Intriguing Parameters of Structural Causal Models
Matej Zečević, Devendra Singh Dhami, Kristian Kersting
Computer Science Department, TU Darmstadt
{matej.zecevic, devendra.dhami, kersting}@cs.tu-darmstadt.de

Abstract
In recent years there has been a lot of focus on adversarial attacks, especially on deep neural networks. Here, we argue that they are more general in nature and can easily affect a larger class of models, e.g., any differentiable perturbed optimizers. We further show that such attacks can be determined by the hidden confounders in a domain, thus drawing a novel connection between such attacks and causality. Establishing this causal perspective is characterized by the influence of the structural causal model’s data generating process on the subsequent optimization thereby exhibiting intriguing parameters of the former. We reveal the existence of such parameters for three combinatorial optimization problems, namely linear assignment, shortest path and a real world problem of energy systems. Our empirical examination also unveils worrisome consequences of these attacks on differentiable perturbed optimizers thereby highlighting the criticality of our findings.

Introduction
Adversarial attacks have gained a lot of traction in recent years (Brendel, Rauber, and Bethge 2018; Ilyas et al. 2018; Guo et al. 2019) as there has been a lot of focus on safety and robustness of machine learning (ML) systems. An interesting observation, though, is that deep neural networks or rather over-parameterized models are the center of attention for most of such adversarial attacks (Zügner, Akbarnejad, and Günnemann 2018; Akhtar and Mian 2018; Chen et al. 2018). We argue that this view is too narrow—adversarial attacks are more general in nature and actually depend on the problem setup. Any mathematical program (MP) solver can suffer from such attacks and, in this work, we focus on differentiable perturbed optimizers (Papandreou and Yuille 2011; Berthet et al. 2020; Gumbel 1954). While, e.g., Bach (Bach 2013) showed how to learn combinatorial problems, adversarial attacks stand as a problem of learning in general. There has been previous works where MPs such as linear programs (LPs) and mixed integer programs (Wu, Wang, and Yu 2020; Tjeng, Xiao, and Tedrake 2019) have been used to compute adversarial attacks but not where such optimization modules themselves have been confronted with the attacks. In fact, and also due to the recent interest in tightly integrating MPs and deep learning (Paulus et al. 2021), this extension of adversarial attacks beyond deep networks already significantly advances our understanding of adversarial attacks i.e., it is not just expressiveness that leads to uninterpretable solutions with counter-intuitive properties.

We draw a novel theoretical connection between adversarial attacks and causality transitively via an intermediate connection between the latter and MPs. Specifically, we prove that hidden confounders can be exploited for the construction of an extended notion of adversarial attacks against perturbed optimizers, which we term Hidden Measure Attacks. To this end, we consider Structural Causal Models (SCM) that are capable of emitting graphical structures, statistical and causal quantities placing them at the center of causal inference and its application to ML in marketing (Hair Jr and Sarstedt 2021), healthcare (Bica, Alaa, and Van Der Schaar 2020) and education (Hoiles and Schaar 2016). Establishing this causal perspective is characterized by the influence of the SCM’s data generating process on the subsequent optimization, thereby exhibiting intriguing parameters of the former exploitable for adversarial attacks.

Overall, we make a number of key contributions: (1) We show that mathematical programs themselves are susceptible to adversarial attacks by considering differentiable perturbed optimizers; (2) We derive a novel theoretical connection between causality, MPs and adversarial attacks, proving that attacks can be constructed via hidden confounders; (3) We study and discuss two classical LP families and one real world applied optimization problem. We make our code repository publicly available.

Background and Related Work
Let us briefly review the background on adversarial attacks as defined in their original setting of classification, then the formalism of mathematical programs alongside two relevant problem classes and finally causal models and the concept of hidden confounders for developing our new unified perspective on adversarial attacks.

Notation. We denote indices by lower-case letters, functions by the general form $g(\cdot)$, scalars or random variables interchangeably by upper-case letters, vectors, matrices and...
Adversarial Attacks. By using a simple optimization procedure, Szegedy et al. (2014) were able to find adversarial examples, which they defined to be imperceptibly perturbed input images such that these new images were no longer classified correctly by the predictive neural model.

Mathematically, they investigated perturbations of the form

$$\mathbf{w} \approx \mathbf{w} \text{ (left side) and cost-wise w.r.t. their optimal solution } \hat{\mathbf{w}}^\top \mathbf{x}(\hat{\mathbf{w}}) \approx \mathbf{w}^\top \mathbf{x}(\mathbf{w}) \text{ i.e., health-wise people in higher need of vaccination are still guaranteed an appointment. However, w.r.t. some hidden measure (Wealth) the adversarial solution } \hat{x}^\top \mathbf{(w)} \text{ drastically deviates i.e., the distribution of vaccines is being skewed towards people of higher wealth. (Best viewed in Color.)}$$

where the inequality represents a possibly strongly significant divergence from the expected semantic meaning of the class to be predicted e.g. instead of confidently predicting an animal like a dog the classifier might now suggest a transportation machine like a plane although the human experimenter’s intuition about the class remains invariant to the performed perturbations. Naturally, the susceptibility of neural networks to such grave mistakes led to a significant increase in research interest regarding robustness to adversarial examples) in neural function approximators evoking the narrative of “attacks” and subsequent “defences” on the inspected classification modules as commonly found in cyber-security (Handa, Sharma, and Shukla 2019).

Mathematical Programming. Selecting the best candidate from some given set with regard to some criterion is a general description of mathematical programming (or optimization), which arguably lies at the core of machine learning and many applications in science and engineering. Classification, e.g., can be considered as a special instance of mathematical programming. An important optimization family are LPs that are concerned with the optimization of an objective function and constraints that are linear in the respective optimization variables. LPs are being applied widely in the real world, e.g., energy systems (Schaber, Steinke, and Hamacher 2012). More formally, an LP $L$ is a specified as follows

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} \mathbf{w}^\top \mathbf{x} \quad \text{subject to } A \mathbf{x} \leq \mathbf{b} \text{ and } \mathbf{x} \geq \mathbf{0},$$

where $\mathbf{w} \in \mathbb{R}^n$ is the weight/cost vector, $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$ are the constraint coefficients and $\mathbf{x}^* \in \mathcal{P} \subseteq \mathbb{R}^n$ is the optimizing solution within the solution polytope $\mathcal{P}$ i.e., the best choice of projecting the decision vector onto the cost vector while adhering to the given constraints.

Table I presents two classical problems that can be expressed as linear programs: the Linear Assignment Problem (LA) and the Shortest Path Problem (SP).

Both problems formulate the optimization variable $\mathbf{x} \in \mathbb{R}^n$ with either $n = |A \times B|$ or $n = |E|$ to be a selector, that is, being worker-job matches for the LA problem

Table I. Present two classical problems that can be expressed as linear programs: the Linear Assignment Problem (LA) and the Shortest Path Problem (SP).

| Problem            | Objective Function | Constraints |
|--------------------|--------------------|-------------|
| Linear Assignment  | $\min \sum_{i,j} c_{ij}x_{ij}$ | $\sum_j x_{ij} = 1 \text{ for all } i$, $\sum_i x_{ij} = 1 \text{ for all } j$, $x_{ij} \geq 0$ |
| Shortest Path      | $\min \sum_{e \in E} d_e x_e$ | $x_e \geq 0$ for all $e$ |

Figure 1: Lead Example: Adversarial Attack on Perturbed Optimizer. A real world inspired example of an attack that causes prioritization bias for the wealthy to become apparent. The adversarial LP parameter $\hat{\mathbf{w}}$ is close to the original both value-wise $\mathbf{w} \approx \mathbf{w}$ (left side) and cost-wise w.r.t. their optimal solution $\hat{\mathbf{w}}^\top \mathbf{x}(\hat{\mathbf{w}}) \approx \mathbf{w}^\top \mathbf{x}(\mathbf{w})$ i.e., health-wise people in higher need of vaccination are still guaranteed an appointment. However, w.r.t. some hidden measure (Wealth) the adversarial solution $\hat{x}^\top \mathbf{(w)}$ drastically deviates i.e., the distribution of vaccines is being skewed towards people of higher wealth. (Best viewed in Color.)
\( \forall i \in A : \sum_{j \in B} x_{ij} = 1 \)
\( \forall j \in B : \sum_{i \in A} x_{ij} = 1 \)
\( x_{ij} \in [0, 1] \)
\( \sum_{(i,j) \in E} x_{ij} - \sum_{(j,i) \in E} x_{ji} = \begin{cases} 
1 & \text{if } i = s \\
-1 & \text{if } i = t \\
0 & \text{else}
\end{cases} \)
\( x_{ij} \in [0, 1] \)

Table 1: Classical Problems formulable as LPs. Linear Assignment (left) and Shortest Path (right).

and edges part of the selected shortest path for the SP problem respectively. Although the original formulation of the LA and SP problems are actually integer LP formulations, which are generally known to be NP-complete as opposed to the less restrictive regular LPs, both problems can be solved in polynomial time. However, extensions of regular SP like the Travelling Salesman or the Canadian Traveller problems are known to be NP- and PSPAC- complete respectively.

Causal Models and Hidden Confounders. Following Peters, Janzing, and Schölkopf [2017], a Structural Causal Model (SCM) is defined as \( \mathcal{C} := (\mathcal{S}, P_{\mathcal{N}}) \) where \( P_{\mathcal{N}} \) is a product distribution over noise variables and \( \mathcal{S} \) is defined to be a set of \( d \) structural equations

\[ X_i := f_i(pa(X_i), N_i), \quad \text{where } i = 1, \ldots, d \]  

with \( pa(X_i) \) representing the parents of \( X_i \) in graph \( G(\mathcal{C}) \). An intervention on SCM \( \mathcal{C} \) as defined in (5) occurs when (multiple) structural equations are replaced through new non-parametric functions \( f^\prime(\cdot) \) thus effectively creating an alternate SCM \( \mathcal{C}^\prime \). An important property of interventions often referred to as “modularity” or “autonomy” states that interventions are fundamentally of local nature, formally \( p(\mathcal{C} | pa(X_i)) = p(\mathcal{C} | pa(X_j)) \) where the intervention of \( \mathcal{C} \) occurred on variable \( X_k \) opposed to \( X_i \). Furthermore, the cause-effect relations of variables imply an invariance to the origin of the mechanism (be it naturally or by intervention, see Pearl, Glymour, and Jewell [2016]).

An SCM \( \mathcal{C} \) induces a causal graph \( G \), induces an observational/associational distribution \( p(\mathcal{C}) \), it can be intervened upon using the do-operator and thus generate interventional distributions \( p(\mathcal{C}_{do(\cdots)}) \) and furthermore given some observations \( \mathcal{Y} \) can also be queried for interventions within a system with fixed noise terms amounting to counterfactual distributions \( p(\mathcal{C} \mid \mathcal{Y} = \mathcal{y}, \mathcal{do}(\cdots)) \). To query for a finite dataset of any given SCM, the procedure is simple: the structural equations are being simulated sequentially following the underlying causal structure starting from the exogenous variables.

Following Spirtes [2010], the set of variables \( X \) on which SCM \( \mathcal{C} \) enacts is called causally sufficient if there exist no hidden confounders (also referred to as common causes). A confounder or common cause is a variable \( \mathcal{C} \) that is causing more than one variable in \( X \) thereby suggesting for causal sufficiency relative to a larger set of “relevant” variables \( \hat{X} \supset X \) for which no pair of variables \( X, Y \) exists such that there is a directed path from \( C \) to \( X \) via \( Y \) and vice versa. Following the definition in Peters, Janzing, and Schölkopf [2017], confounding lies at the core of the asymmetry of interventional and conditional/observational distributions

\[ p(Y = y \mid do(X = x)) \neq p(Y = y \mid X = x), \]  

thus causal reasoning in the setting where interventions are not prohibitive\(^3\) depends on causal identification i.e., adjusting for instrumental variables \( \mathcal{C} \) to filter confounding

\[ p(Y = y \mid do(X = x)) = \sum_{\mathcal{C}} p(Y = y \mid X = x, \mathcal{C} = \mathcal{c}) p(\mathcal{C} = \mathcal{c}) \]  

Unfortunately, obtaining the aforementioned valid adjustment variables \( \mathcal{C} \) in the face of hidden confounding is difficult. A prominent example is Simpson’s paradox [Simpson 1951], which in fact is not a paradox but rather a consequence of hidden confounding. Historically, instances of the hidden confounding as described in Simpson’s paradox have been causal for confusion and erroneous conclusions that are especially worrisome in safety-critical settings like medicine [Charig et al. 1986; Bottou et al. 2013].

Generalized Adversarial Perspective

We first show the shortcoming of the classical adversarial perspective, then establish how causality can provide the missing semantics, then formalize our intuition while establishing our main results and finally conclude with an empirical examination and discussion of existing examples. Fig. 1 will act as the lead example throughout the paper, for which we first introduce motivation/intuition and then formalism.

Differentiable Perturbed Optimizers from the Classical Adversarial Perspective

In the past, MPs have been used defensively for verifying the robustness of neural learners to adversarial examples [Jeng, Xiao, and Tedrake 2019] and offensively for generating actual adversarial examples [Zhou et al. 2020]. Here, we are concerned with a fundamentally different research question: “How do adversarial attacks affect MPs themselves?”. That is, instead of considering MPs as a service to the system to be attacked, we consider the programs themselves to be under attack. We argue that adversarial attacks refer to a more general concept that affects MPs and thus being a property of the problem specification and not per se a property of the expressiveness of deep models or of the classification task. To establish an initial connection between adversarial attacks and MPs we consider general black-box solvers that are differentiable. More specifically, in the following we will consider so-called perturbed optimizers, which inject some noise that is distributed w.r.t. some differentiable probability distribution into the solver. This way these stochastically perturbed optimizers allow for systematically transforming discrete decisions within ML pipelines into differentiable decisions. Perturbed models have been considered for inference tasks within energy models [Papandreou and Yuille 2011] and regularization in online settings [Abernethy et al. 2014]. Initial works in this research direction date back to the Gumbel-max (Gumbel 1954) and were recently generalized to Differentiable Perturbed Optimizers (DPO) featuring end-to-end learnability [Berthet et al. 2020]. It is important to emphasize that the LP structure need not be known to use

\(^3\)Interventions can be costly w.r.t. different measures (finance, ethics etc.) rendering the purely observational setting the arguably more frequent setting in real world situations.
the perturbed maximizers, which is the main reason for us to consider DPO in the remainder of this work.

The key assumption required for an (LP) optimizer $x^*(w) = \arg\max_{x \in P} \langle x, w \rangle$ is the existence of a temperature-controlled random noise vector $\epsilon z$ with positive and differentiable density such that,

$$x^*(\tilde{w}) = x^*_\epsilon(w) = \mathbb{E}_{p_\epsilon(x)}[\arg\max_{x \in P} \langle x, w + \epsilon z \rangle],$$

where $x$ is the optimization variable living in the solution polytope $P$, where $w$ is the LP cost parameterization and $\langle \cdot, \cdot \rangle \in \mathbb{R}$ the inner product. Related work on differentiability of more general cone programs or linear optimization within predict-and-optimize settings (Agrawal et al. 2019; Mandi and Guns 2020) generally rely on the Karush-Kuhn-Tucker conditions, while Eq.[8] is black-box applicable thereby being generally more favorable. To further examine the susceptibility of DPO to the classical notion of adversarials, we will consider a Linear Assignment (LA) problem where the optimal matchings are being viewed as classes from the standard adversarial formulation for classification. Mathematically, the following correspondence is found,

$$J := F(\tilde{w}, w), \quad x := \tilde{w}, \quad y := x^*(w), \quad f_\theta := x^*(\cdot),$$

where classically $J$ is the cost function, $x$ the feature vector, $y$ the class label and $f_\theta$ the predictive model, while $F$ is a distance measure acting on the actual/original $x^*(w)$ and the expected perturbed solution $x^*_\epsilon(\tilde{w})$ and our program solver is considered to be the predictive model. Note that $J, \ldots$ is notation from (Goodfellow, Shlens, and Szegedy 2015), also we use $\tilde{w} = w_\epsilon$ interchangeably to denote the perturbed parameterization. In LA, each optimal matching code $x^*(w) \in [0, 1]^n$ is then considered to be a class and the gradient $\nabla_{\tilde{w}} F$ could be used for performing an “adversarial attack” such that the ‘class’ changes (significantly) while the input remains approximately the same. The major problem faced with this classical perspective/formulation is that there exists no “semantic impact” to be observed for the human inspector akin to a neural network wrongly classifying a dog (small animal) as a plane (big travel machine). In the following, we will motivate and present a causal perspective onto adversarials leading to an extended notion (hidden measure attacks) in which the argued expected semantic impact is being well aligned with a human inspector’s intuition.

### Causality Determines Adversarial Risk of DPO

Intuitively, when attacking DPO in the classical sense, one would expect that any pair of optimal solutions $(x^*_1, x^*_2)$ (by their optimality) induces similar optimization costs $w^T x^*_1 \approx w^T x^*_2$ even if the solution vectors themselves differ. However, this observation, while arguably intuitive, does not negatively prove that adversarial meaning is not to be found within DPO. On the contrary, the possibly huge discrepancy between some pair of optimal solutions when considered in terms of codes as in the LA example (i.e., the structural hamming distance $H_{\text{Hamming 1950}}$ will be significantly larger than zero, $SHD(x^*_1, x^*_2) \gg 0$) seems to suggest that there exists some fundamental difference in solution albeit not for the specific optimization objective at hand. We follow this thought and observe that the underlying causality is indeed indicative and even decisive for defining adversarial risk on DPO. Specificaly, the structural causal model’s (SCM) data generating process implies a translation $\phi$ from the SCM domain to the cost parameterization of the LP and, if $\phi$ is not faithful to the underlying hidden reality $C$ then the existence of hidden confounders implies adversarial risk for DPO. The connection between causality’s SCM and MPs (LPs in our running example) is being illustrated in Fig.[3] The implied consequence of this connection between causality and LPs, i.e. the existence of intruding parameters which are unaware of risk-inducing confounders, is being illustrated in Fig.[3]. We now proceed to elaborate in detail using our concrete lead example from Fig[1]. Following the lead principles from (Pearl 2009; Peters, Janzing, and Schölkopf 2017), there exists some underlying but hidden reality that can be expressed in causal terms, in this case individuals $i \in D(\mathcal{C})$ are being characterized by the structural equations of the SCM $\mathcal{C}$. That is, we have data $D = \{d_i\}_i$ where $d_i := (w_i, h_i, p_i) \in W \times H \times P =: D(\mathcal{C})$ being an individuals wealth, health and prioritization for a vaccine respectively. We argue that a human modeller that tries to design a fair assignment for vaccine spots, modelled as a LA problem (see Tab[1]), will perform a translation from the SCM-generated population $D$ to a cost parameterization $w$ for the specific matching-task LP. Mathematically, this translation is given by a function $\phi : D(\mathcal{C}) \rightarrow \mathbb{R}^n$ where

$\phi(\hat{w}) := w_\epsilon$

Figure 2: Parameters of an SCM. The unobserved SCM $\mathcal{C}$ implies a causal graph $G$, can simulate data $D \in D(\mathcal{C})$ underlying some data distribution $p(D(\mathcal{C}))$ and it also provides for parameters $\phi$ that specify an LP optimization. Here $p$ and $\phi$ are unaware of confounder $W$. (Best viewed in color.)
\(D\) is the domain tuple of the SCM. For instance, wealth is a confounder (or common cause), i.e., the joint density \(p(H, P)\) is skewed towards individuals of higher wealth so that wealthy people can bypass the otherwise fair system of the modeller. Unaware of this hidden confounder, the modeller’s design choices constitute a translation consistent with a sub-graph of the hidden reality, \(\phi\) of \(\phi\) being the parameterization of \(\phi\) in the sense that the causal parameterization \(\phi\) can be exploited in the \(\eta\)-perturbed solution. (Best viewed in color.)

**Hidden Measure Attacks**

To establish a bridge between the intriguing parameters of SCM and adversarial susceptibility of DPO, we now formally define what is meant by SCM parameters as in Fig.2 when inspected in the broader context of MPs.

**Definition 1 (Parameters of SCM).** Given a SCM \(\mathcal{C}\), an induced data set \(D = \{ \text{d} \in D(\mathcal{C}) \}^K\) and some function \(\phi : D(\mathcal{C}) \rightarrow \mathbb{R}^N\), where \(D\) is the domain-tuple of an SCM, the cost parameter \(W \in \mathbb{R}^{K \times N}\) of an LP \(L\) is also being referred to as parameter of the SCM \(\mathcal{C}\) if \(\forall k \in K : \phi(\text{d}_k) = w_k\) holds. Further, \(\phi\) is being referred to as causal parameterization of the SCM \(\mathcal{C}\) and writes as \(\phi \models \mathcal{C}\).

Note that matrix \(W\) is identical to what has been denoted as weight vector \(w\) of the LP upon being flattened. The causal parameterization \(\phi\) provides a numerical description for any data point \(d \in D(\mathcal{C})\) and differs from embeddings in the sense that the \(\phi(d)\) are considered parameters of the LP specification. We argued that there exist different causal parameterization which a modeller can perform to formulate a specific LP instance in the light of how the given reality is being observed, that is, which underlying (possibly implicit) SCM model is being assumed. To this end, we define the following notion of ‘realism’ for modelling LPs from SCM:

**Definition 2 (Causal realism).** Let the set of variables \(X\) modelled by SCM \(\mathcal{C}\) be causally sufficient with \(\phi\) being the causal parameterization. If there exists an extended set of variables \(\tilde{X} \supseteq X\), that is causally insufficient, then \(\phi\) is called causally unrealistic. Otherwise, \(\phi\) is causally realistic.

By Def. 2, it becomes clear that the modeller’s estimated set of relevant variables \(V(\hat{\mathcal{C}}) = \{H, P\}\), where \(\hat{\mathcal{C}}\) is the SCM modelled by SCM, cannot provide for a causally realistic parameterization \(\phi\), since in the given example there exists a hidden confounder: an individual’s wealth \(W\).

**Proposition 1 (Hidden Confounders and Causal Reaslist).** Let \(H\) be a hidden confounder and \(S\) the set of all possible SCMs that do not consider \(H\), that is, \(S := \{ \mathcal{C} \mid H \notin V(\mathcal{C}) \}\). Then the following statement holds: \(\forall \mathcal{C} \in S : \phi \models \mathcal{C} \iff \phi\) is causally unrealistic.

Due to space restrictions, we provide all subsequent proofs within the supplementary material. Prop. 1 implies that the modeller’s chosen parameterization \(\phi\) will always be causally unrealistic as we can always trivially design \(\tilde{X} := \{W, H, P\} \supset V(\hat{\mathcal{C}})\). Later on, an important consequence of the existence of hidden confounders and thereby the existence of causally unrealistic parameterizations will be the possibility of adversarial attacks on DPO. However, before a definition of such attacks is possible, we need to make the important observation that there exist LP parameterizations/instances that allow for multiple optimal solutions i.e., \(|P^*| > 1\) where \(P^* := \{ x \in \mathcal{P} \mid x \text{ is optimal} \}\). We refer to the situation where \(|P^*| > 1\) as tie-breaks, since the solver selects an optimal solution “for breaking the tie”. While the solver might choose among the different possibilities within a tie-break, one can alternatively use use perturbations to obtain a close LP approximation that enforces a decision.

**Proposition 2 (DPO Tie-Break Resolving Perturbations).** Let \(w\) be the parameterization of LP \(L\) with \(|P^*| > 1\) and let \(x^* \in P^*\) be the selected optimal solution by the DPO. Then there exists a function \(F\) and \(\epsilon \in \mathbb{R}_+^+\) such that the perturbed parameterization \(w_\eta = w + \epsilon\eta\) with \(\eta := \nabla w F\) leads to another set of optimal solution(s) \(|P^*_\eta| \neq |P^*|\).

Often times in practice it can be observed that a resolved
We also show Prop. 2 empirically for a simple Shortest Path parameter. More generally, Prop. 2 implies that the gradient $\nabla_w F$ w.r.t. cost parameter $w$ of the LP will contain a signal for resolving the tie-break, which can be used to perform a class change. We also show Prop. 2 empirically for a simple Shortest Path (SP) problem example in Fig. 4. We argue that this class change, in combination with the previously established results on hidden confounders, is enough for extending adversarial attacks to DPO (which solve MPs in general).

**Definition 3 (Hidden Measure Attacks).** Given a DPO $x^* \in \mathcal{P}$ and some $\tau \in \mathbb{R}_+^*$, an adversarial attack $\hat{w} = w + \eta$ via perturbation $\eta$ is also called hidden measure attack if there exists a bijection, called hidden measure, $h : \mathcal{P} \rightarrow \mathbb{R}^+$ such that $|h(x^*(\hat{w})) - h(x^*(w))| > \tau$ holds.

The hidden measure attacks from Def. 3 we denote as HMA. Empirically, we observe arguably high differences ($\tau$) in the hidden measure values. A simple but important corollary suggests the broad possibility of such attacks:

**Corollary 1 (Change in Optimal Solution).** Let $P^*$ be the set of optimal solutions and $h$ a hidden measure. Then the following statement is true: $\forall (x_1, x_2) \in P^* \times P^*: x_1 \neq x_2 \implies h(x_1) \neq h(x_2)$.

Cor. 1 implies that the hidden measure $h$ will always make use of any discrepancy present between pairs of optimal solutions. This simple statement captures the central observation of missing semantics for applying classical adversarial attacks to MPs, as discussed at the beginning of the paper. To finally bridge the gap between causality and adversarials, we provide the required semantics in the form of the hidden measure $h$. This leads us to our main result which suggests that a HMA can be constructed via causality’s confounders.

**Theorem 1 (Confounders Determine HMA).** Let $\phi$ be a causal parameterization of the SCM $\mathcal{C}$. If $\phi$ is causally unrealistic, then there exists a hidden measure $h$ based on the hidden confounders of $\mathcal{C}$.

Given the trivial construction of a hidden measure based on a hidden confounder (Thm. 1), it becomes evident that a modeller’s design choices are crucial for the implied adversarial risk for any subsequent optimization. It is important to note that there can exist hidden measures, not based on confounders, but rather descendants underlying the causal reality of the system. This realizes an important distinction amongst different representations of decision variable:

**Definition 4 (Symmetric Decision Variables).** Let $x := (x_1, \ldots, x_N) \in \mathbb{R}^N$ be a decision variable and let $T_i$ be the type (or description) of the quantity being modelled by $x_i$. If the statement $\forall i, j \in N^2 : T_i = T_j$ holds, then we call the decision variable symmetric. Otherwise, asymmetric.

In Fig. 5 we provide a second HMA example based on an SP problem (Tab. 1) in addition to the lead example from Fig. 1. Both examples define symmetric decision variables (LA: $T_i :=$ individual $i$ to be vaccinated; SP: $T_i :=$ road segment $ij$ to be taken) but the energy system model (discussed below) realizes the asymmetric case. We further argue that these types $T$ are reminiscent of variables in a causal graph, naturally implying a bigger graph for asymmetric decisions.

**Conjecture 1 (Asymmetric Susceptibility).** Let $x_A, x_S$ be (asymmetric decision variables in $\mathbb{R}^N$ respectively and let $H_A, H_S$ be the corresponding numbers of existing hidden measures, then $H_A > H_S$.

The intuition for Conj. 1 follows from both the causal perspective that (a) asymmetric decisions allow for generally more variables in the system to be exploited and (b) descendants in a causal sequence (opposed to only confounders) can also allow for hidden measures.

**Empirical Illustration**

To showcase the theoretical results established, we provide existential proof of HMA examples. We presented a real world inspired example on vaccination scheduling as our lead example (Fig. 1) and thereby only point to a summary within the supplementary material as a compact reference to this specific example, since it has been intensively discussed already. In the following we discuss the other real-world inspired example for a SP problem:

**Travelling from New York City to San Francisco...via Canada?** To further motivate the relevance of adversarial risk outside classification, let us consider the example in Fig. 7. We consider the development of an autonomous car. We let the developmental autonomous car travel within North America from New York City (NY) to San Francisco (SF). Our SP has the intention of reducing overall toll costs for the optimal route, which from experience can be hefty. Our LP cost $w_{ij} \in \mathbb{R}_{>0}$ represents the toll cost when travelling on any road segment from $i$ to $j$. In this example, we know the toll costs for a relevant set of road segments within NA where the Canadian road toll policy is comparably modest. Thereby, having specified the LP, our solver returns...
account for an adversarial attack in that causal descendants (i.e., that photovoltaics are being built pre-dominantly) can ket electricity etc.), thus, the difference in optimal solutions of said technology (see Tab. 2). Opposed to our previous ex-

vantage of one technology implies an increased production solver is asked to balance the usage of the different technolo-

bles a simplified version of the TIMES model (Loulou et al.

2005). We detail the full LP in the Supplementary. The LP

2.05). We detail the full LP in the Supplementary. The LP

Finally, we consider an en-

lem of Energy System Design. 

Discussion of another Example: The Real World Problem of Energy System Design. Finally, we consider an ener-

gy model characterizing the energy portfolio of a single-

security, a single-

family house. It is based on real world data for demand and commonly used equations from energy systems re-

search (Schaber, Steinke, and Hamacher 2012). The exam-

ined model considers photovoltaics, market electricity and heating gas over a year time frame (in hours) and resem-

bles a simplified version of the TIMES model (Loulou et al.

2005). We detail the full LP in the Supplementary. The LP

solver is asked to balance the usage of the different technolo-

gies for matching the required demand such that overall cost is being minimized. Naturally, we observe that a price advantage of one technology implies an increased production of said technology (see Tab. 2). Opposed to our previous ex-

ample, the energy system LP considers an asymmetric opti-

mization variable (e.g. $\bar{v}_i :=$ photovoltaics (PV), $\bar{v}_2 :=$ market electricity etc.), thus, the difference in optimal solutions (i.e., that photovoltaics are being built pre-dominantly) can account for an adversarial attack in that causal descendants

$\bar{x}_U \in [0,1]^{n \times n}$ suggesting a route through the mid-US (where $n$ is the number of cities). By minimally per-

turbing the original parameters, our solver now chooses an alternate solution $\bar{x}_{CA} := x^*(\bar{w})$ suggesting a route across the border via Canada. While evidently the alternate route deviates strongly in terms of selected road segments, mathematically $SHD(x_{US}, x_{CA}) \gg 0$ where $SHD(\cdot, \cdot) \in \mathbb{N}$ is the Structural Hemmen Distance, our model is in fact trust-

fully returning the optimal solution as cost-wise the state-

ment $w^T x_{US} \approx \bar{w}^T x_{CA}$ holds. Nonetheless, with respect to $CO_2$ emissions, which in this case are being represented as the hidden measure $h : \mathcal{P} \to \mathbb{R}$ acting on the LP-Polytope $\mathcal{P}$ that is accessible and being exploited by the adversary, the alternate optimal solution performs significantly worse: $h(x_{CA}) \gg h(x_{US})$ and the HMA is completed.

Figure 5: Another Example: Increased $CO_2$ Emissions. The edges in the graph represent tolls to be paid for travelling a given road segment. The hidden measure attack reveals that travelling via Canada instead of mid-US will amount to the same total travel toll to be paid but the $CO_2$ emissions drastically deviate between the solutions. (Best viewed in Color.)

Table 2: Dominating Technologies. Price perturbations ($w_{PV}$) can boost PV production $Cap_{PV}$ (green) which leads to a significant increase in risk of working injury or fire.

| Dem (µ) | Cap PV | CapGen | Self-Gen | TOT Ex | CAP Ex | Capgen | Capgen | w PV |
|---------|--------|--------|----------|--------|--------|--------|--------|------|
| 3000    | 1.12   | 2.47   | 0.66     | 468.24 | 215.47 | 1.75   | 0.018  | 0.01 |

are involved (opposed to hidden confounders). As example, consider the causal descendants risk of fire or of working injury (falling panels, shattering glas). Nonetheless, we argue that the limitations on PV-production and Market-buy act as discrepancy counter-measures that require the system to balance out different technologies.

Conclusions and Future Work

We presented the first approach that moves adversarial at-

tacks beyond deep network towards perturbed optimizers, instead of using the optimization models merely as a way to model such attacks. To cope with the problems of a classical adversarial perspective, we established a rigorous connection between causality and mathematical programs by extend-

ing the structural causal model’s data generating process to the parameterization of the latter. By this, we could fur-

ther connect causality to these extended adversarial attacks called hidden measure attacks. We proved that confounders can be used to construct such attacks while providing further existen-
tial proof within two classical linear program families. We further considered a pressing real world optimization problem concerned with energy portfolios. Defensive strategies through extended modelling awareness and assumptions naturally follow from this work. Using an equivalence of LPs to ReLU-nets might offer for an interesting connection between causality and deep learning. Also, extending our theoretical framework with results on critical regions from MPLP-litterature poses an interesting future direction.
References

Abernethy, J.; Lee, C.; Sinha, A.; and Tewari, A. 2014. Online linear optimization via smoothing. In Conference on Learning Theory, 807–823. PMLR.

Agrawal, A.; Amos, B.; Barratt, S.; Boyd, S.; Diamond, S.; and Kolter, Z. 2019. Differentiable convex optimization layers. NeurIPS.

Akhtar, N.; and Mian, A. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access.

Bach, F. 2013. Learning with submodular functions: A convex optimization perspective. Foundations and Trends in Machine Learning.

Berthet, Q.; Blondel, M.; Teboul, O.; Cuturi, M.; Vert, J.-P.; and Bach, F. 2020. Learning with differentiable perturbed optimizers. In NeurIPS.

Bica, I.; Alaa, A.; and Van Der Schaar, M. 2020. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In ICML.

Bottou, L.; Peters, J.; Quinonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual reasoning and learning systems. JMLR.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. ICLR.

Charig, C. R.; Webb, D. R.; Payne, S. R.; and Wickham, J. E. 1986. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed).

Chen, S.-T.; Cornelius, C.; Martin, J.; and Chau, D. H. P. 2018. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In ECML-PKDD.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. ICLR.

Gumbel, E. J. 1954. Statistical theory of extreme values and some practical applications: a series of lectures, volume 33. US Government Printing Office.

Guo, C.; Gardner, J.; You, Y.; Wilson, A. G.; and Weinberger, K. 2019. Simple black-box adversarial attacks. In ICML.

Hair Jr, J. F.; and Sarstedt, M. 2021. Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. Journal of Marketing Theory and Practice.

Hamming, R. W. 1950. Error detecting and error correcting codes. The Bell system technical journal, 29(2): 147–160.

Handa, A.; Sharma, A.; and Shukla, S. K. 2019. Machine learning in cybersecurity: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4): e1306.

Hoiles, W.; and Schaar, M. 2016. Bounded off-policy evaluation with missing data for course recommendation and curriculum design. In ICML.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. In ICML.

Loulou, R.; Remme, U.; Kanudia, A.; Lehtila, A.; and Goldstein, G. 2005. Documentation for the times model part ii. Energy Technology Systems Analysis Programme.

Mandi, J.; and Guns, T. 2020. Interior Point Solving for LP-based prediction+ optimisation. NeurIPS.

Papandreou, G.; and Yuille, A. L. 2011. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In 2011 International Conference on Computer Vision, 193–200. IEEE.

Paulus, A.; Rolínek, M.; Musil, V.; Amos, B.; and Martius, G. 2021. CombOptNet: Fit the Right NP-Hard Problem by Learning Integer Programming Constraints. arXiv preprint arXiv:2105.02343.

Pearl, J. 2009. Causality. Cambridge university press.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. Causal inference in statistics: A primer. John Wiley & Sons.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. Elements of causal inference. The MIT Press.

Schafer, K.; Steinke, F.; and Hamacher, T. 2012. Transmission grid extensions for the integration of variable renewable energies in Europe: Who benefits where? Energy Policy.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological).

Spirtes, P. 2010. Introduction to causal inference. JMLR.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In ICLR.

Tjeng, V.; Xiao, K.; and Tedrake, R. 2019. Evaluating robustness of neural networks with mixed integer programming. ICLR.

Wu, K.; Wang, A.; and Yu, Y. 2020. Stronger and faster Wasserstein adversarial attacks. In ICML.

Zhou, N.; Luo, W.; Lin, X.; Xu, P.; and Zhang, Z. 2020. Generating Multi-label Adversarial Examples by Linear Programming. In IJCNN.

Zügner, D.; Akbarnejad, A.; and Günnemann, S. 2018. Adversarial attacks on neural networks for graph data. In KDD.
Appendix - Intriguing Parameters of Structural Causal Models

We make use of this Appendix/Supplement following the main paper to provide details and additional insights we deem important for the reader and for what has been examined in the main paper.

Proofs for Propositions 1 and 2

Prop[1] makes the simple suggestion that if we know that in fact there exist hidden confounders for our SCM $\mathcal{C}$ for which we are not aware of, then any parameterization $\phi$ derived from that SCM will in fact be causally unrealistic since the assumed hidden confounders exist.

**Proposition 1 (Hidden Confounders and Causal Reaslim).** Let $H$ be a hidden confounder and $S$ the set of all possible SCMs that do not consider $H$, that is, $S := \{ \mathcal{C} \mid H \notin V(\mathcal{C}) \}$. Then the following statement holds: $\forall \mathcal{C} \in S : \phi \models \mathcal{C} \iff \phi$ is causally unrealistic.

**Proof.** By construction, we can always define the causally sufficient set of variables $V(\mathcal{C}) \cup H$.

On the other hand, Prop[2] makes a claim about resolving tie-breaks within an optimization. I.e., we can always find a function $F$ and some perturbation parameter $\epsilon$ such that the new optimization will produce a new set of optimal solutions (which usually will also be smaller than the original set).

**Proposition 2 (DPO Tie-Break Resolving Perturbations).** Let $w$ be the parameterization of LP $L$ with $|P^*| > 1$ and let $x^* \in P^*$ be the selected optimal solution by the DPO. Then there exists a function $F$ and $\epsilon \in \mathbb{R}^+$ such that the perturbed parameterization $w_\eta = w + \epsilon \eta$ with $\eta := \nabla_w F$ leads to another set of optimal solution(s) $|P^*_\eta| \neq |P^*|$.

**Proof.** A tie-break $|P^*| > 1$ occurs when there exists exactly one constraint such that its hyperplane $h$ is perpendicular to the objective vector, $h \perp w$. We can always choose $\epsilon > 0$ and $F := ||x^* - x^*_\eta||_2^2$ such that $\eta > 0$. Consequently, the perturbed weight vector $w_\eta$ will be linearly independent of $w$ and we have that $h \not\perp w_\eta$.

Proof for Corollary 1

Corr[1] suggests a very simple statement of great importance to the notion of a hidden measure attack in that any pair of solutions $(x_1^*, x_2^*)$ to some optimization will only be different in terms of their hidden measures $h(x_1^*) - h(x_2^*) > 0$ if they were different to begin with, that is $x_1^* \neq x_2^*$.

**Corollary 1 (Change in Optimal Solution).** Let $P^*$ be the set of optimal solutions and $h$ a hidden measure. Then the following statement is true: $\forall (x_1, x_2) \in P^* \times P^* : x_1 \neq x_2 \implies h(x_1) \neq h(x_2)$.

**Proof.** By definition, any hidden measure $h$ we choose will be a bijective measure and any bijection is also injective.

Proof for Theorem 1

Thm[1] is our key insight regarding the construction of hidden measure attacks. We establish that if $\phi$ is causally unrealistic, then we will always be prone to hidden measure attacks. Thereby, causality gains an importance in how we specify our optimization problems.

**Theorem 1 (Confounders Determine HMA).** Let $\phi$ be a causal parameterization of the SCM $\mathcal{C}$. If $\phi$ is causally unrealistic, then there exists a hidden measure $h$ based on the hidden confounders of $\mathcal{C}$.

**Proof.** We have to show that a hidden measure $h$ can be constructed from hidden confounders of SCM $\mathcal{C}$. It follows from $\phi \models \mathcal{C}$ and $\phi$ being causally unrealistic that there exists a hidden confounder $H$. Thus information from $H$ cannot be contained within the LP cost parameter $w = \phi(d)$. We can always construct a bijection $h : x \mapsto \sum_k h_k$ where $h_k = \tilde{f}_k^H$ is the $H$-value of sample $k$ with $\tilde{f}^H$ being the structural equation of $H$ in an extended SCM $\mathcal{C}$. Let $\eta$ be an adversarial perturbation, then with Corr[1] it follows that $h$ is a bijection and also a hidden measure.

Differentiable Hidden Measures

For performing adversarial attacks (in a whitebox manner), gradients are usually being used thereby assuming that the (modules of the) system are differentiable. Also for HMA, the differentiability of the hidden measure $h$ marks the essence for any targeted attack. Differentiability of the hidden measure can for instance be achieved by providing $h$ approximately via a deep neural net, $h := f(x; \theta)$ where $f$ is the composition of layers (linear combinations plus e.g. ReLU-linearities) and $\theta$ are the weights/biases. Fig[6] shows all relevant high-level-components schematically.

Recap of the Example “A Bias towards the Wealthy for Vaccination Scheduling”

Imagine a vaccination company supported by the government to improve the speed of overall vaccination. The lead modeller has decided to assign each individual a priority list for certain available vaccine spots w.r.t. the knowledge on the individual’s health and recognizes this setting as a Linear Assignment (LA) problem. The intended policy of the modeller can be characterized by higher prioritization for individuals of lower health. This characteristic is eventually resembled by the cost matrix $w$ the modeller comes up with to specify her LP that should then provide for an optimal matching (in reality, $w = \phi(d)$ with $\phi \models \mathcal{C}$ where the SCM $\mathcal{C}$ is the causal pendant to the modeller’s “mental” model). Now, an adversary performs HMA to create an approximate $\hat{w}$ ($\approx w$) and replaces the modeller’s initially designed cost. The modeller unaware of the miniscule change observes the optimal matching $x^*(\hat{w})$ which also does not raise any suspicion as the general policy “lower health, higher priority” is kept intact with adequate cost $w^*x^*(w) \approx \hat{w}^*x^*(\hat{w})$. However, the adversary successfully exploited the differences in wealth of the individuals in consideration (discrepancy under hidden measure, ...
toll cost when travelling on any road segment from

Akin to confounder \( Z \), the hidden measure \( h \) (constructable from hidden confounders, being for instance \( CO_2 \) emissions as in the example of Fig.5) dilutes the information flow of the solver allowing for adversarial attacks on mathematical programs opposed to classification tasks. A differentiable non-parametric function approximator (FA), e.g. a neural network, can be used to model the hidden measure to be used for performing the attack. To learn such an approximator, a supervision setting where the training data set consists of structural equation values for the confounders of the underlying SCM could be considered.

Extended Interpretation for the ”Travelling from NY to SF...via Canada?” Example

To further motivate the relevance of adversarial risk outside classification, let us consider a real world inspired example that is being showcased in Fig.5 which is concerned with a Shortest Path (SP) problem. In the corresponding real world setting, we might consider the development of an autonomous car. We let the developmental autonomous car that is being showcased in Fig.5 which is concerned with a Shortest Path (SP) problem. In the corresponding real world setting, we might consider the development of an autonomous car. We let the developmental autonomous car.

\[
h(x^*(\hat{w})) \gg h(x^*(w)).\]

Now, unintended from the model-\(\hat{w}\), the hidden measure \( h \) (based on confounder wealth \( W \) hidden to the modeler’s \( C \)) has lead to people of higher wealth to take vaccination spots pre-dominantly.

Figure 6: Differentiability of the Hidden Measures akin to Hidden Confounders. Following the idea behind hidden confounders in causality (graph on the right) that considers the dilution of the causal effect between two observed variables, e.g. the causal effect of a given treatment on a patient’s recovery from kidney stones \( T \rightarrow R \) which can lead to false conclusions if not being considerate of the kidney stone size \( Z \). Akin to confounder \( Z \), the hidden measure \( h \) (constructable from hidden confounders, being for instance \( CO_2 \) emissions as in the example of Fig.5) dilutes the information flow of the solver allowing for adversarial attacks on mathematical programs opposed to classification tasks. A differentiable non-parametric function approximator (FA), e.g. a neural network, can be used to model the hidden measure to be used for performing the attack. To learn such an approximator, a supervision setting where the training data set consists of structural equation values for the confounders of the underlying SCM could be considered.

\[
h(x^*(\hat{w})) \gg h(x^*(w)).\]

Now, unintended from the model-\(\hat{w}\), the hidden measure \( h \) (based on confounder wealth \( W \) hidden to the modeler’s \( C \)) has lead to people of higher wealth to take vaccination spots pre-dominantly.

Extended Interpretation for the ”Travelling from NY to SF...via Canada?” Example

To further motivate the relevance of adversarial risk outside classification, let us consider a real world inspired example that is being showcased in Fig.5 which is concerned with a Shortest Path (SP) problem. In the corresponding real world setting, we might consider the development of an autonomous car. We let the developmental autonomous car travel within North America from New York City (NY) to San Francisco (SF). Our SP has the intention of reducing overall toll costs for the optimal route, which from experience can be hefty. Our LP cost \( w_{ij} \in \mathbb{R}_{>0} \) represents the toll cost when travelling on any road segment from \( i \) to \( j \). In this example, we know the toll costs for a relevant set of road segments within NA where the Canadian road toll policy is comparably modest. Our LP model subsequently solves any given SP problem instance, fully parameterized by the directed acyclic graph (DAG) \( w \in \mathbb{R}^{n \times n} \) with \( n \) being the total number of different cities we have specified, returning \( x_{US} := x^*(w) \in [0,1]^{n \times n} \) suggesting a route through the mid-US. By minimally perturbing the original DAG, that is \( \hat{w} \approx w \), our solver now chooses an alternate solution \( x_{CA} := x^*(\hat{w}) \) suggesting a route across the border via Canada.\(^7\) While evidently the alternate route deviates strongly in terms of selected road segments, mathematically \( SHD(x_{US}, x_{CA}) \gg 0 \) where \( SHD(\cdot, \cdot) \in \mathbb{N} \) is the Structural Hemming Distance, our model is in fact trustfully returning the optimal solution as cost-wise the statement \( w^\top x_{US}^* \approx \hat{w}^\top x_{CA}^* \) holds. Nonetheless, the aforementioned deviation in terms of the resulting binary codes lends itself to a severe consequence in terms of adversarial risk i.e., with respect to \( CO_2 \) emissions, which in this case are being represented as a hidden measure \( h : P \rightarrow \mathbb{R} \) acting on the LP-Polytope \( P \) that is accessible and being exploited by the adversary, the alternate optimal solution performs significantly worse: \( h(x_{CA}^*) \gg h(x_{US}^*) \). Conclusively, the adversary has exploited a hidden measure (here \( CO_2 \) emissions) akin to a hidden confounder in causal models in that the original model does not capture the influence of the measure onto the difference in acquired solutions. We therefore argue that in consideration of such a hidden measure one can more generally define adversarial attacks for mathematical programs beyond the original formulation in the classical setting for classification (and deep networks), making the attack a consequence of not the specific methodology being applied to the problem but problem specification itself.

\[^7\]For assuring minimality in the required change (and by that maximizing the ’invisibility’ of the attacker) the given example assumes an initial guess for the DAG to be close to a tie-break within the given SP problem instance, mathematically \( \exists w : |X^*(w)| > 1 \) where \( X^* \) is the set of all optimal solutions.
\[
\begin{align*}
\min_{Cap, p} & \quad c_{PV} \times Cap_{PV} + c_{Bat} \times Cap_{Bat} + \sum_t c_{Ele} \times p_{Ele}(t) + \sum_t c_{Gas} \times p_{Gas}(t) \\
\text{s.t.} & \quad p_{Ele}(t) + p_{PV}(t) + p_{Bat}^{out}(t) - p_{Bat}^{in}(t) - p_{Gas}(t) = D(t), \forall t \\
& \quad p_{Bat}^{S}(t) = p_{Bat}^{S}(t-1) + p_{Bat}^{out}(t) - p_{Bat}^{in}(t), t \in \mathbb{Z}^+, T \\
& \quad 0 \leq p_{PV}(t) \leq Cap_{PV} \times avail_{PV}(t) \times \delta t, \forall t \\
& \quad 0 \leq p_{Bat}^{in}(t), p_{Bat}^{out}(t) \leq Cap_{Bat}, \forall t \\
& \quad 0 \leq p_{Gas}(t) \leq U_{Gas}, \forall t \\
& \quad p_{Bat}^{S}(0) = 0 \\
& \quad 0 \leq p_{Ele}
\end{align*}
\]

Table 3: **Real-world Optimization Modelling Example: 1-year Energy Systems LP for an Average Household.** A large LP that unrolls for 8760 time steps (8760 hours = 1 year). Model based on (Schaber, Steinke, and Hamacher 2012), the quantities represent: Cost for Photovoltaics \(c_{PV} (\text{\(\text{C/kW}\)})\), Battery \(c_{Bat} (\text{\(\text{C/kWh}\)})\), Market Electricity \(c_{Ele} (\text{\(\text{C/kWh}\)})\), Gas \(c_{Gas} (\text{\(\text{C/kWh}\)})\), and the total Demand \(D (\text{\(\text{kWh/Year}\)})\).

| \(c_{PV}\) | \(c_{Bat}\) | \(c_{Ele}\) | \(D\) | \(c_{Gas}\) |
|---|---|---|---|---|
| 0.005 | 300 | 0.25 | 3000 | 0.25 |
| 0.001 | 300 | 0.25 | 3000 | 0.25 |

Table 4: **Parameterization Energy-System.** Cost for Photovoltaics \(c_{PV} (\text{\(\text{C/kW}\)})\), Battery \(c_{Bat} (\text{\(\text{C/kWh}\)})\), Market Electricity \(c_{Ele} (\text{\(\text{C/kWh}\)})\), Gas \(c_{Gas} (\text{\(\text{C/kWh}\)})\), and the total Demand \(D (\text{\(\text{kWh/Year}\)})\).

**Real-world LP: One-family House Energy Portfolio for 1 Year**

We considered an energy model for modelling the energy portfolio of a single-family house based on real world data for demand and commonly used equations from energy systems research (Schaber, Steinke, and Hamacher 2012). The examined model considers photovoltaics (PV), market electricity and heating gas over a year time frame (in hours) and resembles a simplified version of the TIMES model (Loulou et al. 2005). The optimal solution balances the usage of the different technologies for matching the required demand such that overall cost is being minimized. The specific LP template is given in Tab. 3.

Note that \(t \in \{0, \ldots, 8760\}\) with year \(= 8760\) hours rendering the template a very large single LP modelling each hour of the year. However, technologies like PV, in their capacity \((Cap_{PV})\), do not depend on \(t\) which would correspond to the real world intuition that one does not decide and subsequently build new PV for any given hour as it poses a single, fixed-timeframe investment. The limitations on PV-production and Market-buy of electricity act as discrepancy counter-measures that require the system to balance out different technologies i.e., while there will still be dominating technologies under price advantages the maximum skew of the portfolio is naturally being protected from being too drastic as both PV and bought electricity are limited in their “availability” (e.g. solar exposure, roof capacity, law regulations etc.) and thus cannot be naively maximized.

Details for the Reproduction of the HMA Examples

For the LA example, vacc. bias towards the wealthy, we use \(N = 15\) sampling iter. for the perturbation with the temperature parameter \(\sigma = 0.5\) and an attack step \(\epsilon = 0.01\) while for the SP example, travelling from NY to SF via Canada, we use more sampling iterations \((N = 20)\) using a lower temperature \((\sigma = 0.25)\). The energy system model is being parameterized by the quantities in Tab.4.