Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Multi-resolution Theory for Approximating Infinite-*p*-Zero-*n*: Transitional Inference, Individualized Predictions, and a World Without Bias-Variance Tradeoff

Xinran Li & Xiao-Li Meng

To cite this article: Xinran Li & Xiao-Li Meng (2021): A Multi-resolution Theory for Approximating Infinite-*p*-Zero-*n*: Transitional Inference, Individualized Predictions, and a World Without Bias-Variance Tradeoff, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1844210

To link to this article: https://doi.org/10.1080/01621459.2020.1844210

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–15: Precision Medicine
https://doi.org/10.1080/01621459.2020.1844210

Taylor & Francis
Taylor & Francis Group

A Multi-resolution Theory for Approximating Infinite-*p*-Zero-*n*: Transitional Inference, Individualized Predictions, and a World Without Bias-Variance Tradeoff

Xinran Li[a] and Xiao-Li Meng[b]

[a]Department of Statistics, University of Illinois, Champaign, IL; [b]Department of Statistics, Harvard University, Cambridge, MA

ABSTRACT

Transitional inference is an empiricism concept, rooted and practiced in clinical medicine since ancient Greece. Knowledge and experiences gained from treating one entity (e.g., a disease or a group of patients) are applied to treat a related but distinctively different one (e.g., a similar disease or a new patient). This notion of "transition to the similar" renders individualized treatments an operational meaning, yet its theoretical foundation defies the familiar inductive inference framework. The uniqueness of entities is the result of potentially an infinite number of attributes (hence $p = \infty$), which entails zero direct training sample size (i.e., $n = 0$) because genuine guinea pigs do not exist. However, the literature on wavelets and on sieve methods for nonparametric estimation suggests a principled *approximation* theory for transitional inference via a multi-resolution (MR) perspective, where we use the resolution level to index the degree of approximation to ultimate individuality. MR inference seeks a primary resolution indexing an *indirect* training sample, which provides enough matched attributes to increase the relevance of the results to the target individuals and yet still accumulate sufficient indirect sample sizes for robust estimation. Theoretically, MR inference relies on an infinite-term ANOVA-type decomposition, providing an alternative way to model sparsity via the decay rate of the resolution bias as a function of the primary resolution level. Unexpectedly, this decomposition reveals a world without variance when the outcome is a deterministic function of potentially infinitely many predictors. In this deterministic world, the optimal resolution prefers over-fitting in the traditional sense when the resolution bias decays sufficiently rapidly. Furthermore, there can be many "descents" in the prediction error curve, when the contributions of predictors are inhomogeneous and the ordering of their importance does not align with the order of their inclusion in prediction. These findings may hint at a deterministic approximation theory for understanding the apparently over-fitting resistant phenomenon of some over-saturated models in machine learning.

ARTICLE HISTORY
Received July 2019
Accepted October 2020

KEYWORDS
Double descent; Machine learning; Multiple descents; Personalized medicine; Sieve methods; Sparsity; Transition to the similar; Wavelets

1. Motivation and Resolution

1.1. Individualized Predictions and Transitional Inference

Predicting an individual's outcome, such as for personalized medicine, is an alluring proposition. Who would not want to know how a treatment would work for *me* before such treatment even begins? But in order to test the effectiveness of a treatment, we will need some guinea pigs. But who can approximate *me*? Someone with my genetic profiles, age, diet, exercise habit, and medical history? But how detailed should the medical history be? What about family medical history? And how extended should my "family" be?

The arrival of Big Data permits us to look into such questions at deeper levels than before, but it does not make our job easier in any fundamental way. Finding a proxy population to approximate an individual is inherently an ill-defined problem from a mathematical perspective, since each of us is defined by an essentially infinite number of attributes, denoted by $p = \infty$. The implied uniqueness of "me" then renders $n = 0$, that is, there will never be any genuine guinea pig for me. Epistemologically, this need of "transition to the similar" has been pondered by philosophers from Galen to Hume (see, e.g., Hankinson 1987, 1995). For example, Galen, a physician and philosopher in the Roman Empire, wrote (see Hankinson 1987):

> In cases in which there is no history, or in which there is none of sufficient similarity, there is not much hope. And the same thing is true in the case of transference of one remedy from one ailment to another similar to it: one has a greater or smaller basis for expectation of success in proportion to the increase or decrease in similarity of the ailment, whether or not history is involved. And the same goes for the transference from one part of the body to another part: expectation of success varies in direct proportion to the similarity.

Galen's framing is essentially a statistical one, with a nice blend of Bayesian (the reliance on history) and frequentist (the emphasis on proportions regardless of history), albeit long before any of these qualifying terms was invented. Perhaps it is a surprise then that, to the best of our knowledge, there is no statistical theory for this kind of *transitional inference* (Hankinson 1995). We surmise that this absence is largely due to the fact that transitional inference goes outside of our traditional inductive framework since it is not about inferring a population from samples of individuals, but rather about predicting individuals'

CONTACT Xiao-Li Meng *meng@stat.harvard.edu* Whipple V.N. Jones Professor of Statistics, Harvard University, Cambridge, MA 02138.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

outcomes by learning from a proxy population. The notion of *similarity*, central to transitional inference, is also a challenging one to metricize in general.

However, the concept of multi-resolution (MR) analysis in engineering and applied mathematics, such as wavelets (see Daubechies 1992; Meyer 1993), turns out to be rather useful for establishing such a theoretical framework. For wavelets, variations in data are first decomposed according to their resolution levels. For image data, the resolution level is the pixel resolution as we ordinarily define, and the concept of multi-resolution can be easily visualized by the common practice of zooming in and out when taking pictures. Zooming too much or too little both would result in losing seeing the big picture, figuratively and literally. Our central task is then to identify a suitable *primary resolution* to separate signals (i.e., lower-resolution wavelet coefficients) from noise (i.e., higher resolution wavelet coefficients); see Donoho et al. (1995) and especially Johnstone (2011) for a survey. The choice of primary resolution thus determines the unit of our inference, that is, the degree of individualization. The search for the primary resolution is generally a quest for an age-old bias-variance tradeoff: estimating more precisely a less relevant individual assessment versus estimating less precisely a more relevant one.

Because the MR framework permits the resolution level to be potentially infinite, it can also be viewed as the predictive counterpart of the estimation method of sieves for dealing with infinite-dimension models. In order to reveal as early as possible what this framework can offer, we follow a reviewer's suggestion to defer a literature review and comparison with the standard large-p-small-n framework to the end of our article.

1.2. A Fundamental Resolution Decomposition

To set up our MR framework, we consider an outcome variable Y sharing the same probability space $(\Omega, \mathcal{F}, P)$ as an information filtration $\{\mathcal{F}_r, r = 0, 1, \ldots, \}$, where $\mathcal{F}_{r-1} \subset \mathcal{F}_r$, and r indexes our resolution level. Here $\mathcal{F}_0$ corresponds to a population of interest (e.g., those who are infected by a certain virus) from where target individuals come, and $\mathcal{F}_\infty = \cup_{r=0}^{\infty} \mathcal{F}_r$ permits us to define (unique) individuality. For example, $\mathcal{F}_r$ is the σ-field generated by covariates $\{X_0, X_1, \ldots, X_r\}$, and hence determining the primary resolution is the same as determining how many covariates should be used for predicting Y for a given information filtration (see Section 2.5 for the issue of ordering the covariates). Let $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ denote mean and variance, respectively. Denote $\mu_r = \mathbb{E}[Y|\mathcal{F}_r]$ and $\sigma_r^2 = \mathbb{V}[Y|\mathcal{F}_r]$ for all r's, including $r = 0$ and $r = \infty$ (and assume these are well defined). Then by repeatedly applying the iterative law $\mathbb{V}[Y|\mathcal{F}_r] = \mathbb{E}[\mathbb{V}(Y|\mathcal{F}_s)|\mathcal{F}_r] + \mathbb{V}[\mathbb{E}(Y|\mathcal{F}_s)|\mathcal{F}_r]$, where $s > r$, we have the usual ANOVA decomposition (Meng 2014),

$$\sigma_r^2 = \mathbb{E}[\sigma_\infty^2|\mathcal{F}_r] + \sum_{i=r}^{\infty} \mathbb{E}[(\mu_{i+1} - \mu_i)^2|\mathcal{F}_r], \quad \text{for any } r \geq 0. \tag{1}$$

Decomposition (1) reminds us that the usual dichotomy between *variance*, as a measure of random variations, and *bias*, as a measure of systematic differences, is an artificial one, except possibly at the infinite resolution level. That is, the variance at any particular resolution level is merely the accumulation of all the (squared consecutive) systematic differences, that is, biases, at higher resolution levels, plus σ_∞^2, the *intrinsic variance*. Conceptually σ_∞^2 cannot be ascertained from any empirical data, because we can never be sure whether the residual variance from whatever model we fit is due to σ_∞^2 or to a limitation of our always finite amount of data. It therefore seems inconsequential to set $\sigma_\infty^2 = 0$ since we can never prove it false. This proposition should be particularly acceptable to those who believe that the world is ultimately deterministic once all its operating mechanisms are measured and understood (see, e.g., Peat 2002).

However, as we shall reveal in this article, whether or not to set σ_∞^2 to zero has profound implications on the bias-variance tradeoff phenomenon. To the best of our knowledge, the statistical literature has not investigated this phenomenon for chaotic dynamic systems (e.g., Devaney 2018), since when $\sigma_\infty^2 = 0$, the setup here enters the realm of deterministic but potentially chaotic systems. The corresponding findings therefore may be counter-intuitive (initially) to statisticians, but they might provide a bridge to the growing literature in machine learning that casts doubts on the applicability of bias-variance tradeoff, especially the literature surrounding the phenomenon of "double descent" (e.g., Belkin et al. 2019; Belkin, Hsu, and Xu 2019; Hastie et al. 2019; Nakkiran et al. 2019), which we shall explain and extend to "multiple descents" later in this article.

Regardless of how we treat σ_∞^2, declaring that a resolution level R is our primary resolution implies that all the information conveyed by variations at resolution levels higher than R can be effectively ignored when predicting Y. The MR formulation therefore permits us to quantify the degree of individualization, and to be explicit about the two contributing factors of our overall prediction error: (I) the resolution bias due to choosing a finite R; and (II) the estimation error at the given resolution R. The MR framework therefore integrates the model selection step (I) with the model estimation step (II), and hence it does not need to treat the issue of selection post-hoc (e.g., Berk et al. 2013; Lee et al. 2016; Tibshirani et al. 2016). Furthermore, since the filtration $\{\mathcal{F}_r, r = 0, 1, \ldots\}$ forms a cumulative "information basis," the choice of optimal R_n for a given dataset with size n is in the same spirit as finding a *sparse representation* in wavelets, for which there is a large literature (see Poggio and Girosi 1998; Donoho and Elad 2003), though here perhaps it is more appropriate to term it as *parsimonious representation*.

1.3. Time-Honored Intuitions, and Timely New Insights?

Our findings confirm some time-honored intuitions and build new ones. Specifically, in Section 2 we first decompose the total prediction error into three components: the ultimate risk, the resolution bias and the estimation error. We then provide an overview and highlight on how the optimal resolution depends on the decay rates of the resolution bias and the corresponding estimation error under a particular ordering of covariates, respectively, in the stochastic world (i.e., $\sigma_\infty^2 > 0$) and deterministic world (i.e., $\sigma_\infty^2 = 0$). Section 2 concludes with some theoretical insights on the issue of ordering the covariates.

Sections 3 and 4 then establish our general results with an infinite number of continuous and categorical predictors, and illustrates them with linear regression and tree regression,

respectively. In particular, in Sections 3.3 and 4.3, we report some intriguing findings when $\sigma_\infty^2 = 0$, respectively, for these two regression models. In this world without variance, the optimal resolution may rightly prefer the direction of over-fitting in the traditional sense; indeed the optimal resolution level can even approach infinity. But this preference does not violate the time-honored bias-variance tradeoff principle because, without variance, the optimal tradeoff may have to put all its eggs in the basket of bias.

We also find that the predictive error curve can exhibit double descents or even arbitrarily many descents without ever entering the over-parameterized realm. These findings might provide a new angle to investigate very flexible and saturated models, such as deep learning networks, to understand their seemingly magical ability to resist over-fitting. That is, with a huge amount of data, it is conceivable that an exceedingly rich and flexible deterministic model class can learn to practically exhaust all patterns detectable with reasonable chances in reality (which can be far fewer than in theory). In such cases, we would not need $\sigma_\infty^2 > 0$ to absorb the imperfection of the model, effectively rendering it a deterministic system, a system that prefers "over-fitting" in the traditional sense. This is also explored empirically in Section 3.5, where we summarize a simulation study with linear models that investigates the practicality of the MR approach that employs cross validation and other methods for selecting the primary resolution in practice. The details of the study, as well as all the technical proofs in our article, are deferred to the Appendices in the online supplementary materials. Section 5 completes our exploration by making connections to relevant literature and discussing further work.

2. A Multi-Resolution Framework

2.1. Prediction With Potentially Infinitely Many Predictors

To start, let $\odot$ be a member of a target population, which can be as small as a single individual, and $Y(\odot)$ be a univariate response from $\odot$, which can be discrete (e.g., a treatment success indicator) or continuous (e.g., the change of the cholesterol level due to a treatment). Typically the investigators have some prior knowledge about which set of the individual's attributes play more critical roles in determining Y. But, philosophically and practically, no one can be certain about what constitutes the complete set of relevant predictors. Statistically, we can model such a situation by requiring the distribution of $Y(\odot)$ to depend on potentially infinitely many attributes of $\odot$, denoted by $\vec{\boldsymbol{X}}_\infty(\odot) = \{X_0(\odot), X_1(\odot), X_2(\odot), \ldots\}$. In reality we can never observe infinitely many covariates, but the arrival of the digital age has created many situations where we have far more predictors than the sample size. Our job is to seek a small subset of the predictors of the outcome with accuracy that makes our prediction useful.

We use $f_\odot$ to denote the joint probability mass/density function of the response and covariates for the target individual $\odot$. To learn about $f_\odot$, especially the dependence of $Y(\odot)$ on $\vec{\boldsymbol{X}}_\infty(\odot)$, we need to collect a training set $\mathcal{T}_n = \{(y_i, \vec{\boldsymbol{x}}_{i\infty}) : i = 1, 2, \ldots, n\}$, which are (assumed to be) independent and identically distributed (iid) samples from a training (proxy) population. Clearly the phrase "training" implies that we need some assumptions to link $\mathcal{T}_n$ to the target population. The ideal assumption of course is that $f_\odot$ equals the joint probability mass/density function f of $(Y, \vec{\boldsymbol{X}}_\infty)$ for the training population. Whereas all attempts should be made to mimic the target population when we form the training population, it is wise to permit our framework sufficient flexibility to admit cases where f may differ from $f_\odot$ but in an approximately known way. Mathematically, this flexibility can be handled by introducing a weight function

$$w_\odot(Y, \vec{\boldsymbol{X}}_\infty) = \frac{f_\odot(Y, \vec{\boldsymbol{X}}_\infty)}{f(Y, \vec{\boldsymbol{X}}_\infty)} = \frac{f_\odot(Y|\vec{\boldsymbol{X}}_\infty)}{f(Y|\vec{\boldsymbol{X}}_\infty)} \frac{f_\odot(\vec{\boldsymbol{X}}_\infty)}{f(\vec{\boldsymbol{X}}_\infty)}. \quad (2)$$

Normally it is almost inevitable to assume $f_\odot(Y|\vec{\boldsymbol{X}}_\infty) \approx f(Y|\vec{\boldsymbol{X}}_\infty)$, that is, the (stochastic) relationships between the outcome and the predictors for the target population and the training population must be approximately the same, because otherwise our selection of the training sample is a very poor one. Consequently, (2) implies $w_\odot(Y, \vec{\boldsymbol{X}}_\infty) \approx f_\odot(\vec{\boldsymbol{X}}_\infty)/f(\vec{\boldsymbol{X}}_\infty)$, which is easier to estimate since it merely involves adjusting the marginal distribution of the $\vec{\boldsymbol{X}}_\infty$, known as a "covariate shift" in the literature (see, e.g., Bickel, Brückner, and Scheffer 2007; Sugiyama and Kawanabe 2012). However, when $\odot$ is indeed a single individual or beyond the support of the training population, the weight $w_\odot(Y, \vec{\boldsymbol{X}}_\infty)$ is not defined without lowering the resolution level for evaluation; see Meng (2021). We leave the choice of weights for a future study, as our focus in this article is on the choice of optimal resolutions with given weight functions.

To avoid confusion, we use $\mathbb{E}_\odot$ and $\mathbb{E}$ to denote the expectations over the target and the training populations, respectively. To evaluate the prediction performance of a prediction function $\hat{y}(\vec{\boldsymbol{X}}_\infty)$, we can adopt a loss function $\mathcal{L}(y, \hat{y})$, which is problem-dependent. Clearly, we can minimize the expected loss $\mathbb{E}_\odot[\mathcal{L}(Y, \hat{y}(\vec{\boldsymbol{X}}_\infty))]$ via minimizing $\mathbb{E}[\mathcal{L}_\odot(Y, \hat{y}(\vec{\boldsymbol{X}}_\infty))]$, where $\mathcal{L}_\odot(Y, \hat{y}(\vec{\boldsymbol{X}}_\infty)) \equiv \mathcal{L}(Y, \hat{y}(\vec{\boldsymbol{X}}_\infty)) w_\odot(Y, \vec{\boldsymbol{X}}_\infty)$; the subscript $\odot$ indicates its dependence on the utility of prediction and the target population of interest. With this setup, we proceed as follows. At each resolution r, we restrict our prediction to a family of functions $\{g(\vec{\boldsymbol{x}}_r; \boldsymbol{\theta}_r)\}$, where $\vec{\boldsymbol{x}}_r = (x_0, \ldots, x_r)$. For notational simplicity, we suppress the explicit dependence of $g(\cdot)$ on r, but rather use the inputs $\vec{\boldsymbol{x}}_r$ and $\boldsymbol{\theta}_r$ to emphasize such dependence implicitly. Note that $\boldsymbol{\theta}_r$ denotes a generic parameter whose dimension can vary with r. For example, $\dim(\boldsymbol{\theta}_r) = \binom{r+2}{2}$ if $g(\vec{\boldsymbol{x}}_r; \boldsymbol{\theta}_r)$ is a linear function of covariates up to resolution r and of all their quadratic terms and pairwise interactions. Generally, we will choose $g(\cdot)$ such that the family of prediction functions becomes richer as resolution increases. That is, for any $r < r'$, any prediction function $g(\vec{\boldsymbol{x}}_r; \boldsymbol{\theta}_r)$ at resolution r, viewed as a function of $\vec{\boldsymbol{x}}_{r'}$, belongs to the family of prediction functions at resolution r'. At each resolution r, the optimal prediction is then $g(\vec{\boldsymbol{x}}_r; \boldsymbol{\theta}_r^*)$, with $\boldsymbol{\theta}_r^* \equiv \arg\min_{\boldsymbol{\theta}_r} \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_r; \boldsymbol{\theta}_r))]$. A usual estimator for $\boldsymbol{\theta}_r^*$ is obtained by minimizing the empirical risk: $\hat{\boldsymbol{\theta}}_r \equiv \arg\min_{\boldsymbol{\theta}_r} \sum_{i=1}^n \mathcal{L}_\odot(y_i, g(\vec{\boldsymbol{x}}_{ir}; \boldsymbol{\theta}_r))$. Hence, once we choose the primary resolution R, we predict Y by $g(\vec{\boldsymbol{x}}_R; \hat{\boldsymbol{\theta}}_R)$ for an individual with covariate $\boldsymbol{x}_\infty$, and estimate the prediction error $\mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))]$ by the empirical risk $n^{-1} \sum_{i=1}^n \mathcal{L}_\odot(y_i, g(\vec{\boldsymbol{x}}_{iR}; \hat{\boldsymbol{\theta}}_R))$, or by cross-validation.

2.2. A Trio Decomposition of the Prediction Error

To better understand the prediction error at a resolution R, we decompose $\mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))]$ into three parts: the ultimate risk, the resolution bias at resolution R, and the estimation error at resolution R. The *ultimate risk* is $\tau^2 \equiv \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_\infty; \boldsymbol{\theta}^*_\infty))]$, which depends on the families of functions used for prediction. Specifically, it has two sources, one due to model misspecification and the other due to the *intrinsic variation* at the infinite resolution, i.e., $f(Y|\vec{\boldsymbol{X}}_\infty)$. That is, the intrinsic variance $\sigma^2_\infty = \mathbb{V}(Y|\vec{\boldsymbol{X}}_\infty)$ can be positive (or even infinity) in a stochastic world. The *resolution bias* at resolution R then is

$$A(R) = \sum_{r=R+1}^{\infty} \left\{ \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_{r-1}; \boldsymbol{\theta}^*_{r-1}))] - \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_r; \boldsymbol{\theta}^*_r))] \right\}.$$

When the family of prediction functions becomes richer as resolution increases, $A(R)$ is nonincreasing in R and approaches zero as $R \to \infty$, that is, $\lim_{R\to\infty} A(R) = 0$. Finally, the *estimation error* at resolution R,

$$\varepsilon(R, \mathcal{T}_n) = \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))] - \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \boldsymbol{\theta}^*_R))],$$

is nonnegative by the optimality of $\boldsymbol{\theta}^*_R$. From the above, the prediction error at resolution R using training set $\mathcal{T}_n$ can be decomposed as

$$\mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))] = \tau^2 + A(R) + \varepsilon(R, \mathcal{T}_n). \tag{3}$$

As we shall show shortly, theoretically, we can gain good insight by considering the averaged version of this decomposition, that is,

$$\mathbb{E}_n\left[\mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))]\right] = \tau^2 + A(R) + \varepsilon(R, n), \tag{4}$$

where, with slight abuse of notation, $\varepsilon(R, n) = \mathbb{E}_n[\varepsilon(R, \mathcal{T}_n)]$, and $\mathbb{E}_n$ denotes the expectation over all training sets of size n.

It is worthy noting that Equation (3) is an extension of the ANOVA decomposition (1) in expectation, with Equation (1) being a special case with $\mathcal{L}_\odot(y, \hat{y}) = (y - \hat{y})^2$ and $g(\boldsymbol{X}_{\vec{r}}; \boldsymbol{\theta}^*_r) = \mathbb{E}(Y|\vec{\boldsymbol{X}}_r)$ for $r \geq 1$, that is, the prediction functions are correctly specified. Under this special case, the ultimate risk τ^2 reduces to $\mathbb{E}(\sigma^2_\infty)$. We remark that in general $\tau^2 \geq \mathbb{E}(\sigma^2_\infty)$, with equality holds when we correctly specified the prediction functions. Because $\sigma^2_\infty \geq 0$, a zero τ^2 then must imply $\sigma^2_\infty = 0$ (almost surely), that is, a deterministic world without variance. Here, as in Equation (1), $\sigma^2_r = \mathbb{V}[Y|\boldsymbol{X}_{\vec{r}}]$ and $\mu_r = \mathbb{E}[Y|\boldsymbol{X}_{\vec{r}}] = g(\boldsymbol{X}_{\vec{r}}; \boldsymbol{\theta}^*_r)$, which is estimated by $\hat{\mu}_r = g(\boldsymbol{X}_{\vec{r}}; \hat{\boldsymbol{\theta}}_r)$. The resolution bias at resolution R reduces to $\sum_{r=R}^{\infty}[\mathbb{E}(\sigma^2_r) - \mathbb{E}(\sigma^2_{r+1})] = \sum_{r=R}^{\infty}\mathbb{E}(\mu_{r+1} - \mu_r)^2$, and the estimation error to $\mathbb{E}(\hat{\mu}_R - \mu_R)^2$. Consequently, Equation (3) reduces to

$$\mathbb{E}(\sigma^2_R) + \mathbb{E}(\hat{\mu}_R - \mu_R)^2 = \mathbb{E}(\sigma^2_\infty) + \sum_{r=R}^{\infty}\mathbb{E}(\mu_{r+1} - \mu_r)^2 + \mathbb{E}(\hat{\mu}_R - \mu_R)^2, \tag{5}$$

which is equivalent to Equation (1) by further averaging over $\mathcal{F}_r$ (i.e., the conditioning in Equation (1)).

Because in Equations (3) and (4) the ultimate risk is not affected by the resolution (under the assumption that the function form is the same at the infinite resolution), for any training set $\mathcal{T}_n$, the optimal primary resolution that minimizes the prediction error in Equation (3) is

$$\begin{aligned} R_{\mathcal{T}_n,\text{opt}} &= \arg\min_R \mathbb{E}[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))] \\ &= \arg\min_R [A(R) + \varepsilon(R, \mathcal{T}_n)]. \end{aligned}$$

Similarly, the optimal primary resolution that minimizes the prediction error in (4) is

$$\begin{aligned} R_{n,\text{opt}} &= \arg\min_R \mathbb{E}_n\left[\mathbb{E}\left[\mathcal{L}_\odot(Y, g(\vec{\boldsymbol{X}}_R; \hat{\boldsymbol{\theta}}_R))\right]\right] \\ &= \arg\min_R [A(R) + \varepsilon(R, n)]. \end{aligned}$$

Studying $R_{\mathcal{T}_n,\text{opt}}$ or $R_{n,\text{opt}}$ for a particular training set $\mathcal{T}_n$ or a particular size n is generally difficult. We therefore resort to the usual asymptotic strategy. That is, as n goes to infinity, we seek a sequence $\{R_n\}_{n=1}^{\infty}$ such that $A(R_n) + \varepsilon(R_n, \mathcal{T}_n)$ or $A(R_n) + \varepsilon(R_n, n)$ converges to zero (in probability) as fast as possible. We will adopt the notation $a_n \asymp b_n$ if two sequences $\{a_n\}$ and $\{b_n\}$ satisfy $a_n = O(b_n)$ and $b_n = O(a_n)$, and similarly, $\tilde{a}_n \overset{\mathbb{P}}{\asymp} \tilde{b}_n$ if random sequences $\{\tilde{a}_n\}$ and $\{\tilde{b}_n\}$ satisfy $\tilde{a}_n = O_{\mathbb{P}}(\tilde{b}_n)$ and $\tilde{b}_n = O_{\mathbb{P}}(\tilde{a}_n)$, using the usual definition of $O_{\mathbb{P}}$. We also use the notation $a_n \gtrsim b_n$ for $b_n = O(a_n)$.

2.3. Optimal Resolution and Learning Rate in the Stochastic World

Intuitively, there must be a tradeoff in determining the optimal R_n. To control the resolution bias $A(R_n)$, we desire large R_n because of the monotonically decreasing nature of $A(R)$. For $A(R_n)$, we will consider four scenarios, representing four different levels of sparsity. However, to control the estimation error, we want small R_n to reduce the number of model parameters to be estimated. When the intrinsic variance $\sigma^2_\infty > 0$, under some regularity conditions (e.g., our estimation methods are efficient), we have the usual $\varepsilon(R_n, n) \asymp \dim(\boldsymbol{\theta}_{R_n})/n$ asymptotics. Hence we need $\dim(\boldsymbol{\theta}_{R_n}) = o(n)$ to ensure $\varepsilon(R_n, n)$ converges to zero as $n \to \infty$.

Table 1 provides a high-level preview of the general asymptotic results under the above setting, with four (common) choices of the decay rate for $A(r)$. What do these asymptotic results tell us? First, the hard-thresholding cases correspond to the classical parametric setting, with a fixed number (r_0) of predictors. Hence, as long as our resolution level R_n exceeds r_0 (arbitrarily often), we will reach the classical n^{-1} error rate, excluding the ultimate risk (which includes the intrinsic variation).

Second, the rate-optimal resolution R_n—and hence the minimal prediction error—depends critically on both the decay rate $A(r)$ and estimation error $\varepsilon(r, n)$. When $\varepsilon(r, n)$ grows polynomially with the resolution level (e.g., the continuous covariates cases), we can still practically achieve the n^{-1} rate when $A(r)$ decays exponentially, because the price we pay is merely a $\log^\alpha(n)$ term. However, if $\varepsilon(r, n)$ grows exponentially (e.g., with discrete covariates), then although R_n is still practically of

Table 1. Rate-optimal R_n and minimal error $L_n \equiv A(R_n) + \varepsilon(R_n, n)$ in a stochastic world. All c_n's are of $O(1)$ but satisfy different constraints as specified in Theorem 2 (Section 3.2) and Theorem 4 (Section 4.2).

$\varepsilon(r,n)$ \ $A(r)$	Hard Thresholding $1_{\{r<r_0\}}$	Exponential Decay $e^{-\xi r}$ $(\xi>0)$	Polynomial Decay $r^{-\xi}$ $(\xi>0)$	Logarithmic Decay $\log^{-\xi}(r)$ $(\xi>0)$
Polynomial in r	$R_n \asymp c_n \geq r_0$	$c_n \log n$	$c_n n^{1/(\xi+\alpha)}$	$\frac{c_n n^{1/\alpha}}{\log^{\xi/\alpha}(n)}$
r^α/n $(\alpha > 0)$	$L_n \asymp 1/n$	$\log^\alpha(n)/n$	$n^{-\xi/(\xi+\alpha)}$	$[\log(n)]^{-\xi}$
Exponential in r	$R_n \asymp c_n \geq r_0$	$\frac{\log n+\log c_n}{\xi+\log\alpha}$	$c_n \log(n)$	$c_n \log(n)$
α^r/n $(\alpha > 1)$	$L_n \asymp 1/n$	$n^{-\xi/(\xi+\log\alpha)}$	$[\log(n)]^{-\xi}$	$[\log\log(n)]^{-\xi}$

Table 2. Rate-optimal R_n and minimal error $L_n \equiv \mathrm{PE}_n$ in a deterministic world. All c_n's are of $O(1)$ but satisfy different constraints as specified in Theorem 3 (Section 3.3), Theorems 5 and 6 (Section 4.3).

Model \ $A(r)$	Hard Thresholding $1_{\{r<r_0\}}$	Exponential Decay $e^{-\xi r}$	Polynomial Decay $r^{-\xi}$	Logarithmic Decay $\log^{-\xi}(r)$
Linear regression	$n-3 \geq R_n \geq r_0$ $L_n = 0$	$R_n = n - c_n$ $L_n \asymp ne^{-\xi n}$	$c_n n$ $n^{-\xi}$	$c_n n^k, k \in (0,1]$ $[\log(n)]^{-\xi}$
Regression tree with predictors	$R_n \geq r_0$	$\begin{cases} \gtrsim c_n \log(n), & \xi > \log(M) \\ = c_n \log(n), & \xi = \log(M) \\ = c_n \log(n), & \xi < \log(M) \end{cases}$	$c_n \log(n)$	$c_n \log(n)$
X_j's are iid Uniform{1,...,M}	$L_n \asymp (1 - M^{-r_0})^n$	$\begin{cases} \lesssim n^{-1}, & \xi > \log(M) \\ \lesssim n^{-1}\log(n), & \xi = \log(M) \\ n^{-\xi/\log(M)}, & \xi < \log(M) \end{cases}$	$[\log(n)]^{-\xi}$	$[\log\log(n)]^{-\xi}$

Note: like in Table 1, $\xi > 0$. In some cases, the forms of rate-optimal R_n are only sufficient but not necessary for achieving the optimal rate.

$\log(n)$ type, the parametric error rate n^{-1} is no longer achievable even if $A(r)$ decays exponentially. Instead, we can achieve only a nonparametric like error rate in the form of $n^{-\xi/(\xi+\log\alpha)}$, which reduces to n^{-1} only if the decay rate parameter ξ for $A(r)$ goes to infinity.

Third, when $A(r)$ decays polynomially, R_n takes on different rate forms depending on how the estimation error varies with the resolution level r, that is, (A) polynomial in n for polynomial estimation error versus (B) $\log(n)$ for exponential estimation error. More importantly, the difference in the corresponding minimal prediction errors tells us that in case (A), the individualized prediction and learning rate is slow but still practical. However, case (B) belongs to the situation where the individualized learning rate is too slow to be useful. The same is true once the decay rate is logarithmic because then the prediction error rate is no better than that of case (B); see the last column of Table 1. Therefore, among the eight scenarios in Table 1, only the first five (counting first top to bottom then left to right) of these permit practical individualized learning.

Here we give a side note on the asymptotic expression in Table 1. First, a more rigorous expression for the polynomial estimation error is $\varepsilon(r,n) \asymp \max\{r^\alpha, 1\}/n$. We simply use r^α/n not only for descriptive convenience, but also since $r \geq 1$ is required for achieving rate optimal prediction when $A(0) > 0$. Second, the decay rates for resolution biases, for example, $r^{-\xi}$ and $\log^{-\xi}(r)$, may be well-defined only for r larger than a certain value. Whenever such a quantity is not prescribed, we can view it as a finite positive constant. Again, this complication has little relevance for our asymptotic theory for the rate-optimal resolution, which must go to infinity as $n \to \infty$ when $A(r) > 0$ for any finite r.

2.4. Optimal Resolution and Learning Rate in the Deterministic World

The case with $\sigma_\infty^2 = 0$ or more precisely zero ultimate risk, however, behaves rather differently, and will be studied in Sections 3.3 and 4.3 for two popular models. We restrict to specific models because we have not been able to obtain general results parallel to those in Table 1. But even with these specific results, we already see asymptotic behaviors, as revealed in Table 2, that are quite different from those in Table 1. The trivial ones are for hard thresholding, where for the linear model, as long as sample sizes are large enough to solve the linear system, we will have zero error. Similarly, for the regression tree case, where the only possible error is when no exact match of the target case exists with respect to the r_0 important predictors in the training sample of size n. The probability of this occurring is exactly $(1 - M^{-r_0})^n$ under our model assumption that all predictors X_j's are independently and identically distributed as uniform on $\{1, 2, \ldots, M\}$, a mathematically convenient assumption that permits us to obtain analytical results.

The more interesting cases are when $A(r)$ decays exponentially, which permits the optimal R_n to be infinity; for example, for the regression tree model, when the resolution bias decays exponentially with $\xi > \log(M)$, choosing $R_n = \infty$ can lead to prediction error no worse than order n^{-1}. That is, we are not worried about over-fitting because the benefit from exact matching outweighs the imprecision in solving, say, the linear system. This phenomenon does not occur when we restrict ourselves to statistical models with a finite number of predictors, which would force us to adopt an error term to capture the unexplained residual variations in the outcome variable. With an infinite number of predictors, there is at least a theoretical

possibility that collectively they can explain all the variations in the outcome variable. There is no free lunch, however, as this full-explanatory power requires that the predictive model is specified correctly. Nevertheless, the discovery of this phenomena by permitting models with an infinite number of predictors should remind us of the value of exploring this line of thinking, as it might lead to alternative insights into why certain highly saturated black box models (e.g., deep learning networks) can have a seemingly over-fitting resistant nature. We shall explore this line of thinking in Section 3.4, where we show how easily we can go beyond the intriguing "double-descent" phenomenon (e.g., Belkin et al. 2019; Hastie et al. 2019; Nakkiran et al. 2019) in the deterministic world with infinitely many predictors, without even having to actually enter the realm of over-fitting.

2.5. The Impact of Ordering

So far we have assumed that the order of the covariates is predetermined. In reality, the investigators may have some "low resolution" knowledge of the importance of *groups* of the covariates (e.g., age and gender are typically among the predictors to be included in predicting health outcome). However, they often do not possess the refined knowledge to specify the exact order of the covariates in terms of their predictive power (if they did, the problem would be much easier). Mathematically, when the resolution levels change, we can change all the covariates included in the model. But to utilize our partial knowledge, however imprecise, we wish to investigate the dependence of prediction error on the order of the covariates, and in particular the degree of mis-ordering that can fundamentally alter the prediction error rate. That is, how much misspecification of the order can we tolerate before it really matters? Assume that the family of prediction functions becomes richer as resolution increases, and they are invariant to the ordering of the covariates, that is, for any r and any permutation π of $\{0, 1, 2, \ldots, r\}$, the families of functions $\{g(\vec{\boldsymbol{x}}_r; \boldsymbol{\theta}_r)\}$ and $\{g(\vec{\boldsymbol{x}}_{\pi(r)}; \boldsymbol{\theta}_r)\}$ are the same, where $\vec{\boldsymbol{x}}_{\pi(r)} \equiv (x_{\pi(0)}, x_{\pi(1)}, \ldots, x_{\pi(r)})$. Consequently, the ultimate risk $\tau^2 = \mathbb{E}[\mathcal{L}_{\odot}(Y, g(\vec{\boldsymbol{X}}_{\infty}; \boldsymbol{\theta}^*_{\infty}))]$ is invariant to the ordering of covariates. This is most clearly seen under squared loss and correctly specified conditional mean function, where $\tau^2 = \mathbb{E}(\sigma^2_{\infty})$, as discussed prior to arriving at Equation (5). Below we will focus on the resolution bias and estimation error.

We begin by considering a specific ordering of the covariates, $\{X_0, X_1, X_2, \ldots\}$, identified with its resolution bias $A(\cdot)$, estimation error $\varepsilon(\cdot, n)$, and rate-optimal resolution R_n. Let A', ε' and R'_n be their counterparts under a new ordering $\{X'_0, X'_1, \ldots\}$. Generally, the estimation errors $\varepsilon(r_n, n)$ and $\varepsilon'(r_n, n)$ under both orderings (i.e., $r_n = R_n$ or R'_n) are of the same order after some proper scaling of "unit noise," because they involve estimation for the same number of parameters. In the following discussion, we assume $\varepsilon(r_n, n)/[A(r_n) + \tau^2] \asymp \varepsilon'(r_n, n)/[A'(r_n) + \tau^2]$, which reduces to $\varepsilon(r_n, n) \asymp \varepsilon'(r_n, n)$ when $\tau^2 > 0$. As shown later, this assumption is motivated by the linear regression and tree regression models. Then, a sufficient condition for the new order to achieve the optimal rate under the original ordering is that $A'(R_n) = O(A(R_n))$. This condition should be intuitive because all it requires is that the new ordering does not delay the inclusion of covariates which are considered important by the original ordering.

Suppose now that every covariate matters, in the sense that the resolution bias at any finite resolution is positive, regardless of the ordering of covariates. From Section 2.3, for any ordering of covariates, its optimal primary resolution must go to infinity as $n \to \infty$; that is, we exclude the hard-thresholding case (which is too ideal for the kind of individualized learning we address in this article). To measure the difference between $A(\cdot)$ and $A'(\cdot)$, we introduce $M_r(A, A')$ to denote the minimum non-negative integer such that the first $r - M_r(A, A') + 1$ covariates in ordering $A(\cdot)$ is ranked among the first $r + 1$ positions in ordering $A'(\cdot)$, that is, variables $\{X_0, \ldots, X_{r-M_r(A,A')}\}$ are included in $\{X'_0, \ldots, X'_r\}$. Note that $M_r(A, A') \leq r$ because we can assume $X'_0 = X_0$ since they both denote the constant term. It is asymmetric in A and A', and the farther $M_r(A, A')$ is away from zero, the more different A and A' will be. That is, $M_r(A, A')$ is the number of mistakes we make in choosing the first $r + 1$ covariates with respect to the original ordering $A(\cdot)$. The following theorem tells us how many mistakes are acceptable, asymptotically.

Theorem 1. Assume that (a) the family of prediction functions becomes richer as resolution increases, and is invariant to the permutation of the covariates at each resolution; (b) the estimation error rate is invariant to the ordering: $\varepsilon(r_n, n)/[A(r_n) + \tau^2] \asymp \varepsilon'(r_n, n)/[A'(r_n) + \tau^2]$. Then a sufficient condition for $A'(R_n) = O(A(R_n))$ under each decay scenario (as underlined and where $\xi > 0$) is given below.

(i) Exponential Decay: $A(r) \asymp e^{-\xi r}$:
$\limsup_{r\to\infty} M_r(A, A') \leq$ Constant.

(ii) Polynomial Decay: $A(r) \asymp r^{-\xi}$:
$\limsup_{r\to\infty} M_r(A, A')/r < 1$.

(iii) Logarithmic Decay: $A(r) \asymp \log^{-\xi}(r)$:
$M_r(A, A') = r - r^{1/a_r}$ with $a_r = O(1)$.

The qualitative message of Theorem 1 is rather intuitive. The fewer of the important predictors that exist, the surer we need to include them in our prediction model. Although we still need to obtain the necessary conditions, the quantitative messages here can be taken as theoretical guidelines. With exponential decay, the number of forgivable mistakes is very limited, and it cannot be permitted to grow with the resolution level. Under polynomial decay, which still includes the practically learn-able case when the estimation error is also polynomial in resolution r, we can permit the number of mistakes to increase linearly with r (but of course less rapidly than the growth rate of r).

This learn-able case is perhaps the practically most important scenario, since polynomial decay and polynomial estimation error are the kind of cases that we hope to encounter in practice. Exponential decay is likely too much for which to hope in many practical situations, and logarithmic decay is hopeless in terms of individualized learning, as seen in Tables 1 and 2. The result in Theorem 1 with logarithmic decay indicates that we can be almost entirely wrong in our ordering but still maintain the optimal rate. This seemingly too-good-to-be-true result indeed is a negative one, because it is made possible by the fact that there is really not much information in the predictors, so whatever orders one uses will not improve the situation.

3. Prediction With Infinitely Many Continuous Predictors

3.1. Normal Linear Models With Infinitely Many Continuous Covariates

Consider the simple linear regression model with infinitely many covariates, which we assume to hold for both the target and training populations:

$$Y = \boldsymbol{\beta}_\infty^\top \vec{\boldsymbol{X}}_\infty + \eta \equiv \sum_{r=0}^{\infty} \beta_r X_r + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_\eta^2), \quad \eta \perp\!\!\!\perp \vec{\boldsymbol{X}}_\infty,$$
$$\text{where } X_0 = 1, \ \{X_1, X_2, \ldots\} \text{ are jointly normally distributed.} \tag{6}$$

Clearly, for $\mathbb{V}(Y) < \infty$, always the case in practice, there will be restrictions on β_r's. Here we choose the loss function to be $\mathcal{L}_\odot(y, \hat{y}) = \mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$, and the prediction function at resolution r to be linear in the first $r + 1$ covariates, that is, $g(\vec{\boldsymbol{x}}_r, \boldsymbol{\theta}_r) = \boldsymbol{\theta}_r^\top \vec{\boldsymbol{x}}_r$.

Under this setting, the optimal prediction function is $g(\vec{\boldsymbol{x}}_r, \boldsymbol{\theta}_r^*) = \mathbb{E}(Y|\vec{\boldsymbol{X}}_r = \vec{\boldsymbol{x}}_r)$. The estimator $\hat{\boldsymbol{\theta}}_r$ for the true $\boldsymbol{\theta}_r^*$ using empirical risk minimization is the least-squares estimator based on the first $r + 1$ covariates in the training set $\mathcal{T}_n$. Thus, our prediction for a unit with covariates $\boldsymbol{x}_\infty$ using primary resolution r is $g(\vec{\boldsymbol{x}}_r, \hat{\boldsymbol{\theta}}_r) = \hat{\boldsymbol{\theta}}_r^\top \vec{\boldsymbol{x}}_r$. Now we investigate the prediction error at a specific resolution r and in particular its decomposition as in Section 2.2. First, because we consider square loss and specify the prediction function perfectly, the ultimate risk $\tau^2 = \sigma_\infty^2 \equiv \mathbb{V}(Y|\vec{\boldsymbol{X}}_\infty) = \sigma_\eta^2$. Note here because of the additivity of the error term η in (6), σ_∞^2 is a constant. In general, τ^2 and σ_∞^2 are different. In the following we will use $\tau^2 = 0$ to indicate the world without variance.

Second, define $\delta_k^2 \equiv \mathbb{V}(Y|\vec{\boldsymbol{X}}_{k-1}) - \mathbb{V}(Y|\vec{\boldsymbol{X}}_k)$ as the variance of the response explained by the kth covariates in excess to that by the previous ones. Then $A(r) = \sum_{k=r+1}^{\infty} \delta_k^2$. Third, the estimation error is $\varepsilon(r, \mathcal{T}_n) = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)^\top \mathbb{E}(\vec{\boldsymbol{X}}_r \vec{\boldsymbol{X}}_r^\top)(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)$, and its expectation over all training sets of size n is (see the Appendices)

$$\varepsilon(r, n) = \mathbb{E}_n\left[\varepsilon(r, \mathcal{T}_n)\right] = \frac{A(r) + \tau^2}{n - r - 2}\left(\frac{n-2}{n} + r\right). \tag{7}$$

Consequently, the average prediction error in Equation (4) at resolution r is

$$\begin{aligned}
&\mathbb{E}_n\left\{\mathbb{E}[Y - g(\vec{\boldsymbol{X}}_r, \hat{\boldsymbol{\theta}}_r)]^2\right\} \\
&= \tau^2 + \sum_{k=r+1}^{\infty} \delta_k^2 + \mathbb{E}_n\left[(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)^\top \mathbb{E}(\vec{\boldsymbol{X}}_r \vec{\boldsymbol{X}}_r^\top)(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)\right] \\
&= \left[\tau^2 + A(r)\right] \cdot \frac{(n+1)(n-2)}{n(n-r-2)}.
\end{aligned} \tag{8}$$

The prediction error under linear models is also reported in Hastie et al. (2019), where the authors studied ridgeless regression in the growing-p-&-n setting, with p / n assumed to converge to a limit γ. Like most articles in the large-p-small-n literature, they assumed the residual variance, in our notation $A(p)+\tau^2$, is free of p. Under such an assumption, we see from (8) (after replacing r by p), that for any value of $\tau^2 > 0$, the prediction error always explodes when $\gamma = p/n$ approaches 1, yielding the turning point for the "double-descent" phenomenon that we will discuss in Section 3.4.

However, under our MR framework, it is clear that as the number of predictors increases, the variance unexplained, that is, the residual variance will decrease in general. Hence, it makes little statistical sense to assume $A(r)$ will stay as a constant as r changes – if this were the case, what would be the point of including more predictors? By explicitly considering the behavior of the unexplained variance as number of predictors increases, the prediction error can have very different characteristics under different scenarios. In particular, it is quite clear from Equation (8) that when $\tau^2 = 0$, the prediction error may not explode when r / n approaches one, because $A(r)$ is approaching zero as well, creating a limit of the form 0/0, whose value will depend on the rate at which $A(r)$ approaches zero. We will investigate this issue shortly in Section 3.3 when $\tau^2 = 0$, where we reveal the phenomenon for the optimal resolution R to be as close to n as possible, traditionally considered impossible because it is in the region of (nearly) over-fitting.

3.2. General Results Motivated and Illustrated by Linear Regression

Under the linear model (6), when the intrinsic variance is positive, that is, $\tau^2 > 0$, we can show that for any sequence of resolution levels $\{r_n\}$, a necessary condition for $\varepsilon(r_n, n) = o(1)$ is $\lim_{n\to\infty} r_n/n = 0$. Moreover, under this condition, $\varepsilon(r_n, n) \asymp r_n/n$; see the Appendices for a proof. More generally, we expect that $\varepsilon(r_n, n) \asymp \dim(\boldsymbol{\theta}_r)/n$ holds for continuous predictors under regularity conditions.

In general cases with continuous covariates, typically $\dim(\boldsymbol{\theta}_r) \asymp r^\alpha$ for some $\alpha > 0$. The following theorem considers an assumption involving $\varepsilon(r, n) \asymp \dim(\boldsymbol{\theta}_r)/n \asymp r^\alpha/n$. That is, the linear model motivates us to consider this assumption of polynomial estimation error rate in resolution, but the result below is not restricted to the linear model. All proofs are given in the Appendices.

Theorem 2. Let R_n be a rate-optimal resolution, and $L_n = A(R_n)+\varepsilon(R_n, n)$ be the corresponding minimal prediction error (after removing the ultimate risk). Then we have the following asymptotic results under each condition on the decay rate of $A(r)$ (as underlined), but all assume *polynomial estimation error*, that is, $\varepsilon(r, n) \asymp r^\alpha/n$, where $\alpha > 0$. (As in Theorem 1, all $\xi > 0$.)

(i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then $R_n \asymp 1$ with the constraint that $\liminf_{n\to\infty} R_n \geq r_0$; and $L_n \asymp n^{-1}$.

(ii) Exponential Decay: $A(r) \asymp e^{-\xi r}$. Then $R_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{1-\xi a_n} \log^{-\alpha}(n) = O(1)$; and $L_n \asymp n^{-1} \log^\alpha(n)$.

(iii) Polynomial Decay: $A(r) \asymp r^{-\xi}$. Then $R_n \asymp n^{1/(\alpha+\xi)}$; and $L_n \asymp n^{-\xi/(\alpha+\xi)}$.

(iv) Logarithmic Decay: $A(r) \asymp \log^{-\xi}(r)$. Then $R_n = a_n n^{1/\alpha} \log^{-\xi/\alpha}(n)$ with a_n satisfying $a_n = O(1)$ and $\liminf_{n\to\infty} \log(a_n)/\log(n) > -\alpha^{-1}$; and $L_n \asymp \log^{-\xi}(n)$.

This result provides precise descriptions of various restrictions on the deterministic sequence c_n in the first row of Table 1, although their details are mostly secondary to the theoretical and practical insights discussed in Section 2.3. Moreover, Theorem 2, as well as the later theorems, relies only on the rates of $A(r)$ and $\varepsilon(r,n)$, and thus can be applied to general sieves with the same rates of $A(r)$ and $\varepsilon(r,n)$. We remark that in the derivations above we can replace the expected error $\varepsilon(r,n)$ by $\varepsilon(r,\mathcal{T}_n)$, which depends on the actual training set, as in Equation (3). That is, we can seek resolution levels $\{r_n\}$ such that $A(r_n) + \varepsilon(r_n,\mathcal{T}_n)$ converges to zero in probability in the fastest way. The results remain the same if we replace "$\asymp$" by "$\overset{\mathbb{P}}{\asymp}$". Indeed, for the linear model (6) with positive τ^2 we show in the Appendices that (a) for any resolution $\{r_n\}$, $r_n/n = o(1)$ is necessary for the actual estimation error $\varepsilon(r,\mathcal{T}_n)$ to be $o_{\mathbb{P}}(1)$, and (b) when $r_n/n = o(1)$, $\varepsilon(r,\mathcal{T}_n) \overset{\mathbb{P}}{\asymp} r_n/n$. Therefore, Theorem 2 applies with $\alpha = 1$ and "$\asymp$" replaced by "$\overset{\mathbb{P}}{\asymp}$".

3.3. Specific Results for Linear Regression Without Variance

When $\tau^2 = 0$, however, we are entering a rather different world. Under model (6) with zero $\sigma_\eta^2 (= \tau^2)$, the response Y is (almost surely) a deterministic function of the countably many covariates. This is not merely a philosophical contemplation, but a mathematical reality. Indeed, any random variable can be obtained deterministically from a set of uniform variables on the unit interval, and any such uniform variable admits the binary expansion $\sum_{i=1}^{\infty} 2^{-i} U_i$, where $\{U_i, i \geq 1\}$ are iid Bernoulli(1/2); see Zhang (2019) for an investigation of using this deterministic expansion to study statistical independence.

Of course, empirically it is impossible to test whether $\tau^2 = 0$. Hence, one would expect or at least hope that it is inconsequential for practical purposes to set $\tau^2 = 0$ or not, as alluded to in Meng (2014). Therefore we were surprised initially when we saw the critical dependence of our asymptotic results on whether $\tau^2 = 0$ or not. When $\tau^2 = 0$, the asymptotic error $\varepsilon(r,n)$ is no longer dominated by the usual r / n order, but by $A(r)$ itself, as discussed previously. Specifically, contrasting with the case where $\tau^2 > 0$, $r/n = o(1)$ is no longer a necessary requirement for $\varepsilon(r,n)$ to converge to zero, because $A(r)$ can drive the error to zero even if $r/n \to 1$, as seen in Equation (8). This fact leads to different results from Theorem 2, as summarized below. We emphasize that the following theorem, although focuses on the linear model, also holds for cases where the estimation error following the same rate as that in (8).

Theorem 3. Under model (6) with $\tau^2 = 0$ and L^2 loss, the rate-optimal resolution R_n and the corresponding minimal prediction error $L_n = A(R_n) + \varepsilon(R_n, n)$ have the following forms under each condition on the decay rate of $A(r)$, where all $\xi > 0$.

(i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. The optimal resolution is any R_n such that $\liminf_{n\to\infty} R_n \geq r_0$ and $R_n \leq n-3$; and $L_n = 0$ for sufficiently large n.

(ii) Exponential Decay: $A(r) \asymp e^{-\xi r}$. $R_n = n - O(1)$ with $R_n \leq n-3$; and $L_n \asymp n e^{-\xi n}$.

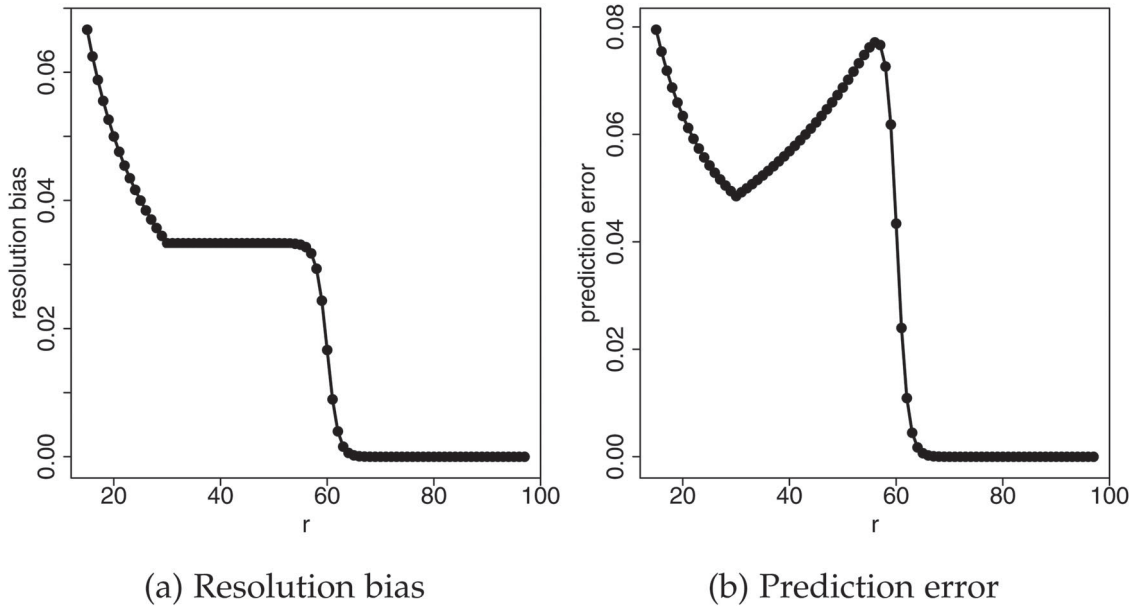
(iii) Polynomial Decay: $A(r) \asymp r^{-\xi}$. $R_n = a_n n$ with a_n satisfying $a_n \asymp 1$ and $\limsup a_n < 1$; and $L_n \asymp n^{-\xi}$.

(iv) Logarithmic Decay: $A(r) \asymp \log^{-\xi}(r)$. Optimal resolution is any R_n such that $\limsup R_n/n < 1$, $\liminf \frac{\log R_n}{\log n} > 0$; and $L_n \asymp \log^{-\xi}(n)$.

The most unexpected finding here is that, unlike the case with $\tau^2 > 0$ where no optimal R_n approaches over-fitting, that is, having R_n close to n, all four cases here permit or even require R_n to be the same order as n. When $A(r)$ has a hard threshold or decays exponentially, we can even allow $R_n = n-3$, almost the largest resolution level by which we can fit an ordinary least squares given sample size n (recall we have $r+1$ unknown parameters at resolution r). When $A(r)$ decays polynomially or logarithmically, we can choose $R_n = cn$ for some constant $c \in (0,1)$. That is, the usual concerns with over-fitting disappear. Another unexpected finding is that the logarithmic case permits $R_n \asymp n^k$ for $k \in (0,1)$, which is smaller than the polynomial case, against our intuition that slower decay should require a larger number of covariates. However, this does not contradict Theorem 2, which applies only to cases with $\tau^2 > 0$.

These unexpected theoretical results compel us to think harder about our intuitions built from the results in Section 3.2, which are consequences from the principle of bias-variance tradeoff. Does the principle fail here, as some declared about the "double-descent" phenomena in machine learning, which apparently can also prefer over-fitted models (e.g., Belkin et al. 2019; Hastie et al. 2019; Nakkiran et al. 2019)? Whereas more research is needed to understand the deterministic regime as identified by Theorem 3, our current understanding is that the bias-variance tradeoff is sound and well. In a world with zero variance, the optimal tradeoff should place all its bets on the bias term. In a deterministic world, the more mathematical constraints imposed for solving a set of equations, the smaller is the set of potential solutions. Without any variance, any specific individual case is a hard mathematical constraint for reconstructing the deterministic relationship between the outcome and the predictors. It is not surprising therefore—retrospectively—that the mathematics is instructing us to use as higher resolution as possible, except for saving some degrees of freedom to take care of the "pseudo-variance" caused by $A(r)$, when it does not decay sufficiently rapidly.

Attempting to understand this preference for over-fitting by the deterministic setup, we realize that the "double-descent" phenomenon may not be due to over-fitting as currently depicted, or at least it can also occur within the "under-fitting" region. In the current literature, "double descents" refers to the phenomenon that as p increases, the prediction error or risk first decreases due to the bias reduction, and then increases due to the inflated variance. However, as p exceeds (effective) data size, the prediction error decreases again, that is, it exhibits a double-descent phenomenon. Many researchers have tried to understand this phenomenon, and most of the studies attribute



(a) Resolution bias (b) Prediction error

Figure 1. Figures plotting the resolution bias in (9), as well as the corresponding prediction error with $\tau^2 = 0$, against the resolution r.

it to over-parameterization and that the fitted model tends to be the smoothest one interpolating all training samples; see, for example, Belkin et al. (2019) and Hastie et al. (2019).

The section below demonstrates that double and indeed multiple descents can occur without over-parameterization. This fact suggests that the issue of ordering covariates discussed in Section 2.5 is an intrinsic one, and that the reasons for the double-descent phenomena in machine learning might be more nuanced than over-parameterization.

3.4. No Surprises: Double- and Multiple-Descent Phenomenon

We first consider a setting which demonstrates a double-descent phenomena within the under-fitting region. We assume that the resolution bias has the following form:

$$A(r) = \begin{cases} r^{-1}, & \text{if } r \le \underline{r}, \\ \frac{1+\exp(\underline{r}-\bar{r})}{\underline{r}} \cdot \frac{1}{1+\exp(r-\bar{r})}, & \text{if } r > \underline{r}, \end{cases} \tag{9}$$

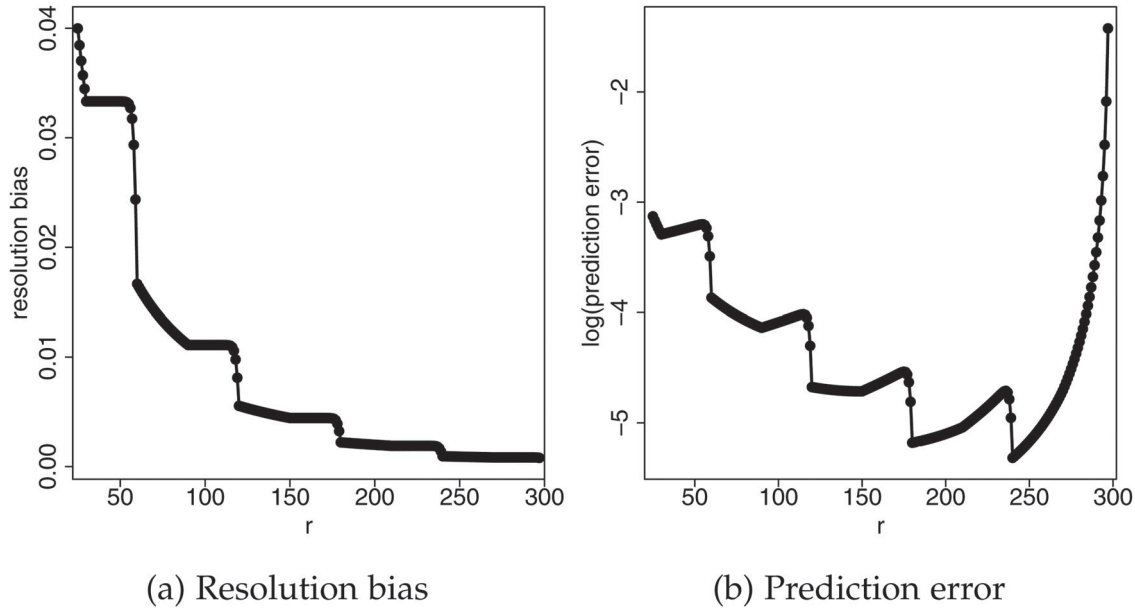
where $\underline{r} \le \bar{r}$ are two positive integers, and the coefficient $\{1 + \exp(\underline{r} - \bar{r})\}/\underline{r}$ for $r > \underline{r}$ is chosen such that $A(r)$ is a continuous function of r. Figure 1(a) plots the resolution bias against the resolution when $\underline{r} = 30$ and $\bar{r} = 60$. Figure 1(b) shows the average prediction loss (8) when $\tau^2 = 0$ and $n = 100$, which clearly demonstrates a "double-descent" phenomenon. Comparing Figures 1a and b, we can see that the double-descent pattern of the prediction error is driven by the varying importance of the added covariates. That is, when we add covariates with little predictive power, we are essentially adding noise to our prediction and hence increase the predictive error, until we add more powerful covariates to (again) bring the error down.

With this insight, it is easy to demonstrate multiple-descent phenomenon for as many descents as we want. For example, we can take

$$A(r) = \begin{cases} \mathbb{1}\{r \le \underline{r}_1\} \cdot r^{-1} + \mathbb{1}\{r > \underline{r}_1\} & \\ \quad \cdot \frac{1+\exp(\underline{r}_1-\bar{r}_1)}{\underline{r}_1} \cdot \frac{1}{1+\exp(r-\bar{r}_1)}, & \text{if } r \le \bar{r}_1, \\ c_2\mathbb{1}\{r \le \underline{r}_2\} \cdot r^{-1} + c_2\mathbb{1}\{r > \underline{r}_2\} & \\ \quad \cdot \frac{1+\exp(\underline{r}_2-\bar{r}_2)}{\underline{r}_2} \cdot \frac{1}{1+\exp(r-\bar{r}_2)}, & \text{if } \bar{r}_1 < r \le \bar{r}_2, \\ c_3\mathbb{1}\{r \le \underline{r}_3\} \cdot r^{-1} + c_3\mathbb{1}\{r > \underline{r}_3\} & \\ \quad \cdot \frac{1+\exp(\underline{r}_3-\bar{r}_3)}{\underline{r}_3} \cdot \frac{1}{1+\exp(r-\bar{r}_3)}, & \text{if } \bar{r}_2 < r \le \bar{r}_3, \\ \ldots & \end{cases} \tag{10}$$

where $\underline{r}_1 \le \bar{r}_1 \le \underline{r}_2 \le \bar{r}_2 \le \underline{r}_3 \le \bar{r}_3 \le \ldots$ and c_k's are chosen such that $A(r)$ is a continuous function of r. Figure 2(a) plots the resolution bias $A(r)$ against the resolution r when $\bar{r}_k = \underline{r}_k + 30 = 60k$ for $k \ge 1$. From Figure 2(a), we can see that, as r increases, the resolution bias keeps repeating the pattern in Figure 1(a), that is, the importance of added covariates keeps fluctuating. Figure 2(b) plots the logarithm of the average prediction error in Equation (8) against the resolution when the sample size $n = 300$ and the intrinsic error $\tau^2 = 0$. Clearly, Figure 2(b) exhibits a multiple-descent phenomenon. However, in contrast to Figure 1(b), the prediction error does not die down in the end. This is because the resolution bias in Figure 1(a) decays exponentially, while that in Figure 2(a) interweaves between exponential and polynomial decays, not covered by our theorems.

From the above discussion, it is not difficult to see that double- or multiple-descent phenomena are driven by the varying decay of resolution bias and inflation of the estimation error. Depending on which of these two terms is dominating, the prediction error can either decrease or increase, and can thus exhibit multiple-descent patterns. A reviewer points out that the multiple-descent phenomenon can also occur when most of the covariates are irrelevant and the relevant ones



(a) Resolution bias (b) Prediction error

Figure 2. Figures plotting the resolution bias in (10), as well as the corresponding prediction error (with $\tau^2 = 0$), against the resolution r.

appear sporadically. Such phenomena are also not restricted to regression settings. For example, in the midst of revising this article, we learned about Liang, Rakhlin, and Zhai (2020), which demonstrated multiple-descent phenomena in kernel machines and neural networks.

We remark that, for any monotonically decreasing function $A(r)$, we can construct a linear model with $A(r)$ as its decay rate, so all the examples above are realizable. Let $X_0 = 1$, $\{X_1, X_2, X_3, \ldots\}$ be iid standard normal random variables, and $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$. Define β_0 to be any constant, and $\beta_r = \sqrt{A(r-1) - A(r)}$, for any $r \geq 1$. Then the corresponding linear model (6) has the desired resolution bias $A(r)$. We will use this construction in the following simulation study.

3.5. Finite Sample Performance—Preliminary Findings

Whereas theoretical results are extremely useful for providing deep understanding and revealing new insights, we must be mindful that they may or may not match the empirical findings with finite samples. As a first step toward a comprehensive (and very challenging) study of our MR framework with finite samples, we conducted a simulation study using the normal linear model in Section 3.1. The simplicity of this model allows us to compute the optimal resolution and minimal prediction error exactly for any given $n(\geq 3)$, which can then be used as benchmarks to investigate the performance of various estimators for the optimal resolution. However, the model is still sufficiently rich and realistic to both confirm some of the asymptotic findings, including the resistance to over-fitting in the absence of intrinsic variation, and to reveal complications with finite samples that are not captured by the asymptotic results.

Due to space limitations, we report only findings on three ways of estimating prediction error curves in finite samples as functions of the resolution r, which then can be minimized for estimating optimal resolution. The three methods are based on cross validation (CV), an unbiased estimator (UE), and an information criteria (IC); see Appendix A9 for details and all other findings. Figure 3 plots the logarithm of averages of the three estimators over 500 Monte Carlo replications against the resolution level r, under different choices of the decay rate $A(r)$ and intrinsic variance τ^2, all with $n = 50$.

We see that UE worked well by being unbiased, CV performed well except when venturing into the over-fitting region, and IC failed badly other than when r is small. The only exception is when there is no bias-variance tradeoff, as depicted in plot (d), where the optimal resolution reaches the sample size, in which case the gross over-fitting tendency of IC brings benefit instead of damage. All six curve shapes are consistent with the theoretical findings in Theorem 2 (for $\tau^2 > 0$) and in Theorem 3 (for $\tau^2 = 0$).

4. Predictions With Infinitely Many Categorical Predictors

4.1. Regression Tree Models With Infinitely Many Categorical Covariates

We now introduce regression tree models with infinitely many categorical covariates, and then use them to illustrate some general results on rate optimal resolution and prediction. Specifically, we assume both target and training populations satisfy

$$X_1, X_2, \ldots \text{ are iid with } \mathbb{P}(X_i = k) = M^{-1} \text{ for } k = 1, 2, \ldots, M, \quad \mathbb{V}(Y) < \infty, \tag{11}$$

and the dependence of Y on $\{X_1, X_2, \ldots\}$ is arbitrary, where $M \geq 2$. That is, (11) is a regression tree in which each covariate increases the depth of the tree by one, and hence it is a tree of (potentially) infinite depth. The loss function is again the square loss: $\mathcal{L}_\odot(y, \hat{y}) = \mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$, and the prediction function at resolution r is fully saturated, that is, it can have different values for different covariates up to resolution r,

$$g(\vec{\boldsymbol{x}}_r, \boldsymbol{\theta}_r) = \sum_{\vec{\boldsymbol{a}}_r \in \{1,2,\ldots,M\}^{r+1}} \mathbb{1}(\vec{\boldsymbol{x}}_r = \vec{\boldsymbol{a}}_r)\boldsymbol{\theta}_r(\vec{\boldsymbol{a}}_r),$$

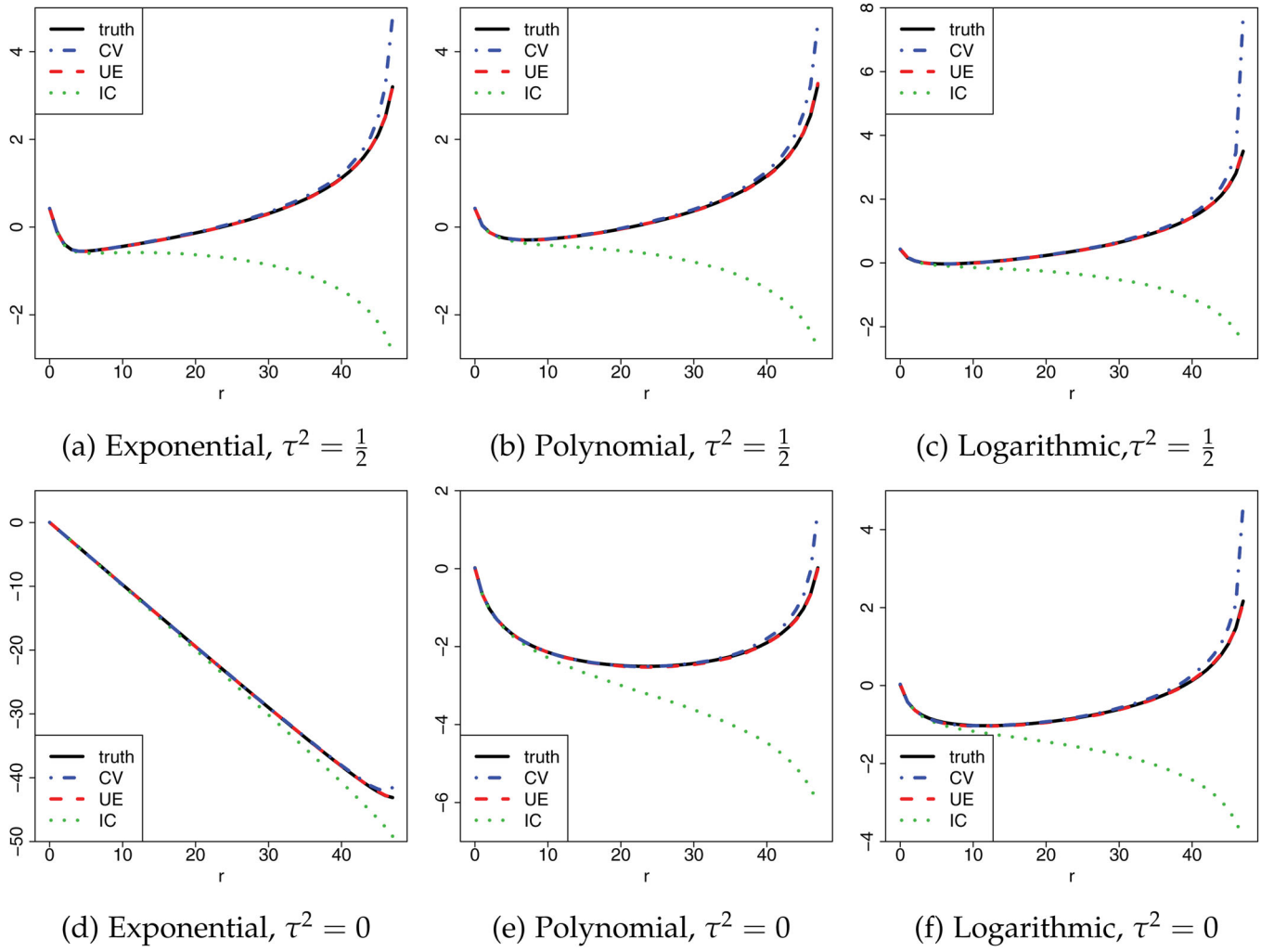


Figure 3. The performance of three strategies CV, UE and IC for estimating prediction error when $n = 50$ and, respectively, with $\tau^2 = 0.5$ (top row) and $\tau^2 = 0$ (bottom row) . The x-axis denotes resolution level r, and the y-axis denotes the logarithm of the true and estimated average prediction error over 500 simulated training sets. The resolution biases follow the decay rates of e^{-r}, r^{-1} and $\{\log(r)\}^{-1}$, respectively, for the three scenarios in (a)–(f).

where the summation is essentially over M^r terms because $X_0 \equiv 1$, $\dim(\boldsymbol{\theta}_r) = M^r$ and $\boldsymbol{\theta}_r(\vec{\boldsymbol{a}}_r)$ denotes the coordinate corresponding to covariate value $\vec{\boldsymbol{a}}_r$.

Given a training set $\mathcal{T}_n$, for each resolution r, we use $n(\vec{\boldsymbol{x}}_r)$ to denote the number of units with covariate value $\vec{\boldsymbol{x}}_r$. When $n(\vec{\boldsymbol{x}}_r) > 0$, minimizing the empirical risk will lead to taking the sample average of the outcome of these $n(\vec{\boldsymbol{x}}_r)$ individuals. The matter is more complicated when $n(\vec{\boldsymbol{x}}_r) = 0$. Here we adopt the "highest-resolution imputation". That is, for each individual of interest, we find training samples that have the same covariates up to a resolution that is as large as possible but is truncated at r, and then use their average response as a prediction for this individual. Note that this estimator is unique conditioning on the given order of the predictors. Consequently, our estimator for the parameter $\boldsymbol{\theta}_r$ has the following form:

$$\hat{\boldsymbol{\theta}}_r(\vec{\boldsymbol{x}}_r) = \begin{cases} \frac{1}{n(\vec{\boldsymbol{x}}_r)} \sum_{i:\vec{\boldsymbol{x}}_{ir}=\vec{\boldsymbol{x}}_r} Y_i, & \text{if } n(\vec{\boldsymbol{x}}_r) > 0, \\ \frac{1}{n(\vec{\boldsymbol{x}}_k)} \sum_{i:\vec{\boldsymbol{x}}_{ik}=\vec{\boldsymbol{x}}_k} Y_i, & \text{if } n(\vec{\boldsymbol{x}}_k) > 0 \text{ and } n(\vec{\boldsymbol{x}}_{k+1}) = 0, \\ & \text{for } 0 \le k < r. \end{cases} \tag{12}$$

This estimator is always well-defined, because $n(\vec{\boldsymbol{x}}_0) = n > 0$. Under model (11), we can derive that (i) the ultimate risk is $\tau^2 + \mathbb{E}[\mathbb{V}(Y|\vec{\boldsymbol{X}}_\infty)]$, (ii) the resolution bias is

$$A(r) = \sum_{k=r+1}^{\infty} \left\{ \mathbb{E}[\mathbb{V}(Y|\vec{\boldsymbol{X}}_{k-1})] - \mathbb{E}[\mathbb{V}(Y|\vec{\boldsymbol{X}}_k)] \right\},$$

and (iii) the estimation error is $\varepsilon(r, \mathcal{T}_n) = \mathbb{E}[\hat{\boldsymbol{\theta}}_r(\vec{\boldsymbol{X}}_r) - \mathbb{E}(Y|\vec{\boldsymbol{X}}_r)]^2$. The expectation of $\varepsilon(r, \mathcal{T}_n)$ over the training sets has three terms, as indicated and simplified below

$$\begin{aligned} \varepsilon(r, n) &= \left[A(r) + \tau^2\right] \cdot \mathbb{E}_n \left[\frac{\mathbb{1}(n(\vec{\mathbf{1}}_r) > 0)}{n(\vec{\mathbf{1}}_r)} \right] \\ &\quad + \sum_{k=0}^{r-1} \left[A(k) + \tau^2\right] \cdot \mathbb{E}_n \left[\frac{\mathbb{1}(n(\vec{\mathbf{1}}_k) > 0, n(\vec{\mathbf{1}}_{k+1}) = 0)}{n(\vec{\mathbf{1}}_k)} \right] \\ &\quad + \sum_{k=0}^{r-1} [A(k) - A(r)] \cdot \mathbb{E}_n \left[\mathbb{1}(n(\vec{\mathbf{1}}_k) > 0, n(\vec{\mathbf{1}}_{k+1}) = 0) \right] \\ &= \mathbb{E}_n \left[\frac{A(\mathcal{K} \wedge r) + \tau^2}{n\left(\vec{\mathbf{1}}_{\mathcal{K} \wedge r}\right)} \right] + \sum_{k=0}^{r-1} [A(k) - A(r)] \\ &\quad \cdot \mathbb{E}_n \left[\mathbb{1}(n(\vec{\mathbf{1}}_k) > 0, n(\vec{\mathbf{1}}_{k+1}) = 0) \right], \end{aligned} \tag{13}$$

where $n(\vec{\mathbf{1}}_k)$ denotes the number of training samples with covariate value $\vec{\boldsymbol{x}}_{ik} = \vec{\mathbf{1}}_k$, $\mathcal{K}$ is the maximum integer k such that $n(\vec{\mathbf{1}}_k) > 0$, and $\mathcal{K} \wedge r = \min\{\mathcal{K}, r\}$. Note that here $n(\vec{\mathbf{1}}_k) \sim \text{Binomial}(n, M^{-k})$ and $n(\vec{\mathbf{1}}_{k+1})|n(\vec{\mathbf{1}}_k) \sim \text{Binomial}(n(\vec{\mathbf{1}}_k), M^{-1})$ for any $k \geq 0$. We stress that it is the assumption that all $\vec{\boldsymbol{x}}_k$'s are uniformly distributed that permits us to replace $n(\vec{\boldsymbol{x}}_k)$ by $n(\vec{\mathbf{1}}_k)$, which greatly simplifies the derivation; see Appendix A5 for deriving error decomposition under model (11).

4.2. General Results Inspired and Illustrated by Regression Tree

Under Equation (11), when $\tau^2 > 0$, we can show that for any sequence $\{r_n\}$, a necessary condition for $\varepsilon(r_n, n) = o(1)$ is that $\lim_{n\to\infty} M^{r_n}/n \to 0$. Moreover, under this condition, the convergence rate of $\varepsilon(r, n)$ is M^r/n, that is, $\varepsilon(r, n) \asymp M^r/n \asymp \dim(\boldsymbol{\theta}_r)/n$. Again, these intuitive results require some rather technical proofs, given in the Appendices.

This inspires us to consider more general cases with categorical covariates in which $\dim(\boldsymbol{\theta}_r) \asymp \alpha^r$ for some $\alpha > 1$; for example, $\alpha = 2$ if the covariates are all binary, and the prediction function $g(\vec{\boldsymbol{x}}_r, \boldsymbol{\theta}_r)$ can have different values for each of the 2^r possible values of $\vec{\boldsymbol{x}}_r$. This contrasts with the previous case featuring continuous covariates in which the dimension of parameters increases polynomially with the resolution. The following theorem is the counterpart of Theorem 2 under the exponential estimation error.

Theorem 4. Same notation and setup as in Theorem 2, except that we now assume *exponential estimation error*: $\varepsilon(r, n) \asymp \alpha^r/n$, for some $\alpha > 1$. As in Theorem 2, all $\xi > 0$.

(i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then $R_n \asymp 1$ with the constraint that $\liminf_{n\to\infty} R_n \geq r_0$, and $L_n \asymp n^{-1}$.

(ii) Exponential Decay: $A(r) \asymp e^{-\xi r}$. Then

$$R_n = [\log(n) + \log(a_n)][\log(\alpha) + \xi]^{-1} \quad \text{with} \quad a_n \asymp 1;$$

and $L_n \asymp n^{-\xi/\{\log(\alpha)+\xi\}}$.

(iii) Polynomial Decay: $A(r) \asymp r^{-\xi}$. Then $R_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{a_n \log(\alpha)-1} \log^{\xi}(n) = O(1)$; and $L_n \asymp \log^{-\xi}(n)$.

(iv) Logarithmic Decay: $A(r) \asymp \log^{-\xi}(r)$. Then $R_n = a_n \log(n)$ with a_n satisfying

$$\liminf_{n\to\infty} \frac{\log(a_n)}{\log\log(n)} > -1, \quad \text{and} \quad \frac{[\log\log(n)]^{\xi}}{n^{1-a_n\log(\alpha)}} = O(1);$$

and $L_n \asymp [\log\log(n)]^{-\xi}$.

4.3. Specific Results for Deterministic Regression Tree

Similar to Section 3.3, we consider the case in which the ultimate risk $\tau^2 = 0$, and we will see again below how this leads to rather different asymptotic behavior. But unlike Section 3.3, even when we restrict ourselves to the regression tree model, the exact asymptotic rate for the estimation error is still difficult to obtain other than when $A(r)$ has a hard-thresholding decay. We therefore adopt a two-step strategy. We first establish an upper bound of the estimation error, yielding a corresponding upper bound for the prediction error, which can then be optimized to obtain the minimal upper-bound rate. We then prove that these minimal upper-bound rates are also the maximal lower-bound rates, except for a couple of cases where our proof fails, and hence whether the upper-bound rates are optimal or sharp is still an open problem.

Specifically, as proved in the appendices, the estimation error can be bounded by

$$\varepsilon(r, n) \leq \frac{2M}{n} \sum_{k=0}^{r} M^k A(r) \equiv \overline{\varepsilon}(r, n).$$

Furthermore, $\overline{\varepsilon}(r, n)$ under varying decay rates for $A(r)$ has the following form:

$$\overline{\varepsilon}(r, n) \asymp \begin{cases} n^{-1}, & \text{if } A(r) \text{ has a hard threshold or} \\ & A(r) \asymp e^{-\xi r} \text{ with } \xi > \log(M), \\ \frac{r}{n}, & \text{if } A(r) \asymp e^{-\xi r} \text{ with } \xi = \log(M), \\ A(r)\frac{M^r}{n}, & \text{if } A(r) \asymp e^{-\xi r} \text{ with } \xi < \log(M), \\ & A(r) \asymp r^{-\xi} \text{ or } A(r) \asymp \log^{-\xi}(r). \end{cases} \tag{14}$$

From (14), compared to $\varepsilon(r, n) \asymp M^r/n$ when $\tau^2 > 0$, we can see that the rate of the estimation error depends also on the resolution bias and converges to zero more quickly; this is similar to the discussion in Section 3.3 under the linear model. Moreover, $M^r/n = o(1)$ is no longer necessary for $\varepsilon(r, n) = o(1)$. In particular and somehow surprisingly, when the resolution bias decays exponentially with rates faster than or equal to M^{-r}, the estimation error behaves like the usual parameter setting as in Theorem 2 with a fixed number of (or r) unknown parameters at resolution r, even the model at each resolution r allows potentially M^r unknown parameters.

The following theorem summarizes sufficient conditions for the prediction error to achieve certain (upper-bound) rates under varying decay rate of the resolution bias.

Theorem 5. Under the model (11) with $\tau^2 = 0$ and L^2 loss, let $L_n = A(R_n) + \varepsilon(R_n, n) \leq A(R_n) + \overline{\varepsilon}(R_n, n) \equiv \overline{L}_n$. The rate-optimal resolution R_n or $\overline{R}_n$ and the corresponding optimal L_n or $\overline{L}_n$, respectively, have the following forms under each $A(r)$, where $\xi > 0$.

(i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then R_n satisfies that $\liminf_{n\to\infty} R_n \geq r_0$; and $L_n \asymp (1 - M^{-r_0})^n$.

(ii) Exponential Decay: $A(r) \asymp e^{-\xi r}$.

  (a) If $\xi > \log(M)$, then $\overline{R}_n$ satisfies $ne^{-\xi\overline{R}_n} = O(1)$; and $\overline{L}_n \asymp n^{-1}$.
  (b) If $\xi = \log(M)$, then $\overline{R}_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{1-a_n\log(M)}/\log(n) = O(1)$; and $\overline{L}_n \asymp n^{-1}\log(n)$.
  (c) If $\xi < \log(M)$, then $\overline{R}_n = a_n \log(n)$ with a_n satisfying $n^{a_n\log(M)-1} \asymp 1$; and $\overline{L}_n \asymp n^{-\xi/\log(M)}$.

(iii) Polynomial Decay: $A(r) \asymp r^{-\xi}$. Then $\overline{R}_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{a_n\log(M)-1} = O(1)$; and $\overline{L}_n \asymp \log^{-\xi}(n)$.

(iv) Logarithmic Decay: $A(r) \asymp \log^{-\xi}(r)$. Then $\overline{R}_n = a_n \log(n)$ with a_n satisfying

$$\liminf_{n\to\infty} \frac{\log(a_n)}{\log\log(n)} > -1, \quad \text{and} \quad n^{a_n \log(M)-1} = O(1);$$

and $\overline{L}_n \asymp [\log\log(n)]^{-\xi}$.

Next we prove that the optimal rates for the upper bounds of prediction errors are also precisely the optimal rates for the true prediction errors, except for the exponential decay case with $\xi \geq \log(M)$, where we can only conjecture but not prove that the results also hold. The following theorem summarizes our results, where for completeness, we include the hard thresholding case, even though Theorem 5 is exact in that case. Specifically, we say l_n is an asymptotic lower bound for the prediction error $A(r) + \varepsilon(r,n)$ and denote it as $A(r) + \varepsilon(r,n) \gtrsim l_n$, if $l_n = O(A(r_n) + \varepsilon(r_n, n))$, for any sequence $\{r_n\}$.

Theorem 6. Under model (11) with $\tau^2 = 0$ and L^2 loss, an asymptotic lower bound for $\varepsilon(r,n)+A(r)$ has the following form under each condition on $A(r)$, where $\xi > 0$.

(i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then $A(r) + \varepsilon(r,n) \gtrsim (1 - M^{-r_0})^n$.
(ii) Exponential Decay: $A(r) \asymp e^{-\xi r}$. Then $A(r) + \varepsilon(r,n) \gtrsim n^{-\xi/\log(M)}$.
(iii) Polynomial Decay: $A(r) \asymp r^{-\xi}$. Then $A(r) + \varepsilon(r,n) \gtrsim \log^{-\xi}(n)$.
(iv) Logarithmic Decay: $A(r) \asymp \log^{-\xi}(r)$. Then $A(r) + \varepsilon(r,n) \gtrsim [\log\log(n)]^{-\xi}$.

Comparing Theorems 5 and 6, we see the upper and lower bounds on L_n match except when $A(r) \asymp e^{-\xi r}$ and $\xi \geq \log(M)$. Also comparing both theorems to Theorem 4 with $\tau^2 > 0$, it is not surprising that the prediction error can achieve the same rate as that in Theorem 4 with polynomial or logarithmic rates. This is because the estimation error when $\tau^2 = 0$ converges to zero more quickly than when $\tau^2 > 0$, as shown in Equation (14). However, in Theorem 5 with polynomial or logarithmic rates, we allow $R_n = \log(n)/\log(M)$, and thus the number of unknown parameters M^{R_n} can be the same order as the sample size n. When the resolution bias $A(r)$ decays exponentially, the prediction error is able to achieve faster rate than that in Theorem 4.

More importantly, when the resolution bias $A(r)$ has a hard threshold or decays exponentially more quickly than M^{-r}, then the prediction error can achieve the usual rate n^{-1}, and the resolution R_n is allowed to even be infinity. In particular, with infinite resolution, for each individual of interest, we are essentially trying to find the training samples that are closest to this individual (in terms of exactly the same covariates up to a certain resolution), and use the average response from these training samples as our prediction. This is similar to the discussion in Section 3.3, where the usual bias-variance tradeoff now puts all its considerations on the bias term in the deterministic world.

Finally, we remark on the construction of model (11) with specific resolution bias $A(r)$ and ultimate risk τ^2. Let β_0 be any constant, and $\beta_k = M/\sqrt{M-1} \cdot \sqrt{A(k-1) - A(k)}$ for $k \geq 1$. Define $Y = \sum_{k=0}^{\infty} \beta_k [\mathbb{1}(X_k = 1) - M^{-1}] + \eta$, where $X_1, X_2, \ldots$ are iid uniform on $\{1, 2, \ldots, M\}$, η has mean zero and variance τ^2, and $\vec{\boldsymbol{X}}_\infty$ and η are independent. Then the corresponding model (11) has the desired resolution bias and ultimate risk.

5. From the Past to Future

5.1. A Logical Consequence of the Large-p-small-n Framework

We appreciate the value of permitting p to vary with n as a *mathematical strategy* for approximations, because it can capture the magnitude of p in relation to n toward determining which approximation terms can or cannot be ignored. But the same cannot be said about the *statistical understanding* of the behavior of the resulting model in real applications. As discussed in Section 3 and further argued below, this is not merely a logical or philosophical issue, but an issue of revealing correctly the actual behavior of our prediction models in practice.

Specifically, for most practical problems, the underlying generative models, however the way nature adopts or we conceptualize them, precede our data collection effort. We therefore can permit our data collection process to be influenced by the generative model, but not vice versa. Nature does not alter its behavior in anticipation of the sample size we may choose. Consequently, when we assume $p > n$ and permit $n \to \infty$, it forces the logical conclusion that $p = \infty$, if p indexes a feature of nature's generative model.

One may argue that p in the large-p-small-n asymptotics should not be conceptualized as an index of nature's behavior, but only as a human's approximation, like our primary resolution R_n. However, in the large-p-small-n framework, it is often assumed that the amount of total variation in the outcome that can be explained by p predictors is a constant when we increase n and hence p because p grows with n (e.g., Belkin et al. 2019; Hastie et al. 2019). But if p is meant to represent the number of predictors we humans use for predicting an outcome, then this assumption of fixed explainability defeats the purpose of using more predictors to improve the explainability of the predictors. When our mathematical formulation prohibits improvements, the resulting theoretical results may mislead us when they are used for building our intuitions, even though they may provide useful mathematical approximations for computational purposes.

As an illustration, let $\delta_i^2 = \mathbb{E}[(\mu_i - \mu_{i-1})^2]$, which measures the incremental contribution of the information in $\mathcal{F}_i$ in excess to that in $\mathcal{F}_{i-1}$ for explaining the variability in Y (over the population as defined by $\mathcal{F}_0$). Taking $r = 0$ in (1), we have

$$\mathbb{V}(Y|\mathcal{F}_0) \equiv \sigma_0^2 = \mathbb{E}[\sigma_\infty^2] + \sum_{i=1}^{\infty} \delta_i^2. \tag{15}$$

This implies that, as i increases, δ_i must be vanishingly small when $\mathbb{V}(Y|\mathcal{F}_0) < \infty$, a trivial condition for virtually all real-life problems. This implies that the value of p in the current large-p-small-n regime cannot possibly be a sensible index of model complexity to be used in linear fashion, because increasing, say, from $p = 2$ to $p = 4$ could be far more consequential than moving from $p = 22$ to 24. Yet it has been a common practice

in the current literature of machine learning or statistics to plot prediction errors against p. It is therefore refreshing to see some recent work for studying and plotting the error against more meaningful indexes, such as a spectral decay in Liang and Rakhlin (2019).

More broadly, the predictability of any set of covariates depends on at least (I) how any of them influence the outcome in the absence of other predictors and (II) how they are related to each other. Neither of the two can be adequately captured in general by merely their size. In this article we therefore adopt the direct measure of the decay rate in prediction error as we increase the resolution level (e.g., employing more predictors). As demonstrated in Theorems 2–6, this resolution decay rate plays a critical role in determining the optimal resolution, as well as in revealing further some problematic aspects of the current large-p-small-n framework.

5.2. Applications to Personalized Treatment

This work was initiated by the need for establishing a statistically principled and scientifically sound theory of personalized treatments (Meng 2014). Therefore, we provide a very brief review of two types of methods in the literature. The first type focuses on modeling the potential outcome of each patient given his or her covariates under each treatment arm, and it uses the resulting predictions to identify optimal treatment regimes; see Murphy (2003), Robins (2004), Zhao, Kosorok, and Zeng (2009), and Künzel et al. (2019). The second type focuses on a posited class of treatment regimes and tries to find the one that maximizes the overall outcome for all units; see Zhao et al. (2012), Laber and Zhao (2015), and Kosorok and Laber (2019).

Our results provide useful theoretical guidance and insight to both types of applications, because they are applicable to different populations of interest or target individuals, as captured by $\mathcal{F}_0$ and $\mathcal{F}_\infty$, respectively. For either approach, the key feature of our framework is the complete avoidance of imposing a relationship between p and n, and hence it is suitable for investigating an arbitrarily large number of covariates. Indeed, as we have seen in Sections 3 and 4, the MR framework can handle predictions with potentially infinitely many covariates.

5.3. The Method of Sieves for infinite-dimension Estimation

The method of sieves (Grenander 1981) deals with infinite-dimension estimation problems, by restricting the parameter estimation to a subset of the parameter space whose dimension grows with the sample size at some judiciously chosen rates (e.g., Geman and Hwang 1982; Shen and Wong 1994; Shen 1997; Johnstone 2011). The sequence of the subsets is then called a sieve, which can be viewed as a counterpart to MR's information filtration indexed by the resolution level r.

Whereas wavelets and sieve methods share similar mathematical constructs, our focus differs from the classical literature on sieves in several ways. First, we focus on prediction instead of parameter estimation. Second, for non/semiparametric estimation, the sieves for certain functional classes are well-understood. Under the MR framework, the resolution bias due to a sieve is generally more complicated, and the order of the covariates or equivalently the choice of sieve plays an important role in prediction error, as shown in Theorem 1. Third, we try to understand both sufficient and necessary conditions for asymptotically optimal prediction (as in Theorems 2–6), where the literature on sieves typically focus on upper bounds for the estimation convergence rate.

5.4. Much More Work is Needed

A most needed theoretical insight is on deciding a reasonable ordering in practice, going beyond the results in Section 2.5. We do not expect any kind of "automated choice" results, in theory or in practice, because of the no-free lunch principle. Since it is impossible to have a direct learning population, judgements and assumptions are inevitable. However, it is possible to obtain relatively general results for some specified (and practically meaningful) problems. Moreover, one may borrow ideas from regularization methods in the large-p-small-n framework, which can explore all possible choices of the subsets of the covariates (e.g., 2^p in Lasso) without any pre-ordering. How to do so effectively within the MR framework is a challenging problem given p potentially is ∞, although the observed number of covariates is always finite in practice.

As mentioned earlier, we were intrigued by the world without variance. We wonder, without ever being able to determine which world we are in, how could we be allowed to see its consequences? The answer seems to lie in the fact that $\sigma^2_\infty = 0$ is a necessary but not sufficient condition for the no bias-variance tradeoff phenomenon. As seen in the bottom row of Figure 3, this phenomenon did not occur when $A(r)$ decays too slowly, for example, polynomially or logarithmically. Note that we can always artificially create infinitely many covariates by certain series expansions of the basic covariates. The observation in the world without variance should motivate us to investigate the performance of nonparametric sieve regression when the response is indeed a deterministic function of the covariates. This observation also suggests the possibility for a black-box procedure to resist (empirically verifiable) over-fitting, when the number of patterns detectable with sufficient frequencies is far fewer than theoretically possible. In such cases, *exhaustive learning* is practically possible with sufficiently large training samples, hence there is no need for "intrinsic variance" to capture model imperfection, avoiding the creation of a petri dish for over-fitting. This possibility suggests a systematic investigation of the deterministic MR framework for complex machine learning models to see if it indeed provides an alternative explanation of the over-fitting resistant nature of these models.

Acknowledgments

The authors thank colleagues, especially James Bailie, Robin Gong, Tengyuan Liang, and Kai Zhang, as well as several meticulous reviewers for encouragements and comments, which have greatly improved both the content and presentation. They also acknowledge partial financial support from NSF grants.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019), "Reconciling Modern Machine-learning Practice and the Classical Bias–variance Trade-off," *Proceedings of the National Academy of Sciences*, 116, 15849–15854. [2,6,8,9,13]

Belkin, M., Hsu, D., and Xu, J. (2019), "Two Models of Double Descent for Weak Features," arXiv preprint arXiv:1903.07571. [2]

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao. (2013), "Valid Post-selection Inference," *The Annals of Statistics*, 41, 802–837. [2]

Bickel, S., Brückner, M., and Scheffer, T. (2007), "Discriminative Learning for Differing Training and Test Distributions," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, New York: ACM. [3]

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia: SIAM. [2]

Devaney, R. (2018), *An Introduction to Chaotic Dynamical Systems*, CRC Press. [2]

Donoho, D. L., and Elad, M. (2003), "Optimally Sparse Representation in General (nonorthogonal) Dictionaries Via l_1 Minimization," *Proceedings of the National Academy of Sciences USA*, 100, 2197–2202. [2]

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society*, Series B, 57, 301–369. [2]

Geman, S., and Hwang, C.-R. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 10, 401–414. [14]

Grenander, U.(1981), *Abstract Inference*, New York: Wiley. [14]

Hankinson, R. J. (1987), "Causes and Empiricism," *Phronesis*, 32, 329–348. [1]

Hankinson, R. J. (1995), "The Growth of Medical Empiricism," in *Knowledge and Scholarly Medical Traditions*, ed. D. Bates,New York: Cambridge University Press. [1]

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019), "Surprises in High-dimensional Ridgeless Least Squares Interpolation," arXiv preprint arXiv:1903.08560. [2,6,7,8,9,13]

Johnstone, I. M. (2011), "Gaussian Estimation: Sequence and Wavelet Models," unpublished manuscript. [2,14]

Kosorok, M. R., and Laber, E. B. (2019), "Precision Medicine," *Annual Review of Statistics and its Application*, 6, 263–286. [14]

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019), "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning," *Proceedings of the National Academy of Sciences*, 116, 4156–4165. [14]

Laber, E. B., and Zhao, Y. Q. (2015), "Tree-based Methods for Individualized Treatment Regimes," *Biometrika*, 102, 501–514. [14]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-selection Inference, With Application to the Lasso," *The Annals of Statistics*, 44, 907–927. [2]

Liang, T., and Rakhlin, A. (2019), "Just Interpolate: Kernel "Ridgeless" Regression Can Generalize," *The Annals of Statistics*, to appear. [14]

Liang, T., Rakhlin, A., and Zhai, X. (2020), "On the Multiple Descent of Minimum-norm Interpolants and Restricted Lower Isometry of Kernels," arXiv preprint arXiv:1908.10292. [10]

Meng, X.-L. (2014), "A Trio of Inference Problems That Could Win You a Nobel Prize in Statistics (If You Help Fund It)," in *Past, Present, and Future of Statistical Science*, eds. X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J.-L. Wang. Boca Raton, Florida: Chapman and Hall/CRC Press. [2,8,14]

Meng, X-L. (2021), "Statistical Paradises and Paradoxes in Big Data (II): Multi-resolution Inference, Simpson's Paradox, and Individualized Treatments," in preparation. [3]

Meyer, Y. (1993), *Wavelets: Algorithms and Applications*, Philadelphia, PA: SIAM. [2]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [14]

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019), "Deep Double Descent: Where Bigger Models and More Data Hurt," arXiv preprint arXiv:1912.02292 . [2,6,8]

Peat, F. D. (2002), *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century*, Washington, D.C.: Joseph Henry Press. [2]

Poggio, T., and Girosi, F. (1998), "A Sparse Representation for Function Approximation," *Neural Computation*, 10, 1445–1454. [2]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," In *Proceedings of the Second Seattle Symposium in Biostatistics*, eds. D. Y. Lin and P. J. Heagerty, New York: Springer. [14]

Shen, X. (1997), "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591. [14]

Shen, X., and Wong, W. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615. [14]

Sugiyama, M., and Kawanabe, M. (2012), *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. Cambridge, Massachusetts and London, England: MIT Press. [3]

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620. [2]

Zhang, K. (2019), "Bet on Independence," *Journal of the American Statistical Association*, 114, 1620–1637. [8]

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009), "Reinforcement Learning Design for Cancer Clinical Trials," *Statistics in Medicine*, 28, 3294–3315. [14]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [14]