Algorithm-independent bounds on the performance of optimization using marginal complexity

Jaron Kent-Dobias • **INFN sezione di Roma I**

Optimization seeks extremal points in a function. When there are superextensively many optima, optimization algorithms are liable to get stuck. Under these conditions, generic algorithms tend to find marginal optima, which have many nearly flat directions. We introduce a technique to count marginal optima in arbitrary settings, and use it to effectively bound the range of endpoints for generic algorithms. We demonstrate the idea using a simple non-Gaussian problem: random nonlinear least most squares.

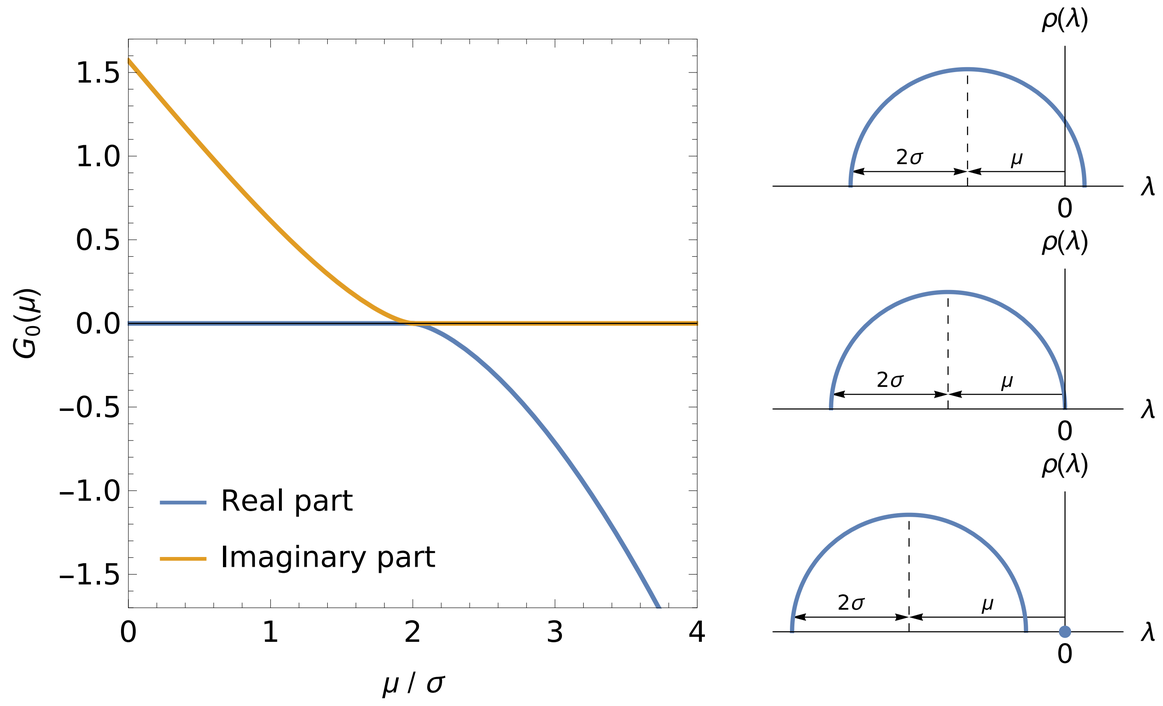
Conditioning on the maximum eigenvalue

An arbitrary function of the maximum eigenvalue of a matrix A can be written

$$g(\lambda_{\max}(A)) = \lim_{\beta\to\infty} \int \frac{d\mathbf{s}\,\delta(N-\mathbf{s}^T\mathbf{s})e^{\beta\mathbf{s}^TA\mathbf{s}}}{\int d\mathbf{s}'\,\delta(N-\mathbf{s}'^T\mathbf{s}')e^{\beta\mathbf{s}'^TA\mathbf{s}'}} g\left(\frac{\mathbf{s}^TA\mathbf{s}}{N}\right) \tag{1}$$

since the measure concentrates on the eigenspace associated with $\lambda_{\max}(A)$. As an example, let $A = B - \mu I$ for GOE B with $\overline{B^2} = \sigma^2/N$. Consider the large-deviation function for the probability that $\lambda_{\max}(A) = \lambda^*$

$$e^{NG_{\lambda^*}(\mu)} = P(\lambda_{\max}(A) = \lambda^*) = \overline{\delta\big(N \times (\lambda_{\max}(A) - \lambda^*)\big)} \tag{2}$$



The result for $\lambda^* = 0$ above shows three regimes:

- when the shift μ is such that the bulk spectrum lies over zero, which requires an N^2 large deviation and $G_0(\mu)$ becomes imaginary
- when the shift μ is such that the bulk spectrum is pseudogapped
- when the shift μ is such that the bulk spectrum is strictly negative, and an isolated eigenvalue is pulled from the spectrum

Conditioning on a pseudogap

Marginal optimal do not simply have a zero eigenvalue; they have a pseudogap. Above, we see that the presence of the pseudogap is characterized by the breakdown of the order-N large-deviation function. In general, one can tune to a pseudogap by starting with sufficiently small μ and increasing it until the solution develops an imaginary component.

In the isotropic examples studied here, there is an easier way. In an isotropic landscape (zero signal to noise), typical spectra do not have an isolated eigenvalue. Therefore, it is always a large deviation with respect to a purely bulk spectrum. We can therefore identify the pseudogap shift $\mu = \mu_{\mathrm{m}}$ by solving

$$0 = \frac{\partial}{\partial\lambda^*} G_{\lambda^*}(\mu_{\mathrm{m}})\Big|_{\lambda^*=0} \tag{3}$$

which indeed gives $\mu_{\mathrm{m}} = 2\sigma$ for the shifted GOE case.

Marginal complexity

Consider an optimization problem over functions H defined on a configuration space itself defined by a set of constraints $\mathbf{g}(\mathbf{x}) = 0$. Introducing a vector $\boldsymbol{\omega}$ of Lagrange multipliers for the constraints, the gradient and Hessian are

$$\nabla H(\mathbf{x},\boldsymbol{\omega}) = \partial H(\mathbf{x}) + \omega_i \partial g_i(\mathbf{x}) \qquad \mathrm{Hess}\, H(\mathbf{x},\boldsymbol{\omega}) = \partial\partial H(\mathbf{x}) + \omega_i \partial\partial g_i(\mathbf{x}) \tag{4}$$

Optima with energy E, shift μ, and maximum eigenvalue λ^* can be counted by integrating the Kac–Rice measure

$$\begin{aligned} dv(\mathbf{x},\boldsymbol{\omega} \mid E,\mu,\lambda^*) = d\mathbf{x}\,d\boldsymbol{\omega}\,\delta\big(\nabla H(\mathbf{x},\boldsymbol{\omega})\big)\,\delta\big(\mathbf{g}(\mathbf{x})\big)\,\big|\det \mathrm{Hess}\, H(\mathbf{x},\boldsymbol{\omega})\big| \\ \times \delta\big(NE - H(\mathbf{x})\big)\,\delta\big(N\mu + \mathrm{Tr}\,\mathrm{Hess}\, H(\mathbf{x})\big)\,\delta\big(N\lambda^* - N\lambda_{\max}(\mathrm{Hess}\, H(\mathbf{x}))\big) \end{aligned} \tag{5}$$

The complexity is the average logarithm of the count, or

$$\Sigma_{\lambda^*}(E,\mu) = \frac{1}{N}\overline{\log \int dv(\mathbf{x},\boldsymbol{\omega} \mid E,\mu,\lambda^*)} \tag{6}$$

For each energy E, we can find the marginal shift $\mu_{\mathrm{m}}(E)$ by again requiring that

$$0 = \frac{\partial}{\partial\lambda^*}\Sigma_{\lambda^*}\big(E,\mu_{\mathrm{m}}(E)\big)\Big|_{\lambda^*=0} \tag{7}$$

yielding the marginal complexity $\Sigma_{\mathrm{m}}(E) = \Sigma_0(E,\mu_{\mathrm{m}}(E))$.

Random nonlinear ~~least~~ most squares

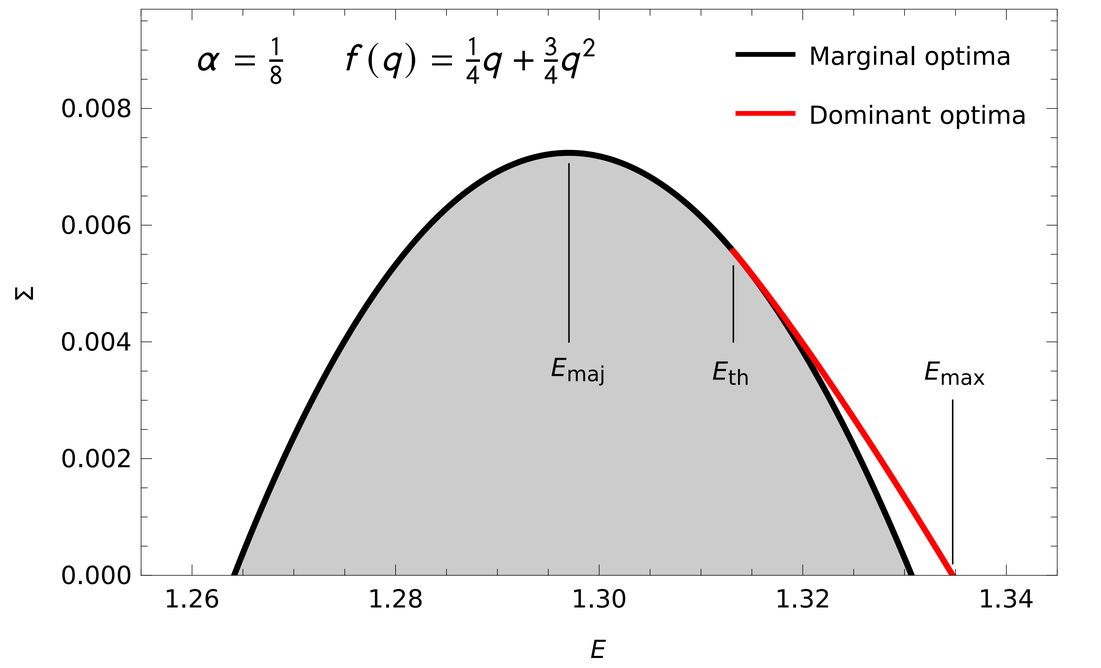
A simple non-Gaussian landscape is the sum of squared Gaussian functions

$$H(\mathbf{x}) = \frac{1}{2}\sum_{k=1}^{\alpha N} V_k(\mathbf{x})^2 \qquad \overline{V(\mathbf{x})} = 0 \qquad \overline{V_i(\mathbf{x})V_j(\mathbf{x}')} = \delta_{ij} f\left(\frac{\mathbf{x}^T\mathbf{x}'}{N}\right) \tag{8}$$

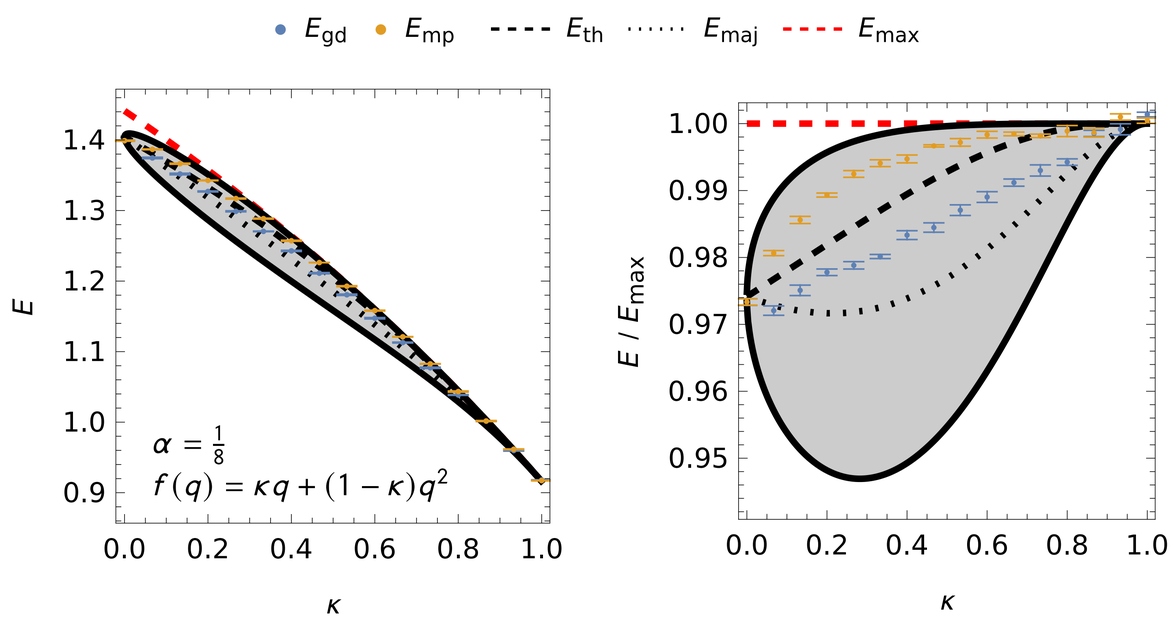
on a spherical configuration space $0 = g(\mathbf{x}) = N - \mathbf{x}^T\mathbf{x}$ with $\mathbf{x} \in \mathbb{R}^N$. Minimizing H is a model of random nonlinear least squares. However, the bottom of the landscape is more complicated than the top: the bottom tends to be either replica symmetric or full RSB, while the top reflects the order of whatever spherical spin-glass is associated with individual functions V.

Therefore, to keep the example simple, we will consider the problem of maximizing H, or nonlinear most squares, which for many natural choices of f has a replica-symmetric complexity of optima.

An example of the complexy is below. The techniques we use can only count minima or maxima in these models, so the dominant complexity is only computed up to the threshold energy E_{th} where saddle points become most common.



We compare the range of marginal optima to the endpoints of gradient descent and message passing optimization algorithms for a range of models. The results are effectively bounded by the range of marginal optima, though the bound is not always tight.



It appears that the energy at which the majority of marginal optima are located E_{maj} may be a useful lower bound on the performance of gradient descent. However, there is no principled reason for this, and the empirical evidence is limited to a small subclass of models right now.