Generalization In Multi-Objective Machine Learning

Anonymous authors
Paper under double-blind review

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 characterization 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.

1 Introduction

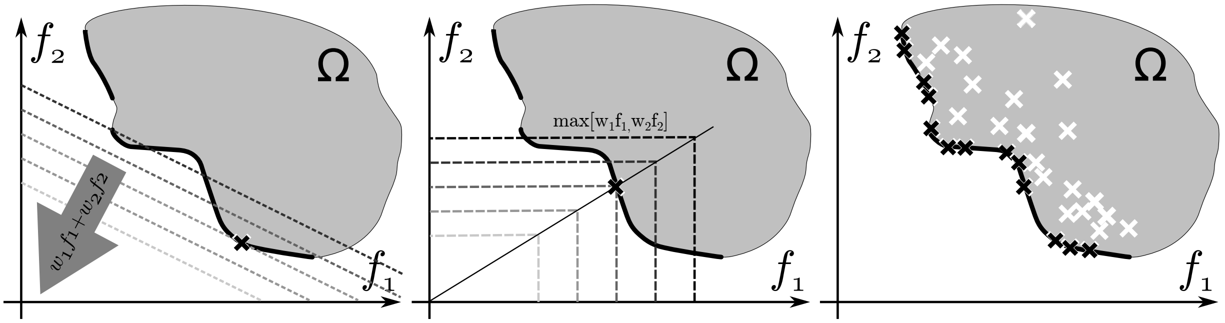
Traditionally, statistical machine learning has concentrated on solving one single-objective optimization problem: to minimize the average loss over a given training set. Additional quantities of interest, such as *model complexity*, had to be either addressed implicitly by the choice of model class, 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 *efficiency*, *robustness* or *fairness* 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 characterize 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 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.

---

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



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

Figure 1: For general multi-objective optimization problems the Pareto-front (bold curves) can be disconnected and non-convex. (a) *Linear 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.

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 (2012) and Nocedal & Wright (1999) for optimization, and Mohri et al. (2018) and Shalev-Shwartz & Ben-David (2014) 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 point $\omega, \omega' \in \Omega$ 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* (Lemaréchal, 2012; Nocedal & Wright, 1999) or *derivative-free* (Audet & Hare, 2017; Bremermann, 1962).

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 *strictly 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* (Zitzler & Thiele, 1999) are often employed. If a single Pareto-optimal solution suffices, *scalarizations* in combination with single-objective optimization can be used (Geoffrion, 1968). 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:

$\mathcal{U}_w^{(p)}(x_1,\dots,x_N) = \big(\sum_{i\in[N]}|w_i x_i|^p\big)^{1/p}$ for $p\in(1,\infty)$, and $\mathcal{U}_w^{(\infty)}(x_1,\dots,x_N) = \max_{i\in[N]}|w_i x_i|$, 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 (Gass & Saaty, 1955). For any non-zero choice of weights, minimizers of this resulting scalarized objective will be Pareto-optimal (Geoffrion, 1968). However, the set of solutions obtainable by varying the weights might not recover the complete Pareto-front, unless the optimization problem is convex (Censor, 1977). In contrast, the choice $p=\infty$ (called *weighted Chebyshev norm*) allows recovering the complete Pareto front when varying the weights in Δ_N (Miettinen, 2012, 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.*[2]

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*, $S=\{z_1,\dots,z_n\}$ to computes 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-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*:

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

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.

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 $\hat{h}^*\in\arg\min_{h\in\mathcal{H}}\widehat{\mathcal{L}}(h)$, an *excess risk bound* holds:

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

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

Multi-objective learning. In multi-objective learning, multiple target objectives, $\mathcal{L}_1,\dots,\mathcal{L}_N$, characterize different properties of interest of the hypotheses. Estimating them from a (single) dataset yields empirical objectives, $\widehat{\mathcal{L}}_1,\dots,\widehat{\mathcal{L}}_N$. 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 Section 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 there are two sets of Pareto-optimal hypotheses:

---

[2] 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 accuracy. For unsupervised learning, one uses $\mathcal{Z}=\mathcal{X}$, and ℓ measures, e.g., the reconstruction error of a clustering or dimensionality reduction step.

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 (Cortes et al., 2020; Deist et al., 2021; Fliege & Svaiter, 2000). Alternatively, one can search for hypotheses along the complete (empirically) Pareto-front, using, e.g., ensemble techniques (Liu & Kadirkamanathan, 1995; Van Veldhuizen & Lamont, 1998), model conditioning (Ruchte & Grabocka, 2021), or hypernetworks (Navon et al., 2021).

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. Hardly any such results exist in the literature, as we discuss in Section 5.

3 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 Section 2. We will discuss instantiations that either improve over related existing work or provide new insights in Sections 4 and 5.

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 individually a generalization bound of the form (1) *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 Section 4, 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.

3.1 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. Then it holds with probability at least $1-\delta$,*

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

Lemma 1 is in fact a straight-forward 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.

3.2 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 $\mathcal{L}_{\mathcal{U}}(h) := \mathcal{U}(\mathcal{L}_1(h), \ldots, \mathcal{L}_N(h))$, $\widehat{\mathcal{L}}_{\mathcal{U}}(h) := \mathcal{U}(\widehat{\mathcal{L}}_1(h), \ldots, \widehat{\mathcal{L}}_N(h))$.

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 $L_{\mathcal{U}}$-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 $h \in \mathcal{H}$:

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

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

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

The proof is a combination of Lemma 1 with the Lipschitz assumptions. It can be found in Appendix A.

Discussion. Theorem 2 establishes *generalization* and *excess bounds* for the situation of scalarization-based multi-objective learning. Their relevance lies not so much in the inequalities (4) and (5) themselves, which have the standard single-objective form, but 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 (Shalev-Shwartz & Ben-David, 2014, Chapter 11). Despite its simplicity, the theorem improves over prior work (Cortes et al., 2020), which proved guarantees that depend on the size of $\mathfrak{U}$. For a more detailed discussion see Section 5.1.

3.3 Pareto excess bounds

We now state formal characterizations 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*

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

The proof relies on Lemma 1 and the definition of (empirical) Pareto-optimality, see Appendix A.

Inequality (6) in Theorem 3 will typically hold only with respect to a single objective, $I = \{i\}$. For multi-objective learning the most relevant question would be if there is an analog 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*

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

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

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

The proofs are provided in Appendix A. Part a) is a consequence of Lemma 1 and the definition of (empirical) Pareto-optimality. Part b) we prove by constructing for each h^* a suitable scalarization $\mathcal{U}$ from which a hypothesis $\hat{h}^*$ can be derived that fulfills (up to generalization terms) $\mathcal{L}_{\mathcal{U}}(h^*) \leq \mathcal{L}_{\mathcal{U}}(\hat{h}^*)$, and $\widehat{\mathcal{L}}_{\mathcal{U}}(\hat{h}^*) \leq \widehat{\mathcal{L}}_{\mathcal{U}}(h^*)$. The main difficulty is to find $\mathcal{U}$ such that the scalarized relations imply related statements about all individual objectives.

As it turns out, without any additional assumption statement b) of Theorem 4 would be false. The following theorem shows this.

Theorem 5. *Let* $N \geq 2$*. Then, for any* $C > 0$ *there exist 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*

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

The proof uses an explicit construction, 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* (Vapnik, 2013). 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.[3]

Finally, 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 $N \geq 2$.

3.4 Summary

In combination, the results of this section establish a complete picture of similarities and differences between the generalization properties of single-objective and multi-objective learning. In particular, it highlights

---

[3] The *ray completeness* assumption is truly quite strong. From the proof it can be seen that a weaker assumption would suffice that relates rays through the empirical Pareto-front with rays through the true Pareto-front. We do not explore this option in this work, though.

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.

4 Applications

Our results of the previous section establish which generalization guarantees carry over from the single-objective to the multi-objective setting. For a variety of learning tasks this provides a new unified and principled framework to study their generalization properties, which previously were either not analyzed at all, or in task-specific or ad-hoc ways. In this section, we sketch some situations in which our results recover or improve existing work or allow new insights. In each situation where our results are applicable, it immediately follows that generalization holds uniformly across all objectives, that learning with arbitrarily many scalarizations will be possible without loss of confidence, and that finding the empirical Pareto-front will be a theoretically justified learning method, though with the caveat detailed above.

Regularization. Explicit regularization methods form a *regularized objective* consisting of an ordinary loss term and one or more (weighted) regularization terms (Tikhonov, 1943; Schölkopf & Smola, 2002; Goodfellow et al., 2016). By varying the weights, hypotheses with different trade-offs between prediction accuracy and hypothesis complexity are sought. Clearly, this *regularized risk minimization* setting is identical to solving a multi-objective learning problem using different linear scalarizations (de Medeiros et al., 2017). Therefore, the found solutions are empirically Pareto-optimal. For accuracy measures, many generalization bounds are known, and the regularization terms typically generalize trivially. Then, our results from Section 3 hold with the same confidence as any original generalization bound on just the original loss term.

Despite the fact that the setting of regularized risk minimization is quite well understood already, one new perspective that the multi-objective view provides stems from Theorem 5 (more precisely, its proof): For every achievable value of the regularizer, some hypothesis will be empirically Pareto-optimal. It is possible, however, that all Pareto-optimal solutions have a much smaller value of the regularizer. This will not negatively affect prediction accuracy, but it might be an undesirable effect for other properties that are related to hypothesis complexity, such as *interpretability*.

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* (Barocas et al., 2019). Because accuracy and fairness can be in conflict with each other, fairness-aware learning is a prototypical candidate for multi-objective learning (Martinez et al., 2020; Wei & Niethammer, 2020; Kamani et al., 2021; Padh et al., 2021). This view also extends naturally to integration of multiple fairness measures (Liu & Vicente, 2020), which might be incompatible with each other (Kleinberg et al., 2016; Chouldechova, 2017; Berk et al., 2021). Generalization bounds for the empirical estimation of unfairness measures have been developed (Woodworth et al., 2017; Konstantinov & Lampert, 2022). Consequently, our results from Section 3 apply, yielding a unified understanding of the generalization properties of fairness-aware learning, e.g., regularization-based (Kamishima et al., 2011) constraint-based (Calders et al., 2009), or Pareto-based (Liu & Vicente, 2020; Navon et al., 2021). The multi-objective view also allows us to conjecture that methods that seek fair hypotheses by other means, such as pre-processing (Kamiran & Calders, 2012) or post-processing (Hardt et al., 2016a), 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* (Madry et al., 2018) adds a robustness-enforcing loss term to the training problem. Generalization bounds for such terms have been derived, e.g. Yin et al. (2019). Consequently, multi-objective learning can be used in this setting with the guarantees and caveats discussed above. *Lipschitz-networks* (Cisse et al., 2017) restrict the hypothesis class to functions with a small Lipschitz constant, typically 1. Afterwards 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 (Weng et al., 2018). 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. (2015); Koltchinskii & Panchenko (2002).

Efficiency. Large machine learning models, in particular deep networks, often have high computational demands, not only at training but also at prediction time (Strubell et al., 2020; Menghani, 2021). Consequently, a number of techniques have been developed that aim at reducing the computational cost. *Parameter sparsification* (Hoefler et al., 2021) and *quantization* (Gholami et al., 2021) 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 (Zhu & Jin, 2019). 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* (Kurtz et al., 2020) is a data-dependent quantity that can also be shown to generalize, see Appendix B. Therefore it as well can be handled in a multi-objective way. *Adaptive computation* methods, such as *ensembles* (Schwing et al., 2011), *classifier cascades* (Viola & Jones, 2001) or *multi-exit architectures* (Huang et al., 2018; Teerapittayanon et al., 2016), evaluate different subsets of a larger model depending on the input sample. For suitable design choices, generalization bounds can be shown for the resulting computation time, see Appendix B. Consequently, in all described cases our results from Section 3 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 (Sener & Koltun, 2018; Lin et al., 2019; Ma et al., 2020; Mahapatra & Rajan, 2020). This setting is of a non-standard form, as each task typically has a dedicated training set. 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* (Zhang & Zhou, 2013) and *extreme classification* (Varma, 2019), 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 (Shi et al., 2012). 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 (Liu, 2009) solutions are typically judged by two measures: *precision* and *recall*. *A priori*, this makes it a promising setting for multi-objective analysis (Cao et al., 2020; Svore et al., 2011). 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)* (Hanley & McNeil, 1982), for which indeed generalization bounds have been derived (Agarwal et al., 2005). However, the AUC is not a scalarization in the sense of Section 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 (Dziugaite & Roy, 2017; McAllester, 1999), or *algorithmic stability* (Bousquet & Elisseeff, 2002; Hardt et al., 2016b). 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.

5 Related work

Solving problems with multiple objectives has a long tradition in artificial intelligence (Aziz et al., 2016; Deb, 2001; Rahwan & Larson, 2008; Zhou et al., 2011), game theory (Fudenberg & Tirole, 1991; Pardalos et al., 2008), and economics (Hochman & Rodgers, 1969; Keeney et al., 1993). Since the 1990s it has also attracted attention from the machine learning community, e.g. Fieldsend & Singh (2005); Goldberg (1989); Jin (2006). 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. (2020); Van Moffaert & Nowé (2014); Ye et al. (2021), as well as exploring the complete Pareto front (Jin & Sendhoff, 2008; Navon et al., 2021; Przybylski & Gandibleux, 2017; Ruchte & Grabocka, 2021; Vamplew et al., 2011; Van Moffaert & Nowé, 2014; Zhu & Jin, 2019). 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* (Teytaud, 2007; Wang & Sandholm, 2003) or *convergence rates* (Stark & Spall, 2003), but not statistical generalization. A notable exception is Cortes et al. (2020), which we discuss in the following section.

5.1 Comparison to Cortes et al. (2020)

In Cortes et al. (2020) the authors study the generalization properties of multi-objective learning for the case of a special scalarization obtained by minimizing over convex combinations. For easier comparison, we state their result in our notation[4]

Theorem 6 (Theorem 3 in Cortes et al. (2020))**.** *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 $\sum_{i=1}^N w_iM_i\leq\beta$ for all $w=(w_1,\dots,w_N)\in W$.*

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

$$\mathcal{L}_W(h)\leq\widehat{\mathcal{L}}_W(h)+2\beta\hat{\mathfrak{R}}_S(\mathcal{H})+M\epsilon+3\beta D\sqrt{\frac{1}{2n}\log\Big[\frac{2|W_\epsilon|}{\delta}\Big]}, \tag{10}$$

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 (10) 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 Section 3. A standard Rademacher-based generalization bound holds for each objective individually, because the loss functions are Lipschitz-continuous. Therefore, Assumption A is fulfilled. By using that the scalarizations $\mathcal{U}_W(x_1,\dots,x_N)=\max_{w\in W}\sum_{i=1}^N w_i|x_i|$ is β-Lipschitz with respect to a suitably defined norm, we obtain

Corollary 1 (Instantiation of Theorem 2 to the setting of Theorem 6)**.** *Make the same assumptions as in Theorem 6. Then, for any $\delta>0$, with probability at least $1-\delta$ the following inequality holds for all $h\in\mathcal{H}$:*

$$\mathcal{L}_W(h)\leq\widehat{\mathcal{L}}_W(h)+2\beta\hat{\mathfrak{R}}_S(\mathcal{H})+3\beta D\sqrt{\frac{\log\frac{2N}{\delta}}{2n}}. \tag{11}$$

The details can be found in Appendix A.

---

[4] Our formulation also has slightly different constants in the generalization term, which we believe to be necessary based on the theorem's proof.

Two differences are apparent. First, our bound is simpler. It holds without need for an ϵ-parameter that additively enters the right hand side of (10), yet also influences the size of the right-most 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 (10) 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 (11) is independent of the size of W and its confidence term grows only logarithmically with respect to N.

A number of further differences between our work and Cortes et al. (2020) exist. First, our results apply to more settings, as we allow for a broader class of generalization bounds rather than just Rademacher-based ones. Second, our results also hold for other scalarizations besides the maximum over linear combinations. Third, our analysis of the relation between empirical and truly Pareo-optimal sets has no counterpart in Cortes et al. (2020).

6 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 a characterization 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 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.

References

Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, Dan Roth, and Michael I.g Jordan. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research (JMLR)*, 6(4), 2005.

Charles Audet and Warren Hare. *Derivative-Free and Blackbox Optimization*. Springer, 2017.

Haris Aziz, Jérôme Lang, and Jérôme Monnot. Computing Pareto optimal committees. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Friedrich L. Bauer, Josef Stoer, and Christoph Witzgall. Absolute and monotonic norms. *Numerische Mathematik*, 3(1):257–264, 1961.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 2:499–526, 2002.

Hans J. Bremermann. Optimization through evolution and recombination. *Self-Organizing Systems*, 93:106, 1962.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *International Conference on Data Mining – Workshops (ICDMW)*, 2009.

Xuezhi Cao, Sheng Zhu, Biao Tang, Rui Xie, Fuzheng Zhang, and Zhongyuan Wang. Ranking with deep multi-objective learning. In *KDD Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, 2020.

Yair Censor. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization*, 4(1): 41–59, 1977.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learing (ICML)*, 2017.

Corinna Cortes, Mehryar Mohri, Javier Gonzalvo, and Dmitry Storcheus. Agnostic learning with multiple objectives. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Talles Henrique de Medeiros, Honovan Paz Rocha, Frank Sill Torres, Ricardo Hiroshi Caldeira Takahashi, and Antônio Pádua Braga. Multi-objective decision in machine learning. *Journal of Control, Automation and Electrical Systems*, 28(2):217–227, 2017.

Kalyamoy Deb. *Multi-Objective Optimization Using Evolutionary Algorithms.* Wiley, 2001.

Timo M. Deist, Monika Grewal, Frank J. W. M. Dankers, Tanja Alderliesten, and Peter A. N. Bosman. Multi-objective learning to predict Pareto fronts using hypervolume maximization. *arXiv preprint arXiv:2102.04523*, 2021.

David L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

Jonathan E. Fieldsend and Sameer Singh. Pareto evolutionary neural networks. *IEEE Transactions on Neural Networks (TNN)*, 16(2):338–354, 2005.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, 2000.

Drew Fudenberg and Jean Tirole. *Game Theory.* The MIT Press, 1991.

Saul Gass and Thomas Saaty. The computational algorithm for the parametric objective function. *Naval Research Logistics Quarterly*, 2(1-2):39–45, 1955.

Arthur M. Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications*, 22(3):618–630, 1968.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021.

David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, 1989.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* The MIT Press, 2016.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016a.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learing (ICML)*, 2016b.

Harold M. Hochman and James D. Rodgers. Pareto optimal redistribution. *The American Economic Review*, 59(4):542–557, 1969.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research (JMLR)*, 22(241):1–124, 2021.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations (ICLR)*, 2018.

Yaochu Jin. *Multi-Objective Machine Learning.* Springer, 2006.

Yaochu Jin and Bernhard Sendhoff. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man and Cybernetics: Systems (TSMCS)*, 38(3):397–415, 2008.

Mohammad Mahdi Kamani, Rana Forsati, James Z. Wang, and Mehrdad Mahdavi. Pareto efficient fairness in supervised learning: From extraction to tracing. *arXiv preprint arXiv:2104.01634*, 2021.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *International Conference on Data Mining – Workshops (ICDMW)*, 2011.

Ralph L. Keeney, Howard Raiffa, and Richard F. Meyer. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs.* Cambridge University Press, 1993.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

Nikola Konstantinov and Christoph H. Lampert. Fairness-aware PAC learning from corrupted data. *Journal of Machine Learning Research (JMLR)*, 2022.

Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Nir Shavit, and Dan Alistarh. Inducing and exploiting activation sparsity for fast inference on deep neural networks. In *International Conference on Machine Learing (ICML)*, 2020.

Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. 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*, 2015.

Claude Lemaréchal. Cauchy and the gradient method. *Documenta Mathematica, Extra Vol., Optimization Stories*, pp. 251–254, 2012.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Guo-Ping Liu and Visakan Kadirkamanathan. Learning with multi-objective criteria. In *International Conference on Artificial Neural Networks (ICANN)*, 1995.

Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *arXiv preprint arXiv:2008.01132*, 2020.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3 (3):225–331, 2009.

Pingchuan Ma, Tao Du, and Wojciech Matusik. Efficient continuous Pareto exploration in multi-task learning. In *International Conference on Machine Learing (ICML)*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in Pareto optimization. In *International Conference on Machine Learing (ICML)*, 2020.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learing (ICML)*, 2020.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory (ALT)*, 2016.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Gaurav Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *arXiv preprint arXiv:2106.08962*, 2021.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization.* Springer, 2012.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* The MIT Press, 2018.

Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the Pareto front with hypernetworks. In *International Conference on Learning Representations (ICLR)*, 2021.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Workshop on Computational Learning Theory (COLT)*, 2015.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* Springer, 1999.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Uncertainty in Artificial Intelligence (UAI)*, 2021.

Panos M. Pardalos, Athanasios Migdalas, and Leonidas Pitsoulis. *Pareto Optimality, Game Theory and Equilibria.* Springer, 2008.

Anthony Przybylski and Xavier Gandibleux. Multi-objective branch and bound. *European Journal of Operational Research (EJOR)*, 260(3):856–872, 2017.

Iyad Rahwan and Kate Larson. Pareto optimality in abstract argumentation. In *Conference on Artificial Intelligence (AAAI)*, 2008.

Michael Ruchte and Josif Grabocka. Scalable Pareto front approximation for deep multi-objective learning. In *International Conference on Data Mining (ICDM)*, 2021.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, 2002.

Alexander G. Schwing, Christopher Zach, Yefeng Zheng, and Marc Pollefeys. Adaptive random forest - how many "experts" to ask before making a decision? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

Chuan Shi, Xiangnan Kong, Philip S. Yu, and Bai Wang. Multi-objective multi-label classification. In *International Conference on Data Mining (ICDM)*, 2012.

David R. Stark and James C. Spall. Rate of convergence in evolutionary computation. In *American Control Conference (ACC)*, 2003.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Conference on Artificial Intelligence (AAAI)*, 2020.

Krysta M. Svore, Maksims N. Volkovs, and Christopher J. C. Burges. Learning to rank with multiple objective functions. In *International World Wide Web Conference (WWW)*, 2011.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. BranchyNet: Fast inference via early exiting from deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Olivier Teytaud. On the hardness of offline multi-objective optimization. *Evolutionary Computation*, 15(4): 475–491, 2007.

Andrey Nikolayevich Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39:195–198, 1943.

Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 84(1):51–80, 2011.

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of Pareto dominating policies. *Journal of Machine Learning Research (JMLR)*, 15(1):3483–3512, 2014.

David A. Van Veldhuizen and Gary B. Lamont. Evolutionary computation and convergence to a Pareto front. In *Late Breaking Papers at the Genetic Programming 1998 Conference.* Stanford University Bookstore, 1998.

Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer, 2013.

Manik Varma. Extreme classification. *Communications of the ACM (CACM)*, 62(11):44–45, 2019.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

Xiaofeng Wang and Tuomas Sandholm. Learning near-pareto-optimal conventions in polynomial time. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2003.

Susan Wei and Marc Niethammer. The fairness-accuracy Pareto front. *Statistical Analysis and Data Mining*, 2020.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learing (ICML)*, 2018.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Workshop on Computational Learning Theory (COLT)*, 2017.

Feiyang Ye, Baijiong Lin, Zhixiong Yue, Pengxin Guo, Qiao Xiao, and Yu Zhang. Multi-objective meta learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learing (ICML)*, 2019.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 26(8):1819–1837, 2013.

Aimin Zhou, Bo-Yang Qu, Hui Li, Shi-Zheng Zhao, Ponnuthurai Nagaratnam Suganthan, and Qingfu Zhang. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, 1(1):32–49, 2011.

Hangyu Zhu and Yaochu Jin. Multi-objective evolutionary federated learning. *IEEE Transactions on Neural Networks and Learned Systems (TNNLS)*, 31(4):1310–1322, 2019.

Eckart Zitzler and Lothar Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation (TEC)*, 3(4):257–271, 1999.

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 $h \in \mathcal{H}$:

$$\begin{aligned} |\mathcal{L}_{\mathcal{U}}(h) - \widehat{\mathcal{L}}_{\mathcal{U}}(h)| &\leq L_{\mathcal{U}} \|\big(\mathcal{L}_1(h) - \widehat{\mathcal{L}}_1(h), \dots, \mathcal{L}_N(h) - \widehat{\mathcal{L}}_N(h)\big)\|_{\mathcal{U}} && (12) \\ &\leq L_{\mathcal{U}} \|\big(|\mathcal{L}_1(h) - \widehat{\mathcal{L}}_1(h)|, \dots, |\mathcal{L}_N(h) - \widehat{\mathcal{L}}_N(h)|\big)\|_{\mathcal{U}} && (13) \\ &\leq L_{\mathcal{U}} \|\big(\mathcal{C}_1(n_1, \mathcal{H}, \delta/N_{\mathrm{nt}}), \dots, \mathcal{C}_N(n_N, \mathcal{H}, \delta/N_{\mathrm{nt}})\big)\|_{\mathcal{U}} && (14) \end{aligned}$$

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 Bauer et al. (1961). In combination, this proves statement a).

Statement b) follows by arguments mirroring the proof of classic *excess risk bounds* (Mohri et al., 2018). Let $\hat{h}^*_{\mathcal{U}} \in \arg\min_{h \in \mathcal{H}} \widehat{\mathcal{L}}_{\mathcal{U}}(h)$.[5] . Then, it holds for arbitrary $h \in \mathcal{H}$ that

$$\begin{aligned} \mathcal{L}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) - \mathcal{L}_{\mathcal{U}}(h) &\leq \mathcal{L}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) - \widehat{\mathcal{L}}_{\mathcal{U}}(\hat{h}^*_{\mathcal{U}}) + \widehat{\mathcal{L}}_{\mathcal{U}}(h) - \mathcal{L}_{\mathcal{U}}(h) && (15) \\ &\leq 2 \|\big(\mathcal{C}_1(n_1, \mathcal{H}, \delta/N_{\mathrm{nt}}), \dots, \mathcal{C}_N(n_N, \mathcal{H}, \delta/N_{\mathrm{nt}})\big)\|_{\mathcal{U}} && (16) \end{aligned}$$

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 and observing that if the inequality holds for all $h \in \mathcal{H}$, it also holds for the infimum.

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 $1-\delta$.

---

[5] If the minimum is not actually attained, we can still prove a variant of the theorem that uses arbitrary close approximation. We not do state this explicitly to avoid cluttered notation.

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 Pareto-optimal hypothesis* $h^* \in \mathcal{H}$ *such that* $\mathcal{L}_i(\hat{h}^*) - \mathcal{L}_i(h^*) > 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{nt})$ *for all for all* $i \in [N]$.

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

$$\widehat{\mathcal{L}}_i(h^*) - \widehat{\mathcal{L}}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) - \mathcal{L}_i(\hat{h}^*) + 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\mathrm{nt}}) < 0. \tag{17}$$

For the first inequality we applied Lemma 1 twice, and the second inequality follows from the assumption. However, (17) establishes that h^* empirically 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):

$$\mathcal{L}_i(\hat{h}^*) - \mathcal{L}_i(h^*) \leq \mathcal{L}_i(\hat{h}^*) - \widehat{\mathcal{L}}_i(\hat{h}^*) + \widehat{\mathcal{L}}_i(h^*) - \mathcal{L}_i(h^*) \leq 2\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\mathrm{nt}}). \tag{18}$$

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 7. *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* $K_1, \dots, K_N$*, a solution* $h \in \mathcal{H}$ *is Pareto-optimal for* $\mathcal{L}_1, \dots, \mathcal{L}_N$ *if and only if it is Pareto-optimal for* $\mathcal{L}_1^{+K_1}, \dots, \mathcal{L}_1^{+K_N}$*. 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 the relative order of objective values, which is not affected by additive shifts.

Lemma 8. *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 $\mathcal{L}_1, \dots, \mathcal{L}_N$. For a more concise notation, we abbreviate $c_i = \mathcal{C}_i(n, \mathcal{H}, \delta/N_{\mathrm{nt}})$.

First, we consider the case where none of the objectives are trivially generalizing, i.e. $c_i > 0$ for all $i \in [N]$. By Lemma 7, we know that h^* is also empirically Pareto-optimal for the shifted objectives $\widehat{\mathcal{L}}_1^{+K_1}, \dots, \widehat{\mathcal{L}}_{N'}^{+K_{N'}}$ with

$$K_i := Cc_i - \widehat{\mathcal{L}}_i(h^*) \qquad \text{for } C = 2 + \max_j [\frac{1}{c_j}(\widehat{\mathcal{L}}_j(h^*) - \min_h \widehat{\mathcal{L}}_j(h))] \tag{19}$$

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 8 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 8 implies $w_1\mathcal{L}_1^{+K_1}(h^*) = \cdots = w_N\mathcal{L}_N^{+K_N}(h^*) = \max_{j\in[N]} w_j\mathcal{L}_N^{+K_j}(h^*)$. The Chebyshev scalarization is a weighted $L^{(\infty)}$-norm and 1-Lipschitz with respect to itself. Therefore, by Theorem 2:

$$\max_{j\in[N]} w_j\mathcal{L}_j^{+K_j}(\hat{h}^*) \leq \max_{j\in[N]} w_j\mathcal{L}_j^{+K_j}(h^*) + 2\max_{j\in[N]} w_j c_j \tag{20}$$

Consequently, we obtain the component-wise inequalities:

$$\forall i \in [N] \qquad w_i\mathcal{L}_i^{+K_i}(\hat{h}^*) \leq w_i\mathcal{L}_i^{+K_i}(h^*) + 2\max_{j\in[N]} w_j c_j \tag{21}$$

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

$$\forall i \in [N] \qquad \mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + \frac{2}{w_i}\max_{j\in[N]} w_j c_j. \tag{22}$$

By construction, $w_jc_j = \frac{1}{C}$ for all $j \in [N]$. Therefore, $\frac{2}{w_i}\max_{j\in[N]} w_jc_j = \frac{2Cc_i}{C} = 2c_i$. 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 $\mathcal{G} = \big\{h \in \mathcal{H} : \widehat{\mathcal{L}}_i(h) = \widehat{\mathcal{L}}_i(\hat{h}^*) \text{ for } i \in \{N_{\text{nt}} + 1, \ldots, N\}\big\}$. Note that also $\mathcal{G} = \big\{h \in \mathcal{H} : \mathcal{L}_i(h) = \mathcal{L}_i(h^*) \text{ for } i \in \{N_{\text{nt}} + 1, \ldots, N\}\big\}$, 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 $i \in [N_{\text{nt}}]$

$$\mathcal{L}_i(\hat{h}^*) \leq \mathcal{L}_i(h^*) + \mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}) \tag{23}$$

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 (23) holds also for these (with $\mathcal{C}_i(n_i, \mathcal{H}, \delta/N_{\text{nt}}) = 0$), which concludes the proof.

Proof of Theorem 5 We construct a concrete counterexample that exploits the classic *overfitting* (or *bias-variance trade-off*) phenomenon of single-objective supervised learning (Vapnik, 2013).

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 $2k$ (Shalev-Shwartz & Ben-David, 2014), 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 $\mathcal{L}_1 = 0$. 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.

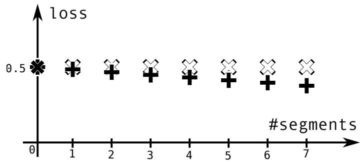


Figure 2: Proof illustration: Theorem 5 for $N = 2$

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.

Proof of Corollary 1 We first show that the scalarization used in Cortes et al. (2020) is Lipschitz-continuous with respect to a suitably constructed norm.

Lemma 9. *The scalarization* $\mathcal{U}_W(x_1, \dots, x_N) = \max_{w \in W} \sum_{i=1}^N w_i |x_i|$ *is* β*-Lipschitz with respect to the* $\frac{1}{M_i}$*-weighted* L^∞*-norm.*

Proof. For all $x_1, \dots, x_N$ and $x'_1, \dots, x'_N$ it holds

$$\mathcal{U}_W(x_1, \dots, x_N) - \mathcal{U}_W(x'_1, \dots, x'_N) = \max_{w \in W} \sum_{i=1}^N w_i |x_i| - \max_{w' \in W} \sum_{i=1}^N w'_i |x'_i| \tag{24}$$

$$\leq \max_{w \in W} \sum_{i=1}^N w_i |x_i - x'_i| = \max_{w \in W} \sum_{i=1}^N w_i \frac{M_i}{M_i} |x_i - x'_i| \tag{25}$$

$$\leq \max_{w \in W} \sum_{i=1}^N w_i M_i \Big(\max_{j=1,\dots,N} \frac{|x_j - x'_j|}{M_j} \Big) = \beta \max_{i=1,\dots,N} \frac{|x_i - x'_i|}{M_i} \tag{26}$$

Because of the symmetry between x and x' on the right hand side, the same inequality also holds for the absolute value of the left hand side. This proves the lemma. ☐

Next, we show that Assumption A holds: for each loss function, ℓ_i, we know that it is M_i-Lipschitz over a domain of radius at most D. Consequently, a standard Rademacher generalization bound for Lipschitz functions (Mohri et al., 2018, Theorem 11.3) yields that with probability at least $1 - \delta$ it holds:

$$\mathcal{L}_i(h) \leq \widehat{\mathcal{L}}_i(h) + 2M_i \hat{\mathfrak{R}}_S(\mathcal{H}) + 3M_i D \sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \tag{27}$$

Now applying Theorem 2 for the scalarization $\mathcal{U}_W$ yields

$$\mathcal{L}_W(h) \leq \widehat{\mathcal{L}}_W(h) + \beta \max_{i=1,\dots N} \frac{1}{M_i} \Big[2M_i \hat{\mathfrak{R}}_S(\mathcal{H}) + 3M_i D \sqrt{\frac{\log \frac{2N}{\delta}}{2n}} \Big]. \tag{28}$$

$$= \widehat{\mathcal{L}}_W(h) + 2\beta \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\beta D \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}. \tag{29}$$

This is exactly the statement of the corollary.

B Appendix – application-specific bounds

In this section we derive some generalization bounds for situations we discuss in Section 4. Our goal is not provide the tightest possible bound, but rather demonstrate how generalization bounds for new objectives can be derived using techniques readily available from the literature.

B.1 Activation sparsity

Let $f : \mathcal{X} \to \mathbb{R}^D$ be a vector-valued function, e.g. the activation values at some internal layer of a neural network. Denote by $\mathcal{F}$ the set of all such functions under consideration, e.g., for all network parameters.

We define the *activation sparsity* objective of $f \in \mathcal{F}$ as its expected *normalized* L^0-*"norm"* (Donoho, 2006) under the data distribution, i.e.

$$\mathcal{L}(f) = \mathbb{E}_{x\sim d}[L(f(x))] \quad \text{with} \quad L(f(x)) = \frac{1}{D}\sum_{j=1}^{D} \ell(f_j(x)) \quad \text{for} \quad \ell(t) = [\![t > 0]\!], \tag{30}$$

where $f_1, \dots, f_d : \mathcal{X} \to \mathbb{R}$ are the components of f. As empirical estimate we use the empirical *ρ-scaled ramp loss*, which for any $\rho > 0$ is defined as

$$\widehat{\mathcal{L}}_\rho(f) = \frac{1}{n}\sum_{i=1}^{n} L_\rho(f(x_i)) \quad \text{for} \quad L_\rho(f(x)) = \frac{1}{D}\sum_{j=1}^{D} \ell_\rho(f_j(x)), \quad \text{with} \quad \ell_\rho(t) = \begin{cases} 1 & \text{for } t \geq 0, \\ \frac{t}{\rho} + 1 & \text{for } -\rho \leq t < 0, \\ 0 & \text{for } t \leq -\rho. \end{cases} \tag{31}$$

Because L_ρ upper bounds L, it follows from standard concentration results, e.g. Mohri et al. (2018), that

$$\mathcal{L}(f) \leq \widehat{\mathcal{L}}_\rho(f) + 2\mathfrak{R}(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{32}$$

where $\mathcal{G} = \{L_\rho \circ f : f \in \mathcal{F}\}$. Because L_ρ is $\frac{\rho}{\sqrt{D}}$-Lipschitz-continuous, it follows from vector-valued concentration results (Maurer, 2016) that

$$\leq \widehat{\mathcal{L}}(f) + \frac{2\sqrt{2}}{D}\sum_{i=1}^{D} \mathfrak{R}(\mathcal{F}_i) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{33}$$

where $\mathfrak{R}(\mathcal{F}_i)$ is the Rademacher complexity of the i-th component functions for $F \in \mathcal{F}$. For typical neural network layers, all $\mathcal{F}_i$ for $i \in [D]$ are identical, and their Rademacher complexity can be bounded in terms of properties of the network weight matrices (Bartlett et al., 2017; Neyshabur et al., 2015).

B.2 Adaptive computation

We formalize the *adaptive computation* setting as a sequence of functions, $f_1, \dots, f_D : \mathcal{X} \to \{0, 1\}$, which are evaluated sequentially: if $f_i(x) = 0$ for some $i \in [D]$, then the computation stops after this stage, otherwise it continues with the next function. For simplicity of notation, we assume that $f_j(x) = 0$ for all $j > i$ then. If $f_i(x) = 0$ for all $i \in [D]$, the process nevertheless stops after f_D. The hypothesis set is the set of all permitted function tuples. We assume that the time it takes to evaluate a function f_i is the same for all such functions and independent of its argument. We denote it by $t_i \geq 0$. Then, for any hypothesis, $h = (f_1, \dots, f_D)$, the overall runtime for any $x \in \mathcal{X}$ is $T_h(x) = t_1 + \sum_{i=1}^{D-1} t_{i+1} f_i(x)$. We define the *adaptive runtime* objective $\mathcal{L}(h) = \mathbb{E}_x[T_h(x)]$ and its empirical counterpart $\widehat{\mathcal{L}}(h) = \frac{1}{n}\sum_{x\in S} T_h(x)$.

To show a generalization bound, we first observe that each f_i is a binary classifier. Denoting by $\mathcal{H}_i$ the set of all such classifiers, we obtain individual generalization bounds of standard form

$$\forall\delta \in (0,1) \quad \Pr\Big\{\forall f_i \in \mathcal{H}_i : \Big|\mathbb{E}[f_i(x)] - \frac{1}{n}\sum_{i=1}^{n} f_i(x)\Big| \leq \mathcal{C}_i(n, \mathcal{H}_i, \delta)\Big\} \geq 1 - \delta, \tag{34}$$

for example with $\mathcal{C}_i(n, \mathcal{H}_i, \delta) = \sqrt{\frac{8d_i \log(n/d_i) + 8\log\frac{4}{\delta}}{n}}$, where d_i is the VC-dimension of $\mathcal{H}_i$ (Shalev-Shwartz & Ben-David, 2014). By a union bound, we obtain that all of these hold simultaneously when δ is replaced by

δ/D in the generalization term. Using this result and the additive form of T_h, we obtain that with probability at least $1-\delta$ it holds jointly for $i \in [D]$ and all tuples $h = (f_1, \ldots, f_D) \in \prod_{i=1}^{D} \mathcal{H}_i$ that

$$|\mathcal{L}(h) - \widehat{\mathcal{L}}(h)| \leq \sum_{i=1}^{D-1} t_{i+1} \mathcal{C}_i(n, \mathcal{H}_i, \delta/D) \tag{35}$$

Because $\mathcal{H} \subset \prod_{i=1}^{D} \mathcal{H}_i$, this implies a generalization bound of the desired form, with generalization term $\mathcal{C}(n, \mathcal{H}, \delta)$ equal to the right hand side of (35).