Robust Causal Domain Adaptation in a Simple Diagnostic Setting

Thijs van Ommen T.VANOMMEN@UU.NL
Information and Computing Sciences, Utrecht University, The Netherlands

Abstract

Causal domain adaptation approaches aim to find statistical relations in a source domain, that will still hold in a target domain, using the assumption that a common causal graph underlies both domains. For many such problems, the available information is insufficient to uniquely identify the target domain distribution, and we find a set of distributions instead. We propose to use a worst-case approach, picking an action that performs well against all distributions in this set. In this paper, we study a specific diagnostic instance of this problem, and find a sufficient and necessary condition that characterizes the worst-case distribution in the target domain. We find that the Brier and logarithmic scores lead to different distributions, and consequently to different recommendations for the decision maker.

Keywords: domain adaptation, causal graph, minimax decision making, robust Bayes, scoring function

1. Introduction

Many approaches in statistics and machine learning rely on the assumption that the training data are drawn from the same distribution as the test data. *Domain adaptation* considers situations in which this assumption is violated: we have access to training data from a *source domain* and want to make decisions in a *target domain*, but the data in the two domains may come from different distributions.

Our interest is in a *proactive* approach, which uses no data at all from the target domain during training. Of course, this is only possible if the source and target domain distributions are in some way related. In *causal* domain adaptation, this relation takes the form of a common *causal graph* [5, 10].

Consider the following motivating example, taken from Subbaswamy et al. [12]. Suppose we want to diagnose lung cancer (X). Lung cancer causes chest pain (Z), while aspirin (Y) relieves chest pain. Further, we know that people who smoke have an increased risk of lung cancer as well as of heart disease, and due to this risk they may be prescribed aspirin. (Our data does not include whether a person smokes.) Now we gather data from one hospital and use it to train a model, which is used successfully for the diagnosis of lung cancer. However, when used at a different hospital, this same model may give worse performance. This can happen because the different hospitals may follow different policies for the prescription of aspirin, so that the data distributions are not the same. Yet, it may be reasonable to expect that the probability of lung cancer, and the effects of lung cancer and aspirin on chest pain, do not vary between hospitals. A causal domain adaptation method would use these aspects of the source domain to make predictions in the target domain.

Subbaswamy et al. [11, Section 4.2] (an earlier version of [12]) suggest addressing this problem using a robust Bayes approach [1]. In this paper, we investigate the underlying mathematical optimization problem for a specific graph, under different probability distributions and loss functions. The graph represents a simple diagnostic setting with only three variables, but as we will see, the optimization problem has some interesting properties.

Following Grünwald and Dawid [2], we model this situation as a zero-sum game between two players, the decision maker and the (imaginary) adversary. In this game (described in more detail in Section 2), the adversary chooses a distribution P for the target domain, constrained by the requirement to be consistent with the invariant parts of the source domain. Then $X = x$, $Y = y$ and $Z = z$ are drawn at random according to P. Seeing only y and z, the decision maker chooses an action $A_{y,z}$ (e.g. treat vs. do not treat). Then x is revealed and the action is evaluated according to a loss function L. The adversary and decision maker aim to respectively maximize and minimize the expected loss

$$\sum_{x,y,z} P(x,y,z) L(x, A_{y,z}) . \qquad (1)$$

If the decision maker plays this game optimally, they have the best guarantees on the expected loss; other strategies are less *robust* because for some distribution P the adversary might pick, they obtain a larger expected loss. Van Ommen et al. [14, 13] recently applied the same worst-case approach to the Monty Hall problem [9], and the present paper can be seen as a generalization of their work to the setting of causal domain adaptation.

Transportability (see e.g. Pearl and Bareinboim [6]) can be seen as an approach to causal domain adaptation. This paper complements that line of work as follows: If it is found that the relations of interest are not transportable from the source to the target domain, our approach can find the safest action to take.

Two other approaches to causal domain adaptation have been proposed recently [3, 8]. Unlike [12] and the present paper, these approaches only search for feature sets S such

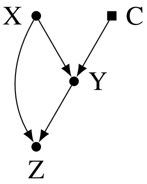

Figure 1: Directed acyclic graph for the Bayesian network that relates the random variables X, Y and Z, and the context variable C.

that a regression model trained on these features will be invariant between the source and target domains. We will see below that this is a strong limitation for some causal graphs, in particular for the one we consider. Their setting is more challenging in a different respect: they do not take the causal graph to be known, and must learn about it from the available data.

2. Problem Definition

In this paper, we will consider a specific setting with four discrete variables, related by a Bayesian network whose structure is depicted in Figure 1.[1] This network structure represents that the distribution of discrete variables X, Y and Z given C can be factored as

$$P(X,Y,Z \mid C) = P(X)P(Y \mid X,C)P(Z \mid X,Y). \tag{2}$$

Here C is a context variable [4] (also called *selection variable* by Subbaswamy et al. [12]; the two concepts are different in general but their meanings coincide for this graph). Data points in the source domain have $C=0$, and those in the target domain have $C=1$. Because Y is a child of C in the graph, we say Y is *mutable*: the distribution $P(Y \mid X, C=1)$ in the target domain may be completely different from the distibution $P(Y \mid X, C=0)$ we saw in the source domain. The two other factors in Equation (2) are the same regardless of C, and so these distributions can be learned using data from the source domain.

We write $\mathcal{P}$ for the set of all joint distributions on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ that can be obtained by varying $P(Y \mid X, C=1)$, while leaving the other factors in Equation (2) the same. Thus $\mathcal{P}$ represents both our aleatory uncertainty about the random outcomes, and our epistemic uncertainty about the mechanism that produces Y.

Our main result (Theorem 1) applies to arbitrary P, but for concreteness, consider the following instantiation of the problem (that we will return to in Section 3). In this example, all variables are binary, and we observe in the source domain that

$$\begin{aligned} P(X=1) &= \tfrac{1}{2}; \\ P(Z=1 \mid Y,X) &= \begin{cases} \frac{1}{2} & \text{if } Y=X; \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \tag{3}$$

Table 1: General form of $P(X,Y,Z \mid C=1)$ consistent with Equation (3).

$P(X,Y,Z \mid C=1)$	$X=0$	$X=1$
$Y=0, Z=0$	$\alpha_0/4$	0
$Y=0, Z=1$	$\alpha_0/4$	$\frac{1}{2}-\alpha_1/2$
$Y=1, Z=0$	0	$\alpha_1/4$
$Y=1, Z=1$	$\frac{1}{2}-\alpha_0/2$	$\alpha_1/4$

Introducing parameters α_0, α_1 to define the unknown $P(Y=x \mid X=x, C=1) = \alpha_x$, we can compute expressions for any joint distribution $P \in \mathcal{P}$ as given in Table 1.

This graph entails that X is independent of C given the empty set, but not entail independence given any other set (this can be verified by checking d-separation [5]). This implies that if we want to learn a distribution $P(X \mid S)$ (for some $S \subseteq \{Y,Z\}$) from the source domain data, the only safe choice is $S=\varnothing$: for any other S, the distribution in the target domain may be very different from the distribution we learned. The choice $S=\varnothing$ would result in a method for predicting X that ignores the information in Y and Z, which in many circumstances can not be expected to result in good predictions (though in practice, this comparison may fall out differently if it is hard to learn $P(Z \mid X,Y)$ from the data).

From now on, we will only be interested in the possible distributions given $C=1$, so we omit C from the notation.

2.1. Loss Functions and Entropy

To make precise what the 'best' decision is, we need to pick a *loss function* $L: \mathcal{X} \times \mathcal{A} \to [0,\infty]$, where $\mathcal{A}$ is the decision maker's *action space*. We will focus on two loss functions in our examples: Brier loss and logarithmic loss. Both loss functions are defined on the same action space $\mathcal{A} = \Delta_{\mathcal{X}}$, the set of all probability distributions over $\mathcal{X}$. Such loss functions (also called *scoring rules*) can be used to elicit the decision maker's belief about $\mathcal{X}$. They are defined as

$$\begin{aligned} L_{\text{Brier}}(x,Q) &= \sum_{x' \in \mathcal{X}} \left(\mathbf{1}_{x'=x} - Q(x')\right)^2; \\ L_{\log}(x,Q) &= -\log Q(x). \end{aligned}$$

[1] The graph in Figure 1 differs from the one in Subbaswamy et al. [12, Figure 1(a)], in that it has a directed edge between X and Y, while in the lung cancer example, the corresponding variables (lung cancer T and aspirin A respectively) share a common confounder. This difference is not relevant in the present paper, so we opted for the simpler graph without latent variables.

For a given loss function L and a distribution $P \in \Delta_{\mathcal{X}}$, Grünwald and Dawid [2] define the generalized entropy as the expected loss when the decision maker, knowing P, chooses Q optimally:

$$H_L(P) := \inf_{Q \in \Delta_{\mathcal{X}}} \sum_x P(x) L(x, Q) = \inf_{Q \in \Delta_{\mathcal{X}}} \mathbf{E}_{X \sim P} L(X, Q). \tag{4}$$

In our setting, we take a further expectation over the two additional variables Y and Z. If the adversary chooses distribution P and the decision maker always chooses Q optimally in response, the game's value is given by

$$\sum_{\substack{(y,z) \in \mathcal{Y} \times \mathcal{Z}: \\ P(y,z) > 0}} P(y,z) H_L(P(\cdot \mid y,z)). \tag{5}$$

Equation (5) considers the game from the adversary's side. In [2, 14], the game is also examined from the decision maker's side: how can the decision maker minimize the worst-case expected loss against any distribution P the adversary might choose? The theoretical results in those papers give conditions under which the game has a Nash equilibrium, in which case knowing the adversary's optimal strategy largely solves the problem of finding an optimal strategy for the decision maker. Finding the precise conditions under which such results hold is beyond the scope of this paper, but we will see that once Equation (5) is solved for a problem instance of interest, finding the decision maker's optimal strategy becomes much easier.

We conclude this section by presenting an intuition for why maximizing Equation (5) is difficult. For any *symmetric* loss function [14] (such as Brier and logarithmic loss), the generalized entropy of a Bernoulli random variable is maximized at probability $\frac{1}{2}$. In our example, $P(X = 1 \mid Y = 0, Z = 1) = P(X = 1 \mid Y = 1, Z = 1) = \frac{1}{2}$ can be simultaneously achieved by setting $\alpha_0 = \alpha_1 = \frac{2}{3}$. (It is clear that the entropy of X given Y and $Z = 0$ is fixed, regardless of the choice of distribution.) However, by reducing α_0 and α_1, the probability that the 'easy-to-predict' rows occur is reduced, at the expense of a smaller entropy for the 'hard-to-predict' rows. The trade-off makes the problem nontrivial.

3. Results

We now present our main result, which shows that under very general conditions, a distribution exists that is optimal for the adversary. Additionally, the theorem gives a necessary and sufficient condition for recognizing such optimal distributions.

This theorem is adapted from Van Ommen et al. [14, Theorem 3]. Its proof can be found in Appendix A.

Theorem 1 (Existence and characterization of P^*) *For H_L finite and continuous, a $P \in \mathcal{P}$ maximizing Equation (5) exists, and P^* is such a maximizer if and only if there exists a $\lambda^* \in \mathbf{R}^{\mathcal{X}}$ such that*

(i) for every $y \in \mathcal{Y}$ with $P^*(y) > 0$,

$$\sum_{\substack{z \in \mathcal{Z}: \\ P^*(y,z) > 0}} P^*(z \mid y) H_L(P^*(\cdot \mid y,z)) = \sum_x P^*(x \mid y) \lambda^*_x; \tag{6}$$

(ii) for every $y \in \mathcal{Y}$, for all $P' \in \Delta_{\mathcal{X}}$, let $P'(x,z \mid y) := P'(x) P(z \mid x,y)$; then

$$\sum_{\substack{z \in \mathcal{Z}: \\ P'(z \mid y) > 0}} P'(z \mid y) H_L(P'(\cdot \mid y,z)) \leq \sum_x P'(x \mid y) \lambda^*_x. \tag{7}$$

The conditions in the theorem can be understood geometrically: the expression on the left-hand side of Equation (7) describes a concave function f mapping $P' \in \Delta_{\mathcal{X}}$ to $\mathbf{R}$, while the right-hand side describes a linear function. By Equation (7), the linear function must be nowhere below f, and for y with $P^*(y) > 0$, by Equation (6), they must touch at $P'(X \mid y) = P^*(X \mid y)$. If this point is in the interior of $\Delta_{\mathcal{X}}$ and f is differentiable there, then the linear function and thus λ^* are completely determined. Specifically, if additionally L is strictly proper, the infimum Q in Equation (4) is attained by $Q = P^*(X \mid y,z)$, and we find that

$$\lambda^*_x = \sum_{\substack{z \in \mathcal{Z}: \\ P^*(y,z) > 0}} P^*(z \mid y) L(x, P^*(\cdot \mid y,z)) \tag{8}$$

(compare [14, Theorem 9]). If f is not differentiable or L is not (strictly) proper, as long as the game has a Nash equilibrium, Q can still be determined from P^* and λ^*, though it is no longer a matter of simply reading it off (see [14, Theorem 7]).

We illustrate Theorem 1 by using it to find analytic expressions for P^* in the case that $P(X)$ and $P(Z \mid X,Y)$ are given by Equation (3), for Brier and logarithmic loss. These optimal distributions are displayed in Tables 2 and 3 respectively.

For both examples, combining Theorem 1 and Equation (8) tells us that P^* is optimal if and only if for all $x \in \mathcal{X}$, the right-hand side of Equation (8) is equal for $y = 0$ compared to $y = 1$. In words, for an optimally playing decision maker who knows P^*, the expected loss given $X = 0, Y = 0$ must equal that given $X = 0, Y = 1$, and similarly for $X = 1$.

For Brier loss, doing so gives a system of polynomial equations whose solution is displayed in Table 2. The loss given $X = 0, Y = 1$ (then $Z = 1$ with probability 1) equals $6 - 4\sqrt{2}$; for $X = 0, Y = 0, Z = 1$ it is twice that, while for $X = 0, Y = 0, Z = 0$ it is zero, and these cases have the same probability so that the expectations over Z are equal.

For logarithmic loss, the terms from the sum in Equation (8) will be of the form $-p_1 \log p_2$. We can rewrite each equation so the sum becomes a product with factors $p_2^{p_1}$.

Table 2: $P^*(X,Y,Z)$ optimal under Brier loss ($\frac{1}{2}-\frac{1}{4}\sqrt{2}\approx 0.146$, $-\frac{1}{2}+\frac{1}{2}\sqrt{2}\approx 0.207$)

$P^*(X,Y,Z)$	$X=0$	$X=1$
$Y=0,Z=0$	$\frac{1}{2}-\frac{1}{4}\sqrt{2}$	0
$Y=0,Z=1$	$\frac{1}{2}-\frac{1}{4}\sqrt{2}$	$-\frac{1}{2}+\frac{1}{2}\sqrt{2}$
$Y=1,Z=0$	0	$\frac{1}{2}-\frac{1}{4}\sqrt{2}$
$Y=1,Z=1$	$-\frac{1}{2}+\frac{1}{2}\sqrt{2}$	$\frac{1}{2}-\frac{1}{4}\sqrt{2}$

Table 3: $P^*(X,Y,Z)$ optimal under logarithmic loss ($\frac{1}{4}-\frac{1}{20}\sqrt{5}\approx 0.138$, $\frac{1}{10}\sqrt{5}\approx 0.224$)

$P(X,Y,Z)^*$	$X=0$	$X=1$
$Y=0,Z=0$	$\frac{1}{4}-\frac{1}{20}\sqrt{5}$	0
$Y=0,Z=1$	$\frac{1}{4}-\frac{1}{20}\sqrt{5}$	$\frac{1}{10}\sqrt{5}$
$Y=1,Z=0$	0	$\frac{1}{4}-\frac{1}{20}\sqrt{5}$
$Y=1,Z=1$	$\frac{1}{10}\sqrt{5}$	$\frac{1}{4}-\frac{1}{20}\sqrt{5}$

For our example, the exponents are 1 or $\frac{1}{2}$, so the equation can then be re-expressed as a quadratic polynomial, and the solution is displayed in Table 3. For arbitrary rational p_2, the degree of the polynomial could be much higher; if p_2 is not rational, then the equations in the system would not even be polynomials.

Knowing P^* is usually enough to also determine an optimal strategy for the decision maker. In particular, for strictly proper loss functions such as Brier and logarithmic loss, the optimal response to any y,z with $P^*(y,z)>0$ will be $P^*(\cdot\,|\,y,z)$. This allows us to compute the decision maker's actions from the expressions in Tables 2 and 3.

We note that the optimal P^* (and in particular $P^*(\cdot\,|\,y,z)$, the relevant part for the decision maker) is different for the two loss functions we considered, so it is important to choose an appropriate loss function before taking a decision in this framework. We also find that finding P^* involved solving systems of polynomial equations (and could even require more general equations), so that Tables 2 and 3 include many irrational numbers, even though in our case, all inputs into the problem from Equation (3) were rational.

In practice, an analytic solution for this problem might have little additional value over an accurate numerical solution. Because finding P^* can be formulated as a concave optimization problem (see the proof of Theorem 1), many numerical optimization tools are available that can efficienty find such a numerical solution.

4. Conclusion

We studied a special case of causal domain adaptation, for a specific graph representing a diagnostic prediction problem. For this problem, we saw how it can be solved in a robust Bayesian way. We observed that the optimal solution may depend on the loss function, even when comparing two strictly proper scoring rules.

Subbaswamy et al. [12] address the question of how to determine invariant parts of P for arbitrary causal graphs, with possibly multiple mutable variables, and even in the case where the graph has latent confounders, or where the target variable (X in our setting) is itself mutable. For such graphs, the optimization problem we discussed in the present paper may take a quite different form. Extending our results along those lines is an important direction for future work.

We did not formally address the question of how an optimal strategy for the decision maker can be found, and did not give conditions under which the two players' strategies form a Nash equilibrium. This and other questions were addressed in a similar setting by Van Ommen et al. [14, Lemma 4 to Theorem 9], and it is likely that they can be adapted to the causal domain adaptation setting, in the same way that we adapted their Theorem 3 in the present paper.

Appendix A. Proof of Theorem 1

Proof of Theorem 1. We rewrite the problem of maximizing Equation (5) to a convex optimization problem where the solution variable μ comes from $\mathbf{R}_{\geq 0}^{\mathcal{X}\times\mathcal{Y}}$ (so μ does not need to sum to one). For convenient notation, we extend μ to $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ by defining $\mu(x,y,z):=\mu(x,y)P(z\,|\,x,y)$. We still use notation for marginal and conditional distributions, e.g. $\mu(y,z):=\sum_x\mu(x,y,z)$ and $\mu(x\,|\,y,z):=\mu(x,y,z)/\mu(y,z)$. Note that the latter defines a probability distribution (i.e. it sums to one) for any y,z with $\mu(y,z)>0$, because any scale factor cancels out.

The following function extends the adversary's objective function given in Equation (5) (the expected generalized entropy of $P\in\mathcal{P}$) to the domain $\mathbf{R}_{\geq 0}^{\mathcal{X}\times\mathcal{Y}}$:

$$\begin{aligned} f_0(\mu) &:= \inf_{(Q_{y,z})_{y\in\mathcal{Y},z\in\mathcal{Z}}} \sum_{\substack{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}:\\ \mu(y,z)>0}} \mu(x,y,z)L(x,Q_{y,z}) \\ &= \sum_{\substack{(y,z)\in\mathcal{Y}\times\mathcal{Z}:\\ \mu(y,z)>0}} \inf_{Q_{y,z}\in\Delta_{\mathcal{X}}} \sum_{x\in\mathcal{X}} \mu(x,y,z)L(x,Q_{y,z}) \\ &= \sum_{\substack{(y,z)\in\mathcal{Y}\times\mathcal{Z}:\\ \mu(y,z)>0}} \mu(y,z)H_L(\mu(\cdot\,|\,y,z)). \end{aligned}$$

Using this concave function (infimum of linear functions), the convex optimization problem is given by

$$\begin{aligned} &\text{maximize} && f_0(\mu) \\ &\text{subject to} && \sum_{y\in\mathcal{Y}} \mu(x,y) = P(x) \qquad \text{for all } x \in \mathcal{X}, \end{aligned}$$

with $\mu \in \mathbf{R}_{\geq 0}^{\mathcal{X}\times\mathcal{Y}}$. Write $\mathcal{P}_\mu \subset \mathbf{R}_{\geq 0}^{\mathcal{X}\times\mathcal{Y}}$ for the set of μ's that satisfy this constraint. We see that $\mu \in \mathcal{P}_\mu$ if and only if $\mu(X,Y,Z) \in \mathcal{P}$, because $\mu(z \,|\, x,y) = P(z \,|\, x,y)$ by definition of $\mu(x,y,z)$. In this sense, $\mathcal{P}_\mu$ is isomorphic to $\mathcal{P}$.

Rockafellar [7, Theorem 27.3] gives conditions under which a convex minimization problem has a solution attaining the infimum. These are satisfied by $\mathcal{P}_\mu$ and $-f_0$: $\mathcal{P}_\mu$ is nonempty, closed, convex, and bounded (thus has no direction of recession), and $-f_0$ is convex, finite for all $\mu \in \mathcal{P}_\mu$ (thus proper) by finiteness of H_L, and lower semi-continuous (thus closed) by continuity of H_L.

By Rockafellar [7, Corollary 28.2.2], a KT-vector λ^* exists, so that for the remaining claims of the theorem, it suffices to show that $P^* = \mu(X,Y,Z)$ (with $\mu \in \mathcal{P}_\mu$) is worst-case optimal and λ^* is a KT-vector if and only if the conditions on (P^*, λ^*) given in the theorem hold.

By Rockafellar [7, Theorem 28.3], $\mu \in \mathbf{R}_{\geq 0}^{\mathcal{X}\times\mathcal{Y}}$ is an optimal solution to the optimization problem and $\lambda^* \in \mathbf{R}^{\mathcal{X}}$ is a KT-vector if and only if $\mu \in \mathcal{P}_\mu$ and at μ, the zero vector is a supergradient to

$$f_0(\mu) - \sum_{x\in\mathcal{X}} \lambda_x^* \left(\sum_{y\in\mathcal{Y}} \mu(x,y) - P(x) \right). \tag{9}$$

The term being subtracted is linear, with gradient $\bar{\lambda} \in \mathbf{R}^{\mathcal{X}\times\mathcal{Y}}$ given by

$$\bar{\lambda}_{x,y} := \frac{\partial}{\partial \mu(x,y)} \sum_{x\in\mathcal{X}} \lambda_x^* \left(\sum_{y\in\mathcal{Y}} \mu(x,y) - P(x) \right) = \lambda_x^*. \tag{10}$$

By Rockafellar [7, Theorem 23.8], 0 is a supergradient to Equation (9) if and only if $\bar{\lambda}$ is a supergradient to f_0 at μ.

For any μ that is not everywhere zero, we have for all $c \geq 0$ that $f_0(c\mu) = cf_0(\mu)$, so that a supporting hyperplane to f_0 at any $\mu \in \mathcal{P}_\mu$ must go through the origin. Then the supporting hyperplane with gradient $\bar{\lambda}$ is defined by the linear function $\nu \mapsto \sum_{x,y} \nu(x,y)\bar{\lambda}_{x,y}$.

If $\sum_{x,y} \nu(x,y)\bar{\lambda}_{x,y}$ defines a supporting hyperplane to f_0 at μ (recall that $P^* = \mu(X,Y,Z)$), then

1. at every $y \in \mathcal{Y}$ with $P^*(y) > 0$, it is a supporting hyperplane to $\sum_{z\in\mathcal{Z}:P^*(y,z)>0} P^*(z \,|\, y) H_L(P^*(\cdot \,|\, y,z))$ at $P^*(\cdot \,|\, y)$, and

2. for every y, $\sum_{z\in\mathcal{Z}:P'(y,z)>0} P'(z \,|\, y) H_L(P'(\cdot \,|\, y,z)) \leq \sum_x P'(x)\bar{\lambda}_{x,y}$ for all $P' \in \Delta_{\mathcal{X}}$.

The converse also holds: we have for all $y \in \mathcal{Y}$ and $P' \in \Delta_{\mathcal{X}}$ that $\sum_{z\in\mathcal{Z}:P'(y,z)>0} P'(z \,|\, y) H_L(P'(\cdot \,|\, y,z)) \leq \sum_{x\in y} P'\bar{\lambda}_{x,y}$, with equality if $P^*(y) > 0$ and $P' = P^*(\cdot \,|\, y)$; taking the convex combination with coefficients $P^*(y)$ shows that the hyperplane defined by $\sum_{x,y} P(x,y)\bar{\lambda}_{x,y}$ is nowhere below f_0 and touches it at $\nu = P^*$.

For $\bar{\lambda}$ of the required form given by Equation (10), this is in turn equivalent to the characterization given in the statement of the theorem. ■

Acknowledgments

The author would like to thank Adarsh Subbaswamy, Joris Mooij, Patrick Forré and Tineke Blom for the interesting discussions about the problem investigated in this paper, and the anonymous reviewers for their useful comments and suggestions.

References

[1] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition, 1985.

[2] Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32:1367–1433, 2004.

[3] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS 2018)*, 2018.

[4] Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint Causal Inference from multiple contexts. *arXiv.org preprint*, https://arxiv.org/abs/1611.10351v3 [cs.LG], March 2018.

[5] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[6] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 247–254, 2011.

[7] R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, New Jersey, 1970.

[8] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[9] S. Selvin. A problem in probability. *The American Statistician*, 29:67, 1975. Letter to the editor.

[10] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, second edition, 2000.

[11] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Learning predictive models that transport. *arXiv preprint arXiv:1812.04597v1*, 2018.

[12] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3118–3127, 2019.

[13] Thijs van Ommen. Computing minimax decisions with incomplete observations. In *Proceedings of the Tenth International Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 358–369, 2017.

[14] Thijs van Ommen, Wouter M. Koolen, Thijs E. Feenstra, and Peter D. Grünwald. Robust probability updating. *International Journal of Approximate Reasoning*, 74:30–57, 2016.