Chapter 8

Trust-region methods

Trust-region method

Problem. Let $f : D \subset \mathbb{R}^n \to \mathbb{R}$ be a $\mathbb{C}^1$ function. To solve

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

it is necessary to find out points (vectors) $x^\star$ such that $\nabla f(x^\star) = 0$.

Strategy (Trust-region methods). The strategy is to start at a given vector $x_0 \in D$ and construct a sequence $\{x_k\}_{k \geq 0}$ where x_{k+1} is the *exact* solution of the problem

$$\min_{p \in \mathbb{R}^n} m_k(p) = f\left(x_k\right) + \nabla f\left(x_k\right) p + \frac{1}{2} p^T B_k p \quad \text{s.t. } ||p|| \leq r_k \quad (1)$$

with B_k being some symmetric matrix. The disc Δ_k centred at x_k and having radius r_k is called the trust-region.

Trust-region method

Remark. If we use Taylor expansion at $x = x_k$, up to order two, we have that

$$f(x) \approx m_k(p) := f(x_k) + (\nabla f(x_k))^T p + \frac{1}{2} p^T Hf(x_k) p.$$

So, using $B_k = Hf(x_k)$ is a natural choice and the method is then called Newton trust-region method.

Lemma. Consider the above notation. If B_k is positive definite and $||B_k^{-1} \nabla f(x_k)|| \leq r_k$ then the solution of (1) is given by

$$p_k = -B_k^{-1} \nabla f(x_k).$$

(The unconstrained minimum of the quadratic map)

Trust-region strategy/algorithm

Problem. Choosing the r_k.

Definition. Suppose we are in x_k. Suppose we are considering $r_k > 0$ to move forward. Let us define the following index

$$\rho_k := \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$$

where p_k is the solution of (1). The numerator is called actual reduction and the denominator is called predicted reduction.

Trust-region strategy/algorithm

Strategy.

(a) If $\rho_k < 0$ we reject the iteration $x_{k+1} = x_k$ and we decrease the trust region $r_k/2$ (since $m_k(0) \geq m_k(p_k)$).

(b) If $\rho_k \approx 1$ we accept the iteration $x_{k+1} = x_k + p_k$ and expand the trust region $2r_k$ (since $f \approx m_k$ in Δ_k)

- We (pre)-fix a bound for the size of r_k and pre-(fix) a bound ($\mu = 1/4$) to decide (a) and (b).

Solving (1)

Problem. To simplify notation we drop the k and x.

$$\min_{p \in \mathbb{R}^n} m(p) = f + (\nabla f)^T p + \frac{1}{2} p^T B\, p \quad \text{s.t. } ||p|| \le r \tag{2}$$

Theorem. The vector $p^\star$ solves (4) if and only if $p^\star$ is feasible and there is a scalar $\lambda \ge 0$ such that the following is satisfied.

$$\begin{aligned} &(B + \lambda\, \mathrm{Id})\, p^\star = -\nabla f(x), \\ &\lambda\,(r - ||p^\star||) = 0, \\ &(B + \lambda\, \mathrm{Id}) \quad \text{is positive semidefinite.} \end{aligned} \tag{3}$$

The Cauchy point (a way to approximate $p^\star$)

Definition. The Cauchy point (for the step k of the process) is

$$p_k^C := \tau_k \hat{p}_k$$

where

(a)

$$\hat{p}_k := \arg\min_{p \in \mathbb{R}^n} f(x_k) + (\nabla f(x_k))^T p \quad \text{s.t } ||p|| \leq r_k$$

(b)

$$\tau_k := \arg\min_{\tau \geq 0} m_k(\tau_k \hat{p}_k) \quad \text{s.t } ||\tau \hat{p}_k|| \leq r_k$$

The Cauchy point (close formula)

Excercise. According to the definition (a) above we obtain

$$\hat{p}_k = -\frac{r_k}{||\nabla f(x_k)||} \nabla f(x_k).$$

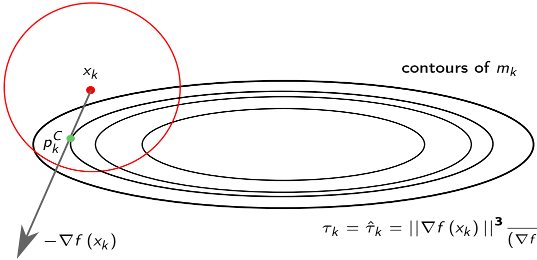
Excercise (More challenging). According to the definition (b) above we obtain

$$\tau_k = \begin{cases} 1 & \text{if} \, (\nabla f(x_k))^T B \nabla f(x_k) \leq 0 \\ \min\{1, \hat{\tau}_k\} & \text{otherwise} \end{cases}$$

with

$$\hat{\tau}_k := ||\nabla f(x_k)||^3 \frac{1}{r_k (\nabla f(x_k))^T B_k \nabla f(x_k)}.$$

The Cauchy point: Illustration



$$\tau_k = \hat{\tau}_k = ||\nabla f(x_k)||^{\mathbf{3}} \frac{r_k}{(\nabla f(x_k))^T B_k \nabla f(x_k)}$$

Excercise. Draw (illustrate) the case $\tau_k = 1$.

Improving the Cauchy point strategy: The dogleg method

Remark. A trust region method will be globally convergent if its steps p_k, $k \geq 0$, give a reduction in the model m_k that is at least δ-proportional ($\delta > 0$) to the one given by the Cauchy step.

Remark. Doing (just) Cauchy point at each step (is fine... but) is just implementing the steepest descent method with a particular choice of the step length.

Remark. It seems we might do better by considering methods for which the matrix B_k is more relevant in the choice of the optimal of the subproblem (1).

Improving the Cauchy point strategy: The dogleg method

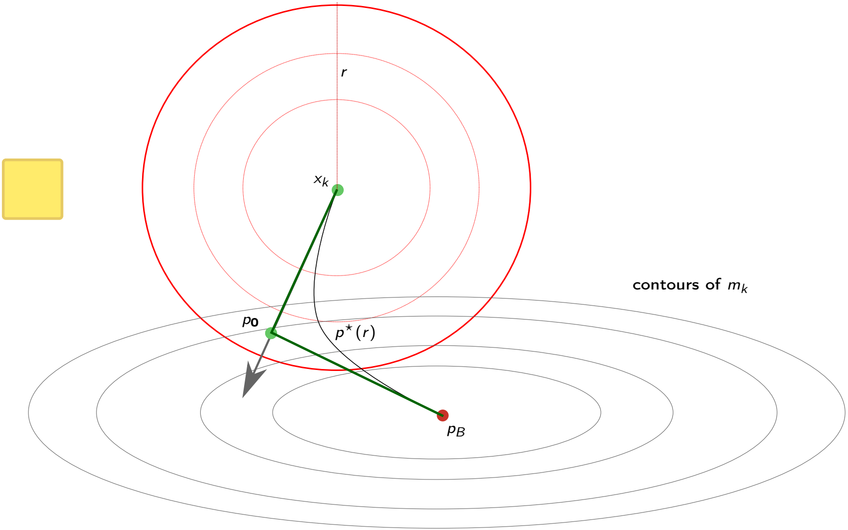
Remark. Assume B is positive definite. We drop the dependence on k and x.

$$\min_{p\in\mathbb{R}^n} m(p) = f + (\nabla f)^T p + \frac{1}{2}p^T B\ p \quad \text{s.t. } ||p|| \leq r \tag{4}$$

Remark 1. Let $p = -B^{-1}\ \nabla f$ the unconstrained minimizer of $m(p)$ (B is positive definite). If it is feasible (that is, $||p|| \leq r$) then the solution of (4) is precisely $p_B = -B^{-1}\ \nabla f$.

Remark 2. If r is small (at least comparable to p_B) then the might neglect the quadratic term (since (4) includes the restriction $||p|| < r$). Hence $p^\star(r) := -\left(\frac{r}{||\nabla f||}\right) \nabla f$. Of course this is a steepest descent method with a particular step-value.

Improving the Cauchy point strategy: The dogleg method



Improving the Cauchy point strategy: The dogleg method

Doglegs method. For small values of r we take the steepest descent method, that is, we follow $-\nabla f$ (the second order terms are not relevant). We do this up to

$$\alpha_0 = -\left(\frac{||\nabla f||^2}{(\nabla f)^T\, B\, \nabla f}\right) \qquad (p_0 := \alpha_0\, \nabla f).$$

(The value α_0 correspond to the minimum of $m(p)$ in the descent direction).

Improving the Cauchy point strategy: The dogleg method

Doglegs method. The dogleg method chooses p to minimize the model function $m(p)$ along the following path (instead of doing so through the exact path). See the figure above.

$$p(\tau) = \begin{cases} \tau p_0 & \text{if } 0 \leq \tau \leq 1 \\ p_0 + (\tau - 1)(p_B - p_0) & \text{if } 1 \leq \tau \leq 2 \end{cases}$$

Lemma. Let B positive definite. Then, the real valued function $||p(\tau)||$ is increasing and $m(p(\tau))$ is a decreasing function.

Corollary. The path $p(\tau)$ crosses $||p|| = r$ exactly once if $||p_B|| \geq r$ and never otherwise. Moreover the dogleg point p is p_B if $||p_B|| \leq r$ and the solution of $||p_0 + (\tau - 1)(p_B - p_0)|| = r^2$ otherwise.

Improving the Cauchy point strategy: The dogleg method

Lemma. Let B positive definite. Then, the real valued function $||p(\tau)||$ is increasing and $m\left(p(\tau)\right)$ is a decreasing function.

Proof. First we consider $\tau \in [0,1]$. Clearly $||p(\tau)|| = \tau ||p_0||$ is an increasing function of τ. Moreover we have that

$$\frac{d}{d\tau} m\left(p(\tau)\right) = (\tau - 1)\left(\frac{||\nabla f||^4}{\left(\nabla f\right)^T B \nabla f}\right) < 0,$$

so $m\left(p(\tau)\right)$ is a decreasing function of τ.

Second we consider $\tau \in [1,2]$. Define

$$h_1(\alpha) = \frac{1}{2}||p\left(1+\alpha\right)||^2 \quad \text{and} \quad h_2(\alpha) = m\left(p\left(1+\alpha\right)\right)$$

with $\alpha \in [0,1]$. Then the proof follows by showing that $h_1'(\alpha) \geq 0$ and $h_2'(\alpha) \leq 0$. We left the (non-trivial) details for the reader.

Global convergence of the trust-region methods

Fact. The global convergence of the trust-region methods (see above) depends on the approximate solution obtaining at least as much decrease in the model function $m(p)$ as the Cauchy point does (or a fixed positive fraction of it).

Proposition (decrease of the Cauchy point) The Cauchy point $p_k^C := \tau_k \hat{p}_k$ (the definitions of τ_k and $\hat{p}_k$ where given above) satisfies

$$m_k(0) - m(p_k^C) \geq \frac{1}{2} ||\nabla f(x_k)|| \min\left(r_k, \frac{||\nabla f(x_k)||}{||B_k||}\right). \quad (5)$$

Global convergence of the trust-region methods

Proof. We drop the dependence on k and x. Remember that

$$\hat{p} = -\frac{r}{||\nabla f||}\nabla f.$$

and

$$\tau = \begin{cases} 1 & \text{if} \quad (\nabla f)^T B \, \nabla f \leq 0 \\ \min\{1, \hat{\tau}\} & \text{otherwise} \end{cases}$$

with

$$\hat{\tau} := ||\nabla f||^3 \frac{r}{(\nabla f)^T B \, \nabla f}.$$

Global convergence of the trust-region methods

Proof. We split the proofs in cases.

Case $(\nabla f)^T B \nabla f \leq 0$. We have

$$m(p^C) - m(0) = m\left(-\frac{r\nabla f}{||\nabla f||}\right) - f = -r||\nabla f|| + \frac{r^2}{2||\nabla f||^2}(\nabla f)^T B\, \nabla f \leq -r||\nabla f||.$$

Case $(\nabla f)^T B \nabla f > 0$ and $||\nabla f||^3 \leq r(\nabla f)^T B\, \nabla f$.

$$\begin{aligned} m(p^C := \hat{\tau}\hat{p}) - m(0) &= -\frac{||\nabla f||^4}{(\nabla f)^T B\, \nabla f} + \frac{||\nabla f||^4}{2\left((\nabla f)^T B\, \nabla f\right)^2} = -\frac{||\nabla f||^4}{2\left((\nabla f)^T B\, \nabla f\right)} \\ &\leq -\frac{||\nabla f||^4}{2||B||||\nabla f||^2} = -\frac{||\nabla f||^2}{2||B||} \end{aligned} \tag{6}$$

Case $(\nabla f)^T B \nabla f > 0$ and $||\nabla f||^3 > r(\nabla f)^T B\, \nabla f$.

Global convergence of the trust-region methods

Theorem Let p_k be any vector such that $||p_k|| < r$. Assume also that

$$m_k(0) - m(p_k) \geq c_2 \left(m_k(0) - m(p_k^C) \right).$$

The p_k satisfies (5) with $c_1 = c_2/2$. In particular, if $p_k = p_k^\star$ is the exact solution of (1), then it satisfies (5) with $c_1 = 1/2$.

Proof. Exercise.