Neural Computing and Applications (2025) 37:24669–24683
https://doi.org/10.1007/s00521-024-10616-1

S.I.: MULTI-OBJECTIVE DECISION MAKING 2023

Generalization in multi-objective machine learning

Peter Súkeník[1] · Christoph Lampert[1]

Received: 14 November 2023 / Accepted: 4 October 2024 / Published online: 14 December 2024

Abstract
Modern machine learning tasks often require considering not just one but multiple objectives. For example, besides the *prediction quality*, this could be the *efficiency*, *robustness* or *fairness* of the learned models, or any of their combinations. Multi-objective learning offers a natural framework for handling such problems without having to commit to early trade-offs. Surprisingly, statistical learning theory so far offers almost no insight into the generalization properties of multi-objective learning. In this work, we make first steps to fill this gap: We establish foundational generalization bounds for the multi-objective setting as well as generalization and excess bounds for learning with scalarizations. We also provide the first theoretical analysis of the relation between the Pareto-optimal sets of the true objectives and the Pareto-optimal sets of their empirical approximations from training data. In particular, we show a surprising asymmetry: All Pareto-optimal solutions can be approximated by empirically Pareto-optimal ones, but not vice versa.

Keywords Machine learning · Multi-objective optimization · Generalization bounds · Pareto-front

1 Introduction

Traditionally, statistical machine learning has concentrated on solving one single-objective optimization problem: to minimize the average loss over a given training set. Additional quantities of interest, such as *model complexity*, had to be either addressed implicitly by the choice of model class, or integrated into the main objective via weighted regularization terms. Recently, however, additional quantities of interest have made it into the focus of the machine learning community, such as the *fairness*, *robustness*, *efficiency* or *interpretability* of the learned models. Optimizing these can be in conflict with the goal of low training loss, and task-specific trade-offs need to be made. Unfortunately, hard-coding such trade-offs can have undesirable consequences, and model-selecting them is a cumbersome process when multiple objectives are involved.

To avoid the need for *a priori* trade-offs, *multi-objective learning* has recently received increasing attention. Using *multi-objective optimization*, it either finds promising trade-off parameters at the same time as training the actual model, or it computes multiple solutions that reflect different trade-offs, ideally along the complete *Pareto-front*.[1] While multi-objective optimization and learning are algorithmically rich fields, their theory is much less well explored. In particular, learning-theoretic results, such as generalization bounds, are almost completely missing.

In this work, we aim at putting multi-objective learning on solid theoretic foundations. Specifically, we present three results of fundamental nature for understanding the properties of learning with multiple objectives. (1) We show that generalization bounds of individual learning objectives carry over also to the situation when learning with multiple objectives simultaneously. (2) We provide generalization and excess bounds that hold uniformly across a broad range of *scalarization* techniques. (3) We analyze in what sense the set of models that are *empirically Pareto-optimal* (i.e., optimal with respect to a training set) approximates the set of models that are actually *Pareto-optimal* (i.e., optimal with respect to the data distribution). Our results provide theoretical justifications for the use of

Peter Súkeník and Christoph Lampert have contributed equally to this work.

✉ Peter Súkeník
peter.sukenik@ista.ac.at

[1] Institute of Science and Technology Austria (ISTA), Am Campus 1, 3400 Klosterneuburg, Austria

[1] We define the technical terms *Pareto-front Pareto-optimal* and *scalarization* in Sect. 2.

scalarization-based as well as Pareto-based multi-objective optimization in a learning context, though with some caveats that have no analog in single-objective learning.

2 Notation and background

In this section, we introduce our notation and provide background information on multi-objective optimization and learning, as well as statistical learning theory. Our description follows standard textbooks, such as Miettinen [1] and Wright [2] for optimization, and Mohri et al. [3] and Shalev-Shwartz and Ben-David [4] for machine learning. More details and derivations can be found there.

2.1 Single- and multi-objective optimization

At the heart of most modern machine learning algorithms lies an optimization step. In standard (single-objective) optimization, one is given an input set, Ω, and an objective function, $f : \Omega \to \mathbb{R}$. Because the objective values are just real numbers, they are totally ordered: Any two points $\omega, \omega' \in \Omega$ are *comparable* in the sense that at least one of the relations $f(\omega) \leq f(\omega')$ or $f(\omega') \leq f(\omega')$ holds. Consequently, it is a natural question to ask which $\omega^* \in \Omega$ achieve the smallest objective value, if any. A plethora of *single-objective optimization* methods have been developed to answer this question, let it be *gradient-based* [2, 5] or *derivative-free* [6, 7].

In *multi-objective optimization*, one is given multiple objective functions, $f_1, f_2, \ldots, f_N : \Omega \to \mathbb{R}$, or equivalently, one vector-valued function, $F : \Omega \to \mathbb{R}^N$ with $F(\omega) = \big(f_1(\omega), \ldots, f_N(\omega)\big)$. We can again define an associated order relation:

Definition 1 For $\omega, \omega' \in \Omega$ we say that ω *weakly dominates* ω' if $f_j(\omega) \leq f_j(\omega')$ for all $j \in [N]$. We say ω *strongly dominates* ω' if additionally $f_j(\omega) < f_j(\omega')$ for at least one $j \in [N]$.

Because of the multi-dimensional nature, these orderings are only partial [?]. There are pairs $\omega, \omega' \in \Omega$ that are *uncomparable*, i.e., neither $F(\omega) \preccurlyeq F(\omega')$, nor $F(\omega') \preccurlyeq F(\omega)$ holds. Consequently, in multi-objective optimization it typically makes no sense to look for absolute *best* solutions. Instead, one searches for *Pareto-optimal* solutions.

Definition 2 A point $\omega^* \in \Omega$ is called *Pareto-optimal* if there is no other point $\omega \in \Omega$ that *strongly dominates* it. The set of all Pareto-optimal points is called *Pareto-optimal set*. The set of corresponding objective value vectors is called *Pareto-front*.

A large number of algorithms have been developed also for multi-objective optimization. When trying to find solutions across the complete Pareto-front, meta-heuristics such as *evolutionary algorithms* [8] are often employed. If a single Pareto-optimal solution suffices, *scalarizations* in combination with single-objective optimization can be used [9]. A *scalarization* function, $\mathcal{U} : \mathbb{R}^N_+ \to \mathbb{R}_+$, combines the individual objective values into a single one. Prominent examples are weighted p-norms: $\mathcal{U}^{(p)}_w(x_1, \ldots, x_N) = \left(\sum_{i \in [N]} |w_i x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$, and $\mathcal{U}^{(\infty)}_w(x_1, \ldots, x_N) = \max_{i \in [N]} |w_i x_i|$, where $w \in W \subset \mathbb{R}^N_+$ is a vector of weights that encode a trade-off between the different objectives.

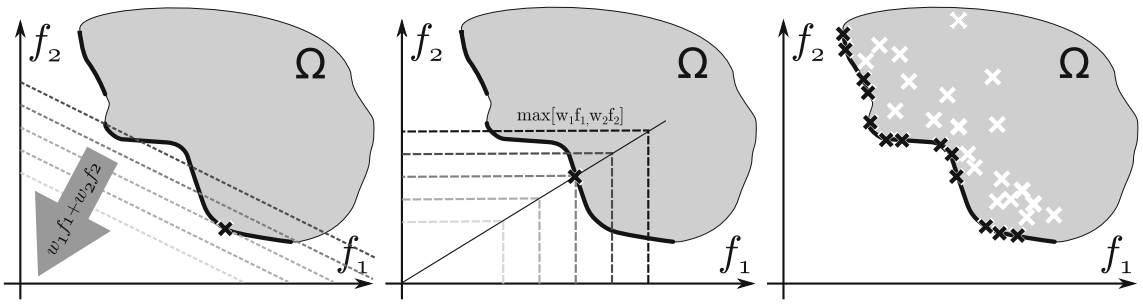
Arguably the most popular choice of scalarization is the L^1-norm with weights in the probability simplex $\Delta_N = \{w \in \mathbb{R}^N_+ : \sum_i w_i = 1\}$. This means, one forms *convex combinations* of the individual objectives [10]. For any nonzero choice of weights, minimizers of this resulting scalarized objective will be Pareto-optimal [9]. However, the set of solutions obtainable by varying the weights might not recover the complete Pareto-front, unless the optimization problem is convex [11]. In contrast, the choice $p = \infty$ (called *weighted Chebyshev norm*) allows recovering the complete Pareto-front when varying the weights in Δ_N [1, Chapter 3.4]. Figure 1 illustrates these concepts.

2.2 Single- and multi-objective learning

Our analysis in this work applies to supervised as well as unsupervised learning. Therefore, we adopt a notation that allows expressing both of these cases in a single concise way. Let $p(z)$ be a fixed but unknown data distribution over a data space $\mathcal{Z}$. We denote by $\mathcal{H}$ a *hypothesis set* and $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}_+$ a *loss function*. For supervised learning with $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$, one uses $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $\ell(z, h) = L(y, h(x))$, where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ measures, e.g., the classification or regression error. For unsupervised learning, one uses $\mathcal{Z} = \mathcal{X}$, and ℓ measures, e.g., the reconstruction error of a clustering or dimensionality reduction step.

Single-objective learning. Standard (single-objective) learning has the goal of identifying a hypothesis with small *risk* (*expected loss*), $\mathcal{L}(h) = \mathbb{E}_{z \in \mathcal{Z}}[\ell(z, h)]$. To approximate this uncomputable quantity, the learner uses a *training set*, $S = \{z_1, \ldots, z_n\}$ to compute the *empirical risk*, $\widehat{\mathcal{L}}(h) = \frac{1}{n} \sum_{i=1}^n \ell(z_i, h)$.

Statistical learning theory studies how well the empirical risk approximates the true risk and under which conditions minimizing the (computable) empirical risk is a good strategy for finding solution with low true risk. Many corresponding results are known. In particular, under well-



(a) Linear scalarization (b) Chebyshev scalarization (c) Ensemble method

Fig. 1 For general multi-objective optimization problems the Pareto-front (bold curves) can be disconnected and non-convex. **a** *Linear (convex) scalarization* can find Pareto-optimal solutions on the convex hull of the front. **b** *Chebyshev scalarization* can find solutions everywhere on the front. **c** *Ensemble methods* compute many solutions, aiming for the complete Pareto-front to be represented

understood conditions on $\mathcal{H}$ and S, one can prove that, with high probability over the sampling of S, the true risk is well approximated by the empirical risk, uniformly across all hypotheses. Mathematically, such a guarantee has the form of a *generalization bound*:

$$\forall \delta \in (0,1) \quad \Pr\Big\{\forall h \in \mathcal{H} : |\mathcal{L}(h) - \widehat{\mathcal{L}}(h)| \leq \mathcal{C}(n, \mathcal{H}, \delta)\Big\} \geq 1 - \delta. \tag{1}$$

The problem-dependent *generalization term* $\mathcal{C}(n, \mathcal{H}, \delta)$ typically consists of a *complexity* component that reflects the expressive power of the hypothesis class, and a *confidence* component that reflects the uncertainty due to finite sampling effects. Ideally, both components will converge to 0 when the number of samples grows to infinity. Among the most well-known complexity components are for instance the Rademacher complexity or VC dimension (see, e.g., [3]).

From bounds of the form (1), one can derive guarantees that, with high probability, solutions obtained by minimizing the empirical risk have close to optimal true risk. Formally, for $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{L}}(h)$, an *excess risk bound* holds:

$$\forall \delta \in (0,1) \quad \Pr\Big\{\mathcal{L}(h^*) \leq \inf_{h \in \mathcal{H}} \mathcal{L}(h) + \mathcal{C}'(n, \mathcal{H}, \delta)\Big\} \geq 1 - \delta, \tag{2}$$

where $\mathcal{C}'(n, \mathcal{H}, \delta)$ is another generalization term as above.

Multi-objective learning. In multi-objective learning, multiple target objectives, $\mathcal{L}_1, \ldots, \mathcal{L}_N$, characterize different properties of interest of the hypotheses. Estimating them from a (single) dataset yields empirical objectives, $\widehat{\mathcal{L}}_1, \ldots, \widehat{\mathcal{L}}_N$. In contrast to the single-objective situation where the objective function is almost always related to a measure of prediction quality, the multi-objective setting provides a principled framework for expressing also other relevant quantities of a machine learning model, such as *efficiency*, *robustness* or *fairness*. Consequently, we allow the objectives to also have other forms than just expected values over per-sample loss functions, and their empirical estimates are not restricted to per-sample averages. As discussed in Sect. 2.1, the multi-objective setting does not induce a total ordering of the hypotheses. Consequently, *a priori* there will be no overall *best* hypothesis anymore. Instead, there are two sets of Pareto-optimal hypotheses:

Definition 3

a) A hypothesis $h \in \mathcal{H}$ is called *empirically Pareto-optimal* if it is Pareto-optimal with respect to the multi-objective optimization problem of minimizing $\widehat{\mathcal{L}}_1(h), \ldots, \widehat{\mathcal{L}}_N(h)$ (with are computed from some training set S). The set of all such hypotheses we call the *empirically Pareto-optimal set*.

b) A hypothesis $h \in \mathcal{H}$ is called (*truly*) *Pareto-optimal* if it is Pareto-optimal with respect to the multi-objective optimization problem of minimizing $\mathcal{L}_1(h), \ldots, \mathcal{L}_N(h)$. The set of all such hypotheses we call the (*truly*) *Pareto-optimal set*.

Analogously to single-objective learning, we are most interested in finding truly optimal hypotheses (here, e.g., the truly Pareto-optimal set), as these can be expected to work well on future data. However, we can only compute solutions to the empirical problem (the empirically Pareto-optimal set). If solutions to the latter problem approximate the former, it is called *multi-objective generalization*.

In recent years, multi-objective learning has received increasing attention in the machine learning community, and a number of algorithms have been proposed for it. In their easiest form, one simply picks a scalarization method and solves the resulting single-objective optimization problem with fixed scalarization weights or one optimizes over those as well [12–14]. Alternatively, one can search for hypotheses along the complete (empirically) Pareto-

front, using, e.g., ensemble techniques [15, 16], model conditioning [17] or hypernetworks [18].

Given the long tradition and algorithmic diversity, one could expect *multi-objective statistical learning theory* also to be a rich field that provides precise quantifications of the relations between true and empirical objective (generalization bounds), as well as relation between the empirical and true Pareto-optimal sets (excess bounds). Surprisingly, this is not the case, and hardly any such results exist in the literature.

3 Related work

Solving problems with multiple objectives has a long tradition in artificial intelligence [19–22], game theory [23, 24] and economics [25, 26]. Since the 1990 s it has also attracted attention from the machine learning community, e.g., Fieldsend and Singh [27], Goldberg [28], Jin [29]. Existing works predominantly study the problem from an algorithmic perspective, in particular proposing and analyzing new optimization techniques. Mirroring the corresponding developments in multi-objective optimization, this includes methods for efficiently finding individual Pareto-optimal solutions, e.g., Cortes et al. [12], Van Moffaert and Nowé [30], Ye et al. [31], as well as exploring the complete Pareto-front [17, 18, 30, 32–35]. Works in both directions implicitly assume that better results of the empirical learning task should translate to better results on future data. So far, this *generalization* aspect was studied only empirically. Theoretical results rather focused on the optimization aspect, e.g., studying *computational complexity* [36, 37] or *convergence rates* [38], but not statistical generalization. A notable exception is Cortes et al. [12], which we discuss in detail in the Sect. 5.3.

4 Main results

In this section, we formally state and discuss our main results: generalization and excess bounds for scalarizations and for Pareto-fronts. For maximal generality, we formulate the results on the generic level introduced in Sect. 2. We will discuss instantiations that either improve over related existing work or provide new insights in Sects. 5 and we provide a high-level overview of potential additional applications in Sect. 6.

Assumptions. Because the multi-objective setting strictly generalizes the single-objective one, multi-objective generalization is not possible unless at least single-objective generalization holds. Therefore, for all our results we adopt the following assumption.

Assumption A For each objective $i \in [N]$ individually, a generalization bound of the form (1) with generalization term $\mathcal{C}_i(n, \mathcal{H}, \delta)$ holds.

Note that Assumption A is technically easy to fulfill, at least for bounded objectives, by setting the required generalization terms, $\mathcal{C}_i(n, \mathcal{H}, \delta)$ for $i \in [N]$, to large enough constants. Our results do hold for such a choice, but their interpretation would mostly not be very interesting. Therefore, whenever we want to interpret results in the light of their approximation quality, we additionally make the following assumption.

Assumption B For each $i \in [N]$ and for each $\delta \in (0, 1)$, it holds that $\mathcal{C}_i(n, \mathcal{H}, \delta) \overset{n\to\infty}{\to} 0$.

As we detail in Sect. 6, Assumption A and Assumption B are fulfilled for many quantities of interest related to the *accuracy*, *fairness*, *robustness* or *efficiency* of machine learning systems. Noteworthy special cases are objectives that are data-independent functions of only the hypothesis, for example, regularization terms. We say that such objectives *generalize trivially*, because they fulfill $\mathcal{L}(h) = \widehat{\mathcal{L}}(h)$ for all datasets and all $h \in \mathcal{H}$, and therefore, generalization bounds of the form (1) hold for them trivially with 0 as generalization term.

4.1 Multi-objective generalization

Our first result states that if generalization bounds hold individually for each objective, then they hold also jointly in the multi-objective setting, where the empirical objectives are computed from a single dataset, at only a minor loss of confidence.

Lemma 1 (*Multi-Objective Generalization Bound*) *Let N_{nt} be the number of non-trivial objectives. Let S be a random dataset of size n. For each $i \in [N]$, let $\widehat{\mathcal{L}}_i$ be an empirical estimate of $\mathcal{L}_i$ based on a subset $S_i \subset S$ of size n_i. The sets S_i are arbitrary. In particular, they are not necessarily disjoint. Then it holds with probability at least $1 - \delta$,*

$$\forall i \in [N],\ \forall h \in \mathcal{H} : |\mathcal{L}_i(h) - \widehat{\mathcal{L}}_i(h)| \leq \mathcal{C}_i(n_i, \mathcal{H}, \delta / N_{\text{nt}}). \tag{3}$$

Lemma 1 is in fact a straightforward consequence of Assumption A, requiring only a union-bound argument as proof. We state it explicitly nevertheless because it has not appeared in this form in the literature so far.

4.2 Generalization and excess bounds for scalarizations

A common way for learning in a multi-objective setting is by performing single-objective learning for one or multiple scalarizations. To keep the notation concise, for any scalarization $\mathcal{U} : \mathbb{R}_+^N \to \mathbb{R}_+$ and $h \in \mathcal{H}$, we abbreviate $\mathcal{L}_\mathcal{U}(h) := \mathcal{U}(\mathcal{L}_1(h), \ldots, \mathcal{L}_N(h))$, $\widehat{\mathcal{L}}_\mathcal{U}(h) := \mathcal{U}(\widehat{\mathcal{L}}_1(h), \ldots, \widehat{\mathcal{L}}_N(h))$. We also remind the reader that a norm $\|\cdot\|$ in $\mathcal{R}^N$ is *monotonic*, if for all $x, y \in \mathcal{R}^N$ if $|x_i| \leq |y_i| \forall i \in [N]$, then $\|x\| \leq \|y\|$.

Theorem 2 (*Generalization and Excess Bounds for Scalarizations*) *Assume the same setting as for Lemma 1. Let $\mathfrak{U} = \{\mathcal{U} : \mathbb{R}^N \to \mathbb{R}_+\}$ be a set of scalarizations, each of which is $\lambda_\mathcal{U}$-Lipschitz continuous with respect to some monotonic norm $\|\cdot\|_\mathcal{U}$. Then, for all $\delta > 0$ the following two statements hold with probability at least $1 - \delta$.*

(a) For all $\mathcal{U} \in \mathfrak{U}$ and $h \in \mathcal{H}$:

$$\left|\mathcal{L}_\mathcal{U}(h) - \widehat{\mathcal{L}}_\mathcal{U}(h)\right| \leq \lambda_\mathcal{U} \left\| \left(\mathcal{C}_1(n_1, \mathcal{H}, \delta/N_{\text{nt}}), \ldots, \mathcal{C}_N(n_N, \mathcal{H}, \delta/N_{\text{nt}})\right) \right\|_\mathcal{U}. \tag{4}$$

(b) For all $\mathcal{U} \in \mathfrak{U}$, for all $\hat{h}_\mathcal{U}^* \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{L}}_\mathcal{U}(h)$ and for all $h \in \mathcal{H}$:

$$\mathcal{L}_\mathcal{U}(\hat{h}_\mathcal{U}^*) \leq \mathcal{L}_\mathcal{U}(h) + 2\lambda_\mathcal{U} \left\| \left(\mathcal{C}_1(n_1, \mathcal{H}, \delta/N_{\text{nt}}), \ldots, \mathcal{C}_N(n_N, \mathcal{H}, \delta/N_{\text{nt}})\right) \right\|_\mathcal{U}. \tag{5}$$

Proof sketch. We provide the main arguments of the proofs here. The complete proofs are provided in Appendix A. a) The Lipschitz property implies that the difference of scalarized objectives is upper bounded by the norm of the differences in objective values. By the norm's monotonicity and Lemma 1, this is again bounded by the norm of the generalization terms. b) from $\widehat{\mathcal{L}}_\mathcal{U}(\hat{h}_\mathcal{U}^*) \leq \widehat{\mathcal{L}}_\mathcal{U}(h)$ it follows that $\mathcal{L}_\mathcal{U}(\hat{h}_\mathcal{U}^*) - \mathcal{L}_\mathcal{U}(h) \leq \mathcal{L}_\mathcal{U}(\hat{h}_\mathcal{U}^*) - \widehat{\mathcal{L}}_\mathcal{U}(\hat{h}_\mathcal{U}^*) + \widehat{\mathcal{L}}_\mathcal{U}(h) - \mathcal{L}_\mathcal{U}(h)$. Using a) we can bound the difference between the first two terms as well as the difference between the last two terms on the right-hand side each by the norm of the generalization terms.

Remark. We formulate the theorem with scalarization-dependent norms $\|\cdot\|_\mathcal{U}$ because in some scenarios using a single (e.g., Euclidean) norm for all scalarizations would yield sub-optimal bounds. For instance, if $\mathcal{U}(x) = \sum_{i=1}^N w_i |x_i|$, then this is itself a norm and it is optimal to choose $\|\cdot\|_\mathcal{U} \equiv \mathcal{U}$ to get the tightest bound.

Discussion. Theorem 2 establishes *generalization* and *excess bounds* for the situation of scalarization-based multi-objective learning. Their relevance lies not only in the inequalities (4) and (5) themselves, which have the standard single-objective form, but also in the fact that these hold *uniformly* over all scalarizations $\mathcal{U} \in \mathfrak{U}$. This implies that one can solve an arbitrary number of scalarized problems without suffering a loss of confidence in the theoretical guarantees. That is in contrast to other situations of repeated learning, e.g., hyperparameter tuning on a validation set, where the statistical guarantees deteriorate with the number of hypotheses considered, because of the *multiple hypothesis testing* phenomenon [4, Chapter 11]. Despite its simplicity, the theorem improves over prior work, Cortes et al. [12], which proved guarantees that depend on the size of $\mathfrak{U}$. For a more detailed discussion, see Sect. 5.3.

4.3 Pareto excess bounds

We now provide a formal analysis of the relation between the set of Pareto-optimal hypotheses and the set of empirically Pareto-optimal hypotheses. First, we show that any two elements of the two Pareto-optimal sets fulfill an excess-type inequality with respect to at least some of the objectives.

Theorem 3 *Assume the same situation as for Lemma 1. Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$: For all Pareto-optimal $h^* \in \mathcal{H}$ and empirically Pareto-optimal $\hat{h}^* \in \mathcal{H}$ there exists a non-empty subset $I \subset [N]$, such that*

$$\forall i \in I : \quad \mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}). \tag{6}$$

Proof sketch. The proof works by contradiction: assume that a pair $(h^*, \hat{h*})$ exists such that for no index set inequality (6) would hold. Then, using Lemma 1, one could show that h^* strongly dominates $\hat{h}^*$ with respect to the empirical objectives, which is a contradiction to the optimality of $\hat{h}^*$. For the formal steps, see Appendix A. Like the sketch, the formal proof does not actually make use of the optimality of h^*. This implies that Theorem 3 holds in fact for all $h \in \mathcal{H}$, making it even more apparent that excess bound with respect to individual objectives are of limited use for studying multi-objective generalization.

For multi-objective learning, the most relevant question is if there is an analog of Theorem 3 for the case of $I = [N]$, i.e., if by finding the empirical Pareto-curve one also approximately recovers the true Pareto-curve with respect to *all* objectives. This is formalized in the following theorem.

Theorem 4 (*Pareto Excess Bound*) *Assume the same setting as for Lemma 1. Then, for any* $\delta > 0$, *it holds with probability at least* $1 - \delta$.

(a) For all Pareto-optimal $h^* \in \mathcal{H}$ there exists an empirically Pareto-optimal $\hat{h}^* \in \mathcal{H}$ with

$$\forall i \in [N]: \quad \mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}). \tag{7}$$

(b) Assume that the Pareto-front is *ray complete*, i.e., for all $R \in \{(r_1, \ldots, r_N) : r_i > 0 \text{ for } i \in [N]\}$, there exists an $h \in \mathcal{P}$ with $\big(\mathcal{L}_1(h), \ldots, \mathcal{L}_N(h)\big) \propto R$. Then, for all empirically Pareto-optimal $\hat{h}^* \in \mathcal{H}$, there exists a Pareto-optimal $h^* \in \mathcal{H}$ with

$$\forall i \in [N]: \quad \mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}). \tag{8}$$

Proof sketch. To prove part a), we make use of the fact that $h^* \in \mathcal{H}$ is dominated with respect to the empirical objectives by some empirically Pareto-optimal $h^* \in \mathcal{H}$, i.e., $\widehat{\mathcal{L}}_i(\hat{h}^*) \leq \widehat{\mathcal{L}}_i(h^*)$ for all $i \in [N]$. Statement (7) follows by applying Lemma 1 to both sides of this inequality and rearranging terms.

The main insight for proving part b) is that $\hat{h}^* \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{L}}_{\mathcal{U}}(h)$ for the Chebyshev scalarization $\mathcal{U}(x_1, \ldots, x_N) = \max_{j \in [N]} w_j x_j$ with weights $w_j = \frac{1}{\widehat{\mathcal{L}}_j(h^*)}$ for $j \in [N]$, as long as $\widehat{\mathcal{L}}_i(h^*) > 0$ for all $i \in [N]$. With $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{U}}(h)$ it follows from Theorem 2 that $\max_{j \in [N]} w_j \mathcal{L}_j(\hat{h}^*) \leq \max_{j \in [N]} w_j \mathcal{L}_j(h^*) + 2 \max_{j \in [N]} w_j \mathcal{C}_j(n_j, \mathcal{H}, \delta/N_{\text{nt}})$. For any $i \in [N]$, it holds that $w_i \mathcal{L}_i(\hat{h}^*) \leq \max_{j \in [N]} w_j \mathcal{L}_j(\hat{h}^*)$, and the assumption of ray completeness ensures that, $w_i \mathcal{L}_i(\hat{h}^*) = \max_{j \in [N]} w_j \mathcal{L}_j(\hat{h}^*)$. In combination, one obtains the same statement as (8), except with a potentially weaker generalization term $\frac{2}{w_i} \max_{j \in [N]} w_j \mathcal{C}_j(n_j, \mathcal{H}, \delta/N_{\text{nt}})$. To obtain the desired result, one creates additively shifted objective functions that result in a learning setting equivalent to the original one, but in which all terms $w_j \mathcal{C}_j(n_j, \mathcal{H}, \delta/N_{\text{nt}})$ for $j \in [N]$ are identical, such that, in particular, $\max_{j \in [N]} w_j \mathcal{C}_j(n_j, \mathcal{H}, \delta/N_{\text{nt}}) = w_i \mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}})$. For complete proofs, see Appendix A.

Discussion. The theorems in this section clarify the relation between the true Pareto-optimal set and its empirical counterpart. When looking a single objective at a time, the relation is nearly trivial: Theorem 3 establishes that any empirically Pareto-optimal hypothesis is not much worse than any truly Pareto-optimal hypothesis with respect to *at least one* of the objectives.

More interesting is the situation when studying all objectives simultaneously. Theorem 4 provides a multi-objective analog of the classical *empirical risk minimization principle* [39]. Solving the empirical multi-objective learning problem makes sense as a learning strategy, because for every truly Pareto-optimal hypothesis there is an empirically Pareto-optimal one that has not much larger (true) objective values, jointly across all of the objectives. Reversely, every empirically Pareto-optimal hypothesis is not substantially worse than some Pareto-optimal one, if we make an additional assumption on the geometry of the Pareto-front.

The employed *ray completeness* assumption is quite restrictive, and we do not expect it to be fulfilled in most real-world situations. For example, it is violated already whenever one of the objectives is bounded from below by a constant bigger than 0.

In the two-objective situation, ray completeness does hold if the Pareto-front is a continuous curve between some point on the $\mathcal{L}_1$-coordinate axis and some point on the $\mathcal{L}_2$-coordinate axis, excluding the origin. An example where such a situation can happen is a classification task in the realizable setting with *classification error* and (suitably defined) *computational cost* as objectives. For a sufficiently rich hypothesis set, the smallest achievable error will be a continuous and monotonically decreasing function of the specified computational budget. Consequently, the Pareto-front will be a continuous curve between a point $(a, 0)$, where a is the classification error of the classifier with minimal budget, and a point $(0, b)$, where b is the smallest computational cost for a classifier achieving minimal classification error. Note that realizability is necessary. Otherwise, the curve would still be monotonic, but the second point in the above construction would not lie on the $\mathcal{L}_2$-coordinate axis. Consequently, ray completeness would not be fulfilled.

While sufficient, ray completeness is certainly not a necessary condition. For example, if both the real objective and the empirical objective are bounded away from zero by the same constant, substracting this constant from both objectives could yield a situation that is equivalent in terms of multi-objective learning, but in which ray completeness might be fulfilled. Furthermore, from the Theorem's proof one can see that a weaker condition would suffice, namely that every ray through the empirical Pareto-front also intersects the actual Pareto-front. This formulation would complicate the condition, though, as it introduces a dependence on the dataset and would nevertheless still not be a mathematically necessary condition. Therefore, we leave the task of identifying a condition that is necessary as well as sufficient to future work.

The following theorem shows that some additional assumption is required for Theorem 4 to hold.

Theorem 5 *Let $N \geq 2$. Then, for any $C > 0$ there exists a learning problem that fulfills Assumptions A and B with $\mathcal{C}_i(n_i, \mathcal{H}, \delta) = 0$ for $i \in [N-1]$, but for which with probability at least $\frac{1}{2}$ there exists an empirically Pareto-optimal $\hat{h}^* \in \mathcal{H}$, such that for all Pareto-optimal $h^* \in \mathcal{H}$, it holds*

$$\forall i \in [N-1]: \quad \mathcal{L}_i(\hat{h}^*) > \mathcal{L}_i(h^*) + C. \tag{9}$$

Proof We prove the theorem by constructing a concrete counterexample that exploits the classic *overfitting* (or *bias-variance trade-off*) phenomenon of single-objective supervised learning [39].

First, we look at the case $N = 2$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \{0, 1\}$, $p(x)$ be the uniform distribution and $p(y|x) = \frac{1}{2}$. Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ be the set of piecewise-constant functions that consist of at most K segments. We choose the number of jumps as $\mathcal{L}_1$ and the 0/1-loss as $\mathcal{L}_2$. Then, Assumption A and Assumption B are fulfilled: $\mathcal{H}$ is known to have VC-dimension $2k$ [4], so a classical generalization bound holds for $\mathcal{L}_2$. $\mathcal{L}_1$ even generalizes trivially.

We observe that every hypothesis in $h \in \mathcal{H}$ fulfills $\mathcal{L}_2(h) = \frac{1}{2}$. Consequently, the two Pareto-optimal solutions, h, are the constant classifiers, which fulfill $\mathcal{L}_2(h) = \frac{1}{2}$ and $\mathcal{L}_1(h) = 0$.

Empirically, however, for sufficiently many points, with high probability, the empirical loss $\widehat{\mathcal{L}}_2(h)$ will be strictly monotonically decreasing with respect to $\widehat{\mathcal{L}}_1(h)$, as more segments allow to better fit the training data. Consequently, the set of empirically Pareto-optimal solutions will contain elements with $\widehat{\mathcal{L}}_1(h) = k$ for any $k \in [K]$, i.e., arbitrarily far from all solutions in the truly Pareto-optimal set. Figure 2 shows a visualization of this situation.

For larger N, we use the analogous construction in $\mathbb{R}^{N-1}$. Hypotheses have at most K jumps in each coordinate dimension. Objectives 1 to $N-1$ are the number of jump per coordinate; objective N is the classification 0/1-loss. □

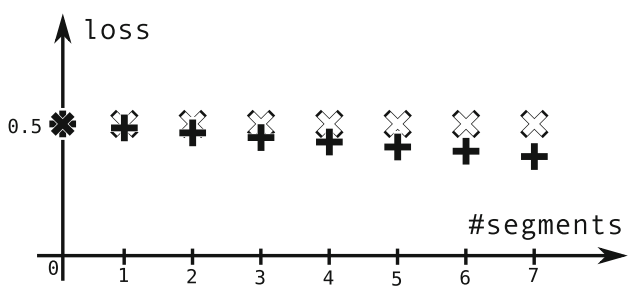


Fig. 2 Illustration of the counterexample proving Theorem 5 for $N = 2$

Discussion. Theorem 5 establishes that the asymmetry between the statements a) and b) of Theorem 4 is an intrinsic property of the multi-objective setting, not a limitation of our proof techniques. There can indeed be hypotheses in the empirically Pareto-optimal set that are not in an excess relation with any hypothesis in the truly Pareto-optimal set. Note that despite the fact that multi-objective learning includes single-objective learning as a special case, there is no contradiction to the classical symmetric result. For $N = 1$, the fact that $I \subset [N]$ is non-empty in Theorem 3 makes its statement identical to Theorem 4 b) without the additional assumption. Theorem 5 holds only for $N \geq 2$.

4.4 Summary

In combination, the results of this section establish a detailed picture of similarities and differences between the generalization properties of single-objective and multi-objective learning. In particular, it highlights a fundamental difference between single-objective learning and Pareto-based multi-objective learning: In the single-objective setting, empirical risk minimization is a good learning strategy, because with growing data, the every minimizer of the empirical risk also has close to optimal true risk. In the multi-objective setting, scalarized learning has the same properties, but the resulting guarantees hold only with respect to the scalarization of the objectives, not each of them individually. Joint statements across all objectives hold as well, thereby justifying Pareto-based learning. However, without additional assumptions excess guarantees can only be given for a subset of the empirical solutions.

5 Applications

Our results of the previous section provide new tools for analyzing learning tasks in which multiple, potentially competing, quantities are of simultaneous interest, such as *fairness*, *robustness*, *efficiency* and *interpretability*. So far, the generalization properties of these quantities have been studied either not at all, or only with task-specific tools. Similarly, we expect that *multi-task*, *multi-label* and *meta-learning*, as well as *AutoML* will potentially be able to benefit from the multi-objective view.

In the rest of this section, we sketch three exemplary applications. In Sect. 5.1 we show how our results on empirical versus true Pareto-optimality can provide new

insights for a well-known method. In Sect. 5.2 we demonstrate how our results on scalarized learning provide a simple and flexible way for constructing new single-objective generalization bounds. In Sect. 5.3, we improve an existing generalization bound for the multi-objective setting, which thanks to our results of Sect. 4 requires only a few lines of proof. For a more high-level discussion of other application scenarios, see Sect. 6.

5.1 Simultaneous feature selection and regression

The classical LASSO method [40] learns a linear regression function by solving the following regularized risk minimization problem

$$\min_{\beta\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i-\beta^\top x_i)^2+\lambda\|\beta\|_{L^1}. \tag{10}$$

Here $\{(x_1,y_1),\ldots,(x_n,y_n)\}\subset\mathbb{R}^d\times\mathbb{R}$ is a given training set and $\lambda\in\mathbb{R}^+$ is a free parameter that trades off the data fidelity on the training set with the sparsity of the coefficient vector β at its optimum. LASSO is particularly popular because it not only learns a regression function but also performs feature selection and therefore can give rise to more interpretable regression models than other regression techniques. The set of all solutions obtained by minimizing (10) for different values of λ is called the *solution path*. A number of efficient techniques for computing it have been developed [41–43].

We can interpret the LASSO problem equivalently as the linear scalarization of a two-objective learning problem. The first objective is the expected squared loss $\mathcal{L}_1(\beta)=\mathbb{E}_{(x,y)}(y-\beta^\top x)^2$, which has $\widehat{\mathcal{L}}_1(\beta)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta^\top x_i)^2$ as empirical counterpart. When the data and coefficient vector come from bounded domains, standard generalization bounds are known to hold that relate $\mathcal{L}_1$ and $\widehat{\mathcal{L}}_1$, see, e.g., Mohri et al. [3, Theorem 11.11]. The second objective is the regularizer, $\mathcal{L}_2(\beta)=\widehat{\mathcal{L}}_2(\beta)=\|\beta\|_{L^1}$, which generalizes trivially.

The (single-objective) generalization properties of LASSO's squared loss term are well understood. The multi-objective view, however, adds insight into its relation with the regularizer, which reflects the sparsity and thereby the interpretability of the solutions. First, we observe that the underlying optimization problem is convex, so each empirically Pareto-optimal solution can be recovered by solving (10) for some value of λ. Therefore, existing *solution path* methods can readily be used to identify the empirical Pareto-front.

Theorem 4 a) now ensures that each truly Pareto-optimal solution can be approximately recovered this way. This means, we can be sure that no solutions exist that are substantially sparser at equal accuracy or more accurate at identical sparsity with respect to the true objectives than some in the solution path.

However, Theorem 5 reminds us that not all solutions found by solving (10) will necessarily be close to truly Pareto-optimal. In particular, this means, while each individual element of the solution path will have optimal sparsity for the empirical accuracy it achieves, its sparsity might be far from optimal compared to other solutions of similar true accuracy. Consequently, if optimal sparsity is important for the task at hand, the solutions on the regularization path should be further evaluated, e.g., using validation data.

The latter comment does not apply if the true underlying regression task is actually linear, such that a coefficient vector, β, exists with vanishing objective, $\mathcal{L}_1(\beta)=0$. Because the same property holds for $\mathcal{L}_2$ (trivially achieved by $\beta=0$), and the regularization path is a connected set [44], the condition of *ray completeness* would be fulfilled. Theorem 4 b) then guarantees that in fact all empirically Pareto-optimal solutions are also approximately truly Pareto-optimal.

5.2 Tilted empirical risk minimization

Tilted empirical risk minimization (TERM) [45] has recently been proposed as a widely applicable technique for making learning problem more *robust* or more *fair*. In its group-based form, TERM consists of minimizing an exponentially weighted risk functional

$$\begin{aligned} J_t(f) &= \frac{1}{t}\log\left(\frac{1}{N}\sum_{i=1}^{N}e^{tR_i(f)}\right) \quad \text{with} \quad R_i(f) \\ &= \frac{1}{n}\sum_{(x,y)\in S_i}\ell(y,f(x)) \end{aligned} \tag{11}$$

for a loss function ℓ and training data given as N potentially overlapping groups, $S_1,\ldots,S_N$. For simplicity of exposition, we assume all groups to be of identical sizes, n. The tilt parameter $t\in\mathbb{R}\setminus\{0\}$ determines how much and in which direction the tilt influences the training process. Positive values of t make J_t behave like a *soft maximum* operation that puts more weight on large values of R_i than a plain average. For $t\to\infty$, TERM converges to $\max_i R_i(f)$. As a consequence, minimizing (11) with $t>0$ encourages *fair* solutions, in the sense that none of the groups should have a much larger values of R_i than the others.

Conversely, negative values of t make J_t behave like a *soft minimum* operation that puts less weight one large values of R_i. For $t\to-\infty$, TERM converges to $\min_i R_i(f)$. Such a behavior is desirable if some of the groups can be expected to be *outliers* with large values for R_i, even if the

others terms are small. Minimizing (11) with $t<0$ can be used to find solutions that are *robust* against such outlier groups.

Theorem 2 readily allows us to make statements about the generalization properties of TERM, not just for one value of t but all possible ones simultaneously. For any value of t, the functional J_t in (11) can be seen as a scalarization of empirical objectives $R_1, \ldots, R_N$. A direct check confirms that each such J_t is 1-Lipschitz with respect to the L^∞ norm. Assuming that generalization bounds hold for each individual group, then Theorem 2 guarantees that a generalization bound of the same structure holds also for J_t, simultaneously across all values of t.

Hierarchical TERM (hTERM) [45] uses the same exponentially weighted functional J_t, but the per-group terms $R_1, \ldots, R_K$ that it combines are not averages across samples as in (11), but TERM losses themselves with individual tilt parameters $\tau_1, \ldots, \tau_K$. From our previous analysis, we know generalization bounds for each of those. Consequently, by the same construction as above, we readily obtain a generalization bound for hTERM, which in fact holds uniformly across all combinations of tilt parameters.

5.3 Agnostic learning with multiple objectives

Cortes et al. [12] study the generalization properties of multi-objective learning for the case of a special scalarization obtained by minimizing over convex combinations. To allow for an easier comparison, we state their result in our notation.[2]

Theorem 6 (*Theorem 3 in Cortes et al. [12]*) *Let $\mathcal{H}$ be a hypothesis set for a supervised learning problem with input set $\mathcal{X}$ and output set $\mathcal{Y}$ that fulfills $\|(h(x'), y') - (h(x), y)\| \leq D$ for some constant $D > 0$ and for all $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$. Let $\ell_i : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ for $i \in [N]$ be loss functions that are M_i-Lipschitz and upper bounded by M. Set $\mathcal{L}_i(h) = \mathbb{E}_{(x,y)}[\ell(y, h(x)]$ and $\widehat{\mathcal{L}}_i(h) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, h(x_i))$ for a dataset $S \overset{i.i.d.}{\sim} p(x, y)$ of size n. For a set of scalarization weights, $W \subset \Delta_N$, let $\mathcal{L}_W(h) = \max_{w \in W} \sum_{i=1}^N w_i \mathcal{L}_i(h)$ and $\widehat{\mathcal{L}}_W(h) = \max_{w \in W} \sum_{i=1}^N w_i \widehat{\mathcal{L}}_i(h)$. Assume that $\sum_{i=1}^N w_i M_i \leq \beta$ for all $w = (w_1, \ldots, w_N) \in W$.*

Then, for any $\epsilon > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$:

$$\mathcal{L}_W(h) \leq \widehat{\mathcal{L}}_W(h) + 2\beta\hat{\mathfrak{R}}_S(\mathcal{H}) + M\epsilon + 3\beta D\sqrt{\frac{1}{2n}\log\left[\frac{2|W_\epsilon|}{\delta}\right]}, \quad (12)$$

where $\hat{\mathfrak{R}}_S(\mathcal{H})$ is the *empirical Rademacher complexity* of the hypothesis class $\mathcal{H}$ with respect to S, and $|W_\epsilon|$ is the size of a minimal ϵ-cover of W.

One can see that inequality (12) precisely matches the form of our excess bound in Theorem 2 for a specific scalarization. Indeed, we can derive an analogous theorem using our results of Sect. 4.

Theorem 7 *Make the same assumptions as in Theorem 6. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ the following inequality holds for all $h \in \mathcal{H}$:*

$$\mathcal{L}_W(h) \leq \widehat{\mathcal{L}}_W(h) + 2\beta\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\beta D\sqrt{\frac{\log\frac{2N}{\delta}}{2n}}. \quad (13)$$

Proof The Lipschitz assumptions imply standard Rademacher-based generalization bounds for each objective individually [3, Theorem 11.3]: With probability at least $1 - \delta$ it holds for all $h \in \mathcal{H}$:

$$\mathcal{L}_i(h) \leq \widehat{\mathcal{L}}_i(h) + 2M_i\hat{\mathfrak{R}}_S(\mathcal{H}) + 3M_i D\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \quad (14)$$

We now apply Theorem 2 for the family of linear scalarizations, $\mathcal{U}_w(x_1, \ldots, x_N) = \sum_{i=1}^N w_i x_i$ with $w \in W$, and we insert that $w_i M_i \leq \beta$. Finally, taking the maximum over w on both sides of the inequality yields (13). □

Because of the power of the introduced multi-objective framework, the proof of Theorem 7 is much shorter than the original one of Theorem 6. Nevertheless, our result has a number of advantages. First, our bound is structurally simpler. It holds without need for an ϵ-parameter that additively enters the right-hand side of (12), yet also influences the size of the rightmost confidence term. Second, the right-hand side of our bound is independent of the size of W, with the confidence term only depending on the number of objectives. As a consequence, our bound is substantially tighter, except for trivially small sets W. For example, for the common case of convex combinations, $W = \Delta_N$, the covering size $|W|_\epsilon$ is of order $(1/\epsilon)^{N-1}$. This makes the generalization term in (12) of order $\sqrt{N/n}$, indicating that to preserve confidence the amount of data has to grow linearly with the number of objectives considered. In contrast, the right-hand side of our bound (13) is independent of the size of W and its confidence term grows only logarithmically with respect to N. Finally, our proof is not only simpler than the original one but also more flexible: It readily extends to other generalization bounds

[2] Our formulation also has slightly different constants in the generalization term, which we believe to be necessary based on the theorem's proof. We multiply the rightmost term by 3 and the $|W_\epsilon|$ by 2.

rather than just Rademacher-based ones, and to other scalarization besides linear combinations.

6 Further application scenarios

In Sect. 5 we highlighted some specific examples in which our proposed multi-objective generalization theory provides new insights into existing methods. In this section, we provide more high-level background and discuss additional quantities that we believe will or will not benefit from a multi-objective analysis.

Fairness. *Algorithmic (group) fairness* asks to create classifiers that are not only accurate but also do not discriminate against certain protected groups in their decisions. Formally, this property can be expressed by different *(un)fairness measures*, such as *demographic parity*, *equality of opportunity* or *equalized odds* [46]. Because accuracy and fairness can be in conflict with each other, fairness-aware learning is a prototypical candidate for multi-objective learning [47–50]. This view also extends naturally to integration of multiple fairness measures [51], which might be incompatible with each other [52–54]. Generalization bounds for the empirical estimation of unfairness measures have been developed [55, 56]. Consequently, our results from Sect. 4 apply, yielding a unified understanding of the generalization properties of fairness-aware learning, e.g., regularization-based [57], constraint-based [58] or Pareto-based [18, 51]. The multi-objective view also allows us to conjecture that methods that seek fair hypotheses by other means, such as preprocessing [59] or post-processing [60], might not reach (empirically) Pareto-optimal solutions. If generalization guarantees do actually hold for these, other ways for proving them would be required.

Robustness. It has been observed that deep network classifiers in continuous domains such as image classification are susceptible to *adversarial examples*, i.e., they are not robust against small perturbation of the input data. Two main research directions have emerged to overcome this limitation: *Adversarial training* [61] adds a robustness-enforcing loss term to the training problem. Generalization bounds for such terms have been derived, e.g., Yin et al. [62]. Consequently, multi-objective learning can be used in this setting with the guarantees and caveats discussed above. *Lipschitz-networks* [63] restrict the hypothesis class to functions with a small Lipschitz constant, typically 1. Afterward, one solves a training problem that tries to enforce a large margin between the predicted class label and the runner-up. From the achieved margins, one can infer how large an input perturbation the classifier can tolerate without changing its decision [64]. We are not aware of existing theoretical studies of such *certified robustness* techniques. However, margin-based loss functions have a long tradition in machine learning, and a number of generalization bounds exist which are applicable in the described situation, such as Kuznetsov et al. [65], Koltchinskii and Panchenko [66].

Efficiency. Large machine learning models, in particular deep networks, often have high computational demands, not only at training but also at prediction time [67, 68]. Consequently, a number of techniques have been developed that aim at reducing the computational cost. *Parameter sparsification* [69] and *quantization* [70] are widely used methods for reducing the number of operations required to evaluate a model. As data-independent properties, they can readily be used as *trivially generalizing* objectives in a multi-objective learning framework [35]. Alternatively, speedup can also be achieved by encouraging as many zero values as possible to occur as part of the internal computation steps of a deep network. Such *activation sparsity* [71] is a data-dependent quantity that can also be shown to generalize using standard techniques. Therefore, it as well can be handled in a multi-objective way. *Adaptive computation* methods, such as *ensembles* [72, 73], *classifier cascades* [74] or *multi-exit architectures* [75, 76], evaluate different subsets of a larger model depending on the input sample. For suitable design choices, generalization bounds for the resulting computation time can be proven, and our results will apply.

Multi-task and multi-label learning. *Multi-task learning* has recently been put forward as a multi-objective task, where each task's loss is treated as a separate objective [77–80]. This setting is of a non-standard form, as each task typically has a dedicated training set, and objectives are not necessarily competing with each others [81]. Nevertheless, our framework can handle this setting as well, making use of the property that we allow the empirical estimates of different objectives to be derived from different subsets of the available data. Pareto-based guarantees are particularly relevant then, because at prediction time, for each sample one is interested in only one of the objectives, namely the one of the task to which this sample belongs. In the related problems of *multi-label learning* [82] and *extreme classification* [83], the goal is to predict multiple outputs (labels) for each sample. Each label has an associated classifier objective, and the losses are estimated either from the total dataset or from (typically overlapping) subsets [84]. Again, our framework is flexible enough to handle this setting. At prediction time, all labels are meant to be predicted, and the quality is typically judged by a task-dependent aggregate measure, making scalarization approaches of particular interest in this setting.

Limitations. Despite its generality, some multi-objective learning settings do not lend themselves to an analysis using our results. For example, in the *learning-to-rank* setting [85] solutions are typically judged by two measures: *precision* and *recall*. *A priori*, this makes it a promising setting for multi-objective analysis [86, 87]. Unfortunately, we are not aware of generalization bounds for the *precision* objective. Given that its value fluctuates heavily in the low-recall regime, it is in fact possible that Assumption B might not be fulfillable. Also in the context of ranking, two other common objectives are *true positive rate (TPR)* and *false positive rate (FPR)*, which together trace out the *receiver operating characteristic (ROC) curve*. TPRs and FPRs can summarized into a single value by the *area under the ROC curve (AUC)* [88], for which indeed generalization bounds have been derived [89]. However, the AUC is not a scalarization in the sense of Sect. 2.1, so Theorem 2 does not apply to it. Finally, besides the uniform generalization bounds of Assumption A, other guarantees of generalization have been developed, e.g., based on PAC-Bayesian theory [90, 91], or *algorithmic stability* [92, 93]. We see no principled reasons why results similar to ours should not hold for such settings as well, but other techniques would be required that lie outside of the scope of this work.

7 Conclusion

In this work, we proved a number of foundational results for the generalization theory of multi-objective learning. In particular, we showed that generalization bounds for the individual objectives imply generalization and excess bounds for multi-objective learning using scalarizations. Our second main result is an analysis of the relation between the Pareto-optimal sets of the empirical and the true learning problem. This justifies the use of Pareto-based methods on empirical data to approximately find all truly Pareto-optimal solutions. However, there is a caveat that some of the solutions found might be close to Pareto-optimal ones only with respect to some of the objectives, not all of them.

We formulated our results on a high level of generality that applies not only to measures of per-sample prediction quality, for which generalization bounds were originally developed, but also many other quantities of interest for modern machine learning systems, such as *fairness*, *robustness* and *efficiency*. While initial results for some of these specific domains exist, we expect that more and stronger guarantees will be possible by more refined objective-specific analyses.

On a technical level, we see two directions for potentially improving our results. First, it would be desirable to have an explicit rather than implicit relationship between Pareto-optimal hypotheses and their best empirically Pareto-optimal approximations. Theorem 4 does not provide this. Even though its proof contains an explicit procedure, it relies on uncomputable quantities, such as the true objective values. Second, given that Theorem 5 establishes that there can be empirically Pareto-optimal hypotheses that do not approximate any truly Pareto-optimal hypothesis with respect to all objectives, it would be desirable to have an algorithmic procedure for testing which hypotheses these are. We see these as interesting directions for future work.

Appendix A

Appendix—proofs of the main results

Proof of Theorem 2 With probability at least $1-\delta$ for the dataset S, the relations of Lemma 1 will hold. By studying only these cases, we again obtaining results that hold with probability at least $1-\delta$.

For statement a), for any $\mathcal{U} \in \mathfrak{U}$ we obtain by the Lipschitz property of the scalarization and Lemma 1 that for all $h \in \mathcal{H}$:

$$|\mathcal{L}_{\mathcal{U}}(h) - \widehat{\mathcal{L}}_{\mathcal{U}}(h)| \leq \lambda_{\mathcal{U}} \| \big(\mathcal{L}_1(h) - \widehat{\mathcal{L}}_1(h), \ldots, \mathcal{L}_N(h) - \widehat{\mathcal{L}}_N(h)\big) \|_{\mathcal{U}} \tag{A1}$$

$$\leq \lambda_{\mathcal{U}} \| \big(|\mathcal{L}_1(h) - \widehat{\mathcal{L}}_1(h)|, \ldots, |\mathcal{L}_N(h) - \widehat{\mathcal{L}}_N(h)|\big) \|_{\mathcal{U}} \tag{A2}$$

$$\leq \lambda_{\mathcal{U}} \| \big(\mathcal{C}_1(n_1, \mathcal{H}, \delta/N_{\mathrm{nt}}), \ldots, \mathcal{C}_N(n_N, \mathcal{H}, \delta/N_{\mathrm{nt}})\big) \|_{\mathcal{U}} \tag{A3}$$

where the last two inequalities hold because of the norms' monotonicity, i.e., the fact that it is non-decreasing under increases of the input vector components [94]. In combination, this proves statement a).

Statement b) follows by arguments mirroring the proof of classic *excess risk bounds* [3]. Let $\hat{h}^*_{\mathcal{U}} \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{L}}_{\mathcal{U}}(h)$. Then, it holds for arbitrary $h \in \mathcal{H}$ that

$$\mathcal{L}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) - \mathcal{L}_{\mathcal{U}}(h) \leq \mathcal{L}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) - \widehat{\mathcal{L}}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) + \widehat{\mathcal{L}}_{\mathcal{U}}(h) - \mathcal{L}_{\mathcal{U}}(h) \tag{A4}$$

$$\leq 2 \| \big(\mathcal{C}_1(n_1, \mathcal{H}, \delta/N_{\mathrm{nt}}), \ldots, \mathcal{C}_N(n_N, \mathcal{H}, \delta/N_{\mathrm{nt}})\big) \|_{\mathcal{U}} \tag{A5}$$

where the first inequality holds because $\widehat{\mathcal{L}}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) \leq \widehat{\mathcal{L}}_{\mathcal{U}}(h)$ by construction of $\hat{h}^*_{\mathcal{U}}$, and the second inequality one follows from applying (4) twice, once for h and once for $\hat{h}^*_{\mathcal{U}}$.

The statement of the theorem now follows by moving the term containing h to the right-hand side.

Proof of Theorem 3 We again only study the case in the inequalities of Lemma 1 are fulfilled, so the results we achieve hold with probability at least $1-\delta$.

We prove the remaining part of the theorem by contradiction. The negation of the statement reads: *there exists an empirically Pareto-optimal hypothesis* $\hat{h}^* \in \mathcal{H}$ *and a hypothesis* $h \in \mathcal{H}$ *such that* $\mathcal{L}_i(\hat{h}^*) - \mathcal{L}_i(h) > 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}})$ *for all* $i \in [N]$.

For these $\hat{h}^*$ and h, it follows that for all $i \in [N]$:

$$\widehat{\mathcal{L}}_i(h) - \widehat{\mathcal{L}}_i(\hat{h}^*) \leq \mathcal{L}_i(h) - \mathcal{L}_i(\hat{h}^*) + 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}) < 0. \tag{A6}$$

For the first inequality, we applied Lemma 1 twice, and the second inequality follows from the assumption. However, (A6) establishes that h empirically strongly dominates $\hat{h}^*$ which is a contradiction to the assumption that $\hat{h}^*$ was empirically Pareto-optimal. □

Proof of Theorem 4 We again only study these case in which the dataset fulfills the inequalities of Lemma 1, so the results we achieve holds with probability at least $1-\delta$.

Statement a) is a consequence of Lemma 1 and the definition of (empirical) Pareto-optimality. Let $h^* \in \mathcal{H}$ be Pareto-optimal. If it is also empirically Pareto-optimal, inequality (7) holds trivially with $\hat{h}^* = h^*$. Otherwise, there exists an empirically Pareto-optimal $\hat{h}^*$ that dominates h^* with respect to the empirical objectives, i.e., in particular $\widehat{\mathcal{L}}_i(\hat{h}^*) \leq \widehat{\mathcal{L}}_i(h^*)$ for all $i \in [N]$. From this, we obtain for all $i \in [N]$, analogously to the proof of Theorem 2b):

$$\begin{aligned}\mathcal{L}_i(\hat{h}^*) - \mathcal{L}_i(h^*) &\leq \mathcal{L}_i(\hat{h}^*) - \widehat{\mathcal{L}}_i(\hat{h}^*) + \widehat{\mathcal{L}}_i(h^*) - \mathcal{L}_i(h^*) \\ &\leq 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}).\end{aligned} \tag{A7}$$

Before proving statement b), we introduce *additively shifted objectives* as the main tool.

Definition 4 For an objective $\mathcal{L}(h)$ with empirical estimate $\widehat{\mathcal{L}}(h)$ and a constant K, we call $\mathcal{L}^{+K}(h) = \mathcal{L}(h) + K$ and $\widehat{\mathcal{L}}^{+K}(h) = \widehat{\mathcal{L}}(h) + K$ their K-additively shifted variants.

Generalization and Pareto-optimality are unaffected by additive shifts.

Lemma 8

a) For any constant K, if a generalization bound of the form (1) holds for an objective $\mathcal{L}$ and its empirical estimate $\widehat{\mathcal{L}}$, then a bound with identical generalization term also holds for $\mathcal{L}^{+K}$ and $\widehat{\mathcal{L}}^{+K}$.

b) For any constants $K_1, \ldots, K_N$, a solution $h \in \mathcal{H}$ is Pareto-optimal for $\mathcal{L}_1, \ldots, \mathcal{L}_N$ if and only if it is Pareto-optimal for $\mathcal{L}_1^{+K_1}, \ldots, \mathcal{L}_1^{+K_N}$. The analogous relation holds for empirically Pareto-optimality.

The proofs are elementary: For a), the additive terms cancel out in the generalization bound. For b) Pareto-optimality depends only on the relative order of objective values, which is not affected by additive shifts.

Lemma 9 *Let* $h^* \in \mathcal{H}$ *be a Pareto-optimal solution with* $\mathcal{L}_i(h) > 0$ *for all* $i \in [N]$. *Then* h^* *is a minimizer to the Chebyshev scalarization* $\mathcal{U}_w^{(\infty)}(h) = \max_{i\in[N]} w_i \mathcal{L}_i(h)$ *with weights* $w_i = \frac{1}{\mathcal{L}_i(h^*)}$ *for* $i \in [N]$. *Furthermore, for any other minimizer,* $h^\dagger$, *of the scalarization it holds that* $\mathcal{L}_i(h^\dagger) = \mathcal{L}_i(h^*)$ *for all* $i \in [N]$. *The analogous result holds for empirically Pareto-optimal hypotheses.*

Proof We prove the lemma by contradiction. First, assume h to be a hypothesis with strictly smaller value for the scalarization. By construction $w_i \mathcal{L}_i(h^*) = 1$ for all $i \in [N]$, therefore $w_i \mathcal{L}_i(h) < 1$ for all $i \in [N]$ must hold. This, however, would imply $\mathcal{L}_i(h) < \mathcal{L}_i(h^*)$ for all $i \in [N]$, which is impossible because h^* is Pareto-optimal. For $h^\dagger$, we know $w_i \mathcal{L}_i(h^\dagger) \leq 1$ and therefore $\mathcal{L}_i(h^\dagger) \leq \mathcal{L}_i(h^*)$ for all $i \in [N]$. Because of h^*'s Pareto-optimality, none of these inequalities can be strict, which proves the statement. The same line of arguments holds in the empirical situation. □

We now turn to the proof of Theorem 4 b). Let $\hat{h}^*$ be an empirically Pareto-optimal solution for $\mathcal{L}_1, \ldots, \mathcal{L}_N$. For a more concise notation, we abbreviate $c_i = \mathcal{C}_i(n, \mathcal{H}, \delta/N_{\text{nt}})$.

First, we consider the case where none of the objectives is trivially generalizing, i.e., $c_i > 0$ for all $i \in [N]$. By Lemma 8, we know that h^* is also empirically Pareto-optimal for the shifted objectives $\widehat{\mathcal{L}}_1^{+K_1}, \ldots, \widehat{\mathcal{L}}_{N'}^{+K_{N'}}$ with

$$\begin{aligned}K_i &:= Cc_i - \widehat{\mathcal{L}}_i(h^*) \quad \text{for } C \\ &= 2 + \max_j \left[\frac{1}{c_j}\left(\widehat{\mathcal{L}}_j(h^*) - \min_h \widehat{\mathcal{L}}_j(h)\right)\right]\end{aligned} \tag{A8}$$

An explicit calculation confirms that $\widehat{\mathcal{L}}_i^{+K_i}(h) \geq 2c_i > 0$, which by assumption implies $\mathcal{L}_i^{+K_i}(h) \geq c_i > 0$, for all $i \in [N]$. By Lemma 9 we know that h^* is a minimizer of the Chebyshev scalarization with weights $w_i = \frac{1}{\mathcal{L}_i^{+K_i}(h^*)} = \frac{1}{Cc_i}$ for all $i \in [N']$. Let h^* be a minimizer of the scalarization of the true objectives with same weights w_i. The assumption of *ray completeness* together with Lemma 9 implies

$$w_1 \mathcal{L}_1^{+K_1}(h^*) = \ldots = w_N \mathcal{L}_N^{+K_N}(h^*) = \max_{j\in[N]} w_j \mathcal{L}_N^{+K_j}(h^*)$$

. The Chebyshev scalarization is a weighted $L^{(\infty)}$-norm and 1-Lipschitz with respect to itself. Therefore, by Theorem 2:

$$\max_{j\in[N]} w_j \mathcal{L}_j^{+K_j}(\hat{h}^*) \leq \max_{j\in[N]} w_j \mathcal{L}_j^{+K_j}(h^*) + 2\max_{j\in[N]} w_j c_j \tag{A9}$$

Consequently, we obtain the component-wise inequalities:

$$\forall i \in [N] \qquad w_i \mathcal{L}_i^{+K_i}(\hat{h}^*) \leq w_i \mathcal{L}_i^{+K_i}(h^*) + 2\max_{j\in[N]} w_j c_j \tag{A10}$$

Now, inserting the definition K_i, subtracting $w_i K_i$ from both sides and dividing by w_i we obtain

$$\forall i \in [N] \qquad \mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + \frac{2}{w_i}\max_{j\in[N]} w_j c_j. \tag{A11}$$

By construction, $w_j c_j = \frac{1}{C}$ for all $j \in [N]$. Therefore, $\frac{2}{w_i}\max_{j\in[N]} w_j c_j = \frac{2Cc_i}{C} = 2c_i$. Because $c_i = \mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}})$ this concludes the proof.

For the general situation, assume that there are N_{nt} non-trivially and $N - N_{\text{nt}}$ trivially generalizing objectives. If $M = 0$, then $\mathcal{L}_i(h) = \widehat{\mathcal{L}}_i(h)$ for all $i = 1, \ldots, N$ and for all $h \in \mathcal{H}$. Then, Pareto-optimal and empirically Pareto-optimal sets coincide, and $\hat{h}^* = h^*$ fulfills the statement of the theorem.

Otherwise, assume without loss of generality that the objectives are ordered such that, $\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}) > 0$ for $i \in [N_{\text{nt}}]$ and $\mathcal{C}_i(n_i, \mathcal{H}, \delta/N'_{\text{nt}}) = 0$ for $i \in \{N_{\text{nt}} + 1, \ldots, N\}$. Let $\mathcal{G} = \left\{h \in \mathcal{H} : \widehat{\mathcal{L}}_i(h) = \widehat{\mathcal{L}}_i(\hat{h}^*) \text{ for} i \in \{N_{\text{nt}} + 1, \ldots, N\}\right\}$. Note that also $\mathcal{G} = \left\{h \in \mathcal{H} : \mathcal{L}_i(h) = \mathcal{L}_i(h^*) \text{ for} i \in \{N_{\text{nt}} + 1, \ldots, N\}\right\}$, because $\mathcal{L}_{N_{\text{nt}}+1}, \ldots, \mathcal{L}_N$ are trivially generalizing. $\mathcal{G}$ is a subset of $\mathcal{H}$ that is non-empty (because $\hat{h}^* \in \mathcal{G}$). Consequently, the inequalities of Lemma 1 and Theorem 2 hold also as statements for all $g \in \mathcal{G}$ rather than $h \in \mathcal{H}$. Because $\hat{h}^*$ is empirically Pareto-optimal within $\mathcal{H}$ with respect to $\widehat{\mathcal{L}}_1, \ldots, \widehat{\mathcal{L}}_N$, it is also empirically Pareto-optimal in $\mathcal{G}$ with respect to $\widehat{\mathcal{L}}_1, \ldots, \widehat{\mathcal{L}}_M$. Applying the result from the case without trivially generalizing objectives to this situation, we obtain that there exists $h^* \in \mathcal{G}$ such that for all $i \in [N_{\text{nt}}]$

$$\mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + \mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}) \tag{A12}$$

For $i \in \{N_{\text{nt}} + 1, \ldots, N\}$, we have $\mathcal{L}_i(\hat{h}^*) = \mathcal{L}_i(h^*)$, because $h^* \in \mathcal{G}$. Consequently, inequality (A12) holds also for these (with $\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}) = 0$), which concludes the proof.

Funding Open access funding provided by Institute of Science and Technology (IST Austria).

Data availability Our manuscript does not include or refer to any data sources.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article. No funding was received to assist with the preparation of this manuscript.

References

1. Miettinen K (2012) Nonlinear multiobjective optimization. Springer, Berlin
2. Nocedal J, Wright SJ (1999) Numerical optimization. Springer, Berlin
3. Mohri M, Rostamizadeh A, Talwalkar A (2018) Foundations of machine learning. The MIT Press, Cambridge
4. Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: from theory to algorithms. Cambridge University Press, Cambridge
5. Lemaréchal C (2012) Cauchy and the gradient method. Doc Math Extra 251(254):10
6. Audet C, Hare W (2017) Derivative-free and blackbox optimization. Springer, Berlin
7. Bremermann HJ (1962) Optimization through evolution and recombination. Self-Organ Syst 93:106
8. Zitzler E, Thiele L (1999) Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans Evolut Comput(TEC) 3(4):257–271
9. Geoffrion AM (1968) Proper efficiency and the theory of vector maximization. J Math Anal Appl 22(3):618–630
10. Gass S, Saaty T (1955) The computational algorithm for the parametric objective function. Naval Res Logist Quart 2(1–2):39–45
11. Censor Y (1977) Pareto optimality in multiobjective problems. Appl Math Optim 4(1):41–59
12. Cortes C, Mohri M, Gonzalvo J, Storcheus D (2020) Agnostic learning with multiple objectives. In: Conference on Neural Information Processing Systems (NeurIPS)
13. Deist TM, Grewal M, Dankers FJWM, Alderliesten T, Bosman PAN (2021) Multi-objective learning to predict Pareto fronts using hypervolume maximization. arXiv preprint arXiv:2102.04523
14. Fliege J, Svaiter BF (2000) Steepest descent methods for multicriteria optimization. Math Methods Oper Res 51(3):479–494

15. Liu G-P, Kadirkamanathan V (1995) Learning with multi-objective criteria. In: International Conference on Artificial Neural Networks (ICANN)
16. Van Veldhuizen DA, Lamont GB (1998) Evolutionary computation and convergence to a Pareto front. In: Late Breaking Papers at the Genetic Programming 1998 Conference. Stanford University Bookstore
17. Ruchte M, Grabocka J (2021) Scalable Pareto front approximation for deep multi-objective learning. In: International Conference on Data Mining (ICDM)
18. Navon A, Shamsian A, Fetaya E, Chechik G (2021) Learning the Pareto front with hypernetworks. In: International Conference on Learning Representations (ICLR)
19. Aziz H, Lang J, Monnot J (2016) Computing Pareto optimal committees. In: International Joint Conference on Artificial Intelligence (IJCAI)
20. Deb K (2001) Multi-Objective Optimization Using Evolutionary Algorithms. Wiley, Amsterdam
21. Rahwan I, Larson K (2008) Pareto optimality in abstract argumentation. In: Conference on Artificial Intelligence (AAAI)
22. Zhou A, Qu B-Y, Li H, Zhao S-Z, Suganthan PN, Zhang Q (2011) Multiobjective evolutionary algorithms: A survey of the state of the art. Swarm Evol Comput 1(1):32–49
23. Fudenberg D, Tirole J (1991) Game Theory. The MIT Press, Cambridge
24. Pardalos PM, Migdalas A, Pitsoulis L (2008) Pareto Optimality, Game Theory and Equilibria. Springer, Berlin
25. Hochman HM, Rodgers JD (1969) Pareto optimal redistribution. Am Econ Rev 59(4):542–557
26. Keeney RL, Raiffa H, Meyer RF (1993) Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press, Cambridge
27. Fieldsend JE, Singh S (2005) Pareto evolutionary neural networks. IEEE Trans Neural Netw (TNN) 16(2):338–354
28. Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley
29. Jin Y (2006) Multi-Objective Machine Learning. Springer, Berlin
30. Van Moffaert K, Nowé A (2014) Multi-objective reinforcement learning using sets of Pareto dominating policies. J Mach Learn Res (JMLR) 15(1):3483–3512
31. Ye F, Lin B, Yue Z, Guo P, Xiao Q, Zhang Y (2021) Multi-objective meta learning. In: Conference on Neural Information Processing Systems (NeurIPS)
32. Jin Y, Sendhoff B (2008) Pareto-based multiobjective machine learning: An overview and case studies. IEEE Trans Syst Man Cybernet Syst (TSMCS) 38(3):397–415
33. Przybylski A, Gandibleux X (2017) Multi-objective branch and bound. Eur J Oper Res (EJOR) 260(3):856–872
34. Vamplew P, Dazeley R, Berry A, Issabekov R, Dekker E (2011) Empirical evaluation methods for multiobjective reinforcement learning algorithms. Mach Learn 84(1):51–80
35. Zhu H, Jin Y (2019) Multi-objective evolutionary federated learning. IEEE Trans Neural Netw Learned Syst (TNNLS) 31(4):1310–1322
36. Teytaud O (2007) On the hardness of offline multi-objective optimization. Evol Comput 15(4):475–491
37. Wang X, Sandholm T (2003) Learning near-pareto-optimal conventions in polynomial time. In: Conference on Neural Information Processing Systems (NeurIPS)
38. Stark DR, Spall JC (2003) Rate of convergence in evolutionary computation. In: American Control Conference (ACC)
39. Vapnik V (2013) The Nature of Statistical Learning Theory. Springer, Berlin
40. Tibshirani RJ (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58(1):267–288
41. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499
42. Gaines BR, Kim J, Zhou H (2018) Algorithms for fitting the constrained lasso. J Comput Graph Stat 27(4):861–871
43. Osborne MR, Presnell B, Turlach BA (2000) On the LASSO and its dual. J Comput Graph Stat 9(2):319–337
44. Tibshirani RJ (2013) The lasso problem and uniqueness. Electron J Stat 7:1456–1490
45. Li T, Beirami A, Sanjabi M, Smith V (2021) Tilted empirical risk minimization. In: International Conference on Learning Representations (ICLR)
46. Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning. https://fairmlbook.org/
47. Martinez N, Bertran M, Sapiro G (2020) Minimax Pareto fairness: A multi objective perspective. In: International Conference on Machine Learing (ICML)
48. Wei S, Niethammer M (2020) The fairness-accuracy Pareto front. Stat Anal Data Mining 15:287
49. Kamani MM, Forsati R, Wang JZ, Mahdavi M (2021) Pareto efficient fairness in supervised learning: From extraction to tracing. arXiv preprint arXiv:2104.01634
50. Padh K, Antognini D, Lejal-Glaude E, Faltings B, Musat C (2021) Addressing fairness in classification with a model-agnostic multi-objective algorithm. In: Uncertainty in Artificial Intelligence (UAI)
51. Liu S, Vicente LN (2022) Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. CMS 19(3):513–537
52. Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. In: Innovations in Theoretical Computer Science Conference (ITCS)
53. Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5(2):153–163
54. Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res 50(1):3–44
55. Woodworth B, Gunasekar S, Ohannessian MI, Srebro N (2017) Learning non-discriminatory predictors. In: Workshop on Computational Learning Theory (COLT)
56. Konstantinov N, Lampert CH (2022) Fairness-aware PAC learning from corrupted data. J Mach Learn Res (JMLR). 23:1
57. Kamishima T, Akaho S, Sakuma J (2011) Fairness-aware learning through regularization approach. In: International Conference on Data Mining – Workshops (ICDMW)
58. Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: International Conference on Data Mining – Workshops (ICDMW)
59. Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33
60. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Conference on Neural Information Processing Systems (NeurIPS)
61. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR)
62. Yin D, Kannan R, Bartlett P (2019) Rademacher complexity for adversarially robust generalization. In: International Conference on Machine Learing (ICML)
63. Cisse M, Bojanowski P, Grave E, Dauphin Y, Usunier N (2017) Parseval networks: Improving robustness to adversarial examples. In: International Conference on Machine Learing (ICML)
64. Weng L, Zhang H, Chen H, Song Z, Hsieh C-J, Daniel L, Boning D, Dhillon I (2018) Towards fast computation of certified robustness for ReLU networks. In: International Conference on Machine Learing (ICML)

65. Kuznetsov V, Mohri M, Syed U (2015) Rademacher complexity margin bounds for learning with a large number of classes. In: ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels
66. Koltchinskii V, Panchenko D (2002) Empirical margin distributions and bounding the generalization error of combined classifiers. Ann Stat 30(1):1–50
67. Strubell E, Ganesh A, McCallum A (2020) Energy and policy considerations for modern deep learning research. In: Conference on Artificial Intelligence (AAAI)
68. Menghani G (2021) Efficient deep learning: A survey on making deep learning models smaller, faster, and better. arXiv preprint arXiv:2106.08962
69. Hoefler T, Alistarh D, Ben-Nun T, Dryden N, Peste A (2021) Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. J Mach Learn Res (JMLR) 22(241):1–124
70. Gholami A, Kim S, Dong Z, Yao Z, Mahoney MW, Keutzer K (2022) Chap. 13 A survey of quantization methods for efficient neural network inference. In: Thiruvathukal GK, Lu Y-H, Kim J, Chen Y, Chen B (eds) Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence. CRC Press, Boca Raton
71. Kurtz M, Kopinsky J, Gelashvili R, Matveev A, Carr J, Goin M, Leiserson W, Moore S, Shavit N, Alistarh D (2020) Inducing and exploiting activation sparsity for fast inference on deep neural networks. In: International Conference on Machine Learing (ICML)
72. Schwing AG, Zach C, Zheng Y, Pollefeys M (2011) Adaptive random forest - how many "experts" to ask before making a decision? In: Conference on Computer Vision and Pattern Recognition (CVPR)
73. Lampert CH (2012) Dynamic pruning of factor graphs for maximum marginal prediction. In: Conference on Neural Information Processing Systems (NeurIPS)
74. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Conference on Computer Vision and Pattern Recognition (CVPR)
75. Huang G, Chen D, Li T, Wu F, Maaten L, Weinberger K (2018) Multi-scale dense networks for resource efficient image classification. In: International Conference on Learning Representations (ICLR)
76. Teerapittayanon S, McDanel B, Kung H-T (2016) BranchyNet: Fast inference via early exiting from deep neural networks. In: International Conference on Learning Representations (ICLR)
77. Sener O, Koltun V (2018) Multi-task learning as multi-objective optimization. In: Conference on Neural Information Processing Systems (NeurIPS)
78. Lin X, Zhen H-L, Li Z, Zhang Q-F, Kwong S (2019) Pareto multi-task learning. In: Conference on Neural Information Processing Systems (NeurIPS)
79. Ma P, Du T, Matusik W (2020) Efficient continuous Pareto exploration in multi-task learning. In: International Conference on Machine Learing (ICML)
80. Mahapatra D, Rajan V (2020) Multi-task learning with user preferences: Gradient descent with controlled ascent in Pareto optimization. In: International Conference on Machine Learing (ICML)
81. Ruchte M, Grabocka J (2021) Multi-task problems are not multi-objective. arXiv preprint arXiv:2110.07301
82. Zhang M-L, Zhou Z-H (2013) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng (TKDE) 26(8):1819–1837
83. Varma M (2019) Extreme classification. Commun ACM (CACM) 62(11):44–45
84. Shi C, Kong X, Yu, PS, Wang B (2012) Multi-objective multi-label classification. In: International Conference on Data Mining (ICDM)
85. Liu T-Y (2009) Learning to rank for information retrieval. Found Trends Inf Retr 3(3):225–331
86. Cao X, Zhu S, Tang B, Xie R, Zhang F, Wang Z (2020) Ranking with deep multi-objective learning. In: KDD Workshop on Deep Learning Practice for High-Dimensional Sparse Data
87. Svore KM, Volkovs MN, Burges CJC (2011) Learning to rank with multiple objective functions. In: International World Wide Web Conference (WWW)
88. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
89. Agarwal S, Graepel T, Herbrich R, Har-Peled S, Roth D, Jordan MI (2005) Generalization bounds for the area under the ROC curve. J Mach Learn Res (JMLR) 6(4)
90. Dziugaite GK, Roy DM (2017) Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In: Uncertainty in Artificial Intelligence (UAI)
91. McAllester DA (1999) Some PAC-Bayesian theorems. Mach Learn 37(3):355–363
92. Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Learn Res (JMLR) 2:499–526
93. Hardt M, Recht B, Singer Y (2016) Train faster, generalize better: Stability of stochastic gradient descent. In: International Conference on Machine Learing (ICML)
94. Bauer FL, Stoer J, Witzgall C (1961) Absolute and monotonic norms. Numer Math 3(1):257–264