Consider the following linear problem: we have a set of $P$ sparse $N$-dimensional Gauss-Bernoulli signals, created from the distribution $P(x_i^n) = (1 - \rho)\delta(x_i^n) + \rho \mathcal{P}(x_i^n)$, where $i = 1, \ldots, N$, $l = 1, \ldots, P$. For each signal we perform $M \times N$ measurement, summarized by a $M \times N$ measurement matrix with iid elements $F_{il}$ generated from a Gaussian ensemble with zero mean and unit variance. We have only access to the (noisy) results of these measures, that is, to $x_i^n = \sum_{l=1}^{P} F_{il} x_l + \xi_{il}$, where $\xi_{il}$ is a Gaussian additive noise with variance $\Delta$. It is possible to find both the vectors $x^n$ and the matrix (dictionary) $F^0$ (up to a permutation of $N$ elements)? This is the dictionary learning problem. A related situation is when one knows at least a noisy version of the matrix $F^0$, defined by $F^0 = (F^0 + \sqrt{\eta} X)/(1 + \eta)$ where $X$ is a random matrix with the same statistics as $F^0$. We denote by $P(F^0, F')$ the joint probability of $F^0$ and $F'$. Recovering $F^0$ and the signals, knowing this time $F'$ and the $P$ vectors $\bar{y}_n$ is the blind calibration problem, which is equivalent to dictionary learning when $\eta \to \infty$.

We define $\alpha = M/N, \pi = P/N$. We denote by $D$ the MSE (mean squared error) on the dictionary and by $E$ the MSE on the signal, and we ask: for which values of the parameters $\alpha, \rho, \pi$ is there enough information to determine $F^0$ and the $\bar{y}_n$ from the knowledge of the $\bar{y}_n$ in dictionary learning? How does this change for finite value of $\eta$ when one knows $F^0$? While rigorous results were able to show learnability only from exponential many samples [1], recent work suggested that $P = O(N)$ samples should be sufficient [2].

We answer this question using a heuristic approach from statistical physics called the replica method [3]. We estimate the Bayes optimal MMSE (minimal mean square error) on the dictionary and by $E$ the MSE on the signal, and we ask: for which values of the parameters $\alpha, \rho, \pi$ is there enough information to determine $F^0$ and the $\bar{y}_n$ in dictionary learning? How does this change for finite value of $\eta$ when one knows $F^0$? While rigorous results were able to show learnability only from exponential many samples [1], recent work suggested that $P = O(N)$ samples should be sufficient [2].

We estimate the Bayes optimal MMSE (minimal mean square error) by computing the mutual information of the matrix and signals elements. Our computation is different from the replica analysis of dictionary learning by [2], who did not compute the Bayes optimal MMSE, but minimized $\bar{y} - F\bar{x}, \bar{z}$, and assumed the so-called replica symmetry hypothesis that is incorrect in that case, but correct in our case. Our analysis (see [6] for a similar computation in compressed sensing) indicates that in the limit of large signals, when $N \to \infty$, the MMSE is given by $\Phi(E, D)$, so that the “potential” $\Phi(E, D)$ given by

\[
\Phi(E, D) = -\frac{\alpha}{\Delta} \log (\Delta + E + D(\rho - E))^{-\Delta} + \frac{\alpha}{\pi} \left( \int D\bar{z} \log \left( \frac{\mathcal{P}(\bar{z})}{\mathcal{P}(\bar{z}^0)} \right) \mathcal{P}(\bar{z}) \right)
\]

where $f(x)$ denotes an average of a function $f$ of the random variable $x$ with density $Q(x)$, $\Delta = \Delta(E, D)$, $\bar{m}_x = \frac{1}{N} \sum_{i=1}^{N} F_{il}x_i$, $\bar{m}_F = \frac{\Delta(E, D)}{\Delta(E, D) + 2|\Delta|}$, and the conditional probabilities are computed according to the Bayes rule.

The analysis of this expression in the zero-noise case ($\Delta = 0$) allows to demonstrate our first main result: in both the blind calibration and dictionary learning problems, the global maximum is given by $D = E = 0$ as long as $\alpha > \rho$ and $\pi > \pi^* = \alpha/(\alpha - \rho)$. In this regime, it is thus possible to recover the matrix and the signal exactly if one can compute the Bayes-optimal values $F, \bar{x}$.

Bayes-optimal learning, however, remains an extremely hard computational problem. Message-passing based sampling algorithms like those of e.g. [3], [4] attempt to perform steepest ascent in the function $\Phi(E, D)$. Whenever there exist secondary maxima of this function, the sampling algorithms will be trapped for exponentially long time. We will call $\pi^*(\eta)$ (the spinodal transition) the value above which $\Phi(E, D)$ does not have secondary maxima. Examples of this transition are shown in Fig. 1. Generalization of the AMP approach [4] designed in compressed sensing was successfully applied to the dictionary learning problem in the region $\pi > \pi^*(\eta)$ [5], [6].

To conclude, there exist three regions in the phase diagram for blind calibration and dictionary learning, corresponding to impossible, intractable, and perhaps-tractable learning. The potential $\phi(E, D)$ is very general and can be used to study many similar problems, such as matrix completion, or sparse matrix decomposition.

REFERENCES