How much gradient noise does a gradient-based
linesearch method tolerate?

by S. Gratton, Ph. L. Toint and A. Tröltzsch

Report NAXYS-04-2012   21 February 2012

How much gradient noise does a gradient-based linesearch method tolerate?

Serge Gratton[*], Philippe L. Toint[†], Anke Tröltzsch[‡]

21 February 2012

Abstract

Among numerical methods for smooth unconstrained optimization, gradient-based linesearch methods, like quasi-Newton methods, may work quite well even in the presence of relatively high amplitude noise in the gradient of the objective function. We present some properties on the amplitude of this noise which ensure a descent direction for such a method. Exploiting this bound, we also discuss conditions under which global convergence can be guaranteed.

1 Introduction

In the solution of smooth unconstrained optimization problems, one may anticipate that the presence of noise in the gradient of the objective function may create specific numerical difficulties, possibly jeopardizing convergence. Global convergence to local critical points has nevertheless been proved for a class of trust-region methods for smooth unconstrained optimization [1, 2, 4, 5] in which the gradient values are approximated rather than computed exactly. The situation is less clear for linesearch methods [6], but it may be observed that, assuming accurate objective function values, gradient-based linesearch methods, such as quasi-Newton methods, often work well for many problems, even in the presence of relatively high amplitude noise in the gradient. The purpose of this short paper is to shed light on why this is the case by attempting to model the gradient noise which is allowed by such a method.

[*]ENSEEIHT, 2, rue Charles Camichel, 31000 Toulouse, France. Email: serge.gratton@enseeiht.fr

[†]Namur Research Center for Complex Systems (NAXYS), FUNDP-University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@fundp.ac.be

[‡]CERFACS, 42 avenue G. Coriolis, 31057 Toulouse, France. Email: anke.troeltzsch@cerfacs.fr

2 Getting sufficient decrease in the presence of a noisy gradient

We consider the smooth nonlinear unconstrained minimization problem, i.e.

$$\min_{x\in\mathbb{R}^n} f(x) \tag{2.1}$$

where f is a continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}$. We also restrict our attention to linesearch methods of the form given by Algorithm 2.1, in which the k-th iterate is denoted by x_k and $g_k \stackrel{\text{def}}{=} \nabla_x f(x_k)$.

Algorithm 2.1: Linesearch minimization

Step 0: A starting point x_0 is given. Compute $f(x_0)$ and g_0, and set $k = 0$.

Step 1: Determine a search direction d_k such that $g_k^T d_k < 0$.

Step 2: Perform a linesearch along d_k, yielding x_{k+1}, $f(x_{k+1})$ and g_{k+1}.

Step 3: Increment k and go back to Step 1.

The hope is that the sequence $\{x_k\}$ generated by this algorithm asymptotically approaches a first-order critical point of problem (2.1), in the sense that

$$\lim_{k\to\infty} \|g_k\| = 0, \tag{2.2}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$. We then say that the algorithm is globally convergent.

2.1 Deterministic properties

It is well-known that the descent condition stated in Step 1 of this algorithm is generally insufficient to guarantee (2.2) from arbitrary starting points, even if the linesearch used in Step 2 satisfies the standard conditions (see, for instance, page 116 ff. in [6]). However, this desired convergence property may be ensured if one is ready to strengthen the descent condition of Step 1 and assume that the angle formed by the search direction d_k and the steepest descent direction $-g_k$ is uniformly bounded away from 90 degrees. This was formalized by Zoutendijk [11] as the condition that

$$\cos\varphi_k \geq \delta > 0, \quad \text{for all } k, \tag{2.3}$$

where δ is a positive constant and

$$\cos \varphi_k = \frac{-g_k^T d_k}{\|g_k\|\|d_k\|} \tag{2.4}$$

is the cosine of the said angle. Thus, if condition (2.3) holds, the direction d_k has sufficient decrease to guarantee (2.2) with a proper linesearch scheme. Let us know assume that g_k is unknown, but that a noisy gradient $\tilde{g}_k$ is available instead, and review the consequences of this simple change.

Steepest descent method

Following (2.3), the noisy variant of the steepest descent method chooses the direction $d_k = -\tilde{g}_k$. Therefore, if condition

$$\frac{g_k^T \tilde{g}_k}{\|g_k\|_2 \|\tilde{g}_k\|_2} \geq \delta, \tag{2.5}$$

holds, the negative noisy gradient direction remains a sufficient descent direction. Furthermore, defining Δg_k, the noise in the gradient, by $\tilde{g}_k \overset{\text{def}}{=} g_k + \Delta g_k$, we obtain the following property.

Proposition 2.1 *Assuming that, for all $k \geq 0$, the norm of the gradient noise satisfies*

$$\|\Delta g_k\|_2 \leq \frac{1-\delta}{1+\delta}\|g_k\|_2, \tag{2.6}$$

then condition (2.5) holds for all k and the noisy steepest descent algorithm is globally convergent.

In the interest of readability we will drop the iteration count k in the following proofs.

Proof. We know that

$$\|g\|_2\|\Delta g\|_2 \geq |g^T \Delta g| \geq -g^T \Delta g.$$

By combining these with (2.6) multiplied by $\|g\|_2$, we obtain that

$$-g^T \Delta g + \delta\|g\|_2\|\Delta g\|_2 \quad \leq \quad \|g\|_2^2 - \delta\|g\|_2^2$$

and then that

$$\delta\|g\|_2^2 + \delta\|g\|_2\|\Delta g\|_2 \quad \leq \quad \|g\|_2^2 + g^T \Delta g.$$

Using now that $\|g + \Delta g\|_2 \leq \|g\|_2 + \|\Delta g\|_2$ by the triangle inequality and that $\|g\|_2^2 = g^T g$, we deduce that

$$\delta\|g\|_2\|g + \Delta g\|_2 \leq g^T g + g^T \Delta g,$$

which, in turn, gives that $\cos \varphi \geq \delta > 0$. $\qquad\square$

General quasi-Newton method

If we now consider using the noisy gradient $\tilde{g}_k$ in a quasi-Newton framework, that is choosing $d_k = -H_k^{-1}\tilde{g}_k$ for some symmetric positive definite matrix H_k, we may verify that condition (2.3) becomes

$$\frac{g_k^T H_k^{-1}\tilde{g}_k}{\|g_k\|_2 \|H_k^{-1}\tilde{g}_k\|_2} \geq \delta, \tag{2.7}$$

In this case, it is possible to derive the following property.

Proposition 2.2 *Assuming that the matrices H_k are positive definite with uniformly bounded condition numbers κ, that $\delta \leq \frac{1}{\sqrt{\kappa}}$, and that the noise satisfies*

$$\|\Delta g_k\|_2 \leq \frac{1 - \delta\sqrt{\kappa}}{1 + \delta\sqrt{\kappa}} \frac{\|g_k\|_2}{\sqrt{\kappa}} \tag{2.8}$$

for all $k \geq 0$, then condition (2.7) holds and the corresponding quasi-Newton algorithm is globally convergent.

Proof. We know from Property 2.1 that

$$\text{if } \|\Delta \hat{g}\|_2 \leq \frac{1 - \delta}{1 + \delta}\|\hat{g}\|_2, \quad \text{then} \quad \frac{\hat{g}^T(\hat{g} + \Delta\hat{g})}{\|\hat{g}\|_2 \|(\hat{g} + \Delta\hat{g})\|_2} \geq \delta.$$

If we now substitute $H^{-\frac{1}{2}}g = \hat{g}$ and $H^{-\frac{1}{2}}\Delta g = \Delta\hat{g}$, we obtain that

$$\text{if } \|H^{-\frac{1}{2}}\Delta g\|_2 \leq \frac{1-\delta}{1+\delta}\|H^{-\frac{1}{2}}g\|_2, \quad \text{then} \quad \frac{g^T H^{-1}(g + \Delta g)}{\|H^{-\frac{1}{2}}g\|_2 \|H^{-\frac{1}{2}}(g + \Delta g)\|_2} \geq \delta.$$

The right hand side of this statement can be reformulated by considering the singular values of $H^{\frac{1}{2}}$ and $H^{-\frac{1}{2}}$ and using the facts that $\sigma_{min}(H^{-\frac{1}{2}})\|g\| \leq \|H^{-\frac{1}{2}}g\|_2$, that $\sigma_{min}(H^{\frac{1}{2}})\|H^{-1}(g+\Delta g)\|_2$ and that $\sigma_{min}(H^{-\frac{1}{2}})\sigma_{min}(H^{\frac{1}{2}}) = 1/\kappa(H^{\frac{1}{2}})$. As a consequence, after dividing the complete if-statement by $\|H^{-\frac{1}{2}}\|_2$ and using that $\|H^{-\frac{1}{2}}\Delta g\|_2 \leq \|H^{-\frac{1}{2}}\|_2\|\Delta g\|_2$, we obtain that

$$\text{if } \|\Delta g\|_2 \leq \frac{1-\delta}{1+\delta}\frac{\|H^{-\frac{1}{2}}g\|_2}{\|H^{-\frac{1}{2}}\|_2}, \quad \text{then} \quad \frac{1}{1/\kappa(H^{\frac{1}{2}})}\frac{g^T H^{-1}(g + \Delta g)}{\|g\|_2\|H^{-1}(g + \Delta g)\|_2} \geq \delta,$$

which gives, using the definition of $\cos\varphi$ from (2.4), that

$$\text{if } \|\Delta g\|_2 \leq \frac{1-\delta}{1+\delta}\frac{\|H^{-\frac{1}{2}}g\|_2}{\|H^{-\frac{1}{2}}\|_2}, \quad \text{then} \quad \cos\varphi \geq \frac{\delta}{\kappa(H^{\frac{1}{2}})}.$$

By substituting $\delta' = \delta/\kappa(H^{\frac{1}{2}})$, we thus deduce that

$$\text{if } \|\Delta g\|_2 \leq \frac{1 - \delta'\kappa(H^{\frac{1}{2}})}{1 + \delta'\kappa(H^{\frac{1}{2}})}\frac{\|H^{-\frac{1}{2}}g\|_2}{\|H^{-\frac{1}{2}}\|_2}, \quad \text{then} \quad \cos\varphi \geq \delta'.$$

4

Multiplying the numerator and denominator of the right hand-term of the if-statement by $\|H^{\frac{1}{2}}\|_2$ and using then that $\|H^{\frac{1}{2}}\|_2\|H^{-\frac{1}{2}}g\|_2 \geq \|H^{\frac{1}{2}}H^{-\frac{1}{2}}g\|_2 = \|g\|_2$ and that $\|H^{\frac{1}{2}}\|_2\|H^{-\frac{1}{2}}\|_2 = \kappa(H^{\frac{1}{2}})$, we obtain that

$$\text{if} \quad \|\Delta g\|_2 \leq \frac{1 - \delta'\kappa(H^{\frac{1}{2}})}{1 + \delta'\kappa(H^{\frac{1}{2}})} \frac{\|g\|_2}{\kappa(H^{\frac{1}{2}})}, \quad \text{then} \quad \cos\varphi \geq \delta'.$$

The proof is completed by observing that $\kappa(H^{\frac{1}{2}}) = \sqrt{\kappa}$. $\qquad\square$

After having proved these properties and having observed that the required upper bound on the noise is (unsurprisingly) depending on the condition number of the involved Hessian matrix of the optimization problem, we may now have a closer look and perform a small test. The idea is to check the sign of the supposed descent direction (2.7) for different random values of Δg with different amplitudes. We observed that condition (2.7), of course, always held for Δg smaller than the bound from (2.8) but, somewhat surprisingly, that it also held very often for Δg bigger than the bound given by (2.8). This observation suggests that the necessary condition (2.8) might be often too stringent.

We illustrate this issue on a small test example (Powell badly scaled function [9]) with $x^* = (1.1 \cdot 10^{-5}, 9.1)$ and a uniform random gradient noise with standard deviation equal to 10^{-3}), assuming for the moment that Property 2.2 is a tight upper bound, which means that no noise would be allowed for an ill-conditioned problem and thus that quasi-Newton methods would break down in a noisy situation. In this setting, it seems possible to regularize the problem whenever the condition number is too big to allow for a higher amplitude noise in the problem. Applying this strategy to an existing line search BFGS method gave interesting results, as can be seen in Figure 2.1. We implemented an automatic regularization technique which checks at every iteration whether Property 2.2 holds or not. If it is not satisfied for the given noise level, the condition number is decreased by applying $H_k = H_k + \lambda I$ for an increasing λ until Property 2.2 holds. The iterates obtained by using such a regularization step are displayed in the right-hand side figure, while the left-hand side figure shows the minimization of the problem without taking care of the gradient noise.

As it can be seen in Figure 2.1, both methods terminated prematurely due to the noise but the regularized version stopped significantly earlier than the unregularized one, in contrast with what can be expected from Property 2.2. In fact, the regularization had a negative effect on the convergence path because steps become close to steepest descent steps and get very short due to the conditioning of the matrix.

This experience therefore confirms the overly stringent nature of Property 2.2, and suggests that another approach could be of interest.
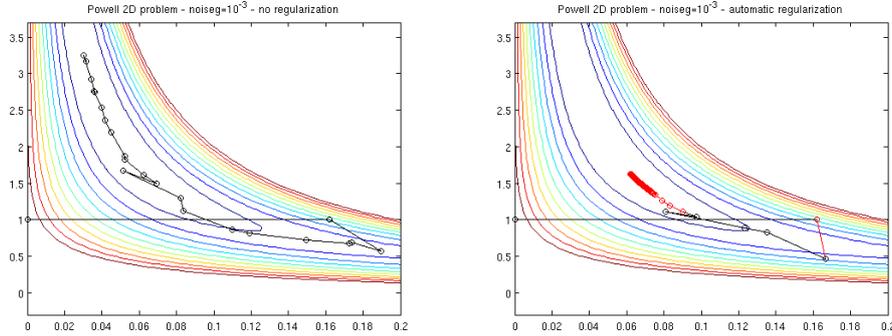
Figure 2.1: Convergence paths for Powell 2D problem with (right) and without (left) regularization

3 Statistical approach for a quasi-Newton method

One such approach is of statistical nature. If we now assume that the gradient noise is Gaussian $\Delta g \sim \mathcal{N}(0, \sigma^2 I)$ with a normal distribution with zero mean and standard deviation σ, can we find out the largest σ ensuring a (user specified) probability to obtain a descent direction?

Before investigating sufficient decrease, we start by considering the case where only strict descent is required (which corresponds to the condition in Step 1 of the linesearch algorithm, or to the case where $\delta = 0$ and the inequality is strict in (2.3)).

Strict descent direction
The condition for a quasi-Newton direction with an inexact gradient to be a strict descent direction is obviously that

$$-g_k^T H_k^{-1} \tilde{g}_k < 0. \tag{3.9}$$

The question is therefore to find for which standard deviation σ does (3.9) hold with a given probability. In other words, how big can the noise become in average such that (3.9) is still very likely to hold? As we know the distribution of Δg, we can rewrite (3.9) as

$$\textbf{(StD)} \quad \begin{array}{l} g^T H^{-1} \Delta g \geq -g^T H^{-1} g, \\ \text{where the left hand side is normally distributed with} \\ g^T H^{-1} \Delta g \sim \mathcal{N}(0, \sigma^2 \|H^{-1}g\|_2^2). \end{array} \tag{3.10}$$

The strict inequality from (3.9) can be relaxed in the statistical approach as the probability that exactly $-g_k^T H_k^{-1} \tilde{g}_k = 0$ is zero.

We now look for $\sigma_{\max}$, the largest σ ensuring (StD) with a given proba-

bility. From (3.10) we obtain that

$$P_{\text{StD}} \stackrel{\text{def}}{=} P[(\text{StD}) \text{ holds}] = \frac{2}{\sqrt{\pi}} \int_{-\frac{g^T H^{-1} g}{\sqrt{2}\sigma_{\max} \|H^{-1} g\|_2}}^{+\infty} e^{-t^2} \, dt. \qquad (3.11)$$

This can be expressed in terms of the complementary Gauss error function erfc as

$$P_{\text{StD}} = \frac{1}{2} \operatorname{erfc} \left[\frac{1}{\sqrt{2}\sigma_{\max}} \left(-\frac{g^T H^{-1} g}{\|H^{-1} g\|_2} \right) \right], \qquad (3.12)$$

or, in terms of the inverse complementary Gauss error function erfcinv, as

$$\operatorname{erfcinv}(2\, P_{\text{StD}}) = \frac{1}{\sqrt{2}\sigma_{\max}} \left(-\frac{g^T H^{-1} g}{\|H^{-1} g\|_2} \right). \qquad (3.13)$$

We may then extract the value of $\sigma_{\max}$ from this equality and obtain that

$$\sigma_{\max} = \frac{1}{\sqrt{2} \operatorname{erfcinv}(2\, P_{\text{StD}})} \left(-\frac{g^T H^{-1} g}{\|H^{-1} g\|_2} \right). \qquad (3.14)$$

This indicates that one could expect the allowed noise level $\sigma_{\max}$ to depend on the conditioning of the matrix H, the amplitude of the gradient and of course on the required probability of descent. We present a numerical illustration of these findings below.

Sufficient descent direction
Consider now the sufficient descent direction (2.7) and the question to find the largest σ ensuring (2.7) for a given probability. From the known distribution of Δg, we rewrite (2.7) as

$$\textbf{(SuD)} \qquad g^T H^{-1} g \geq \delta \|g\|_2 \|H^{-1}(g + \Delta g)\|_2 - g^T H^{-1} \Delta g. \qquad (3.15)$$

and we are now interested in finding the largest $\sigma_{\max}$ such that (3.15) holds for a given probability P_{SuD}.

Unfortunately, the distribution of the right-hand side of (3.15) can no longer be easily computed, and we therefore have resort to simulation to compute a corresponding value of $\sigma_{\max}$. This can be done by imposing a first σ_0, e.g., the one obtained from (StD), generating 10^7 examples of the random variable $\delta \|g\|_2 \|H^{-1}(g + \Delta g)\|_2 - g^T H^{-1} \Delta g$, and computing a CDF (cumulative distribution function) from which we then compute the probability that (SuD) holds. If this is lower than the desired probability, σ_0 is decreased by using

$$\sigma_{k+1} = \sigma_k - \frac{\sigma_k}{10}$$

and a new simulation is started, until a suitable $\sigma_{\max}$ is found.

Numerical illustration

We now wish to assess our analysis and apply it to a 10-dimensional test problem where the gradient and Hessian are defined as follows

$$g = [1, 2, \ldots, 10]/10, \qquad H = \begin{pmatrix} 1^t & & \ldots & 0 \\ & 2^t & & \\ \vdots & & \ddots & \vdots \\ 0 & & \ldots & 10^t \end{pmatrix}$$

where $t = 0, \ldots, 8$ represents the order of magnitude of κ, the condition number of H. We then compute how much gradient noise can be tolerated when the problem becomes more and more ill-conditioned. In Property 2.2 and in (SuD), we set $\delta = 10^{-5}$ for the experiments. The results are reported in Figures 3.2-3.4.
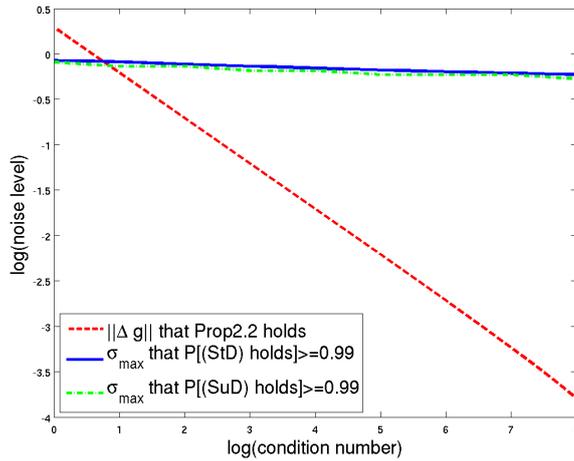

Figure 3.2: Critical noise level $\|\Delta g\|$ for Property 2.2 and $\sigma_{\max}$ for $P_{\text{StD}} = P_{\text{SuD}} = 0.99$

On each of these figures, we see the inversely proportional connection between the condition number of H and the bound on the noise level Δg such that Property 2.2 holds (dashed line). The other curves show the behaviour of the standard deviation $\sigma_{\max}$ of the Gaussian noise Δg. More specifically, the plain line shows the values $\sigma_{\max}$ ensuring that (StD) (strict descent) holds with probability 0.99. The dash-dotted line shows the values $\sigma_{\max}$ ensuring that (SuD) (simulated sufficient descent) holds with probability 0.99. We can see on Figure 3.2 that the plain line (StD) and the dash-dotted line (SuD) are very close together, but we observed that the bigger δ, the larger the distance between the two curves because the allowed noise level $\sigma_{\max}$ in (SuD) decreases if δ is increased in the condition.
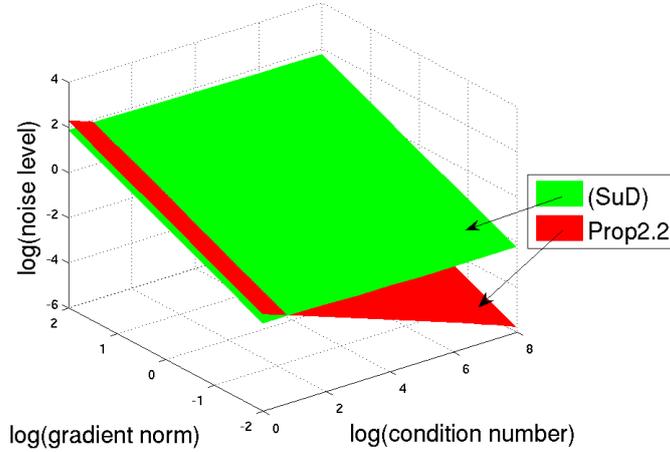
8

Figure 3.3: Critical noise level $\|\Delta g\|$ for Property 2.2 and $\sigma_{\max}$ for $P_{\mathrm{SuD}} = 0.99$ for varying gradient norm

In the example description above, a normalized gradient $\|g\|_\infty = 1$ is used. In the following, we are interested in the influence of the size of the gradient on the tolerated noise level in the optimization method. The case of (StD) is not represented on Figure 3.3 for visibility reasons, given that it is nearly undistinguishable from that of (SuD) for $\delta = 10^{-5}$. If we look at Figure 3.3 along the axis representing the log of the condition number, we see the pattern of Figure 3.2 with Property 2.2 and (SuD).

The effect of the gradient norm (with $\|g\|_\infty = 10^{-2}, \ldots, 10^2$) can be appreciated by looking at the same figure along the other horizontal axis, which shows the propotional dependence of the tolerated noise on the size of the gradient. This shows that, for well-conditioned problems, the tolerated noise level is nearly as big as the infinity norm of the gradient itself.

On Figure 3.4, all statistical properties are tightened to hold with a probability 0.9999. As expected, the curves show that the noise has to be smaller than for the smaller probability (see Figure 3.2).

Furthermore, we can observe, in these experiments, that our statistical approach allows for an amplitude of noise wich is sometimes considerably higher than the deterministic one, and also that the dependency on the condition number of H is very marginal in the statistical approach.

In practice. . .
Of course, the question remains of how to interpret the obtained probabilities of descent in terms of convergence of the corresponding minimization algorithm, assuming that the gradient noise at different iterations is independent. Theoretical global convergence guarantees of the type (2.2) now
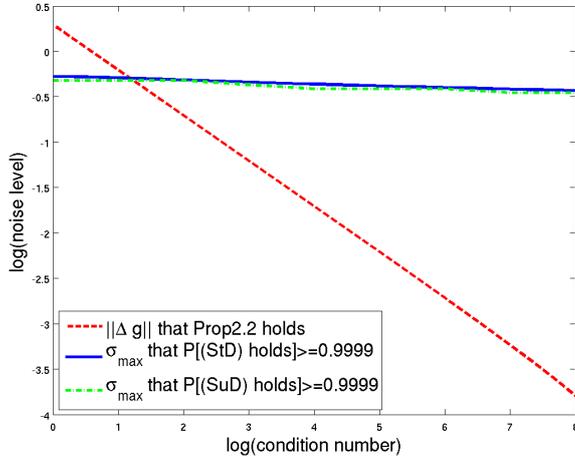
9

Figure 3.4: Critical noise level $\|\Delta g\|$ for Property 2.2 and $\sigma_{\max}$ for $P_{\mathrm{StD}} = P_{\mathrm{SuD}} = 0.9999$

depend on the infinite product of sufficient descent probablilities at each iteration. Obviously, if sufficient descent occurs with the same probability at every iteration and if this probability is less than one, then the product is zero and convergence almost certainly fails. However, it is possible to choose a sequence of increasing probablities of sufficient descent such that their infinite product converges to a value reasonably close to one, implying global convergence with probability equal to this value (see [10] for a discussion of suitable conditions on infinite products). But this theoretical perspective is again somewhat too pessimistic, as algorithms are never run for infinitely many iterations. A first reason is that practical methods always include some stopping criterion, and that complexity theory results (see [8, 7, 3] for examples) indicate that these stopping criteria may be achieved in a finite (albeit large) number of iterations. More significantly, stopping rules nearly always include a limit on the number of iterations which is often relatively small. This limit is typically even smaller (a few tens) in the frequent case where the only purpose of the minimization is to obtain significant decrease of the objective function (rather than convergence to a solution). If p is an iteration independent probability of sufficient descent resulting from our above discussion, the probability that the corresponding minimization algorithm finds a descent direction for the first m iterations is p^m, which might still be acceptable, in this perspective, for small m and p relatively close to one.

10

4   Numerical example in aerodynamics

We finally illustrate our theory on some real data from an aerodynamical shape optimization problem. In this test case, we consider a function measuring the pressure drag for a parameterized wing shape, which is obtained by approximately solving the Navier-Stokes equations by simulation. We sampled 100 function and gradient values by only changing one variable (the position of a normalized bump on the upside of the wing) in the interval $(0.7, 0.9)$. Figure 4.5 displays the adjoint state gradient (dotted line) and the gradient approximated by finite differences (plain line). By examining the zero ordinate, we note that the finite differences gradient reports a critical point close to 0.82, while the adjoint gradient places it close to 0.78. Assuming that the finite differences gradient gives a fairly good approximation of the real gradient, the picture reveals a very significant error/noise in the adjoint state gradient.
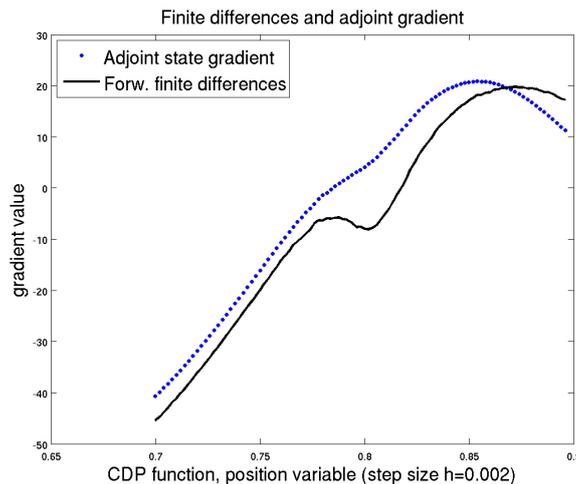

Figure 4.5: Finite differences gradient and adjoint state gradient of aerodynamic example

However, the adjoint can be computed much more cheaply (in terms of computing time) than a finite differences gradient, especially in higher dimensions. As a consequence, practitioners prefer to use the adjoint even though it clearly inherits a considerable amount of noise from the simulation process. In Figure 4.6, the attained noise level of the adjoint state gradient (computed error with respect to the finite differences gradient) is depicted as a dotted line, the amount of tolerated noise to ensure deterministic descent (Property 2.2) as a dashed line, and the amount of tolerated noise to ensure that (SuD) holds with probability 0.99 as a dotted line. Thus, we expect a gradient-based algorithm to converge if the values of the gradient error are
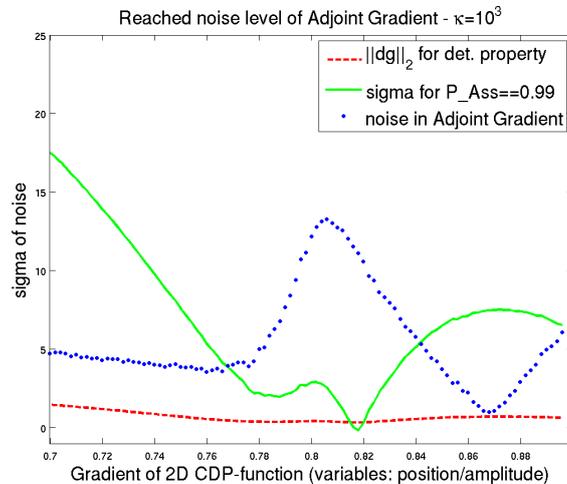
11

Figure 4.6: Error in adjoint state gradient and error bounds of quasi-Newton method

below one of these curves. In our example, this gives a hypothetically safe region up to an abscissa of 0.77 and beginning again from 0.84. The interval $[0.77, 0.84]$ indicates the region where we expect the iterates of the optimization algorithm to get into trouble. Note that the deterministic bound is violated for all reported values of the abscissa. Interestingly, these observations provide a nice *a posteriori* explanation of the unexpected results of a comparison of gradient-based optimization packages which we ran for other purposes (with a maximum number of iterations of 100), and in which the approximate solutions returned by the solvers for this one-dimensional problem ranged from $x^* = 0.7862$ to $x^* = 0.8180$ with corresponding function values from $f^* = 111.515$ to $f^* = 111.705$.

5   Conclusions

We addressed the question of how much noise in the gradient is tolerable for a quasi-Newton method without jeopardizing its descent properties. We established a deterministic property on the noise which ensures a sufficient descent direction for a globally convergent algorithm using an inexact gradient in this method. This property shows a strong dependence between the condition of the problem and the allowed noise. However, a simple experimental check with random noise of different amplitude has shown that this property indeed covers the worst case but seems to be overstringent in the average case. We also developed an alternative model assuming that the gradient noise has Gaussian distribution with mean zero and variance σ^2 in order to analyze what level of noise can be accomodated on average.

12

Our theory and experiments confirmed dependencies between the amplitude of the gradient norm and Hessian condition number and the tolerated noise in the gradient. However, the dependence on the allowed noise level on Hessian conditioning is much weaker in the statistical context than in the deterministic one.

Finally, we used our analysis to explain the unexpectedly good behaviour of quasi-Newton methods on a practical problem arising from aerodynamics.

Our development assumes exact function values and noisy gradient values. The extension to the more general situation where both, function and gradient information, are inexact is of course also of interest.

Acknowledgement

References

[1] R. G. Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM Journal on Numerical Analysis*, 28:251–265, January 1991.

[2] R. G. Carter. Numerical experience with a class of algorithms for non-linear optimization using inexact function and gradient information. *SIAM Journal on Scientific Computing*, 14(2):368–388, 1993.

[3] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming A*, 2010. DOI: 10.1007/s10107-009-0337-y.

[4] A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints. *SIAM Journal on Optimization*, 3(1):164–221, 1993.

[5] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, PA, USA, 2000.

[6] Jr. J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Classics in Applied Mathematics, 16).* Classics in Applied Mathematics. SIAM, Philadelphia, 1996.

[7] S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444, 2008.

[8] Yu. Nesterov. *Introductory Lectures on Convex Optimization.* Applied Optimization. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[9] M. J. D. Powell. A hybrid method for nonlinear equations. In P. Rabinowitz, editor, *Numerical Methods for Nonlinear Algebraic Equations.* Gordon and Breach, 1970.

[10] W. Rudin. *Real and complex analysis.* McGraw-Hill Book Co., New York, third edition, 1987.

[11] G. Zoutendijk. Nonlinear Programming, Computational Methods. In J. Abadie, editor, *Integer and Nonlinear Programming*, pages 37–86, North-Holland, Amsterdam, 1970.