COUNTING STATIONARY POINTS OF THE LOSS FUNCTION IN THE SIMPLEST CONSTRAINED LEAST-SQUARE OPTIMIZATION*

Yan V. Fyodorov, Rashel Tublin

King's College London, Department of Mathematics
London WC2R 2LS, United Kingdom

(Received November 30, 2019)

We use the Kac–Rice method to analyze statistical features of an "optimization landscape" of the loss function in a random version of the Oblique Procrustes Problem, one of the simplest optimization problems of the least-square-type on a sphere.

DOI:10.5506/APhysPolB.51.1663

1. Introduction

One of the simplest optimization problems of the least-square-type arising in the Multiple Factor Data Analysis is the following:

Oblique Procrustes Problem [1]: *For a given pair of $M\times N$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ find such $N\times N$ matrix $\boldsymbol{X}$ that the equality $\boldsymbol{B}=\boldsymbol{A}\boldsymbol{X}$ holds as close as possible and columns $\boldsymbol{x}_i\in\mathbb{R}^N$, $i=1,\dots N$ are all of the same fixed length:* $||\boldsymbol{x}||_2:=\sqrt{\sum_i x_i^2}=\text{const}$.

For $M>N$, the associated system of linear equations is over-complete and a solution can be found separately for each column $\boldsymbol{x}$ by minimizing the loss/cost function

$$H(\boldsymbol{x})=\frac{1}{2}||A\boldsymbol{x}-\boldsymbol{b}||^2:=\frac{1}{2}\sum_{k=1}^{M}\left[\sum_{j=1}^{N}A_{kj}\boldsymbol{x}_j-b_k\right]^2\,,\qquad ||\boldsymbol{x}||_2:=\text{const}\,. \tag{1}$$

* Presented at the conference *Random Matrix Theory: Applications in the Information Era*, Kraków, Poland, April 29–May 3, 2019.

The problem was first analysed in that setting by Browne in 1967 [1] and then independently by numerical mathematicians (see *e.g.* [2, 3]) who used the Lagrange multiplier to take care of the spherical constraint. Introducing the Lagrangian $\mathcal{L}_{\lambda,\boldsymbol{s}}(\boldsymbol{x}) = \mathcal{H}(\boldsymbol{x}) - \frac{\lambda}{2}(\boldsymbol{x},\boldsymbol{x})$, with real λ being the Lagrange multiplier, the stationary conditions $\nabla\mathcal{L}_{\lambda,\boldsymbol{s}}(\boldsymbol{x}) = 0$ yield a linear system

$$A^T\left[A\boldsymbol{x} - \boldsymbol{b}\right] = \lambda\boldsymbol{x}\,, \quad \Rightarrow \boldsymbol{x} = \left(A^TA - \lambda I_N\right)^{-1} A^T\boldsymbol{b}\,. \tag{2}$$

We find it convenient to use the normalization such that the radius of the sphere is $||\boldsymbol{x}||_2 := \sqrt{N}$, with the spherical constraint yielding the equation for λ in the form of

$$\boldsymbol{b}^T A\,\frac{1}{\left(A^TA - \lambda I_N\right)^2}\,A^T\boldsymbol{b} = N\,, \tag{3}$$

which is equivalent to a polynomial equation of degree $2N$ in λ. Each real solution for the Lagrange multiplier λ_i corresponds to a stationary point $\boldsymbol{x}_i$ of the loss function $H(\boldsymbol{x}) = \frac{1}{2}||A\boldsymbol{x} - \boldsymbol{b}||^2$ on the sphere $\boldsymbol{x}^2 = N$ and one can show that the order $\lambda_1 < \lambda_2 < \ldots < \lambda_{\mathcal{N}}$ implies $H(\boldsymbol{x}_1) < H(\boldsymbol{x}_2) < \ldots < H\left(\boldsymbol{x}_{\mathcal{N}}\right)$ [1]. Thus, the minimal loss is given by $\mathcal{E}_{\min} = H\left(\boldsymbol{x}_1\right)$.

Actually, the loss function (1) is one of the simplest examples of the "optimization landscape", interest in which governs developing various search algorithms efficiently converging to the global minimum. To consider a "typical" landscape, it makes sense to assume that the parameters of the model, *i.e.* the matrix A and the vector b, are random. Geometrical and topological properties of random landscapes have general and intrinsic mathematical interest, see *e.g.* [4], and have attracted considerable attention in recent years due to their relevance in the area of "deep learning" and optimization, see *e.g.* [5, 6]. Fruitful analogies with spin glasses where "energy landscapes" have been under intensive investigation for some time, see [7–14], play an important role in guiding the intuition in this area. In this context, the goal of the present research is to investigate the simplest landscape Eq. (1) by counting the stationary points via the Lagrange multipliers λ_i, $i = 1, \ldots, \mathcal{N} \leq 2N$ and eventually find the minimal loss $\mathcal{E}_{\min}$. For concreteness and analytical tractability, we assume the entries A_{kj} of $M \times N$, $M > N$ matrix A to be i.i.d. normal real variables such that $A^TA = W$ is $N \times N$ Wishart with the probability density

$$P_{N,M}(W) = C_{N,M}\,\mathrm{e}^{-\frac{N}{2}\mathrm{Tr}W}\,(\det W)^{\frac{M-N-1}{2}}\,. \tag{4}$$

We will also assume for convenience that the vector $\boldsymbol{b}$ is normally distributed: $\boldsymbol{b} = \sigma\,\xi$ with $\sigma > 0$ and the components of $\xi = (\xi_1, \ldots, \xi_M)^T$ are mean zero standard normals.

2. Qualitative considerations and the Kac–Rice method

Equation Eq. (3) for the Lagrange multiplier can be conveniently written in terms of N nonzero eigenvalues $s_1, \ldots, s_N$ of $M \times M$ matrix $W^{(a)} = AA^T$ and the associated eigenvectors $\boldsymbol{v}_i$

$$\sum_{i=1}^{N} \frac{s_i}{(\lambda - s_i)^2} \left(\boldsymbol{\xi}^T \boldsymbol{v}_i\right)^2 = \frac{N}{\sigma^2}\,. \tag{5}$$

The left-hand side is a positive function of λ having a single minimum between every consecutive pair of eigenvalues of $W^{(a)}$. This implies there are 0 or 2 solutions of (5) (and 1 solution with probability zero) for λ between every consecutive pair of eigenvalues, plus two more solutions: one in $\lambda \in (-\infty, s_1)$ and another one in $\lambda \in (s_N, \infty)$. Note that the latter two solutions exist for any value of $\sigma \in [0, \infty]$, whereas by changing σ, one changes the number of solutions available between consecutive eigenvalues. In particular, in the limit of vanishing noise (*i.e.* $\sigma \to 0$ hence $||\boldsymbol{b}||_2 = 0$), every stationary point solution for the Lagrange multiplier corresponds to an eigenvalue s_n of the Wishart matrix, with, $\boldsymbol{x} = \pm \boldsymbol{e}_n$ being the associated eigenvectors (hence there are $2N$ stationary points). On the other hand, when $\sigma \to \infty$, the ratio N/σ^2 in the right-hand side becomes smaller than the global minimum of the left-hand side in $[s_1, s_N]$. Then only two stationary points remain outside that interval. Obviously, in every particular realization, the number of stationary points will gradually change between the two limits as a function of growing σ, forming a staircase $\mathcal{N}_{\rm st}(\sigma)$. Let us illustrate this on a simple example in the case of small $N = 5$, see Fig. 1.

This is exactly the "gradual topology trivialization" phenomenon discussed (as the function of magnetic field) for the standard GOE-based spherical model in [15] (see also [11, 12]) by adopting formulas derived in the general case by Auffinger *et al.* [7, 8]. It is quite easy to see from (5) that the trivialization happens on the scale $\sigma^2 \sim 1/N$ as only for such values the left-hand side is of the same order as the right-hand side for a generic $\lambda \in [s_l, s_{l+1}]$ (in our normalization, the typical distance $|s_l - s_{l+1}| = O(1/N)$). When averaged over the realizations, the staircase is replaced by smoothly decreasing function $\langle \mathcal{N}_{\rm st}(\sigma) \rangle$ which we will find explicitly using the Kac–Rice approach, and investigate its asymptotics as $N \to \infty$.

The number $\mathcal{N}_{\rm st}[a, b]$ of real solutions of the Lagrange equation (2), *i.e.* $A^T [A\boldsymbol{x} - \boldsymbol{b}] - \lambda \boldsymbol{x} = 0$ on the sphere $\boldsymbol{x}^2 = N$ such that $\lambda \in [a, b]$ can be counted by employing the Kac–Rice-type formula

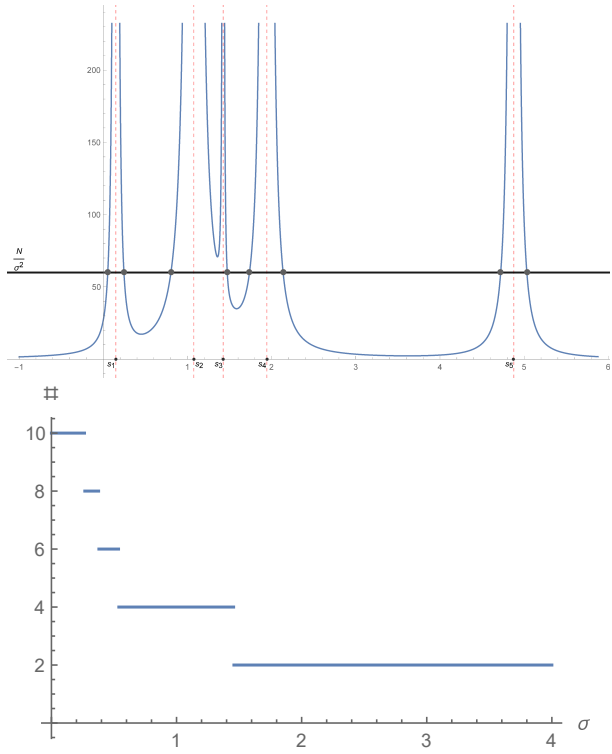

Fig. 1. Graphical representation of Eq. (5) for $N = 5$.

$$\mathcal{N}_{\rm st}[a,b] \;=\; \int_a^b \mathrm{d}\lambda \int \delta\left[A^T\left(A\boldsymbol{x}-\boldsymbol{b}\right)-\lambda\boldsymbol{x}\right]\delta\left(\boldsymbol{x}^2-N\right) \times\left|\det\begin{pmatrix} A^TA-\lambda I_N & \boldsymbol{x} \\ -2\boldsymbol{x}^T & 0\end{pmatrix}\right|\mathrm{d}\boldsymbol{x}\,. \tag{6}$$

Using Gaussianity of both the matrix entries $A_{ij} \sim \mathcal{N}(0,1)$ and the vector components $\boldsymbol{b} \sim \mathcal{N}_M(0, I_M\sigma^2)$ and introducing the parameter $\delta = \frac{1}{2}\ln(1+\sigma^2)$, one can eventually find the mean number of solutions as

$$\mathbb{E}\left\{\mathcal{N}_{\rm st}[a,b]\right\} = \int_a^b p(\lambda)\,\mathrm{d}\lambda$$

with the density $p(\lambda)$ for $\lambda > 0$ given by

$$p(\lambda \geq 0) = 2\sqrt{\frac{N}{\pi}}\,\frac{\mathrm{e}^{-\frac{M+N-1}{2}\delta}}{\sqrt{\sinh\delta}}\,K_{\frac{M-N}{2}}\left(\frac{N\lambda}{2\sinh\delta}\right)\mathrm{e}^{\frac{N\lambda}{2}\coth\delta}\,\langle\rho_N(\lambda)\rangle\,\sqrt{\lambda}\,, \tag{7}$$

where $K_\nu(z)$ is the Bessel–Macdonald function, and $\langle \rho_N(\lambda) \rangle$ stands for the mean eigenvalue density of $N \times N$ real Wishart matrices W distributed according Eq. (4). Such a density for any values $M \geq N$ can be found in [16]. For negative values of the Lagrange multiplier λ, we have instead

$$
\begin{aligned}
p(\lambda < 0) \;=\; & \frac{N! N^{(M-N)/2}}{2^{(M+N-3)/2}} \frac{1}{\Gamma\left(\frac{N}{2}\right) \Gamma\left(\frac{M}{2}\right)} \\
& \times \frac{\mathrm{e}^{-(M+N-1)\delta/2}}{\sqrt{\sinh \delta}} \mathrm{e}^{-\frac{1}{2} N|\lambda|(\coth \delta - 1)} |\lambda|^{(M-N)/2} \\
& \times \left[\sum_{j=0}^{N-1} \binom{M-1}{N-1-j} \frac{1}{j!} (N|\lambda|)^j \right] K_{\frac{M-N}{2}} \left(\frac{N|\lambda|}{2 \sinh \delta} \right) . \qquad (8)
\end{aligned}
$$

These formulas are exact, and we can compare them with the direct numerical simulations in Fig. 2 for moderate matrix sizes.

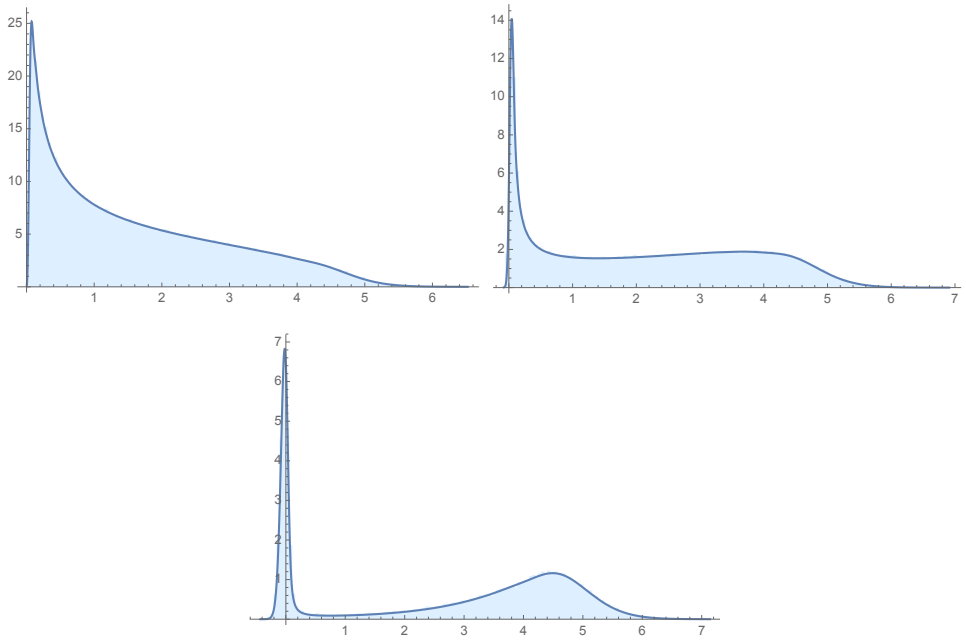

Fig. 2. (Color online) Evolution of the density $p(\lambda)$ for $N = 20$, $M = 30$ as the function of variance $\sigma^2 = 0.005; 0.25; 0.70$. The gray/blue histograms correspond to 10 000 realizations.

Our next goal is to investigate the limit N and $M \to \infty$.

2.1. Asymptotic analysis

2.1.1. "Bulk" scaling regime: extensive number of stationary points

As N and $M \to \infty$ in such a way that $1 < \mu = M/N < \infty$, the number of stationary points in the loss function landscapes shows three different regimes depending on the magnitude of the parameter $\delta = \frac{1}{2}\ln(1+\sigma^2)$. The first regime is associated with the "bulk scaling" corresponding to small enough $\delta \sim 1/N$, so that $\gamma = \frac{\delta N}{4} < \infty$. For such a regime, one finds that the total number of solutions $\mathcal{N}$ is *extensive*, namely

$$\lim_{N\to\infty} \frac{\mathbb{E}\{\mathcal{N}\}}{N} = \int_{s_-}^{s_+} p_{\rm B}(\lambda)\,{\rm d}\lambda > 0\,, \qquad s_\pm = \left(\sqrt{\mu}\pm 1\right)^2\,, \tag{9}$$

where the density function $p_{\rm B}(\lambda)$ is expressed via the Marchenko–Pastur [17] limiting eigenvalue density $p_{\rm MP}(\lambda)$ for the Wishart ensemble as (see Fig. 3 below)

$$\begin{aligned} p_{\rm B}(\lambda) &= 2\,p_{\rm MP}(\lambda)\,\exp\left[-\frac{\gamma}{\lambda}(\lambda-s_-)(s_+-\lambda)\right], \\ p_{\rm MP}(\lambda) &= \frac{1}{2\pi\lambda}\sqrt{(\lambda-s_-)(s_+-\lambda)}\,. \end{aligned} \tag{10}$$

For $\gamma = 0$, we obviously have $\mathbb{E}\{\mathcal{N}\} = 2N$, whereas for $\gamma \gg 1$, we have asymptotically

$$\lim_{N\to\infty} \frac{\mathbb{E}\{\mathcal{N}\}}{N}\Big|_{\gamma\gg 1} \approx \frac{1}{4\sqrt{\pi}}\frac{1}{\gamma^{3/2}} \ll 1\,.$$

Evaluating the above for $\gamma \sim N^{2/3}$ (*i.e.* $\delta \sim N^{-1/3} \gg 1/N$) indicates that the mean number of stationary points for such a γ becomes of the order of unity as $N \gg 1$ defining a different scaling regime, *cf.* [15].

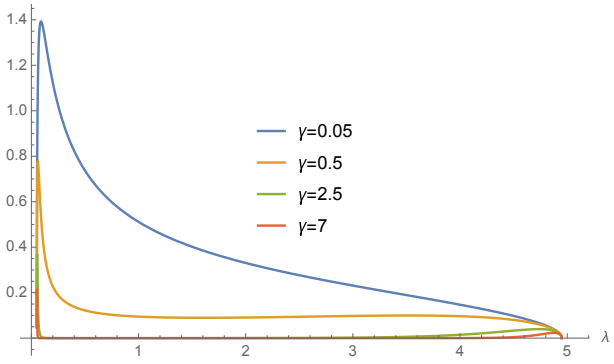

Fig. 3. Evolution of the density $p_{\rm B}(\lambda)$ in the "bulk scaling" regime.

2.1.2. "Edge" scaling regime: finite number of stationary points

The density of Lagrange multipliers for $\delta \sim N^{-1/3}$ is dominated by the vicinities of the spectral edges

$$|\lambda - s_\pm| \sim N^{-2/3} \left(\frac{4s_\pm^2}{s_+ - s_-} \right)^{1/3} \xi \,,$$

where the Marchenko–Pastur law is no longer valid and has to be replaced by a more precise "edge density" given by [18]

$$p_{\rm MP}(\lambda) \longrightarrow \left(\frac{s_+ - s_-}{4N s_\pm^2} \right)^{1/3} \rho_{\rm edge}(\xi) \,, \tag{11}$$

with

$$\rho_{\rm edge}(\zeta) = \left[{\rm Ai}'(\zeta) \right]^2 - \zeta \left[{\rm Ai}(\zeta) \right]^2 + \frac{1}{2} {\rm Ai}(\zeta) \left(1 - \int_\zeta^\infty {\rm Ai}(\eta)\, {\rm d}\eta \right) \,, \tag{12}$$

where ${\rm Ai}(\zeta) = \frac{1}{2\pi i} \int_\Gamma {\rm e}^{\frac{v^3}{3} - v\zeta}$ is the Airy function solving ${\rm Ai}''(\zeta) - \zeta {\rm Ai}(\zeta) = 0$.

Introducing the scaling parameter $\omega = N^{1/3}\delta \left(\frac{s_+ - s_-}{4} \right)$, one then finds that the total number of stationary points in this regime is finite as $N \to \infty$

$$\lim_{N\to\infty} \mathbb{E}\{\mathcal{N}\} = 2 \int_{-\infty}^{\infty} \left[\exp\left(-\frac{\omega^3}{3s_-} + \frac{\omega\zeta}{s_-^{1/3}} \right) + \exp\left(-\frac{\omega^3}{3s_+} + \frac{\omega\zeta}{s_+^{1/3}} \right) \right] \rho_{\rm edge}(\zeta)\, {\rm d}\zeta \,. \tag{13}$$

In particular, that number tends to just $\lim_{N\to\infty} \mathbb{E}\{\mathcal{N}\} = 2$ as long as $\omega \to \infty$, indicating that for any fixed and finite variance $0 < \sigma^2 < \infty$, only two stationary points typically exist: one maximum and one minimum, *cf.* [15].

Comparison with results numerical simulations is shown in Fig. 4.

2.2. Large Deviations for the smallest Lagrange multiplier

For large $N \to \infty$, fixed $1 < \mu = M/N < \infty$ and fixed finite $\sigma^2 > 0$, the probability density for the smallest Lagrange multiplier $\lambda_{\rm min}$ has the *Large Deviation* form

$$p(\lambda_{\rm min} < s_-) \sim {\rm e}^{-\frac{N}{2}\boldsymbol{\Phi}(\lambda_{\rm min})} \,, \qquad \boldsymbol{\Phi}(\lambda) = L_1(\lambda) + L_2(\lambda) + \frac{(\mu+1)}{2} \ln\left(1+\sigma^2\right) \,, \tag{14}$$

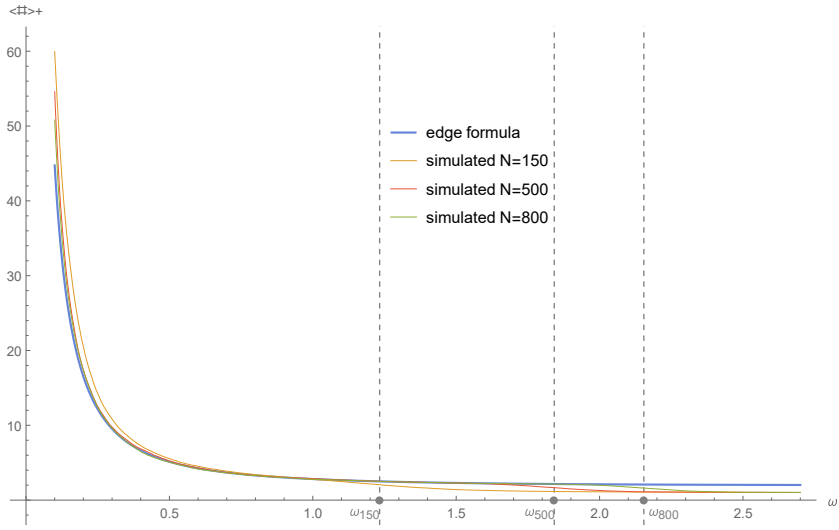

Fig. 4. Counting stationary points in the edge regime.

where $s_- = (\sqrt{\mu}-1)^2$ is the Marchenko–Pastur left edge and for $\kappa = \frac{(\mu-1)\sigma^2}{2\sqrt{1+\sigma^2}}$, we defined

$$L_1(\lambda) = (\mu-1)\left\{\frac{\sqrt{\lambda^2+\kappa^2}}{\kappa} - \ln\left(\kappa+\sqrt{\lambda^2+\kappa^2}\right) - \lambda\frac{\sqrt{(\mu-1)^2+\kappa^2}}{(\mu-1)\kappa}\right\}$$

and

$$L_2(\lambda) = -\sqrt{(\lambda-s_-)(\lambda-s_+)} - 2\ln\frac{\left(\mu+1-\lambda+\sqrt{(\lambda-s_-)(\lambda-s_+)}\right)}{2\sqrt{\mu}}$$
$$+2(\mu-1)\ln\frac{\left(\mu-1+\lambda+\sqrt{(\lambda-s_-)(\lambda-s_+)}\right)}{2\sqrt{\mu}}. \qquad (15)$$

Comparison with the probability density of the smallest solution of Eq. (5) found numerically is shown in Fig. 5.

One then finds that $\Phi(\lambda)$ is minimized for

$$\lambda = \lambda_* = \left(\sqrt{\mu}-\sqrt{1+\sigma^2}\right)\left(\sqrt{\mu}-\frac{1}{\sqrt{1+\sigma^2}}\right) \qquad (16)$$

providing the most probable/typical value of the smallest Lagrange multiplier. Substituting this value to Eq. (2) and then to Eq. (1) gives eventually the most probable value of the minimal loss/error

$$\lim_{N\to\infty}\frac{\mathcal{E}_{\min}}{N} = \frac{1}{2}\left[\sqrt{\mu\,(1+\sigma^2)}-1\right]^2. \qquad (17)$$

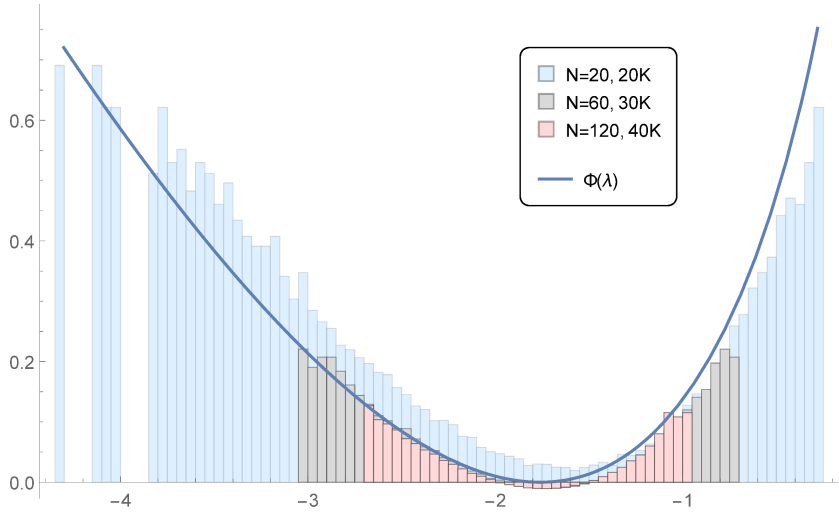

Fig. 5. The Large Deviation function for the smallest Lagrange multiplier *vs.* simulations for different matrix sizes N and different number of samples.

2.3. Open questions

In conclusion, 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 $\mathcal{E}_{\min}$. The following questions remain open: *(i)* fluctuations of the counting function, *(ii)* large/small deviations of the minimal loss $\mathcal{E}_{\min}$, *(iii)* gradient search dynamics on the sphere, *(iv)* understanding the landscapes for "least-square" optimization of more general type, *e.g.* involving nonlinearities *etc.*, *cf.* [19]. We hope to address some of these issues in future publications.

REFERENCES

[1] M.W. Browne, *Psychometrica* **32**, 125 (1967).

[2] W. Gander, *Numer. Math.* **36**, 291 (1980).

[3] G.H. Golub, U. von Matt, *Numer. Math.* **59**, 561 (1991).

[4] Y.V. Fyodorov, A. Lerario, E. Lundberg, *J. Geom. Phys.* **95**, 1 (2015).

[5] A. Choromanska *et al.*, «The Loss Surfaces of Multilayer Networks», in: Proceedings of the 18$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA.

[6] V. Ros, G. Ben Arous, G. Biroli, C. Cammarota, *Phys. Rev. X* **9**, 011003 (2019).

[7] A. Auffinger, G. Ben Arous, *Ann. Probab.* **41**, 4214 (2013).

[8] A. Auffinger, G. Ben Arous, J. Cerny, *Commun. Pure Appl. Math.* **66**, 165 (2013).

[9] A.J. Bray, D.S. Dean, *Phys. Rev. Lett.* **98**, 150201 (2007).

[10] Y.V. Fyodorov, *Phys. Rev. Lett.* **92**, 240601 (2004); *Erratum ibid.* **93**, 149901 (2004).

[11] Y.V. Fyodorov, *Markov Process. Relat. Fields* **21**, 483 (2015).

[12] Y.V. Fyodorov, *J. Stat. Mech.: Theor. Exp.* **2016**, 124003 (2016).

[13] Y.V. Fyodorov, C. Nadal, *Phys. Rev. Lett.* **109**, 167203 (2012).

[14] Y.V. Fyodorov, I. Williams, *J. Stat. Phys.* **129**, 1081 (2007).

[15] Y.V. Fyodorov, P. Le Doussal, *J. Stat. Phys.* **154**, 466 (2014).

[16] G. Livan, M. Novaes, P. Vivo, «Introduction to Random Matrices: Theory and Practice», *Springer International Publishing*, Cham 2018.

[17] V.A. Marchenko, L.A. Pastur, *Mathematics of the USSR-Sbornik* **72**, 507 (1967).

[18] P.J. Forrester, *J. Phys. A: Math. Theor.* **45**, 145201 (2012).

[19] Y.V. Fyodorov, *J. Stat. Phys.* **175**, 789 (2019).