of an $O(f(n))$ universal model for $\mathcal{H}$ relative to loss function **L**. For example, the squared loss function is mixable as long as it is defined relative to a compact set of outcomes $\mathcal{X} = [-R, R]$ rather than the full real line. Unfortunately, the important $0/1$-loss is *not* mixable. Indeed, if $\mathcal{H}$ consists of a fixed number of N experts, and if we allow the prediction algorithm to randomize (i.e. use a biased coin to determine whether to predict 0 or 1), then the optimal universal $0/1$-loss predictor has worst-case regret (in the worst-case over all types of experts and all sequences x^n) of $\text{ORDER}(\sqrt{n})$, whereas the log loss predictor has a much smaller worst-case regret $\ln N$, independently of n and x^n (Cesa-Bianchi, Freund, Helmbold, Haussler, Schapire, and Warmuth 1997). The latter fact can be seen by noting that the worst-case regret of $\bar{P}_{\text{Bayes}}(\cdot \mid \mathcal{M}_{\mathcal{H}})$ with the uniform prior is bounded by $\ln N$. The upshot is that there exist important nonmixable loss functions **L** such as the $0/1$-loss, which have the property that universal prediction with respect to **L** *cannot* be seen as universal prediction with respect to log loss.[13]

Mixability We now give an informal definition of mixability.[14] As we shall see, mixability cannot be obtained for simple loss functions. Thus, let **L** be a loss function that is not simple, so that $Z(\beta) = Z_a(\beta)$, defined as below (17.23), depends on a. We now define a function $C(\beta) := \sup_{a \in \mathcal{A}} Z_a(\beta)$ and use this to define a defective distribution (Chapter 3, page 94)

$$P_a(x) := \frac{1}{C(\beta)} e^{-\beta \mathbf{L}(x,a)}. \tag{17.27}$$

Now set, for fixed β, $\mathcal{P}_{\mathcal{A}}$ as the set of distributions P_a on $\mathcal{X}$ given by (17.27), so that $\mathcal{P}_{\mathcal{A}}$ contains one distribution for each $a \in \mathcal{A}$. Now let $\overline{\mathcal{P}}_{\mathcal{A}}$ be the convex closure of $\mathcal{P}_{\mathcal{A}}$, i.e. the set of all distributions on $\mathcal{X}$ that can be written as mixtures of elements of $\mathcal{P}_{\mathcal{A}}$.

We say that **L** is mixable if we can choose a $\beta > 0$ such that for *any* mixture $P_{\text{mix}} \in \overline{\mathcal{P}}_{\mathcal{A}}$, there exists an $a \in \mathcal{A}$ such that for all $x \in \mathcal{X}$,

$$-\ln P_{\text{mix}}(x) \geq \beta \mathbf{L}(x,a) + \ln C(\beta). \tag{17.28}$$

Since $P_{\text{mix}}(x) = C(\beta)^{-1} \int e^{-\beta \mathbf{L}(x,a)} w(a) da$ for some prior w on $\mathcal{A}$, (17.28) can be rewritten in the following more common form: for every prior w, there should be an a such that for all x,

$$-\frac{1}{\beta} \ln \int e^{-\beta \mathbf{L}(x,a)} w(a) da \geq L(x,a).$$

[13] Nevertheless, some universal predictors that achieve the minimax optimal 0/1-regret to within a constant, are still based on entropification-related ideas. The important difference is that in such algorithms, the β used in (17.23) varies as a function of n. To get good worst-case performance, one needs to take $\beta = O(1/\sqrt{n})$.

[14] Vovk's technical definition is more complicated.

Note that if $\mathbf{L}$ were simple, this would be impossible to achieve since then $C(\beta) = Z(\beta)$ and (17.28) expresses that for all x, $P_a(x) \geq P_{\text{mix}}(x)$, which cannot hold if $P_a(x) \neq P_{\text{mix}}(x)$. Since for nonsimple loss functions, we have $C(\beta) > Z(\beta)$, there sometimes does exist a β for which (17.28) holds after all.

Now define P_h as before, but with $Z(\beta)$ replaced by $C(\beta)$, and for a given set of predictors $\mathcal{H}$, define $\mathcal{M}_{\mathcal{H}} = \{P_h \mid h \in \mathcal{H}\}$. If the mixability condition (17.28) holds, we can modify an $f(n)$-universal code $\bar{P}$ for $\mathcal{M}_{\mathcal{H}}$ into an $O(f(n))$-universal prediction strategy $\bar{h}$ for the loss function $\mathbf{L}$, as long as the predictions $\bar{P}(\cdot \mid x^n)$ can be written as mixtures over the elements of $\mathcal{M}_{\mathcal{H}}$. Thus, unlike in the original entropification approach, we can now also use Bayesian universal codes $\bar{P}_{\text{Bayes}}$. To see this, suppose that $\bar{P}$ is an $f(n)$-universal code for $\mathcal{M}_{\mathcal{H}}$ such that for all n, x^n, $\bar{P}(\cdot \mid x^n) \in \overline{\mathcal{P}}_{\mathcal{A}}$. For each n, x^n, we first set P_{mix} in (17.28) to $\bar{P}(\cdot \mid x^n)$, and then we set $\bar{h}(x^n)$ equal to the a for which (17.28) holds. From (17.28) it is immediate that, for each n, x^n, each $h \in \mathcal{H}$,

$$\beta \sum_{i=1}^{n} \mathbf{L}(x_i, \bar{h}(x^{i-1})) + n \ln C(\beta) \leq - \sum_{i=1}^{n} \ln \bar{P}(x_i \mid x^{i-1}) =$$
$$- \ln \bar{P}(x^n) \leq$$
$$- \ln P_h(x^n) + f(n) \leq \beta \sum_{i=1}^{n} \mathbf{L}(x_i, h(x^{i-1})) + n \ln C(\beta) + f(n), \quad (17.29)$$

from which it follows that

$$\sum_{i=1}^{n} \mathbf{L}(x_i, \bar{h}(x^{i-1})) \leq \sum_{i=1}^{n} \mathbf{L}(x_i, h(x^{i-1})) + \beta^{-1} f(n).$$

As an example, if $\mathcal{X} = \{0,1\}$, $\mathcal{A} = [0,1]$ and the squared loss is used, then the best achievable β is given by $\beta = 1/2$, and an $f(n)$-universal model relative to $\mathcal{P}_{\mathcal{H}}$ with respect to log loss becomes a $2f(n)$-universal model relative to $\mathcal{H}$ with respect to squared loss. This type of correspondence was initiated by Vovk (1990). Further examples of such correspondences, as well as many other relations between log loss and general universal prediction, are discussed by Yamanishi (1998) in the context of his notion of *extended stochastic complexity*.

MDL Is Not Just Prediction The analysis above suggests that MDL should simply be thought of as the special case of the sequential universal prediction framework, instantiated to log loss, and that all references to data compression may be dropped. This reasoning overlooks three facts. First, Theorem 15.1 tells us that in statistical contexts, there is something special about log loss: in contrast to many other loss functions, with probabilistic predictions, it leads to consistent (prequential) estimators $\bar{P}(\cdot \mid X^n)$. Thus, if a