Compressed Sensing and Overparametrized Networks: Overfitting Peaks in a Model of Misparametrized Sparse Regression in the Interpolation Limit

Partha P Mitra
Cold Spring Harbor Laboratory
Cold Spring Harbor
NY, NY 11724
mitra@cshl.edu
November 26, 2019

Abstract

Current practice in machine learning is to employ deep nets in an overparametrized limit, with the nominal number of parameters typically exceeding the number of measurements. This resembles the situation in compressed sensing, or in sparse regression with $l_1$ penalty terms, and provides a theoretical avenue for understanding phenomena that arise in the context of deep nets. One such phenomenon is the success of deep nets in providing good generalization in an interpolating regime with zero training error. Traditional statistical practice calls for regularization or smoothing to prevent "overfitting" (poor generalization performance). However, recent work shows that there exist data interpolation procedures which are statistically consistent and provide good generalization performance[4] ("perfect fitting").

In this context, it has been suggested that "classical" and "modern" regimes for machine learning are separated by a peak in the generalization error ("risk") curve, a phenomenon dubbed "double descent"[3]. While such overfitting peaks do exist and arise from ill-conditioned design matrices, here we challenge the interpretation of the overfitting peak as demarcating the regime where good generalization occurs under overparametrization. We propose a model of Misparametrized Sparse Regression (MiSpaR) and analytically compute the GE curves for $l_2$ and $l_1$ penalties. We show that the overfitting peak arising in the interpolation limit is dissociated from the regime of good generalization. The analytical expressions are obtained in the so called "thermodynamic" limit. We find an additional interesting phenomenon: increasing overparametrization in the fitting model increases sparsity, which should intuitively improve performance of $l_1$ penalized regression. However, at the same time, the relative number of measurements decrease compared to the number of fitting parameters, and eventually overparametrization does lead to poor generalization. Nevertheless, $l_1$ penalized regression can show good generalization performance under conditions of data interpolation even with a large amount of overparametrization. These results provide a theoretical avenue into studying inverse problems in the interpolating regime using overparametrized fitting functions such as deep nets.

1 Introduction

Modern machine learning has two salient characteristics: large numbers of measurements $m$, and non-linear parametric models with very many fitting parameters $p$, with both $m$ and $p$ in the range of $10^6 - 10^9$ for many applications. Fitting data with such large numbers of parameters stands in contrast to the inductive scientific process where models with small numbers of parameters are normative. Nevertheless, these large-parameter models are successful in dealing with real life complexity, raising interesting theoretical questions about the generalization ability of models with large numbers of parameters, particularly in the overparametrized regime $\mu = p/m > 1$.

Classical statistical procedures trade training (TE) and generalization error (GE) by controlling the model complexity. Sending TE to zero (for noisy data) is expected to increase GE[10]. However deep nets seem to over-parametrize and drive TE to zero (data interpolation) while maintaining good GE[18, 5]. Over-parametrization has the benefit that global minima of the empirical loss function
The effects of sparsity and fitting model entropy near the peak (see in particular Figs 4.3 and 5.2 in [7]). Classical statistical folk wisdom where the GE curve is thought to be U-shaped, understanding the significance of such peaks is an open question, and motivates the current paper. Parenthetically, similar over-fitting peaks were reported almost twenty years ago (cf. statistical physics approach to learning) and attributed to increased fitting model entropy near the peak depending on the value of the undersampling parameter $\alpha$. For the simulations with $\lambda \to 0$, in the $l_2$ case a pseudoinverse was used. For the $l_1$ case a numerically small value $\lambda = 10^{-5}$ was used, and it was checked that the results do not change on decreasing $\lambda$.

The proliferation and become easier to find[12, 15]. These observations have led to recent theoretical activity[4, 5, 11]. Regression and classification algorithms have been shown that interpolate data but also generalize optimally[4]. An interesting related phenomenon has been noted: the existence of a peak in GE with increasing fitting model complexity[2, 1, 8, 9]. In [2] it was suggested that this peak separates a classical regime from a modern (interpolating) regime where over-parametrization improves performance. While the presence of a peak in the GE curve is in stark contrast with the classical statistical folk wisdom where the GE curve is thought to be U-shaped, understanding the significance of such peaks is an open question, and motivates the current paper. Parenthetically, similar over-fitting peaks were reported almost twenty years ago (cf. statistical physics approach to learning) and attributed to increased fitting model entropy near the peak (see in particular Figs 4.3 and 5.2 in [7]).

### 1.1 Summary of Results

1. We introduce a model, Misparametrized (or Misspecified) Sparse Regression (MiSpaR), which separates the number of measurements $m$, the number of model parameters $n$ (which can be controlled for sparsity by a parameter $\rho$), and the number of fitting degrees of freedom $p$.  

2. We obtain analytical expressions for the GE and TE curves for $l_2$ penalized regression in the "high-dimensional" or "thermodynamic" asymptotic regime $m, p, n \to \infty$ keeping the ratios $\mu = p/m$ and $\alpha = m/n$ fixed. We are also able to analytically compute GE for $l_1$ penalized regression, and exhibit explicit expressions for $\mu < 1$ and $\mu >> 1$ as $\lambda \to 0$.  

3. We show that for $\lambda \to 0$ and for $\sigma > 0$, the overfitting peak appears at the data interpolation point $\mu = 1$ ($p = m$) for both $l_2$ and $l_1$ penalized interpolation ($GE \sim (1 - \mu)^{-1}$ near $\mu = 1$), but does not demarcate the point at which "good generalization" first occurs, which for small

---

[A similar misspecified model has been studied in [9] with $l_2$ regularization, but this paper did not study the effects of sparsity and $l_1$ penalized regression.]
\( \sigma \) corresponds to the point \( p = n (\mu \alpha = 1) \) (Figure 1). The region of good generalization can start before or after the overfitting peak. The overfitting peak is suppressed for finite \( \lambda \).

4. For infinitely large overparametrization, generalization does not occur: \( GE(\mu \to \infty) = 1 \) for both \( l_2 \) and \( l_1 \) penalized interpolation. However, for small values of the sparsity parameter \( \rho \) and measurement noise variance \( \sigma^2 \), there is a large range of values of \( \mu \) where \( l_1 \) regularized interpolation generalizes well, but \( l_2 \) penalized interpolation does not (Fig. 1).

This range is given by
\[
1 < \log(\mu) < \frac{1}{\sigma^2}, \quad \frac{1}{\rho}, \quad \text{with} \quad \sigma^2, \rho/\alpha < < 1.
\]
In this regime the sparsity penalty is effective, and suppresses noise-driven mis-estimation of parameters for the \( l_1 \) penalty. This shows how generalization properties of penalized interpolation depend strongly on the inductive bias, and are not properties of data interpolation per se. This has important implications for the usage of deep nets for solving inverse problems.

5. For \( \sigma = 0 \) and for \( \mu > 1 \), \( GE(l_2) > 0 \). In contrast, if \( \alpha \) is greater than a critical value \( \alpha_c(\rho) \) that depends on \( \rho \), then for \( l_1 \) penalized interpolation \( GE_1 = 0 \) for a range of overparameterization \( \frac{1}{\alpha} \leq \mu \leq \mu_c \). The maximum overparameterization \( \mu_c \) for which \( GE_1 = 0 \) depends on \( \frac{1}{\alpha} \). For small values of \( \frac{1}{\alpha} \), \( \mu_c \approx \sqrt{\frac{\alpha}{2\rho} e^{\frac{\alpha}{\rho}}} \).

For \( \mu > \mu_c \), \( GE_1(\mu) \) rises quadratically from zero \((GE_2(\mu > \mu_c) \propto (\mu - \mu_c)^2 \) for small \( \mu - \mu_c \) and \( GE_1(\mu \to \infty) = 1 \).

6. For \( \sigma = 0 \) and \( \alpha > \alpha_c(\rho) \), \( GE_1 \) goes to zero linearly at \( \mu \alpha = 1 \) \((GE_1 \propto (\frac{1}{\alpha} - \mu) \) for \( \frac{1}{\alpha} - \mu \) small). When \( \alpha = \alpha_c(\rho) \), \( GE_1 = 0 \) only at the single point \( \mu_c = \frac{1}{\alpha_c(\rho)} \). In this case \( GE_1 \) goes to zero with a nontrivial \( \frac{1}{\alpha} \) power \( GE_1(\mu \leq \frac{1}{\alpha_c(\rho)}) \propto (\frac{1}{\alpha_c(\rho)} - \mu)^2 \) on the left, but rises quadratically on the right \( GE_1(\mu \geq \frac{1}{\alpha_c(\rho)}) \propto (\mu - \frac{1}{\alpha_c(\rho)})^2 \). For \( \alpha < \alpha_c(\rho) \), \( GE_1 > 0 \) for all values of \( \mu \).

2 Model: Misparametrized Sparse Regression

Usually in linear regression the same (known) design matrix \( x_{ij} \) is used both for data generation and for parameter inference. In MiSpaR the generative model has a fixed number \( n \) of parameters \( \beta_j \), which generate \( m \) measurements \( y_i \), but the number of parameters \( p \) in the inference model is allowed to vary freely, with \( p < n \) corresponding the under-parametrized and \( p > n \) the over-parametrized case. For the under-parametrized case, a truncated version of the design matrix is used for inference, whereas for the over-parametrized case, the design matrix is augmented with extra rows.

In addition, we assume that the parameters in the generative model are sparse, and consider the effect of sparsity-inducing regularization in the interpolation limit. Combining misparametrization with sparsity is important to our study for two reasons

- Dissociating data interpolation (which happens when \( \mu = 1, \lambda \to 0 \)) from the regime where good generalization can occur (this is controlled by the undersampling \( \alpha \) as well as by the model sparsity \( \rho \)).

- We are able to study the effect of different regularization procedures on data interpolation in an analytically tractable manner and obtain analytical expressions for the generalization error.

**Generative Model ("Teacher")** We assume that the (known/measured) design variables are i.i.d. Gaussian distributed\(^2\) from one realization of the generative model to another with variance \( 1/n \). This choice of variance is important to fix normalization. Other choices have also been employed in the literature (notably \( x_{ij} \sim N(0,1/m) \)) - this is important to keep in mind when comparing with literature formulae where factors of \( \alpha \) may need to be inserted appropriately to

\(^2\)Note that these choices are convenient, but could be relaxed. It is only required that \( x_{ij} \) are i.i.d. and have finite second moment. The asymptotic results only depend on \( V(x_{ij}) \)
obtain a match.

\[ y_i = \sum_{j=1}^{n} x_{ij} \beta_j + n_i \]

\[ n_i \sim N(0, \sigma) \quad \beta_j \sim (1 - \rho) \delta_{\beta_j} + \rho \pi(\beta) \]

\[ \pi(\beta) \sim N(0, 1) \] unless otherwise specified

\[ x_{ij} \sim N \left( 0, \frac{1}{n} \right) \quad i = 1 \ldots m \text{ measurements} \quad j = 1 \ldots n \text{ generative parameters} \]

Undersampling: \( \alpha = m/n \)  Sparsity: \( \rho \)  Overparametrization: \( \mu = p/m \)

Here \( \pi(\beta) \) is the distribution of the non-zero model parameters. We assume this distribution to be Gaussian as this permits closed form evaluation of integrals appearing in the \( l_1 \) case. Note that we term \( \mu = p/m \) as overparametrization (referring to the case where \( \mu > 1 \)) and we term \( \alpha = m/n \) as undersampling (referring to the case where \( \alpha < 1 \)).

**Inference Model ("Student")** The design matrix used for inference is mis-parametrized or mis-specified: under-specified (or partially observed) when \( \mu \alpha < 1 \equiv p < n \); over-specified, with extra, effect-free rows in the design matrix when \( \mu \alpha > 1 \equiv p > n \)

\[ x_{ij}^{inf} = x_{ij}, \quad j = 1 \ldots p \quad \text{if } p \leq n \]

\[ x_{ij}^{inf} = x_{ij}^{extra}, \quad j = n + 1 \ldots p \quad \text{if } p > n \]

Parameter inference is carried out by minimizing a penalized mean squared error

\[ \hat{\beta} = \arg \min_{\beta} \frac{1}{2} |Y - X^{inf} \beta|^2 + \lambda V(\beta) \]

Note that for \( p > n \), the model parameters \( \beta \) are augmented by \( p - n \) zero entries. We consider \( l_2 \) and \( l_1 \) penalties (correspondingly \( V(\beta) = \frac{1}{2} |\beta|^2 \) or \( |\beta|_1 \)) and the interpolation limit is obtained by taking \( \lambda \to 0 \). For the \( l_2 \) penalty (ridge regression), \( \hat{\beta}_2 = (X^+ X + \lambda)^{-1} X^+ Y \). The training and generalization errors are defined as the expected values of the normalized MSEs on training and test sets,

\[ TE = \frac{E[|Y - X^{inf} \hat{\beta}|^2]}{E[|Y|^2]} \quad \text{and} \quad GE = \frac{E[|Y - X^{extra} \hat{\beta}|^2]}{E[|Y|^2]} \] Note that the expectation \( E \) is taken simultaneously over the parameter values, the design matrix and measurement noise.

We obtain exact analytical expressions for the risk (generalization error) in the (thermodynamic) limit where \( n, p, m \) all tend to infinity, but the ratios \( \alpha = m/n, \mu = p/m \) are held finite. Similar "thermodynamic" or "high-dimensional" limiting procedures are used in statistical physics, eg in the study of random matrices and spin-glass models in large spatial dimensions[14, 13]. Such limits are also well-studied in modern statistics[17] (for example to understand phase-transition phenomena in the LASSO algorithm[6]). While there is a large literature on the LASSO phase transition, we were unable to find any computations of the GE curves that span across the underparametrized and overparametrized regimes in a systematic model as presented here.

We derive analytical formulae for TE and GE with \( l_2 \) or ridge regularization. For \( l_1 \) regularization, explicit formulae are given in some parameter regimes. More generally for the \( l_1 \) case we obtain a pair of simultaneous nonlinear equations in two variables which implicitly define the MSE. These can be solved numerically to obtain the GE. The nonlinear equations are given in closed form without hidden parameters and do not require integration.

**Analytical Formulae:** \( TE_2, GE_2 \) are the training and generalization errors for the \( l_2 \) penalized case, and \( GE_1 \) the generalization error for the \( l_1 \) penalized case. Due to lack of space we do not present the analytical formulae for \( \lambda > 0 \) as these expressions are complex, but the corresponding analytical expressions were used to generated the theory curves in Fig.1 for the case \( \lambda > 0 \). The derivations employ the cavity mean field theory approach [16]. Here \( \sigma^2_{\beta} = \sigma^2 + \rho (1 - \mu \alpha) \).

Note that the formulae for GE agree for the \( l_2 \) and \( h \) cases in the underparametrized regime \( \mu < 1 \), but diverge in the overparametrized regime: infinitesimal \( l_1 \) regularization provides no better generalization than the pseudoinverse based procedure unless there is overparametrization. Further note that "good generalization" (GE small) begins when \( \mu \alpha > 1 \) not at the overfitting peak (\( \mu = 1 \)).
\[ T_{E_2}(\sigma > 0) \begin{cases} \frac{\sigma^2_{\mathcal{E}_2}(1-\mu)}{\sigma^2 + \mu} \quad & \text{if } \mu \geq 1 \\ \frac{\sigma^2_{\mathcal{E}_2}}{\sigma^2 + \mu} \quad & \text{if } \mu < 1 \end{cases} \]

\[ G_{E_2}(\sigma = 0) \begin{cases} \frac{1 - \mu}{1 - \mu} \quad & \text{if } \sigma \geq 0 \\ \frac{1}{\sigma^2 + \mu} \quad & \text{if } \sigma < 0 \end{cases} \]

Condition (i): \[ 1 < \log(\mu) \ll \min\left(\frac{1}{\sigma^2}, \frac{2}{\rho}\right) \]

Condition (ii): \[ [\alpha > \alpha_c(\rho) \wedge [\mu < \mu_c(\alpha/\rho)] \]

Acknowledgement

This work was supported by the Crick-Clay Professorship (CSHL) and the H N Mahabala Chair Professorship (IIT Madras). Help from Dr Jaikishan Jaikumar in preparing the figures shown in the paper is gratefully acknowledged.

References


