On the loss function landscape in the simplest constrained least-square optimization [1]

Yan V Fyodorov

Department of Mathematics

Project supported by the EPSRC grant EP/N009436/1

"XIV Brunel-Bielefeld RMT Workshop " London, 14th of December 2018

[1] Based on: YVF & Rashel Tublin , under preparation.

Background:

The simplest optimization problem of the least-square type on the sphere $x \in \mathbb{R}^N;\ \|x\|^2 = \text{const}$ arises in the Multiple Factor Data Analysis and is known as the Oblique Procrustes Problem:

For a given pair of $M \times N$ matrices A and B find such $N \times N$ matrix X that the equality $B = AX$ holds as close as possible and columns $x_i \in \mathbb{R}^N;\ i = 1;\ldots;N$ are of unit length.

For $M > N$ this system of linear equations is overcomplete and a solution can be found separately for each column x by minimizing the loss/cost function

$$H(x) = \tfrac{1}{2}\|Ax - b\|^2 := \tfrac{1}{2}\sum_{k=1}^{M}\left[\sum_{j=1}^{N} A_{kj}\,x_j - b_k\right]^2; \quad \|x\|^2 = \text{const}$$

The problem was first analysed in that setting by M. W. BROWNE in 1967, and then independently by numerical mathematicians (e.g. W. GANDER 1981) who used the Lagrange multiplier to take care of the spherical constraint. Introducing the Lagrangian $L_{\lambda;s}(x) = H(x) - \frac{\lambda}{2}(x;x)$, with real λ being the Lagrange multiplier, the stationary conditions $\nabla L_{\lambda;s}(x) = 0$ yields linear system:

$$A^T[Ax - b] = \lambda x; \quad \Rightarrow \quad x = (A^TA - \lambda I_N)^{-1}A^Tb$$

Setting of the problem:

The spherical constraint $x^2 = N$ yields the equation for in the form:

$$b^T A \frac{1}{(A^T A \quad I_N)^2} A^T b = N$$

which is equivalent to a polynomial equation of degree $2N$ in . Each real solution for the Lagrange multiplier $_i$ corresponds to a stationary point x_i of the loss function $H(x) = \frac{1}{2}||Ax \quad b||^2$ on the sphere $x^2 = N$ and one can show that the order $_1 < \ _2 < \ldots < \ _N$ implies $H(x_1) < H(x_j) < \ldots < H(x_N)$. Thus the minimal loss is given by $E_{min} = H(x_1)$.

Our goal: To count the stationary points via the Lagrange multipliers

$$_i;\ i = 1;\ldots;N \qquad 2N$$

and eventually nd the minimal loss E_{min} after assuming the entries A_{kj} of $M \quad N;\ M > N$ matrix A to be i.i.d. normal real variables such that $A^T A = W$ is $N \quad N$ Wishart with the probability density

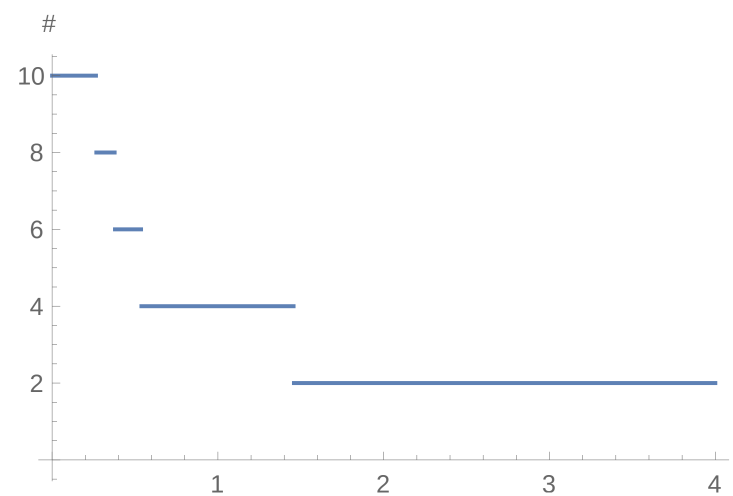
$$P_{N;M}(W) = C_{N;M}\ e^{\ \frac{N}{2}\mathrm{Tr}W} (\det W)^{\frac{M \quad N \quad 1}{2}}$$

We will also assume for convenience that the vector b is normally distributed: $b =$
with $>$ 0 and the components of $= (\ _1;\ldots;\ _M)^T$ are mean zero standard normals.

Qualitative considerations:

The equation for the Lagrange multiplier can be conveniently written in terms of N nonzero eigenvalues $s_1; \dots ; s_N$ of M M matrix $W^{(a)} = AA^T$ and the associated eigenvectors v_i:

$$\sum_{i=1}^{N} \frac{s_i}{(\quad s_i)^2} (\quad {}^T v_i)^2 = \frac{N}{2}$$



Case $N = 5$

Counting Lagrange multipliers via the Kac-Rice formula:

The number $N_{st}[a;b]$ of real solutions of the equation $A^T[Ax \quad b] \quad x = 0$ on the sphere $x^2 = N$ such that $\quad 2 \; [a;b]$ can be counted by employing the Kac-Rice type formula

$$N_{st}[a;b] = \int_a^b d \quad \int \quad A^T(Ax \quad b) \quad x \quad x^2 \quad N$$

$$\det \begin{pmatrix} A^T A & I_N & x \\ 2x^T & & 0 \end{pmatrix} dx$$

Using Gaussianity of both the matrix entries $A_{ij} \quad N \quad (0;1)$ and the vector components $b \quad N_M(0; I_M \quad ^2)$ and introducing the parameter $\quad = \frac{1}{2}\ln(1 + \quad ^2)$ one can eventually nd the mean number of solutions as

$$E\{N_{st}[a;b]\} = \int_a^b$$

Counting Lagrange multipliers via the Kac-Rice formula :

For negative values of the Lagrange multiplier we have instead:

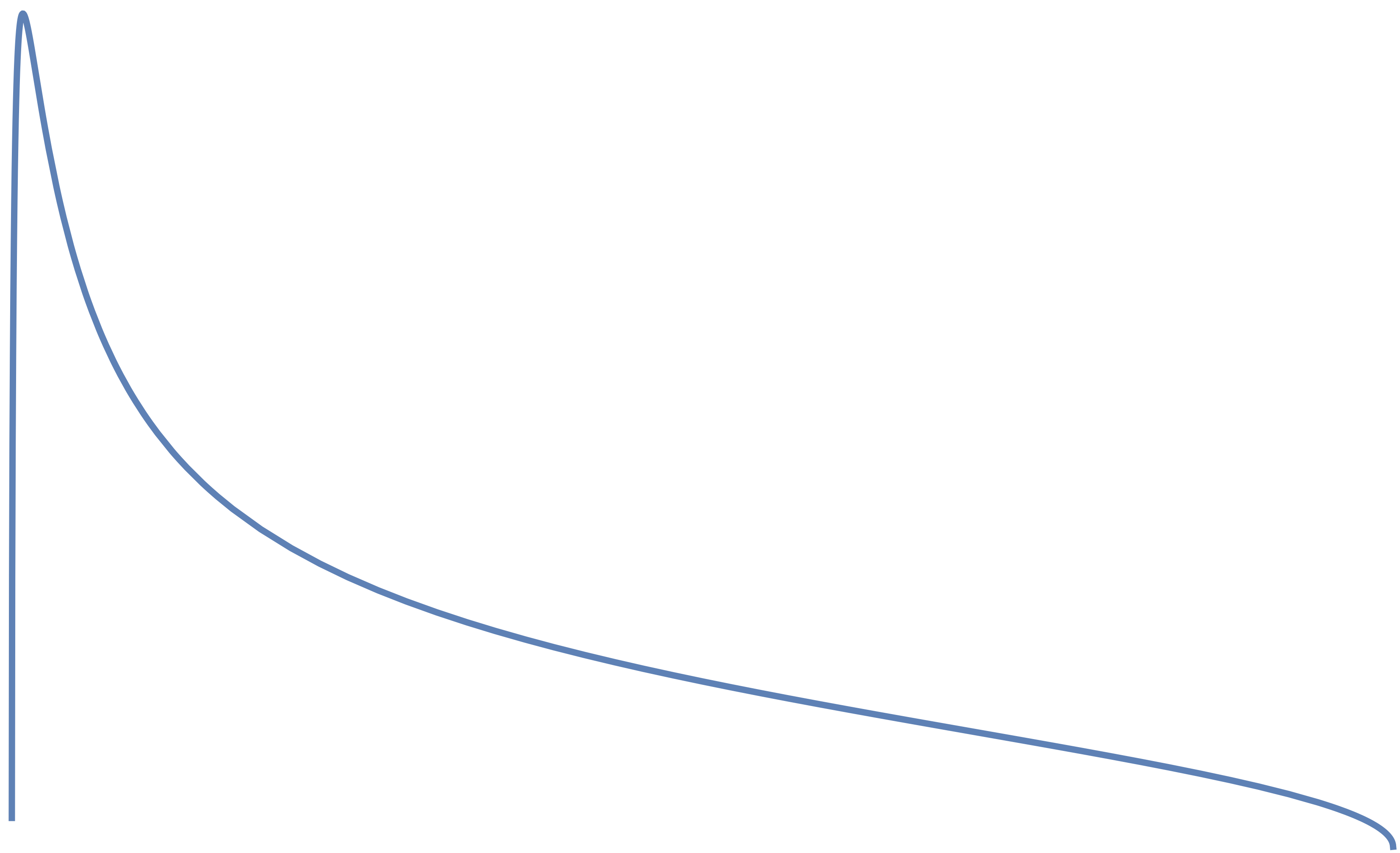
$$p(\ <\ 0) = \frac{N!N^{(M\ \ N)=2}}{2^{(M+N\ \ 3)=2}}\,\frac{1}{(\frac{N}{2})\ (\frac{M}{2})}\,\frac{e^{\ \ (M+N\ \ 1)=2}}{p\ \ \overline{\sinh}}\,e^{\ \frac{1}{2}N\,j\,j^2 j}$$

N jj [(=F22 12.0548 Tf 50 7 Td9 m 7 Td [(p)]TJ ET q 1 0 0 1 395.8 Tf323.634 cm []0 d 0 J 0.826 w 0 50 7 Td399.851S Q BT

"Bulk" Scaling Regime: extensive number of stationary points:

As N & M ! 1 in such a way that $1 <$ $= M{=}N <$ 1 the number of stationary points in the loss function landscapes shows three different regimes depending on the magnitude of the parameter $= \frac{1}{2} \ln (1 + {}^{2})$.

"Bulk" Scaling Regime: for small enough $1{=}N$ so that



Evolution of the density p_B (   ) in the 'bulk scaling' regime.

"Edge" Scaling Regime: nite number of stationary points:

The density of Lagrange multipliers for $N^{1=3}$ is dominated by the vicinities of the spectral edges

$$j \qquad s \;\; j \qquad N^{2=3} \quad \frac{4s^2}{s_+ \quad s}^{1=3}$$

where the

Counting stationary points in the edge regime.

Large Deviations for the smallest Lagrange multiplier:

For large N ! 1 , xed 1 < = M=N < 1 and xed nite ^{2} > 0 the probability density for the smallest Lagrange multiplier $_{min}$ has the Large Deviation form:

$$p(\ _{min} < s\ ) \quad e^{\frac{N}{2}\ (\ _{min})}; \qquad (\ ) = L_1(\ ) + L_2(\ ) + \frac{(\ +1)}{2}\ln(1+\ ^2),$$

where s = (p — 1)^2 is the 'Marchenko-Pastur' left edge and for $= \frac{(p\ 1)}{2}\ \frac{2}{\overline{1+\ ^2}}$

$$L_1(\ ) = (\qquad 1)\ \frac{p\ \overline{\ ^2 + \ ^2}}{\ } \quad \ln \quad + p\ \overline{\ ^2 + \ ^2} \qquad \frac{p\ \overline{(\ \ 1)^2 + \ ^2}}{(\ \ 1)}$$

$$L_2(\ ) = \quad p\ \overline{(\qquad s\ )(\qquad s_+)} \quad 2\ln \frac{(\ +1\quad + p\ \overline{(\ \ s\ )(\ \ s_+)})}{2 p\ —}$$

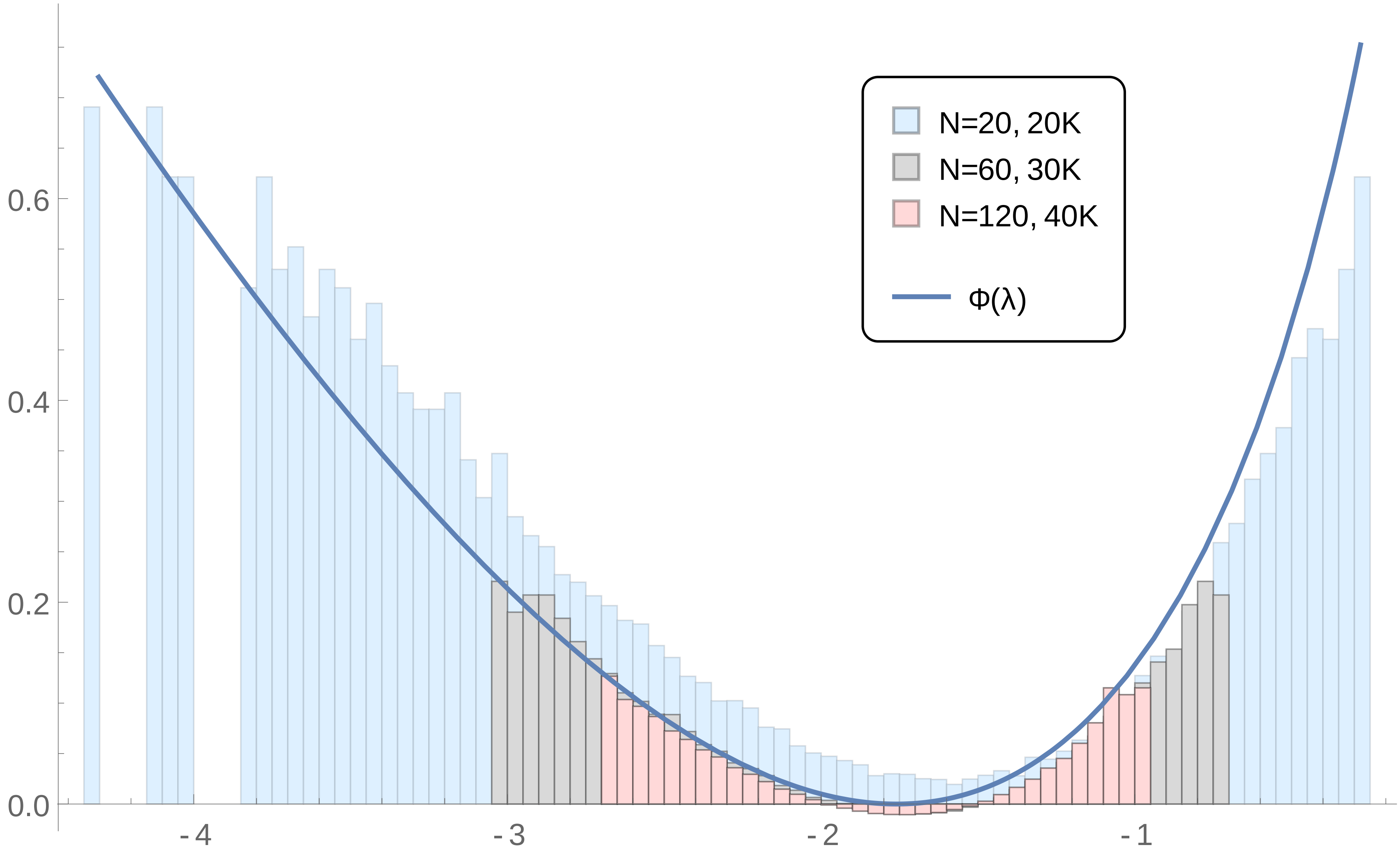
$$+2(\qquad 1)\ln \frac{(\ \ 1+\ \ + p\ \overline{(\ \ s\ )(\ \ s_+)})}{2 p\ —}$$

One nds that ( ) is minimized for

$$= \quad = (\ p\ — \quad p\ \overline{1+\ ^2})\quad p\ — \quad p\ \frac{1}{\overline{1+\ ^2}}$$

which eventually implies the most probable value of the minimal loss/error :

$$\lim_{N\ !1}\ \frac{E_{min}}{N} = \frac{1}{2}\left[p\ \overline{(1+\ ^2)} \quad 1 \right]^2$$



N=20, 20K
N=60, 30K
N=120, 40K
Φ(λ)
0.6
0.4
0.2
0.0
-4
-3
-2
-1

Conclusions:

We counted the mean number of stationary points of the simplest 'least-square ' optimization problem on a sphere via the Lagrange multipliers in various scaling regimes, and found the typical minimal loss E_{min} .

Open questions:

- Fluctuations of the counting function,
- large/small deviations of the minimal loss E_{min}
- Gradient search dynamics on the sphere
- Landscape for a nonlinear 'least-square' optimization, etc.

THANK YOU!