

Decision Theory & The Empirical Loss Principle

10716, Spring 2020

Pradeep Ravikumar

1 Preliminaries

Let $\theta \in \Theta$ denote the “state of nature”. We do not observe this state of nature, instead, we observe a random variable $X \in \mathcal{X}$ with distribution $P(\cdot; \theta)$ specified by the state of the nature. Our goal is to figure out some aspect of the state of nature (or perhaps simply the state of nature itself) just given the random sample X . Let \mathcal{A} denote the set of possible outputs, typically this will simply be Θ when we want to estimate the entire state of nature.

Let $\delta : \mathcal{X} \mapsto \mathcal{A}$ be an estimator. How good is this estimator? To answer this, we need the notion of a loss function $L : \Theta \times \mathcal{A} \mapsto \mathbb{R}$, so that $L(\theta^*, a)$ quantifies the cost of estimate $a \in \mathcal{A}$ when the state of nature is θ^* . While this quantifies the cost of an *estimate* $a \in \mathcal{A}$, we can also use this to quantify the cost of an *estimator* δ as:

$$R(\theta^*, \delta) = \mathbb{E}_{X \sim P(\cdot; \theta^*)} L(\theta^*, \delta(X)).$$

This quantity is also called the **risk** of the estimator δ . Now that we can evaluate the goodness of any estimator, we can then ask: what is the best possible estimator? As it turns out, for a fixed state of nature θ^* , for most risk functions, there is a very simple estimator $\delta(\cdot)$ that is optimal, namely: $\delta(X) = \theta^*$, which incurs the risk $R(\theta^*, \theta^*)$.

To circumvent this, we would need to define more global notions of optimality: minimax optimality, and Bayesian optimality.

Minimax Risk and Estimators. Let Γ be a set of candidate estimators. Then, the minimax risk is specified as:

$$r(\Theta, \Gamma) := \inf_{\delta \in \Gamma} \sup_{\theta^* \in \Theta} R(\theta^*, \delta),$$

and any estimator δ_{MM} that achieves this minimax risk is said to be a minimax optimal estimator. The main caveats with this notion are practical: it is typically difficult to certify, and moreover is also not practically constructive i.e. it requires solving a min-max problem over all candidate estimators, which is typically intractable. There are also criticisms of this notion as being overly conservative: if δ_1 is much much better than δ_2 for most states of nature θ , except for one θ' where it is marginally worse, minimax-optimality might well pick δ_2 .

Bayesian Risk and Estimators. Given a prior π over states of nature, we can define the Bayesian risk:

$$r(\pi, \delta) = \int_{\theta^* \in \Theta} R(\theta^*, \delta) \pi(\theta^*) d\theta^*,$$

and the estimator δ_π minimizing this Bayesian risk,

$$\delta_\pi \in \arg \inf_{\delta} r(\pi, \delta),$$

is said to be the optimal Bayesian estimator given prior π . Unlike the minimax case, the Bayes risk is easier to evaluate, or at least approximate, and moreover it is also easier to compute or at least approximate the optimal Bayesian estimator δ_π , by solving for the conditional Bayesian risk:

$$\delta_\pi(x) \in \arg \inf_{a \in \mathcal{A}} \int_{\theta^* \in \Theta} L(\theta^*, a) \pi(\theta^* | x) d\theta^*,$$

which unlike the minimax optimal case is a more tractable, typically even a finite dimensional estimation problem. The main caveat is that it requires the specification of a prior π , which essentially specifies the linear combination weights of how to combine the risks at different states of nature; and moreover such a linear combination might not capture the true notion of global risk. Notwithstanding the concerns with the specification of the prior, for medium to higher dimensional problems, the computations above again get intractable, so that this is not always a practical estimator for many modern data settings.

2 Estimation Principles

The minimax risk is associated with the so-called minimax principle:

Minimax Principle: An estimator δ_1 is preferred to another estimator δ_2 if its worst case risk is lower: $\max_{\theta^* \in \Theta} R(\theta^*, \delta_1) < \max_{\theta^* \in \Theta} R(\theta^*, \delta_2)$.

The minimax optimal estimator is that which achieves the minimum worst case risk, though this term is used even for estimators that achieve the minimax risk upto some absolute constants that do not depend on key quantities such as the sample size, or problem dimension or complexity parameters.

Similarly the Bayesian risk is associated with the so-called Bayes risk principle:

Bayes Risk Principle: An estimator δ_1 is preferred to another estimator δ_2 if its Bayes risk is lower: $r(\pi, \delta_1) < r(\pi, \delta_2)$.

The Bayes estimator is that which achieves the Bayes risk. A related principle is the so-called conditional Bayes principle:

Conditional Bayes Principle: Given samples X , choose a decision or action $a \in \mathcal{A}$ which minimizes

$$\rho(P(\theta^*|X), a) := \int_{\theta^* \in \Theta} L(\theta^*, a) P(\theta^*|X) d\theta^*,$$

where $P(\theta^*|X)$ is the posterior distribution of the state of nature θ^* given the samples X .

It can be seen that the Bayes risk principle and the conditional Bayes principle yield the same answers since $r(\pi, \delta) = \int_X \rho(P(\theta^*|X), \delta(X)) P(X) dX$, which can be minimized over estimators $\delta(\cdot)$ by, for each sample X , setting $\delta(X)$ to the minimizer of $\rho(P(\theta^*|X), a)$ which is precisely the conditional Bayes Principle.

While these are the most natural principles given the global optimality notions of minimax and Bayesian risks, as noted above, these are typically not practical, and indeed are not the typical class of estimators used in practice in statistical ML. How then do we come up with practical estimators that are principled from a statistical decision theory standpoint?

But before a study of principled and practical statistical ML estimators, it is instructive to consider the criticism of the notion of expected risk $R(\theta^*, \delta)$ itself: that it does not evaluate an estimator in light of the given set of observations; rather it computes its expected performance over all possible sample sets. This global performance might not be indicative of the local performance given the specific set of observations. This is best illustrated by the following example. Suppose the state of nature is $\theta \in \mathbb{R}$, given which the observation $X \in \mathbb{R}$ has the following distribution: $P_\theta(X = \theta + 1) = P_\theta(X = \theta - 1) = 1/2$. Suppose we are interested in estimating the state of nature θ^* , so that the action space $\mathcal{A} = \Theta$, and that we have the zero-one loss so that $L(\theta^*, \theta) = \mathbb{I}[\theta^* \neq \theta]$. Suppose we see two samples $X = (X_1, X_2)$ from $P(\cdot; \theta)$, and that the estimator is given by: $\delta(X) = \frac{1}{2}(X_1 + X_2)\mathbb{I}(X_1 \neq X_2) + (X_1 - 1)\mathbb{I}(X_1 = X_2)$.

Its risk is then given by $R(\theta^*, \delta) = \mathbb{P}[\delta(X) \neq \theta^*] = 0.25$, for all $\theta^* \in \Theta$. Let $E(X) = \mathbb{I}[X_1 = X_2]$ be the event that the two samples are identical. It can then be seen that conditioned on $E = 1$, the risk of the estimator is zero, since it necessarily is then the case that $\theta^* = (X_1 + X_2)/2$. While conditioned on $E = 0$, the risk of the estimator is 0.5, since θ^* could be either of $X_1 - 1$ or $X_1 + 1$ with equal chance. Thus, a global or unconditional risk $R(\theta^*, \delta)$ of 0.25 is misleading in both these cases, especially so when the observed sample X is such that $E(X) = 1$, and the estimator is actually always correct. This leads to the so-called conditionality principle, a weaker and easier stated version of which is as follows:

(Weak) Conditionality Principle. In order to estimate the state of nature θ , suppose we can perform two experiments E_1 or E_2 . Suppose J is a binary random variable, such that $J = 1$ or 2 with equal probability of $1/2$. Consider the mixed experiment, where we first sample the value of J , and then perform the experiment E_J . Then the information about θ obtained from the mixed experiment E_J should only depend on the experiment E_j that is actually performed.

This is also illustrated by the following example. Suppose that an engineer uses a voltmeter that makes 5 observations each ranging from 75 to 99 volts. He then asks the statistician to estimate the true voltage. The next day, the engineer says, by the way, it seems the voltmeter truncates voltages at 100, but that should not matter since all measured voltages were below 100. The statistician however is worried: not so fast, he says. This changes the distribution of the observations (since it is a truncated random variable), and hence the risk measure computations, and therefore he will have to redo his calculations. Wait, says the engineer, I should let you know I always carry a backup bulkier voltmeter that I use when I see observations equal to 100 (which might indicate truncation), but since I didn't see any values that were equal to 100, I didn't use it. Phew, says the statistician, I don't have to redo my calculations after all. The next day the engineer says, sorry again, it seems the backup bulkier voltmeter has not been working for the past many months, but surely that shouldn't matter, since we didn't have to use it. Oh but it does, says the Statistician with a sigh, we now have a truncated distribution again, so I have to redo the calculations after all.

In this example, the statistician with his focus on computing the expected risk is violating the conditionality principle, but also our common sense or rationality intuitions that the estimator should not have to worry about the observations being truncated at 100 conditioned on the fact that the observations were all lower than 100.

But how to operationalize this conditionality principle? The biggest advance towards this was the likelihood principle. Recall the definition of the likelihood function $\ell(\theta) = P(X|\theta)$, which is the density of the observed samples given state of nature θ , as a function of $\theta \in \Theta$.

Likelihood Principle. In making inferences about the state of nature θ after observing X , all relevant information is contained in the likelihood function $\ell(\theta) = P(X|\theta)$.

The likelihood principle was advocated in the 1950s by R. A. Fisher, and G. A. Barnard. Its importance has been bolstered by technical arguments such as Birnbaum [REF], who show that it is implied by the weak conditionality principle, and the sufficiency principle (which requires that estimators should be functions of sufficient statistics of the state of nature or parameters thereof). It is to be noted that the likelihood principle by itself is not actionable: in the sense that it is not clear how to construct an estimator that satisfies the likelihood principle. The Bayes estimator is one estimator that does follow the likelihood principle. When we wish to estimate the state θ itself, then the MLE also satisfies the likelihood principle, as we will see presently. But for more general parameters, it is not clear how to satisfy the likelihood principle, and in particular, how to reconcile the likelihood function with the loss function $L(\theta^*, a)$.

We next consider a new principle, that we will call the empirical loss principle.

Empirical Loss Principle. In making inferences about the state of nature θ after observing $X := \{X_i\}_{i=1}^n$, only use information $\{\ell(X_i, a)\}_{i=1}^n$, for some $\ell : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$.

This essentially entails that estimators only considers the losses of actions on individual samples, for some loss function $\ell : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$. We note that this is a more general variant of the likelihood principle, under the setting where the action space $\mathcal{A} = \Theta$, wherein we set $\ell(X_i, \theta) = P(X_i|\theta)$.

But for general settings, it is again not clear how to construct an estimator that satisfies the empirical loss principle. In particular, how do we reconcile the decision theoretic loss function $L : \Theta \times \mathcal{A} \mapsto \mathbb{R}$. In particular, how can we reconcile the decision-theoretic loss $L(\theta^*, a)$ in terms of an “empirical loss principle” based loss $\ell(X, a)$ that only depends on the observations X rather than on the unknown parameter θ^* directly?

3 Decomposable Losses and ERM

Consider the ℓ_p loss: $L(\theta^*, \theta) = \|\theta^* - \theta\|_p$. It is not obvious how to construct an estimator that satisfies the empirical loss principle with respect to this decision theoretic loss, since that would entail approximating this loss function just using samples from P_{θ^*} , and it is unclear if such an approximation is always possible. But as it turns out, there is a specific sub-class of decision-theoretic loss functions, that we will call **decomposable loss functions**, which are indeed simple to estimate given samples.

Definition 1 A loss function $L : \Theta \times \mathcal{A} \mapsto \mathbb{R}$ is said to be decomposable iff:

$$L(\theta^*, a) = \mathbb{E}_{X \sim P(\cdot; \theta^*)} \ell(X, a), \forall \theta^* \in \Theta, a \in \mathcal{A},$$

for some loss function $\ell : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$.

Thus, for decomposable decision-theoretic loss functions, we can express the loss $L(\theta^*, a)$ as the expectation of a term $\ell(X, a)$ that is entirely as a function of the sample X rather than the unknown state of nature θ^* . A key advantage to this is that we could then compute the so-called empirical loss:

$$\hat{L}_n(\theta^*, a) = \frac{1}{n} \sum_{i=1}^n \ell(X^{(i)}, a),$$

entirely using samples $\{X^{(i)}\}_{i=1}^n \sim \mathcal{P}(\cdot; \theta^*)$.

In what might perhaps be confusing terminology from a statistical decision theory standpoint, this empirical loss is typically called empirical risk in ML. Which then leads to the class of

empirical risk minimizers:

$$\begin{aligned}\hat{\delta}_{\text{ERM}} &:= \arg \min_{a \in \mathcal{A}} \hat{L}_n(\theta^*, a) \\ &= \arg \min_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \ell(X^{(i)}, a).\end{aligned}$$

It can be seen that empirical risk minimizers satisfy the empirical loss principle, and moreover, are statistical decision theoretically principled. Indeed, they could perhaps even be viewed as more principled than minimax optimal estimators (which might not satisfy the conditionality principle), and Bayes optimal estimators (that require the additional specification of a prior). The caveat of course is that they are only applicable to decomposable decision-theoretic loss functions. We next look at some examples of decomposable loss functions, and empirical risk minimizers.

3.1 Examples

MLE. Consider a family of distributions $\{P_\theta\}_{\theta \in \Theta}$. Suppose we are given samples $\{X_i\}_{i=1}^n \sim P_{\theta^*}$ for some $\theta^* \in \Theta$, and we wish to estimate θ^* given the n samples. Here the decision or action space $\mathcal{A} = \Theta$. Suppose $L(\theta^*, \theta) = KL(P_{\theta^*}, P_\theta)$. This can be seen to be decomposable since: $KL(P_{\theta^*}, P_\theta) = \mathbb{E}_{X \sim P_{\theta^*}} \log P_{\theta^*}(X)/P_\theta(X)$.

The corresponding empirical risk minimizer with respect to this loss is then given by:

$$\begin{aligned}\hat{\theta}_n &= \arg \inf_{\theta} \frac{1}{n} \sum_{i=1}^n \log P_{\theta^*}(X_i)/P_\theta(X_i) \\ &= \arg \inf_{\theta} \frac{1}{n} \sum_{i=1}^n -\log P_\theta(X_i),\end{aligned}$$

which is the Maximum Likelihood Estimator or MLE. This estimator thus satisfies the empirical loss principle as well as clearly, the likelihood principle. This example also makes clear that the likelihood principle, which entails that estimators only use the likelihood function as a summary of the data, is a special case of the empirical loss principle.

Binary Classification. Let $X \in \mathcal{X}$ denote the so-called input random variable, and $Y \in \{-1, +1\}$ a binary output random variable, jointly distributed as $(X, Y) \sim P \in \mathcal{P}$. Given observations $\{(X_i, Y_i)\}_{i=1}^n \sim P$, we wish to obtain a classifier $f : \mathcal{X} \mapsto \{-1, +1\}$ that minimizes the so-called mis-classification error: $L(P, f) = \mathbb{P}[f(X) \neq Y]$. Let $\ell(f, (X, Y)) = \mathbb{I}[f(X) \neq Y]$. It can then be seen that $L(P, f) = \mathbb{E}_{(X, Y) \sim P} \ell(f, (X, Y))$,

so that the mis-classification error is a decomposable decision-theoretic loss function. The corresponding empirical risk minimizer would then be given as:

$$\begin{aligned}\hat{f}_n &= \arg \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f, (X_i, Y_i)) \\ &= \arg \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}[f(X_i) \neq Y_i].\end{aligned}$$

The classifier

$$\begin{aligned}f^* &= \arg \inf_f L(P, f) \\ &= \text{sign}(P(Y = 1|\cdot) - 1/2)\end{aligned}$$

that minimizes the loss $L(P, f)$ directly is called the Bayes optimal classifier. This might again seem like a terminological mis-step in ML, since this does not seem Bayesian at all: f^* is just minimizing the pointwise loss $L(P, f)$, where the state of nature is $P \in \mathcal{P}$, and the decision or action is $f \in F$.

This does not seem Bayesian at all. What might a fully Bayesian treatment look like? Suppose we have some prior π over distributions in \mathcal{P} , and given the samples $(X, Y) = ((X_i, Y_i))_{i=1}^n$, we could then compute the posterior $\pi(\cdot|(X, Y))$, and compute the Bayes optimal estimate as:

$$f_{\text{Bayes}} = \arg \min_f \rho(\pi(\cdot|(X, Y)), f),$$

which can be seen to be much more complicated than what is called the Bayes optimal classifier in ML.

But there is a simple and principled reason why f^* above is called the Bayes classifier. Instead of taking $P \in \mathcal{P}$ to be the state of nature, we will consider $Y \in \{-1, 1\}$ to be the random state of nature, and set the action space as $\mathcal{A} = \{-1, 1\}$. The observations $X \in \mathcal{X}$, given the state Y , are then distributed as $P_{X|Y}(\cdot|Y)$. The decision rule $\delta : \mathcal{X} \mapsto \{-1, +1\}$ is then precisely a binary classifier that maps the inputs in \mathcal{X} to outputs in $\{-1, +1\}$. The decision theoretic loss is then set to $L(Y, a) = \mathbb{I}[Y \neq a]$. The corresponding risk for a decision rule $\delta(\cdot)$ is then given as:

$$\begin{aligned}R(Y, \delta) &= \mathbb{E}_{X \sim P(\cdot|Y)} L(Y, \delta(X)) \\ &= \mathbb{E}_{X \sim P(\cdot|Y)} \mathbb{I}(Y \neq \delta(X))\end{aligned}$$

so that with a prior distribution $P_Y(\cdot)$ over the state Y , we get the Bayes risk:

$$\begin{aligned}r(P_Y, \delta) &= \mathbb{E}_{Y \sim P_Y} \mathbb{E}_{X \sim P_{X|Y}(\cdot|Y)} \mathbb{I}(Y \neq \delta(X)) \\ &= \mathbb{E}_{(X, Y) \sim P} \mathbb{I}(Y \neq \delta(X)), \quad \text{where } P(X, Y) = P_Y(Y)P_{X|Y}(X|Y),\end{aligned}$$

which is precisely the mis-classification error $L(P, \delta)$. Thus, the classifier minimizing the mis-classification error f^* above is precisely the Bayes optimal classifier under this decision theoretic setup. Unlike typical Bayesian estimation settings however, both the prior P_Y , as well as the observation distribution $P_{X|Y}$ is unknown here, so that this is not actually actionable, but rather provides a characterization of the ideal classifier (i.e. assuming knowledge of $P(X, Y)$).

We note that there exist many other popular loss functions for binary classification that are not decomposable; see [REF] for a study of such non-decomposable loss functions. For instance, consider the precision loss function which is the fraction of true positives to the total number of predicted positives, so that

$$L(P, f) = P(Y = 1 | f(X) = 1) = \frac{P(Y = 1, f(X) = 1)}{P(f(X) = 1)},$$

which is not of the form $\mathbb{E}_{(X,Y) \sim P} \ell(f, (X, Y))$ for any $\ell(\cdot)$.

3.2 Plugin estimators

A close cousin of empirical risk minimization based estimators are so-called plugin estimators. We can distinguish between two classes of plugin estimators. In the first, we compute a plugin estimate of the loss itself, so that we approximate $L(P, f)$ by $L(P_n, f)$, where P_n is the empirical distribution given samples $\{X_i\}_{i=1}^n \sim P$. This then allows us to compute:

$$\hat{f}_{\text{PLUGIN;I}} = \arg \inf_{f \in \mathcal{F}} L(P_n, f). \quad (1)$$

Note that this does not require that the loss function be decomposable. For instance, for the precision loss above, this would entail solving for:

$$\arg \inf_{f \in \mathcal{F}} \frac{\sum_{i=1}^n \mathbb{I}[f(X_i) = Y_i = 1]}{\sum_{i=1}^n \mathbb{I}[f(X_i) = 1]}.$$

For decomposable losses, it can be seen that Eqn (1) directly reduces to empirical risk minimizers (ERMs).

The second class of plugin estimators is to first characterize the ideal estimator $f^*(P) = \arg \inf_f L(P, f)$, and then directly compute the plugin estimate: $\hat{f}_{\text{PLUGIN;II}} = f^*(P_n)$. This is not always a good idea in this exact form. For instance, with the zero-one loss function, we get $\hat{f}_{\text{PLUGIN;II}} = \text{sign}(P_n(Y = 1 | \cdot) - 1/2)$, where $P_n(Y = 1 | X)$ is simply the empirical conditional distribution, but which for continuous inputs X , will thus likely reduce to random guessing. Much more common is a related variant of computing a smoothed variant \tilde{P}_n (or for instance, fitting some statistical model such as logistic regression), and then using the plugin estimate $f^*(\tilde{P}_n)$.

4 Characterization of Decomposable Losses

A natural question then is: what classes of loss functions $L(\theta^*, a)$ are decomposable, and hence amenable to ERM like estimators? Note that in its general form, this loss could depend arbitrarily on the state of nature θ^* , and the action a . Denote the distribution over the observations given the state of nature by P_{θ^*} , and fix the action a . Assuming that the observation distributions are identifiable (i.e. we can recover θ from P_θ), the loss $L(\theta^*, a)$ could then be viewed as a functional of the distribution P_{θ^*} . For the loss to be decomposable as defined earlier, it would entail that:

$$L(\theta^*, a) = \mathbb{E}_{Z \in P_{\theta^*}} \ell_a(Z),$$

where we have used the overloaded notation $\ell_a(Z) := \ell(a, Z)$.

Generalizing this requirement, we can ask the following general question: given any distribution P , what loss functionals $\mathcal{L}(P)$ can be expressed as the expectation of some auxiliary loss evaluated at a random variable with distribution P ? In other words, when can we write:

$$\mathcal{L}(P) = \mathbb{E}_{Z \sim P}(\ell(Z)), \quad (2)$$

for some auxiliary loss function $\ell(\cdot)$ of a random variable with the same distribution as the argument to the loss function $\mathcal{L}(\cdot)$. As we saw earlier, not all possible loss functionals can have this form, but classical results from Utility Theory [REF] state some very reasonable sufficient conditions under which any loss functional will necessarily have the above form.

Let \mathcal{P} be some class of distributions. Then a loss functional $\mathcal{L} : \mathcal{P} \mapsto \mathbb{R}$ can be viewed as a preference order on \mathcal{P} , where $P_1 \prec P_2$ iff $\mathcal{R}(P_1) < \mathcal{R}(P_2)$, for $P_1, P_2 \in \mathcal{P}$, and so on.

Axiom 1. If $P_1, P_2 \in \mathcal{P}$, then either $P_1 \prec P_2$, $P_1 \succ P_2$, or $P_1 \sim P_2$.

This states that preferences exist among all elements of \mathcal{P} , which in the context of loss functionals, entails that it be defined over all of \mathcal{P} , which vacuously holds if \mathcal{P} is set to lie within the domain of the loss functional.

Axiom 2. If $P_1 \preceq P_2$, and $P_2 \preceq P_3$, then $P_1 \preceq P_3$.

This is a natural transitivity axiom, that also vacuously holds for loss functional based preferences.

Axiom 3. If $P_1 \prec P_2$, then $\alpha P_1 + (1-\alpha)P_3 \prec \alpha P_2 + (1-\alpha)P_3$, for any $\alpha \in (0, 1)$, and $P_3 \in \mathcal{P}$.

This states that if P_1 is preferred to P_2 , then given two random situations which are identical except that in one P_1 occurs with probability α , while in the other P_2 occurs with probability α ; the first random situation is to be preferred.

Axiom 4. If $P_1 \prec P_2 \prec P_3$, there exist $\alpha, \beta \in (0, 1)$ s.t.

$$\alpha P_1 + (1 - \alpha) P_3 \prec P_2, \quad \text{and} \quad P_2 \prec \beta P_1 + (1 - \beta) P_3.$$

This axiom loosely states that there are no infinitely bad or good distributions. In the context of loss functional based preferences, a sufficient condition for Axiom 4 to hold is that the loss functional be bounded over \mathcal{P} .

Theorem 2 (Degroot, 76) *Suppose the preference order corresponding to a loss functional $\mathcal{L} : \mathcal{P} \mapsto \mathbb{R}$, over distributions in some set of distributions \mathcal{P} , satisfies the four axioms above. Then, the loss functional has the form $\mathcal{L}(P) = \mathbb{E}_{Z \sim P} \ell(Z)$, for all $P \in \mathcal{P}$.*

Thus the class of decomposable loss functionals encompasses all “rational” loss functionals that satisfy the very reasonable axioms above.

5 Statistical Analysis: Pointwise Bounds

So far we have focused on decision theoretic principles that constrain, and at times even specify suitable classes of estimators. We now consider the related question of how to validate or certify a given class of estimators. Statistical decision theory also provides the framework to do so in a global sense, ranging over all possible states of nature $\theta^* \in \Theta$. In particular, some of the notions we looked at earlier, of Bayesian and minimax risk, could be used to quantify such global performance of the class of estimators.

But we might also be interested in *pointwise bounds* for any given state of nature θ^* such as:

$$L(\theta^*, \hat{a}) \leq \inf_{a \in \mathcal{A}} L(\theta^*, a) + r_{n, \theta^*},$$

with high probability given θ^* , and for some rate r_{n, θ^*} that depends on the number of samples n and θ^* . Such bounds are mainly of interest in frequentist analyses, where we assume the state of nature θ^* is fixed, but from a technical standpoint, one could always obtain uniform variants of these pointwise bounds (that hold uniformly over $\theta^* \in \Theta$), to then obtain bounds on the minimax risk and Bayesian risk.

Also, contrast the strong guarantee above with the much weaker (and also non-constructive) frequentist guarantee of admissibility: decision rule $\hat{\delta}$ is admissible if

$$\forall \delta' \exists \theta^* \in \Theta \text{ s.t. } R(\theta^*, \hat{\delta}) \leq R(\theta^*, \delta').$$

5.1 ERMs: Pointwise Bounds via Uniform Laws

For the specific estimator class of ERMs, there is a particular technical tool that helps us obtain such pointwise bounds, namely, so-called uniform laws. Given any estimate a , these provide uniform guarantees of the deviation of the empirical risk from the true risk:

$$r_{n;\text{UNIF}} := \sup_{\theta^* \in \Theta} |\widehat{L}_n(\theta^*, a) - L(\theta^*, a)|.$$

Given such a uniform law bound, we could then provide guarantees on the empirical risk minimizer (ERM):

$$\widehat{a}_{\text{ERM}} := \arg \min_{a \in \mathcal{A}} \widehat{L}_n(\theta^*, a),$$

by a simple chaining argument:

$$\begin{aligned} L(\theta^*, \widehat{a}_{\text{ERM}}) - L(\theta^*, a^*) &\leq L(\theta^*, \widehat{a}_{\text{ERM}}) - \widehat{L}_n(\theta^*, \widehat{a}_{\text{ERM}}) \\ &\quad + \widehat{L}_n(\theta^*, \widehat{a}_{\text{ERM}}) - \widehat{L}_n(\theta^*, a^*) \\ &\quad + \widehat{L}_n(\theta^*, a^*) - L(\theta^*, a^*) \\ &\leq 2r_{n;\text{UNIF}}, \end{aligned}$$

since the first and third terms are bounded by the uniform bound, and the second term is bounded by zero, since by construction, \widehat{a}_{ERM} is the minimizer of the empirical risk.

Such uniform laws are also used to derive the so-called **generalization bounds**, where we bound the difference between empirical risk and true risk (note that we use the ML terminology here, and mean the losses rather than the expectation of these over the dataset) for the ERM estimator specifically, so that:

$$L(\theta^*, \widehat{a}_{\text{ERM}}) \leq \widehat{L}_n(\theta^*, \widehat{a}_{\text{ERM}}) + r_{n;\text{GEN}},$$

where it can be seen that $r_{n;\text{GEN}} \leq r_{n;\text{UNIF}}$.