Risk-Monotonicity in Statistical Learning

Zakaria Mhammedi*
Massachusetts Institute of Technology
mhammedi@mit.edu

Abstract

Acquisition of data is a difficult task in many applications of machine learning, and it is only natural that one hopes and expects the population risk to decrease (better performance) monotonically with increasing data points. It turns out, somewhat surprisingly, that this is not the case even for the most standard algorithms that minimize the empirical risk. Non-monotonic behavior of the risk and instability in training have manifested and appeared in the popular deep learning paradigm under the description of double descent. These problems highlight the current lack of understanding of learning algorithms and generalization. It is, therefore, crucial to pursue this concern and provide a characterization of such behavior. In this paper, we derive the first consistent and risk-monotonic (in high probability) algorithms for a general statistical learning setting under weak assumptions, consequently answering some questions posed by [53] on how to avoid non-monotonic behavior of risk curves. We further show that risk monotonicity need not necessarily come at the price of worse excess risk rates. To achieve this, we derive new empirical Bernstein-like concentration inequalities of independent interest that hold for certain non-i.i.d. processes such as Martingale Difference Sequences.

1 Introduction

Guarantees on the performance of machine learning algorithms are desirable, especially given the widespread deployment. A traditional performance guarantee often takes the form of a generalization bound, where the expected risk associated with hypotheses returned by an algorithm is bounded in terms of the corresponding empirical risk plus an additive error which typically converges to zero as the sample size increases. However, interpreting such bounds is not always straightforward and can be somewhat ambiguous. In particular, given that the error term in these bounds goes to zero, it is tempting to conclude that more data would monotonically decrease the expected risk of an algorithm such as the Empirical Risk Minimizer (ERM). However, this is not always the case; for example, [33] showed that increasing the sample size by one, can sometimes make the test performance worse in expectation for commonly used algorithms such as ERM in popular settings including linear regression. This type of non-monotonic behavior is still poorly understood and indeed not a desirable feature of an algorithm since it is expensive to acquire more data in many applications.

Non-monotonic behavior of risk curves [47]—the curve of the expected risk as a function of the sample size—has been observed in many previous works [18, 43, 48, 20] (see also [33, 52] for nice accounts of the literature). At least two phenomena have been identified as being the cause behind such behavior. The first one, coined peaking [30, 19], or double descent according to more recent literature [6, 49, 7, 17, 15, 37, 40, 41, 16, 12, 11, 14, 42], is the phenomenon where the risk curve peaks at a certain sample size $n$. This sample size typically represents the cross-over point from an over-parameterized to under-parameterized model. For example, when the number of data points is less than the number of parameters of a model (over-parameterized model), such as Neural

*Work done while at the Australian National University.

Networks, the expected risk can typically increase until the number of data points exceeds the number of parameters (under-parameterized model). The second phenomenon is known as dipping \[^{32,31}\], where the risk curve reaches a minimum at a certain sample size \(n\) and increases after that—never reaching the minimum again even for very large \(n\). This phenomenon typically happens when the algorithm is trained on a surrogate loss that differs from the one used to evaluate the risk \[^{8}\].

It is becoming more apparent that the two phenomena just mentioned (double descent and dipping) do not fully characterize when non-monotonic risk behavior occurs \[^{34}\]. \[^{33}\] showed that non-monotonic risk behavior could happen outside these settings and formally prove that the risk curve of ERM is non-monotonic in linear regression with prevalent losses. The most striking aspect of their findings is that the risk curves in some of the cases they study can display a perpetual “oscillating” behavior; there is no sample size beyond which the risk curve becomes monotone—see Figure 1. In such cases, the risk’s non-monotonicity cannot be attributed to the peaking/double descent phenomenon. Moreover, they rule out the dipping phenomenon by studying the ERM on the actual loss (not a surrogate loss).

The findings of \[^{33}\] stress our current lack of understanding of generalization. This was echoed more particularly by \[^{33}\], who posed the following question as part of a COLT open problem:

\[
\text{How can we provably avoid non-monotonic behavior?}
\]

While excess risk bounds are typically monotonic, this does not guarantee the monotonicity of the actual risk. In this work, we study under which assumptions on the learning problem there exist consistent and risk monotonic algorithms. We also aim to quantify the price to pay, in terms of corresponding excess risk rates, for achieving risk monotonicity.

**Contributions.** In this work, we answer some questions posed by \[^{53}\] by presenting an algorithm that is both consistent and risk-monotonic in high probability under weak assumptions on the learning problem. Our algorithm is technically a “wrapper” that takes as input any base learning algorithm \(\mathcal{B}\) and makes up a new algorithm \(\mathcal{A}\) that is risk monotonic in high probability and enjoys essentially the same excess risk rate as \(\mathcal{B}\). Crucially, our results show that risk monotonicity need not come at the expense of worse excess risk rates. In fact, we show that fast rates are achievable under a Bernstein condition (Definition \[^{5}\]).

Our results hold under the general statistical learning setting with a bounded loss. We even go beyond the standard i.i.d. assumption on the loss process. Our relaxed technical condition on the loss process, which is formalized in Assumption \[^{1}\] below, is reminiscent of the condition characterizing Martingale Difference Sequences (MDS). In a nutshell, we will assume a setting where the instance random variables \(Z_1, Z_2, \ldots\) and the loss \(\ell\) satisfy, for all hypotheses \(h\), \(\mathbb{E}[\ell(h, Z_t) \mid Z_1, \ldots, Z_{t-1}] = L(h)\), for some risk function \(L\). This is trivially satisfied in the i.i.d. case, where \(L\) corresponds to the standard risk function. In general, this condition may be satisfied even if \(Z_1, Z_2, \ldots\) are dependent or have different marginal distributions. We argue that our relaxed assumption on the loss process is the weakest assumption under which studying risk-monotonicity still makes sense.

To achieve risk monotonicity under our loss process assumption, we derive a new concentration inequality/generalization bound of PAC-Bayesian flavor for MDS (see Proposition \[^{5}\]). This concentration inequality may be thought of as an empirical Freedman’s inequality \[^{23}\] or as an extension of the empirical Bernstein inequality \[^{35}\] to MDS. Our concentration inequalities also have the advantage of being time-uniform with the optimal dependence on the number of samples. Here, time-uniform means that the inequalities hold for all sample sizes simultaneously. While standard concentration inequalities can be turned into time-uniform ones using a union bound over the number of samples.
$n$, the resulting bounds will have a sub-optimal $\ln n$ factor instead of the optimal $\ln \ln n$ that we are able to get. Finally, our concentration bounds are easily derived using the guarantee of a recent parameter-free online learning algorithm—\textsc{FreeGrad} \cite{zhang2019}. Our approach opens up the door for obtaining new concentration inequalities through the design of online learning algorithms.

**Approach Overview.** Our approach to deriving the new concentration inequalities is based on the guarantee of the recent \textsc{FreeGrad} algorithm. The algorithm operates in rounds, where at each round $t$, \textsc{FreeGrad} outputs $\tilde{w}_t$ in some convex set $\mathcal{W}$, say $\mathbb{R}^d$, then observes a vector $g_t \in \mathbb{R}^d$, typically the sub-gradient of a loss function at the iterate $\tilde{w}_t$. The algorithm guarantees a regret bound of the form $\sum_{t=1}^T g_t^\top (\tilde{w}_t - w) \leq O(\|w\|\sqrt{Q_T})$, for all $w \in \mathcal{W}$, where $Q_T := \sum_{t=1}^T \|g_t\|^2$. What is more, \textsc{FreeGrad}'s outputs ($\tilde{w}_t$) ensure the following (see \cite[Theorem 5]{zhang2019}):

$$\tilde{w}_t^\top g_t + \Phi(S_t, Q_t) \leq \Phi(S_{t-1}, Q_{t-1}), \quad \forall t \geq 1,$$

where $S_t := \|\sum_{i=1}^t g_i\|$, $Q_t := \sum_{i=1}^t \|g_i\|^2$, and $\Phi(S, V) := \exp\left(\frac{S^2}{2V}\right) - \frac{1}{2} \ln \frac{S^2}{V}$. Instantiating this guarantee in 1d with $(g_t)$ set to an MDS $(X_t)$ and taking (conditional) expectation in \cite{zhang2019} shows that $\Phi_t := \Phi(\sum_{i=1}^t X_i, \sum_{i=1}^t X_i^2)$ is a non-negative supermartingale, from which concentration results can be obtained via Ville’s inequality (a generalization of Markov’s inequality—see Lemma\cite{zhang2019}). Our proof technique is similar to the one introduced in \cite{zhang2019}, with the difference that we use the specific shape of \textsc{FreeGrad}'s potential function to build our supermartingale, which leads to a desirable empirical variance term in the final concentration bound.

On the side of risk monotonicity, given $n$ samples, the key idea behind our approach is to iteratively generate a sequence of distributions $P_1, P_2, \ldots$ leading up to $P_n$ over hypotheses, where we only allow consecutive distributions, say $P_{k-1}$ and $P_k$ to differ if we can guarantee (with high enough confidence) that the risk associated with $P_k$ is lower than that of $P_{k-1}$. To test for this, we compare the average empirical losses of hypotheses sampled from $P_{k-1}$ versus ones sampled from $P_k$, taking into account the potential gap between empirical and population expectations. Applying our new concentration bounds to quantify this gap not only allows us to achieve risk monotonicity under a non-i.i.d. loss process but also enables us to achieve fast excess risk rates under the Bernstein condition. For the latter, it was crucial to have an empirical loss variance term in the concentration inequality.

**Related Works.** Much work has already been done in efforts to mitigate the non-monotonic behavior of risk curves \cite{zhang2019, zhang2021, zhang2022}. For example, in the supervised learning setting with the zero-one loss, \cite{zhang2019} introduced the “memorize” algorithm that predicts the majority label on any test instance $x$ that was observed during training; otherwise, a default label is predicted. \cite{zhang2019} showed that this algorithm is risk-monotonic. However, it is unclear how their result could generalize beyond the particular setting they considered. Risk-monotonic algorithms are also known for the case where the model is correctly specified (see \cite{zhang2022} for an overview); in this paper, we do not make such an assumption.

Closer to our work is that of \cite{zhang2019} who, like us, also used the idea of only updating the current predictor for sample size $n$ if it has a lower risk than the predictor for sample size $n-1$. They determine whether this is the case by performing statistical tests on a validation set (or through cross-validation). They introduce algorithm wrappers that ensure that the risk curves of the final algorithms are monotonic with high probability. However, their results are specialized to the 0-1 loss and they do not answer the question by \cite{zhang2021} on the existence of learners that guarantee a monotonic risk in expectation.

On the side of concentration bounds for non-i.i.d. processes, our results are somewhat similar to those found in e.g. \cite{zhang2019, zhang2021}. However, our technique for deriving them, which relies on the guarantee of a parameter-free online learning algorithm, is entirely different. The theoretical link between online regret and concentration inequalities was previously drawn—see e.g. \cite{zhang2019, zhang2021}. However, our approach is slightly different as we use the monotonicity of an online algorithm’s potential function to get our concentration results. Thus, new concentration inequalities may be derived similarly by modifying the explicit potential function directly. Our approach is more similar to that of \cite{zhang2019} who also derived concentration inequalities using guarantees of online betting algorithms that bet fractions smaller than one of their wealth at each round.

\footnote{One can not improve on the $\ln \ln n$ factor by the law of iterated logarithm \cite{zhang2019}.}
We will not assume that the random variables \( Z_i \) are independent and identically distributed. Instead, we will make the following weaker assumption on the loss process \( \ell(h, Z_t) \):

**Assumption 1 (Process Assumption).** There exists a risk function \( L : \mathcal{H} \to [0, 1] \) such that the sequence of random variables \( Z_1, Z_2, \ldots \) satisfy \( L(Q) = E_{h,Z} \left[ L(h, Z_t) \right], \) for all \( Q \in \triangle(\mathcal{H}) \), where \( L \) is as in Assumption 1.

A learning algorithm \( A \) is a map from \( \bigcup_{i=1}^{\infty} Z^i \) to \( \triangle(\mathcal{H}) \); given data \( Z_{1:t} \), the output \( A(Z_{1:t}) \) of the algorithm is a distribution over hypotheses in \( \mathcal{H} \). This definition includes deterministic algorithms for which the distribution \( A(Z_{1:t}) \) is a Dirac at some \( h = h(Z_{1:t}) \). We will use the notation \( A(\cdot | Z_{1:t}) := A(Z_{1:t})(\cdot) \).

Throughout, we will present an algorithm wrapper that takes any base algorithm \( B \) as input and makes a risk-monotonic algorithm out of it with essentially the same excess risk rate. The results of this paper are useful for base algorithms that output distributions that are absolutely continuous w.r.t. our choice of prior \( P_0 \); that is \( B(Z_{1:t}) \ll P_0 \) for all \( t \geq 1 \). In practice, if \( \mathcal{H} = \mathbb{R}^d \), \( P_0 \) may be a multivariate Gaussian around the origin and \( B(Z_{1:t}) \) may also be a multivariate Gaussian around the ERM \( \hat{h}_n = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell(h, Z_i) \), in which case \( B(Z_{1:t}) \ll P_0 \) holds for all \( t \geq 1 \). We now define the notion of risk monotonicity we will work with:

**Definition 1 (Risk Monotonicity).** For \( \delta \in (0, 1) \) and \( N \geq 1 \), we say that a learning algorithm \( A : \bigcup_{i=1}^{\infty} Z^i \to \triangle(\mathcal{H}) \) is \((\delta, N)\)-risk-monotonic if, with probability at least \( 1 - \delta \),

\[
\forall t \geq N, \quad E_{A(h|Z_{1:t})} \left[ E_{h} \left[ \ell(h, Z_{t+1}) \right] \right] \leq E_{A(h|Z_{1:t})} \left[ E_{h} \left[ \ell(h, Z_{t+1}) \right] \right].
\]

---

1. We inherit this restriction from the PAC-Bayesian approach that we use to quantify generalization. In Appendix B we show how this restriction can be removed in the i.i.d. setting (see also Remark 7).
We will go beyond the notation of consistency and study the rate of convergence of the risk of our algorithm \( A \) on the next sample, conditioned on the past data, to decrease with the size of the data. We note that if Assumption 1 does not hold and \( \mathbb{E}[(\ell(h, Z_t))] \) depends on \( t \) in an arbitrary fashion, then requiring \( \mathbf{3} \) would be too strong. Thus, we will restrict our attention to processes that satisfy Assumption 2. We stress that the condition in this assumption is weaker than i.i.d., and allows the random variables \( Z_1, Z_2, \ldots \) to have different (conditional) distributions as long as the moment constraint \( \mathbb{E}_{t-1}[(\ell(h, Z_t))] = L(h) \) is satisfied for all \( h \in \mathcal{H} \) and \( t \geq 1 \).

The notion of monotonicity presented in \( \mathbf{33, 53} \) concerned only i.i.d. random variables, and requires the risk to be monotonic in expectation as opposed to in high probability. In particular, the strongest notion of monotonicity in \( \mathbf{33} \), which they refer to as global \( \mathbf{Z} \)-monotonicity, can be expressed as

\[
\forall t \geq 1, \quad \mathbb{E}[E_A(h|Z_t)](L(h)) \leq \mathbb{E}[E_A(h|Z_t)](L(h)),
\]

where \( L(h) : \mathbb{E}_{P_Z(z)}[\ell(h, z)] \) and \( Z_1, Z_2, \ldots \) i.i.d. \( P_Z \). We will show that achieving risk-monotonicity in expectation up to a small fast rate term is as easy as achieving \( (\delta, N) \)-risk-monotonicity—at least for bounded losses. In fact, we will show how our risk-monotonic (in the sense of Def. 1) algorithm can easily be turned into one that satisfies (5) up to a fast rate term.

Monotonicity alone is rather easy to achieve; it suffices to output a fixed hypothesis \( h \in \mathcal{H} \) regardless of the training dataset. In this case, the risk would be constant, and so risk monotonicity is achieved by definition. In practice, it is important to generate hypotheses with low risk, and so a fixed hypothesis that does not dependent on data is likely to be useless. Formally, we want algorithms that are risk monotonic and consistent:

**Definition 2** (Consistency). Under Assumption 2, we say that algorithm \( A \) is consistent if for any \( \epsilon > 0 \),

\[
\lim_{n \to \infty} \mathbb{P} \left[ \left| \mathbb{E}[E_A(h|Z_n)](L(h)) - \inf_{h \in \mathcal{H}} L(h) \right| > \epsilon \right] = 0.
\]

We will go beyond the notion of consistency and study the rate of convergence of the risk of our algorithm to the optimal risk. We do this under the assumption that the loss process \( \ell(h, Z_t) \) satisfies the Bernstein condition for \( h \in \mathcal{H} \):

**Definition 3** (Bernstein Condition). For \( \beta \in [0, 1] \) and \( B > 0 \), the \((\beta, B)\)-Bernstein condition holds if the random variables \( Z_1, Z_2, \ldots \) and the loss \( \ell \) satisfy, for all \( t \geq 1 \) and \( h \in \mathcal{H} \),

\[
\mathbb{E}_{t-1}[\left( \ell(h, Z_t) - \ell(h^*, Z_t) \right)^2] \leq B \mathbb{E}_{t-1}[\ell(h, Z_t) - \ell(h^*, Z_t)]^\beta,
\]

for \( h^* \in \arg \inf_{h \in \mathcal{H}} \mathbb{E}_{t-1}[\ell(h, Z_t)] \).

The Bernstein condition \( \mathbf{3, 4, 5, 21, 29} \) essentially characterizes the easiness of the learning problem. In particular, it implies that the conditional variance of the excess-loss random variable \( \ell(h, Z_t) - \ell(h^*, Z_t) \) vanishes when the risk associated with the hypothesis \( h \in \mathcal{H} \) gets closer to the \( \mathcal{H} \)-optimal risk \( L(h^*) \). For bounded loss functions, the Bernstein condition with \( \beta = 0 \) always holds, and so the results of this paper are always true for \( \beta = 0 \). The Bernstein condition with \( \beta = 1 \) corresponds to the easiest learning setting. The case where \( \beta \in (0, 1) \) interpolates naturally between these two extremes, where intermediate excess-risk rates are achievable. We refer the reader to \( \mathbf{29} \) Section 3) for examples of learning settings where a Bernstein condition holds.

**Additional useful definitions.** For \( \rho > 1 \) and \( \delta \in (0, 1) \), we define

\[
c := \sum_{k=1}^{\infty} \frac{1}{k \ln^2(k + 1)} \approx 3.2; \quad \phi_\rho(n) := c \sqrt{\rho + 1} \ln(n) + 1 \ln^2(\ln(n) + 2); \quad (4)
\]

and

\[
n_\delta := \sup \{ n \in \mathbb{N} : 8 \ln(\phi_\rho(n)/\delta) > n \} . \quad (5)
\]

Note that \( n_\delta \) is not too large as a function of \( 1/\delta \). In fact, the definitions of \( \phi_\rho \) and \( n_\delta \) imply that \( n_\delta \leq O(\ln(1/\delta)) \). In the next section, we present some new concentration inequalities of independent interest that will be useful in achieving risk-monotonicity under Assumption 1 while maintaining good excess risk rates.

### 3 New (PAC-Bayesian) Concentration Inequalities

In this section, we present some new concentration inequalities of PAC-Bayesian flavor that will be crucial to deriving our risk monotonic algorithm wrapper under Assumption 1. These concentration
inequalities hold for non-i.i.d. data (which we require to accommodate Assumption 1), and are so-called time-uniform; the inequalities hold for all sample sizes simultaneously given a fixed confidence level. We explain below the advantage that this has in our setting.

We start by a new concentration inequality for Martingale Difference Sequences, from which we derive the bound we need under Assumption 1. First, we give the formal definition of an MDS:

**Definition 4.** Let \((F_t)_{t \in \mathbb{N}}\) be a filtration w.r.t the underlying probability space \((\Omega, F, P)\), i.e. \((F_t)_{t \in \mathbb{N}}\) is a sequence of non-decreasing sub-\(\sigma\)-algebras of \(F\). A sequence of random variables \((X_t)\) is an MDS w.r.t. \((F_t)_{t \in \mathbb{N}}\) if for all \(t \geq 1\), \(X_t\) is \(F_t\)-measurable; \(E[|X_t|] < \infty\); and \(E[X_t | F_{t-1}] = 0\) a.s.

With this in hand, we present our first concentration inequality:

**Proposition 5 (PAC-Bayes for MDS).** Let \(\rho > 1\) and \(\phi_\rho\) be as in [4]. Further, let \(\{X_h^n\}\) be a family of random variables taking values in \([-1, 1]\) and \((F_t)_{t \in \mathbb{N}}\) be a filtration such that \((X_h^n)\) is an MDS w.r.t to \((F_t)_{t \in \mathbb{N}}\), for all \(h \in \mathcal{H}\). Then, for any distribution \(P_0\) on \(\mathcal{H}\) and all \(\delta \in (0, 1)\), we have

\[
P\left[ \forall n \geq 1, \forall P, \frac{E_{P(h)}[|S_h^n|]^2}{2(\rho + 1)E_{P(h)}[V_h^n] + 2E_{P(h)}[|S_h^n|]} \leq \text{KL}(P | P_0) + \ln \frac{\phi_\rho(n)}{\delta} \right] \geq 1 - \delta,
\]

where \(S_h^n := \sum_{t=1}^n X_t^h\) and \(V_h^n := 1 + \sum_{t=1}^n (X_t^h)^2\).

The proof of the theorem is in Appendix 5. The bound in Proposition 5 does not look like the typical PAC-Bayesian bound. However, simple algebra reveals that for any \(C > 0\),

\[
\frac{E_{P(h)}[|S_h^n|]^2}{2(\rho + 1)E_{P(h)}[V_h^n] + 2E_{P(h)}[|S_h^n|]} \leq C \implies E_{P(h)}[|S_h^n|] \leq C_1 n \text{ and all } \delta \in (0, 1),
\]

Combining this with the fact that \(E_{P(h)}[|S_h^n|] \leq E_{P(h)}[|S_h^n|]\) (by Jensen’s inequality), and (6), we obtain, under the same conditions as Proposition 5 that

\[
P\left[ \forall n \geq 1, \forall P \in \Delta(\mathcal{H}), E_{P(h)}[|S_h^n|] \leq 2C_n(P) + \sqrt{2(\rho + 1)E_{P(h)}[V_h^n]} \cdot C_1 \right] \geq 1 - \delta.
\]

where \(C_n(P) := \text{KL}(P | P_0) + \ln(\phi_\rho(n)/\delta)\). When \(\mathcal{H}\) is a singleton, the concentration inequality in [7] can be viewed as an empirical version of Freedman’s inequality [23] for MDS. The inequality is also reminiscent of the PAC-Bayesian empirical Bernstein inequality due to [50]. In addition to it holding for MDS, another advantage of (7) is that it is time-uniform—it holds for all sample sizes \(n\), simultaneously. While standard concentration inequalities that hold for a fixed sample size can be turned into time-uniform ones by applying a union bound over sample sizes, the resulting inequalities will have sub-optimal \(\ln n\) factors under the main square-root error term. In contrast, the concentration inequality in [7] has a \(\ln(\phi_\rho(n)) = O(\ln \ln n)\) term, which matches the optimal dependence in \(n\) according to the law of iterated logarithm [13]. MDS is a central concept in probability theory [44] and machine learning [10], and so our result in Proposition 5 is of independent interest.

Interestingly, the concentration inequality in Proposition 5 is derived using a guarantee of a parameter-free online algorithm—FREEGRAD. This opens the door for new ways of deriving such concentration inequalities through online learning algorithms, adding to existing results due to [45][22][28].

Using the result of Proposition 5, we are now going to derive a time-uniform “Empirical Bernstein” concentration inequality that holds under Assumption 1.

**Theorem 6.** Let \(\delta \in (0, 1), \rho > 1\), and \(\phi_\rho\) be as in [4]. Further, for \(P_0 \in \Delta(\mathcal{H})\) define \(\epsilon_n(P) := 2(\rho + 1)/n \cdot \text{KL}(P | P_0) + \ln 2\phi_\rho(n)/\delta\) and \(\mathcal{P}_n := \{P : 1 - \epsilon_n(P) > 0\}\). Under Assumption 1 we have

\[
P\left[ \forall n \geq n_{\delta}, \forall P \in \mathcal{P}_n, \frac{1}{n} \sum_{t=1}^n E_{P(h)}[\ell(h, Z_t) - L(h)] \leq \frac{\sqrt{\epsilon_n(P) \cdot \mathbb{V}_n(P) + \epsilon_n(P)/\rho + 1}}{1 - \epsilon_n(P)} \right] \geq 1 - \delta,
\]

where \(\mathbb{V}_n(P) := 1/n \sum_{t=1}^n E_{P(h)}[(\ell(h, Z_t) - 1/n \sum_{t=1}^n E_{P(h)}[\ell(h, Z_t)])^2] + \frac{1}{n}\) and \(n_{\delta}\) is as in (5).

The restriction that \(P \in \mathcal{P}_n\) in the theorem merely ensures that the denominator in the concentration bound remains positive. The set \(\mathcal{P}_n\) is guaranteed to be non-empty for all \(n \geq n_{\delta}\). In this case, \(P\) is in
We now move on to describing our risk monotonic procedure that makes use of our new concentration inequality. Let us first introduce our new concentration inequality in Theorem 6:

**Theorem 6** (Risk Monotonicity in Statistical Learning). Assume access to a base learning algorithm $\mathcal{B} : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \rightarrow \Delta(\mathcal{H})$ such that for $\rho > 1$, prior distribution $P_0$ on $\mathcal{H}$, and all $k \geq 1$,

$$\mathcal{B}(Z_{1:k}) \in \mathcal{Q}_k \coloneqq \{ Q \in \Delta(\mathcal{H}) : 16(\rho + 1) \text{KL}(Q \| P_0) \leq k \}. \quad (8)$$

With the right choice of gap sequence $(\xi_n)$, the procedure we present in Algorithm 1 takes as input a base learning algorithm $\mathcal{B} : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \rightarrow \Delta(\mathcal{H})$ together with samples $Z_{1:n}$ whose generating process satisfies Assumption 1, and returns a distribution over $\mathcal{H}$ that has a monotonic risk as a function of $n$ with high probability. By leveraging our new concentration inequality in Theorem 6 to specify the gap sequence $(\xi_n)$, we further show that achieving risk monotonicity need not deteriorate rates of convergence to the optimal risk $\inf_{h \in \mathcal{H}} L(h)$. In fact, we show that it is possible to attain fast rates under the Bernstein condition (Definition 3). To arrive at this result, it was crucial for our concentration inequality to have an empirical variance term.

**Algorithm 1** A Risk Monotonic Algorithm Wrapper

**Require:**
- A base learning algorithm $\mathcal{B} : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \rightarrow \Delta(\mathcal{H})$.
- Samples $Z_1, \ldots, Z_n$.
- A sequence of gap functions $(\xi_k)$, where $\xi_k = \xi_k(Z_{1:k}, \mathcal{B})$ for all $k$. \hfill $(\xi_k)$ will be set according to (9) [resp. (11)] for risk-monotonicity in probability [resp. expectation].

1: for $k = 1, \ldots, n$ do
2: \hspace{1em} if $\frac{1}{k} \sum_{i=1}^{k} \mathbb{E}_{B(h|Z_{1:k})}[\ell(h, Z_i)] - \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}_{P_{k-1}(h)}[\ell(h, Z_i)] \leq -\xi_k$ then
3: \hspace{2em} Set $P_k = \mathcal{B}(Z_{1:k})$.
4: \hspace{1em} else
5: \hspace{2em} Set $P_k = P_{k-1}$.
6: Return $P_n$.

To simplify our analysis, we will focus our attention on base algorithms that are restricted in the following way:

**Assumption 2** (Base Algorithm Restriction). We assume access to a base algorithm $\mathcal{B} : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \rightarrow \Delta(\mathcal{H})$ such that for $\rho > 1$, prior distribution $P_0$ on $\mathcal{H}$, and all $k \geq 1$,

$$\mathcal{B}(Z_{1:k}) \in \mathcal{Q}_k \coloneqq \{ Q \in \Delta(\mathcal{H}) : 16(\rho + 1) \text{KL}(Q \| P_0) \leq k \}. \quad (8)$$
In other words, we are restricting our attention to algorithms whose posteriors given samples of size $k$ do not have a KL divergence to the prior $P_0$ that grow too quickly with $k$. We make this assumption to satisfy the technical conditions needed for the concentration bound in Theorem 6 to be non-vacuous. Assumption 2 is reasonable even for large Neural Network models [55].

In practice, if $\mathcal{H}$ is $\mathbb{R}^d$, an option for $B$ is the algorithm that outputs a multivariate Gaussian distribution around a regularized ERM. The KL-divergence between the outputs of $B$ and $P_0$ can be controlled by tuning the regularisation parameter(s) and/or the covariance matrix of the multivariate Gaussian.

**Remark 7.** If one only cares about i.i.d. loss processes, Assumption 2 can be removed by using generalization bounds based on the standard Bernstein concentration inequality, e.g. [50, 36], which would also enable our risk decomposition in Theorem 9 below (and thus fast rates). We make Assumption 2 only to bound the denominator in our (non-i.i.d) concentration in Theorem 6 away from zero, which is not needed for other concentration bounds in the i.i.d. setting such as those in [50, 36] (see Appendix D for more detail).

To specify the sequence of gaps $(\xi_n)$ that our wrapper Algorithm 1 requires, we will use our concentration bound in Theorem 6. We recall that using this concentration bound instead of other existing ones allows us to I) achieve risk monotonicity under a weaker condition than i.i.d. on the loss process (Assumption 1 in this case); and II) to achieve potentially fast excess risk rates under the Bernstein condition. The latter is made possible by the fact that the concentration bound in Theorem 6 has an empirical loss variance term that allows a particularly useful decomposition of the excess risk under the Bernstein condition (see Theorem 9 below).

To give a concise expression of the gaps $(\xi_n)$, we let $Q_k := B(Z_{1:k}) \times P_{k-1}$, where $(P_k)$ are the intermediate distributions generated internally by Algorithm 1. With this, and the convention that $1/0 = +\infty$, we define

$$\xi_k = \frac{\sqrt{\epsilon_k \cdot \overline{V}_k} + 2\epsilon_k}{\sqrt{1 - \epsilon_k}}, \quad \text{where} \quad \epsilon_k = \frac{2 \text{KL}(Q_k \parallel P_0 \times P_0) + 2 \ln \frac{\phi_0(k)}{\delta}}{k \cdot (\rho + 1)^{-1}}; \quad \text{and} \quad \left(9\right)$$

$$\overline{V}_k = \frac{1}{k} \sum_{i=1}^{k} E_{Q_k(h, h')} \left[ (\ell(h, Z_t) - \ell(h', Z_t))^2 \right] - \left( \frac{1}{k} \sum_{i=1}^{k} E_{Q_k(h, h')} \left[ \ell(h, Z_t) - \ell(h', Z_t) \right] \right)^2.$$

We note that our Assumption 2 ensures that $\epsilon_n < 1/2$ (which in turn ensures that the gaps $(\xi_n)$ are not too large), for all $n \geq n_{\delta}$, where $n_{\delta}$ is as in (5). This follows from the fact that $\text{KL}(Q_k \parallel P_0 \times P_0) = \text{KL}(B(Z_{1:k}) \parallel P_0) + \text{KL}(P_{k-1} \parallel P_0)$, and that there exists $m < k$ such that $P_{k-1} = B(Z_{1:m}) \in Q_m \subseteq Q_k$ (by definition of $(P_k)$ in Algorithm 1). Furthermore, $Q_n = \emptyset$ for all $n \geq n_{\delta}$.

Before presenting our results for this section, we also note that Algorithm 1 assumes we can evaluate expectations over the output distributions of $B$. Such expectations can be approximated via Monte Carlo sampling, and any approximation errors need be added to the gaps $(\xi_n)$ to maintain the guarantees we present. Alternatively, one can avoid estimating expectations by applying recent derandomization techniques, see e.g. [46, 51], or by using other, non-PAC-Bayesian generalization bounds (see Appendix D).

We now state the guarantees of Algorithm 1. We start by the statement of risk-monotonicity (the proof is postponed to Appendix C).

**Theorem 8 (Risk Monotonicity).** Let $\delta \in (0, 1)$ and $n_{\delta}$ be as in (5). Under Assumptions 7 and 2 Algorithm 1 with $(\xi_n)$ as in (9) is $(\delta, n_{\delta})$-risk-monotonic according to Definition 7.

We now show that risk monotonicity need not come at a worse excess risk rate. Under the Bernstein condition, we have the following excess risk decomposition for the output of Algorithm 1.

**Theorem 9 (Risk Decomposition).** Let $B > 1$, $\beta \in [0, 1]$, and suppose that the $(\beta, B)$-Bernstein condition holds. Further, for $\varphi > 1$ and $\delta \in (0, 1)$, let $n_{\delta}$ and $\epsilon_n$ be as in (5) and (9), respectively. Then, under Assumptions 7 and 2 the outputs $(P_k)$ of Alg. 1 satisfy, with probability at least $1 - 2\delta$,

$$\forall n \geq n_{\delta}, \quad L(P_n) - L(h^*_\beta) \leq 3 \left( L(B(Z_{1:n}) - L(h^*_\beta)) + O(\epsilon_n) \right)^{1/\varphi}, \quad \left(10\right)$$

\[\text{Besides, when the KL divergence is very large, it is often very hard to infer non-vacuous generalization bounds (which one may view as a prerequisite to achieving risk monotonicity) using a PAC-Bayesian approach.}\]
We stress that this risk decomposition was only made possible by the fact that our concentration bound in Theorem 6 has an empirical loss variance term.

Theorem 9 shows that the excess risk of Algorithm 1 is at most a constant times the excess risk of the base algorithm B, plus a potentially lower-order term $O(\frac{1}{n})$. To appreciate what this additional term is doing, consider the case of a finite hypothesis class $\mathcal{H}$. In this case, the definition of $\epsilon_n$ in (9) implies that $\epsilon_n \leq O\left(\frac{1}{n} \ln \frac{\|\mathcal{H}\| \ln n}{s}\right)$. Thus, $\frac{\epsilon_n}{n}$ interpolates between the fast $\frac{1}{n} \ln \frac{\|\mathcal{H}\| \ln n}{s}$ rate under the Bernstein condition with $\beta = 1$ and the standard (up to log-log-factors) rate $\frac{1}{2} \ln \frac{\|\mathcal{H}\| \ln n}{s}$ under the Bernstein condition with $\beta = 0$, which we recall always holds for bounded losses. What is more, if $\mathcal{H}$ is finite and algorithm B is the ERM, i.e. if $B(Z_{1:k})$ is a Dirac at $\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \sum_{i=1}^n \ell(h, Z_i)$, then we have the following explicit excess risk rate for Algorithm 1:

**Proposition 10.** Under the setting of Theorem 9, if $\mathcal{H}$ is finite, $P_0$ is set to the uniform prior over $\mathcal{H}$, and $B(Z_{1:k})$ is a point mass around the ERM $h_n \in \arg\min_{h \in \mathcal{H}} \sum_{i=1}^n \ell(h, Z_i)$, then Alg. 1 satisfies, with probability at least $1 - 2\delta$,

$$
\forall n \geq n_0 \vee (16(\rho + 1) \ln \|\mathcal{H}\|), \quad L(P_n) - L(h_n) \leq O\left(\frac{\ln \frac{\|\mathcal{H}\| \ln n}{s}}{n}\right)^{\frac{1}{\tau}}.
$$

The story is not much different for a continuous set $\mathcal{H}$. The standard excess risk rate one would expect from algorithm B is $\sqrt{\frac{(KL(B(Z_{1:k}) \| P_0) + \ln(1/\delta))}{k}}$, which can dominate the right-most term in our risk decomposition (10) since $5$

$$
(\epsilon_n)^{\frac{1}{\tau}} \leq O\left(\max_{k \leq n} KL(B(Z_{1:k}) \| P_0) + \ln \frac{n \ln n}{s}\right)^{\frac{1}{\tau}}.
$$

Together, the above inequality and Theorem 9 show that fast rates for algorithm 1 are achievable whenever the Bernstein condition holds with $\beta > 0$ and the base algorithm B itself achieves a fast rate (when $\mathcal{H}$ is finite and B is the ERM). Proposition 10 shows that it is sufficient that $\beta > 0$. Thus, risk monotonicity need not come at the price of a worse rate of convergence to the optimal risk. Finally, we note that the factor 3 in our risk decomposition (10) is just an artifact of our analysis. In fact, by slightly modifying our proof of Theorem 9, one can show that any factor in the interval $(1, 3]$ is achievable at the cost of a larger lower-order term in (10).

**Risk monotonicity in expectation.** The original open problem due to [53] was around risk monotonicity in expectation—as in [5]. Our algorithm wrapper also allows us to almost achieve this notion of monotonicity with a slight modification of the sequence $\{\xi_k\}$ in (9); we essentially set the confidence level $\delta$ as a function of $k$. In fact, the next result shows that risk monotonicity in expectation is achievable up to an additive fast-rate term:

**Theorem 11.** Let $b \geq 1$, $\rho > 1$, and $\phi_b$ be as in (4). Under Assumptions 7 and Algorithm 1 with the sequence of gaps $\{\xi_k\}$ given by

$$
\xi_k = \xi_k := \frac{1}{1 - \epsilon_k}\left(\sqrt{\epsilon_k \cdot \mu_k + \frac{2\epsilon_k}{\rho + 1}}\right),
$$

with $\epsilon_k := \frac{2^{(\rho + 1)}}{\frac{1}{n}(KL(B(Z_{1:k}) \| P_0) + 2\ln(\phi_b(n)) + 2b \ln k)}$ and $(P_k)$ being as in Algorithm 1 satisfies, for all $t \geq N := \sup \{n : 8 \ln(n^b \phi_b(n)) > n\}$, $\{P_{t,1} \} \leq \{P_{t,1}(k) \}$ and $L(h) = \{L(h) \} + 1/t^b$.

We note that $N$ need not be too large since $\phi_b(n) \leq O(\ln n)$. However, while taking $b$ large will make the additive $1/t^b$ term in (12) small, it will increase the sample size $N$ beyond which (12) holds.
Ignoring the additive $1/t^b$ term, the notion of risk-monotonicity in theorem 11 corresponds to the notion of weak $Z$-monotonicity in [33]—one of their strongest notions of risk monotonicity since it holds for all data generating distributions (granted Assumption 1 holds, or the data is i.i.d.). We can modify Algorithm 1 to make it globally $Z$-monotonic, where (12) would hold for all $t \geq 1$ (instead of $t \geq N$), by forcing the outputs $P_k$ of the algorithm to be $P_0$ for all $k < N$.

With the same choice of gap sequence $(\xi_k)$ in Theorem 11 it is easy to show (following almost the same proof steps as in our previous results) that the excess risk rates in Theorem 9 and Proposition 10 will hold with probability $1/n^b$ (instead of $\delta$) for every sample size $n$.

5 Discussion and Future Work

The primary goal of this paper was to answer the fundamental question around the existence of a consistent, risk-monotonic algorithm in the general statistical learning setting. We answer this in the affirmative for a notation of monotonicity that holds with high probability and further show that there is virtually no cost for achieving this when it comes to excess-risk rates. We believe this is an important milestone in the search for risk-monotonic algorithms. It remains to see if risk-monotonicity in expectation is achievable; i.e. achieving (12) without the additive term.

From a computational perspective, the main setback of Algorithm 1 is that returning the distribution $P_n = P_n(Z_{1:n})$ requires $n$-calls to the base algorithm $B$. This implies that it may run $n$ times slower than $B$ in the worst case. However, when the base algorithm outputs distributions centered at ERMs, it may be possible to efficiently generate the sequence of distributions $(B(Z_{1:k}))_k$ by leveraging the fact that ERM solutions for sample sizes $k$ and $k + 1$ can be close to each other.

When the loss $\ell$ is convex in the first argument, there is no need for a randomized algorithm (i.e. we do not need $B(Z_{1:k}) \ll P_0$), and it is possible to efficiently generate a sequence of predictors $(\hat{h}_n)$ with monotonic risk using an online convex optimization algorithm as the base algorithm $B$. However, in general, it is unclear whether risk-monotonicity can be achieved without the (greedy) for-loop procedure of Algorithm 1. We note also that if one only wants a decreasing risk after some sample size $s \in \mathbb{N}$, then computing the distributions $(P_k)_{k<s}$ is unnecessary. In this case, the for-loop in Algorithm 1 need only start at $k = s$; the resulting hypotheses would satisfy the monotonicity condition in Definition 1 for all $n \geq s \lor n_\delta$.

Through Theorem 9, we showed that the excess risk of Algorithm 1 is at most three times that of the base algorithm plus a lower-order term. It remains to evaluate the empirical performance of the algorithm to identify for which applications the additional cost in the excess risk is worth it to achieve risk monotonicity.

Some important questions remain open along the axes of assumptions. In particular, can we remove the boundedness condition on the loss while retaining risk-monotonicity? It might be possible to achieve this for unbounded convex losses using the non-exponential weighted aggregation techniques recently suggested by [2]. Lifting the boundedness assumption may be key in resolving another COLT open problem [24] regarding achievable risk rates of log-loss Bayesian predictors. Our results build foundations for these avenues, which are promising subjects for future work.

Finally, it would be interesting to explore what other concentration inequalities for non-i.i.d. processes can be derived using other parameter-free online learning algorithms. An obvious starting point is to look at MATRIX-FREEGRAD [39].

---

Randomization is also not needed if one is only interested in i.i.d. processes (see Appendix D).
Acknowledgments and Disclosure of Funding

We would like to thank Hisham Husain for instrumental discussions during the early phases of the project. We thank Olivier Bousquet and his colleagues for pointing out a mistake in the original proof of Theorem 11. We also thank anonymous reviewers for their valuable feedback. This work was supported by the Australian Research Council and Data61.

References


