Replica Method for the Machine Learning Theorist

Blake Bordelon, Haozhe Shan, Abdul Canatar, Boaz Barak, Cengiz Pehlevan *[Boaz's note: Blake and Haozhe were students in the ML theory seminar; in that seminar we touched on the replica method in the lecture on inference and statistical physics but here Blake and Haozhe (with a little help from the rest of us) give a great overview of the method and its relations to ML. See also all seminar posts.]*

I. Analysis of Optimization Problems with Statistical Physics

In computer science and machine learning, we are often interested in solving optimization problems of the form

$$\min_{x \in \mathcal{S}} H(x, \mathcal{D})$$

where H is an objective function which depends on our decision variables $x \in \mathcal{S}$ as well as on a set of problem-specific parameters $\mathcal{D}$. Frequently, we encounter problems relevant to machine learning, where $\mathcal{D}$ is a random variable. The replica method is a useful tool to analyze *large problems* and their *typical* behavior over the distribution of $\mathcal{D}$.

Here are a few examples of problems that fit this form: 1. In supervised learning, H may be a training loss, x a set of neural network weights and $\mathcal{D}$ the data points and their labels 2. We may want to find the most efficient way to visit all nodes on a graph. In this case $\mathcal{D}$ describes nodes and edges of the graph, x is a representation of the set of chosen edges, and H can be the cost of x if x encodes a valid path and ∞ (or a very large number) if it doesn't encode a valid path. 3. Satisfiability: $x \in \{0,1\}^N$ is a collection booleans which must satisfy a collection of constraints. In this case the logical constraints (clauses) are the parameters $\mathcal{D}$. $H(x)$ can be the number of constraints violated by x. 4. Recovery of structure in noisy data: x is our guess of the structure and $\mathcal{D}$ are instances of the observed noisy data. For example PCA attempts to identify the directions of maximal variation in the data. With the replica method, we could ask how the accuracy of the estimated top eigenvector degrades with noise.

II. The Goal of the Replica Method

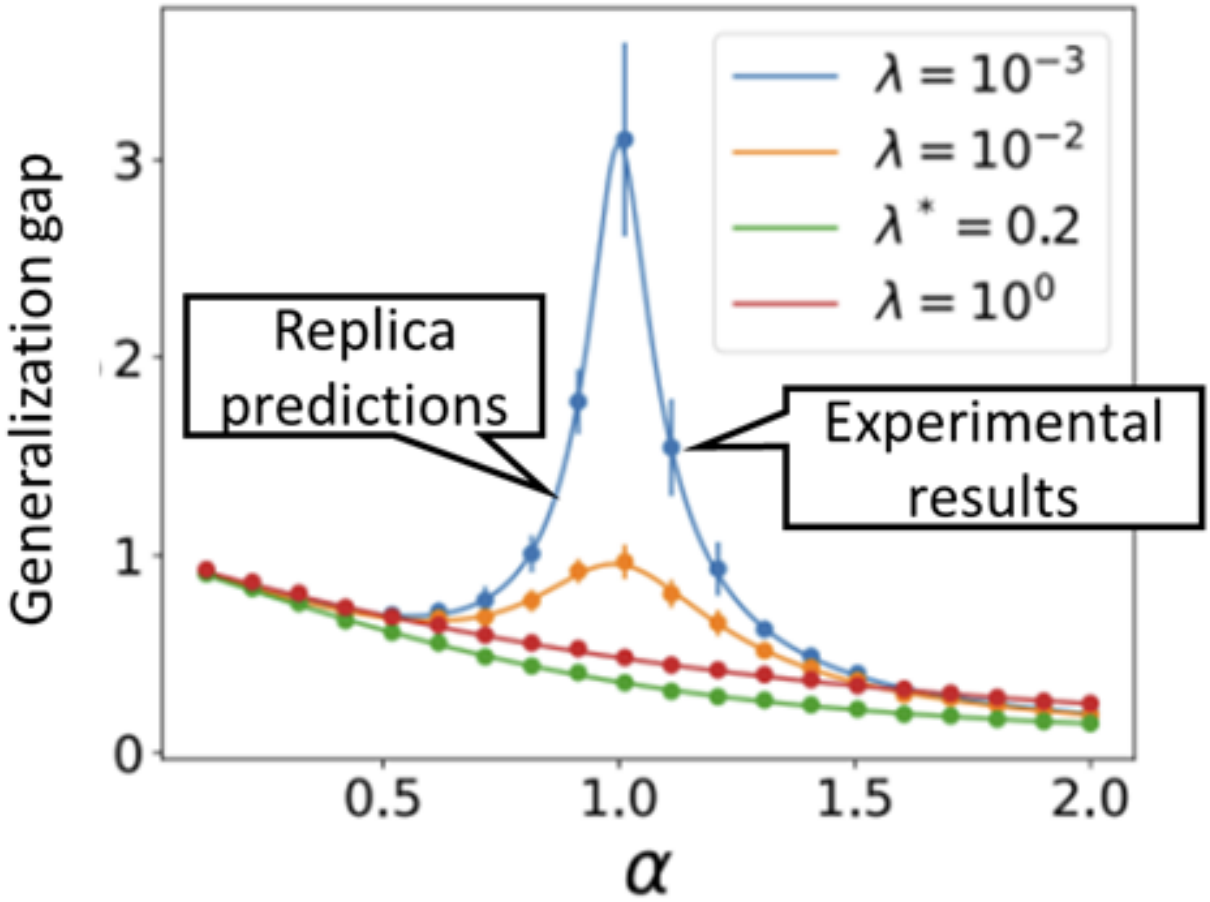
The **replica method** is a way to calculate the value of some statistic (**observable** in physics-speak) $O(x)$ of x where x is a "typical" minimizer of $H(x, \mathcal{D})$ and $\mathcal{D}$ is a "typical" value for the parameters (which are also known in physics-speak as the **disorder**).

In Example 1 (supervised learning), the observable may be the generalization error of a chosen algorithm (e.g. a linear classifier) on a given dataset. For Example 2 (path), this could be the cost of the best path. For Example 3 (satisfiability), the observable might be whether or not a solution exists at all for the problem. In Example 4 (noisy data), the observable might be the quality of decoded data (distance from ground truth under some measure).

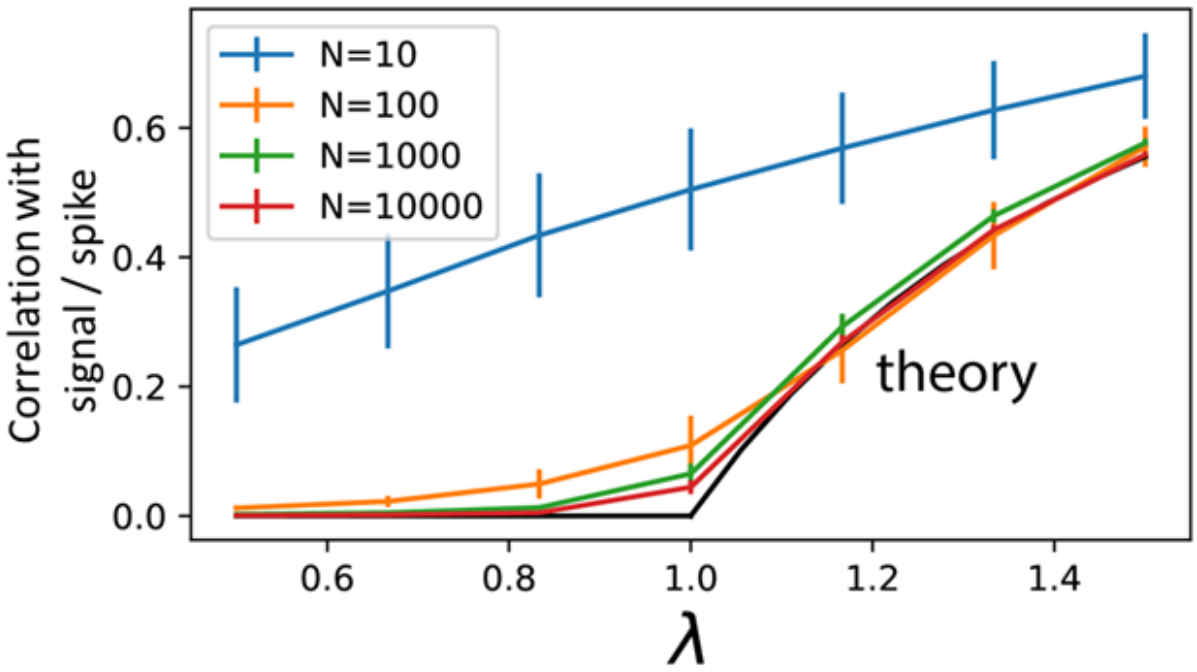
An observable like generalization error obviously depends on $\mathcal{D}$, the problem instance. However, can we say something more general about this *type of problem*? In particular, if $\mathcal{D}$ obeys some probability distribution, is it possible to characterize the the typical observable over different problem instances $\mathcal{D}$?

For instance, in Example 1, we can draw all of our training data from a distribution. For each random sample of data points $\mathcal{D}$, we find the set of x which minimize $H(x, \mathcal{D})$ and compute a generalization error. We then repeat this procedure many times and average the results. Sometimes, there are multiple x that minimize H for a given sample of $\mathcal{D}$; this requires averaging the observable over all global minima for each $\mathcal{D}$ first, before averaging over different $\mathcal{D}$.

To give away a "spolier", towards the end of this note, we will see how to use the replica method to give accurate predictions of performance for noisy least square fitting and spiked matrix recovery.



Generalization gap in least squares ridge regression, figure taken from Canatar, Bordelon, and Pehlevan



Performance (agreement with planted signal) as function of signal strength for spiked matrix recovery, as the dimension grows, the experiment has stronger agreement with theory. See also Song Mei's exposition

A. What do we actually do?

Now that we are motivated, let's see what quantities the replica method attempts to obtain. In general, given some observable $O(x, \mathcal{D})$, the average of O over a minimizer x chosen at random from the set $\arg\min H(x, \mathcal{D})$, and take the average of this quantity over the choice of the disorder $\mathcal{D}$. In other words, we want to compute the following quantity:

$$\text{Desired quantity } = \mathbb{E}_{\mathcal{D}} \mathbb{E}_{x \in \arg\min H(x,\mathcal{D})} [O(x, \mathcal{D})]$$

The above equation has two types of expectation- over the disorder D, and over the minimizers x. The physics convention is to * use $\langle f \rangle_{\mathcal{D}}$ for the expectation of a function $f(\mathcal{D})$ over the disorder $\mathcal{D}$ * use $\int g(x)\mu(x)dx$ for the expectation of a function $g(x)$ over x chosen according to some measure μ.

Using this notation, we can write the above as

$$\text{Desired quantity } = \left\langle \int p^*(x; \mathcal{D}) O(x, \mathcal{D}) dx \right\rangle_{\mathcal{D}}$$

where $\langle \rangle_{\mathcal{D}}$ denotes an average over the probability measure for problem parameters $\mathcal{D}$, and $p^*(x; \mathcal{D})$ is a uniform distribution over the set of x that minimize $H(x, \mathcal{D})$ with zero probability mass placed on sub-optimal points.

The ultimate goal of the replica method is to express

$$\text{Desired quantity } = \text{solution of optimization on constant number of variables}$$

but it will take some time to get there.

B. The concept of "self-averaging" and concentration

Above, we glossed over an important distinction between the "typical" value of $f(\mathcal{D}) = \int p^*(x;\mathcal{D})O(x,\mathcal{D})dx$ and the *average* value of $f(\mathcal{D})$. This is OK only when we have *concentration* in the sense that with high probability over the choice of $\mathcal{D}$, $f(\mathcal{D})$ is close to its expected value. We define this as the property that with probability at least $1-\epsilon$, the quantity $f(\mathcal{D})$ is within a $1 \pm \epsilon$ multiplicative factor of its expectation, whwere ϵ is some quantity that goes to zero as the system size grows. A quantity $f(\cdot)$ that concentrates in this sense is called **self averaging**.

For example, suppose that $X = \sum_{i=1}^n X_i$ where each X_i equals 1 with probabilty $1/2$ and 0 with probability $1/2$ independently. Standard Chernoff bounds show that with high probability $X \in [n/2 \pm O(\sqrt{n})]$ or $\frac{X}{\mathbb{E}X} \in \left(1 + O(\frac{1}{\sqrt{n}})\right)$. Hence X is a self averaging quantity.

In contrast the random variable $Y = 2^X$ is not self averaging. Since $Y = \prod_{i=1}^n 2^{X_i}$ and these random variables are independent, we know that $\mathbb{E}Y = \prod_{i=1}^n \mathbb{E}2^{X_i} = \left(\frac{1}{2}2^1 + \frac{1}{2}2^0\right)^n = (3/2)^n$. However, with high probability a typical value of Y will be of the form $2^{n/2 \pm O(\sqrt{n})} = \sqrt{2}^{n \pm O(\sqrt{n})}$. Since $\sqrt{2} < 3/2$ we see that the typical value of Y is exponentially smaller than the expected value of Y.

The example above is part of a more general pattern. Often even if a variable Y is not self averaging, the variable $X = \log Y$ will be self-averaging. Hence if we are interested in the typical value of Y, the quantity $\exp\left(\mathbb{E}[\log Y]\right)$ is more representative than the quantity $\mathbb{E}[Y]$.

C. When is using the replica replica method a good idea?

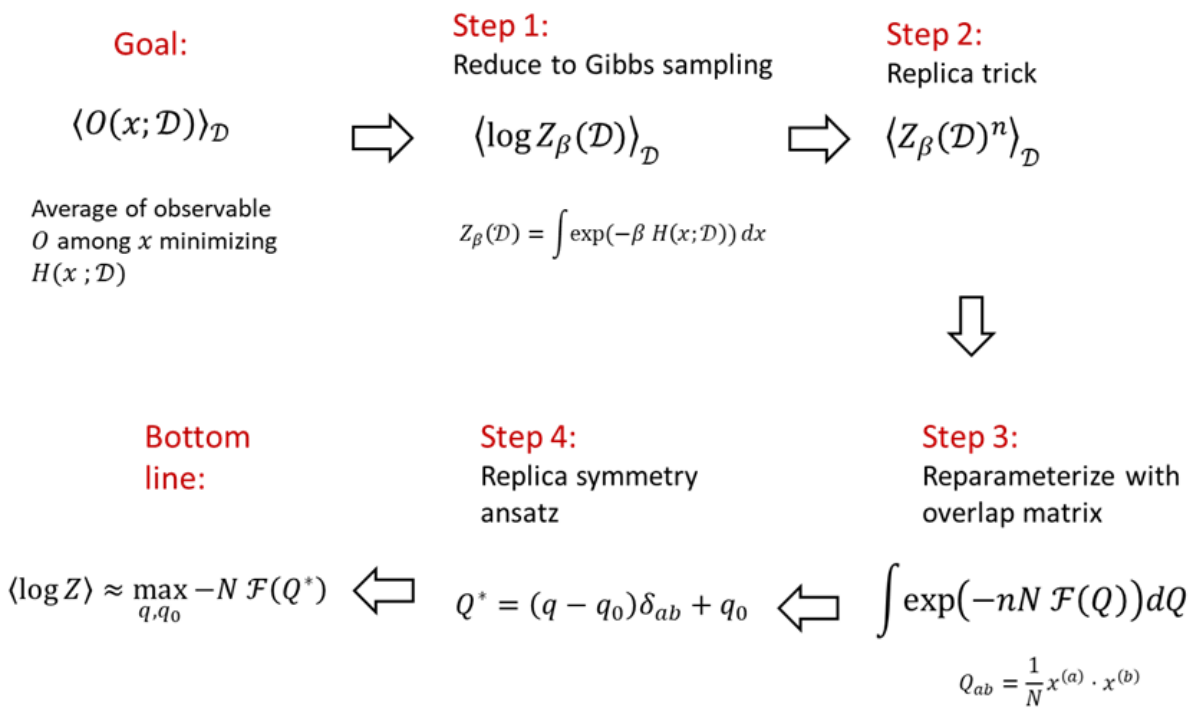
Suppose that we want to compute a quantity of the form above. When is it a good idea to use the replica method to do so? Generally, we would want it to satisfy the following conditions:

1. The learning problem is high dimensional with a large budget of data. The replica method describes a *thermodynamic limit* where the system size and data budget are taken to infinity with some fixed ratio between the two quantities. Such a limit is obviously never achieved in reality, but in practice sufficiently large learning problems can be accurately modeled by the method.
2. The loss or the constraints are convenient functions of x and $\mathcal{D}$. Typically H will be a low degree polynomial or a sum of local functions (each depending on small number of variables) in x.
3. Averages over the disorder in $\mathcal{D}$ are tractable analytically. That is, we can compute marginals of the distribution over $\mathcal{D}$.
4. The statistic that we are interested in is self-averaging.

If the above conditions aren't met, it is unlikely that this problem will gain much analytical insight from the replica method.

III. The Main Conceptual Steps Behind Replica Calculations

We now describe the conceptual steps that are involved in calculating a quantity using the replica method. They are also outlined in this figure:



Step 1:"Softening" Constraints with the Gibbs Measure

The uniform measure on minimizers $p^*(x;\mathcal{D})$ is often difficult to work with. To aid progress, we can think of it as a special case of what is known as the Gibbs measure, defined as $p_\beta(x;\mathcal{D})dx = \frac{1}{Z(\mathcal{D})} \exp\left(-\beta H(x,\mathcal{D})\right) dx$

where $Z(\mathcal{D}) = \int \exp\left(-\beta H(x,\mathcal{D})\right) dx$ is the normalization factor, or **partition function**. β is called the **inverse temperature**, a name from thermodynamics. It is easy to see that when $\beta \to \infty$ (i.e., when the temperature tends to the absolute zero), the Gibbs measure converges to a uniform distribution on the minimizers of H: $p_\beta \to p^*$.

Hence we can write

Desired quantity $= \langle \int p^*(x;\mathcal{D}) O(x,\mathcal{D}) dx \rangle_{\mathcal{D}} = \langle \lim_{\beta\to\infty} \int p_\beta(x;\mathcal{D}) O(x,\mathcal{D}) dx \rangle_{\mathcal{D}}$

Physicists often exchange the order of limits and expectations at will, which generally makes sense in this setting, and so assume

Desired quantity $= \lim_{\beta\to\infty} \langle \int p_\beta(x;\mathcal{D}) O(x,\mathcal{D}) dx \rangle_{\mathcal{D}}$

Thus general approach taken in the replica method is to derive an expression for the average observable for any β and then take the $\beta \to \infty$ limit. The quantity $\int p_\beta(x;\mathcal{D}) O(x,\mathcal{D}) dx$ is also known as the *thermal average* of O, since it is taken with respect to the Gibbs distribution at some positive temperature.

To compute the thermal average of O, we define the following *augmented partition function*:

$Z(\mathcal{D}, J) = \int_S \exp\left(-\beta H(x,\mathcal{D}) + JO(x;\mathcal{D})\right) dx.$

One can then check that

$\frac{d}{dJ} \log Z(\mathcal{D}, J)\Big|_{J=0} = \frac{1}{Z} \int O(x,\mathcal{D}) \exp(-\beta H(x,\mathcal{D})) dx = \int p_\beta(x;\mathcal{D}) O(x,\mathcal{D}) dx$

Hence our desired quantity can be obtained as

Desired quantity $= \lim_{\beta\to\infty} \frac{\partial}{\partial J} \langle \log Z(\mathcal{D}, J)\rangle_{\mathcal{D}}(0)$

or (assuming we can again exchange limits at will):

Desired quantity $= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left[\lim_{\beta\to\infty} \langle \log Z(\mathcal{D}, \epsilon)\rangle_{\mathcal{D}} - \lim_{\beta\to\infty} \langle \log Z(\mathcal{D}, 0)\rangle_{\mathcal{D}}\right]$

We see that ultimately computing the desired quantity reduces to computing quantities of the form

$\langle \log Z'(\mathcal{D})\rangle_{\mathcal{D}}$ ( )

for the original or modified partition function Z'. Hence our focus from now on will be on computing ( ). Averaging over $\mathcal{D}$ is known as "configurational average" or **quenched average**. All together, we obtain the observable, first thermal averaged to get $O^*(\mathcal{D})$ (averaged over $p^*(x;\mathcal{D})$) and then quenched averaged over $\mathcal{D}$.

> *What Concentrates?*: It is not just an algebraic convenience to average $\log Z$ instead of averaging Z itself. When the system size N is large, $\frac{1}{N}\log Z$ concentrates around its average. Thus, the typical behavior of the system can be understood by studying the quenched average $\frac{1}{N}\langle \log Z\rangle$ The partition function Z itself often does not concentrate and in general the values $\frac{1}{N}\log\langle Z\rangle$ (known as the "annealed average") and $\frac{1}{N}\langle \log Z\rangle$ (known as the "quenched average") could differ subtantially. For more information, please consult Mezard and Montanari's excellent book, Chapter 5.

Step 2: The Replica Trick

Hereafter, we use $\langle\cdot\rangle$ to denote the average and drop the dependence of $\mathcal{D}$ and J. To compute $\langle \log Z(\mathcal{D}, J)\rangle_{\mathcal{D}}$, we use an identity

$$\langle \log Z\rangle = \lim_{n\to 0} \frac{1}{n} \log \langle Z^n\rangle .$$

For the limit to make sense, n should be any real number. However, the expression for Z^n is only easily computable for natural numbers. This step is non-rigorous: *we will obtain an expression for* $\log\langle Z^n\rangle$ *for natural number* n*, and then take the* $n \to 0$ *limit after the fact.*

Recall that under the Gibbs distribution $p_\beta(\mathcal{D})$, the probability density on state x is equal to $\exp(-\beta H(x,\mathcal{D}))/Z(\mathcal{D})$. Denote by $p_\beta(\mathcal{D})^n$ the probability distribution over a tuple $\vec{x} = (x^{(1)}, \ldots, x^{(n)})$ of n independent samples (also known as **replicas**) chosen from $p_\beta(\mathcal{D})$.

Since the partition function Z is an integral (or sum in the discrete case) of the form $\int \exp(-\beta H(x;\mathcal{D}))dx$, we can write $Z(\mathcal{D})^n$ as the integral of $\prod_{a=1}^n \exp(-\beta H(x^{(a)};\mathcal{D})) = \exp\left(-\beta\sum_{a=1}^n H(x^{(a)},\mathcal{D})\right)$ where $x^{(1)}, \ldots, x^{(n)}$ are independent variables.

Now since each $x^{(a)}$ is weighed with a factor of $\exp(-\beta H(x^{(a)});\mathcal{D})$, this expression can be shown as equal to taking expectation of some exponential function $\exp(\sum_{a=1}^n G(x^{(a)};\mathcal{D}))$ over a tuple $(x^{(1)}, \ldots, x^{(n)})$ of n independent samples of **replicas** all coming from the same Gibbs distribution $p_\beta(\mathcal{D})$ corresponding to the same instance $\mathcal{D}$. (The discussion on G is just for intuition - we will not care about the particular form of this G, since soon average it over $\mathcal{D}$.)

Hence

$\langle Z^n\rangle = \left\langle \int_{\vec{x}\sim p_*(\mathcal{D})^n} \exp\left(-\sum_{a=1}^n G(x^{(a)},\mathcal{D})dx\right)\right\rangle_{\mathcal{D}}$.

Step 3: The Order Parameters

The above expression is an expectation of an integral, and so we can switch the order of summation, and write it also as

$\langle Z^n\rangle = \int_{\vec{x}\sim p_*(\mathcal{D})^n} \left\langle \exp\left(-\sum_{a=1}^n G(x^{(a)},\mathcal{D})\right)\right\rangle_{\mathcal{D}} dx.$

It turns out that for natural energy functions (for example when H is quadratic in x such as when it corresponds to Mean Squared Error loss), for any tuple of $x^{(1)}, \dots, x^{(n)}$, the expectation over $\mathcal{D}$ of $\exp\left(-\beta \sum_{a=1}^{n} H(x^{(a)}, \mathcal{D})\right)$ only depends on the angles between the $x^(a)$'s. That is, rather than depending on all of these N-dimensional vectors, it only depends on the n^2 coefficients $Q_{ab} = \frac{1}{N} x^{(a)} \cdot x^{(b)}$. The $n \times n$ matrix Q is known as the **overlap matrix** or **order parameters** and one can often find a nice analytical function $\mathcal{F}$ whose values are bounded (independently of N) such that

$\left\langle \exp\left(-\sum_{a=1}^{n} G(x^{(a)}, \mathcal{D})\right)\right\rangle_{\mathcal{D}} = \exp(-nN\mathcal{F}(Q))$.

Hence we can replace the integral over $x^{(1)}, \dots, x^{(n)}$ with an integral over Q and write

$\langle Z^n \rangle = \int dQ \exp\left(-nN\mathcal{F}(Q)\right)$

where the measure dQ is the one induced by the overlap distribution of a tuple $\vec{x} \sim p_\beta(\mathcal{D})$ taken for a random choice of the parameters $\mathcal{D}$.

Since Q only ranges over a small (n^2 dimensional set), at the large N limit, the integral is dominated by the maximum of its integrand ("method of steepest descent" / "saddle point method"). Let Q^* be the global minimum of $\mathcal{F}$ (within some space of matrices). We have

$\lim_{N\to\infty} \langle Z^n \rangle = \exp(-nN\mathcal{F}(Q^*))$.

Once we arrive at this expression, the configurational average of $-\log Z$ is simply $N\mathcal{F}(Q^*)$. These steps constitute the replica method. The ability to compute the configurational average by creating an appropriate Q is one of the factors determining whether the replica method can be used. For example, in the supervised learning example, H is almost always assumed to be quadratic in x; cross-entropy loss, for instance, is generally not amendable.

> *Bad Math Warning*: there are three limits, $\beta \to \infty$, $N \to \infty$, and $n \to 0$. In replica calculations, we assume that we take take these limits in whichever order that is arithmetically convenient.

IV. Constraints on Q: Replica Symmetry and Replica Symmetry Breaking

Unfortunately, the description of Q in Step 3 is a gross simplification. Q cannot be any matrix – it needs to satisfy particular constraints. In particular, it cannot be a matrix that appears with probability tending to zero with N as the overlap matrix of a tuple of n replicas from the Gibbs distribution. Hence, we need to understand the space of potential matrices Q that could arise from the probability distribution, and Q^* is the global minimum under these constraints.

The most important constraint on Q is the **replica symmetry** (RS), or the lack thereof (**replica symmetry breaking**, or **RSB**). Recall that Q encodes the overlap between $\{x^{(a)}\}$, where each element is a Gibbs random variable. On a high level, the structure of Q describes the geometry of the Gibbs distribution. An in depth description of the relationship between the two is beyond the scope of this post (check out Mean Field Models for Spin Glasses by Michel Talagrand). We will give some intuitions that apply in the zero-temperature limit.

A. What is the symmetry ansatz and when is it a good idea?

The **replica symmetric ansatz** studies the following special form of Q matrix

$$Q_{ab} = (q - q_0)\delta_{ab} + q_0$$

where δ_{ab} is the Kroneker delta. In other words, this ansatz corresponds to the guess that if we pick n random replicas $x^{(1)}, \dots, x^{(n)}$ then they will satisfy that $\|x^{(a)}\|^2 \approx q$ for all $a = 1, \dots n$, and $x^{(a)} \cdot x^{(b)} \approx q_0$ for $a \neq b$. This ansatz is especially natural for problems with unique minimizers for a fixed problem instance $\mathcal{D}$. In such a problem we might imagine that the n replicas are all random vectors that are have the same

correlation with the true minimizer $x^{(0)}$ and since they are random and in high dimension, this correlation explains their correlation with one another (see example below).

> **What have we done?** It is worthwhile to pause and take stock of what we have done here. We have reduced computing $\langle \log Z \rangle$ into finding an expression for $\langle Z^n \rangle$ and then reduced this to computing $\mathbb{E} \exp(-nN\mathcal{F}(Q))$ whre the expectation is taken over the induced distribution of the $n \times n$ overlap matrix. Now for every fixed n, we reduce the task to optimizing over just two parameters q and q_0. Once we find the matrix Q^* that optimizes this bound, we obtain the desired quantity by taking $\lim_{n\to 0} \frac{1}{n} \log \exp(-nN\mathcal{F}(Q^*) = N\mathcal{F}(Q^*)$.

B. An illustration:

Annealed langevin dynamics on a convex and non-convex objective below illustrate how the geometry of the learning problem influences the structure of the overlap matrix Q.

A convex problem https://www.youtube.com/watch?v=5Ekqn4qNX34

A non-convex problem https://www.youtube.com/watch?v=euyd9z0___YY

We see that even in low-dimensional problems, the structure of the loss landscape influences the resulting Q matrices. Since the replica method works only in high dimensions, these animations cannot be taken too seriously as a justification of the symmetry ansatz, but below we discuss in what kinds of models we could expect the symmetry ansatz to be a good idea.

C. Replica Symmetry in High Dimensions

We will now discuss a simple model where the replica symmetry ansatz is especially natural. For a fixed problem instance $\mathcal{D}$, suppose that the x_a vectors are distributed in a point cloud about some mean vector $\mu \in \mathbb{R}^N$.

$$x_a = \mu + \epsilon_a$$

where ϵ_a are zero-mean noise independently sampled across different replicas with covariance $\left\langle \epsilon_{a,i}\epsilon_{b,j} \right\rangle = \frac{\sigma^2}{N}\delta_{ab}\delta_{ij}$. This is equivalent to stipulating a Gibbs measure with energy $\beta H(x) = -\log p_\epsilon(x-\mu)$ where $p_\epsilon(\cdot)$ is the distribution of each noise variable. In this case, the Q matrix has elements

$$Q_{ab} = |\mu|^2 + \mu^\top \epsilon_a + \mu^\top \epsilon_b + \epsilon_a^\top \epsilon_b$$

By the central limit theorem, these sums of independently sampled random variables are approximately Gaussian (remember $N \to \infty$), so we can estimate how Q behaves in the large N limit

$$\langle Q_{ab} \rangle = |\mu|^2 + \sigma^2 \delta_{ab} \ , \ \text{Var } Q_{ab} = O(1/N)$$

This implies that in the thermodynamic $N \to \infty$ limit, the Q matrix concentrates around a replica symmetric structure. Note that the emergence of this RS structure relied on the fact that high dimensional random vectors are approximately orthogonal. For many supervised learning problems such as least squares fitting, this toy model is actually relevant by specifically taking $\epsilon \sim \mathcal{N}(0, \sigma^2/N)$.

V. Example Simple Problem: Learning Curve Phase Transition in Least Squares Fitting

To show these tools in action we will first study the simplest possible example with that has an interesting outcome. We will study the generalization performance of ridge regression on Gaussian distributed random features. In particular we will study a thermodynamic limit where the number of samples P and the number of features N are both tending to infinity $P, N \to \infty$ but with finite ratio $\alpha = P/N$. We will observe a phase transition the point $\alpha = 1$, where the learning problem transitions from over-parameterized ($P < N$) to under-parameterized ($P > N$). In the presence of noise this leads to an overfitting peak which can be eliminated through explicit regularization.

A. Some References

Hertz, Krogh, and Thorbergsson first studied this problem and noted the phase transition at $\alpha = 1$. Advani and Ganguli examine this model as a special case of M-estimation. Analysis of this model can also be obtained as a special case of kernel regression, for which general learning curves were obtained by Canatar, Bordelon, and Pehlevan with the replica method. Similar overfitting peaks were recently observed in nonlinear two layer neural networks by Belkin, Hsu, Ma, and Mandal and modeled with the replica method by d'Ascoli, Refinetti, Biroli, and, Krzakala allowing them to clarify the two possible types of overfitting peaks in random feature models. This problem can also be studied with tools from random matrix theory as in the work of Mei and Montanari and several others.

B. Problem Setup

Our problem instance is a dataset $\mathcal{D} = \{(x^\mu, y^\mu)\}_{\mu=1}^P$ with $x^\mu \in \mathbb{R}^N$ drawn i.i.d. from a Gaussian distribution $x_k^\mu \sim \mathcal{N}(0, 1/N)$. The target values y^μ are generated by a noisy linear teacher

$$y^\mu = \sum_{k=1}^N w_k^* x_k^\mu + \sigma \epsilon^\mu$$

where $||w^*||^2 = N$ and noise is Gaussian distributed $\epsilon^\mu \sim \mathcal{N}(0, 1)$. We will compute, not the generalization error for a particular problem instance $\mathcal{D}$, but the *average* performance over random datasets! The energy function we study is the ridge regression loss function

$$H(w, \mathcal{D}) = \frac{1}{\lambda} \sum_{\mu=1}^P \left(\sum_{k=1}^N w_k x_k^\mu - y^\mu \right)^2 + \sum_k w_k^2$$

The ridge parameter λ controls the trade-off between training accuracy and regularization of the weight vectors. When $\lambda \to 0$, the training data are fit perfectly while the $\lambda \to \infty$ limit gives $w = 0$ as the minimizer of H. The $\beta \to \infty$ limit of the Gibbs distribution corresponds to studying the performance of the ridge regression solution which minimizes H. The generalization error is an average over possible test points drawn from the same distribution

$$E_g = \left\langle \left(\sum_k (w_k - w_k^*) x_k \right)^2 \right\rangle_x = \frac{1}{N} \sum_{k=1}^N (w_k - w_k^*)^2$$

C. Partition Function

We introduce a partition function for the Gibbs distribution on $H(w, \mathcal{D})$

$$Z(\mathcal{D}) = \int dw \exp\left(-\frac{\beta}{2\lambda}\sum_{\mu=1}^{P}(w^\top x^\mu - y^\mu)^2 + \frac{\beta}{2}||w||^2\right).$$

D. Replicated Partition Function

We can rewrite the integral through a simple change of variables $\Delta = w - w^*$ since $y^\mu = w^{*\top}x^\mu + \sigma\epsilon^\mu$. Δ represents the discrepancy between the learned weights w and the target weights w^*. We will now replicate and average over the training data $\mathcal{D}$, ie compute $\langle Z^n\rangle$.

$$\langle Z(\mathcal{D},J)^n\rangle_{\mathcal{D}} = \int\prod_{a=1}^{n} d\Delta_a \left\langle \exp\left(-\frac{\beta}{2\lambda}\sum_{\mu=1}^{P}\sum_{a=1}^{n}(\Delta_a\cdot x^\mu - \sigma\epsilon^\mu)^2 - \frac{\beta}{2}\sum_{a=1}^{n}||\Delta_a + w^*||^2\right)\right\rangle_{\mathcal{D}}$$

> :warning: Notice that by writing these integrals, we are implicitly assuming that n is an integer. Eventually, we need to take $n \to 0$ limit to obtain the generalization error E_g from $\langle \log Z\rangle$. After computation of $\langle Z^n\rangle$ at integer n, we will get an analytic expression of n which we will allow us to non-rigorously take $n \to 0$.

The randomness from the dataset $\mathcal{D}$ is present in the first term only appears through mean-zero Gaussian variables $\gamma_a^\mu = \Delta_a\cdot x^\mu - \epsilon^\mu\sigma$ which have covariance structure

$$\langle\gamma_a^\mu\gamma_b^\nu\rangle = \delta_{\mu\nu}\left[\frac{1}{N}\Delta_a\cdot\Delta_b + \sigma^2\right] = \delta_{\mu\nu}\left[Q_{ab} + \sigma^2\right] \implies \langle\gamma\gamma^\top\rangle = Q + \sigma^2 11^\top \in \mathbb{R}^{n\times n}$$

where $1 \in \mathbb{R}^n$ is the vector of all ones and we introduced overlap order parameters Q_{ab} defined as

$$Q_{ab} = \frac{1}{N}\Delta_a\cdot\Delta_b\ .$$

The average over the randomness in the dataset $\mathcal{D}$ is therefore converted into a routine Gaussian integral. Exploiting the independence over each data point, we break the average into a product of P averages.

$$\left\langle \exp\left(-\frac{\beta}{2\lambda}\sum_{a,\mu}(\gamma_a^\mu)^2\right)\right\rangle_{\{\gamma_a^\mu\}} = \left\langle \exp\left(-\frac{\beta}{2\lambda}\sum_a\gamma_a^2\right)\right\rangle_{\{\gamma_a\}}^P$$

Each average is a multivariate Gaussian integral of the form

$$\int\frac{d\gamma_1 d\gamma_2 ... d\gamma_n}{\sqrt{(2\pi)^n\det(Q+\sigma^2 I)}}\exp\left(-\frac{1}{2}\sum_{ab}\gamma_a\gamma_b\left(Q+\sigma^2 11^\top\right)^{-1}_{ab} - \frac{\beta}{2\lambda}\sum_a\gamma_a^2\right) = \det\left(I + \frac{\beta}{\lambda}Q + \frac{\beta}{\lambda}\sigma^2 11^\top\right)^{-1/2}$$

This integral can be derived by routine integration of Gaussian functions, which we derive in the Appendix.

E. Enforcing Order Parameter Definition

To enforce the definition of the order parameters, we insert delta-functions into the expression for $\langle Z^n\rangle$ which we write as Fourier integrals over dual order parameters $\hat{Q}_{ab}$

$$\delta(NQ_{ab} - \Delta_a\cdot\Delta_b) = \frac{1}{2\pi}\int d\hat{Q}_{ab}\exp\left(i\hat{Q}_{ab}(NQ_{ab} - \Delta_a\cdot\Delta_b)\right)$$

This trick is routine and is derived in the Appendix of this post.

After integration over $\mathcal{D}$ and Δ_a, we are left with an expression of the form

$$\langle Z^n \rangle = \int \prod_{ab} dQ_{ab} \prod_{ab} d\hat{Q}_{ab} \exp\left(-PG_E(Q) + iN \sum_{ab} Q_{ab}\hat{Q}_{ab} - NG_S(\hat{Q}) \right)$$

where G_E is a function which arises from the average over γ_a^μ and G_S is calculated through integration over the Δ_a variables.

> :warning: The functions G_E and G_S have complicated formulas and we omit them here to focus on the conceptual steps in the replica method. Interested readers can find explicit expressions for these functions in the references above.

F. Replica Symmetry

To make progress on the integral above, we will make the replica symmetry assumption, leveraging the fact that the ridge regression loss is convex and has unique minimizer for $\lambda > 0$. Based on our simulations and arguments above, we will assume that the Q and $\hat{Q}$ matrices satisfy *replica symmetry*

$$Q_{ab} = q\delta_{ab} + q_0 \ , \ \hat{Q}_{ab} = \hat{q}\delta_{ab} + \hat{q}_0$$

G. Saddle Point Equations and Final Result

After the replica symmetry ansatz, the replicated partition function has the form

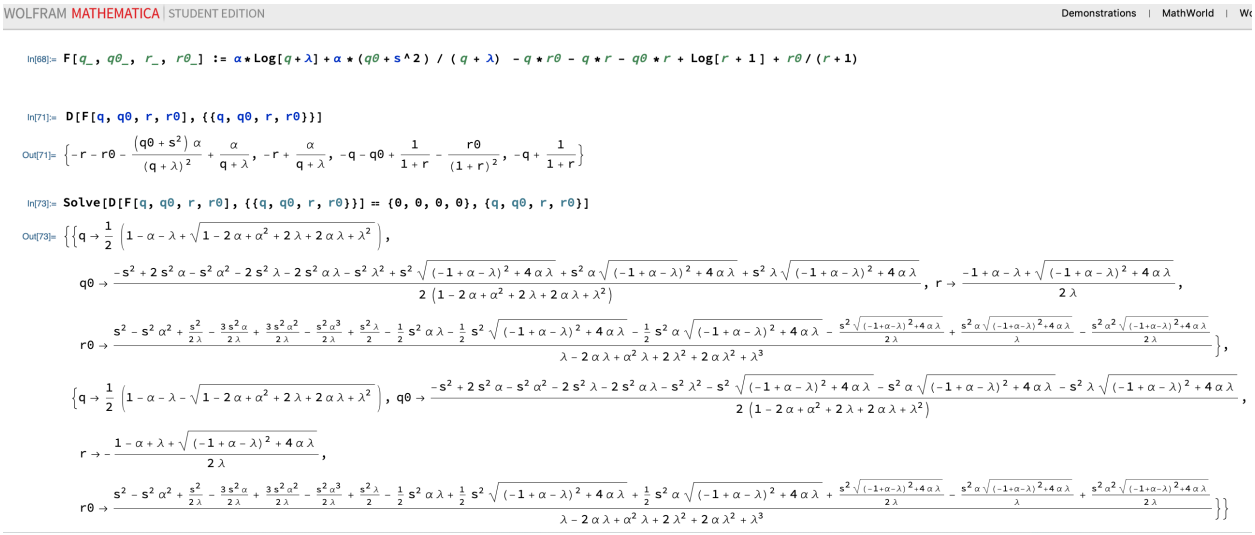
$$\langle Z^n \rangle = \int dq d\hat{q} dq_0 d\hat{q}_0 \exp(-nN\mathcal{F}(q, \hat{q}, q_0, \hat{q}_0))$$

In the $N \to \infty$ limit, this integral is dominated by the order parameters $q, \hat{q}, q_0, \hat{q}_0$ which satisfy the saddle point equations

$$\frac{\partial \mathcal{F}}{\partial q} = 0 \ , \ \frac{\partial \mathcal{F}}{\partial \hat{q}} = 0 \ , \ \frac{\partial \mathcal{F}}{\partial q_0} = 0 \ , \ \frac{\partial \mathcal{F}}{\partial \hat{q}_0} = 0$$

> :warning: Notice that n is small (we are working in $n \to 0$ limit to study $\log Z$) but N is large (we are studying the "thermodynamic" $N \to \infty$ limit). The order of taking these limits matters. It is important that we take $N \to \infty$ first before taking $n \to 0$ so that, at finite value of n, the integral for $\langle Z^n \rangle$ is dominated by the saddle point of $\mathcal{F}$.

We can solve the saddle point equations symbolically with Mathematica (see this notebook) in the $\beta \to \infty$ limit. We notice that Q must scale like $O(1/\beta)$ and $\hat{Q}$ must scale like $O(\beta)$. After factoring out the dependence on the temperature, we can compute the saddle point conditions through partial differentiation.



This symbolically gives us the order parameters at the saddle point. For example, the overlap parameter $q = \frac{1}{2}[1-\lambda-\alpha+\sqrt{(1-\lambda-\alpha)^2+4\lambda}]$. After solving the saddle point equations, the generalization error can be written entirely in terms of the first order parameter q at the saddle point. For replica a, the generalization error is merely $||\Delta_a||^2 = ||w_a - w^*||^2 = Q_{aa} = q + q_0$. Thus

$$E_g = q + q_0 = \frac{(q+\lambda)^2 + \sigma^2\alpha}{(q+\lambda+\alpha)^2 - \alpha}$$

Where $q + \lambda = \frac{1}{2}\left[1 + \lambda - \alpha + \sqrt{(1+\lambda-\alpha)^2 + 4\lambda\alpha}\right]$ at the saddle point.

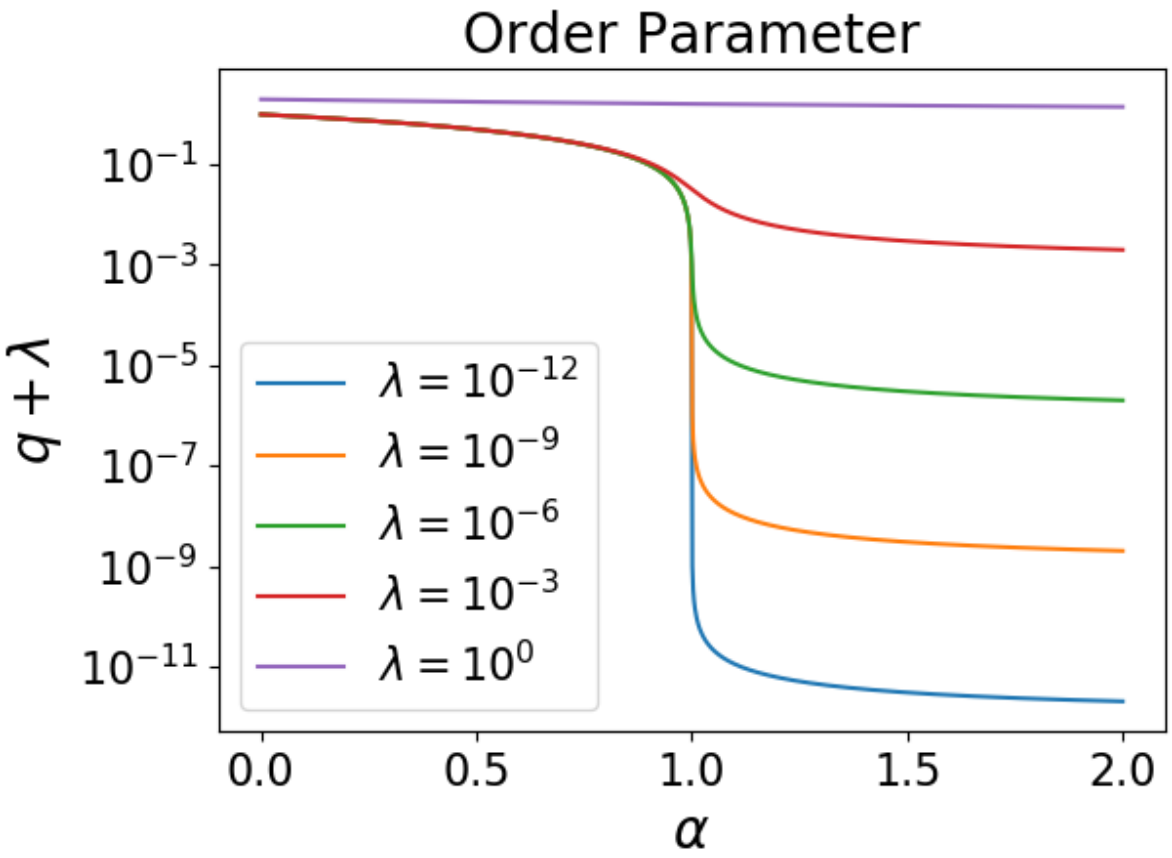
H. Noise Free Estimation

When $\sigma^2, \lambda \to 0$ the generalization error decreases linearly with α: $E_g = 1 - \alpha$ for $\alpha < 1$ and $E_g = 0$ for $\alpha > 1$. This indicates the target weights are perfectly estimated when the number of samples equals the number of features $P \to N$. A finite ridge parameter $\lambda > 0$ increases the generalization error when noise is zero $\sigma^2 = 0$. Asymptotically, the generalization error scales like $E_g \sim \frac{\lambda^2}{\alpha^2}$ for large α.

###I. Phase transition and overfitting peaks

In the presence of noise $\sigma^2 > 0$, the story is different. In this case, the generalization error exhibits a peak at $\alpha \approx 1 + \lambda$ before falling at a rate $E_g \sim \sigma^2/\alpha$ at large α. In this regime, accurate estimation requires reducing the variance of the estimator by increasing the number of samples.

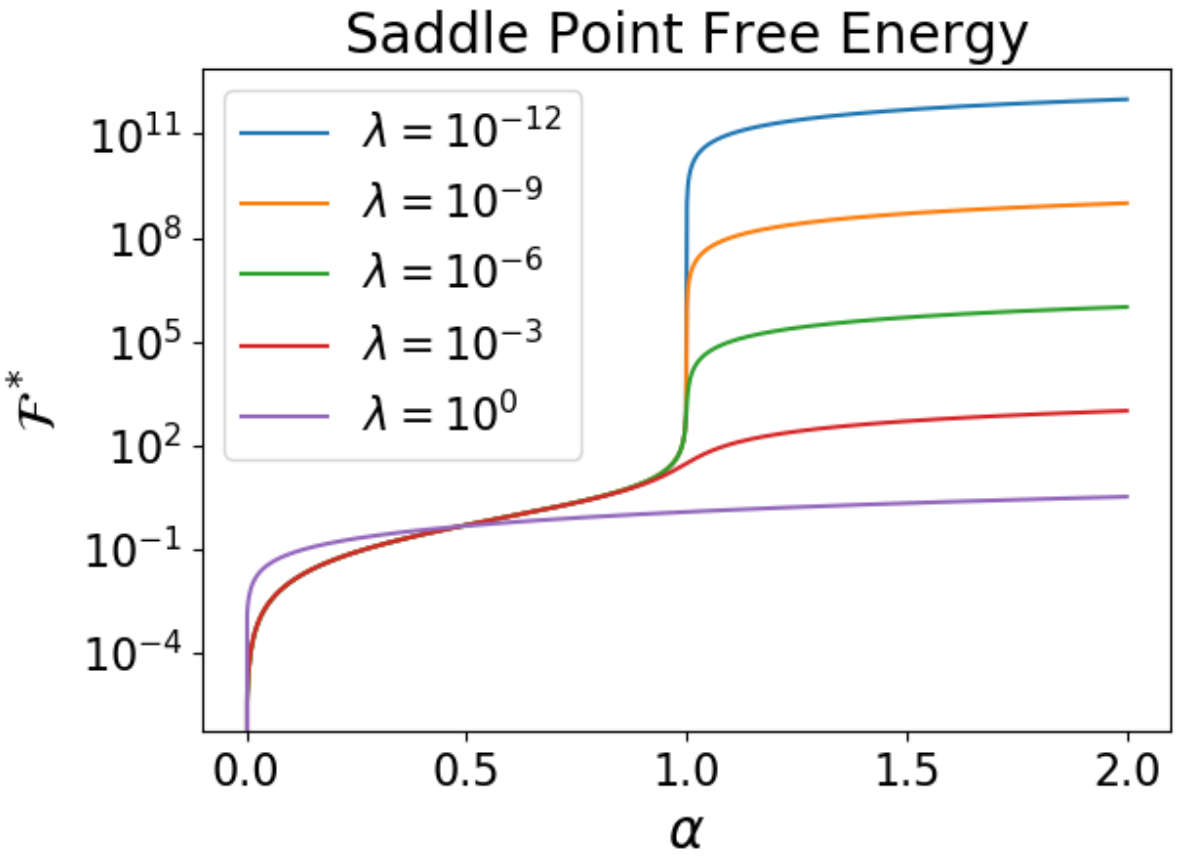
In small $\lambda \to 0$ limit, the order parameter behaves like $q + \lambda \sim 1 - \alpha$ for $\alpha < 1$ and $q + \lambda \sim \lambda$ for $\alpha > 1$.



The free energy $\mathcal{F}$ exhibits a discontinuous first derivative as $\alpha \to 1$, a phenomenon known as first-order phase transition. Let $\mathcal{F}^*$ be the value of the free energy at the saddle point $(q^*, \hat{q}^*, q_0^*, \hat{q}_0^*)$. Then we find

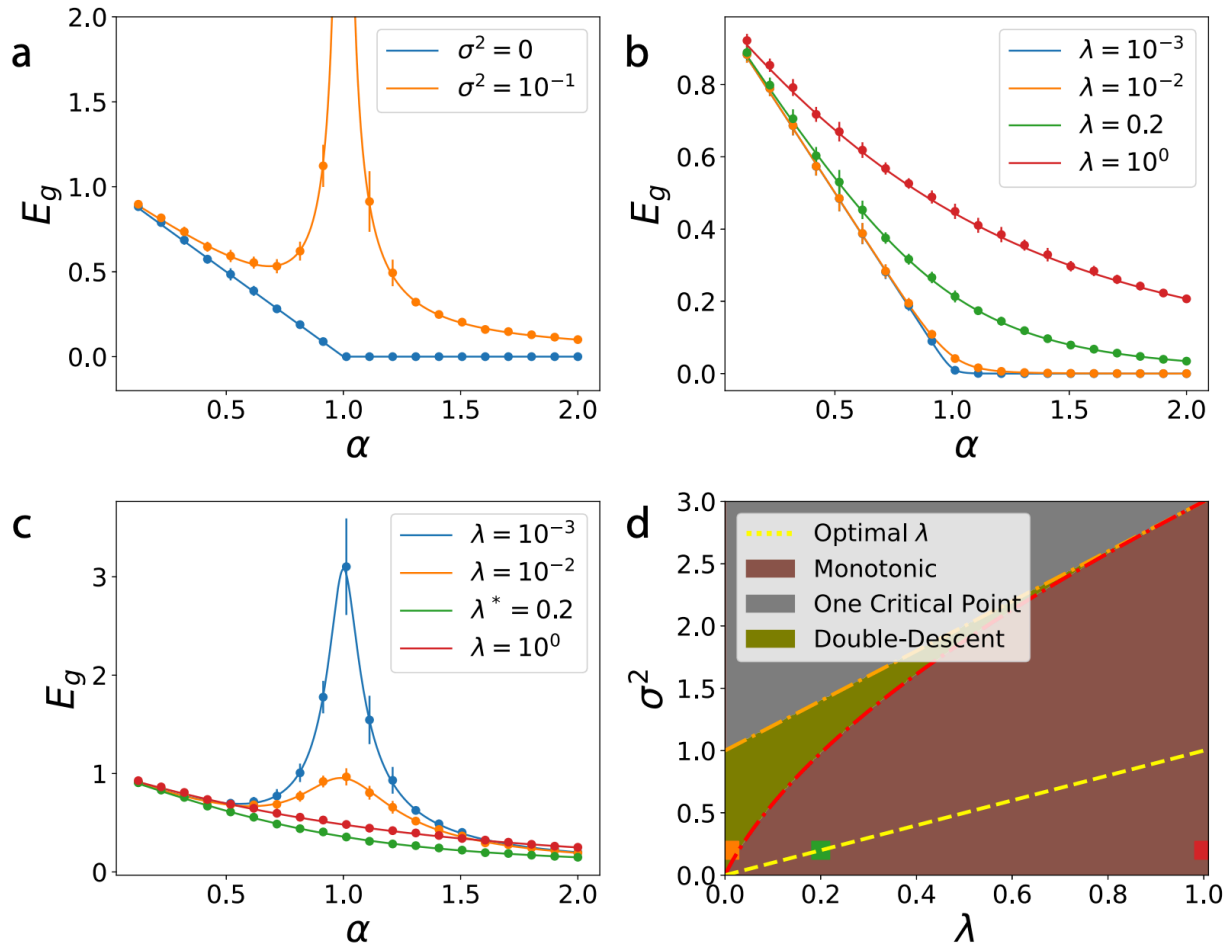
$$\frac{\partial \mathcal{F}^*}{\partial \alpha} \sim \frac{\sigma^2}{2\lambda}\Theta(\alpha - 1) + \mathcal{O}_\lambda(1) \ , \ (\lambda \to 0, \alpha \to 1)$$

which indicates a discontinous first derivative in the $\lambda \to 0$ limit as $\alpha \to 1$. We plot this free energy $\mathcal{F}^*(\alpha)$ for varying values of λ, showing that as $\lambda \to 0$ a discontinuity in the free energy occurs at $\alpha \to 1$. The non-zero ridge parameter $\lambda > 0$ prevents the strict phase transition at $\alpha = 1$.



J. Putting it all together

Using the analysis of the saddle point, we are now prepared to construct a full picture of the possibilities. A figure below from this paper provies all of the major insights. We plot experimental values of generalization error E_g in a $N = 800$ dimensional problem to provide a comparison with the replica prediction.



(a) When $\lambda = 0$, the generalization error either falls like $1 - \alpha$ if noise is zero, or it exhibits a divergence at $\alpha \to 1$ if noise is non-zero.

(b) When noise is zero, increasing the explicit ridge λ increases the generalization error. At large α, $E_g \sim \frac{\lambda^2}{\alpha^2}$.

(c) When there is noise, explicit regularization can prevent the overfitting peak and give optimal generalization. At large α, $E_g \sim \frac{\sigma^2}{\alpha}$.

(d) In the (λ, σ^2) plane, there are multiple possibilities for the learning curve $E_g(\alpha)$. Monotonic learning curves $E_g'(\alpha) < 0$ are guaranteed provided λ is sufficiently large compared to σ^2. If regularization is too small, then two-critical points can exist in the learning curve, ie two values α^* where $E_g'(\alpha^*) = 0$ (sample wise double descent). For very large noise, a single local maximum exists in the learning curve $E_g(\alpha)$, which is followed by monotonic decreasing error.

VI. Example Problem 2: Spiked Matrix Recovery

Detailed calculations can be found in this excellent introduction of the problem by Song Mei.

Suppose we have a N-by-N rank-1 matrix, $\lambda \mathbf{u}\mathbf{u}^T$, where $\mathbf{u}$ is a norm-1 column vector constituting the signal that we would like to recover. The input $\mathbf{A}$ we receive is corrupted by symmetric Gaussian i.i.d. noise, i.e.,

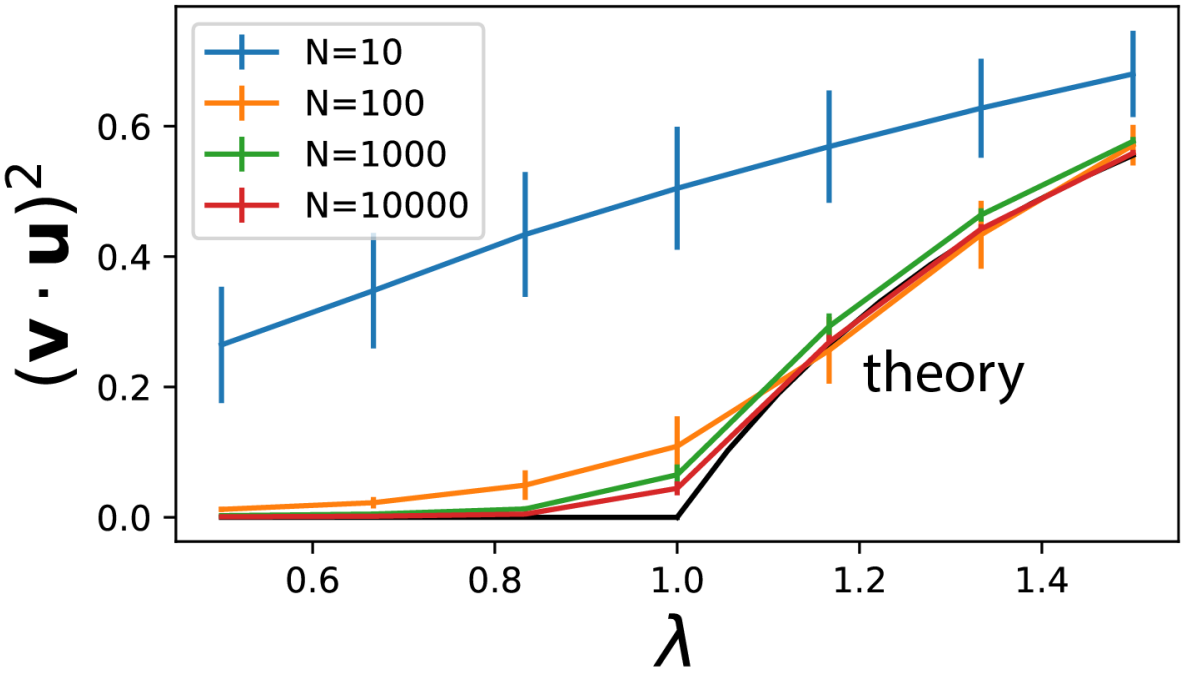
$$\mathbf{A} = \lambda \mathbf{u}\mathbf{u}^T + \mathbf{W}$$

where $W_{ij} = W_{ji} \sim \mathcal{N}(0, 1/N), W_{ii} \sim \mathcal{N}(0, 2/N)$ ($\mathbf{W}$ is drawn from a Gaussian Orthogonal Ensemble). At large N, eigenvalues of $\mathbf{W}$ are distributed uniformly on a unit disk in the complex plane. Thus, the best estimate (which we call $\mathbf{v}$) of $\mathbf{u}$ from $\mathbf{A}$ is the eigenvector associated with the largest eigenvalue. In other words

$$\mathbf{v} = \arg \max_{\mathbf{x} \in \mathbb{S}^{N-1}} \mathbf{x}^T \mathbf{A} \mathbf{x}.$$

The *observable* of interest is how well the estimate, $\mathbf{v}$, matches $\mathbf{u}$, as measured by $(\mathbf{v} \cdot \mathbf{u})^2$. We would like to know its average over different $\mathbf{W}$.

In the problem setup, λ is a constant controlling the signal-to-noise ratio. Intuitively, the larger λ is, the better the estimate should be (when averaged over $\mathbf{W}$). This is indeed true. Remarkably, for large N, $\mathbf{v} \cdot \mathbf{u}$ is almost surely 0 for $\lambda < 1$. For $\lambda \geq 1$, it grows quickly as $1 - \lambda^{-2}$. This discontinuity at $\lambda = 1$ is a **phase transition**. This dependence on λ can be derived using the replica method.



In the simulations above, we see two trend with increasing N. First, the average curve approaches the theory, which is good. In addition, trial-to-trial variability (as reflected by the error bars) shrinks. This reflects the fact that our observable is indeed self-averaging.

Here, we give a brief overview of how the steps of a replica calculation can be set up and carried out.

Step 1

Here, $\mathbf{W}$ is the problem parameter ($\mathcal{P}$) that we average over. The minimized function is

$$H(\mathbf{x}, \mathbf{W}) = -\mathbf{x}^T(\lambda \mathbf{u}\mathbf{u}^T + \mathbf{W})\mathbf{x} = -\lambda(\mathbf{x} \cdot \mathbf{u})^2 - \mathbf{x}^T \mathbf{W} \mathbf{x}.$$

This energy function already contains a "source term" for our observable of interest. Thus, the vanilla partition function will be used as the augmented partition function. In addition, this function does not scale

with N. To introduce the appropriate N scaling, we add an N factor to H, yielding the partition function

$$Z(\mathbf{W}, \lambda) = \int_{\mathbb{S}^{N-1}} d\mathbf{x} \exp\left(-\beta N H(\mathbf{x}, \mathbf{W})\right).$$

It follows that (again using angular brackets to denote average over $\mathbf{W}$)

$$\langle (\mathbf{x} \cdot \mathbf{u})^2 \rangle = \frac{1}{\beta N} \frac{d}{d\lambda} \langle \log Z(\mathbf{W}, \lambda) \rangle \Big|_{\lambda=0}.$$

Since we are ultimately interested in this observable for the best estimate, $\mathbf{v}$, at the large N, limit, we seek to compute

$$\lim_{N\to\infty} \mathbb{E}[(\mathbf{v} \cdot \mathbf{u})^2] = \lim_{N\to\infty} \lim_{\beta\to\infty} \frac{1}{\beta N} \frac{d}{d\lambda} \langle \log Z(\mathbf{W}, \lambda) \rangle.$$

Why don't we evaluate the derivative only at $\lambda = 0$? Because $\lambda(\mathbf{x} \cdot \mathbf{u})^2$ is not a source term that we introduced. Another way to think about it is that this result needs to be a function of λ, so of course we don't just evaluate it at one value.

Step 2

Per the replica trick, we need to compute

$$\langle Z^n \rangle = \left\langle \prod_{a=1}^{n} \left(\int D\mathbf{x}^{(a)} \right) \exp\left(\beta N \sum_{a=1}^{n} \lambda (\mathbf{x}^{(a)} \cdot \mathbf{u})^2 + \mathbf{x}^{(a)T} \mathbf{W} \mathbf{x}^{(a)} \right) \right\rangle \tag{1}$$

$$= \int D\mathbf{W} \prod_{a=1}^{n} \left(\int D\mathbf{x}^{(a)} \right) \exp\left(\beta N \sum_{a=1}^{n} \lambda (\mathbf{x}^{(a)} \cdot \mathbf{u})^2 + \mathbf{x}^{(a)T} \mathbf{W} \mathbf{x}^{(a)} \right). \tag{2}$$

> :warning: Hereafter, our use of "$\langle Z^n \rangle =$" is loose. When performing integrals, we will ignore the constants generated and only focus on getting the exponent right. This is because we will eventually take log of $\langle Z^n \rangle$ and take the derivative w.r.t. λ. A constant in λ in front of the integral expression for $\langle Z^n \rangle$ does not affect this integral. This is often the case in replica calculations.

where $D\mathbf{x}$ is a uniform measure on $\mathbb{S}^{N-1}$ and $D\mathbf{W}$ is the probability measure for $\mathbf{W}$, as described above.

We will not carry out the calculation in detail in this note as the details are problem-specific. But the overall workflow is rather typical of replica calculations:

1. Integrate over $\mathbf{W}$. This can be done by writing $\mathbf{W}$ as the sum of a Gaussian i.i.d. matrix with its own transpose. The integral is then over the i.i.d. matrix and thus a standard Gaussian integral. After this step, we obtain an expression that no longer contains $\mathbf{W}$,a major simplification.
2. Introduce the order parameter $\mathbf{Q}$. After the last integral, the exponent only depends on $\mathbf{x}^{(a)}$ through $\mathbf{u} \cdot \mathbf{x}^{(a)}$ and $\mathbf{x}^{(\mathbf{a})} \cdot \mathbf{x}^{(\mathbf{b})}$. These dot products can be described by a matrix $\mathbf{Q} \in \mathbb{R}^{N+1 \times N+1}$, where we define $Q_{0,a} = Q_{a,0} = \mathbf{u} \cdot \mathbf{x}^{(a)}$ and $Q_{a\geq 1, b\geq 1} = \mathbf{x}^{(a)} \cdot \mathbf{x}^{(b)}$.
3. Replace the integral over $\mathbf{x}$ with one over $\mathbf{Q}$. A major inconvenience of the integral over $\mathbf{x}$ is that it is not over the entire real space but over a hypersphere. However, we can demand $\mathbf{x}^{(a)} \in \mathbb{S}^{N-1}$ by requiring $Q_{aa} = 1$. Now, we rewrite the exponent in terms of $\mathbf{Q}$ and integrate over $\mathbf{Q}$ instead, but we add many Dirac delta functions to enforce the definition of $\mathbf{Q}$. We get an expression in the form

$$\langle Z^n \rangle = \int d\mathbf{Q} \exp(f(Q)) \prod_{i=1}^{N} \delta(\mathbf{u} \cdot \mathbf{x}^{(i)} - Q_{0i}) \prod_{1\leq i\leq j\leq N} \delta(\mathbf{x}^{(j)} u \cdot \mathbf{x}^{(i)} - Q_{ji}).$$

4. After some involved simplifications, we have

$$\langle Z^n \rangle = \int d\mathbf{Q} \exp(Ng(\mathbf{Q}) + C)$$

where C does not depend on $\mathbf{Q}$ and $g(\mathbf{Q})$ is $O(1)$. By the saddle point method,

$$\langle Z^n \rangle = \max_{\mathbf{Q}} \exp(Ng(\mathbf{Q})),$$

where $\mathbf{Q}$ needs to satisfy the various constraints we proposed (e.g., its diagonal is all 1 and it is symmetric).

Step 3

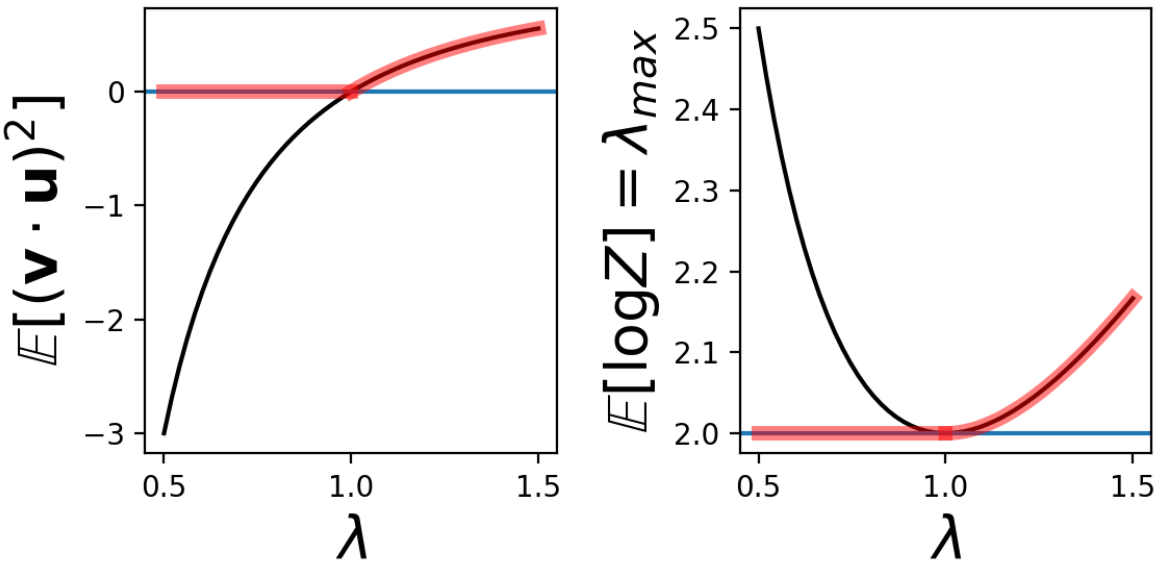
The optimization over $\mathbf{Q}$ is not trivial. Hence, we make some guesses about the structure of $\mathbf{Q}^*$, which maximizes the exponent. This is where the *replica symmetry (RS) ansatz* comes in. Since the indices of $\mathbf{x}^{(a)}$ are arbitrary, one guess is that for all $a, b \geq 1$ $Q_{a \neq b}$ has the same value ,q. In addition, for all a, $Q_{0,a} = \mu$. This is the RS ansatz – it assumes an equivalency between replicas. Rigorously showing whether this is indeed the case is challenging, but we can proceed with this assumption and see if the results are correct.

The maximization of $g(\mathbf{Q})$ is now over two scalars, μ and q. Writing the maximum as μ^*, q^* and using the replica identity

$$\langle \log Z \rangle = \lim_{n \to 0} \log \langle Z^n \rangle / n = \lim_{n \to 0} Ng(\mu^*, q^*).$$

Setting the derivative of g w.r.t. them to zero yields two solutions. >:warning: Bad Math Warning: Maximizing g w.r.t. μ, q requires checking that the solutions are indeed global minima, a laborious effort that has done for some models. We will assume them to be global minima.

For each solution, we can compute $\lim_{N \to \infty} \lim_{\beta \to \infty} \frac{1}{\beta N} \langle \log Z \rangle$, which will become what we are looking for after being differentiated w.r.t. λ. We obtain an expression of $\langle \log Z \rangle$. The two solutions yield $\langle \log Z \rangle = \lambda + 1/\lambda$ and $\langle \log Z \rangle = 2$, respectively. Differentiating each w.r.t. λ to get $\langle (v \cdot u)^2 \rangle$, we have $1 - \lambda^{-2}$ and 0.



Which one is correct? We decide by checking whether the solutions (black line and blue line above) are sensible ("physical"). It can be verified that $\lim_{N \to \infty} \lim_{\beta \to \infty} \frac{1}{\beta N} \langle \log Z \rangle = \langle \lambda_{\max} \rangle$, which is the largest

eigenvalue of $\mathbf{A}$. Clearly, it should be non-decreasing as a function of λ. Thus, for $\lambda \leq 1$, we choose the 0 solution, and for $\lambda \geq 1$ the $\lambda + 1/\lambda$ solution. Thus, $\langle (v \cdot u)^2 \rangle$ is 0 and $1 - \lambda^{-2}$ in the two regimes, respectively.

Appendix: Gaussian Integrals and Delta Function Representation

We frequently encounter Gaussian integrals when using the replica method and it is often convenient to rely on basic integration results which we provide in this Appendix.

Single Variable The simplest Gaussian integral is the following one dimensional integral

$$I(a) = \int_{-\infty}^{\infty} \exp\left(-\frac{a}{2}x^2\right) dx$$

We can calculate the square of this quantity by changing to polar coordinates $x = r\cos\phi, y = r\sin\phi$

$$I(a)^2 = \int_{\mathbb{R}^2} \exp\left(-\frac{a}{2}\left(x^2+y^2\right)\right) dxdy = \int_0^{2\pi} d\phi \int_0^{\infty} r \exp\left(-\frac{a}{2}r^2\right) dr = 2\pi a^{-1}$$

We thus conclude that $I(a) = \sqrt{2\pi/a}$. Thus we find that the function $\sqrt{\frac{a}{2\pi}}e^{-\frac{a}{2}x^2}$ is a normalized function over $(-\infty, \infty)$. This integral can also be calculated with Mathematica or Sympy. Below is the result in Sympy.

```
from sympy import *
from sympy.abc import a, b, x, y
x = Symbol('x')
integrate( exp( -a/2 * x**2 ) , (x, -oo,oo))
```

$$\begin{cases} \frac{\sqrt{2}\sqrt{\pi}}{\sqrt{a}} & \text{for } |\arg(a)| \leq \frac{\pi}{2} \\ \int_{-\infty}^{\infty} e^{-\frac{ax^2}{2}}\, dx & \text{otherwise} \end{cases}$$

This agrees with our result since we were implicitly assuming a real and positive ($\arg a = 0$).

We can generalize this result to accomodate slightly more involved integrals which contain both quadratic and linear terms in the exponent. This exercise reduces to the previous case through simple completion of the square

$$I(a,b) = \int_{-\infty}^{\infty} \exp\left(-\frac{a}{2}x^2 \pm bx\right) dx = \exp\left(\frac{b^2}{2a}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{a}{2}\left[x \mp \frac{b}{a}\right]^2\right) dx = \exp\left(\frac{b^2}{2a}\right)\sqrt{2\pi/a}$$

We can turn this equality around to find an expression

$$\exp\left(\frac{b^2}{2a}\right) = \sqrt{\frac{a}{2\pi}} \int \exp\left(-\frac{a}{2}x^2 \pm bx\right) dx$$

Viewed in this way, this formula allows one to transform a term quadratic in b in the exponential function into an integral involving a term linear in b. This is known as the Hubbard-Stratanovich transformation. Taking b be imaginary ($b = ik$ for real k), we find an alternative expression of a Gaussian function

$$\exp\left(-\frac{k^2}{2a}\right) = \sqrt{\frac{a}{2\pi}} \int \exp\left(-\frac{a}{2}x^2 \pm ikx\right) dx$$

Delta Function Integral Representation A delta function $\delta(z)$ can be considered as a limit of a normalized mean-zero Gaussian function with variance taken to zero

$$\delta(z) = \lim_{a \to 0} \sqrt{\frac{1}{2\pi a}} \exp\left(-\frac{1}{2a}z^2\right)$$

We can now use the Hubbard-Stratanovich trick to rewrite the Gaussian function

$$\exp\left(-\frac{1}{2a}z^2\right) = \sqrt{\frac{a}{2\pi}} \int \exp\left(-\frac{a}{2}x^2 \pm izx\right) dx$$

Thus we can relate the delta function to an integral

$$\delta(z) = \lim_{a \to 0} \frac{1}{2\pi} \int \exp\left(-\frac{a}{2}x^2 \pm ixz\right) dx = \frac{1}{2\pi} \int \exp\left(\pm ixz\right) dx$$

This trick is routinely utilized to represent delta functions with integrals over exponential functions during a replica calculation. In particular, this identity is often used to enforce defintions of the order parameters Q_{ab} in the problem. For example, in the least squares problem where $Q_{ab} = \frac{1}{N}\Delta_a \cdot \Delta_b$ we used

$$\delta(NQ_{ab} - \Delta_a \cdot \Delta_b) = \frac{1}{2\pi} \int d\hat{Q}_{ab} \exp\left(iNQ_{ab}\hat{Q}_{ab} - i\hat{Q}_{ab}\Delta_a \cdot \Delta_b\right)$$

Multivariate Gaussian integrals We commonly encounter integrals of the following form

$$I(M) = \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\sum_{ab} M_{ab}x_a x_b\right) dx_1 dx_2 ... dx_n$$

where matrix M_{ab} is symmetric and positive definite. An example is the data average in the least squares problem studied in this blog post where $M = (Q + \sigma^2 11^\top)^{-1} + \beta I$. We can reduce this to a collection of one dimensional problems by computing the eigendecomposition of $M = \sum_\rho \lambda_\rho u_\rho u_\rho^\top$. From this decomposition, we introduce variables

$$z_\rho = \sum_{a=1}^{n} u_{\rho,a} x_a$$

The transformation from x to z is orthogonal so the determinant of the Jacobian has absolute value one. After changing variables, we therefore obtain the following decoupled integrals

$$I(M) = \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\sum_{\rho} \lambda_\rho z_\rho^2\right) dz_1 ... dz_n = \prod_{\rho=1}^{n} \int \exp\left(-\frac{\lambda_\rho}{2} z_\rho^2\right) dz_\rho = \prod_{\rho=1}^{n} \sqrt{\frac{2\pi}{\lambda_\rho}}$$

Using the fact that the determinant is the product of eigenvalues $\det M = \prod_\rho \lambda_\rho$, we have the following expression for the multivariate Gaussian integral

$$I(M) = (2\pi)^{n/2} \det(M)^{-1/2}$$

.