \left\{\lambda_j \hat{p}_{nj}\left(X_i\right)\right\} - \sum_{j=1}^{J-1} \lambda_j c_j;$$

the second is the doubly robust (double / debiased machine learning) problem:

$$\underset{\lambda \in \mathbb{R}_+^J, \lambda_0 = 1}{\arg\min} \frac{1}{n} \sum_{i=i}^{n} \lambda_j Y_{ij} 1\left(\lambda_j \hat{p}_{nj}\left(X_i\right) > \max_{l \neq j} \left\{\lambda_l \hat{p}_{nl}\left(X_i\right)\right\}\right) - \sum_{j=1}^{J-1} \lambda_j c_j.$$

In the following asymptotic convergence rate discussion, we will assume that the second is numerical solvable. We do not need to worry too much about the convexified problem, since it is convex, even if it is not smooth. Here for brevity, we also assume measurability for all functions and sets involved.

Let $\mathbb{M}_n$ be the objective function of the convexified problem or the doubly robust problem and M be the population optimal value. We denote the out of sample expectation of $\mathbb{M}_n$ as $\tilde{M}$. If we apply the sample splitting scheme, we have

$$\begin{aligned}
&\mathbb{E} \sup_{\|\lambda, \lambda_0\| < \delta} \left|\left(\mathbb{M}_n - M\right)(\lambda) - \left(\mathbb{M}_n - M\right)(\lambda_0)\right| \\
&= \mathbb{E}_{\mathbb{T}_n} \mathbb{E}_{\mathbb{O}_n} \sup_{\|\lambda, \lambda_0\| < \delta} \left|\left(\mathbb{M}_n - \tilde{M}\right)(\lambda) - \left(\mathbb{M}_n - \tilde{M}\right)(\lambda_0) + \left(\tilde{M} - M\right)(\lambda) + \left(M - \tilde{M}\right)(\lambda_0)\right| \\
&\leqslant \underbrace{\mathbb{E}_{\mathbb{T}_n} \mathbb{E}_{\mathbb{O}_n} \sup_{\|\lambda, \lambda_0\| < \delta} \left|\left(\mathbb{M}_n - \tilde{M}\right)(\lambda) - \left(\mathbb{M}_n - \tilde{M}\right)(\lambda_0)\right|}_{①} \\
&+ \underbrace{\mathbb{E}_{\mathbb{T}_n} \sup_{\|\lambda, \lambda_0\| < \delta} \left[\left|\left(\tilde{M} - M\right)(\lambda)\right| + \left|\left(M - \tilde{M}\right)(\lambda_0)\right|\right]}_{②}.
\end{aligned}$$

We first consider the ② term. Note that max is a Lipschitz function and λ is bounded. We obtain that if $\tilde{M}$ corresponds to the convexified problem, then $② = O\left(\mathbb{E}_{\mathbb{T}_n} \mathbb{E} \left\|\hat{p}_n(X) - p(X)\right\|\right)$; if $\tilde{M}$ corresponds to the double robust problem, then under mild regularity conditions, $② = o\left(n^{-\frac{1}{2}}\right)$.

For the ① term, we can first consider the inner expectation. Take the convexified problem as example, we need to bound

$$\mathbb{E}_{\mathbb{O}_n}\left[\sup_{\|\lambda, \lambda_0\| < \delta} \left|\left(\mathbb{M}_n - \tilde{M}\right)(\lambda) - \left(\mathbb{M}_n - \tilde{M}\right)(\lambda_0)\right| \left|\hat{p}_n(\cdot)\right.\right],$$

where $\mathbb{M}_n = \frac{1}{n}\sum_{i=1}^n \max_{j=0,\ldots,J-1}\{\lambda_j \hat{p}_n(X_i)\}$ (the penalty terms $\sum_{j=1}^{J-1}\lambda_j c_j$ are simply cancelled out). By a trick popularized by Bartlett and Mendelson (2002) and might be dated back partially to Hoffmann-Jørgensen (1974), there is

$$\begin{aligned}
&\mathbb{E}_{\mathbb{O}_n}\left[\sup_{\|\lambda,\lambda_0\|<\delta}\left|\left(\mathbb{M}_n - \tilde{M}\right)(\lambda) - \left(\mathbb{M}_n - \tilde{M}\right)(\lambda_0)\right|\left|\hat{p}_n(\cdot)\right.\right] \\
&\leqslant \frac{1}{\sqrt{n}}\mathbb{E}_{\mathbb{O}_n}\mathbb{E}_{\xi}\left[\sup_{\|\lambda,\lambda_0\|<\delta}\left|\mathbb{M}_n^{\xi}(\lambda) - \mathbb{M}_n^{\xi}(\lambda_0)\right|\left|\hat{p}_n(\cdot), \{X_i\}_{i=1}^n\right.\right]
\end{aligned}$$

for the symmetrized process $\mathbb{M}_n^{\xi}(\lambda) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \max_{j=0,\ldots,J-1}\{\lambda_j \hat{p}_{nj}(X_i)\}$. ξ_i here are i.i.d. random variables independent to $\{X_i\}_{i=1}^n$ such that $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$.

The symmetrized process $\mathbb{M}_n^{\xi}(\lambda)$ is sub-Gaussian with the kernel

$$\frac{1}{n}\sum_{i=1}^n\left(\max_{j=0,\ldots,J-1}\{\lambda_j \hat{p}_{nj}(X_i)\} - \max_{j=0,\ldots,J-1}\{\lambda'_j \hat{p}_{nj}(X_i)\}\right)^2. \tag{30}$$

Since $\hat{p}_n$ is uniformly bounded, we have that $\max_{j=0,\ldots,J-1}\{\lambda_j \hat{p}_{nj}\}$ is uniformly Lipschitz with respect to $\hat{p}_n$. In fact, the kernel less of equal to $\frac{1}{n}\sum_{i=1}^n \|\lambda - \lambda'\| = \|\lambda - \lambda'\|$. So we actually can say that for arbitrary $\hat{p}_n$, our symmetrized process are all sub-Gaussian with respect to $\|\cdot\|$.

Consider the covering number of δ-closed ball in $\mathbb{R}^{J-1}$. The volume of δ-closed ball is $C_1\delta^{J-1}$, and the volume of $\frac{1}{2}\epsilon$-closed ball is $C_1\left(\frac{\epsilon}{2}\right)^{J-1}$. Therefore, for the packing number we should have

$$D\left(\frac{1}{2}\epsilon, \|\cdot\|\right) \times \epsilon^{J-1} \leqslant \left(\delta + \frac{1}{2}\epsilon\right)^{J-1}.$$

By the fact that

$$N(\epsilon, \|\cdot\|) \leqslant D(\epsilon, \|\cdot\|) \leqslant N\left(\frac{1}{2}\epsilon, \|\cdot\|\right),$$

we obtain that $N(\epsilon, \|\cdot\|) \leqslant \left(\frac{2\delta}{\epsilon} + 1\right)^{J-1}$.

Now, by standard chaining for sub-Gaussian process (see, for example Theorem 2.3.7 in Giné and Nickl (2021)), we obtain

$$\begin{aligned}
\mathbb{E}_{\xi}\left[\sup_{\|\lambda,\lambda_0\|<\delta}\left|\mathbb{M}_n^{\xi}(\lambda) - \mathbb{M}_n^{\xi}(\lambda_0)\right|\left|\hat{p}_n(\cdot), \{X_i\}_{i=1}^n\right.\right] &\leqslant \left(16\sqrt{2}+2\right)\int_0^{\delta}\sqrt{\log\left(2N(\epsilon, \|\cdot\|)\right)}d\epsilon \\
&\leqslant \left(16\sqrt{2}+2\right)\int_0^{\delta}\left[\sqrt{\log 2} + \sqrt{\log\left(N(\epsilon, \|\cdot\|)\right)}\right]d\epsilon.
\end{aligned}$$

We can then calculate to get

$$\int_0^{\delta} \sqrt{\log N\left(\epsilon, \|\|\right)} d\epsilon = \delta \int_0^1 \sqrt{\log\left(\frac{1}{u}\right)} du \leqslant 0.886227\delta.$$

We obtain that $\phi_n(\delta) := ① + ② = \frac{K\delta}{\sqrt{n}} + O\left(\frac{\mathbb{E}_{\mathbb{T}_n}\mathbb{E}\|\hat{p}_n(X) - p(X)\|}{\sqrt{n}}\right)$ for the convexified problem estimator.[5] It is not hard to see that for $1 < \alpha < 2$, $\frac{\phi_n(\delta)}{\delta^{\alpha}}$ is decreasing.

For the convergence rate r_n for $\hat{\lambda}_n$, if the first step estimator has a speed of order $n^{-\beta}$, then we need that $r_n^2\left(\frac{K}{r_n} + n^{\frac{1}{2}-\beta}\right) \leqslant \sqrt{n}$. This implies that $r_n^2 \leqslant Cn^{\beta}$, i.e. $r_n = O\left(n^{\frac{\beta}{2}}\right)$. Based on Theorem 6.3.2, we obtain that our three step estimator at least has a convergence rate of the same order of the first step estimator.

Next, we consider the convexifed problem in depth with some strong high level conditions. In this draft, we intend to keep this discussion at a heuristic level. Denote the objective function of the convexified problem as $Q_n(\lambda)$. For fixed $\hat{p}_n$, let $\bar{\lambda}_n$ be the minimum point of $\mathbb{E}Q_n(\lambda)$ and λ^* be the minimum point of the population problem. We assume all these solutions are unique. Define

$$G_n(v) = n\left(Q_n\left(\bar{\lambda}_n + \frac{v}{\sqrt{n}}\right) - Q_n\left(\bar{\lambda}_n\right)\right)$$

which is minimized at $\hat{v} = \sqrt{n}\left(\hat{\lambda} - \bar{\lambda}_n\right)$. Then denote $\Gamma_n(v) = \frac{1}{2}v^T H_n\left(\bar{\lambda}_n\right)v + o(1)$ which is the second order approximation of the population objective function at the minimum point. Define

$$R_{i,n} = \max_{j=0,\ldots,J-1}\left\{\left(\bar{\lambda}_{nj} + \frac{v_j}{\sqrt{n}}\right)\hat{p}_{nj}(X_i)\right\} - \max_{j=0,\ldots,J-1}\left\{\bar{\lambda}_{nj}\hat{p}_{nj}(X_i)\right\} - \frac{\langle c, v\rangle}{\sqrt{n}} - \frac{\langle D_{\hat{p}_n,i}, v\rangle}{\sqrt{n}},$$

where

$$D_{\hat{p}_n,ij} = \hat{p}_{nj}(X_i)\,1\left(\bar{\lambda}_{nj}\hat{p}_{nj}(X_i) > \max_{l\neq j}\left\{\bar{\lambda}_{nl}\hat{p}_{nl}(X_i)\right\}\right) - c_j,\ j = 0,\ldots,J-1, c_0 = 0.$$

Note that $|R_{i,n}|$ is bounded by $\max_{j=0,\ldots,J-1}\hat{p}_n(X_i)\frac{\|v\|}{\sqrt{n}}$, and can be nonzero only on $\bigcup_{j,k}A_{jk}$,

$$\begin{aligned}
A_{jk} &= \left\{\bar{\lambda}_{nj}\hat{p}_{nj}(x) \geqslant \bar{\lambda}_{nk}\hat{p}_{nk}(x) \wedge \left(\bar{\lambda}_{nk} + \frac{v_k}{\sqrt{n}}\right)\hat{p}_{nk}(x) \geqslant \left(\bar{\lambda}_{nj} + \frac{v_j}{\sqrt{n}}\right)\hat{p}_{nj}(x)\right\} \\
&= \left\{\bar{\lambda}_{nj}\hat{p}_{nj}(x) \geqslant \bar{\lambda}_{nk}\hat{p}_{nk}(x) \geqslant \bar{\lambda}_{nj}\hat{p}_{nj}(X_i) + \frac{v_j}{\sqrt{n}}\hat{p}_{nj}(X_i) - \frac{v_k}{\sqrt{n}}\hat{p}_{nk}(X_i)\right\} \\
&= \left\{0 \geqslant \bar{\lambda}_{nk}\hat{p}_{nk}(x) - \bar{\lambda}_{nj}\hat{p}_{nj}(x) \geqslant \frac{v_j}{\sqrt{n}}\hat{p}_{nj}(X_i) - \frac{v_k}{\sqrt{n}}\hat{p}_{nk}(X_i)\right\}.
\end{aligned}$$

[5] There are some difficulty in analyzing the doubly robust problem estimator. In this case, it seems that we can not directly get $\phi_n = \frac{K\delta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$ because the kernel of the symmetrized process is not (30), but a noncontinuous function involving Y_i and indicator functions.

If $\bar{\lambda}_{nk}\hat{p}_{nk}(x) - \bar{\lambda}_{nj}\hat{p}_{nj}(x)$ is regular enough uniformly in all possible $\hat{p}$, i.e. we assume a uniform margin assumption, then we obtain $\mathbb{P}(A_{jk}) = O\left(\frac{\|v\|}{\sqrt{n}}\right)$ for a uniform $O(\cdot)$. Then by bounding the second order moment, we have $\sum_{i=1}^{n}(R_{i,n}(v) - \mathbb{E}R_{i,v}(v)) = o_{\mathbb{P}}(1)$.

By convex convergence (see for example Corollary 6.1.4), we have

$$G_n = \frac{1}{2}v^T H_n\left(\bar{\lambda}_n\right) + \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} D_{\hat{p}_n,i}\right)^T v + o_{\mathbb{P}}(1) + o(1).$$

By the uniform margin assumption, the $o_{\mathbb{P}}(1)$ above is uniform with respect to $\hat{p}$ and arbitrary compact set of v. We may further expect Γ_n to have a uniformly precise second order approximation to let the $o(1)$ term to also be uniformly small.

Now, consider an argument in Pollard (1991). Define $\eta_n = -H_n^{-1}\left(\bar{\lambda}_n\right)W_n$, where $W_n = n^{\frac{1}{2}}\sum_{i=1}^{n} D_{\hat{p}_n,i}$. Fix $\delta > 0$, for all v such that $\|v - \eta_n\| > \delta$, write $v = \eta_n + \beta\ell$ for $\beta > \delta$ and ℓ a unit vector. Also let $v^\star = \eta_n + \delta\ell$. Then by convexity

$$\begin{aligned}\frac{\delta}{\beta}G_n(v) + \left(1 - \frac{\delta}{\beta}\right)G_n(\eta_n) &\geqslant G_n(v^\star)\\ &= -\frac{1}{2}W_n^T H_n^{-1}\left(\bar{\lambda}_n\right)W_n + \delta^2\ell^T H_n\ell - W_n^T H_n^{-1}\left(\bar{\lambda}_n\right)\delta\ell - o_{\mathbb{P}}(1)\\ &= G_n(\eta_n) + \delta^2\ell^T H_n\ell - o_{\mathbb{P}}(1).\end{aligned}$$

So we obtain that

$$\inf_{\|v-\eta_n\|>\delta} G_n(v) \geqslant G_n(\eta_n) + \left(\delta^2\ell^T H_n\ell - o_{\mathbb{P}}(1)\right).$$

This implies that $\mathbb{P}(\|\hat{v} - \eta_n\| > \delta) \to 0$, i.e. $\hat{\lambda} - \bar{\lambda}_n = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$.

However, we cannot expect $\hat{\lambda}$ converges to λ^* in the same rate. Actually, if all constraints are binding in the population problem, by functional implicit function theorem for Fréchet differentiability, we can show that $\left\|\bar{\lambda}_n - \lambda^*\right\| = O(\|\hat{p} - p\|)$. So if our asymptotic linearity representation

$$\sqrt{n}\left(\hat{\lambda} - \bar{\lambda}_n\right) = -H_n^{-1}\left(\bar{\lambda}_n\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{\hat{p}_n,i} + o_{\mathbb{P}}(1)$$

would not blow up, we can expect that $\hat{\lambda}$ has the same convergence rate to λ^* as the rate of the convergence of $\hat{p}$. Note that we do not need to assume Donskerness here.

References

Abadeh, Soroosh Shafieezadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn, "Distributionally robust logistic regression," *Advances in neural information processing systems*,

2015, *28*.

Adjaho, Christopher and Timothy Christensen, "Externally valid treatment choice," *arXiv preprint arXiv:2205.05561*, 2022, *1* (1).

Ai, Chunrong, Yue Fang, and Haitian Xie, "Data-driven Policy Learning for a Continuous Treatment," *arXiv preprint arXiv:2402.02535*, 2024.

Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer, 2005.

Andersen, Per Kragh and Richard D Gill, "Cox's regression model for counting processes: a large sample study," *The Annals of Statistics*, 1982, pp. 1100–1120.

Armentano, Diego, Jean-Marc Azaïs, and José Rafael León, "On a general Kac–Rice formula for the measure of a level set," *The Annals of Applied Probability*, 2025, *35* (3), 1828–1851.

Armstrong, Timothy B and Shu Shen, "Inference on optimal treatment assignments," *The Japanese Economic Review*, 2023, *74* (4), 471–500.

Athey, Susan and Stefan Wager, "Policy learning with observational data," *Econometrica*, 2021, *89* (1), 133–161.

Attouch, Hedy, Giuseppe Buttazzo, and Gérard Michaille, *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*, SIAM, 2014.

Audibert, Jean-Yves and Alexandre B Tsybakov, "Fast learning rates for plug-in classifiers," *Annals of statistics*, 2007, *35* (2), 608–633.

Bartlett, Peter L and Shahar Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of machine learning research*, 2002, *3* (Nov), 463–482.

Ben-Michael, Eli, Kosuke Imai, and Zhichao Jiang, "Policy Learning with Asymmetric Counterfactual Utilities," *Journal of the American Statistical Association*, 2024, pp. 1–14.

Bertsekas, Dimitri and Steven E Shreve, *Stochastic optimal control: the discrete-time case*, Vol. 5, Athena Scientific, 1996.

Bhattacharya, Debopam and Pascaline Dupas, "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 2012, *167* (1), 168–196.

Biase, Fausto Di and Steven G Krantz, "Foundations of Fatou theory and a tribute to the work of EM Stein on boundary behavior of holomorphic functions," *The Journal of Geometric Analysis*, 2021, *31* (7), 7184–7296.

Blanchet, Jose and Karthyek Murthy, "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, 2019, *44* (2), 565–600.

Boucheron, Stéphane, Olivier Bousquet, and Gábor Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: probability and statistics*, 2005, *9*, 323–375.

Chang, Kung-ching, *Lecture Notes for Functional Analysis I (in Chinese)*, Peking University Press, 2021.

Chen, Ruidi and Ioannis Ch Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *Journal of Machine Learning Research*, 2018, *19* (13), 1–48.

Chen, Xiaohong and Wayne Yuan Gao, "Semiparametric Learning of Integral Functionals on Submanifolds," *arXiv preprint arXiv:2507.12673*, 2025.

_ , Oliver Linton, and Ingrid Van Keilegom, "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 2003, *71* (5), 1591–1608.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 2018, *21*, C1–C68.

_ , Iván Fernández-Val, and Alfred Galichon, "Quantile and probability curves without crossing," *Econometrica*, 2010, *78* (3), 1093–1125.

_ , _ , and Ye Luo, "The sorted effects method: discovering heterogeneous effects beyond their averages," *Econometrica*, 2018, *86* (6), 1911–1938.

_ , Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins, "Locally robust semiparametric estimation," *Econometrica*, 2022, *90* (4), 1501–1535.

Dellacherie, Claude and Paul-André Meyer, *Probabilities and potential, A*, Vol. 29 of *North-Holland Mathematics Studies*, North-Holland, 1978.

Devroye, Luc, László Györfi, and Gábor Lugosi, *A Probabilistic Theory of Pattern Recognition*, Vol. 31 of *Stochastic Modelling and Applied Probability*, Springer New York, 1996.

Duchi, John C and Hongseok Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, 2021, *49* (3), 1378–1406.

Dudley, Richard M, *Real analysis and probability*, Cambridge University Press, 2002.

Edwards, Darrin C, Charles E Metz, and Robert M Nishikawa, "The hypervolume under the ROC hypersurface of" near-guessing" and" near-perfect" observers in N-class classification tasks," *IEEE transactions on medical imaging*, 2005, *24* (3), 293–299.

Esfahani, Peyman Mohajerin and Daniel Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, 2018, *171* (1), 115–166.

Evans, Lawrence C and Ronald F Garzepy, *Measure theory and fine properties of functions, revised edition*, Chapman and Hall/CRC, 2015.

Fan, Jianqing, Nancy E Heckman, and Matt P Wand, "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *Journal of the American Statistical Association*, 1995, *90* (429), 141–150.

_ , Tien-Chung Hu, and Young K Truong, "Robust non-parametric function estimation," *Scandinavian journal of statistics*, 1994, pp. 433–446.

Fan, Ky, "A generalization of Tychonoff's fixed point theorem," *Mathematische Annalen*, 1961, *142* (3), 305–310.

_ , "Some Properties of Convex Sets Related to Fixed Points Theorems*.," *Mathematische Annalen*, 1984, *266*, 519–538.

Fan, Yanqin, Hyeonseok Park, and Gaoqian Xu, "Quantifying distributional model risk in marginal problems via optimal transport," *Mathematics of Operations Research*, 2025.

_ , Yuan Qi, and Gaoqian Xu, "Policy Learning with α-Expected Welfare," *arXiv preprint arXiv:2505.00256*, 2025.

Feng, Kai, Han Hong, and Denis Nekipelov, "Statistical Inference of Optimal Allocations I: Regularities and their Implications," *arXiv preprint arXiv:2403.18248*, 2025.

Gantmakher, Feliks Ruvimovich, *The theory of matrices*, Vol. 1, Chelsea Publishing Company, 1977.

Gao, Rui and Anton Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research*, 2023, *48* (2), 603–655.

Giné, Evarist and Richard Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge university press, 2021.

Hoffmann-Jørgensen, Jørgen, "Sums of independent Banach space valued random variables," *Studia Mathematica*, 1974, *52* (2), 159–186.

Jin, Ying, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang, "Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality," *The Annals of Statistics*, 2025, *53* (4), 1483–1512.

Kido, Daido, "Distributionally robust policy learning with wasserstein distance," *arXiv preprint arXiv:2205.04637*, 2022.

Kim, Jeankyung and David Pollard, "Cube root asymptotics," *The Annals of Statistics*, 1990, *18* (1), 191–219.

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616.

_ and _ , "Equality-minded treatment choice," *Journal of Business & Economic Statistics*, 2021, *39* (2), 561–574.

_ and Guanyi Wang, "Individualized Treatment Allocation in Sequential Network Games," *arXiv preprint arXiv:2302.05747*, 2023.

_ and _ , "Who should get vaccinated? Individualized allocation of vaccines over SIR network," *Journal of Econometrics*, 2023, *232* (1), 109–131.

Knaster, Bronisław, Kazimierz Kuratowski, and Stefan Mazurkiewicz, "Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexe," *Fundamenta Mathematicae*, 1929, *14* (1), 132–137.

Kumaresan, Somaskandan, *Topology of metric spaces*, Alpha Science Int'l Ltd., 2005.

LeCam, Lucien, "Convergence in distribution of stochastic processes," *University of California Publications in Statistics*, 1957, *2*, 207–236.

Lehmann, Erich Leo and Joseph P Romano, *Testing statistical hypotheses*, 4 ed., Vol. 3 of *Springer Texts in Statistics*, Springer, 2022.

Liang, Annie, Jay Lu, Xiaosheng Mu, and Kyohei Okumura, "Algorithm design: A fairness-accuracy frontier," *arXiv preprint arXiv:2112.09975*, 2021. Journal of Political Economy, forthcoming.

Liu, Yiqi and Francesca Molinari, "Inference for an Algorithmic Fairness-Accuracy Frontier," *arXiv preprint arXiv:2402.08879*, 2024.

Luedtke, Alex and Antoine Chambaz, "Performance guarantees for policy learning," *Annales de l'IHP Probabilites et statistiques*, 2020, *56* (3), 2162.

Luedtke, Alexander R and Mark J van der Laan, "Optimal individualized treatments in resource-limited settings," *The International Journal of Biostatistics*, 2016, *12* (1), 283–303.

Mammen, Enno and Alexandre B Tsybakov, "Smooth discrimination analysis," *The Annals of Statistics*, 1999, *27* (6), 1808–1829.

Manski, Charles F, "Statistical treatment rules for heterogeneous populations," *Econometrica*, 2004, *72* (4), 1221–1246.

Mbakop, Eric and Max Tabord-Meehan, "Model selection for treatment choice: Penalized welfare maximization," *Econometrica*, 2021, *89* (2), 825–848.

Munkres, James R, *Topology*, Prentice Hall, Upper Saddle River, NJ, 2000.

Niculescu, Constantin P and Lars-Erik Persson, *Convex Functions and Their Applications: A Contemporary Approach*, 2 ed., Springer, 2018.

Pollard, David, "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 1991, *7* (2), 186–199.

Qian, Min and Susan A Murphy, "Performance guarantees for individualized treatment rules," *Annals of statistics*, 2011, *39* (2), 1180.

Qu, Zhaonan and Yongchan Kwon, "Distributionally Robust Instrumental Variables Estimation," *arXiv preprint arXiv:2410.15634*, 2024.

Rai, Yoshiyasu, "Statistical inference for treatment assignment policies," 2018. Unpublished Manuscript.

Rockafellar, R Tyrrell and Roger J-B Wets, *Variational analysis*, Vol. 317 of *Grundlehren der mathematischen Wissenschaften*, Springer Science & Business Media, 2009.

Rockafellar, Ralph Tyrell, *Convex Analysis*, Princeton University Press, 1970.

Sasaki, Yuya, "What do quantile regressions identify for general structural functions?," *Econometric Theory*, 2015, *31* (5), 1102–1116.

Schaefer, HH and MP Wolff, *Topological Vector Spaces*, 2 ed., Vol. 3 of *Graduate Texts in Mathematics*, Springer New York, 1999.

Schwartz, Laurent, "Radon measures on arbitrary topological spacesand cylindrical measures," *Tata Institute of Fundamental Research*, 1973.

Semenova, Vira, "Debiased machine learning of set-identified linear models," *Journal of Econometrics*, 2023, *235* (2), 1725–1746.

Shapiro, Alexander, "Distributionally robust stochastic programming," *SIAM Journal on Optimization*, 2017, *27* (4), 2258–2275.

Shen, Yi, Pan Xu, and Michael M Zavlanos, "Wasserstein distributionally robust policy evaluation and learning for contextual bandits," *arXiv preprint arXiv:2309.08748*, 2023.

Sinha, Aman, Hongseok Namkoong, Riccardo Volpi, and John Duchi, "Certifying some distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*, 2017.

Stoye, Jörg, "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 2009, *151* (1), 70–81.

Tetenov, Aleksey, "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 2012, *166* (1), 157–165.

Tsybakov, Alexander B, "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, 2004, *32* (1), 135–166.

van der Vaart, AW and Jon A Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, 2 ed., Springer Nature, 2023.

Villani, Cédric, *Optimal transport: old and new*, Vol. 338 of *Grundlehren der mathematischen Wissenschaften*, Springer, 2009.

Viviano, Davide, "Policy targeting under network interference," *Review of Economic Studies*, 2025, *92* (2), 1257–1292.

Zhan, Ruohan, Zhimei Ren, Susan Athey, and Zhengyuan Zhou, "Policy learning with adaptively collected data," *Management Science*, 2024, *70* (8), 5270–5297.

Zhang, Luhao, Jincheng Yang, and Rui Gao, "A short and general duality proof for Wasserstein distributionally robust optimization," *Operations Research*, 2025, *73* (4), 2146–2155.

Zhao, Chaoyue and Yongpei Guan, "Data-driven risk-averse stochastic optimization with Wasserstein metric," *Operations Research Letters*, 2018, *46* (2), 262–267.

Zhao, Yingqi, Donglin Zeng, A John Rush, and Michael R Kosorok, "Estimating individualized treatment rules using outcome weighted learning," *Journal of the American Statistical Association*, 2012, *107* (499), 1106–1118.

Zhen, Jianzhe, Daniel Kuhn, and Wolfram Wiesemann, "A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle," *Operations Research*, 2025, *73* (2), 862–878.

Zhou, Xin, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok, "Residual weighted learning for estimating individualized treatment rules," *Journal of the American Statistical Association*, 2017, *112* (517), 169–187.

Zhou, Zhengyuan, Susan Athey, and Stefan Wager, "Offline multi-action policy learning: Generalization and optimization," *Operations Research*, 2023, *71* (1), 148–183.

T Technical addendum

T.1 Preliminary definitions and results

T.1.1 Functional Analysis and Topology

Theorem T.1.1. *Banach–Alaoglu theorem*
Let X be a normed space. Then the operator norm unit ball of the dual X^* is compact for the weak topology.*

Theorem T.1.2. *Metrizability for the weak* topology*
Let X be a separable normed space. Then the operator norm unit ball of the dual X^* is metrizable for the weak topology.*

Theorem T.1.3. *A general form of Arzelà-Ascoli theorem*
Let X be a σ-compact Hausdorff space, $\{f_n\}$ is a function sequence $f_n : X \mapsto \mathbb{R}^k$. If $\{f_n\}$ is pointwise bound and equicontinuous, then $\{f_n\}$ has a subsequence that converges to a continuous function with respect to topology of compact convergence.

Theorem T.1.4. *Tychonoff theorem*
Let $\{X_i : i \in I\}$ be a collection of compact spaces. Then $\prod_{i\in I} X_i$ is compact in product topology.

Definition T.1.1. (Radon space) A separable metric space X is a Radon space if every Borel probability measure $\mu \in \mathscr{P}(X)$ is inner regular, i.e. for all $B \in \mathscr{B}$ and for all $\epsilon > 0$, there exists a compact subset K_ϵ of B such that $\mu(B \backslash K_\epsilon) \leqslant \epsilon$.

Definition T.1.2. (Continuous convergence) Let $\mathcal{X}, \mathcal{Y}$ be two metric spaces, and $\{f_n\}$ be a sequence of mappings $f_n : \mathcal{X} \to \mathcal{Y}$. Given $f : \mathcal{X} \to \mathcal{Y}$ then f_n convergence to f continuously if $f_n(x_n) \to f(x)$ whenever $\{x_n\} \subset \mathcal{X}, x \in \mathcal{X}, x_n \to x$.

Lemma T.1.5. *Equivalence between uniform convergence and continuous convergence*
Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces, and $\{f_n\}, f$ be mappings from $\mathcal{X}$ to $\mathcal{Y}$.

1. *If $\mathcal{X}$ is compact and f is continuous then $f_n \to f$ continuously if and only if $f_n \to f$ uniformly in $\mathcal{X}$.*
2. *If $f_n \to f$ continuously then f is continuous.*

As a consequence, if $\mathcal{X}$ is compact then $f_n \to f$ continuously if and only if $f_n \to f$ uniformly in $\mathcal{X}$ and f is continuous.

Proof of Lemma T.1.5.

1. If part: Let d be the metric on $\mathcal{Y}$. For $f_n \to f$ uniformly and $x_n \to x$ we have

$$\begin{aligned} d\left(f_n\left(x_n\right), f\left(x\right)\right) &\leqslant d\left(f_n\left(x_n\right), f\left(x_n\right)\right) + d\left(f\left(x_n\right), f\left(x\right)\right) \\ &\leqslant \sup_{x\in X} d\left(f_n\left(x_n\right), f\left(x_n\right)\right) + d\left(f\left(x_n\right), f\left(x\right)\right). \end{aligned}$$

As $n \to \infty$, the first term goes to 0 by uniform convergence and the second term goes to 0 by continuity.

Only if part: Prove by contradiction. If $f_n \to f$ continuously but not uniformly. Then there is a subsequence $\{n_k\}$ and $\epsilon > 0$ such that for all n_k

$$\sup_{x\in X} d\left(f_{n_k}\left(x\right), f\left(x\right)\right) > 2\epsilon.$$

By the definition of sup there is a sequence $\{x_k\}$ such that

$$d\left(f_{n_k}\left(x_k\right), f\left(x_k\right)\right) > \epsilon. \tag{31}$$

Since $\mathcal{X}$ is a compact metric space there is a convergent subsequence $\{x_{k'}\}$ of $\{x_k\}$ with $x_{k'} \to x_0$. Continuous convergence and continuity of f require

$$d\left(f_{n_{k'}}\left(x_{k'}\right), f\left(x_{k'}\right)\right) \leqslant d\left(f_{n_{k'}}\left(x_{k'}\right), f\left(x_0\right)\right) + d\left(f\left(x_0\right), f\left(x_{k'}\right)\right) \to 0,$$

which violates (31).

2. Let $d_{\mathcal{X}}$ be the metric on $\mathcal{X}$. Suppose $\{x_n\}$ is an arbitrary sequence that converges to x. For all k, consider sequence $\{f_m\left(x_{k,m}\right)\}$, where $d_{\mathcal{X}}\left(x_{k,m}, x_k\right) < \frac{1}{m}$. By continuous convergence,

$$\lim_{m\to\infty} f_m\left(x_{k,m}\right) = f\left(x_k\right).$$

Therefore there exist $n_k \in \mathbb{N}$, such that for all $n > n_k$,

$$d\left(f_n\left(x_{k,n}\right), f\left(x_k\right)\right) < \frac{1}{k}.$$

Next consider a sequence $\{y_n\}$,

$$\{y_n\} = \left\{ \underbrace{x_{1,n_1}}_{n_1-terms}, \underbrace{x_{2,n_1+n_2}}_{n_2-terms}, \ldots, \underbrace{x_{k,\sum_{j=1}^{k} n_j}}_{n_k-terms}, \ldots \right\}.$$

Since $y_n \to x$ as $n \to \infty$ by continuous convergence $f_n\left(y_n\right) \to f\left(x\right)$. Write $\sum_{j=1}^{k} n_j$ as $\sum^{k} n_j$.

Note that $y_{\sum^k n_j} = x_{k,\sum^k n_j}$, $f_{\sum^k n_j}\left(y_{\sum^k n_j}\right) = f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right)$, we have

$$\lim_{k\to\infty} f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right) = f(x).$$

Then by

$$\begin{aligned} d\left(f(x_k), f(x)\right) &\leqslant d\left(f(x_k), f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right)\right) + d\left(f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right), f(x)\right) \\ &\leqslant \frac{1}{k} + o(1), \end{aligned}$$

f is continuous. ■

Definition T.1.3. Let $\mathcal{X}$ be a linear space, $E \subset \mathcal{X}$. A set-valued map $D : E \rightrightarrows \mathcal{X}$ is called a Knaster-Kuratowski-Mazurkiweicz (KKM) map, if $\operatorname{conv}(\mathscr{I}) \subset \bigcup_{x\in\mathscr{I}} D(x)$ for all finite $\mathscr{I} \subset \mathcal{X}$.

Theorem T.1.6 (Knaster-Kuratowski-Mazurkiewicz (KKM) theorem (Knaster et al., 1929))**.** *Let $x_1, \ldots, x_n$ be n points in $\mathbb{R}^d$ and $D_1, \ldots, D_n$ be n closed subset of $\mathbb{R}^d$. If $conv(\{x_i : i \in I\}) \subset \bigcup_{i\in I} D_i$ for all $I \subset \{1, \ldots, n\}$, then $\bigcap_{i=1}^n D_i \neq \varnothing$.*

Theorem T.1.7 (Fan-Knaster-Kuratowski-Mazurkiewicz (FKKM) theorem, Lemma 1 in Fan (1961))**.** *Let $\mathcal{X}$ be a Hausdorff topological vector space, E be an open subset of $\mathcal{X}$ and D a closed-valued KKM map. If there exists $x_0 \in E$ such that $D(x_0)$ is compact, then $\bigcap_{x\in E} D(x) \neq \varnothing$.*

Proof. Define $D_0(x) = D(x) \cap D(x_0)$, we want to verify the finite intersection property of $\{D_0(x) : x \in E\}$. We prove by contradiction. Assume that there exists $\{x_1, \ldots, x_n\} \subset E$ such that $\bigcap_{i=1}^n D_0(x_i) = \varnothing$, i.e. $\bigcap_{i=1}^n D(x_i) = \varnothing$. Consider the finite dimensional subspace $L = \operatorname{Span}(\{x_1, \ldots, x_n\})$. By Theorem 1.3.2 in Schaefer and Wolff (1999), L is normable. Denote the metric induced by a norm of L as d. Since $\bigcap_{i=1}^n D(x_i) = \varnothing$, with the convention $\frac{+\infty}{+\infty} = 1$, we can define for all $\forall x \in L$,

$$t(x) = \sum_{i=1}^n d\left(x, L \cap D(x_i)\right) > 0, \quad \text{and} \quad w_i(x) = \frac{1}{t(x)} d\left(x, L \cap D(x_i)\right), i = 1, \ldots, n.$$

Define $C = \operatorname{conv}(\{x_1, \ldots, x_n\})$ and $\varphi : x \mapsto \sum_{i=1}^n w_i(x) x_i$. By the Brouwer fixed point theorem, there exits $\bar{x}$ such that $\bar{x} = \varphi(\bar{x})$. Let $I(\bar{x}) = \{i : \beta_i(\bar{x}) > 0\}$, by define of $\beta_i(\cdot)$, we have $\bar{x} \notin \bigcup_{i\in I(\bar{x})} D(x_i)$. However, we also have that $\bar{x} \in \operatorname{conv}(\{x_i : i \in I(\bar{x})\})$, which is impossible because D is a KKM map. ■

Theorem T.1.8 (Generalized FKKM theorem, Theorem 3 and Theorem 4 in Fan (1984)). *Let $\mathcal{X}$ be a Hausdorff topological vector space, $\mathcal{Y}$ be a convex subset of $\mathcal{X}$ and $\varnothing \neq E \subset \mathcal{Y}$.*

1. *Assume that $A : E \rightrightarrows \mathcal{Y}$ is relatively-open-valued and satisfies $\bigcup_{x\in E} A(x) = \mathcal{Y}$. If there exists $E_0 \neq \varnothing$ such that $\mathcal{Y}\backslash\bigcup_{x\in E_0} A(x)$ is compact, and E_0 is contained in a compact convex subset C of $\mathcal{Y}$, then there exists a nonempty finite set $\{x_1, \ldots, x_n\} \subset E$ such that $conv(\{x_1, \ldots, x_n\}) \cap \bigcap_{i=1}^n A(x_i) \neq \varnothing$.*

2. *Assume that $D : E \rightrightarrows \mathcal{Y}$ is relatively-closed-valued and is a KKM map. If there exists $E_0 \neq \varnothing$ such that $\bigcap_{x\in E_0} D(x)$ is compact, and E_0 is contained in a compact convex subset C of $\mathcal{Y}$, then $\bigcap_{x\in E} D(x) \neq \varnothing$.*

PROOF.

step 1 *1.* when $\mathcal{Y}\backslash\bigcup_{x\in E_0} A(x) = \varnothing$. Let $D(x) = \mathcal{Y}\backslash A(x)$ and $D_0(x) = D(x) \cap C$. We prove by contradiction. Assume that the statement *1.* is false. Then for all finite $\{x_1, \ldots, x_n\} \subset E_0$, we have $\operatorname{conv}(\{x_1, \ldots, x_n\}) \subset C \cap \bigcup_{i=1}^n D(x_i) = \bigcup_{i=1}^n (D(x_i) \cap C)$. Since C is compact and $D(x)$ is relatively closed in $\mathcal{Y}$ for all $x \in E_0$, we have $D(\cdot) \cap C$ is closed and compact, i.e. D_0 is a KKM map on E_0. By Theorem T.1.7, we obtain that $\bigcap_{x\in E_0} D_0(x) \neq \varnothing$ and thus $\bigcap_{x\in E_0} D(x) \neq \varnothing$. This implies that $\bigcup_{x\in E_0} A(x) = \mathcal{Y}\backslash\bigcap_{x\in E_0} D(x) \neq \mathcal{Y}$, a contradiction.

step 2 *1.* when $\mathcal{Y}\backslash\bigcup_{x\in E_0} A(x) \neq \varnothing$. We still prove by contradiction. For an arbitrary $\{x_1, \ldots, x_n\} \subset E$, denote $E_1 = E_0 \cup \{x_1, \ldots, x_n\}$. We claim that $\operatorname{conv}(E_1)$ is compact. To see this, note that all $x \in E_1$ can be written as a convex combination $x = \lambda_0 c + \sum_{i=1}^n \lambda_i x_i$, where $c \in C$, $\lambda_j \geqslant 0$ for all $j = 0, \ldots, n$ and $\sum_{i=0}^n \lambda_j = 1$. The simplex

$$\Delta := \left\{ (\lambda_0, \ldots, \lambda_n) : \lambda_i \geqslant 0, i = 0, \ldots, n, \sum_{i=1}^n \lambda_i = 1 \right\}$$

is compact in $\mathbb{R}^{n+1}$. The product $C \times \Delta$ is compact in product topology by the Tikhonov theorem. The map $(c, \lambda_0, \ldots, \lambda_n) \mapsto \lambda_0 c + \sum_{i=1}^n \lambda_i x_i$ is continuous in a topological vector space. We obtain that $\operatorname{conv}(E_1)$ is compact by the fact the continuous image of a compact set is compact. Now, let $D(x) = \mathcal{Y}\backslash A(x)$ and $D_1(x) = D(x) \cap \operatorname{conv}(E_1)$. Similar to **step** 1, we have that for all $\{x'_1, \ldots, x'_m\} \subset E_1$, $\operatorname{conv}(\{x'_1, \ldots, x'_m\}) \subset \bigcup_{i=1}^m D_1(x_i)$ and $D_1(\cdot)$ is closed and compact, i.e. D_1 is a KKM map on E_1. By Theorem T.1.7, $\bigcap_{x\in E_1} D_1(x) \neq \varnothing$ which implies that $\bigcap_{x\in E_1} D(x) = \bigcap_{i=1}^m \left(\left(\bigcap_{x\in E_0} D(x) \right) \cap D(x_i) \right) \neq \varnothing$. By assumption $\bigcap_{x\in E_0} D(x)$ is compact, so is also $\left(\bigcap_{x\in E_0} D(x) \right) \cap D(\cdot)$. We can use the finite intersection property to conclude that $\bigcap_{x\in E} \left(\left(\bigcap_{x\in E_0} D(x) \right) \cap D(x) \right) \neq \varnothing$. This implies that $\bigcup_{x\in E} A(x) = \mathcal{Y}\backslash\bigcap_{x\in E} D(x) \neq \mathcal{Y}$, a contradiction.

step 3 Equivalence between *1.* and *2.* For *1.* to *2.*, note that if D satisfies the conditions in *2.* and $\bigcap_{x\in E} D(x) = \varnothing$, then $A : \mathcal{Y}\backslash D$ satisfies the conditions in *1.*. This implies that there exists nonempty $\{x_1, \ldots, x_n\}$ such that there exits $x_0 \in \operatorname{conv}(\{x_1, \ldots, x_n\})$ such that $x_0 \in \mathcal{Y}\backslash\bigcup_{i=1}^n D(x_i)$

and $x_0 \in \bigcup_{i=1}^{n} D(x_i)$, which is impossible. For *2.* to *1.*, we can use the special case Theorem T.1.7 and reduce to **step** 2. ■

T.1.2 Measure Theory

Lemma T.1.9. *Compact measurable selection, Proposition 7.33 in Bertsekas and Shreve (1996) Let Y be a metrizable space, X a compact separable metrizable space, D a closed subset of $Y \times X$, and let $f : D \to \{-\infty\} \cup \mathbb{R}$ be upper semicontinuous. Let $f^* : Y_{\#}D \to \{-\infty\} \cup \mathbb{R}$ be given by*

$$f^*(y) = \max_{x \in X_{\#}D} f(x, y).$$

Then $Y_{\#}D$ is closed in Y, f^* is upper semicontinuous and there exists a Borel measurable function $\varphi : Y_{\#}D \to X$ such that $\{(\varphi(y), y) : y \in Y_{\#}D\} \subset D$ and

$$f(\varphi(y), y) = f^*(y), \ \forall y \in Y_{\#}D.$$

Definition T.1.4. (Outer measure) For a set O, an outer measure is a function

$$\mu^* : 2^O \to [0, \infty] \quad \text{such that}$$

(a) $\mu^*(\varnothing) = 0$.

(b) For arbitrary subsets A, B, $A \subset B \subset O$, $\mu^*(A) \leqslant \mu^*(B)$.

(c) For arbitrary subsets $B_1, B_2, \ldots$ of O,

$$\mu^*\left(\bigcup_{i=1}^{\infty} B_i\right) \leqslant \sum_{i=1}^{\infty} \mu^*(B_i).$$

If an outer measure μ^* is defined on a metric space (O, d) and satisfies

$$d(A, B) > 0 \Rightarrow \mu^*(A \cup B) = \mu^*(A) + \mu^*(B), \quad \forall A, B \subset O.$$

Then μ^* is called a metric outer measure.

Definition T.1.5. (Hausdorff outer measure) Let d be the metric on O. For arbitrary subset $E \subset O$, define its diameter as diam $E = \sup_{x_1, x_2 \in E} d(x_1, x_2)$ with diam $\varnothing = 0$. Let

$$\alpha_k = \frac{\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2} + 1\right)}, \quad k \in \mathbb{N} = \{0, 1, \ldots\},$$

where $\Gamma(\cdot)$ is the gamma function. For arbitrary $E \subset O$, and for $\delta > 0$, define

$$\mathcal{H}_{k,\delta}^{*}(E) = \inf\left\{\sum_{j\geqslant 1}\alpha_k\left(\frac{\text{diam } B_j}{2}\right)^k : E \subset \bigcup_{j\geqslant 1} B_j, \text{diam } B_j \leqslant \delta\right\}.$$

The k-dimensional Hausdorff outer measure of E is defined as $\mathcal{H}_k^*(E) = \lim_{\delta\downarrow 0}\mathcal{H}_{k,\delta}^*(E)$, and can be verified to be a metric outer measure on the metric space (O, d).

Definition T.1.6. (Hausdorff measure) For $k \in \mathbb{N}$, let $\mathcal{M}_k = \mathcal{M}_k(O)$ be the σ-algebra generated by $\mathcal{H}_k^*$ by the Carathéodory criterion (See Theorem T.1.10). The restriction of $\mathcal{H}_k^*$ to $\mathcal{M}_k$, $\mathcal{H}_k = \mathcal{H}_k^*|_{\mathcal{M}_k}$, is called Hausdorff measure on $\mathcal{M}_k$.

Theorem T.1.10. *Carathéodory criterion*
Let μ^* be an outer measure on a set O. A subset $E \subset O$ is said to be μ^*-measurable if

$$\mu^*(T) = \mu^*(T \cap E) + \mu^*(T \cap E^c) \quad \forall T \subset O.$$

Let $\mathcal{M}$ be the collection of all μ^*-measurable sets. Then $\mathcal{M}$ is a σ-algebra, and the restriction of μ^* to $\mathcal{M}$: $\mu = \mu^*|_{\mathcal{M}}, \mu(E) = \mu^*(E), E \in \mathcal{M}$ satisfies:

(a) $(O, \mathcal{M}, \mu)$ is a complete measure space.

(b) If $E \subset O$ and $\mu^*(E) = 0$, then $E \in \mathcal{M}$ and thus $\mu(E) = 0$.

From now on, we may also call a μ^*-measurable set a μ-measurable set. If μ^* is a metric outer measure, then

(c) $\mathcal{B}(O) \subset \mathcal{M}$, i.e. $\mathcal{M}$ contains all Borel sets of O and thus $(O, \mathcal{B}(O), \mu)$ where μ is implicitly further restricted to $\mathcal{B}(O)$ is a Borel measure space.

T.1.3 Classical Analysis and More

Theorem T.1.11. *Rademacher theorem*
Let $m, n \in \{1, 2, \ldots\}$, $E \subset \mathbb{R}^n$ be an open set, $f : E \to \mathbb{R}^m$ be a Lipschitz function, then f is differentiable $\mathcal{L}_n$ a.e. and the gradient ∇f is a measurable function.

Theorem T.1.12. *Whitney extension theorem*
Let $m \in \{1, 2, \ldots\}$, $E \subset \mathbb{R}^m$ be a closed set, $f : E \to \mathbb{R}$, $g : E \to \mathbb{R}^m$ be continuous functions. Denote

$$R(x, a) = \frac{f(x) - f(a) - g(a)(x - a)}{|x - a|}, \; x, a \in E, x \neq a.$$

If for all compact set $C \subset E$,

$$\sup\{|R(x,a)| \,|0 < |x-a| \leqslant \delta, x, a \in C\} \to 0,$$

as $\delta \downarrow 0$. Then there exists a C^1 function $\overline{f} : \mathbb{R}^m \to \mathbb{R}$ such that

$$\overline{f}|_E = f, \quad \overline{f}'|_E = f'.$$

Theorem T.1.13. *Lusin theroem*
Let μ be a Borel regular measure over a metric space X, $m \in \{1,2,\ldots\}$, $f : X \to \mathbb{R}^m$ be a μ measurable function, $E \subset X$ be a μ measurable set, $\mu(E) < \infty$. Then for arbitrary $\epsilon > 0$, there exists a compact set $C \subset E$ such that $\mu(E \backslash C) < \epsilon$ and $f|_C$ is continuous.

Theorem T.1.14. *Egoroff theorem*
Let μ be a Borel regular measure over a metric space X, $m \in \{1,2,\ldots\}$, $\{f_n\}$ be a sequence of μ measurable functions $f_n : X \to \mathbb{R}^m$, $f : X \to \mathbb{R}^m$ be a μ measurable function. If

$$f_n(x) \to f(x), \ \mu \text{ a.e. } x \in E,$$

where $E \subset X$ is μ measurable, $\mu(E) < \infty$. Then for arbitrary $\epsilon > 0$, there exists a μ measurable set $S \subset E$ such that $\mu(E \backslash S) < \epsilon$ and

$$f_n \to f, \text{ uniformly on } S.$$

Definition T.1.7. (Vitali cover) Let $E \subset \mathbb{R}^m$, $m \in \{1,2,\ldots\}$. If $\mathcal{V}$ is a collection of closed balls or closed cubes in $\mathbb{R}^m$ such that for all $x \in E$ and arbitrary $\epsilon > 0$, there exists $B \in \mathcal{V}$ such that $x \in B$ and $\text{diam}B < \epsilon$, then $\mathcal{V}$ is called a Vitali cover of E.

Theorem T.1.15. *Vitali covering theorem*
Let $E \subset \mathbb{R}^m$, $\mathcal{V}$ is a Vitali cover of E. Then, there exists an at most countable disjoint subset $\{B_j\} \subset \mathcal{V}$, such that

$$\mathcal{L}_m^* \left(E \backslash \bigcup_j B_j \right) = 0.$$

Theorem T.1.16. *Isodiametric inequality*
For all set $E \subset \mathbb{R}^m$, $m \in \{1,2,\ldots\}$,

$$\mathcal{L}^*(E) \leqslant \alpha_m \left(\frac{\text{diam } E}{2} \right)^m.$$

Theorem T.1.17. *Coincidence between Spherical Hausdorff and Hausdorff outer measures*
For all set $E \subset \mathbb{R}^m$, $m \in \{1, 2, \ldots\}$,

$$\mathcal{H}_m^{S*} = \mathcal{H}_m^*,$$

where $\mathcal{H}_m^{S*}$, the spherical Hausdorff outer measure is defined as

$$\mathcal{H}_m^{S*} = \lim_{\delta \downarrow 0} \inf \left\{ \sum_{j \geqslant 1} \alpha_m \left(\frac{\text{diam } B_j}{2} \right)^m : E \subset \bigcup_{j \geqslant 1} B_j, \text{diam } B_j \leqslant \delta, B_j \text{ is a closed ball} \right\}.$$

Definition T.1.8. (Clarke subgradients) Consider a locally Lipschitz function $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^n$ is an open subset. For each $x \in \Omega$, define

$$\begin{aligned} f^\circ (x; v) &:= \limsup_{\substack{y \to x \\ \lambda \downarrow 0}} \frac{f(y + \lambda v) - f(y)}{\lambda} \\ &= \limsup_{\substack{\varepsilon \to 0 \\ \epsilon \downarrow 0}} \left\{ \frac{f(y + \lambda v) - f(y)}{\lambda} : y \in \Omega \cap B(x, \varepsilon), \lambda \in (0, \epsilon) \right\}, \forall v \in \mathbb{R}^n, y + \lambda v \in \Omega, \end{aligned}$$

and the Clarke subgradient[6] of f at x:

$$\partial f(x) := \{ \xi \in \mathbb{R}^n : f^\circ (x; v) \geqslant \langle v, \xi \rangle, \forall v \in \mathbb{R}^n \}.$$

Theorem T.1.18. *Characterization of Clarke subgradients in $\mathbb{R}^n$*
Let $f : \Omega \to R$ be a locally Lipschitz function, where $\Omega \subset \mathbb{R}^n$ is an open set, then

$$\partial f(x) = conv \left\{ \lim_{k \to \infty} \nabla f(x_k) : x_k \to x, \nabla f(x_k) \text{ exists} \right\}.$$

Definition T.1.9. (Clarke Jacobian) Let $F : \Omega \to \mathbb{R}^m$ be a locally Lipschitz function, where $\Omega \subset \mathbb{R}^n$ is an open set and $m > 1$. The Clarke Jacobian of F at $x \in \Omega$, denoted as $J_c F(x)$, is

$$J_c F(x) := \text{conv} \left\{ \lim_{k \to \infty} JF(x_k) : x_k \to x, JF(x_k) \text{ exists} \right\}.$$

Theorem T.1.19. *Nonsmooth implicit function theorem*
Let $F : \Omega \to \mathbb{R}^m$ be a locally Lipschitz (C^1) function, where $\Omega \subset \mathbb{R}^{n+m}$ is an open set. Assume that (x_0, y_0) is such that

1. *$F(x_0, y_0) = 0$.*
2. *$J_{c,y} F(x_0, y_0)$ is full rank in the sense that all matrices in $J_{c,y} F(x_0, y_0)$ is full rank, where*

[6] This is also called Clarke subdifferential in some literatures, or generalized gradient and generalized directional derivative by Francis Clarke.

$J_{c,y}F(x_0,y_0)$ consists of all $m\times m$ component matrices in $J_cF(x_0,y_0)$ written as $[A_{m\times n},B_{m\times m}]$.

Then there exists an $(n+m)$-dimensional interval $I=I_x^n\times I_y^m\subset\Omega$, where for some positive vectors α and β,

$$I_x=\{x\in\mathbb{R}^n:|x-x_0|<\alpha\},\quad I_y=\{y\in\mathbb{R}^m:|y-y_0|<\beta\},$$

where $|x-x_0|<\alpha$ means that $|x_i-x_{0,i}|<\alpha_i$ for $\alpha=(\alpha_1,\ldots,\alpha_n)$, and a Lipschitz ($C^1$) function $\xi:I_x^n\to I_y^m$ such that for all $(x,y)\in I_x^n\times I_y^m$,

$$F(x,y)=0\Leftrightarrow y=\xi(x).$$

Definition T.1.10. (Sub (sup) differentiability) Let Ω be an open set of $\mathbb{R}^n$, and $f:\Omega\to\mathbb{R}$ a function. Then f is said to be subdifferentiable at x, with subgradient p, if

$$f(x')\geqslant f(x)+\langle p,x'-x\rangle+o(\|x'-x\|).$$

The convex set of all subgradients p at x will be denoted by $\nabla^-f(x)$. If $(-f)$ is subdifferentiable, then f is said to be supdifferentiable, and the convex set of the negated subgradients for $(-f)$ at x is denoted as ∇^+f.

Theorem T.1.20. *Morse-Sard theorem*
Let $f\in C^r(\Omega,\mathbb{R}^m)$, where $\Omega\in\mathbb{R}^n$ is an open set, $r>\max(n-m,0)$, then the set of critical values of f is of zero Lebesgue measure and is meager.

Theorem T.1.21. *Morse-Sard theorem in Sobolev spaces*
Let $f\in W_{loc}^{n-m+1,p}(\Omega,\mathbb{R}^m)$, where $\Omega\in\mathbb{R}^n$ is an open set, $p>m\geqslant 1$. Then the set of critical values of f is of zero Lebesgue measure.

Theorem T.1.22. *(Hausdorff dimension of critical values)*

Theorem T.1.23. *Partition of unity*
Let $E_1,\ldots,E_k$ be open sets in $\mathbb{R}^n$ and K a compact subset of $\bigcup_j K_j$. Then one can find $\phi_j\in C_0^\infty(E_j)$ so that $\phi_j\geqslant 0$ and $\sum_1^k\phi_j\leqslant 1$ with equality in a neighborhood of K.

Definition T.1.11. (Lusin property (N)) Let $(X,\mathcal{F},\mu)$ and $(Y,\mathcal{E},\nu)$ be two measure spaces. A function $f:X\to Y$ has the Lusin Property (N) if for all $N\subset X$ such that $\mu(N)=0$, there holds $\nu(f(N))=0$.

Definition T.1.12. (Dini derivative) Let $f:\mathbb{R}\to\mathbb{R}$, denote

$$D^+f(x)=\limsup_{h\downarrow 0}\frac{f(x+h)-f(x)}{h},\quad D_+f(x)=\liminf_{h\downarrow 0}\frac{f(x+h)-f(x)}{h},$$
$$D^-f(x)=\limsup_{h\uparrow 0}\frac{f(x+h)-f(x)}{h},\quad D_-f(x)=\liminf_{h\uparrow 0}\frac{f(x+h)-f(x)}{h}.$$

They are called upper right, lower right, upper left and lower left Dini derivatives.

Theorem T.1.24. *Dini derivative to Lusin (N)*
Let $E\subset\mathbb{R}$ and $f:E\to\mathbb{R}$. If $D^+f(x)$ is finite $\forall x\in E$, then f has Lusin property (N) on E.

Theorem T.1.25. *Lusin (N) + BV = AC*
Let $f:[a,b]\to\mathbb{R}$ be a continuous function of bounded variation, where $a<b$, $a,b\in\mathbb{R}$. Then f is absolutely continuous if and only if it fulfills Lusin property (N) on $[a,b]$.

T.1.4 Probability

Theorem T.1.26. *Hoeffding inequality*
Let X_i, $i\in\{1,\dots,n\}$ be independent centred random variables taking values in $[a_i,b_i]$ for $-\infty<a_i<0\leqslant b_i<+\infty$, respectively. Denote $S_n=\sum_{i=1}^n X_i$, then for all $t>0$,

$$\mathbb{E}e^{tS_n}\leqslant\exp\left(\frac{t^2\sum_{i=1}^n(b_i-a_i)^2}{8}\right),$$

and for all $t\geqslant 0$,

$$\mathbb{P}\{S_n\geqslant t\}\leqslant\exp\left(-\frac{2t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right),\quad \mathbb{P}\{S_n\leqslant -t\}\leqslant\exp\left(-\frac{2t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right).$$

Theorem T.1.27. *Bennett, Prokhorov and Bernstein inequalities*
Let X_i, $i\in\{1,\dots,n\}$ be independent centred random variables taking variables bounded by $c<\infty$ in absolute value. Denote $\sigma^2=\frac{1}{n}\sum_{i=1}^n\mathbb{E}X_i^2$ and $S_n=\sum_{i=1}^n X_i$. Then, for all $t\geqslant 0$,

$$\begin{aligned}\mathbb{P}\{S_n\geqslant t\}&\leqslant\exp\left(-\frac{n\sigma^2}{c^2}\left[\left(1+\frac{tc}{n\sigma^2}\right)\log\left(1+\frac{tc}{n\sigma^2}\right)-\frac{tc}{n\sigma^2}\right]\right)&&(Bennett)\\&\leqslant\exp(-\frac{3t}{4c}\log\left(1+\frac{2tc}{3n\sigma^2}\right))&&(Prokhorov)\\&\leqslant\exp\left(-\frac{t^2}{2\left(n\sigma^2+\frac{tc}{3}\right)}\right)&&(Bernstein)\end{aligned}$$

Theorem T.1.28. *Talagrand inequality with Bousquet upper tail*

Let (S,Σ) be a measurable space, and let $X_i,\ldots,X_n$ be independent S-valued random variables. Let $\mathcal{F}$ be a (countable) set of measurable centred real-valued functions on S such that for all $f\in\mathcal{F}$, $\|f\|_\infty\leqslant U<\infty$. Denote $S_j=\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^j f(X_i)\right|$, $j\in\{1,\ldots,n\}$, $\sigma^2=\frac{1}{n}\sum_{i=1}^n\sup_{f\in\mathcal{F}}\mathbb{E}f^2(X_i)$ and $v_n=2U\mathbb{E}S_n+n\sigma^2$. Then,

$$\begin{aligned}\mathbb{P}\{S_n-\mathbb{E}S_n\geqslant t\}&\leqslant\mathbb{P}\left\{\max_{1\leqslant j\leqslant n}S_j-\mathbb{E}S_n\geqslant t\right\}\\&\leqslant\exp\left(-\frac{v_n}{U^2}\left[\left(1+\frac{tU}{v_n}\right)\log\left(1+\frac{tU}{v_n}\right)-\frac{tU}{v_n}\right]\right)\\&\leqslant\exp\left(-\frac{3t}{4U}\log\left(1+\frac{2tU}{3v_n}\right)\right)\\&\leqslant\exp\left(-\frac{t^2}{2\left(v_n+\frac{tU}{3}\right)}\right).\end{aligned}$$

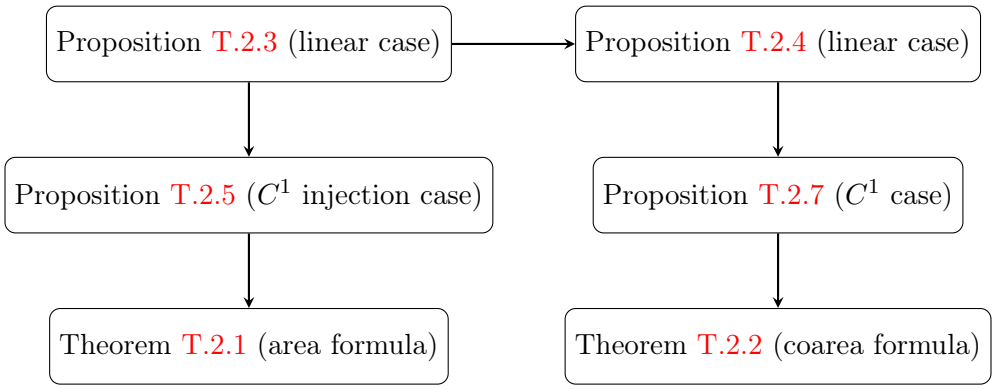
The Banach–Alaoglu theorem T.1.1 is standard in many references involving functional analysis, see for example Theorem 2.4.8 in Attouch et al. (2014). For the form of Arzelà-Ascoli theorem T.1.3 used here, see Exercise 47.4 of Munkres (2000). The Tychonoff theorem is also standard and can be found in Theorem 37.3 in Munkres (2000). For the Rademacher theorem T.1.11, see Theorem 2.10.43 in Federer (1969). For the Whitney extension theorem T.1.12, see Theorem 6.10 in Evans and Garzepy (2015) or Theorem 3.1.14 in Federer (1969). For the form of the Lusin theorem T.1.13 and the Egoroff theorem T.1.14 used here, see Theorem 2.3.5 and Theorem 2.3.7 in Federer (1969), respectively. For Vitali covering theorem T.1.15, see Theorem 1.10 in Falconer (1986). For the isodiametric inequality T.1.16, see Theorem 2.4 in Evans and Garzepy (2015) or Corollary 2.10.33 of Federer (1969). For the relationship between the spherical Hausdorff outer measure and the Hausdorff outer measure, see 2.10.6 in Federer (1969). For the nonsmooth implicit function theorem T.1.19, see Clarke (1976) and Theorem 11 of Hiriart-Urruty (1979). The Morse-Sard theorem in Sobolev spaces T.1.21 originates from de Pascale (2001). For more information about the Morse-Sard type theorem we refer to Figalli (2008) and its renowned infinite dimensional version Sard-Smale theorem from Smale (1965). The form of the partition of unity theorem T.1.23 is taken from Theorem 2.2 in Grigoryan (2009). See also Theorem 3.5 of Grigoryan (2009) for a manifold version. Most of these results have been rewritten to comply with the convention and styles used in Evans and Garzepy (2015) and Lu (2019). The proof of part 1 of Lemma T.1.5 comes from Resnick (1987), we remove the separable space requirement in the premise. Actually, if $\mathcal{X}$ is a compact metric space, then $\mathcal{X}$ is separable, a countable dense subset of $\mathcal{X}$ can be constructed by the total boundedness of $\mathcal{X}$. Theorem T.1.24 with respect to the upper right Dini derivative is a weaker form of Theorem IX.4.6 of Saks (1937). For the characterization of absolutely continuous function theorem T.1.25, see Theorem VII.6.7 of Saks (1937). Some elementary information about the Dini derivatives can be found in Hagood and Thomson (2006). The form of the Hoeffding inequality, the Bennett, Prokhorov and Bernstein inequalities and the Talagrand inequality is taken

from Giné and Nickl (2021) Theorem 3.1.2, Theorem 3.1.7 and Theorem 3.3.9.

T.2 Proofs of area and coarea formulas

Figure 1 illustrates the ordering of the proofs for area and coarea formulas. Besides Federer (1969) and Evans and Garzepy (2015), we also borrow lots of material from Jerrard (2013) and Lu (2019), especially from the latter. Corollaries T.2.6 and T.2.8 are used to illustrate the applicability of the area and coarea formulas T.2.1 and T.2.2.

Figure 1: Ordering of proofs for area and coarea formulas



Theorem T.2.1. *Area formula*
Let $E \subset \mathbb{R}^k$ be an open set, $k \in \{1, 2, \ldots, n\}$, $\psi : E \mapsto \mathbb{R}^n$ be a C^1 or Lipschitz function. Then for all measurable $S \subset E$,

$$\int_S J\psi(x)\, d\mathcal{L}_k x = \int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y. \tag{32}$$

If a measurable function $f : S \mapsto \mathbb{R}$ is nonnegative or if the left hand side of (33) is finite, then the following equality holds,

$$\int_S f(x)\, J\psi(x)\, d\mathcal{L}_k x = \int_{\psi(S)} \left[\int_{S\cap\psi^{-1}(y)} f(x)\, d\mathcal{H}_0 x\right] d\mathcal{H}_k y. \tag{33}$$

Theorem T.2.2. *Coarea formula*
Let $E \subset \mathbb{R}^n$ be an open set, $k \in \{1, 2, \ldots, n\}$, $\varphi : E \mapsto \mathbb{R}^k$ be a C^1 or Lipschitz function, then for all measurable $S \subset E$,

$$\int_S J\varphi(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y. \tag{34}$$

If a measurable function $f : S \mapsto \mathbb{R}$ is nonnegative or if the left hand side of (35) is finite, then the following equality holds,

$$\int_S f(x) J\varphi(x) \, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \left[\int_{S \cap \varphi^{-1}(y)} f(x) \, d\mathcal{H}_{n-k} x \right] d\mathcal{H}_k y. \tag{35}$$

Proposition T.2.3. *Let $k \in \{1, 2, \ldots, n\}$, matrix $A \in \mathbb{R}^{n \times k}$, then*

(i) if $\operatorname{rank}(A) < k$, that is $\det\left(A^T A\right) = 0$, then $\mathcal{H}_k(A(E)) = 0$ for all $E \subset \mathbb{R}^k$.

(ii) if $\operatorname{rank}(A) = k$, that is $\det\left(A^T A\right) > 0$, then for all $E \subset \mathbb{R}^k$,

$$E \text{ is Lebesgue measurable} \Leftrightarrow A(E) \text{ is } \mathcal{H}_k \text{ measurable.}$$

When either of the sides holds, $\mathcal{H}_k(A(E)) = \sqrt{\det(A^T A)} \mathcal{L}_k(E)$.

Proposition T.2.4. *Let $k \in \{1, 2, \ldots, n\}$, matrix $A \in \mathbb{R}^{k \times n}$, then*

(i) if $\operatorname{rank}(A) < k$, that is $JA = \sqrt{\det(AA^T)} = 0$, then $\mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right) = 0$ for $\mathcal{L}_k$ a.e. $y \in \mathbb{R}^k$ for all $E \subset \mathbb{R}^n$.

(ii) if $\operatorname{rank}(A) = k$, that is $\det\left(AA^T\right) > 0$, then for all measurable $E \subset \mathbb{R}^n$, $y \in \mathbb{R}^k \mapsto \mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right)$ is $\mathcal{L}_k$ measurable, and

$$JA \cdot \mathcal{L}_n(E) = \int_E JA(x) \, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right) d\mathcal{L}_k y.$$

Proof Outline of Proposition T.2.3.

step 1 The Hausdorff outer measure $\mathcal{H}_k^*$ is invariant under orthogonal transformations, i.e. if A is a orthogonal transformations then

$$\mathcal{H}_k^*(A(E)) = \mathcal{H}_k^*(E), \forall E \subset \mathbb{R}^k$$

step 2 Let $A = (I_k, 0)^T$, prove that

$$\mathcal{H}_k^*(A(E)) = \mathcal{H}_k^*(E \times \{0\}) = \mathcal{H}_k^*(E) = \mathcal{L}_k^*(E)$$

where $\mathcal{L}_k^*$ is the k-dimensional Lebesgue outer measure.

step 3 There exists orthogonal $T \in \mathbb{R}^{n \times n}$,

$$\mathcal{H}_k^*(A(E)) = \mathcal{H}_k^*(TA(E)) = \mathcal{H}_k^*((A_1(E)) \times \{0\}) = \mathcal{L}_k^*(A_1(E)),$$

where $A_1 \in \mathbb{R}^{k\times k}$, $A \in \mathbb{R}^{n\times k}$, and $TA = \left(A_1^T, 0\right)^T$.
step 4 Proof of

$$\mathcal{L}_k^* \left(A_1 \left(E\right)\right) = |\det\left(A_1\right)| \mathcal{L}_k^* \left(E\right) = \sqrt{\det\left(A^T A\right)} \mathcal{L}_k^* \left(E\right).$$

step 5 For all $E \subset \mathbb{R}^k$, E is Lebesgue measurable $\Leftrightarrow$ $A\left(E\right)$ is $\mathcal{H}_k$ measurable. ■

Remark T.2.1. We provide two supplementary details to assist with understanding the proof outlines of Proposition T.2.3 and the following results.

1. Hausdorff outer measure $\mathcal{H}_k^*$ has the invariance under the orthogonal transformations. This is a direct result of the fact that Hasudroff outer measure keeps distance inequality, i.e. if $k, l, m, n \in \mathbb{Z}^+$, $E \subset \mathbb{R}^l$, map $f : E \to \mathbb{R}^n$ and map $g : E \to \mathbb{R}^m$ satisfies

   $$|f\left(x\right) - f\left(y\right)| \leqslant C|g\left(x\right) - g\left(y\right)| \quad \forall x, y \in E,$$

   then $\mathcal{H}_k^* \left(f\left(E\right)\right) \leqslant C^k \mathcal{H}_k^* \left(g\left(E\right)\right)$. This property will also be used in the following discussion.

2. A cube is a subset of $\mathbb{R}^k$ in the form of

   $$\prod_{i=1}^k \left(a_i, b_i\right), \quad \prod_{i=1}^k \left(a_i, b_i\right], \quad \prod_{i=1}^k \left[a_i, b_i\right), \quad \prod_{i=1}^k \left(a_i, b_i\right) \quad a_i < b_i, \forall i \in \{1, 2, \ldots, k\}.$$

   Let $E \in \mathbb{R}^n$ be a nonempty open set, then there exists a sequence of disjoint left open and right closed cubes $\{Q_k\}_{k=1}^\infty$, such that

   $$E = \bigcup_{k=1}^\infty Q_k = \bigcup_{k=1}^\infty \bar{Q}_k.$$

Proposition T.2.5. *Let $E \subset \mathbb{R}^k$ be an open set, $k \in \{1, 2, \ldots, n\}$, $\psi : E \to \mathbb{R}^n$ be a C^1 injection and $\det\left(\psi'\left(x\right)^T \psi'\left(x\right)\right) > 0$ for all $x \in E$, then for all $D \subset \psi\left(E\right)$,*

$$D \text{ is } \mathcal{H}_k \text{ measurable} \Leftrightarrow \psi^{-1}\left(D\right) \text{ is Lebesgue measurable.}$$

When D is $\mathcal{H}_k$ measurable,

$$\mathcal{H}_k\left(D\right) = \int_{\psi^{-1}(D)} \sqrt{\det\left(\psi'\left(x\right)^T \psi'\left(x\right)\right)} d\mathcal{L}_k x.$$

Proposition T.2.5 plays the most essential role in a proof of more general area formula Theorem T.2.1. Proposition T.2.5 can also be powerful when used alone.

Proof Outline of Proposition T.2.5.

step 1 Estimates of $|\psi(x)-\psi(y)|$: Let $k\in\{1,2,\ldots,n\}$, $E\subset\mathbb{R}^k$ be an open set, $\psi:E\to\mathbb{R}^n$ be a C^1 injection such that $\det\left(\psi'(x)^T\psi'(x)\right)>0$ for all $x\in E$, then:

1. if $K\subset E$ is convex and compact, then there exists $0<c<C$ such that

$$c|x-y|\leqslant|\psi(x)-\psi(y)|\leqslant C|x-y| \tag{36}$$

   for all $x,y\in K$.

2. For arbitrary $x_0\in E$ and for all $0<\epsilon<1$, there exists $\delta>0$, such that open ball $B(x_0,\delta)\subset E$ and

$$(1-\epsilon)|\psi'(x_0)(x-y)|\leqslant|\psi(x)-\psi(y)|\leqslant(1+\epsilon)|\psi'(x_0)(x-y)| \tag{37}$$

   for all $x,y\in B(x_0,\delta)$.

step 2 Prove that, for all $D\subset\psi(E)$,

$$D \text{ is } \mathcal{H}_k \text{ measurable} \Leftrightarrow \psi^{-1}(D) \text{ is Lebesgue measurable,}$$

by estimates (36), the fact that Hausdorff outer measure keeps distance inequality, and the fact that if D is a $\mathcal{H}_k$ measurable set with $\mathcal{H}_k(D)<\infty$, then there exist a Borel set P and a $\mathcal{H}_k$ zero measure set Z such that $D=P\cup Z$[7].

step 3 By (37) and Proposition T.2.3, closed cube $Q\subset E$ satisfies: for all $x_0\in E$, for all $0<\epsilon<1$, there exists $\delta>0$, such that, if $\operatorname{diam}(Q)<\delta$ and $x_0\in Q$ then

$$\begin{aligned}(1-\epsilon)^k\sqrt{\det\left(\psi'(x_0)^T\psi'(x_0)\right)}\mathcal{L}_k(Q)&\leqslant\mathcal{H}_k(\psi(Q))\\&\leqslant(1+\epsilon)^k\sqrt{\det\left(\psi'(x_0)^T\psi'(x_0)\right)}\mathcal{L}_k(Q)\end{aligned} \tag{38}$$

step 4 Prove that, for all closed cube $Q\subset E$,

$$\mathcal{H}_k(\psi(Q))=\int_Q\sqrt{\det\left(\psi'(x)^T\psi'(x)\right)}d\mathcal{L}_kx$$

by (38).

[7]This Borel set, zero measure set construction of $\mathcal{H}_k$ measurable set is also a common useful result. Actually, the complete result states that there exist Borel sets P_1,P_2 and $\mathcal{H}_k$ zero measure sets Z_1,Z_2, such that $D=P_1\cup Z_1=P_2\backslash Z_2$.

step 5 Prove that for all bounded open set E_b such that $\overline{E_b} \subset E$, if $O \subset E_b$ is $\mathcal{L}_k$ measurable then

$$\mathcal{H}_k\left(\psi\left(O\right)\right) = \int_O \sqrt{\det\left(\psi'\left(x\right)^T \psi'\left(x\right)\right)} d\mathcal{L}_k x.$$

Conclude using the fact that any open set $E \subset \mathbb{R}^k$ can be decomposed to a countable disjoint union of bounded cube. ■

Proof of Theorem T.2.1.

A classical proof based on Proposition T.2.5 can be separated into three fundamental parts.

part 1 In case that ψ is not necessarily bijective, while still requiring that $J\psi\left(x\right) > 0$ for all $x \in E$, by the implicit function theorem, for all $x \in E$ there exist a neighborhood U such that ψ is bijective in U. Take a Vitali cover $\mathcal{V}$ of E such that ψ is bijective in every closed ball $B \in \mathcal{V}$. Then by the Vitali covering theorem, there exists an at most countable disjoint subset $\{B_j\} \subset \mathcal{V}$, such that $\mathcal{L}_k^*\left(E \backslash \bigcup_j B_j\right) = 0$. From the definition of $\mathcal{H}_0$,

$$\sum_j 1\left(y \in \psi\left(S \cap B_j\right)\right) = \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right).$$

By the property of Lebesgue integral and Proposition T.2.5,

$$\int_S J\psi\left(x\right) d\mathcal{L}_k x = \sum_j \int_{\psi(S)} 1\left(y \in \psi\left(S \cap B_j\right)\right) d\mathcal{H}_k y = \int_{\psi(S)} \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y.$$

Then note that

$$\begin{aligned}
&\int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\
&= \int_{\psi(S)} \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y + \int_{\psi(S) \backslash \psi\left(S \backslash \bigcup_j B_j\right)} \mathcal{H}_0\left(\left(S \backslash \bigcup_j B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\
&\qquad + \int_{\psi\left(S \backslash \bigcup_j B_j\right)} \mathcal{H}_0\left(\left(S \backslash \bigcup_j B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\
&= \int_{\psi(S)} \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y + 0 + 0.
\end{aligned}$$

When $J\psi\left(x\right) = 0$, let $\epsilon > 0$, define $\psi_\epsilon : E \to \mathbb{R}^{k+n}$ as

$$x \mapsto \left(\epsilon x, \psi\left(x\right)\right),$$

$Crit(\psi) = \{x \in E | J\psi(x) = 0\}$. Note that $J\psi_\epsilon(x) > 0$ for all $x \in E$,

$$\begin{aligned}\int_{Crit(\psi)} J\psi_\epsilon(x)\, d\mathcal{L}_k x &= \int_{\psi_\epsilon(Crit(\psi))} \mathcal{H}_0\left(Crit(\psi) \cap \psi_\epsilon^{-1}(y)\right) d\mathcal{H}_k y \\ &\geqslant \int_{\psi_\epsilon(Crit(\psi))} d\mathcal{H}_k y = \mathcal{H}_k\left(\psi_\epsilon\left(Crit(\psi)\right)\right).\end{aligned}$$

By the fact that Hausdorff outer measure keeps distance inequality,

$$\mathcal{H}_k\left(\psi\left(Crit(\psi)\right)\right) \leqslant \mathcal{H}_k\left(\psi_\epsilon\left(Crit(\psi)\right)\right),$$

since the coordinate projection from $\psi_\epsilon(Crit(\psi))$ to $\psi(Crit(\psi))$ satisfies a distance inequality with $C = 1$. Therefore,

$$\mathcal{H}_k\left(\psi\left(Crit(\psi)\right)\right) \leqslant \int_{Crit(\psi)} J\psi_\epsilon(x)\, d\mathcal{L}_k x,$$

the right hand side converges to 0 as $\epsilon \downarrow 0$ if E is bounded. Then conclude that $\mathcal{H}_k\left(\psi\left(Crit(\psi)\right)\right) = 0$ for (not necessarily bounded) open set $E \subset \mathbb{R}^k$ by the cube decomposition in the **step** 5 of the sketch of the proof of Proposition T.2.5. Now, one can conclude that (32) is true for C^1 function ψ.

part 2 To verify (32) when ψ is a Lipschitz function but does not necessarily belong to C^1, a Lusin type approximation of ψ can be used. To continue, we use Rademacher theorem T.1.11, Whitney extension theorem T.1.12, Lusin theroem T.1.13 and Egoroff theorem T.1.14.

Assume first E is bounded, by Rademacher theorem T.1.11, ψ is differentiable $\mathcal{L}_k$ *a.e.* and the gradient $\psi' \leqslant \text{Lip}\psi$ is measurable, where

$$\text{Lip}\psi = \sup\left\{\frac{|\psi(x_1) - \psi(x_2)|}{|x_1 - x_2|} | x_1, x_2 \in E, x_1 \neq x_2\right\}.$$

Apply Lusin theorem to ψ', there exists a compact set $C \subset E$ such that $\mathcal{L}_k(E \backslash C) < \frac{1}{2}\epsilon$ and $\psi'|_C$ is continuous. Let

$$R(x, a) = \frac{\psi(x) - \psi(a) - \psi'(a)(x - a)}{|x - a|},\ x, a \in C, x \neq a,$$

since ψ is differentiable, for all $a \in C$,

$$R(a) = \sup\{|R(x, a)| | 0 < |x - a| \leqslant \delta, x \in C\} \to 0,$$

as $\delta \downarrow 0$. Then by Egoroff theorem and regularity of Lebesgue measure, there exists a compact set $C' \subset C$ such that $\mathcal{L}_k(C \backslash C') < \frac{1}{2}\epsilon$ and $R(a)$ converge to 0 uniformly on C'. Now, we can apply Whitney extension theorem to ψ and ψ' (actually, to each component function of ψ and its

gradient), i.e. there exists a C^1 function $\overline{\psi}$ such that

$$\begin{aligned} \overline{\psi}|_{C'}(x) = \psi|_{C'}(x), \quad \overline{\psi}'|_{C'}(x) = \psi'|_{C'}(x), \\ \mathcal{L}_k\left(\left\{x|\overline{\psi}|_E(x) \neq \psi(x)\right\}\right) < \epsilon, \quad \mathcal{L}_k\left(\left\{x|\overline{\psi}'|_E(x) \neq \psi'(x)\right\}\right) < \epsilon. \end{aligned} \tag{39}$$

Now, one can conclude that (32) is true for Lipschitz function ψ and set $S \cap C'$.

part 3 The final step is to verify a Lusin property (N) for

$$\int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y.$$

Specifically, for arbitrary measurable $S \subset E$,

$$\int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y \leqslant (\mathrm{Lip}\psi)^n \mathcal{L}_k(S).$$

To see this, let

$$\mathcal{Q}_m = \left\{Q|Q = \prod_{i=1}^{k} (a_i, b_i], a_i = \frac{c_i}{m}, b_i = \frac{c_i+1}{m}, c_i \in \mathbb{Z}\right\},$$

and

$$g_m(y) = \sum_{Q \in \mathcal{Q}_m} 1\left(y \in \psi(S \cap Q)\right).$$

Since $\mathbb{R}^k = \bigcup_{Q \in \mathcal{Q}_m}$, and $g_m(y)$ is the number of cubes $Q \in \mathcal{Q}_m$ such that

$$\mathcal{H}_0\left(S \cap Q \cap \psi^{-1}(y)\right) > 0.$$

Therefore, for all $y \in \mathbb{R}^n$,

$$g_m(y) \uparrow \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right),$$

as $m \to \infty$. Then by the monotone convergence theorem

$$\begin{aligned} \int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y &= \lim_{m\to\infty} \int_{\psi(S)} g_m(y)\, d\mathcal{H}_k y \\ &= \lim_{m\to\infty} \sum_{Q \in \mathcal{Q}_m} \mathcal{H}_k\left(\psi(S \cap Q)\right) \\ &\leqslant \lim_{m\to\infty} \sum_{Q \in \mathcal{Q}_m} (\mathrm{Lip}\psi)^k \mathcal{L}_k(S \cap Q) \\ &= (\mathrm{Lip}\psi)^k \mathcal{L}_k(S). \end{aligned}$$

Now note that,

$$\int_{S\cap C'} J\psi\left(x\right) d\mathcal{L}_k x \leqslant \int_{\psi(S)} \mathcal{H}_0\left(S\cap\psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \leqslant \int_{S\cap C'} J\psi\left(x\right) d\mathcal{L}_k x + \left(\text{Lip}\psi\right)^k \mathcal{L}_k\left(S\backslash C'\right),$$

where $\mathcal{L}_k\left(S\backslash C'\right) < \epsilon$. Note that $J\psi$ is bounded on E due to the Lipschitz continuity of ψ (By Rademacher theorem, $J\psi$ exists for $\mathcal{L}_k$ *a.e.* $x \in E$), i.e. there exists a constant M, such that

$$\int_{S\backslash C'} J\psi\left(x\right) d\mathcal{L}_k x \leqslant M\mathcal{L}_k\left(S\backslash C'\right).$$

Therefore,

$$\begin{aligned}\int_{S\cap C'} J\psi\left(x\right) d\mathcal{L}_k x &\leqslant \int_S J\psi\left(x\right) d\mathcal{L}_k x \\ &\leqslant \int_{S\cap C'} J\psi\left(x\right) d\mathcal{L}_k x + M\mathcal{L}_k\left(S\backslash C'\right),\end{aligned}$$

and (32) follows from the arbitrariness of ϵ. The case when open set $E \subset \mathbb{R}^k$ is unbounded follows from the cube decomposition. ∎

Corollary T.2.6. *A Sard type lemma*
Let $k \in \{1,2,\ldots,n\}$, $E \subset \mathbb{R}^k$ be an open set, $\psi : E \to \mathbb{R}^n$ be a C^1 function and $Crit\left(\psi\right) = \{x \in E | J\psi\left(x\right) = 0\}$, then $\mathcal{H}_k\left(\psi\left(Crit\left(\psi\right)\right)\right) = 0$.

Proof. Since ψ is a C^1 function, $J\psi : E \to \mathbb{R}$ is continuous. Therefore, $Crit\left(\psi\right) = \{x \in E | J\psi\left(x\right) \geqslant 0\}\backslash\{x \in E | J\psi\left(x\right) > 0\}$ is $\mathcal{L}_k$ measurable. By (32),

$$\begin{aligned}\int_{Crit(\psi)} J\psi\left(x\right) d\mathcal{L}_k x &= \int_{\psi(Crit(\psi))} \mathcal{H}_0\left(Crit(\psi)\cap\psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\ &\geqslant \int_{\psi(Crit(\psi))} d\mathcal{H}_k y = \mathcal{H}_k\left(\psi\left(Crit\left(\psi\right)\right)\right).\end{aligned}$$

The integral in the left hand side of the first equality is 0. ∎

Proof Outline of Proposition T.2.4.
step 1 To prove *(i)*, note that

$$\mathcal{L}_k^*\left(A\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(U\circ\Sigma\circ V\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(U\circ\Sigma\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(\Sigma\left(\mathbb{R}^n\right)\right),$$

by SVD and the fact that Lebesgue outer measure is invariant under orthogonal transformation.

Since $\operatorname{rank}(A) < k$, $\operatorname{rank}(\Sigma) < k$, therefore,

$$\mathcal{L}_k^*(\Sigma(\mathbb{R}^n)) = \mathcal{L}_k^*\left(\mathbb{R}^{\operatorname{rank}(\Sigma)}\right) = 0.$$

step 2 By SVD / PD, $A = WPV$, where $V \in \mathbb{R}^{n\times n}$ is orthogonal, $P : \mathbb{R}^n \to \mathbb{R}^k$ is the coordinate projection of the first k dimensions, and $W \in \mathbb{R}^{k\times k}$ is symmetric[8].
step 3 By the Fubini-Tonelli theorem, $y \in \mathbb{R}^k \mapsto \mathcal{L}_{n-k}\left(\left\{(x_1,\ldots,x_{n-k}) : x \in V(E) \cap (P)^{-1}(y)\right\}\right)$ is $\mathcal{L}_k$ measurable and

$$\mathcal{L}_n(E) = \mathcal{L}_n(V(E)) = \int_{\mathbb{R}^k} \mathcal{L}_{n-k}\left(\left\{(x_1,\ldots,x_{n-k}) : x \in V(E) \cap (P)^{-1}(y)\right\}\right) d\mathcal{L}_k y.$$

Then, note that

$$\begin{aligned}\mathcal{L}_{n-k}\left(\left\{(x_1,\ldots,x_{n-k}) : x \in V(E) \cap (P)^{-1}(y)\right\}\right) &= \mathcal{H}_{n-k}\left(V(E) \cap (P)^{-1}(y)\right)\\ &= \mathcal{H}_{n-k}\left(E \cap A^{-1} \circ W(y)\right)\end{aligned}$$

by $A^{-1} = V^{-1} \circ (P)^{-1} \circ W^{-1}$.
step 4 By Proposition T.2.3 and a standard approximation procedure,

$$JM \cdot \int_{\mathbb{R}^k} f(M(x))\, d\mathcal{L}_k x = \int_{\mathbb{R}^k} f(y)\, d\mathcal{L}_k y.$$

for all $M \in \mathbb{R}^{k\times k}$ invertible and f nonnegative $\mathcal{L}_k$ measurable. Therefore,

$$JW \cdot \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(E \cap A^{-1} \circ W(y)\right) d\mathcal{L}_k y = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right) d\mathcal{L}_k y.$$

■

Proposition T.2.7. *Let $E \subset \mathbb{R}^n$ be an open set, $k \in \{1,2,\ldots,n\}$, $\varphi : E \to \mathbb{R}^k$ be a C^1 function, then for all measurable S, $S \cap \varphi^{-1}(y)$ is $\mathcal{H}_{n-k}$ measurable for $\mathcal{H}_k$ a.e. y, $y \in \mathbb{R}^k \mapsto \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ is $\mathcal{H}_k$ measurable, and*

$$\int_S J\varphi(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y.$$

Remark T.2.2.

1. We should note that for open set $E \subset \mathbb{R}^n$, $k \in \{1,2,\ldots,n\}$, $\varphi : E \to \mathbb{R}^k$ be at least continuous, $S \subset E$ measurable, then, $\varphi(S)$ **is not necessarily $\mathcal{H}_k$ measurable**. Actually,

[8] We do not distinguish between linear transformation and its matrix.

if $\varphi : E \subset \mathbb{R}^n \to \mathbb{R}^m, m, n \in \{1, 2, \ldots\}$, φ is continuous, then

$$\varphi \text{ map } \mathcal{L}_n \text{ measurable subset of } E \text{ to } \mathcal{L}_m \text{ measurable set in } \mathbb{R}^m \Leftrightarrow$$
$$\varphi \text{ map } \mathcal{L}_n \text{ zero measure subset of } E \text{ to } \mathcal{L}_m \text{ zero measure set in } \mathbb{R}^m.$$

Even if φ is more smooth than continuous, the right hand side of above relationship will not be automatically satisfied.

2. Although $S \cap \varphi^{-1}(y)$ may not be $\mathcal{H}_{n-k}$ measurable for all $y \in \mathbb{R}^k$, $E \cap \varphi^{-1}(y)$ is $\mathcal{H}_{n-k}$ measurable for all $y \in \mathbb{R}^k$, since E is open and $\varphi^{-1}(y)$ is a Borel set.

Proposition T.2.7 states one of the most essential idea of more general coarea formula Theorem T.2.2. Besides, we should first verify that the integrand on the right hand side of (34) is well defined. A classical proof of Proposition T.2.7 can be separated into two fundamental parts.

Proof of Proposition T.2.7.

part 1 We start from verifying a Lusin property (N) for

$$\int_{\mathbb{R}^k} \mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y.$$

First of all, the outer integral of f is defined as

$$\int^*_{\mathbb{R}^k} f(x)\, d\mathcal{H}_k x = \inf\left\{\int_{\mathbb{R}^k} g(x)\, d\mathcal{H}_k x | g \text{ is } \mathcal{H}_k \text{ measurable }, f \leqslant g \text{ a.e.}\right\}.$$

For arbitrary measurable $S \subset E$, the required Lusin property (N) is provided by Eilenberg inequality which states that

$$\int^*_{\mathbb{R}^k} \mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \leqslant \frac{\alpha_{n-k}\alpha_k}{\alpha_n} (\text{Lip}\varphi)^k \mathcal{H}_n(S) \tag{40}$$

holds. Besides, we can also show that $\mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ is well defined and $\mathcal{H}_k$ measurable, and therefore the outer integral in (40) is actually redundant.

To verify (40), we use the isodiametric inequality T.1.16 and the coincidence between spherical Hausdorff outer measure and Hausdorff outer measure T.1.17. By the equivalence in Theorem T.1.17, for all $l > 0$, there exists an at most countable collection of closed balls $\{B^l_j\}$ such that $S \subset \bigcup_j B^l_j$, diam $B^l_j < \frac{1}{l}$ for all j, and

$$\sum_i \alpha_n \left(\frac{\text{diam } B^l_j}{2}\right)^n \leqslant \mathcal{H}_n(S) + \frac{1}{j}.$$

Define

$$g_j^l(y)=\alpha_{n-k}\left(\frac{\operatorname{diam}\ B_j^l}{2}\right)^{n-k}1\left(y\in\varphi\left(B_j^l\right)\right),$$

since B_j^l is a closed ball, g_j^l is $\mathcal{H}^k$ measurable. Note that $\{B_j^l\}$ covers $A\cap\varphi^{-1}(y)$ for all y,

$$\mathcal{H}^*_{n-k,\frac{1}{l}}\left(S\cap\varphi^{-1}(y)\right)\leqslant\sum_j g_j^l(y).$$

Then by the Fatou lemma, the isodiametric inequality, and the fact that Hausdorff outer measure keeps distance inequality,

$$\begin{aligned}
\int_{\mathbb{R}^k}^*\mathcal{H}^*_{n-k}\left(S\cap\varphi^{-1}(y)\right)d\mathcal{H}_k y&=\int_{\mathbb{R}^k}^*\lim_{l\to\infty}\mathcal{H}^*_{n-k,\frac{1}{l}}\left(S\cap\varphi^{-1}(y)\right)d\mathcal{H}_k y\\
&\leqslant\int_{\mathbb{R}^k}\liminf_{l\to\infty}\sum_j g_j^l(y)\,d\mathcal{H}_k y\\
&\leqslant\liminf_{l\to\infty}\int_{\mathbb{R}^k}\sum_j g_j^l(y)\,d\mathcal{H}_k y\\
&\leqslant\liminf_{l\to\infty}\sum_j\int_{\mathbb{R}^k}g_j^l(y)\,d\mathcal{H}_k y\\
&\leqslant\liminf_{l\to\infty}\sum_j a_{n-k}\left(\frac{\operatorname{diam}\ B_j^l}{2}\right)^{n-k}\mathcal{H}_k\left(\varphi\left(B_j^l\right)\right)\\
&=\liminf_{l\to\infty}\sum_j a_{n-k}\left(\frac{\operatorname{diam}\ B_j^l}{2}\right)^{n-k}\mathcal{L}_k\left(\varphi\left(B_j^l\right)\right)\\
&\leqslant\liminf_{l\to\infty}\sum_j a_{n-k}\left(\frac{\operatorname{diam}\ B_j^l}{2}\right)^{n-k}\alpha_k\left(\frac{\operatorname{diam}\ \varphi\left(B_j^l\right)}{2}\right)^{k}\\
&\leqslant\frac{\alpha_{n-k}\alpha_k}{\alpha_n}(\mathrm{Lip}\varphi)^k\liminf_{l\to\infty}\sum_j\left(\frac{\operatorname{diam}\ B_j^l}{2}\right)^{n}\\
&\leqslant\frac{\alpha_{n-k}\alpha_k}{\alpha_n}(\mathrm{Lip}\varphi)^k\mathcal{H}_n(S).
\end{aligned}$$

Next, we should verify that $S\cap\varphi^{-1}(y)$ is $\mathcal{H}_{n-k}$ measurable for $\mathcal{H}_k$ *a.e.* y, and $y\in\mathbb{R}^k\mapsto\mathcal{H}_{n-k}\left(S\cap\varphi^{-1}(y)\right)$ is $\mathcal{H}_k$ measurable. First, assuming that S is compact, we can write

$$\mathcal{H}^*_{n-k}\left(S\cap\varphi^{-1}(y)\right)=\lim_{\delta\downarrow0}\mathcal{H}^*_{n-k,\delta}\left(S\cap\varphi^{-1}(y)\right)=\sup_{\delta>0}\mathcal{H}^*_{n-k,\delta}\left(S\cap\varphi^{-1}(y)\right).$$

Note that in this case, $S\cap\varphi^{-1}(y)$ is a Borel set and thus $\mathcal{H}_{n-k}$ measurable, therefore it suffices to verify that $y\mapsto\mathcal{H}^*_{n-k,\delta}\left(S\cap\varphi^{-1}(y)\right)$ is $\mathcal{H}_k$ measurable for all $\delta>0$. Actually, $\mathcal{H}^*_{n-k,\delta}\left(S\cap\varphi^{-1}(y)\right)$

is upper semicontinuous. To see this, note that the spherical Hausdorff outer measure can also be defined by open balls, and thus for arbitrary $\epsilon > 0$, there exits an at most countable collection of open balls $\{B_j\}$ such that $S \cap \varphi^{-1} \subset \bigcup_j B_j$, for all j, diam $B_j \leqslant \delta$, and

$$\sum_j \alpha_{n-k} \left(\frac{\text{diam } B_j}{2} \right)^{n-k} \leqslant \mathcal{H}^*_{n-k,\delta} \left(S \cap \varphi^{-1}(y) \right) + \epsilon.$$

The compactness of S implies that if $|y - y'|$ small enough, then $S \cap \varphi^{-1}(y') \subset \bigcup_j B_j$. Therefore,

$$\limsup_{y' \to y} \mathcal{H}^*_{n-k,\delta} \left(S \cap \varphi^{-1}(y') \right) \leqslant \mathcal{H}^*_{n-k,\delta} \left(S \cap \varphi^{-1}(y') \right) + \epsilon, \tag{41}$$

then the upper semicontinuity follows from the arbitrariness of ϵ. Second, let S be just measurable, then by the regularity of Lebesgue measure, there exists a sequence of compact sets $C_1 \subset C_2 \subset \ldots$ such that $S \backslash \bigcup_{i=1}^{\infty} C_i$ is of zero Lebesgue measure. By the cube decomposition of open set and the Eilenberg inequality,

$$\int_{\mathbb{R}^k}^* \mathcal{H}^*_{n-k,\delta} \left(\left(S \backslash \bigcup_{i=1}^{\infty} C_i \right) \cap \varphi^{-1}(y) \right) d\mathcal{H}_k y = 0,$$

i.e. $\mathcal{H}^*_{n-k,\delta} \left(\left(S \backslash \bigcup_{i=1}^{\infty} C_i \cap \varphi^{-1}(y) \right) \right) = 0$, $\mathcal{H}_k$ *a.e.* y. The *a.e.* measurability of $S \cap \varphi^{-1}(y)$ and measurability of $\mathcal{H}^*_{n-k} \left(S \cap \varphi^{-1}(y) \right)$ follow from

$$\mathcal{H}^*_{n-k} \left(S \cap \varphi^{-1}(y) \right) = \mathcal{H}^*_{n-k} \left(\bigcup_{i=1}^{\infty} C_i \cap \varphi^{-1}(y) \right) + \mathcal{H}^*_{n-k} \left(\left(S \backslash \bigcup_{i=1}^{\infty} C_i \cap \varphi^{-1}(y) \right) \right),$$

and now we can use $\mathcal{H}_{n-k} \left(S \cap \varphi^{-1}(y) \right)$ instead of $\mathcal{H}^*_{n-k} \left(S \cap \varphi^{-1}(y) \right)$.

part 2 Assuming that $J\varphi(x) > 0$ for all $x \in E$. We use an estimates of $|\varphi(x) - \varphi(y)|$ similar to (37): Let $k \in \{1, 2, \ldots, n\}$, $E \subset \mathbb{R}^n$ be an open set, $\varphi : E \to \mathbb{R}^k$ be a C^1 function such that $J\varphi(x) > 0$ for all $x \in E$, then for arbitrary $x_0 \in E$ and for all $\epsilon > 0$, there exists $\delta > 0$, such that open ball $B(x_0, \delta) \subset E$ and

$$|\varphi(x) - \varphi(y) - \varphi'(x_0)(x - y)| \leqslant \epsilon |x - y| \tag{42}$$

for all $x, y \in B(x_0, \delta)$.

Let $x_0 \in E$, since $J\varphi(x_0) > 0$, without loss of generality, assuming that the first k columns $\left\{ \frac{\partial}{\partial x_1}\varphi(x_0), \frac{\partial}{\partial x_2}\varphi(x_0), \ldots, \frac{\partial}{\partial x_k}\varphi(x_0) \right\}$ are linear independent. Define $\Phi(x) = (\varphi(x), x_{k+1}, \ldots, x_n)$, by the implicit function theorem, Φ is a bijection on a neighborhood of x_0. By definition,

$$\begin{aligned} \Phi(x) &= \left(\Phi(x) - \Phi'(x_0)(x - x_0) \right) + \Phi'(x_0)(x - x_0) \\ &= \Phi'(x_0) \left[\left(\Phi'(x_0) \right)^{-1} \left(\Phi(x) - \Phi'(x_0)(x - x_0) \right) + (x - x_0) \right], \end{aligned}$$

denote the term in square brackets on the right hand side of the second equality as $g(x)$, then by estimate (42), for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$(1-\epsilon)|x-y| \leqslant |g(x)-g(y)| \leqslant (1+\epsilon)|x-y|, \tag{43}$$

for all $x, y \in B(x_0, \delta)$. Therefore, $\varphi = A \circ g$ on $B(x_0, \delta)$, where $A = P\Phi'(x_0)$, $P : \mathbb{R}^n \to \mathbb{R}^k$ is the coordinate projection of the first k dimensions. By (43), $1-\epsilon \leqslant |g'| \leqslant 1+\epsilon$, by definition,

$$(J\varphi(x))^2 = \det\left(\varphi'(x)\varphi'(x)^T\right) = \det\left(A \circ g'(x) \circ g'(x)^T \circ A^T\right).$$

By SVD / PD, $g'(x) \circ g'(x)^T = Q^T C Q$, where C is diagonal with $(1-\epsilon)^2 \leqslant c_{ii} \leqslant (1+\epsilon)^2$, $i \in \{1, 2, \ldots, n\}$, $Q^T Q = I_n$; $A = PU^T$, where $P \in \mathbb{R}^{k\times k}$ is symmetric and $U \in \mathbb{R}^{n\times k}$ is orthogonal. Therefore,

$$\det\left(A \circ g'(x) \circ g'(x)^T \circ A^T\right) = \det\left(PU^T Q^T C Q U P^T\right) = (\det(A))^2 \det\left(U^T Q^T C Q U\right).$$

Note that QU is also orthogonal, then

$$(1-\epsilon)^{2k} \leqslant \det\left(U^T Q^T C Q U\right) \leqslant (1+\epsilon)^{2k}.$$

As a result,

$$(1-\epsilon)^k JA \leqslant J\varphi(x) \leqslant (1+\epsilon)^k JA \tag{44}$$

on $B(x_0, \delta)$. Compare (44) and (38), then see that the ideas of the **step** 4 and **step** 5 of the proof sketch of Proposition T.2.5 can be used here.

When $J\varphi(x) = 0$, let $\epsilon > 0$, define $\varphi_\epsilon : E \times \mathbb{R}^k \to \mathbb{R}^k$ as

$$(x, z) \mapsto \varphi(x) + \epsilon z,$$

$Crit(\varphi) = \{x \in E | J\varphi(x) = 0\}$. Note that $J\varphi_\epsilon(x, z) > 0$ for all $(x, z) \in E \times \mathbb{R}^k$, for arbitrary $w \in \mathbb{R}^k$,

$$\begin{aligned}\int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y &= \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y - \epsilon w)\right) d\mathcal{H}_k y \\ &= \frac{1}{\alpha_k} \int_{B(0,1)} \int_{\mathbb{R}^{n-k}} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y - \epsilon w)\right) d\mathcal{H}_k y d\mathcal{H}_k w.\end{aligned}$$

Let $P' : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ be the coordinate projection of the last k dimensions, $S' = S \times B(0, 1) \subset$

$\mathbb{R}^{n+k}$. Then note that for all $y \in \mathbb{R}^k$, $w \in B(0,1)$,

$$S' \cap \varphi_\epsilon^{-1}(y) \cap \left(P'\right)^{-1}(w) = \left(S \cap \varphi^{-1}(y - \epsilon w)\right) \times \{w\}.$$

Therefore, by the Eilenberg inequality and the Fubini-Tonelli theorem,

$$\begin{aligned}
&\int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \\
&\qquad = \frac{1}{\alpha_k} \int_{B(0,1)} \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S' \cap \varphi_\epsilon^{-1}(y) \cap \left(P'\right)^{-1}(w)\right) d\mathcal{H}_k y d\mathcal{H}_k w \\
&\qquad \leqslant \frac{\alpha_{n-k}}{\alpha_n} \int_{\mathbb{R}^k} \left(\mathrm{Lip} P'\right)^k \mathcal{H}_n\left(S' \cap \varphi_\epsilon^{-1}(y)\right) d\mathcal{H}_k y \\
&\qquad \leqslant \frac{\alpha_{n-k}}{\alpha_n} \int_{S'} J\varphi_\epsilon(x,z)\, d\mathcal{H}_{n+k}(x,z) \leqslant \frac{\alpha_{n-k}\alpha_k}{\alpha_n} \mathcal{L}_n(S) \sup_{(x,z)\in S'} J\varphi_\epsilon(x,z).
\end{aligned}$$

The last inequality above uses the fact that $\mathcal{H}_{n+k} = \mathcal{L}_{n+k}$ is the completion of $\mathcal{L}_n \times \mathcal{L}_k$. If E is bounded, the right hand side of the last inequality converges to 0 as $\epsilon \downarrow 0$. Then conclude for (not necessarily bounded) open set $E \subset \mathbb{R}^n$ by the cube decomposition. ∎

Proof of Theorem T.2.2.

We show that Theorem T.2.2 follows from Proposition T.2.7, using the Rademacher-Whitney-Lusin-Egoroff framework as in **part** 2 of the proof sketch of the Theorem T.2.1. Without loss of generality, let E be bounded. Suppose we already find a C^1 function $\overline{\varphi}$ and compact set C', such that (39) holds for an $\epsilon > 0$. Now, by the Eilenberg inequality,

$$\begin{aligned}
\int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y &= \int_{S\cap C'} J\varphi(x)\, d\mathcal{L}_n x + \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(\left(S\backslash C'\right) \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \\
&\leqslant \int_{S\cap C'} J\varphi(x)\, d\mathcal{L}_n x + \frac{\alpha_{n-k}\alpha_k}{\alpha_n} \left(\mathrm{Lip}\varphi\right)^k \mathcal{L}_n\left(S\backslash C'\right) \\
&\leqslant \int_{S\cap C'} J\varphi(x)\, d\mathcal{L}_n x + M_1\epsilon,
\end{aligned}$$

where M_1 is a constant that does not depend on ϵ. Since φ is Lipschitz, there exists another constant M_2, such that $J\varphi < M_2$, and thus,

$$\int_{S\backslash C'} J\varphi(x)\, d\mathcal{L}_n x \leqslant M_2\epsilon.$$

Then conclude by the same discussion as **part** 3 of the proof sketch Theorem T.2.1. ∎

Corollary T.2.8. *Another Sard type lemma*

Let $k \in \{1, 2, \ldots, n\}$, $E \subset \mathbb{R}^n$ be an open set, $\varphi : E \to \mathbb{R}^k$ be a C^1 function. Then for $\mathcal{L}_k$ a.e. $y \in$

$\mathbb{R}^k$,

$$\mathcal{H}_{n-k}\left(Crit\left(\varphi\right)\cap\varphi^{-1}\left(y\right)\right)=0 \tag{45}$$

and $\varphi^{-1}\left(y\right)\backslash Crit(\varphi)$ *can be locally parameterized by implicit functions.*

PROOF. Since φ is a C^1 function, $J\varphi : E\to\mathbb{R}$ is continuous. Therefore, $Crit\left(\varphi\right)=\{x\in E : J\varphi\left(x\right)\geqslant 0\}\backslash\{x\in E : J\varphi\left(x\right)>0\}$ is $\mathcal{L}_n$ measurable. Let $y\in\mathbb{R}^k$, by (34),

$$\int_{Crit(\varphi)} J\varphi\left(x\right)d\mathcal{L}_n x=\int_{\mathbb{R}^k}\mathcal{H}_{n-k}\left(Crit\left(\varphi\right)\cap\varphi^{-1}\left(y\right)\right)d\mathcal{H}_k y.$$

The integral in the left hand side of the above equality is 0, thus by the property of Lebesgue integral and the fact that $\mathcal{L}_k=\mathcal{H}_k$ in $\mathbb{R}^k$, (45) holds for $\mathcal{L}_k$ *a.e.* $y\in\mathbb{R}^k$.

For arbitrary $x\in\varphi^{-1}\left(y\right)\backslash Crit(\varphi)$ where y satisfies (45), $J\varphi\left(x\right)>0$. Note that $\varphi'\left(x\right)$ is row full rank. Therefore, the implicit function theorem can be applied in a neighborhood of x. Specifically, pick k linear independent columns of $\varphi'\left(x\right)$, without loss of generality, assuming that the first k columns $\left\{\frac{\partial}{\partial x_1}\varphi\left(x\right),\frac{\partial}{\partial x_2}\varphi\left(x\right),\ldots,\frac{\partial}{\partial x_k}\varphi\left(x\right)\right\}$ are linear independent. Let U be a neighborhood of x, define

$$\Psi\left(x'_1,\ldots,x'_k,x'_{k+1},\ldots,x'_n\right)=\varphi\left(x'_1,\ldots,x'_k,x'_{k+1},\ldots,x'_n\right)-y,$$

where $y=\varphi\left(x\right)$, then $\nabla_{x_1,x_2,\ldots,x_k}\Psi\left(x_1,\ldots,x_k,x_{k+1},\ldots,x_n\right)$ is full rank, and thus there exists a C^1 implicit function g in a open subset $B_{x_1,\ldots,x_k}\times B_{x_{k+1},\ldots,x_n}\subset U$, where

$$\begin{aligned}B_{x_1,\ldots,x_k}&=\left\{\left(x'_1,\ldots,x'_k\right)\in\mathbb{R}^k : \left|\left(x'_1,\ldots,x'_k\right)-\left(x_1,\ldots,x_k\right)\right|<\alpha\right\},\\ B_{x_{k+1},\ldots,x_n}&=\left\{\left(x'_{k+1},\ldots,x'_n\right)\in\mathbb{R}^{n-k} : \left|\left(x'_{k+1},\ldots,x'_n\right)-\left(x_{k+1},\ldots,x_n\right)\right|<\beta\right\},\end{aligned}$$

such that $\Psi\left(x'_1,\ldots,x'_k,x'_{k+1},\ldots,x'_n\right)=0\Leftrightarrow\left(x'_1,\ldots,x'_k\right)=g\left(x'_{k+1},\ldots,x'_n\right)$ for all $x'\in B_{x_1,\ldots,x_k}\times B_{x_{k+1},\ldots,x_n}$. Now, define

$$\psi\left(x'_{k+1},\ldots,x'_n\right)=\left(g\left(x'_{k+1},\ldots,x'_n\right),x'_{k+1},\ldots,x'_n\right)^T,$$

then ψ is a C^1 diffeomorphism and thus a homeomorphism. ∎

Addendum References

Attouch, Hedy, Giuseppe Buttazzo, and Gérard Michaille, *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*, SIAM, 2014.

Clarke, Francis, "On the inverse function theorem," *Pacific Journal of Mathematics*, 1976, *64* (1), 97–102.

de Pascale, Luigi, "The Morse–Sard theorem in Sobolev spaces," *Indiana University mathematics journal*, 2001, pp. 1371–1386.

Evans, Lawrence C and Ronald F Garzepy, *Measure theory and fine properties of functions, revised edition*, Chapman and Hall/CRC, 2015.

Falconer, Kenneth J, *The geometry of fractal sets*, Vol. 85 of *Cambridge Tracts in Mathematics*, Cambridge University Press, 1986.

Federer, Herbert, *Geometric Measure Theory*, Springer, 1969.

Figalli, Alessio, "A simple proof of the Morse-Sard theorem in Sobolev spaces," *Proceedings of the American Mathematical Society*, 2008, *136* (10), 3675–3681.

Giné, Evarist and Richard Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge university press, 2021.

Grigoryan, Alexander, *Heat kernel and analysis on manifolds*, Vol. 47 of *AMS/IP Studies in Advanced Mathematics*, American Mathematical Soc., 2009.

Hagood, John W and Brian S Thomson, "Recovering a function from a Dini derivative," *The American Mathematical Monthly*, 2006, *113* (1), 34–46.

Hiriart-Urruty, Jean-Baptiste, "Tangent cones, generalized gradients and mathematical programming in Banach spaces," *Mathematics of Operations Research*, 1979, *4* (1), 79–97.

Jerrard, Robert L., "Lecture note for Geometric Measure Theory," 2013. University of Toronto, http://www.math.toronto.edu/rjerrard/1501/gmt.html.

Lu, Xuguang, "Lecture Notes for Mathematical Analysis," September 2019. Tsinghua University.

Munkres, James R, *Topology*, Prentice Hall, Upper Saddle River, NJ, 2000.

Resnick, Sidney I, *Extreme values, regular variation, and point processes*, Vol. 4 of *Springer Series in Operations Research and Financial Engineering*, Springer Science & Business Media, 1987.

Saks, Stanisław, *Theory of the Integral*, G.E. Stechert, 1937.

Smale, S, "An Infinite Dimensional Version of Sard's Theorem," *American Journal of Mathematics*, 1965, *87* (4), 861–866.