# Mean-field theory for certain large convex optimization problems

1. **Simplified derivations for high-dimensional convex learning problems associated with statistical learning**
   Authors: David G. Clark, and Haim Sompolinsky
   arXiv:2412:01110

2. **Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models**
   Authors: Jason W. Rocks, and Pankaj Mehta
   Phys. Rev. Research 4, 013201 (2022)

*Recommended with a Commentary by Anirvan M. Sengupta, Rutgers University and Flatiron Institute*

Several interesting convex optimization problems, coming from statistical learning, could be phrased as min-max problems over a bipartite graphs. These are the graphs in which the vertices could be divided into two sets (sometimes called the left set and the right set), such that edges only connect vertices from distinct sets. When there are a large number of variables (and constraints) in that convex problem, in other words the numbers of vertices in the left set and the right set are both large, one could apply a mean-field theory technique called cavity method, originally invented for spin glasses on typically non-bipartite graphs.

Recently, Clark and Sompolinsky [1] discussed three such problems: The Gardner perceptron capacity problem, the manifold capacity problem, and kernel ridge regression. It is a useful pedagogical paper. A predecessor of such two step cavity method, without explicitly mentioning the bipartite structure, is the work of Shamir and Sompolinsky [5]. Unfortunately, the manuscript does not cite a long series of other papers setting up cavity method for bipartite graphs from the intervening years. See, for example, the recent paper by Rocks and Mehta [4], my second recommendation, taking up similar problems and applying similar methods.

To give a flavor of the ideas, I will pick one of the simplest problem of this kind is ridge regression with (quenched) random features $\{\xi_{\mu i}\}_{\{\mu=1,\ldots,P;\, i=1,\ldots,N\}}$ with

$$\langle \xi_{\mu i} \rangle = 0, \langle \xi_{\mu i} \xi_{\nu j} \rangle = \frac{1}{N} \delta_{\mu\nu} \delta_{ij}.$$

This is essentially the third problem we just discussed by Clark and Sompolinsky, with $\phi_i(x_\mu) \equiv \xi_{\mu i}$. One major difference is the scaling of $\xi_{\mu i}$ covariances as $\frac{1}{N}$. We want to keep the regression weights $O(1)$ and not $O(1/\sqrt{N})$. The presentation follows [3].

Let us say we have some observations $\{y_\mu\}_{i=1}^P$ generated by a teacher model

$$y_\mu = \sum_{i=1}^N \xi_{\mu i} a_i. \tag{1}$$

A student model is trying to learn the parameters $\{a_i\}_{i=1}^N$. If there is enough data ($P \geq N$), and the $P \times N$ matrix $\Xi$, with entries $\xi_{\mu i}$, has rank $N$, then just minimizing the square loss $\sum_{\mu=1}^P (y_\mu - \sum_{i=1}^N \xi_{\mu i} w_i)^2$ will do. However, we may have an overparametrized regression problem: $P << N$. In that case, the square loss may be augmented by quadratic penalties for the weights to find a unique set of weights $\{w_i\}_{i=1}^N$.

The resulting optimization problem looks like:

$$\min_w \big[\frac{1}{2}\sum_{\mu=1}^P (y_\mu - \sum_{i=1}^N \xi_{\mu i} w_i)^2 + \frac{\gamma}{2}\sum_{i=1}^N w_i^2/\rho_i\big]. \tag{2}$$

This is a convex problem, with $\rho_i > 0$ for all $i$. The penalty terms want $w_i = 0$ for all $i$. The loss function would be satisfied if $w_i = a_i$. We want to know how much mistake we get in each weight, measured by $\Delta_i = a_i - w_i$. We want to know how big these errors are.

Using the expression of $y_\mu$ from Eq. 1 and using it in Eq. 2, we get the optimization problem for $\{\Delta_i\}$:

$$\min_\Delta \big[\frac{1}{2}\sum_{\mu=1}^P (\sum_{i=1}^N \xi_{\mu i}\Delta_i)^2 + \frac{\gamma}{2}\sum_{i=1}^N (a_i - \Delta_i)^2/\rho_i\big]. \tag{3}$$

Using the Hubbard-Stratonovich trick and introducing auxilliary variables $\{\lambda_\mu\}_{\mu=1}^P$, the previous optimization problem can be rewritten as

$$\min_\Delta \max_\lambda \big[ -\frac{1}{2}\sum_{\mu=1}^P \lambda_\mu^2 + \sum_{\mu=1}^P\sum_{i=1}^N \lambda_\mu \xi_{\mu i}\Delta_i + \frac{\gamma}{2}\sum_{i=1}^N (\Delta_i - a_i)^2/\rho_i\big].$$

One can now imagine a bipartite graph with $P$ vertices on the left side, connected to $N$ vertices on the right side. The $\lambda_\mu$ variable lives on the $\mu$-th left vertex and the $\Delta_i$ variable live on the $i$-th right vertex, and these two variables are coupled by the quenched random variable $\xi_{\mu i} \sim O(1/\sqrt{N})$ associated with the $(\mu i)$ edge. The random $\xi_{\mu i}$ is analogous to $J_{ij}$ in disordered spin models. In addition to such interaction terms, there are one-body terms, like local field energies in spin models. We can imagine as if there are two types of continuous spins: $\Delta_i$'s and $\lambda_\mu$'s. Of course, a key difference with spin model energy minimization is that we are maximizing in $\lambda_\mu$'s while minimizing in $\Delta_i$'s.

The idea in the cavity method to approximate this optimization problem by another with a collection self-consistent effective one-body terms

$$\min_\Delta \max_\lambda \big[\sum_\mu \phi_\mu(\lambda_\mu) + \sum_i \psi_i(\Delta_i)\big].$$

Now let us pay attention to a single $\lambda_\mu$ variable. What is a effective one site objective function of $\lambda_\mu$? Looks like we have

$$\phi_\mu(\lambda_\mu) \stackrel{?}{=} -\frac{1}{2}\lambda_\mu^2 + \Big(\sum_{i=1}^N \xi_{\mu i}\Delta_i\Big)\lambda_\mu := -\frac{1}{2}\lambda_\mu^2 + h_\lambda \lambda_\mu$$

where, for large $N$, we could replaced the 'effective $\lambda$ field' $\left(\sum_{i=1}^{N} \xi_{\mu i}\Delta_i\right)$ by a quenched random variable $h_\lambda \sim \mathcal{N}(0, \frac{1}{N}\sum_i \Delta_i^2)$. This is the naive mean-field theory.

The key thing to improve this naive mean-field theory is to incorporate the contribution of the analogue of Onsager reaction terms. These terms are very important for cavity method and Thouless-Anderson-Palmer equations in spin glass theory [2]. If a $\mu$-vertex with a $\lambda_\mu$ value is added, it changes each $\Delta_i$ a bit from the solution where we do not have that vertex (the $\mu$-cavity). The net feedback of these changes of many $\Delta_i$'s on $\lambda_\mu$ is mimicked by an additional term in $\phi_\mu(\lambda_\mu)$. Similar considerations applies to $\psi_i(\Delta_i)$ as well. One has to then get the distribution $\lambda_\mu$'s and the distribution of $\Delta_i$'s to satisfy a set of self-consistency conditions. The ultimate goal is to calculate the average loss for the model with $\{w_i\}$ on a new input-output pair $((\xi_{1\,(N+1)}, \ldots, \xi_{P\,(N+1)}, y_{N+1})$, telling us how well the regression model generalizes.

A few brief words on the Gardner capacity problem and on the manifold capacity one. These two problems have to do with classification as opposed to regression. The first problem starts with $P$ random $N$ dimensional points with random $\pm 1$ labels on each point. We are asked to find a hyperplane through the origin which separates $+1$ labeled points from the $-1$ labeled ones. As $\alpha = P/N$ increases, there is a capacity, an $\alpha_c$, such that, for $\alpha > \alpha_c$, the probability of finding such that a plane goes to zero, when $N, P = \alpha N$ both go to infinity. This was originally a problem posed and solved by Cover. Gardner sets up the problem where the separating of is done with a minimum margin: the points has to be at least distance $\kappa$ away from the hyperplane. This is a more demanding constraint, resulting in a $\kappa$-dependent $\alpha_c$. In the manifold capacity problem, the points are replaced by extended shapes with non-zero widths, making the problem more complex.

In summary, many convex problems related to statistical learning have a set of $N$ weights and a set of $P$ 'constraints' (low error in prediction or correct classification). This structure leads to a bipartite graph. When $N, P$ are large, one can gain many insights by using mean field theory on such problems. The cavity method avoids some mysteries associated with the replica trick. It can even be implemented as an approximate numerical approach for concrete realizations of the problem.

# References


[1] David G Clark and Haim Sompolinsky. Simplified derivations for high-dimensional convex learning problems. *arXiv preprint arXiv:2412.01110*, 2024.

[2] Konrad H Fischer and John A Hertz. *Spin glasses*. Number 1. Cambridge university press, 1993.

[3] Mohammad Ramezanali, Partha P Mitra, and Anirvan M Sengupta. The cavity method for phase transitions in sparse reconstruction algorithms. *arXiv preprint arXiv:1501.03194*, 2015.

[4] Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical review research*, 4(1):013201, 2022.

[5] Maoz Shamir and Haim Sompolinsky. Thouless-anderson-palmer equations for neural networks. *Physical Review E*, 61(2):1839, 2000.