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

Yan V Fyodorov

Department of Mathematics

KING'S College LONDON

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 $\mathbf{x} \in \mathbb{R}^N,\ \mathbf{x}^2 = const$ arises in the **Multiple Factor Data Analysis** and is known as the **Oblique Procrustes Problem**:

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

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

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

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 $\mathcal{L}_{\lambda,\mathbf{s}}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \frac{\lambda}{2}(\mathbf{x}, \mathbf{x})$, with real λ being the Lagrange multiplier, the stationary conditions $\nabla\mathcal{L}_{\lambda,\mathbf{s}}(\mathbf{x}) = 0$ yields linear system:

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

Setting of the problem:

The spherical constraint $\mathbf{x}^2 = N$ yields the equation for λ in the form:

$$\mathbf{b}^T A \frac{1}{\left(A^T A - \lambda I_N\right)^2} A^T \mathbf{b} = N$$

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** $\mathbf{x}_i$ of the loss function $H(\mathbf{x}) = \frac{1}{2}||A\mathbf{x} - \mathbf{b}||^2$ on the sphere $\mathbf{x}^2 = N$ and one can show that the order $\lambda_1 < \lambda_2 < \ldots < \lambda_{\mathcal{N}}$ implies $H(\mathbf{x}_1) < H(\mathbf{x}_j) < \ldots < H(\mathbf{x}_{\mathcal{N}})$. Thus the **minimal loss** is given by $\mathcal{E}_{min} = H(\mathbf{x}_1)$.

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

$$\lambda_i,\ i = 1, \ldots, \mathcal{N} \leq 2N$$

and eventually find the **minimal loss** $\mathcal{E}_{min}$ after assuming the entries A_{kj} of $M \times N$, $M > N$ matrix A to be i.i.d. normal real variables such that $A^T A = W$ is $N \times N$ **Wishart** with the probability density

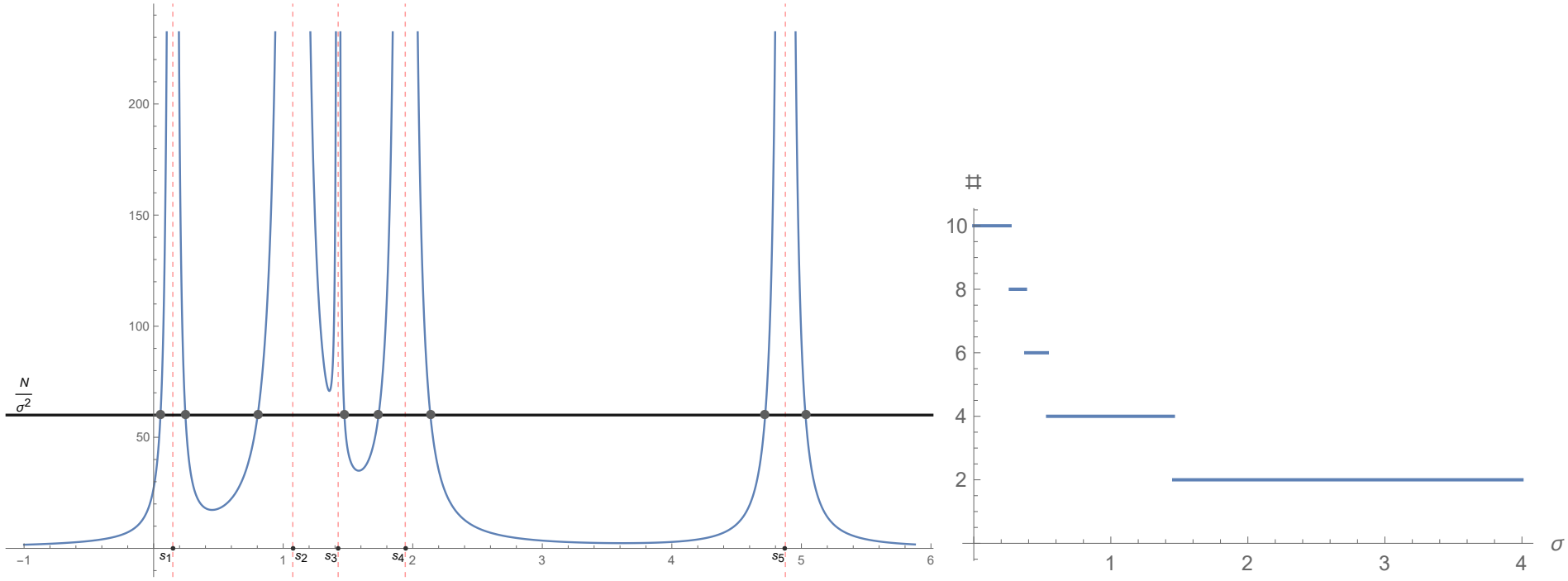
$$P_{N,M}(W) = C_{N,M} e^{-\frac{N}{2}\mathrm{Tr}W} \left(\det W\right)^{\frac{M-N-1}{2}}$$

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

Qualitative considerations:

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

$$\sum_{i=1}^{N} \frac{s_i}{(\lambda - s_i)^2} (\xi^T \mathbf{v}_i)^2 = \frac{N}{\sigma^2}$$



Case $N = 5$

Counting Lagrange multipliers via the Kac-Rice formula:

The number $\mathcal{N}_{st}[a,b]$ of real solutions of the equation $A^T\left[A\mathbf{x}-\mathbf{b}\right]-\lambda\mathbf{x}=0$ on the sphere $\mathbf{x}^2=N$ such that $\lambda\in[a,b]$ can be counted by employing the **Kac-Rice** type formula

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

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

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

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

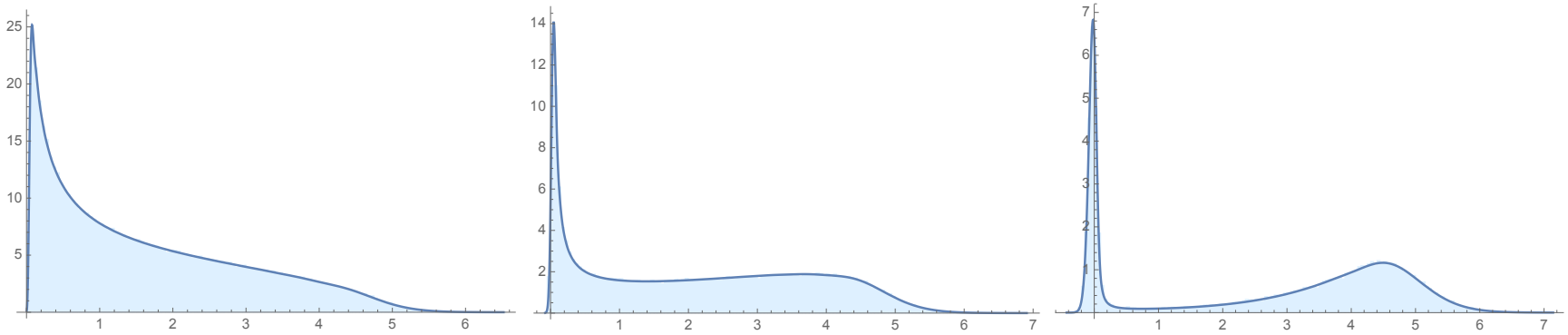
$$p(\lambda\geq 0)=2\sqrt{\frac{N}{\pi}}\frac{e^{-\frac{M+N-1}{2}\delta}}{\sqrt{\sinh\delta}}K_{\frac{M-N}{2}}\left(\frac{N\lambda}{2\sinh\delta}\right)e^{\frac{N\lambda}{2}\coth\delta}\left\langle\rho_N(\lambda)\right\rangle\sqrt{\lambda}$$

where $K_\nu(z)$ is the Bessel-Macdonald function, and $\left\langle\rho_N(\lambda)\right\rangle$ stands for the mean eigenvalue density of $N\times N$ Wishart matrix $W=A^TA$ presented for any M,N in **Introduction to Random Matrices: Theory and Practice** by **G. Livan**,**M. Novaes** and **P. Vivo** (Springer 2018).

Counting Lagrange multipliers via the Kac-Rice formula :

For negative values of the Lagrange multiplier λ we have instead:

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



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 blue histograms correspond to 10000 realizations.

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

As $N \,\&\, M \to \infty$ 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 $\delta = \frac{1}{2}\ln(1+\sigma^2)$.

"Bulk" Scaling Regime: for small enough $\delta \sim 1/N$ so that $\gamma = \frac{\delta N}{4} < \infty$ one finds that the total number of solutions $\mathcal{N}$ is **extensive**: $\mathcal{N} \sim N$ so that

$$\lim_{N\to\infty} \frac{\mathbb{E}\{\mathcal{N}\}}{N} = \int_{s_-}^{s_+} p_B(\lambda)\, d\lambda > 0, \quad s_\pm = (\sqrt{\mu} \pm 1)^2$$

$$p_B(\lambda) = 2\, p_{MP}(\lambda)\, \exp\left[-\tfrac{\gamma}{\lambda}(\lambda - s_-)(s_+ - \lambda)\right]$$

where

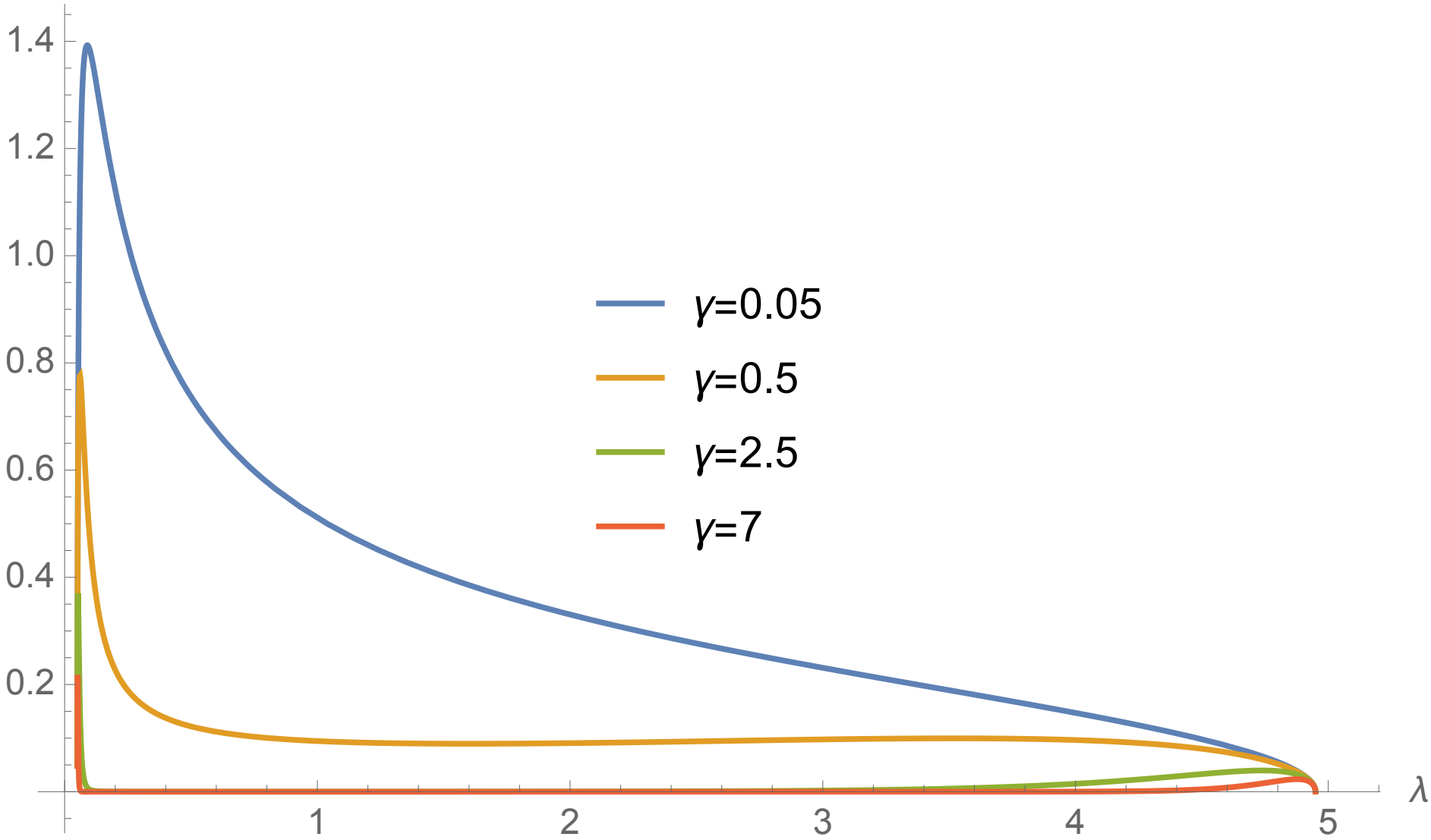
$$p_{MP}(\lambda) = \tfrac{1}{2\pi\lambda}\sqrt{(\lambda - s_-)(s_+ - \lambda)}$$

is the **Marchenko-Pastur** limiting eigenvalue density for the Wishart ensemble.

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

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

This indicates that for $\gamma \sim N^{2/3}$ (i.e. $\delta \sim N^{-1/3} >> 1/N$) the number of stationary points becomes of order of unity as $N \to \infty$ defining a different scaling regime.



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

"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 replaced by the **edge density**, see **Forrester '12**:

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

with

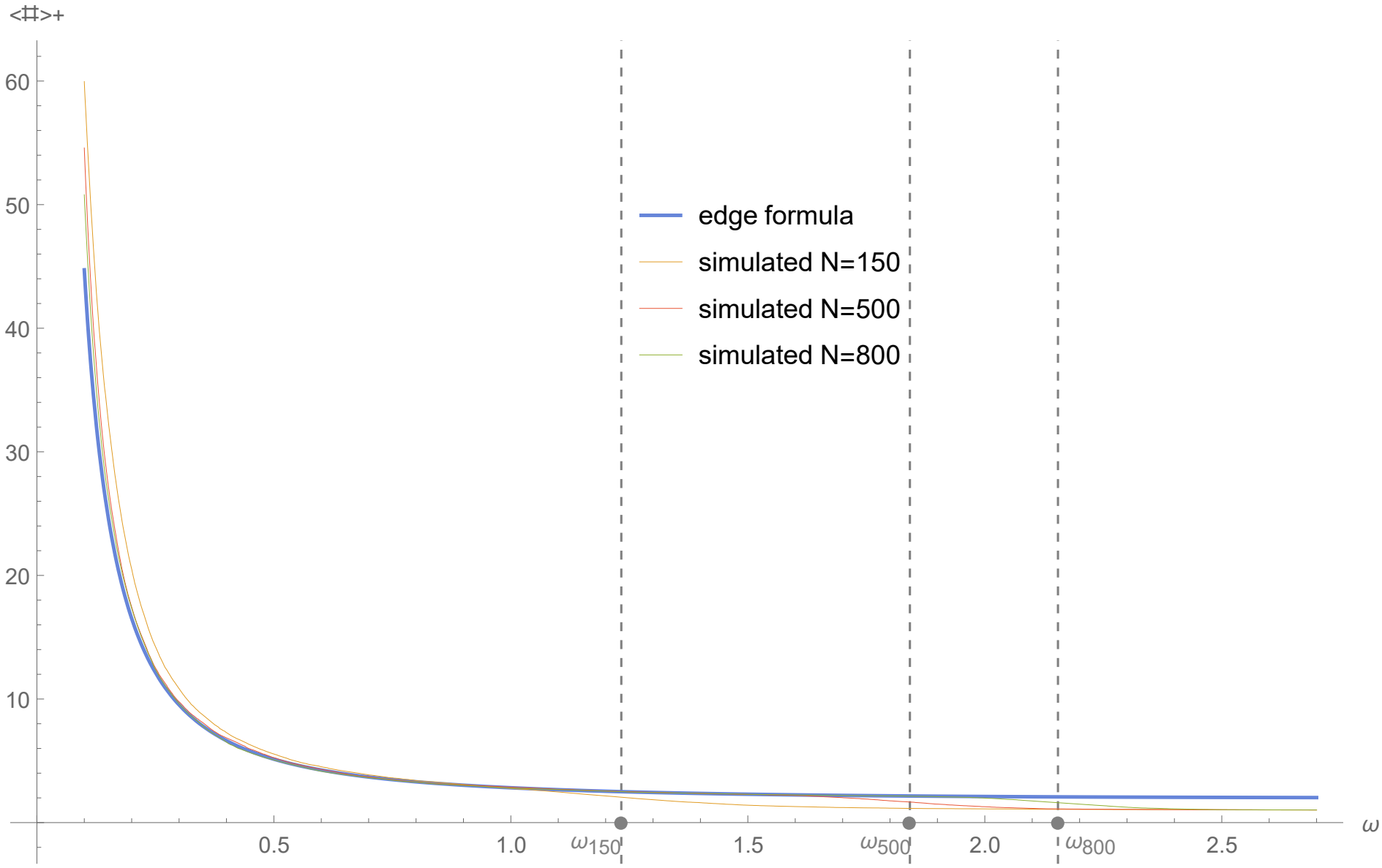
$$\rho_{edge}(\zeta) = [Ai'(\zeta)]^2 - \zeta [Ai(\zeta)]^2 + \frac{1}{2} Ai(\zeta) \left(1 - \int_{\zeta}^{\infty} Ai(\eta)\, d\eta \right)$$

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

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

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

In particular, that number tends to just **two** 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**.



Counting stationary points in the edge regime.

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 λ_{min} has the **Large Deviation** form:

$$p(\lambda_{min} < s_-) \sim e^{-\frac{N}{2}\mathbf{\Phi(\lambda_{min})}}, \quad \mathbf{\Phi}(\lambda) = \mathbf{L_1}(\lambda) + \mathbf{L_2}(\lambda) + \frac{(\mu+1)}{2}\ln(1+\sigma^2),$$

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

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

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

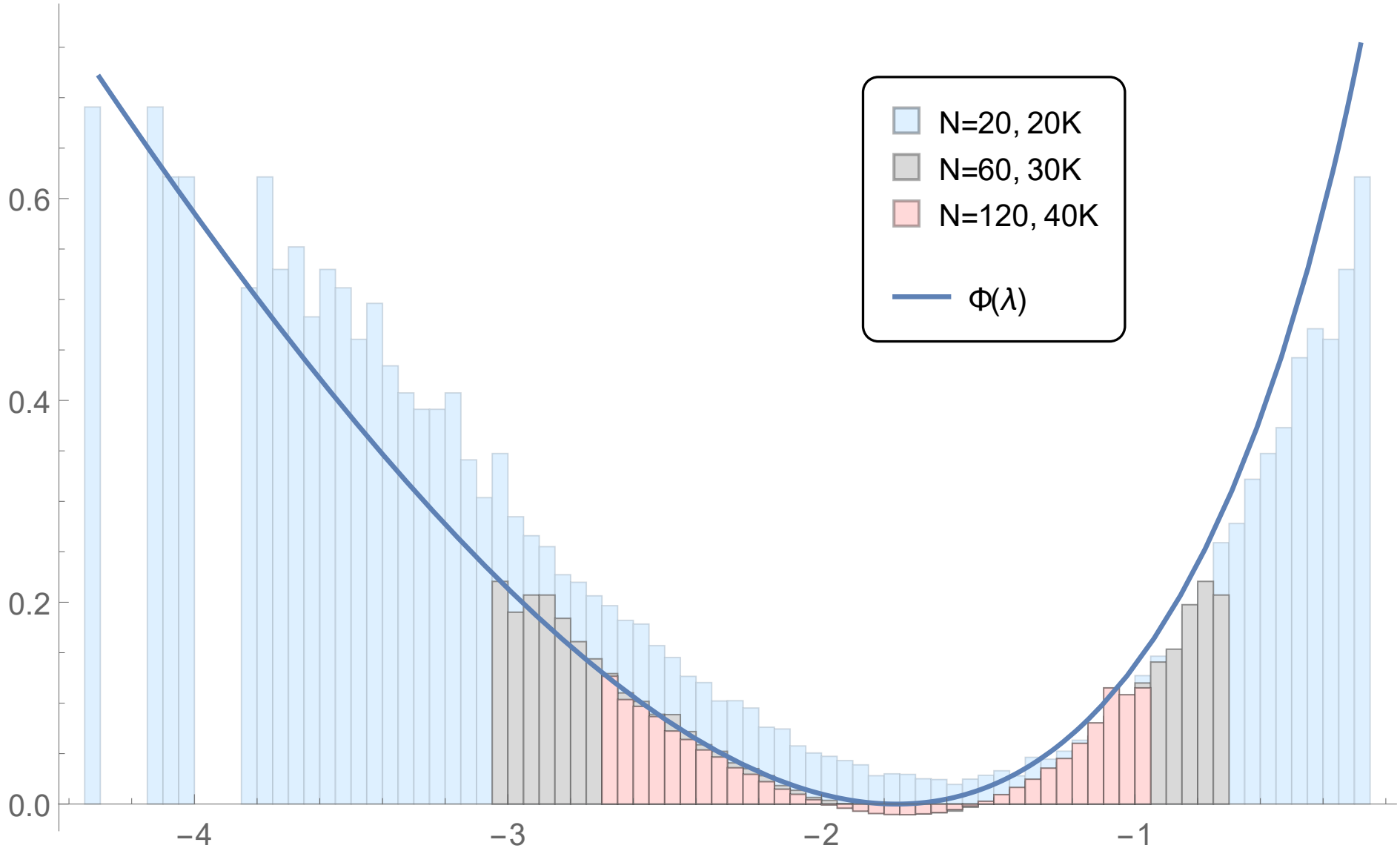
$$+2(\mu-1)\ln\frac{\left(\mu-1+\lambda+\sqrt{(\lambda-s_-)(\lambda-s_+)}\right)}{2\sqrt{\mu}}$$

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

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

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

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



The large deviation function for the smallest Lagrange multiplier vs. simulations

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 $\mathcal{E}_{min}$.

- **Open questions:**
  - Fluctuations of the counting function,
  - **large/small deviations** of the minimal loss $\mathcal{E}_{min}$
  - Gradient search dynamics on the sphere
  - Landscape for a **nonlinear** 'least-square' optimization, etc.

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 $\mathcal{E}_{min}$.

- **Open questions:**
  - Fluctuations of the counting function,
  - **large/small deviations** of the minimal loss $\mathcal{E}_{min}$
  - Gradient search dynamics on the sphere
  - Landscape for a **nonlinear** 'least-square' optimization, etc.

THANK YOU!