On the Interaction between Norm and Dimensionality: Multiple Regimes in Learning

Percy Liang PLIANG@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720, USA

Nati Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

Abstract

A learning problem might have several measures of complexity (e.g., norm and dimensionality) that affect the generalization error. What is the interaction between these complexities? Dimension-free learning theory bounds and parametric asymptotic analyses each provide a partial picture of the full learning curve. In this paper, we use high-dimensional asymptotics on two classical problems—mean estimation and linear regression—to explore the learning curve more completely. We show that these curves exhibit multiple regimes, where in each regime, the excess risk is controlled by a subset of the problem complexities.

1. Introduction

Most analyses of learning algorithms proceed by identifying a measure of complexity of the learning problem and then provide either bounds or asymptotic expressions for the generalization error (risk) in terms of that complexity. For instance, for linear models, bounds based on Rademacher complexity (Bartlett & Mendelson, 2001), covering numbers (Pollard, 1984), or online learning (Cesa-Bianchi & Lugosi, 2006) depend on the *norm* (in relation to the variance of data and the noise) and not the dimensionality. On the other hand, classical parametric asymptotic analyses (van der Vaart, 1998; Liang et al., 2010) provide answers that depend only on the *dimensionality* and not the norm. There seems to be some tension here: If the sample complexity depends asymptotically only on the dimensionality, how can it be bounded in terms of only the norm?

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Haifa, Israel, 2010.

What we really want to understand is the true behavior of the excess risk $E_n(B, d)$ as a function of the sample size n, norm B, and dimensionality d. In this paper, we analyze the excess risk for two classical problems—mean estimation (Section 2) and linear regression (Section 3)—by performing a high-dimensional asymptotic analysis. In particular, we allow the complexity (B, d) of the problem to grow with the sample size n, so that the excess risk $E_n(B, d)$ converges to a non-vanishing asymptotic limit $\mathcal{E}(\tilde{B}, \tilde{d})$, where $\tilde{B}$ and $\tilde{d}$ are the rescaled complexities. We then study this limiting function as $\tilde{B}$ and $\tilde{d}$ vary to see the interaction between norm and dimensionality. We show how the excess risk can have multiple regimes, where in each regime, the excess risk is controlled by a subset of the relevant complexities. Furthermore, we find that the transitions between regimes are smooth, even asymptotically.

Notation For a vector $v \in \mathbb{R}^d$, we write $v^{\otimes} = vv^\top$. Let $X_n = O_p(1)$ denote that the sequence of random variables $(X_n)_{n \geq 1}$ is bounded in probability, that is, for every $\epsilon > 0$, there exists $M < \infty$ such that $\sup_n P(X_n > M) \leq \epsilon$. We write $X_n = O_p(Y_n)$ to mean $\frac{X_n}{Y_n} = O_p(1)$. Let $X_n \xrightarrow{P} 0$ denote convergence in probability, that is, for every $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$. When we use big-O notation, only universal constants are hidden, never parameters of the learning problem.

2. Constrained Mean Estimation

A classical problem in statistics is estimating the mean of a multivariate Gaussian from i.i.d. samples. We consider a variant of this problem where the norm of mean vector is constrained to a Euclidean ball. Even in this simple problem, we will see that two learning regimes emerge: a *random regime* controlled by the norm and a *unregularized regime* controlled by the dimensionality.

2.1. Setup

The mean estimation problem is defined as follows: Let $\mu^* \in \mathbb{R}^d$ be the unknown mean vector that we wish to estimate, and let $B = \|\mu^*\|_2$ denote its norm. We obtain n i.i.d. training points: $X^{(i)} \sim \mathcal{N}(\mu^*, \sigma^2 I_{d\times d})$ for $i = 1, \ldots, n$. Let the tuple $\Psi = (B, d, \sigma^2)$ specifies an instance of the mean estimation problem, which includes the norm B, dimensionality d, and data variance σ^2.

Given a vector $\mu \in \mathbb{R}^d$, we measure its generalization error (risk) using squared loss:

$$\epsilon(\mu) \stackrel{\text{def}}{=} \mathbb{E}_{X\sim\mathcal{N}(\mu^*,\sigma^2 I)}\left[(X-\mu)^2\right]. \tag{1}$$

Define the following estimator which minimizes the empirical risk subject to a norm constraint:

$$\hat{\mu}_n \stackrel{\text{def}}{=} \operatorname*{argmin}_{\|\mu\|_2=B} \sum_{i=1}^{n} (X^{(i)} - \mu)^2. \tag{2}$$

Our goal is to study the *excess risk* of $\hat{\mu}_n$, defined as follows:

$$E_n(\Psi) \stackrel{\text{def}}{=} \epsilon(\hat{\mu}_n) - \epsilon(\mu^*). \tag{3}$$

2.2. Preliminary Analysis

We first derive a closed form solution for the estimator $\hat{\mu}_n$. Consider the Lagrangian of the constrained optimization problem in (2): $L(\mu, \lambda) = \sum_{i=1}^{n}(X^{(i)} - \mu)^2 + \lambda\|\mu\|^2$. Differentiating L with respect to μ and setting it to zero, we get $\mu = \frac{\bar{X}}{1+\lambda}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$ is the empirical mean. To satisfy the constraint $\|\mu\|_2 = B$, we must have $\hat{\mu}_n = \frac{B\bar{X}}{\|\bar{X}\|_2}$. The estimator $\hat{\mu}_n$ is just the unconstrained estimator $\bar{X}$ projected onto the radius-B sphere.

Next, we decompose the risk in (1) into two orthogonal parts: $\epsilon(\mu) = \mathbb{E}[(X-\mu^*)^2] + (\mu-\mu^*)^2 = d\sigma^2 + (\mu-\mu^*)^2$.

Plugging the derived expressions for $\hat{\mu}_n$ and $\epsilon(\mu)$ into (3) yields the following expression for the excess risk:

$$E_n(\Psi) = \left(\frac{B\bar{X}}{\|\bar{X}\|_2} - \mu^*\right)^2. \tag{4}$$

Note that $\bar{X}$ is distributed as $\mathcal{N}(\mu^*, \frac{\sigma^2}{n})$.

2.3. Asymptotic Analysis

To analyze the excess risk $E_n(\Psi)$, we turn to asymptotics to simplify the form of $E_n(\Psi)$. In particular, we consider a sequence of problems $\Psi_n = (B_n, d_n, \sigma_n^2)$ so that the excess risk $E_n(\Psi_n)$ converges to some non-vanishing quantity $\mathcal{E}(\tilde{\Psi})$ as $n \to \infty$. The allowed sequences Ψ_n are given in the following definition:

Definition 1 (Limiting Problem Specification)**.** *A sequence of mean estimation problems $\Psi_n = (B_n^2, d_n, \sigma_n^2)$ has a limit $\tilde{\Psi} = (\tilde{B}^2, \tilde{d}, \tilde{\sigma}^2)$ if*

$$B_n^2 \to \tilde{B}^2, \quad \frac{d_n\sigma_n^2}{n} \to \tilde{d}, \quad \sigma_n^2 \to \tilde{\sigma}^2 \tag{5}$$

as $n \to \infty$.

Intuitively, the limiting problem specification $\tilde{\Psi}$ captures the essence of the mean estimation problem. The following proposition gives a precise handle of the excess risk in this limit:

Proposition 1. *Suppose a sequence of mean estimation problems $\Psi_n = (B_n, d_n, \sigma_n^2)$ has a limit $\tilde{\Psi} = (\tilde{B}, \tilde{d}, \tilde{\sigma}^2)$. Then the excess risk (3) has the following asymptotic limit:*

$$E_n \stackrel{\text{def}}{=} E_n(\Psi_n) \xrightarrow{P} \mathcal{E}(\tilde{\Psi}), \tag{6}$$

where the asymptotic excess risk is

$$\mathcal{E}(\tilde{\Psi}) = 4\tilde{B}^2 \sin^2\left(\frac{1}{2}\arctan\sqrt{\frac{\tilde{d}}{\tilde{B}^2}}\right). \tag{7}$$

Note that $\mathcal{E}(\tilde{\Psi})$ is a non-random function; this is because in high dimensions, the excess risk concentrates. Before proving the proposition, let us establish some intuitions about the regimes that $\mathcal{E}(\tilde{\Psi})$ exhibit by varying $\tilde{\Psi}$:

- Random Regime ($\tilde{B}^2 \ll \tilde{d}$): When the rescaled dimensionality $\tilde{d}$ is large, the arctan term tends to $\frac{\pi}{2}$; also, $\sin^2(\frac{\pi}{4}) = \frac{1}{2}$, so the asymptotic excess risk is $\mathcal{E} \approxeq 2\tilde{B}^2$. In this regime, the norm $\tilde{B}$ dominates the excess risk and the dimensionality $\tilde{d}$ is irrelevant.

  Geometrically, the estimator $\hat{\mu}_n$ essentially produces a random point on a $(d_n - 1)$-dimensional sphere, whose squared distance from μ_n^* concentrates around $2B_n^2$.
- Unregularized Regime ($\tilde{d} \ll \tilde{B}^2$): When the rescaled dimensionality is small, then $4\sin^2(\frac{1}{2}\arctan(x)) \approxeq x^2$, so the excess risk is $\mathcal{E} \approxeq \tilde{d}$. Here, the dimensionality $\tilde{d}$ dominates the excess risk and the norm $\tilde{B}$ is irrelevant.

  This regime is very closely related to parametric asymptotics. The maximum likelihood estimator $\bar{X}_n$ has excess risk exactly $\frac{\sigma_n^2}{n} \cdot \chi^2_{d_n}$, where $\chi^2_{d_n}$ denotes a χ^2 random variable with d_n degrees of freedom, which has mean d_n. When B_n^2 is large, the sphere looks locally flat, so the projection of $\bar{X}_n$ onto its surface simply removes an insignificant degree of freedom.

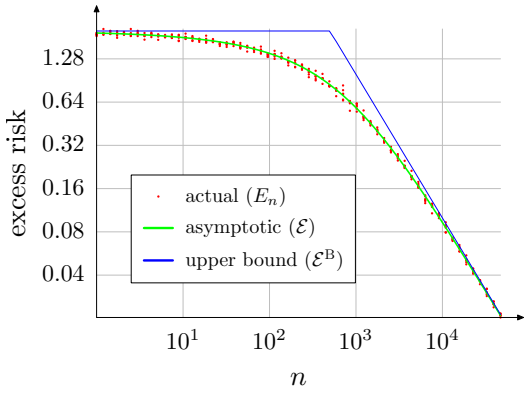


Figure 1. On constrained mean estimation: Log-log plot of the excess risk for $d_n = 1000, B_n^2 = 1, \sigma_n^2 = 1$. The corresponding limiting values are $\tilde{B}^2 = B_n^2 = 1$ and $\tilde{d} = \frac{d_n \sigma_n^2}{n} = \frac{1000}{n}$. One can clearly see the two regimes marked by their slopes (0 and -1, respectively): In the random regime, the norm $\tilde{B}$ controls the excess risk, while in the unregularized regime, the dimensionality $\tilde{d}$ does. The bound $\mathcal{E}^{\mathrm{B}} = \min\{2\tilde{B}^2, \tilde{d}\}$ represents the ends of $\mathcal{E}$ quite accurately, but misses the smooth transition in the middle.

Based on the previous discussion, we can actually stitch together the following upper bound,

$$\mathcal{E}^{\mathrm{B}}(\tilde{\Psi}) \stackrel{\mathrm{def}}{=} \min\left\{2\tilde{B}^2, \tilde{d}\right\} \geq \mathcal{E}(\tilde{\Psi}), \tag{8}$$

which clearly marks out the two regimes.

We can see how well the asymptotics represent the actual excess risk E_n for finite n by plotting the learning curve (Figure 1), increasing n while keeping Ψ_n fixed. Note that increasing n decreases the rescaled dimensionality $\tilde{d}$. From Figure 1, we see that while the bound $\mathcal{E}^{\mathrm{B}}$ matches the asymptotic $\mathcal{E}$ at the ends, there is a noticeable gap when transitioning between the two regimes. In particular, the bound is piecewise linear whereas asymptotic curve is smooth, tracking the empirical excess risk much more closely. We note that the transition is smooth even in the asymptotic limit; this is due to the smoothness of the loss function.

Proof of Proposition 1. Without loss of generality, assume $\mu_n^* = (B_n, 0, \ldots, 0)^\top$ because the problem is rotationally invariant.

First, let us decompose the pre-projected estimator $\bar{X}_n$ into two components: (1) the component in the subspace of μ_n^*, which has length $U_n \stackrel{\mathrm{def}}{=} |\bar{X}_{n1}|$, and (2) the component orthogonal to μ_n^*, which has length $V_n \stackrel{\mathrm{def}}{=} \sqrt{\sum_{j=2}^{d_n} \bar{X}_{nj}^2}$. Figure 2 depicts the setup: the excess risk E_n we want to compute is denoted geometrically.

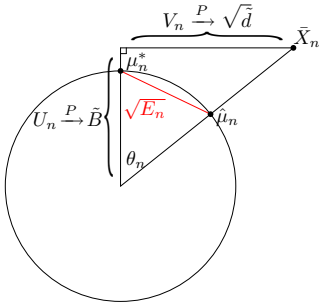


Figure 2. Geometric depiction of the excess risk E_n (used in the proof of Proposition 1). The constrained estimator $\hat{\mu}_n$ is obtained by projecting $\bar{X}_n$ down to the radius-B_n sphere. The key is that in high dimensions, the two random components U_n and V_n both concentrate.

We can obtain E_n using basic trigonometry. First, compute the angle $\theta_n = \arctan\left(\frac{U_n}{V_n}\right)$. Bisecting θ_n and converting angles back to lengths yields $\sqrt{E_n} = 2B_n \sin(\frac{1}{2}\theta_n)$. Putting everything together, we have

$$E_n = 4B_n^2 \sin^2\left(\frac{1}{2}\arctan\left(\frac{U_n}{V_n}\right)\right). \tag{9}$$

Now we compute the limits of U_n and V_n. First, U_n includes the small deviation along the first component: $U_n = B_n + O_p(\sqrt{\frac{\sigma_n^2}{n}}) \xrightarrow{P} \tilde{B}$. V_n includes the deviations along the other $d-1$ components, which amounts to $V_n = \sqrt{\frac{\sigma_n^2}{n}\chi_{d_n-1}^2} \xrightarrow{P} \sqrt{\tilde{d}}$. Since E_n depends on U_n and V_n only via smooth trigonometric functions (9), we can apply the continuous mapping theorem to obtain $E_n \xrightarrow{P} \mathcal{E}$ as desired. □

3. Regularized Linear Regression

We now turn to norm-regularized linear regression. We first analyze a componentwise estimator (which treats each parameter separately), showing that even in this simple case, the asymptotic excess risk exhibits three regimes, not two as in mean estimation. For the full least squares estimator, we use a combination of upper bounds to hypothesize the existence of four regimes.

3.1. Setup

We assume that data are generated as follows: $X \sim \mathcal{N}(0, \Sigma_{d\times d})$ and $Y = \langle X, \beta^* \rangle + W$, where $\beta^* \in \mathbb{R}^d$ is the true parameter and $W \sim \mathcal{N}(0, \sigma^2)$ is independent noise. Let $p_{\beta^*}(X, Y)$ denote the resulting distribution. We consider a *linear regression problem* to be fully specified by the tuple $\Psi = (\Sigma, \beta^*, \sigma^2)$.

Given a predictor $\beta \in \mathbb{R}^d$, we are interested in its

generalization error (risk), which averages the squared loss over test points:

$$\epsilon(\beta) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)\sim p_{\beta^*}}[(Y - \langle X, \beta\rangle)^2]. \tag{10}$$

Given n i.i.d. training points $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$ drawn from p_{β^*}, define the *regularized least-squares estimator*:

$$\hat{\beta}_n^\lambda \stackrel{\text{def}}{=} \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \left(Y^{(i)} - \left\langle X^{(i)}, \beta\right\rangle\right)^2 + \lambda\|\beta\|^2, \tag{11}$$

where $\lambda \geq 0$ is the regularization parameter. We are interested in analyzing the excess risk of this estimator:

$$E_n^\lambda(\Psi) \stackrel{\text{def}}{=} \epsilon(\hat{\beta}_n^\lambda) - \epsilon(\beta^*). \tag{12}$$

We also consider the excess risk of an *oracle estimator* which chooses the optimal λ (in a data-dependent way) to minimize the excess risk:

$$E_n^*(\Psi) \stackrel{\text{def}}{=} \inf_{\lambda\geq 0} E_n^\lambda(\Psi). \tag{13}$$

Assumption 1 (Diagonal Covariance)**.** *Assume that the covariance matrix of the data is diagonal, that is,* $\Sigma = \operatorname{diag}(\tau_1^2, \ldots, \tau_d^2)$.

This assumption is made without loss of generality for the estimators we have defined so far, which are rotationally invariant. This is not true for the componentwise estimator that we will introduce later.

3.2. Preliminary Analysis

Define the empirical covariance $\hat{\Sigma} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n X^{(i)\otimes}$ and let $\hat{S} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n X^{(i)}W^{(i)}$. First, we solve (11) in the standard way by differentiating and setting the result to zero to get $\hat{\beta}_n^\lambda = (\hat{\Sigma} + \lambda I)^{-1}(\hat{\Sigma}\beta^* + \hat{S})$. Next, since the noise W is independent of X, we can rewrite (10) as $\epsilon(\beta) = \frac{\sigma^2}{n} + \operatorname{tr}\{\Sigma(\beta - \beta^*)^\otimes\}$. Applying these two derived expressions to (12), we get $E_n^\lambda(\Psi) = \operatorname{tr}\{\Sigma(\hat{\beta}_n^\lambda - \beta^*)^\otimes\}$.

Using some algebra, we can write the parameter error as $\hat{\beta}_n^\lambda - \beta^* = -\lambda(\hat{\Sigma}+\lambda I)^{-1}\beta^* + (\hat{\Sigma}+\lambda I)^{-1}\hat{S}$. Putting the last two equations together yields:

$$E_n^\lambda(\Psi) = \operatorname{tr}\{\Sigma[\lambda(\hat{\Sigma}+\lambda I)^{-1}\beta^* - (\hat{\Sigma}+\lambda I)^{-1}\hat{S}]^\otimes\}. \tag{14}$$

Unregularized Estimator Before we consider regularization, let us comment on the unregularized estimator (when $\lambda = 0$). In this case, the excess risk $E_n^0(\Psi)$ in (14) simplifies to $E_n^0(\Psi) = \operatorname{tr}\{\hat{\Sigma}^{-1}\hat{S}^\otimes\hat{\Sigma}^{-1}\Sigma\}$. We can compute the expectation of $E_n^0(\Psi)$ in closed form. First, conditioned on $X^{(1)}, \ldots, X^{(n)}$, we can integrate out the $W^{(1)}, \ldots, W^{(n)}$ in $\hat{S}$ by independence; this yields $\mathbb{E}[\hat{S}^\otimes \mid X^{(1)}, \ldots, X^{(n)}] = \hat{\Sigma}\frac{\sigma^2}{n}$. Next, the inverse covariance matrix $\hat{\Sigma}^{-1}$ has an inverse Wishart distribution ($\hat{\Sigma}^{-1} \sim \mathcal{W}^{-1}(\frac{1}{n}\Sigma, n)$), which has mean $\frac{n\Sigma^{-1}}{n-d-1}$. Putting everything together, we obtain $\mathbb{E}[E_n^0(\Psi)] = \frac{d\sigma^2}{n-d-1}$.

It is interesting to note that the excess risk does not depend on β^* and Σ, but only on the dimensionality d. The norm of β^* does not play a role at all because the unregularized estimator is shift-invariant. The covariance of the data Σ does not play a role due to the following intuition: the larger Σ is, the easier it is to estimate β, but the harder it is to predict. The two forces cancel out exactly.

3.3. Componentwise Estimator Asymptotics

In this section, we introduce and analyze a simple estimator that still provides additional insight into the interaction between norm and dimensionality. Define the *componentwise least-squares estimator*, which estimates each component of $\hat{\beta} \in \mathbb{R}^d$ separately, as follows:

$$\hat{\beta}_j^\lambda = \frac{\hat{\tau}_j^2\beta_j^* + \hat{S}_j}{\hat{\tau}_j^2 + \lambda} \quad \forall j = 1, \ldots, d, \tag{15}$$

where $\hat{\tau}_j^2 = \hat{\Sigma}_{jj}$.

The componentwise estimator consistently estimates β^* regardless of whether Σ is diagonal. When Σ is diagonal, the excess risk is just the sum across the components, where component involves a one-dimensional regression problem. Without regularization, the expected excess risk of the componentwise estimator is $\frac{d\sigma^2}{n-2}$. Note this is smaller than the excess risk of the full unregularized estimator, which is $\frac{d\sigma^2}{n-d-1}$. We effectively gain an effective sample of size $d-1$ by exploiting knowledge of the eigenstructure of Σ.

In this section, we analyze the excess risk of the componentwise estimator using asymptotics. The lack of covariance structure simplifies the math considerably. Consider a sequence of regression problems $\Psi_n = (\tau_n^2, \beta_n^*, \sigma_n^2)$. We do not yet commit to a particular scaling, but we do impose the following constraints:

Assumption 2 (Constraints on Limiting Problem Specification)**.** *Assume that* $\limsup_n \frac{d_n\sigma_n^2}{n} < \infty$ *and* $\limsup_n \sum_{j=1}^{d_n} |\beta_{nj}^*|\tau_{nj} < \infty$.

We now derive the asymptotic excess risk of the oracle componentwise estimator:

Proposition 2. *Consider a sequence of regression problems* $\Psi_n = (\tau_n^2, \beta_n^*, \sigma_n^2)$ *satisfying Assumption 2.*

Define

$$\mathcal{E}_n^\lambda \stackrel{\text{def}}{=} \sum_{j=1}^{d_n} \underbrace{\frac{\lambda^2 \beta_{nj}^{*2} \tau_{nj}^2}{(\tau_{nj}^2+\lambda)^2}}_{\text{squared bias}} + \underbrace{\frac{\sigma_n^2 \tau_{nj}^4}{n(\tau_{nj}^2+\lambda)^2}}_{\text{variance}}. \tag{16}$$

Suppose that there exists a function $\mathcal{E}^\lambda$ such that $\sup_{\lambda\geq 0}|\mathcal{E}_n^\lambda - \mathcal{E}^\lambda| \to 0$. Then the excess risk $E_n^* \stackrel{\text{def}}{=} E_n^*(\Psi_n)$ of the oracle componentwise estimator (13) has the following asymptotic limit $\mathcal{E}^*$:

$$E_n^* \stackrel{\text{def}}{=} \inf_{\lambda\geq 0} E_n^\lambda \xrightarrow{P} \inf_{\lambda\geq 0} \mathcal{E}^\lambda \stackrel{\text{def}}{=} \mathcal{E}^*. \tag{17}$$

For a fixed λ, we will show that the excess risk E_n^λ converges to a non-random asymptotic excess risk $\mathcal{E}^\lambda$, using $\mathcal{E}_n^\lambda$ as an intermediate quantity that intuitively removes the randomness from E_n^λ. The concentration of E_n^λ around $\mathcal{E}_n^\lambda$ must be established.

What is new in regression is the minimization over λ. To establish $\inf_{\lambda\geq 0} E_n^\lambda \xrightarrow{P} \inf_{\lambda\geq 0}\mathcal{E}^\lambda$ (i.e., switching inf with lim), we need some sort of uniformity over λ, which occupies most of the following proof.

Proof of Proposition 2. To prove that $E_n^* \xrightarrow{P} \mathcal{E}^*$, it suffices to show that $\sup_{\lambda\geq 0}|E_n^\lambda - \mathcal{E}_n^\lambda| \xrightarrow{P} 0$. If we have that, the proposition can be established as follows: Let $\lambda_n \in \operatorname{argmin}_{\lambda\geq 0} E_n^\lambda$ and $\lambda^* \in \operatorname{argmin}_{\lambda\geq 0}\mathcal{E}^\lambda$. Note that $E_n^* = E_n^{\lambda_n}$ and $\mathcal{E}^* = \mathcal{E}^{\lambda^*}$. For any $\epsilon > 0$, we have $E_n^{\lambda_n} \leq E_n^{\lambda^*} \approxeq_{\frac{\epsilon}{2}} \mathcal{E}_n^{\lambda^*} \approxeq_{\frac{\epsilon}{2}} \mathcal{E}^{\lambda^*} \leq \mathcal{E}^{\lambda_n} \approxeq_{\frac{\epsilon}{2}} \mathcal{E}_n^{\lambda_n} \approxeq_{\frac{\epsilon}{2}} E_n^{\lambda_n}$ for sufficiently large n with high probability, where $a \approxeq_\epsilon b$ denote $|a-b| \leq \epsilon$. This ensures $|\mathcal{E}^* - E_n^*| \leq \epsilon$.

Now, we need to show that $\sup_{\lambda\geq 0}|E_n^\lambda - \mathcal{E}_n^\lambda| \xrightarrow{P} 0$, i.e., that the residual $|E_n^\lambda - \mathcal{E}_n^\lambda|$ goes to zero at a rate that does not depend on λ. Specializing (14) to the componentwise estimator and expanding yields:

$$E_n^\lambda = \sum_{j=1}^{d_n} \frac{\tau_{nj}^2(\lambda^2\beta_{nj}^{*2} - 2\lambda\beta_{nj}^*\hat{S}_{nj} + \hat{S}_{nj}^2)}{(\hat{\tau}_{nj}^2+\lambda)^2}. \tag{18}$$

For each $j = 1,\ldots,d_n$, let $\hat{R}_{nj1} \stackrel{\text{def}}{=} \frac{\hat{S}_{nj}}{\sigma_n\tau_{nj}} = O_p(n^{-\frac{1}{2}})$, $\hat{R}_{nj2} \stackrel{\text{def}}{=} \frac{\hat{\tau}_{nj}^2 - \tau_{nj}^2}{\tau_{nj}^2} = O_p(n^{-\frac{1}{2}})$, and $\hat{H}_{nj} \stackrel{\text{def}}{=} \frac{n\hat{S}_{nj}^2}{\sigma_n^2\tau_{nj}^2} - 1 = O_p(1)$. Importantly, these variables (1) do not depend on the problem specification Ψ_n and (2) capture all the randomness in E_n^λ. Using these variables, define:

$$F_n^\lambda(R) = \sum_{j=1}^{d_n} \frac{\lambda^2\beta_{nj}^{*2}\tau_{nj}^2 - 2\lambda\beta_{nj}^*\sigma_n\tau_{nj}^3 R_{j1} + \frac{\sigma_n^2}{n}\tau_{nj}^4}{(\tau_{nj}^2(1+R_{j2})+\lambda)^2}, \tag{19}$$

$$G_n^\lambda(R) = \sum_{j=1}^{d_n} \frac{\frac{\sigma_n^2}{n}\tau_{nj}^4\hat{H}_{nj}}{(\tau_{nj}^2(1+R_{j2})+\lambda)^2}. \tag{20}$$

We have constructed F_n^λ and G_n^λ so that $\mathcal{E}_n^\lambda = F_n^\lambda(0)$ and $E_n^\lambda = F_n^\lambda(\hat{R}_n) + G_n^\lambda(\hat{R}_n)$, which can be verified with some algebra. Intuitively, $F_n^\lambda(0)$ captures the non-random problem-dependent part of the excess risk; $\hat{R}_n$ and $G_n^\lambda(\hat{R}_n)$ contribute the random problem-independent part.

Let A_n be the event that $\|\hat{R}_n\|_\infty \leq \frac{1}{2}$. On event A_n, Lemma 1 below will show that $\|\nabla F_n^\lambda(\hat{R}_n)\|_1 \leq M$ for some constant M independent of Ψ_n and λ. Note: norms on the matrices are element-wise.

Lemma 1. *For all $R \in \mathbb{R}^{d_n\times 2}$ such that $\|R\|_\infty \leq \frac{1}{2}$ and $\lambda \geq 0$, we have $\|\nabla F_n^\lambda(R)\|_1 \leq M$, where M is a constant independent of Ψ_n and λ.*

Proof. Let $Q_{nj} \stackrel{\text{def}}{=} \tau_{nj}^2(1+R_{j2}) + \lambda$. For each j, we have $\frac{\partial F_n^\lambda(R)}{\partial R_{j2}} = -2Q_{nj}^{-3}\tau_{nj}^2(\lambda^2\beta_{nj}^{*2}\tau_{nj}^2 - 2\lambda\beta_{nj}^*\sigma_n\tau_{nj}^3 R_{j1} + \frac{\sigma_n^2}{n}\tau_{nj}^4)$. Using the fact that Q_{nj} is larger than $\frac{1}{2}\tau_{nj}^2$ and λ, we obtain $|\frac{\partial F_n^\lambda(R)}{\partial R_{j2}}| \leq 4\beta_{nj}^{*2}\tau_{nj}^2 + 8\beta_{nj}^*\sigma_n\tau_{nj} + 16\frac{\sigma_n^2}{n}$. Similarly, we can bound $|\frac{\partial F_n^\lambda(R)}{\partial R_{j1}}| \leq 4\beta_{nj}^*\sigma_n\tau_{nj}$. Sum the right-hand sides over $j = 1,\ldots,d_n$. By Assumption 2, this sum is bounded above by a quantity independent of Ψ_n and λ. □

By the mean value theorem, $F_n^\lambda(\hat{R}_n) - F_n^\lambda(0) = \nabla F_n^\lambda(c\hat{R}_n)^\top \hat{R}_n$ for some $c \in [0,1]$, where, abusing notation, $\hat{R}_n$ is treated as a vector. Applying the lemma with Hölder's inequality and taking a sup over λ yields $\sup_{\lambda\geq 0}|F_n^\lambda(\hat{R}) - F_n^\lambda(0)| \leq M\|\hat{R}_n\|_\infty$. Note this is all still conditioned on A_n.

Now we want to bound $\sup_{\lambda\geq 0}|G_n^\lambda(\hat{R}_n)|$. It suffices to take $\lambda = 0$, which is where G_n^λ attains its maximum value. Simplifying, we get $G_n^0(\hat{R}_n) = \frac{\sigma_n^2}{n}\sum_{j=1}^{d_n}\frac{\hat{H}_{nj}}{(1+\hat{R}_{nj2})^2}$. Note that each summand converges in distribution to a χ^2 distribution minus 1 (which has mean zero), independent of Ψ_n and λ.

To finish the proof, fix $\epsilon > 0$. With high probability, we can take n large enough (in a way that does not depend on Ψ_n or λ) such that (1) $\|\hat{R}\|_\infty < \frac{\epsilon}{2M}$ and event A_n holds by applying a standard tail bound plus

a union bound over the $2d_n = O(n)$ elements of $\hat{R}_n$; and (2) $|\frac{1}{d_n}\sum_{j=1}^{d_n}\frac{\hat{H}_{nj}}{(1+\hat{R}_{nj2})^2}| \leq \frac{\epsilon}{2\tilde{d}}$ by the law of large numbers.

This ensures that both $\sup_{\lambda\geq 0}|F_n^\lambda(\hat{R}) - F_n^\lambda(0)| \leq \frac{\epsilon}{2}$ and $\sup_{\lambda\geq 0}|G_n^\lambda(\hat{R})| \leq \frac{\epsilon}{2}$, implying that $\sup_{\lambda\geq 0}|E_n^\lambda - \mathcal{E}_n^\lambda| \leq \epsilon$. □

3.4. Learning Regimes

To get some concrete intuition for Proposition 2, let us specialize the problem specification:

Assumption 3 (Two-Part Regression Structure)**.** *Let the true parameter vector be* $\beta_n^* = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{d_n}$ *and the data covariance be* $\Sigma = \mathrm{diag}(B_n^2, \frac{C_n^2}{d_n-1}, \ldots, \frac{C_n^2}{d_n-1}) \in \mathbb{R}^{d_n\times d_n}$ *for some* $B_n^2, C_n^2 > 0$.

The idea is that B_n^2 captures the squared norm of the signal in the data (which exists only on the first component), and C_n^2 captures the squared norm of irrelevant components. The norm of β_n^* can always taken to be one without loss of generality, since the true measure of complexity is the product of the norm of the predictor with the norm of the data.

Definition 2 (Limiting Problem Specification)**.** *A sequence of linear regression problems* $\Psi_n = (B_n^2, C_n^2, d_n, \sigma_n^2)$ *converges to a limit* $\tilde{\Psi} = (\tilde{B}^2, \tilde{C}^2, \tilde{d}, \tilde{\sigma}^2)$ *if*

$$B_n^2 \to \tilde{B}^2,\ \frac{C_n^2\sigma_n^2}{n} \to \tilde{C}^2,\ \frac{d_n\sigma_n^2}{n} \to \tilde{d},\ \hat{\sigma}_n^2 \to \tilde{\sigma}^2 \quad (21)$$

as $n \to \infty$.

Note that we allow both the dimensionality d_n and the squared norm C_n^2 of the irrelevant components tend to ∞. The presence of C_n^2 will create a new intermediate learning regime.

Specializing the asymptotic excess risk from (16) to this problem specification:

$$\mathcal{E}_n^\lambda = \frac{B_n^2\lambda^2}{(B_n^2+\lambda)^2} + \frac{\sigma_n^2 B_n^4}{n(B_n^2+\lambda)^2} + \sum_{j=2}^{d_n}\frac{\sigma_n^2\frac{C_n^4}{(d_n-1)^2}}{n(\frac{C_n^2}{d_n-1}+\lambda)^2},$$

which can be shown to converge uniformly across $\lambda \geq 0$ to

$$\mathcal{E}^\lambda = \underbrace{\frac{\tilde{B}^2\lambda^2}{(\tilde{B}^2+\lambda)^2}}_{\text{squared bias } \stackrel{\text{def}}{=} \mathcal{U}^\lambda} + \underbrace{\frac{\frac{\tilde{C}^4}{\tilde{d}}}{(\frac{\tilde{C}^2}{\tilde{d}}+\lambda)^2}}_{\text{variance } \stackrel{\text{def}}{=} \mathcal{V}^\lambda}. \quad (22)$$

Recall that we are ultimately interested in the excess risk of the oracle estimator $\mathcal{E}^* = \inf_{\lambda\geq 0}\mathcal{E}^\lambda$, which by

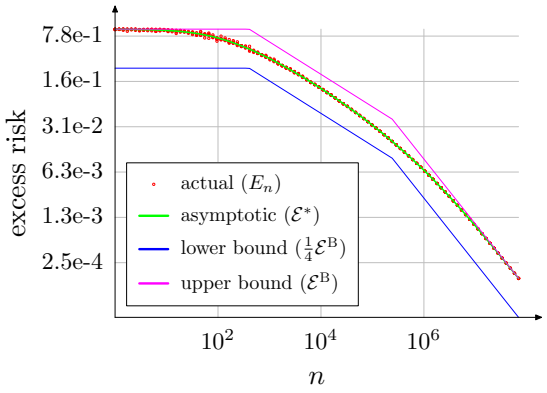


Figure 3. Componentwise estimator for linear regression: Log-log plot of the learning curve for $d = 100, B^2 = 1, C^2 = 10, \sigma^2 = 100$. There are three regimes, each characterized by a different slope (0 corresponding to a constant excess risk in the *random* regime, $-\frac{1}{2}$ corresponding to a rate of $\frac{1}{\sqrt{n}}$ in the *regularized* regime, and -1 corresponding to $\frac{1}{n}$ in the *unregularized* regime).

Proposition 2 is the limit of the excess risk E_n^* of the oracle estimator. The following proposition sheds light into the multi-regime structure of $\mathcal{E}^*$:

Proposition 3 (Bounds on Regimes)**.** *Let*

$$\mathcal{E}^{\mathrm{B}} \stackrel{\text{def}}{=} \min\left\{\tilde{B}^2, \frac{2\tilde{C}^2}{\tilde{B}\sqrt{\tilde{d}}}, \tilde{d}\right\}. \quad (23)$$

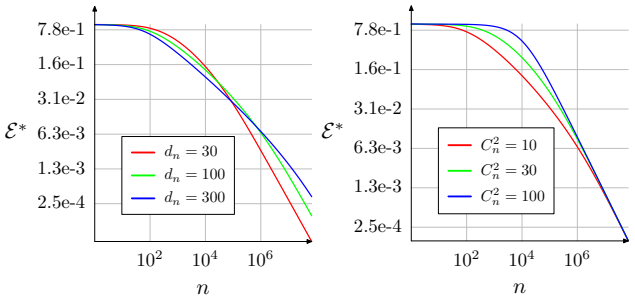
The asymptotic excess risk $\mathcal{E}^*$ *of the oracle componentwise estimator defined in (15) is bounded by* $\mathcal{E}^{\mathrm{B}}$ *to within a factor of four:*

$$\frac{1}{4}\mathcal{E}^{\mathrm{B}} \leq \mathcal{E}^* \leq \mathcal{E}^{\mathrm{B}}. \quad (24)$$

We can plot the learning curve as a relationship between the sample size and excess risk, for a fixed specification of the regression problem. Figure 3 shows the actual excess risk E_n, the asymptotic excess risk $\mathcal{E}^*$, and bounds $\frac{1}{4}\mathcal{E}^{\mathrm{B}}, \mathcal{E}^{\mathrm{B}}$. Note that $\tilde{C}^2$ and $\tilde{d}$ scale inversely with n, so that the three regimes scale as 1, $\frac{1}{\sqrt{n}}$, and $\frac{1}{n}$, respectively.

The bound (23) indicates three regimes corresponding to each of the three terms:

- Random Regime: In this regime, λ should be large so that the variance $\mathcal{V}^\lambda \to 0$ and the squared bias $\mathcal{U}^\lambda \to \tilde{B}^2$ (see (22)). This corresponds to simply guessing $\hat{\beta}_n = 0$. Only the squared norm of the signal $\tilde{B}^2$ controls the excess risk.
- Regularized Regime: In this new regime, λ must be optimized to balance the squared bias $\mathcal{U}^\lambda$ and variance $\mathcal{V}^\lambda$ terms. The squared norm of the irrelevant



(a) Varying dimensionality d_n (b) Varying norm C_n^2

Figure 4. Log-log plots of asymptotic excess risk $\mathcal{E}^*$ (same default parameters as in Figure 3), where we study the impact of varying the norm C_n^2 and dimensionality d_n. Varying the dimensionality d_n affects both the regularized and unregularized regimes (a); varying the squared norm C_n^2 only affects the regularized regime (b).

components $\tilde{C}^2$ dominates the excess risk, favorably scaled down by large $\tilde{B}$ and $\sqrt{\tilde{d}}$. The implied dependence of the excess risk on n is $\frac{1}{\sqrt{n}}$.

- Unregularized Regime: λ should be small so that $\mathcal{U}^\lambda \to 0$ and $\mathcal{V}^\lambda \to \tilde{d}$. Here, the dimensionality $\tilde{d}$ controls the excess risk, yielding an excess risk of order $\frac{1}{n}$, independent of the norm.

Figure 4 shows the impact of varying the problem parameters on the various regimes. As one would expect from (23), changing the norm C^2 only affects the intermediate regime, whereas changing the dimensionality d affects both the variance and the intermediate regime.

Proof of Proposition 3. We first prove the upper bound $\mathcal{E}^* \le \mathcal{E}^{\mathrm{B}}$. To do this, we need to show that $\mathcal{E}^*$ (defined in terms of (22)) is upper bounded by each of the three terms in (23).

To get $\mathcal{E}^* \le \tilde{B}^2$, take $\lambda \to \infty$ (infinite regularization). In this limit, the squared bias term $\mathcal{U}^\lambda$ dominates and converges to $\tilde{B}^2$.

To show $\mathcal{E}^* \le \frac{2\tilde{C}^2}{\tilde{B}\sqrt{\tilde{d}}}$, observe that $\mathcal{E}^\lambda = \mathcal{U}^\lambda + \mathcal{V}^\lambda$, where $\mathcal{U}^\lambda \le \frac{\lambda^2}{\tilde{B}^2}$ and $\mathcal{V}^\lambda \le \frac{\frac{\tilde{C}^4}{\tilde{d}}}{\lambda^2}$ (22). Optimize the sum of the two bounds with respect to λ:

$$\mathcal{E}^* \le \inf_{\lambda \ge 0} \left\{ \frac{\lambda^2}{\tilde{B}^2} + \frac{\frac{\tilde{C}^4}{\tilde{d}}}{\lambda^2} \right\} = \frac{2\tilde{C}^2}{\tilde{B}\sqrt{\tilde{d}}}, \tag{25}$$

which is attained by setting $\lambda^2 = \frac{\tilde{B}\tilde{C}^2}{\sqrt{\tilde{d}}}$.

Finally, to show that $\mathcal{E}^* \le \tilde{d}$, simply set $\lambda = 0$ (corresponding to no regularization), to get that $\mathcal{E}^\lambda = \tilde{d}$.

Now we show the lower bound $\frac{1}{4}\mathcal{E}^{\mathrm{B}} \le \mathcal{E}^*$. Using the fact that $a \ge b \ge 0$ implies $\frac{1}{(a+b)^2} \ge \frac{1}{4a^2}$, we have the following relations:

(i) $\lambda > \tilde{B}^2$ implies $\mathcal{U}^\lambda \ge \frac{1}{4}\tilde{B}^2$,
(ii) $\lambda \le \tilde{B}^2$ implies $\mathcal{U}^\lambda \ge \frac{1}{4}\lambda^2$,
(iii) $\lambda > \frac{\tilde{C}^2}{\tilde{d}}$ implies $\mathcal{V}^\lambda \ge \frac{1}{4}\frac{\frac{\tilde{C}^4}{\tilde{d}}}{\lambda^2}$, and
(iv) $\lambda \le \frac{\tilde{C}^2}{\tilde{d}}$ implies $\mathcal{V}^\lambda \ge \frac{1}{4}\tilde{d}$.

Take any $\lambda \ge 0$. The plan is to construct $\mathcal{L}^\lambda$ out of two lower bounds on $\mathcal{U}^\lambda$ and $\mathcal{V}^\lambda$, respectively, based on which of the above relations are satisfied. In doing so, we ensure that $\mathcal{L}^\lambda \le \mathcal{E}^\lambda$. We will also show that $\frac{1}{4}\mathcal{E}^{\mathrm{B}} \le \min_{\lambda' \ge 0} \mathcal{L}^{\lambda'} \le \mathcal{L}^\lambda$. Since this holds for all $\lambda \ge 0$, we will have that $\frac{1}{4}\mathcal{E}^{\mathrm{B}} \le \inf_{\lambda \ge 0} \mathcal{E}^\lambda = \mathcal{E}^*$.

Now we consider the four cases for λ: If $\lambda > \tilde{B}^2$ (i) and $\lambda > \frac{\tilde{C}^2}{\tilde{d}}$ (iii), we have $\inf_{\lambda'} \mathcal{L}^{\lambda'} = \frac{1}{4}\tilde{B}^2$ with $\lambda' \to \infty$. If $\lambda > \tilde{B}^2$ (i) and $\lambda \le \frac{\tilde{C}^2}{\tilde{d}}$ (iv), $\inf_{\lambda'} \mathcal{L}^{\lambda'} = \frac{1}{4}\tilde{B}^2 + \frac{1}{4}\tilde{d}$. If $\lambda \le \tilde{B}^2$ (ii) and $\lambda > \frac{\tilde{C}^2}{\tilde{d}}$ (iii), $\inf_{\lambda'} \mathcal{L}^{\lambda'} = \frac{1}{4}\frac{2\tilde{C}^2}{\tilde{B}\sqrt{\tilde{d}}}$, with $\lambda' = \frac{\tilde{B}\tilde{C}^2}{\sqrt{\tilde{d}}}$, based on an earlier derivation. Finally, if $\lambda \le \tilde{B}^2$ (ii) and $\lambda \le \frac{\tilde{C}^2}{\tilde{d}}$ (iv), $\inf_{\lambda'} \mathcal{L}^{\lambda'} = \frac{1}{4}\tilde{d}$ with $\lambda' = 0$. □

The regularized regime does not always exist. For example if the dimensionality is relatively small ($\tilde{d} \le \frac{2\tilde{C}^2}{\tilde{B}^2}$), then based on $\mathcal{E}^{\mathrm{B}}$ in (23), the excess risk of the regularized regime ($\frac{2\tilde{C}^2}{\tilde{B}\sqrt{\tilde{d}}}$) will be larger than the geometric average of the excess risks of the other regimes ($\tilde{B}\sqrt{\tilde{d}}$). In this case, we jump directly from the random regime to the unregularized regime.

3.5. Full Estimator

So far, we have analyzed the componentwise estimator. This section offers a partial characterization of the learning curve for the full estimator (11).

Clearly, $\mathcal{E}^* \le \tilde{B}^2$ by taking $\lambda \to \infty$ (the random regime), and $\mathcal{E}^* \le \tilde{d}$ by taking $\lambda = 0$ (the unregularized regime). To analyze the intermediate regime, define the *constrained estimator* to be one which chooses λ so that $\|\hat{\beta}_n^\lambda\| \le 1$, and λ_n denote this λ. Let $\mathcal{E}^{\mathrm{C}}$ be the corresponding asymptotic excess risk and note that the oracle asymptotic excess risk $\mathcal{E}^* \le \mathcal{E}^{\mathrm{C}}$.

Having not yet been able to derive an exact asymptotic form for $\mathcal{E}^{\mathrm{C}}$, we instead offer some speculations based on upper bounds for stochastic optimization (online learning). Let $\hat{\beta}_n^{\mathrm{SGD}}$ be the estimator obtained by running one pass of stochastic gradient descent over the

training data. Then in expectation over the sample we have

$$E_n^{\text{SGD}} \leq 2\frac{B_n^2 + C_n^2}{n} + 2\sqrt{2\frac{(B_n^2 + C_n^2)\sigma_n^2}{n}}.$$

This follows from Srebro et al. (2010), which is based on Theorem 1 of Shalev-Shwartz (2007). While this bound holds only for stochastic gradient descent, we strongly suspect that it also holds for the constrained estimator (the regularized empirical risk minimizer).

Putting together all the pieces yields the following coarse approximate form to the risk:

$$\mathcal{E}^{\text{C}} \approxeq \min\left\{\tilde{B}^2, O\left(\frac{\tilde{C}^2}{\tilde{\sigma}^2} + \tilde{C}\right), \tilde{d}\right\}. \tag{26}$$

We emphasize that (26) is purely speculative. Nevertheless, it can help us understand what might change from the componentwise analysis. Comparing (26) with (23), we see an additional factor of $\frac{\tilde{C}}{\tilde{B}\sqrt{\tilde{d}}}$ that might correspond to the benefit of specializing to a diagonal covariance. We also notice the additional additive term $\frac{\tilde{C}^2}{\tilde{\sigma}^2}$, which behaves as $\frac{1}{n}$ and is the relevant term when $\tilde{\sigma}^2$ is large. This additional additive term gives rise to a fourth regime. While the componentwise estimator has three regimes, (26) suggests the full estimator has four regimes, with two regimes controlled only by the norm, independent of the dimensionality. Further work is necessary to confirm these hypotheses.

4. Discussion

Our broad goal is to obtain an accurate picture of the learning curve. There are a plethora of approaches in the literature that tackle pieces of the curve. Classical parametric asymptotics, a dominant approach in statistics, let the sample size $n \to \infty$ while fixing the problem specification Ψ. Hence, they consider the limit of the learning curve where the excess risk $E_n \xrightarrow{P} 0$. These analyses thus focus on the local fluctuations of estimators around a limiting value. As a result, norm constraints do not enter into the asymptotic risk, even with considering higher-order asymptotics (e.g., Liang et al. (2010)).

On the other hand, finite sample complexity bounds (e.g., Bartlett & Mendelson (2001)) provide statements for any sample size n and problem specification Ψ_n. These focus on controlling structural complexities. Thus, they are well suited for handling norm constraints and typically yield dimension-free results. However, these are only upper bounds and can be far from being tight.

Both analyses provide complementary but incomplete views of the learning curve. In this paper, we used high-dimensional asymptotics to obtain an asymptotically exact analysis also when E_n is away from zero, albeit for simple problems. The key is that as Ψ_n grows with n, the appropriate ratios between sample size and complexity are maintained, while still allowing us reap the benefits of asymptotics, namely concentration. Such ideas have been around in statistics since Kolmogorov's work in the 1960s, and more recently have played an important role in high-dimensional sparse settings (e.g., Wainwright (2009)). Related ideas can also be found in statistical physics approaches for studying learning curves (Haussler et al., 1994).

The particular focus in this paper has been on understanding how multiple problem complexities interact to generate multiple regimes in learning curves. We have so far characterized the regimes for two problems—mean estimation and componentwise linear regression–as a starting point. We hope future work will help shed light on learning curves in more general settings.

References

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. In *Computational Learning Theory*, pp. 224–240, 2001.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, Cambridge, UK, 2006.

Haussler, D., Kearns, M., Seung, H. S., and Tishby, N. Rigorous learning curve bounds from statistical mechanics. In *Computational Learning Theory*, pp. 76–87, 1994.

Liang, P., Bach, F., Bouchard, G., and Jordan, M. I. Asymptotically optimal regularization in smooth parametric models. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2010. MIT Press.

Pollard, D. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

Shalev-Shwartz, S. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

Srebro, N., Sridharan, K., and Tewari, A. Stochastic optimization and online learning with smooth loss functions. Technical report, TTI Chicago, 2010.

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998.

Wainwright, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.