8 Error analysis: jackknife & bootstrap

As discussed before, it is no problem to calculate the expectation values and *statistical error estimates* of “normal” observables from Monte Carlo. However, often we have to calculate functions which depend (possibly non-linearly) on the *expectation values* of some quantities.
As a common example, we may have 2 observables, a and b, and we want the *correlation coefficient*

$$\rho = \frac{\sum_i (a_i - \langle a \rangle)(b_i - \langle b \rangle)}{\sqrt{\sum_i (a_i - \langle a \rangle)^2 \sum_j (b_j - \langle b \rangle)^2}} = \frac{\langle (a - \langle a \rangle)(b - \langle b \rangle) \rangle}{\sqrt{\langle (a - \langle a \rangle)^2 \rangle \langle (b - \langle b \rangle)^2 \rangle}}$$

where, as usual, $\langle a \rangle = 1/N \sum_i a_i$.

Now, if we write the measurements a_i, b_i in a file, it is of course no problem to calculate ρ. However, *what is the error of our result for ρ?* We cannot construct a sigle-configuration quantity ρ_i, which we could plug in the autocorrelation analysis.

One methdod to do the analysis would be to divide the measurements $(a, b)_i$ in M blocks (bins), with block length $\gg \tau$, the autocorrelation time. Then we can calculate ρ_m for each block, and use the naive error formula for the error. However, this is not optimal: the value of ρ_m may vary a lot from block to block.

Jackknife and bootstrap methods are nowadays standard ways to calculate the error in this case.
– R.G. Miller, *the jackknife – a review*, Biometrika 61 (1974) pg. 1–17.
– B. Efron, *Computers and the theory of statistics: thinking the unthinkable*, SIAM Review, vol 21, No. 4 460 (1979)
– B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics (1982)
– Moore, D. S., G. McCabe, W. Duckworth, and S. Sclove (2003): Bootstrap Methods and Permutation Tests, `http://bcs.whfreeman.com/pbs/cat_140/chap18.pdf`

8.1 Reminder: standard error propagation

Often we need to calculate some function $f(\langle a \rangle)$ of the *expectation value* of some quantity a. It is of course easy to calculate $\langle a \rangle$ and insert that into $f(x)$, but what are the error bars of the result?

It should be noted that, as a rule, we cannot simply define $f_i = f(a_i)$ and use these as measurements of f; normally even

$$f(\langle a \rangle) \neq \langle f(a) \rangle = \frac{1}{N} \sum_i f_i .$$

For example, let a_i's be random numbers from uniform distribution $0 \leq a_i \leq 1$, and $f(x) = 1/x$. Now $f(\langle a \rangle) = 2$, but $\langle f(a) \rangle = \int_0^1 1/ada = \infty$.

Let us denote the average and error after N measurements as $\langle a \rangle$ and δa. As before, the error (1-σ) is defined as the variance of the gaussian probability distribution of $\langle a \rangle$ (the distribution will be gaussian if N is large enough).

If now we have the situation that δa is small enough so that the expansion

$$f(\langle a \rangle \pm \delta a) = f(\langle a \rangle) \pm (\delta a) f'(\langle a \rangle) + \ldots$$

is a good approximation when truncated to the first term, we obtain the result for the error of $f(\langle a \rangle)$:

$$\delta f = f'(\langle a \rangle)\, \delta a$$

What if f depends on more than 1 random variable, for example $f = f(\langle a \rangle, \langle b \rangle)$? If the random variables are *statistically independent* and the linearity above holds wrt. all of these, the errors are added *quadratically:*

$$(\delta f)^2 = \left[\frac{\partial f}{\partial \langle a \rangle}\right]^2 (\delta a)^2 + \left[\frac{\partial f}{\partial \langle b \rangle}\right]^2 (\delta b)^2$$

(This result comes from independent gaussian distributions). This generalises to any number of random variables. For example, if we add the averages of 2 independent random variables $\langle a \rangle + \langle b \rangle$, the error will be $\delta = \sqrt{(\delta a)^2 + (\delta b)^2}$.

However, very often we do not have independent statistical variables. In Monte Carlo simulations we measure different quantitites from the same configurations; thus, the measurements are naturally correlated. In this case the expression for the error above has to be modified to take this into account (and we obtain cross-correlation matrix). This rapidly becomes cumbersome, and it is recommended to use jackknife or bootstrap analysis methods, which take these into account automatically.

8.2 Jackknife

Jackknife method is a systematic way of obtaining the “standard deviation” error of a set of stochastic measurements:

1. Calculate average $\bar{\rho}$ (or some function f) from the full dataset
2. Divide data $(a, b)_i$ into M blocks, with block length $\gg \tau$. This is done in order to get rid of autocorrelations; if there are no correlations, block length can be 1.
3. For each $m = 1 \ldots M$, take away block m and calculate the average $\bar{\rho}_{(m)}$ using the data from all other blocks.

4. Estimate the error of ρ by calculating the deviation of $\bar\rho_{(m)}$'s from $\bar\rho$:

$$\boxed{\delta\rho = \sqrt{\frac{M-1}{M}\sum_{m=1}^{M}(\bar\rho_{(m)}-\bar\rho)^2}}$$

The factor $(M-1)/M$ is there to give the correct result if we look at the errors of simple observables. For example, assume that we have a random observable a_i. Dividing these into M blocks, we can calculate an average from each block as a_m. Now we obtain the jackknife-blocked average as $a_{(m)} = 1/(M-1)\sum_{m'\neq m} a_{m'}$. Thus,

$$\sqrt{\frac{M-1}{M}\sum_m (a_{(m)} - \langle a\rangle)^2} = \sqrt{\frac{\sum_m (a_m - \langle a\rangle)^2}{(M-1)M}}$$

i.e. we obtain the standard (blocked) error estimate.

Why does jackknife work? $\bar\rho_{(m)}$ contains almost the full set of data, thus, they are quite close to the full dataset value. Indeed, *each jackknifed block $\sim$ a new MC average of length $N-m$* (However, these are naturally not independent!). This is why jackknife and bootstrap are often called **resampling** methods: they construct pseudo-independent 'new' simulation results.

Note: parametrically,

$$\bar\rho_{(m)} - \bar\rho \sim \delta\rho/\sqrt{M}\ , \tag{3}$$

i.e. the distribution of the results of these new "simulations" is *narrower* than $\delta\rho$, the expected error. Thus, we cannot really consider jackknife sets to be "new" simultations (these would have distribution with width $\delta\rho$).

As in the example in Sec. 8.1, we often want an estimate of some function of the expectation value of a, $f(\langle a\rangle)$, where a is some quantity which we measure from simulation. Note that this is in general very different from $\langle f(a)\rangle$!

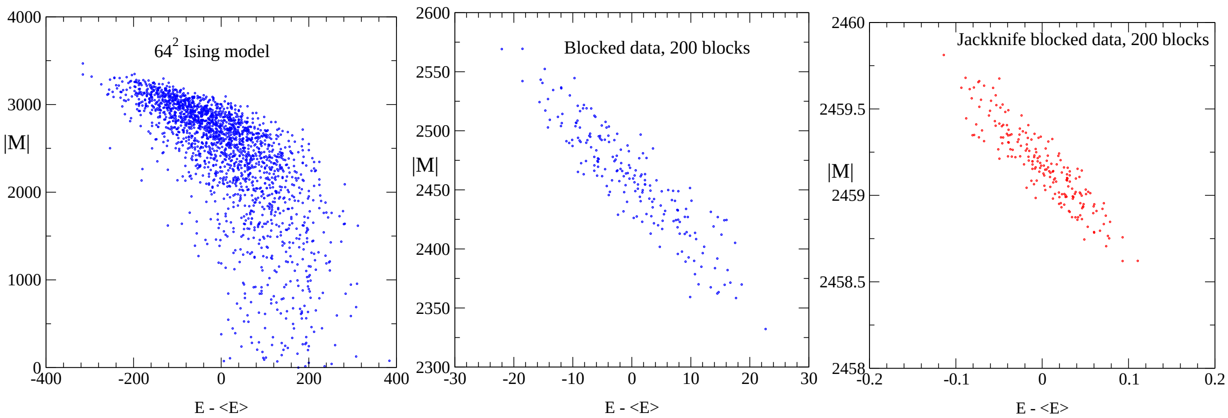
The jackknife error estimate is naturally

$$\delta f = \sqrt{\frac{M-1}{M}\sum_m (f_{(m)} - \langle f\rangle)^2} \approx |f'(\langle a\rangle)|\delta a \tag{4}$$

where $f_{(m)} = f(a_{(m)})$. This is typically a good approximation, because $f_{(m)}$ is close to $\langle f\rangle$. (Here $\langle f\rangle = \sum_m f_{(m)}/M$ *usually*; it could also be $\langle f\rangle = f(\langle a\rangle)$).

Jackknife blocked data makes it easy to 'chain' the blocks through consecutive functions: for example, if we want $g(f(\langle a\rangle))$, where a is some statistical measurable, we can form $a_{(m)}$, average of the observable a over jackknifed block m. Now we can calculate $f_{(m)} = f(a_{(m)})$ and $g_{(m)} = g(f_{(m)})$, from which the error is

$$\delta g(f(\langle a \rangle)) = \sqrt{\frac{M-1}{M} \sum_m (g_{(m)} - \langle g \rangle)^2}$$



Above: 64^2 Ising at β_c;
a) sample of ($E - \langle E \rangle, |M|$)-data,
b) block averages with 200 blocks of length 2000,
c) jackknife block averages with same blocks.

Each jackknife block is very close to the final value, and the distribution of the jackknifed data is within the error ellipse (scaled down by $\sqrt{200}$) of the $(\langle E \rangle, \langle |M| \rangle)$-data.

Feeding these blocks through some (non-linear) function, the distribution of the results gives us directly the error of the function value.

Let us calculate the correlation coefficient of $a = (E - \langle E \rangle)$ and $b = (|M| - \langle |M| \rangle)$,

$$\rho = \frac{\langle ab \rangle}{\langle a \rangle \langle b \rangle}$$

Doing this with 200 jackknife blocks as above, we have to calculate $\langle a \rangle$, $\langle b \rangle$ and $\langle ab \rangle$ for each of the 200 jackknife sets, each containing $199 \times 2000 = 398000$ points. Lot of operations!

In this case, I obtain $\rho = -0.7139(15)$.

One could also (incorrectly) assume that each of the $\langle \cdot \rangle$'s in ρ are statistically independent. Then I would obtain, using standard independent error propagation,[5] $\rho = -0.7139(67)$. The error is 4 times too large!

[5] Independent errors: $\delta^2 f(a,b) = [\partial f/\partial a]^2 \delta^2 a + [\partial f/\partial b]^2 \delta^2 b$

8.3 Bootstrap

Bootstrap method is closely related to jackknife, but it mimics the resampling more closely. It works as follows:

1. Divide data $(a, b)_i$ into M blocks, with block length $\gg \tau$ (independent blocks).

2. From the set of the M blocks, pick randomly M blocks, not trying to avoid double sampling. Thus, some blocks may not get selected at all, some once, some twice etc.

3. Calculate the quantity of interest over the selected data – for example, the correlation coefficient ρ^*.

4. Repeat steps 2 and 3 a large number of times, say N_B times, each time using a statistically independent selection of the blocks to generate the bootstrap sample. The new correlation coefficients are $\rho_1^*, \rho_2^*, \ldots, \rho_{N_B}^*$.

5. Find values a and b so that these bracket the central 68% of the ρ^* values:

   $$\frac{\#\{\rho_i^* < a\}}{N_B} = 0.16 \qquad\qquad \frac{\#\{\rho_i^* > b\}}{N_B} = 0.16$$

   In effect, by generating a large number (N_B) of bootstrap samples one is generating the distribution function of the final result. The values a and b define the "1-σ" cumulants of this distribution, between which 68% of the probability mass resides. For gaussian distribution this gives directly the gaussian variance.

6. The bootstrap estimate of the standard deviation can be now given as

   $$\delta\rho = \frac{b - a}{2} .$$

   Or, more accurately, we can give asymmetric errors

   $$\rho = \rho_0{}^{+(b-\rho_0)}_{-(\rho_0-a)}$$

   where $\rho_0 = \langle\rho\rangle$.

7. The points 5-6 above can be usually substituted with the bootstrap estimate of the standard deviation:

   $$\delta\rho = \sqrt{\frac{1}{N_B - 1}\sum_i (\rho_i^* - \langle\rho^*\rangle)^2}$$

   where $\langle\rho^*\rangle = \sum_i \rho_i^*/N_B$.

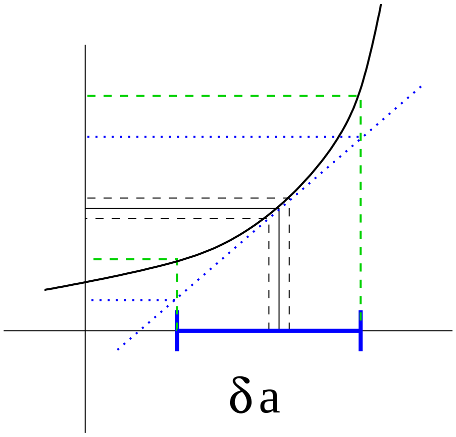
In jackknife/bootstrap literature there is no initial blocking – successive datapoints are considered to be statistically independent (no autocorrelations)! However, this is a minor modification: by blocking initially we obtain statistically independent measurements.

What is a sufficient number of blocks? One should have at least several tens of samples, preferably hundreds. N_B in the bootstrap can go to $\sim$ 1000.

Note: let us have measurements a_i, with expectation value $\langle a \rangle$ and "error" δa. Let us block this into M "normal" blocks a_m, jackknife block averages $a_{(m)}$ and N_B bootstrap averages a_i^*. Now

1. $a_m - \langle a \rangle \sim \delta a \sqrt{M}$
2. $a_{(m)} - \langle a \rangle \sim \delta a / \sqrt{M}$
3. $a_i^* - \langle a \rangle \sim \delta a$

Thus, jackknife "resamples" the distribution on a very narrow range, whereas bootstrap gives the "right" range.
For example, we may have a function $f(\langle a \rangle)$ which happens to be significantly non-linear in range δa.
For jackknife, $\delta f \sim |f'(\langle a \rangle)| \delta a$.



Example: reweighting

$$\langle O \rangle_\beta = \frac{\sum_i O_i \, e^{-(\beta-\beta_c)E_i}}{\sum_i e^{-(\beta-\beta_c)E_i}}$$

If the number of bootstrap samples is very large, it can become expensive to evaluate the sums over the measurements (for each bootstrap block, we have a sum over the full number N measurements!). This can be optimized by precalculating sums over the original blocks.

Bootstrap and jackknife are also used in *fits* to the data (not discussed here).

8.4 Bias estimation

- For concreteness, let us again have a function $f(\langle a \rangle)$. The true value $f(\langle a \rangle)$ is obtained when we calculate the average of a using infinitely many samples a_i.
- Let us now denote the average over N samples $a_1 \ldots a_N$ as $\langle a \rangle_N$, from which we obtain an estimate $f(\langle a \rangle_N)$ Now, if we take the *average over many samples of size* N, we can calculate the expectation value $\langle f(\langle a \rangle_N) \rangle$.

- **Bias** is now defined as

$$\text{Bias} = \langle f(\langle a\rangle_N)\rangle - f(\langle a\rangle) \tag{5}$$

  i.e. the deviation of the expectation value using N measurements from the one using ∞ measurements.

- Note that the bias is a well-defined quantity, the expectation values do not have "errors" (like a single $f(\langle a\rangle_N)$ has).
- Many quantities do not have a bias – for example, simple observables like magnetization M. However, a function of $\langle M\rangle$ like $\langle M\rangle^2$ has a bias!
- The original purpose for the jackknife method was bias reduction (Quenouille, J.Roy.Statist.Soc.Ser.B, 1949).

- Calculate jackknife block averages $a_{(n)}$ from data a_i, and let $f_{(n)} = f(a_{(n)})$ (let me assume here that the original data a_i are already "blocked", if needed).
- Let $\langle f_{(\cdot)}\rangle = \frac{1}{N}\sum_n f_{(n)}$.
- Now the Queinouille estimate for bias is

  $\text{Bias} = (N-1)(\langle f_{(\cdot)}\rangle - f(\langle a\rangle))$.

- This leads to the **bias-corrected jackknifed estimate** of f:

$$\tilde{f}(\langle a\rangle) = f(\langle a\rangle) - \text{Bias} = Nf(\langle a\rangle) - (N-1)\langle f_{(\cdot)}\rangle \tag{6}$$

- The usual rationale for bias correction goes as follows: if we assume that

$$\langle f(\langle a\rangle_N)\rangle = f(\langle a\rangle) + \frac{c_1}{N} + \frac{c_2}{N^2} + \ldots \tag{7}$$

  Now, because each jackknife blocked set contains only $N-1$ measurements,

$$\langle f_{(\cdot)}\rangle = \langle f(\langle a\rangle_{N-1})\rangle. \tag{8}$$

  Now we can eliminate the "error" term $\propto 1/N$, giving us the bias correction result.

- While in principle the bias correction can be performed, in practice it is often of very limited use. Usually the bias correction is completely overwhelmed by the statistical error of the sample.

9 Example: 2nd order phase transition and finite size scaling

One of the most common physical problems studied in simulations are *phase transitions* in various forms (ferromagnetism, Ising model, crystal melting, QCD . . .).

Most (but not all!)[6] phase transitions can be described by an **order parameter**. Mathematically, this is zero in one phase (usually called the disordered phase), non-zero in the other phase (ordered phase). Thus, it cannot be an analytic function at the transition point. (Examples: magnetization in Ising model, Polyakov line in Ising gauge).

Normally, transitions are either 1st or 2nd order. The name comes from the number of derivatives of the free energy $F = -T \log Z$ we need before we see discontinuous behaviour.

- F itself (zeroth derivative) is always continuous.
- First order — the order parameter (and almost any thermodynamical quantity) has a discontinuous jump:

  – latent heat: discontinuity in energy density
- Second order — second derivatives of F, i.e. various susceptibilities are divergent.

As a concrete example, the Ising model partition function with external field H and $\beta = 1/T$

$$Z = \sum_{\{s_x=\pm 1\}} \exp\left[-\beta\left(\tfrac{1}{2}\sum_{<xy>}(1-s_x s_y) + H\sum_x s_x\right)\right] \tag{9}$$

gives magnetization M and magnetic susceptibility χ_M as 1st and 2nd derivatives wrt. H:

$$\begin{aligned} M &= \frac{1}{V}(dF/dH)_{H=0} \\ \chi_M &= V(\langle M^2\rangle - \langle M\rangle^2) = \frac{1}{V}(d^2F/dH^2)_{H=0} \end{aligned}$$

(limit $V \to \infty$ implied here.)

[6] A common-day transition without an order parameter is the liquid-vapour 1st order phase transition, for example, boiling of water. There is no exact order parameter, and the two phases can be analytically connected. The transition line ends in a *critical point*, where the transition has 3-dim. Ising model universal behaviour.

Second order transitions are classified by their **critical exponents**, which characterize the behaviour at the critical point. The most important here are

$$\begin{array}{lrcl} \text{Magnetization} & M & \sim & |T-T_c|^{\beta} \\ \text{Mag. susceptibility} & \chi_M & \sim & |T-T_c|^{-\gamma} \\ \text{Heat capacity} & C_V = \frac{1}{V}\frac{d\langle E\rangle}{dT} & \sim & |T-T_c|^{-\alpha} \\ \text{Correlation length} & \xi & \sim & |T-T_c|^{-\nu} \end{array} \tag{10}$$

For the 2d Ising model, these exponents are known exactly: $\alpha = 0$, $\beta = 0.125$, $\gamma = 1.75$, $\nu = 1$.

However, as already mentioned before, on a finite lattice we have finite number of degrees of freedom and everything is analytic! This causes several problems:

- What is a good order parameter? The order parameters are always either zero or non-zero. For example, both

  $$M = \left\langle \frac{1}{V} \sum_i s_i \right\rangle \qquad\qquad |M| = \left\langle \left| \frac{1}{V} \sum_i s_i \right| \right\rangle$$

  are real order parameters in infinite volume, but $\langle M \rangle = 0$ and $\langle |M| \rangle > 0$ on a finite lattice. Of these, $\langle |M| \rangle$ is usable, since it is almost zero in the symmetric phase.

- How to locate the true phase transition?

- How to measure the critical exponents?

9.1 Finite size scaling in Ising model

For concreteness, let us consider Ising model. In infinite volume, the correlation length ξ (domain size) diverges near the transition point as

$$\xi \propto |t|^{-\nu}, \tag{11}$$

with $t = T - T_c \approx \beta_c - \beta$.

However, because the system in simulations has a finite size L^d, when the correlation length is $\xi \approx L$, the system already becomes *effectively ordered*. Thus, we can argue that the system has a **pseudocritical point** when

$$[\beta_c(\infty) - \beta_c(V)]^{-\nu} \propto L \quad \Rightarrow \quad \beta_c(V) = \beta_c(\infty) - \text{const.} \times L^{-1/\nu} \tag{12}$$

How to locate this point (if we don't know ν or $\beta_c(\infty)$)? Consider, for example, magnetic susceptibility which diverges in infinite volume as

$$\chi_{|M|} = \frac{1}{V}(\langle M^2 \rangle - \langle |M| \rangle^2) \propto |t|^{-\gamma} \tag{13}$$

We can now (somewhat arbitrarily) argue that on a finite volume $\chi_{|M|}$ has a *maximum* at the pseudocritical point $\beta_c(V)$. At this point the maximum value should be

$$\chi_{M,\max} \propto (\beta_c(V) - \beta_c(\infty))^{-\gamma} \propto L^{\gamma/\nu} \tag{14}$$

The above (not extremely robust) argument gives us a prescription how to determine the true critical point $\beta_c = \beta_c(\infty)$ and even to estimate the critical exponents ν and γ:

1. Using various volumes V, locate the maximum of $\chi_{|M|}$.
2. Make a (power law) fit to the maximum location of $\chi_{|M|}$:

   $$\beta_{\max} = \beta_c - c_1 \times L^x \tag{15}$$

   Fit has 3 parameters, β_c, c_1, x, where x should be equal to $-1/\nu$.
3. The exponent γ/ν can be estimated from the maximum value $\chi_{\max} \propto L^{\gamma/\nu}$.

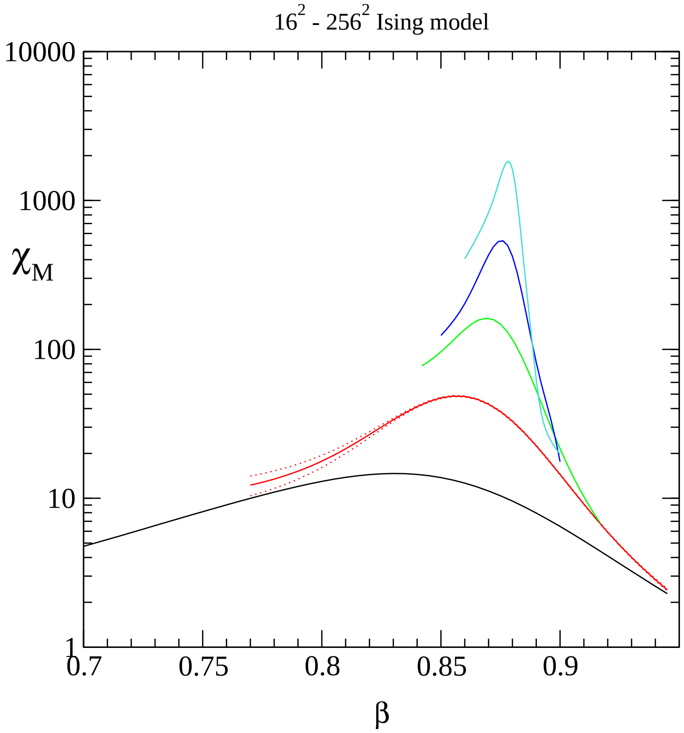
WARNING: while this process gives pretty good estimate of the infinite volume critical point β_c, the exponents can be more difficult to obtain reliably. This is especially so if we would use the heat capacity C_V instead of $\chi_{|M|}$, this is due to the fact that the critical exponent α is usually much smaller than γ.

NOTE:
Above, somewhat misleadingly, we actually used susceptibility $\chi_{|M|} = \frac{1}{V}(\langle M^2\rangle - \langle|M|\rangle^2)$. This is strictly speaking not equivalent to the ’true’ magnetic susceptibility $\chi_M = \frac{1}{V}(\langle M^2\rangle - \langle M\rangle^2)$. In the broken phase these are equal, but in the symmetric phase these differ by a constant. The critical exponents are equal, however.

9.2 Example: 2d Ising model results

2d Ising model with volumes 16^2, 32^2, 64^2, 128^2 and 256^2. $\chi_{|M|}$ reweighted to a range of β-values around the critical point:

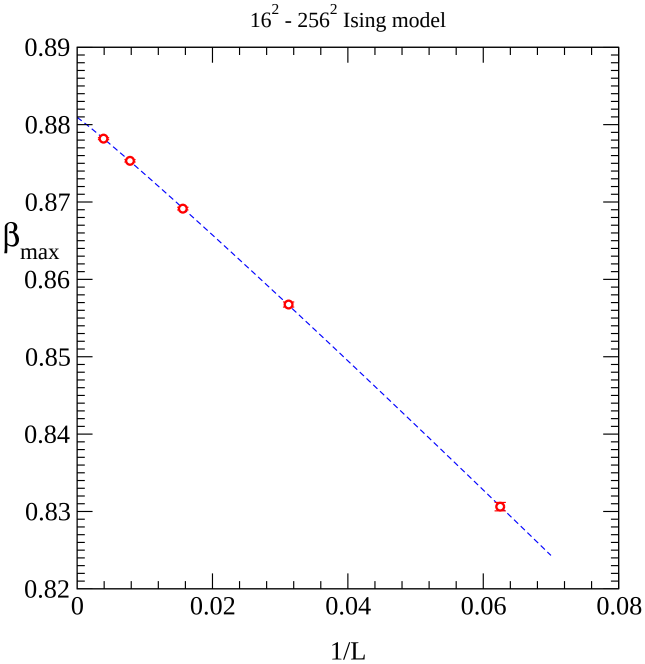
The peak of $\chi|M|$ clearly grows (like L^z, note log-scale) and the location moves to larger β.



Power law fit to the location of the maxima: $\beta_{max} = \beta_c - cL^x$.
(χ^2/d.o.f $= 0.33/2$, confidence level 0.84.)

Results:
$\beta_c = 0.88093(24)$; $x = -1.05(2)$.

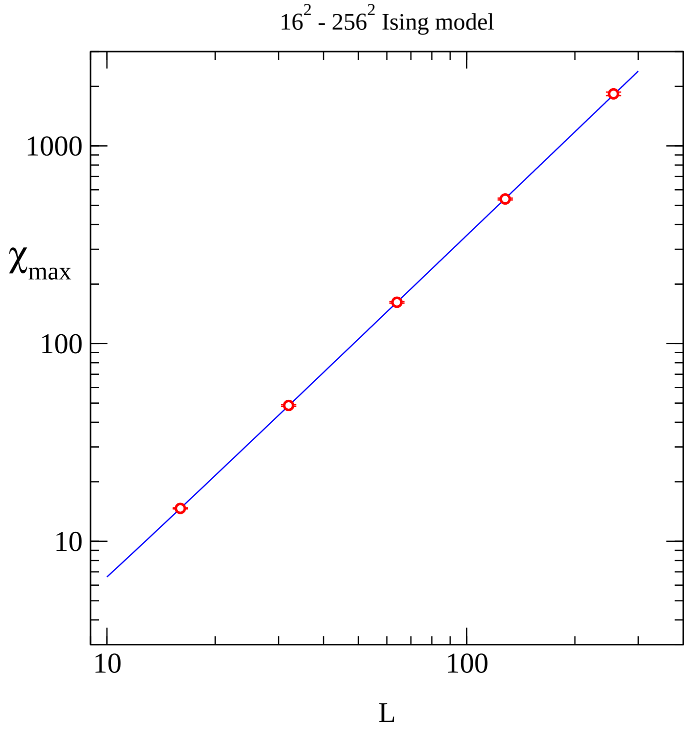
Right results:
$\beta_c = \ln(1+\sqrt{2}) \approx 0.88137$
$x = -1/\nu = -1$.
We are very close, but still $\sim 2\sigma$ off the correct values. This is most likely due to too small volume (16^3).



If we drop 16^3 and fix the exponent $x = -1$, we obtain $\beta_c = 0.88132(13)$, which is perfectly compatible with the right result.

Power law fit to the value of the maximum: $\chi_{\max} = cL^z$.
If we exclude 16^2, we obtain
$z = 1.740(8)$,
which is compatible with the right value
$z = \gamma/\nu = 1.75$

(Using also 16^2 makes the fit a bit worse.)



9.3 Critical exponent β

Finite size scaling can be used to determine other exponents too: for example, let us consider the spin-spin correlation function just at the critical point:

$$\langle s_x s_y \rangle = G(|x-y|) \propto |x-y|^{-(d-2+\eta)}, \qquad |x-y| \to \infty, \tag{16}$$

where d is the dimensionality of the system and η is the "anomalous" exponent of the correlation length (for 2d Ising, $\eta = 1/4$(?)).

We define the root mean square magnetization

$$M_{\rm rms} = \sqrt{\langle M^2 \rangle} = \sqrt{\sum_{x,y} s_x s_y / V^2}. \tag{17}$$

At the infinite volume critical temperature T_c (or β_c), we can approximate behaviour of the correlation function on a *finite volume* as

$$\sum_x \langle s_x s_y \rangle \propto \int_0^{L/2} dr\, r^{d-1}\, G(r) \propto \int_0^{L/2} dr\, r^{1-\eta} \propto L^{2-\eta}. \tag{18}$$

Thus, $M_{\rm rms}$ becomes

$$M_{\rm rms}^{T=T_c} \propto \sqrt{L^{2-d-\eta}} \propto L^{-\beta/\nu}. \tag{19}$$

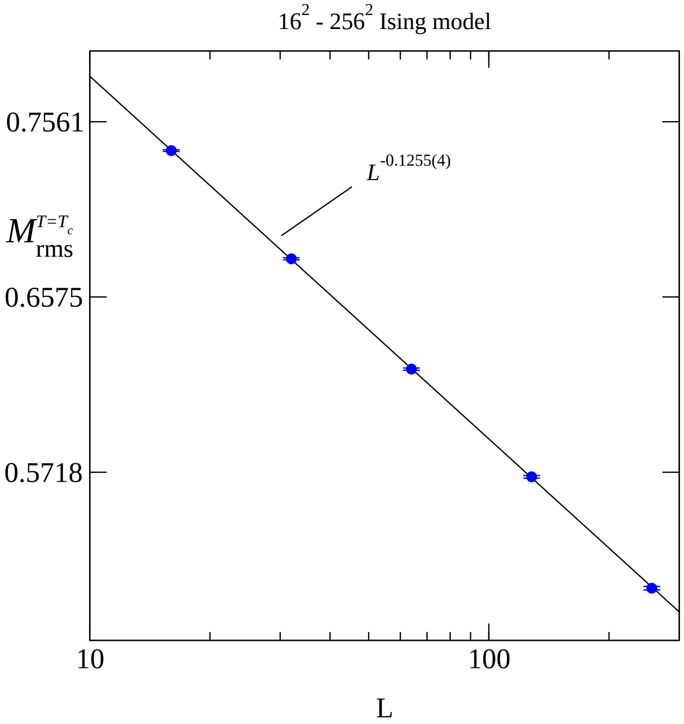
In the last stage we used scaling law $2 - \eta = \gamma/\nu$ and the so-called *hyperscaling* $d\nu = 2\beta + \gamma$.

Hyperscaling is not always valid: this happens, for instance, in systems above their marginal dimensionality d^* where the mean field values for the critical exponents become valid. For simple spin models with local action (Ising, for example) $d^* = 4$.

Example: 2d Ising, again, and determine $M_{\rm rms}$ from lattices $16^2 - 256^2$.

Power law fit: $M_{\rm rms} = cL^z$.
The result from the fit is
$z = -0.1255(4)$,
which is compatible with the known value
$z = -\beta/\nu = -0.125$

[Note: the plot is of log-log type]


16^2 - 256^2 Ising model
$M_{\rm rms}^{T=T_c}$
$L^{-0.1255(4)}$
L