# Effect of Cost Function Complexity and Dimensionality on Newton's Method Convergence Rate

Scott Reid
*Stanford University*

(Dated: December 16, 2017)

## INTRODUCTION

Today, Newton's Method is commonly used in machine learning and optimization to locate local minima of cost functions. Because it is such a general tool, the cost functions that is implemented on can range widely in terms of both their dimensionality and their complexity. However, little is understood about how Newton's Method performs on cost functions with different dimensionalities and complexities. In the simplest case, where the cost function is a paraboloid, one can show that Newton's Method converges in a single iteration to the minimum. We can increase the complexity of our cost functions by considering perturbative deviations from the paraboloid cost function. Intuitively, one should expect that as we increase the cost function complexity, the convergence rate will decline. However, as of now, there is no simple way to parametrize the complexity of a cost function, and thus, little is known about the way that cost function complexity affects the convergence of Newton's Method.

In this paper, I introduce a method of generating perturbed cost functions which allow us to systematically deviate from the simple paraboloid case. These perturbed cost functions are parametrized by a parameter $N$, which describes the number density of gaussian pertubation sites and is thus related to the complexity of the generated cost functions. By empirically fitting a model to convergence rate data, I find that the convergence rate depends strongly on both $N$ and $D$ (the dimensionality of the cost function). This dependence on $N$ and $D$ leads me to the hypothesis that the convergence rate of Newton's Method depends directly on the mean spacing between pertubations.

## I. GENERATING COMPLEX COST FUNCTIONS

I generate a family of cost functions where the complexity and dimensionality can be systematically varied. I do this by multiplying a $D$-dimensional parabola by a randomly generated pertubation function. The unperturbed cost function is thus the $D$-dimensional parabola. For $\boldsymbol{x} \in \mathbb{R}^d$, $J_o(\boldsymbol{x})$ is given by

$$J_o(\boldsymbol{x}) = ||\boldsymbol{x}||^2. \tag{1}$$

This cost function has a single global minimum at $\boldsymbol{x} = \boldsymbol{0}$ with $J_o(\boldsymbol{x}) \geq 0$.

I define my pertubation functions $P_N(\boldsymbol{x})$, where $N$ is the number of pertubations, as follows

$$P_N(\boldsymbol{x}) = 1 + A\sum_{i=1}^{N}(-1)^i \exp((\boldsymbol{x} - \boldsymbol{\mu}_i)^2/2\sigma^2) \tag{2}$$

The pertubation is the sum of $N$ gaussians with equal magnitude $A$ and variances $\sigma^2$. The centers of the gaussians $\boldsymbol{\mu}_i$ are randomly sampled from a uniform distribution over the ball of unit radius in $\mathbb{R}^d$ centered about the origin. The inclusion of 1 in the sum ensures that when $N = 0$, $P_0(\boldsymbol{x}) = 1$ and hence $P_0(\boldsymbol{x}) \times J_o(\boldsymbol{x}) = J_o(\boldsymbol{x})$.

I wanted to ensure that the value of $\boldsymbol{x}$ which minimizes the perturbed cost function remained $\boldsymbol{0}$. To ensure this, I restricted the amplitudes $|A| < 1$ and defined $\tilde{P}_N(\boldsymbol{x}) = \min(1 + A, \max(1 - A, P_N(\boldsymbol{x})))$. This ensures that $\tilde{J}_N(\boldsymbol{x}) \geq (1 - |A|)J_o(\boldsymbol{x})$. Since $(1 - |A|)$ is positive, we can be sure that $\tilde{J}_N(\boldsymbol{x})$ is minimized by $\boldsymbol{x} = \boldsymbol{0}$ with a minimum value of 0.

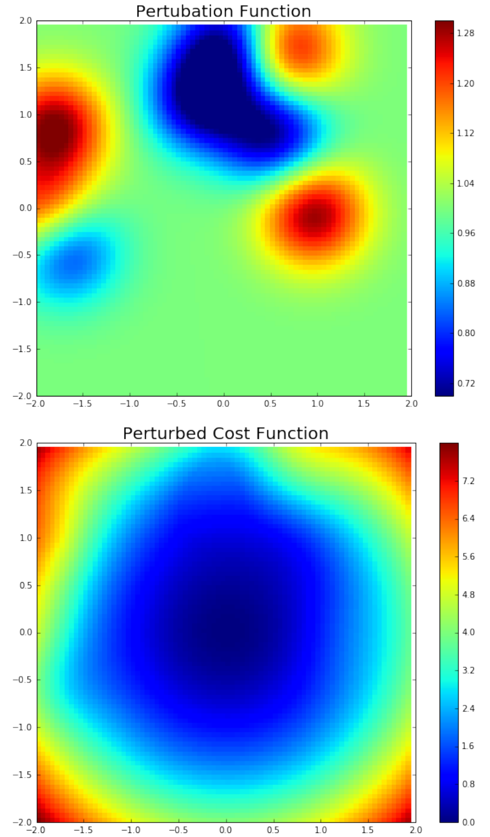

**Fig. 1** An example of a pertubation function and perturbed cost function in 2 dimensions for $N = 8$.

## II. MEASURING CONVERGENCE

In each iteration of Newton's method, the position of $\boldsymbol{x}$ is updated to $\boldsymbol{x}^{(\boldsymbol{t}+\boldsymbol{1})} := \boldsymbol{x}^{(\boldsymbol{t})} - H^{-1}\nabla J(\boldsymbol{x})$ where $H$ is the hessian matrix. With each iteration $t$ of Newton's method, the value $J(\boldsymbol{x}^{(t)})$ is reduced from its value during the previous iteration. We assume that the cost as a function of iteration exponentially decays, i.e. $J(\boldsymbol{x}^{(\boldsymbol{t})}) \propto e^{-\gamma t}$. With this assumption, and under the condition that the minimum cost is zero, the convergence rate $\gamma$ is defined as:

$$\gamma = \log \frac{J(\boldsymbol{x}^{(\boldsymbol{t}-\boldsymbol{1})})}{J(\boldsymbol{x}^{(\boldsymbol{t})})} \tag{3}$$

Our assumption that the cost exponentially decays with iteration number matches well with measured convergence, as seen in the example plot below. In practice, I measure the convergence rate $\gamma$ by comparing the cost after one iteration with the initial cost. This allows for faster calculations because newton's method only needs to be iterated a single time.

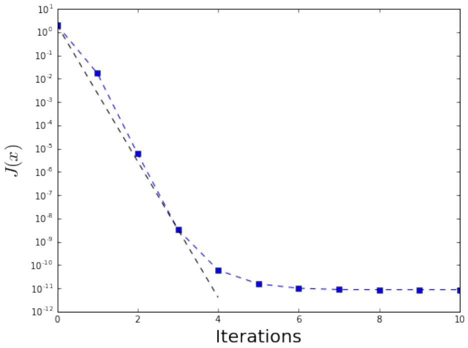

**Fig. 2** An example of a Newton's Method convergence for $d = 2$ and $N = 4$. Until the limit of computer accuracy is reached, the cost vs. iteration number is well modeled by exponential decay with a convergence rate $\gamma \approx 6.7$ (in black)

## III. METHODS AND RESULTS

To measure the effect of dimensionality and complexity on the convergence of Newton's Method, I measured the convergence rate $\gamma$ as defined in section II for a variety of cost functions. At each dimensionality $d \in [1, \dots 5]$ and each cost function pertubation number density $N \in [0, \dots 8]$, I average $\gamma$ measured for 200 cost functions with $\boldsymbol{x}^{(0)} = 1$.

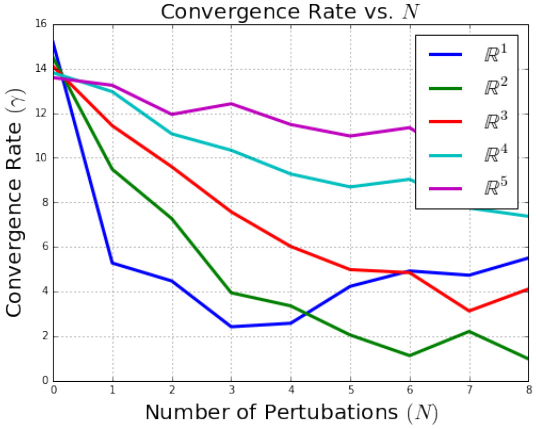

**Fig. 3** Convergence rate $\gamma$ for cost functions in $\mathbb{R}^1 - \mathbb{R}^5$ plotted against $N$. The upturn in the $\mathbb{R}^1$ is likely due to the onset of non-convexity.

As is seen in figure 3, the convergence rate at each dimensionality seems to exponentially decay with $N$, the cost function pertubation number density. The rate of this decay with dimensionality $\Gamma(D)$ is dependent on the dimensionality. We can empirically fit this behavior as:

$$\gamma(D, N) \approx \gamma_o(D) e^{-\Gamma(D)N} \tag{4}$$

From this expression, we can define $\Gamma(D)$ as:

$$\Gamma(D) = \log \frac{\gamma(D, N)}{\gamma(D, N+1)} \tag{5}$$

$\Gamma(D)$ can be understood as *the rate at which the pertubation number density affects the convergence rate of Newton's Method in $D$ dimensions.*

I plot $\Gamma(D)$ for $D \in [1, \dots 12]$. Values of $\Gamma(D)$ are calculated by comparing $\gamma(D, N)$ and $\gamma(D, N+1)$ as described by equation (5) for $N \in [1, \dots 5]$, averaged over 1500 trials.

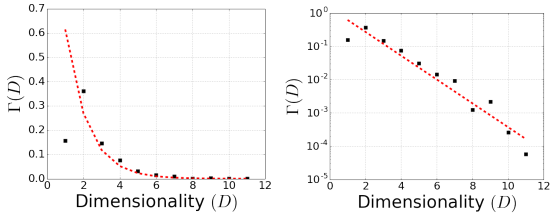

**Fig. 4** $\Gamma(D)$ is plotted both on linear and logarithmic axis against a curve of best fit $\Gamma(D) \approx 1.4 \times (2.3)^{-D}$

As discussed in the caption to figure 4, an empirical best fit for $\Gamma(D)$ is $\Gamma(D) \approx 1.4 \times (2.3)^{-D}$. This exponential fit performs extremely well. The exponential behavior is interpreted in the next section.

## IV. INTERPRETATION OF RESULTS

In section III, I show that the convergence rate $\gamma$ as a function of dimensionality $D$ and number of pertubations $N$ can be modeled as:

$$\gamma(D, N) \propto \exp\big(-1.4 \times (2.3)^{-D} \times N\big) \tag{6}$$

This expression, although it produces a very good fit to the experimental data (see figure 4), seems rather ad hoc, particularly in the absence of any theory that would justify such an expression. However, as I argue below, the unusual structure of the expression is likely due to the fact that we have chosen the wrong way to classify the complexity of the cost function surface. During this study, I have implicity assumed that $N$, the number density of pertubations, is the correct way to classify the cost function complexity. However, as I will argue below, expression (6) suggests that we should classify the cost function complexity in terms of the inverse of the average spacing between pertubations $n \propto N^{1/D}$.

Using expression (6), let us try to answer the following question: *what density of pertubations $N$ is required to reduce the convergence rate by a factor of $f$ in $D$ dimensions?* We find that:

$$N_f(D) = -\frac{\ln f}{1.4} \times 2.3^D \tag{7}$$

Thus, an *exponentially higher* density of pertubations is required to reduce the convergence rate by the same factor in, for example, 5 dimensions compared to 2 dimensions. However, we note that $n_f(D) \equiv N_f(D)^{1/D}$ is close to constant (based on our empirical model) for convergence rate reduction factor $f$.

It is interesting that $n \equiv N^{1/D}$ determines the convergence rate reduction factor. Since $N$ is linearly proportional the pertubation density (the number of pertubations in a unit volume), $n \equiv N^{1/D}$ is related the inverse of the average distance between pertubations by dimensional analysis. Based on this fact, it is likely that the convergence rate of Newton's Method is related more directly to the average distance between pertubations than to the number density of pertubations.

## CONCLUSION

In conclusion, I found that we can generate arbitrary cost functions in $\mathbb{R}^D$ which, in addition to be parametrized by the dimensionality $D$, are also parametrized by the number density of gaussian pertubations $N$. I analyzed the convergence rate of Newton's Method as a function of both $N$ and $D$ and found an empirical model to fit the measured convergence rates:

$$\gamma(D,N) \approx \gamma_o(D)e^{\Gamma(D)N} \tag{8}$$

$\Gamma(D)$ was found to exponentially decrease as a function of $D$. This behavior suggests that a more natural way of parametrizing the complexity of the generated cost functions is through the mean distance between pertubations. Further work will be required to establish if the convergence rate varies in a more natural way on the average spacing between pertubations. As it stands, this paper establishes that there is a strong connection between dimensionality, cost function complexity, and the convergence rate of Newton's Method.

One potential application of these results is the following diagnostic tool for measuring the complexity of a given cost function. In the real world, cost functions do not appear in the form that I generate them in this paper. However, expression (8) can be inverted to give the number density of pertubations $N$:

$$N = \frac{\ln \frac{\gamma_o(D)}{\gamma_{meas}}}{\Gamma(D)} \tag{9}$$

Thus, $N$, which is a good way of classifying the complexity of the cost function, can be deduced from the measured convergence rate. Experimentalists, engineers, and data scientists can calculate $\gamma_{meas}$ by iterating Newton's Method a single time from a variety of starting points. A similar systematic analysis of other optimization techniques (such as Gradient Descent), with cost functions generated in the same way as in this paper, can give information about which techniques are most appropriate as a function of the cost function dimensionality and complexity parameter $N$. Thus, by measuring $N$ with Newton's Method, engineers can select the most appropriate optimization technique for their given problem.

## RELATED WORKS

As far as I can tell, the methods that I followed in this paper are fairly novel. I could not find any papers which asked similar questions about how dimensionality and complexity affect the convergence rate of Newton's Method. As a result, there were no papers which directly influenced my paper. Instead, I list a few very general papers and books that are related to optimization and Newton's Method.

## REFERENCES

[1] Bertsekas, Dimitri P. *Nonlinear programming.* Athena scientific Belmont, 1999.

[2] Decker, DW and Keller, HB and Kelley, CT. *Convergence rates for Newtons method at singular points.* SIAM Journal on Numerical Analysis, 20(2):296–314, 1983.

[3] Kelley, Carl T. *Iterative methods for optimization.* SIAM, 1999.

[4] Smith, Steven T. *Optimization techniques on Riemannian manifolds.* Fields institute communications, 3(3):113–135, 1994.

[5] Epureanu, Bogdan I and Greenside, Henry S. *Fractal basins of attraction associated with a damped Newton's method.* SIAM review, 3(3):102–109, 1998.