Chapter 1

Variational methods

(Lesson 21 of 27/04/20)
Compiled: May 18, 2020

Exactly solvable models are rare. For example, the Ising Model, describing in a very simplified manner a discrete set of local interacting binary variables, has been exactly solved only for $d = 1$ in general, and for $d = 2$ only in absence of an external field ($h = 0$). The latter, in particular, requires long and sophisticated derivations.
Even for other models, the trend is the same: whenever we wish to study *emergent phenomena* the problem usually becomes analytically intractable.

One possibility is then to resort to **numerical simulations**. However, these are often time-consuming, require significant computational power, and can be hard to interpret - as interesting "high level" characteristics (such as the conditions for phase transitions) are drowned in lots of irrelevant "low-level" data.

So we may resort to **approximate computations** instead. The idea is to find a simple model that is able to capture, at least *qualitatively*, features from a more complex one, while still admitting an exact solution. This can then give hints on *what to look for* in a full numerical simulation, thus allowing a deeper understanding.

One quick way to compute approximations is through **variational methods**. In essence, we consider some parametrized pdf $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and tweak the parameters $\boldsymbol{\theta}$ so that it becomes "closer and closer" to the target pdf $f(\boldsymbol{x})$ of the full model. If we choose a sufficiently *simple* form for $f_{\boldsymbol{\theta}}$, we will be able to perform exact computations, while still retaining some sort of "correspondance" with the more complex model.

In the following, we will first introduce a notion of "**distance**" between pdf*s* (**relative entropy**), giving a mathematical meaning to the notion of "closeness" between probability distributions. Then we will explicitly state the *variational method* as a **minimization problem**, and, using the Ising Model as an example, we will see a popular choice for the parametrization of $f_{\boldsymbol{\theta}}$: the **mean-field approximation**.

1.0.1 Relative Entropy

Given two (discrete) probability distributions $\{p_i\}_{i\in\mathcal{D}}$ and $\{q_i\}_{i\in\mathcal{D}}$, with $p_i, q_i > 0$ and $\sum_i p_i = \sum_i q_i = 1$, we define the **relative entropy** (or Kullback–Leibler divergence) of $\{p_i\}$ with respect to $\{q_i\}$ as follows:

$$S_R(\{p_i\},\{q_i\}) = -\sum_{i\in\mathcal{D}} p_i \ln\frac{p_i}{q_i} \leq 0 \tag{1.1}$$

In a sense, relative entropy measures the *closeness* between the two distributions - as it is maximum ($S_R = 0$) when the two coincide, i.e. $p_i = q_i\ \forall i$. Note, however, that S_R is not a *distance function* in the proper sense, as it does not satisfy the triangular inequality.

Proof that $S_R \leq 0$

The fact that $S_R = 0$ is the maximum point of S_R, i.e. $S_R \leq 0$, can be proven as follows. First we define an auxiliary function $f(x)$ over $(0,\infty)$:

$$f(x) = -x\ln x \qquad x > 0$$

Such function $f(x)$ is **concave**. In fact:

$$f'(x) = -1 - \ln x$$
$$f''(x) = -\frac{1}{x} < 0 \qquad x > 0$$

So, we may apply Jensen's inequality. For any choice of a set of non-negative numbers $\{\lambda_i\}$ summing to 1, the following relation holds:

$$f\left(\sum_i \lambda_i x_i\right) \geq \sum_i f(x_i)\lambda_i \qquad \sum_i \lambda_i = 1 \ \wedge\ \lambda_i \geq 0$$

And letting $\lambda_i = q_i$ and $x_i = p_i/q_i$ completes the proof:

$$S_R = \sum_i q_i f\left(\frac{p_i}{q_i}\right) \leq f\left(\sum_i \cancel{q_i}\frac{p_i}{\cancel{q_i}}\right) = f(1) = 0$$

with the equality holding if and only if $p_i = q_i$.

1.0.2 Approximation as an optimization problem

Let's consider, for simplicity, a system with **discrete** states $\{\boldsymbol{\sigma_i}\}_{i\in\mathcal{D}}$, each with energy $\mathcal{H}(\boldsymbol{\sigma_i})$, and an associated probability q_i given by a Boltzmann distribution:

$$\rho(\boldsymbol{\sigma_i}) \equiv q_i = \frac{e^{-\beta\mathcal{H}(\boldsymbol{\sigma_i})}}{Z} = e^{-\beta(\mathcal{H}(\boldsymbol{\sigma})-F)} \qquad Z = \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta\mathcal{H}(\boldsymbol{\sigma})} \equiv e^{-\beta F}$$

where F is the system's **free energy** function.

In general, the $\{q_i\}$ are difficult to explicitly compute, because Z is generally a sum over a huge number of terms (2^V in the case of the Ising Model) with no analytical form.

So, the idea is to approximate ρ with another "easier" distribution ρ_0, the **variational ansatz**, which is parametrized as a Boltzmann distribution with a different Hamiltonian $\mathcal{H}_0$ (and so also a different free energy F_0):

$$\rho_0(\boldsymbol{\sigma_i}) \equiv p_i = \frac{e^{-\beta\mathcal{H}_0(\boldsymbol{\sigma_i})}}{Z_0} = e^{-\beta(\mathcal{H}_0(\boldsymbol{\sigma})-F_0)} \qquad Z_0 = \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta\mathcal{H}_0(\boldsymbol{\sigma})} \equiv e^{-\beta F_0} \tag{1.2}$$

The *closeness* of $\{p_i\}$ to $\{q_i\}$ is given by their **relative entropy** (**??**):

$$\begin{aligned} 0 \leq \sum_i p_i \ln \frac{p_i}{q_i} &= \sum_{\{\boldsymbol{\sigma}\}} \frac{e^{-\beta\mathcal{H}_0(\boldsymbol{\sigma})}}{Z_0} \ln \frac{e^{-\beta\mathcal{H}_0(\boldsymbol{\sigma})}}{\underbrace{Z_0}_{e^{-\beta F_0}}} \frac{\overbrace{Z}^{e^{-\beta F}}}{e^{-\beta\mathcal{H}(\boldsymbol{\sigma})}} = \\ &= \frac{1}{Z_0} \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta H_0(\boldsymbol{\sigma})} \beta[\mathcal{H}(\boldsymbol{\sigma}) - \mathcal{H}_0(\boldsymbol{\sigma}) - F + F_0] = \\ &= \beta\langle\mathcal{H} - \mathcal{H}_0\rangle_0 - \beta(F - F_0) \end{aligned} \tag{1.3}$$

where $\langle\cdots\rangle_0$ denotes the average according to the ansatz distribution:

$$\langle f(\boldsymbol{\sigma})\rangle_0 \equiv \frac{1}{Z_0} \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta\mathcal{H}_0(\sigma)} f(\boldsymbol{\sigma})$$

The expression (**??**) is called the **Gibbs-Bogoliubov-Feynman inequality**[1], and holds as an equality if and only if $\rho = \rho_0 \Leftrightarrow \mathcal{H} = \mathcal{H}_0$.

Rearranging (**??**):

$$\beta F \leq \beta F_0 + \beta\langle\mathcal{H} - \mathcal{H}_0\rangle_0 = \beta\langle\mathcal{H}\rangle_0 + \beta(F_0 - \langle\mathcal{H}_0\rangle_0) \tag{1.4}$$

Note that F_0 does not depend on $\boldsymbol{\sigma}$, as it's $\propto \ln Z_0$, and so we can bring it inside the average, and expand it:

$$\beta(F_0 - \langle\mathcal{H}_0\rangle_0) = \beta\langle F_0 - \mathcal{H}_0\rangle_0 = \sum_{\{\boldsymbol{\sigma}\}} \rho_0(\boldsymbol{\sigma})\, \beta(F_0 - \mathcal{H}_0(\boldsymbol{\sigma}))$$

Then, from (**??**) note that:

$$\rho_0(\boldsymbol{\sigma}) = e^{-\beta(\mathcal{H}_0(\boldsymbol{\sigma})-F_0)} \Rightarrow \ln\rho_0(\boldsymbol{\sigma}) = \beta(F_0 - \mathcal{H}_0(\boldsymbol{\sigma}))$$

and substituting above:

$$\beta(F_0 - \langle\mathcal{H}_0\rangle_0) = -\frac{1}{k_B} \underbrace{\left(-k_B \sum_{\{\boldsymbol{\sigma}\}} \rho_0(\boldsymbol{\sigma}) \ln\rho_0(\boldsymbol{\sigma})\right)}_{S[\rho_0]} = -\frac{S[\rho_0]}{k_B} \tag{1.5}$$

where $S[\rho_0]$ is the **information entropy** of ρ_0:

$$S[\rho_0] = -k_B \sum_{\{\boldsymbol{\sigma}\}} \rho_0(\boldsymbol{\sigma}) \ln\rho_0(\boldsymbol{\sigma})$$

[1] ∧Physically, it is completely equivalent to the second law of thermodynamics.

Thus, substituting (**??**) back in the inequality (**??**) leads to:

$$\beta F \leq \beta \langle \mathcal{H} \rangle_0 - \frac{S[\rho_0]}{k_B} = \beta \langle \mathcal{H} \rangle_0 - \beta T S[\rho_0] \tag{1.6}$$

And dividing by β:

$$F \leq F_V \equiv \langle \mathcal{H} \rangle_0 - T S[\rho_0]$$

where F_V is called the **Variational Free Energy** (VFE).
So, the true free energy F is always less or equal to the variational one F_V. An optimal estimate of F is obtained by minimizing F_V with respect to ρ_0.

Clearly, if we do not require any constraint on ρ_0, thus allowing arbitrary complexity, then the minimum is obtained when $\rho_0 = \rho$: the most accurate approximation of a model is the model itself. Realistically ρ is mathematically intractable, and we need to *bound* the "complexity" of ρ_0, with the effect that it won't be able to perfectly replicate ρ, and so the minimum for F_V will be larger than F (but hopefully still somewhat close).

One possible way to constrain the "complexity" of ρ_0 is to *force it* to be separable:

$$\rho_0(\boldsymbol{\sigma}) = \prod_x \rho_x(\sigma_x) \tag{1.7}$$

In this way, all degrees of freedom of the system become **decoupled**. In a sense, correlations and complex behaviours are "averaged" between each component - and in fact the approximation in (**??**) is known as the **mean field** ansatz.

1.1 Mean Field Ising Model

Consider a d-dimensional nearest-neighbour Ising Model, where we allow each spin to interact with a **local** magnetic field b_x, leading to the Hamiltonian:

$$\mathcal{H}(\boldsymbol{\sigma}) = -J \sum_{\langle x,y \rangle} \sigma_x \sigma_y - \sum_x b_x \sigma_x$$

To understand its behaviour, we use the **mean-field** approximation (**??**), and choose a parametrization inspired by the non-interacting Ising Model (**??**, pag. **??**):

$$\rho_0(\boldsymbol{\sigma}) = \prod_x \rho_x(\sigma_x) \qquad \rho_x(\sigma_x) = \frac{1 + m_x \sigma_x}{2} \quad m_x \in [-1, 1] \tag{1.8}$$

where the $\{m_x\}$ are the *variational parameters* that will be *tweaked* to make $\rho_0(\boldsymbol{\sigma})$ closer to the real probability distribution $\rho(\boldsymbol{\sigma})$ of the Ising Model, by minimizing the **variational free energy** F_V. The constraint $m_x \in [-1, 1]$ comes from requiring all probabilities to be non-negative $\rho_x(\sigma_x) \geq 0$.
Before proceeding, note that (**??**) is already normalized:

$$\sum_{\sigma_x = \pm 1} \rho_x(\sigma_x) = \frac{1 + m_x}{2} + \frac{1 - m_x}{2} = \frac{1}{2} + \frac{1}{2} = 1$$

and that each *variational parameter* m_x corresponds to the **local magnetization** of spin σ_x *in the mean-field model*:

$$\begin{aligned}\langle\sigma_x\rangle_0 &= \sum_{\{\boldsymbol{\sigma}\}} \rho_0(\boldsymbol{\sigma})\sigma_x = \sum_{\{\boldsymbol{\sigma}\}} \prod_y \frac{1+m_y\sigma_y}{2}\sigma_x = \\ &\underset{(a)}{=} \sum_{\sigma_x=\pm1} \left(\prod_{y\neq x} \underbrace{\sum_{\sigma_y=\pm1} \frac{1+m_y\sigma_y}{2}}_{1} \right) \frac{1+m_x\sigma_x}{2}\sigma_x = \\ &= \sum_{\sigma_x=\pm1} \sigma_x \frac{1+m_x\sigma_x}{2} = \frac{1+m_x}{2} - \frac{1-m_x}{2} = m_x \end{aligned} \tag{1.9}$$

where in (a) we split the product in the case $y \neq x$ and $y = x$. Also note that the average is over ρ_0 and not the "true" pdf ρ.

Choice of parametrization. The distribution $\rho_x(\sigma_x)$ in (**??**) is the most general discrete distribution for a binary variable such as σ_x, just rewritten to highlight the average m_x.
In fact, consider a generic **binary** variable σ. Its distribution is:

$$\mathbb{P}[\sigma = +1] = p_+ \qquad \mathbb{P}[\sigma = -1] = p_-$$

Due to normalization, $p_+ + p_- = 1$, and so there is only **one free parameter** needed to completely specify the pdf:

$$\mathbb{P}[\sigma = +1] = p \qquad \mathbb{P}[\sigma = -1] = 1 - p$$

If we then rewrite p as function of the average $\langle\sigma\rangle = m$, we get:

$$m = \sum_{\sigma=\pm1} \sigma\mathbb{P}[\sigma] = p - (1-p) = 2p + 1 \Rightarrow p = \frac{1+m}{2}$$

And so:

$$\mathbb{P}[\sigma = +1] = \frac{1+m}{2} \qquad \mathbb{P}[\sigma = -1] = \frac{1-m}{2}$$

Which can be rewritten more compactly as:

$$\rho(\sigma) = \frac{1+m\sigma}{2}$$

So we are not making any additional hypothesis other than that of a separable $\rho(\boldsymbol{\sigma})$ (given by the mean field approximation).

For simplicity, we work with βF_V, denoting $\beta J \equiv K$ and $\beta b_x \equiv h_x$. From the variational principle (**??**):

$$\beta F \leq \min_{\boldsymbol{m}} \beta F_V(\boldsymbol{m}, \boldsymbol{h}) = \min_{\boldsymbol{m}} \left(\beta\langle\mathcal{H}\rangle_0 - \frac{S[\rho_0]}{k_B} \right) \tag{1.10}$$

The average of $\mathcal{H}$ according to the ansatz is:

$$\langle\mathcal{H}\rangle_0 = \langle -J \sum_{\langle x,y\rangle} \sigma_x\sigma_y - \sum_x b_x\sigma_x\rangle_0 = -J \sum_{\langle x,y\rangle} \langle\sigma_x\sigma_y\rangle_0 - \sum_x b_x\langle\sigma_x\rangle_0$$

We already computed $\langle\sigma_x\rangle_0 = m_x$ in (**??**). For the two-point correlation, as ρ_0 is separable and thus σ_x and σ_y are decoupled, we get:

$$\langle\sigma_x\sigma_y\rangle_0 = \langle\sigma_x\rangle_0\langle\sigma_y\rangle_0 = \sum_{\sigma_x}\frac{1+m_x\sigma_x}{2}\sigma_x\sum_{\sigma_y}\frac{1+m_y\sigma_y}{2}\sigma_y = m_x m_y$$

Thus:

$$\langle\mathcal{H}(\boldsymbol{\sigma})\rangle_0 = -J\sum_{\langle x,y\rangle} m_x m_y - \sum_x b_x m_x = \mathcal{H}(\boldsymbol{m}) \tag{1.11}$$

This is valid more in general when applying the mean field approximation to even more complex Hamiltonians, as it is a consequence of the separability of ρ_0.

On the other hand, the entropy of ρ_0 can be directly computed. Noting that $\rho_x(\sigma_x)$ is exactly the same pdf we used in the non-interacting Ising Model, we can borrow the results (**??**) and (**??**, pag. **??**) from there:

$$\begin{aligned}-\frac{S[\rho_0]}{k_B} &= \sum_{\{\boldsymbol{\sigma}\}}\rho_0(\boldsymbol{\sigma})\ln\rho_0(\boldsymbol{\sigma}) = \sum_x\sum_{\sigma_x}\frac{1+m_x\sigma_x}{2}\ln\frac{1+m_x\sigma_x}{2} = \\ &= \sum_x\left(\frac{1+m_x}{2}\ln\frac{1+m_x}{2}+\frac{1-m_x}{2}\ln\frac{1-m_x}{2}\right)\equiv\sum_x s_0(m_x)\end{aligned} \tag{1.12}$$

where we defined a *local entropy* s_0 as:

$$s_0(m) \equiv \frac{1+m}{2}\ln\frac{1+m}{2}+\frac{1-m}{2}\ln\frac{1-m}{2}$$

Substituting these results (**??**) and (**??**) back in (**??**) we arrive to:

$$\begin{aligned}\beta F_V(\boldsymbol{m},\boldsymbol{h}) &= \beta H(\boldsymbol{m}) + \sum_x s_0(m_x) = \\ &= -K\sum_{\langle x,y\rangle} m_x m_y - \sum_x h_x m_x + \sum_x\left[\frac{1+m_x}{2}\ln\frac{1+m_x}{2}+\frac{1-m_x}{2}\ln\frac{1-m_x}{2}\right]\end{aligned} \tag{1.13}$$

where the first line holds for a generic Hamiltonian $\mathcal{H}(\boldsymbol{\sigma})$, and the second is specific for the Ising Model we are studying.

Then, we minimize $F_V(\boldsymbol{m},\boldsymbol{h})$ with respect to $\boldsymbol{m}$, denoting the minimum as $F_V(\boldsymbol{M},\boldsymbol{h})$:

$$\frac{\partial}{\partial m_x}\beta F_V\Big|_{\boldsymbol{m}=\boldsymbol{M}} \stackrel{!}{=} 0 \tag{1.14}$$

$$\begin{aligned}0 &\stackrel{!}{=} \frac{\partial}{\partial m_x}\left[-K\sum_{\langle x,y\rangle} m_x m_y - \sum_x h_x m_x + \sum_x\left(\frac{1+m_x}{2}\ln\frac{1+m_x}{2}+\frac{1-m_x}{2}\ln\frac{1-m_x}{2}\right)\right]_{\boldsymbol{m}=\boldsymbol{M}} = \\ &= -K\sum_{y\in\langle x,y\rangle} M_y - h_x + \frac{1}{2}\ln\frac{1+M_x}{2} + \cancel{\frac{1+M_x}{2}\frac{2}{1+M_x}\frac{1}{2}} - \frac{1}{2}\ln\frac{1-M_x}{2} - \cancel{\frac{1-M_x}{2}\frac{2}{1-M_x}\frac{1}{2}} = \\ &= -K\sum_{y\in\langle x,y\rangle} M_y - h_x + \frac{1}{2}\ln\left(\frac{1+M_x}{\cancel{2}}\frac{\cancel{2}}{1-M_x}\right)\end{aligned}$$

where the sum is over all nodes y neighbouring x, i.e. the ones included in some pair of neighbours $\langle y, x\rangle$ involving x.
Using the identity (**??**, pag. **??**)

$$\tanh^{-1} M_x = \frac{1}{2} \ln \frac{1+M_x}{1-M_x}$$

and rearranging leads to:

$$M_x(\boldsymbol{h}, K) = \tanh \left[K \sum_{y \in \langle y,x \rangle} M_y + h_x \right] \tag{1.15}$$

1.1.1 Physical meaning of the variational parameters M_x

It would be interesting to associate some physical meaning to the variational solution, and in particular understand what the M_x represent.

So, we found that:

$$\min_{\boldsymbol{m}} F_V(\boldsymbol{m}, \boldsymbol{h}) \equiv F_V(\boldsymbol{M}, \boldsymbol{h})$$

with the $\boldsymbol{M}$ given by solving the N equations (**??**), one for each node.

The *magnetization* given by the variational free energy is:

$$\langle \sigma_x \rangle_V \underset{(??)}{=} -\frac{\partial}{\partial h_x} [\beta F_V(\boldsymbol{M}, \boldsymbol{h})] = -\beta \left[\underbrace{\sum_y \frac{\partial F_V}{\partial m_y}(\boldsymbol{m}, \boldsymbol{h})}_{0\ (??)} \frac{\partial m_y}{\partial h_x} - \underbrace{\frac{\partial F_V}{\partial h_x}(\boldsymbol{m}, \boldsymbol{h})}_{M_x\ (??)} \right]_{\boldsymbol{m}=\boldsymbol{M}} =$$

$$= M_x \tag{1.16}$$

Note that the variational free energy F_V **is not** the *ansatz free energy* F_0, and so $\langle \sigma_x \rangle_V$ and $\langle \sigma_x \rangle_0$ are different averages, and (**??**) should not be confused with (**??**).

So, M_X is the best estimate of the *true magnetization* σ_x, as it is obtained with the F_V *closest* to the real F.

1.1.2 Uniform case

Suppose the magnetic field is uniform $h_x \equiv h$. In this case, the system is **translationally invariant**. So, it is reasonable to consider the *ansatz* where also all the local magnetizations are the same: $m_x \equiv m$, and search for a single value of m.

Given these assumptions, (**??**) becomes:

$$\beta F_V(m, h) = -Km^2 \sum_{\langle x,y \rangle} 1 - mh \sum_x 1 + \left[\frac{1+m}{2} \ln \frac{1+m}{2} + \frac{1-m}{2} \ln \frac{1-m}{2} \right] \sum_x 1$$

Then $\sum_x 1$ is just the number of nodes N, and $\sum_{\langle x,y \rangle} 1$ is the number of possible pairs, which is Nd for a d-dimensional cubic lattice (each node contributes with one pair for every possible *direction*). Dividing by N:

$$\beta \frac{F_V(m, h)}{h} = -Kdm^2 + \frac{1+m}{2} \ln \frac{1+m}{2} + \frac{1-m}{2} \ln \frac{1-m}{2} - hm \tag{1.17}$$

The equation for M_X (**??**) becomes:

$$M(h,K) = \tanh\left[KM \sum_{y\in\langle y,x\rangle} 1 + h\right]$$

The sum is over all *neighbours* of x, which are $2d$ for a d-dimensional cubic lattice (2 for every *direction*), leading to:

$$M(h,K) = \tanh(2dKM + h) \tag{1.18}$$

Let's start with the case of no external field $h = 0$. In this case, the variational free energy (**??**) is an **even** function of m: *Case 1.* $h = 0$

$$F_V(m,0) = F_V(-m,0)$$

We can then study the solutions of (**??**):

$$M = \tanh(2dKM) \qquad M(K,0) \equiv M(K) \tag{1.19}$$

Clearly $M = 0$ is always a solution. Depending on $2dK$, there can be two more solutions, as can be seen by plotting each side and looking for intersections (**??**).

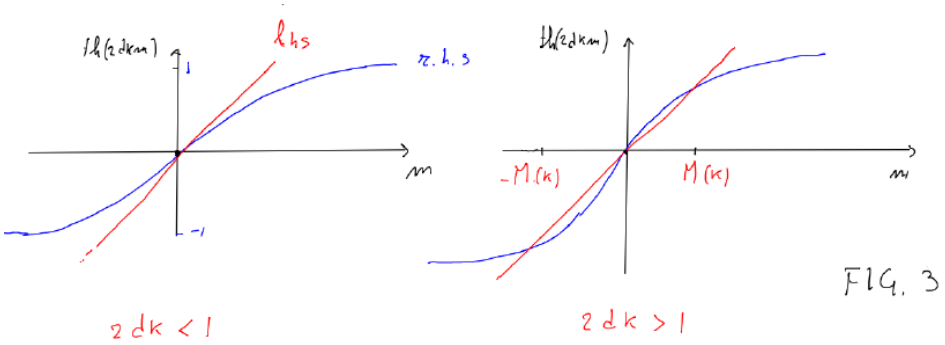

Figure (1.1) – Solutions of (**??**) are intersections of the two curves.

The plots in (**??**) can be obtained by expanding $\tanh x$ in Taylor series around $x = 0$. The first three derivatives are:

$$\frac{\mathrm{d}}{\mathrm{d}x}\tanh x = 1 - \tanh^2 x$$
$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\tanh x = -2\tanh x(1 - \tanh^2 x)$$
$$\frac{\mathrm{d}^3}{\mathrm{d}x^3}\tanh x = -2(1 - \tanh^2 x) + 4\tanh^2 x(1 - \tanh^2 x)$$

So:

$$\tanh x = \tanh 0 + x\frac{\mathrm{d}}{\mathrm{d}x}\tanh x\Big|_{x=0} + \frac{x^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\tanh x\Big|_{x=0} + \frac{x^3}{3!}\frac{\mathrm{d}^3}{\mathrm{d}x^3}\tanh x\Big|_{x=0} + \cdots =$$
$$= x - \frac{2x^3}{3\cdot 2\cdot 1} + O(x^5) = x - \frac{x^3}{3} + O(x^5)$$

For small x, $\tanh x$ is linear, and in particular $\tanh(2dKM)$ is a line passing through the origin with slope $2dK$. It that slope is **less** than te one of $y = M$, i.e. 1, then the only intersection is at $M = 0$ (left of fig. **??**)