1 25th of October 2018 — A. Frangioni

We need to choose a descending direction, we have more choices than the opposite of the gradient.

We want the derivative to be almost zero, given a tolerance value.

Let's see another variant of line search.

1.0.1 Line search: second order approaches

Theorem 1.1. *Given $f \in C^2$, $\exists \varphi''(\alpha) = d^T \nabla^2 f(x + \alpha d) d$ and it's continuous.*

Proof. Via chain rule. □

Since we are looking for a point where the derivative $\varphi'(\alpha) = 0$, we may use the second derivative to write a model and, assuming to trust the model, it can be studied.

Definition 1.1 (Model−Newton tangent method)**.** *Our model, in this case is*

$$\varphi'(\alpha) \approx \varphi'(\alpha^k) + \varphi''(\alpha^k)(\alpha - \alpha^k)$$

In this context, solving $\varphi'(\alpha) = 0$ implies finding those α such that $\alpha = \alpha^k - \varphi'(\alpha^k)\varphi''(\alpha^k)$

Algorithm 1.1 Pseudocode for Newton method.

```
procedure LSNM(φ', φ'', α, ε)
    while (|φ'(α)|> ε) do
        α ← α − φ'(α)/φ''(α);
    end while
end procedure
```

We need to understand when and why $\varphi''(\alpha) \neq 0$ and when and why this method converges.

Theorem 1.2. *Let $\varphi \in C^3$ such that $\varphi'(\alpha_*) = 0$ and $\varphi''(\alpha_*) \neq 0$. $\exists\ \delta > 0$ s.t. if $\alpha^0 \in [\alpha_* - \delta, \alpha_* + \delta]$ then $\{\alpha^k\} \to \alpha_*$, with $p = 2$.*

Proof. We are in the hypothesis that the function φ is three times differentiable and we would like to prove that $\alpha^{k+1} - \alpha_* \to 0$

We want to compute how much the error is, if compared to the error at the previous iteration. Since $\varphi(\alpha_*) = 0$ we can use a dirty trick.

1. Since $\alpha^{k+1} \overset{(1)}{=} \alpha^k - \frac{\varphi'(\alpha_*)}{\varphi''(a^k)}$ and $\varphi'(\alpha_*) \overset{(2)}{=} 0$, we obtain:

$$\begin{aligned} \alpha^{k+1} - \alpha_* &\overset{(1)}{=} \alpha^k - \alpha_* - \frac{(\varphi'(\alpha^k))}{\varphi''(\alpha^k)} \\ &\overset{(2)}{=} \alpha^k - \alpha_* - \frac{(\varphi'(\alpha^k) - \varphi'(\alpha_*))}{\varphi''(\alpha^k)} \\ &= \frac{[\varphi'(\alpha^k) - \varphi'(\alpha_*) + \varphi''(\alpha^k) \cdot (\alpha^k - \alpha_*)]}{\varphi''(\alpha^k)} \end{aligned} \tag{1}$$

Where the term inside the square parenthesis is the first order model, centered in α_* computed in α^k.

2. Now we can use the first form of Taylor's formula, which says$\exists\ \beta \in [\alpha^k, \alpha^*]$ s.t. $\varphi'(\alpha_*) \overset{(3)}{=} \varphi'(\alpha^k)+\varphi''(\alpha^k)(\alpha^k-\alpha_*)+\varphi'''(\beta)\frac{(\alpha^k-\alpha_*)^2}{2}$. Let's see what happens to $\alpha^{k+1}-\alpha_*$:

$$\begin{aligned} \alpha^{k+1}-\alpha_* &\overset{(5)}{=} \frac{[\varphi'(\alpha^k)-\varphi'(\alpha_*)+\varphi''(\alpha^k)\cdot(\alpha^k-\alpha_*)]}{\varphi''(\alpha^k)} \\ &\overset{(3)}{=} \frac{-\varphi'''(\beta)}{2\varphi''(\alpha^k)}\cdot(\alpha^k-\alpha_*)^2 \end{aligned} \tag{2}$$

3. We can say that the quantity $2\varphi''(\alpha^k)$ doesn't become too small and that the numerator $\varphi'''(\beta)$ doesn't become too big. This is proved since $\exists\delta > 0$ s.t. $\varphi''(\alpha) \geq k_2 > 0$ and also $|\varphi'''(\beta)| \leq k_1 < \infty$. We can go on bounding the difference between α^{k+1} and α_* as follows: for $\alpha,\ \beta \in [\,\alpha_*-\delta\,,\,\alpha_*+\delta$

$$\begin{aligned} |\alpha^{k+1}-\alpha_*| &\overset{(6)}{=} \frac{-\varphi'''(\beta)}{2\varphi''(\alpha^k)}\cdot(\alpha^k-\alpha_*)^2 \\ &= |\alpha^{k+1}-\alpha_*| \leq [\frac{k_1}{2k_2}](\alpha^k-\alpha_*)^2 \end{aligned} \tag{3}$$

We may notice that $[\frac{k_1}{2k_2}]$ may be very large, but it's multiplied for $(\alpha^k-\alpha_*)^2$, which means that if we start close enough to α^* it's ok.

$$\begin{aligned} |\alpha^{k+1}-\alpha_*| = |\alpha^{k+1}-\alpha_*| &\leq [\frac{k_1}{2k_2}](\alpha^k-\alpha_*)^2 \\ = |\alpha^{k+1}-\alpha_*| &\leq [\frac{k_1}{2k_2}](\alpha^k-\alpha_*)\cdot(\alpha^k-\alpha_*) \end{aligned} \tag{4}$$

Where $[\frac{k_1}{2k_2}](\alpha^k-\alpha_*) < 1$, so $\frac{k_1(\alpha^k-\alpha_*)}{2k_2} \leq 1 \Longrightarrow \left|\alpha^{k+1}-\alpha_*\right| < \left|\alpha^k-\alpha_*\right|$ At this point, if we start from a point a^0 close enough to α^* (according to this formula) then $\{\,\alpha^k\,\} \to \alpha_*$ and the convergence is quadratic.

□

In the end, we may conclude that if we start from the right point we converge with a quadratic speed.

Problem: This solution makes us compute all the derivatives until the third one. We will now see a solution to this issue.

1.0.2 Exact line search: zeroth-order approaches

Can we do line search without computing derivatives at all? Following this approach we can circumvent the problem of the existence of derivatives. In the case of derivatives definite, it's better if we don't have to compute them.

Key idea: The more derivatives we have, the smallest number of points we need (second derivative → two points, third derivative → zero points). The opposite holds as well.

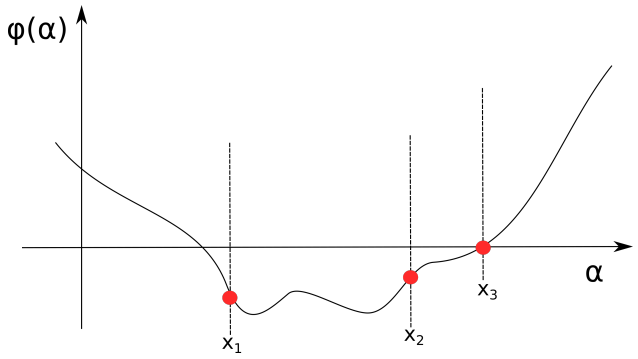

Figure 1.1: The interesting interval is $[x_1, x_2]$, since $\varphi(x_2) > \varphi(x_1)$ and we are allowed to exclude the interval $[x_3, +\infty)$ since the value in x_3 is bigger than $\varphi(x_2)$.

Obs: We have to minimize a function we know nothing about, except for its value in a point where we compute it. We would like to reduce to the interval which has as extremes the smallest point.

How can we choose these points? The idea is to choose the points that imply that the interval shrinks as fast as possible.

Obs: We have no guarantee that the interval that we are discarding doesn't contain a very deep minimum.

Elegant solution via golden ratio:

$$r = (\sqrt{5} - 1)/2 \ (\approx 0.618), \ r : 1 = (1 - r) : r$$

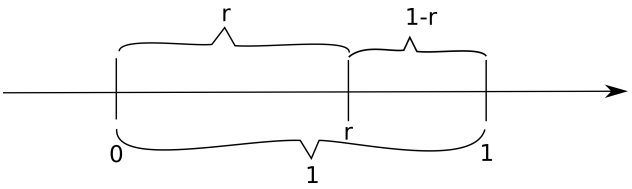

Figure 1.2: The relationship between r and $1 - r$ is $r \ : \ 1 = (1 - r) \ : \ r$.

Algorithm 1.2 Pseudocode for non differentiable functions for local minimum detection.

procedure LSGRM$(\varphi, \alpha, \varepsilon)$
  $\alpha_{i-} \leftarrow 0$; $\alpha_+ \leftarrow \alpha$;
  $\alpha'_- \leftarrow (1-r)\,\alpha$;
  $\alpha'_+ = r\,\alpha$;
  while $(\alpha_+ - \alpha_- \leq \varepsilon)$ **do** // note: not the same ε
    if $\varphi(\alpha'_-) > \varphi(\alpha'_+)$ **then**
      $\alpha_- \leftarrow \alpha'_-$;
      $\alpha'_- \leftarrow \alpha \leftarrow \alpha'_+$;
      $\alpha'_+ \leftarrow r(\alpha_+ - \alpha_-)$;
    else
      $\alpha_+ \leftarrow \alpha'_+$;
      $\alpha'_+ \leftarrow \alpha \leftarrow \alpha'_-$;
      $\alpha'_- \leftarrow (1-r)(\alpha_+ - \alpha_-)$;
    end if
  end while
end procedure

1.0.3 Inexact line search: Armijo-Wolfe

Key idea: Take your favourite line search (it also suggest a simpler one), and run it, but you don't have to wait for the derivative to become zero.

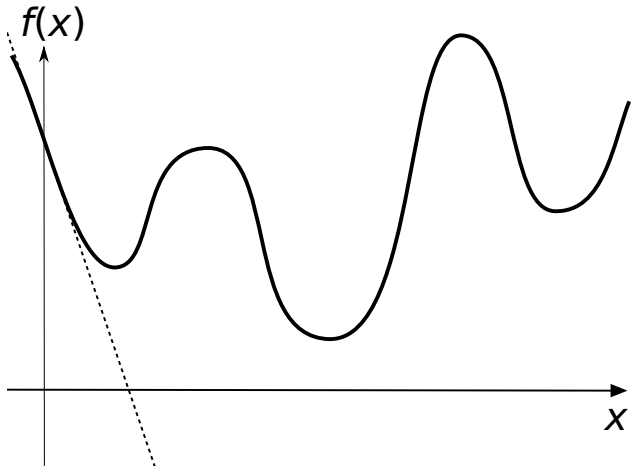

Figure 1.3: The dotted line represents the line which has the derivative as slope.

Definition 1.2 (Armijo condition)**.** *Let* $0 < m_1 < (\ll)1$

$$\varphi(\alpha) \leq \varphi(0) + m_1 \alpha \varphi'(0) \ (A)$$

Problem of Armijo condition: small steps satisfy the armijo condition, but make convergence very slow.

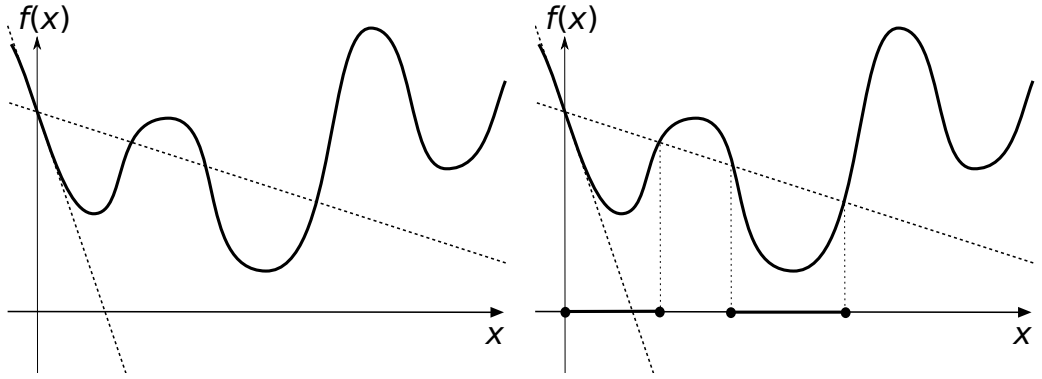

Figure 1.4: In the left picture Armijo condition chooses a new line which slope is still negative, but less stip than the original one. In the right one, the ranges where to search are highlighted.

We need another condition, in order to have a lower bound for the step size:

Definition 1.3 (Goldstein condition)**.** *Let* $m_1 < m_2 < 1$

$$\varphi(\alpha) \geq \varphi(0) + m_2 \alpha \varphi'(0) \ (G)$$

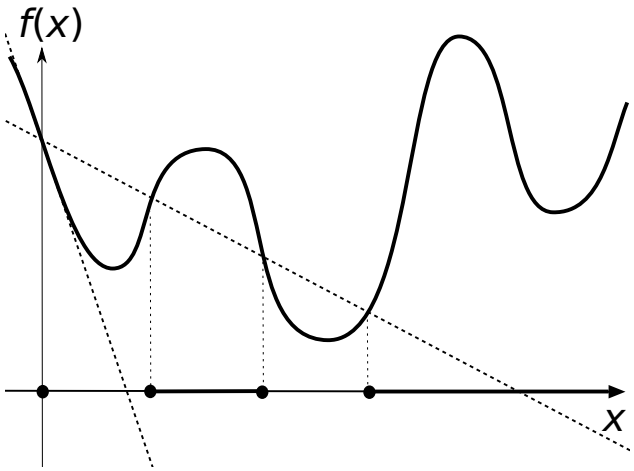

Figure 1.5: Goldstein condition chooses a new line which slope is still negative, less stip than the original one, but stipper than the one obtained by Armijo.

Problem: the point that satisfies both Goldstein and Armijo may not contain a local minimum.

To circumvent this problem another condition comes to help us.

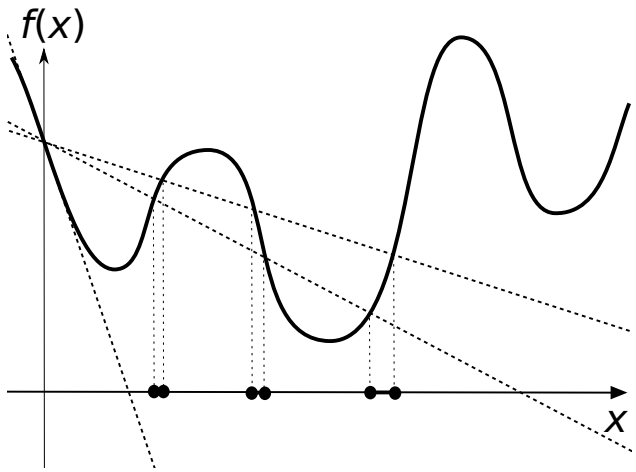

Figure 1.6: Here are the intervals that satisfy both Armijo and Goldstein conditions.

Definition 1.4 (Wolfe condition)**.** *Let* $m_1 < m_3 < 1$

$$\varphi'(\alpha) \geq m_3 \varphi'(0) \ (W)$$

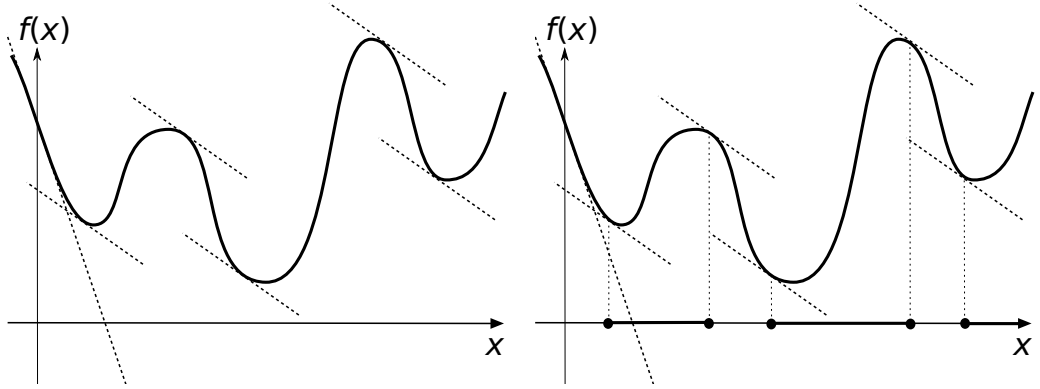

Figure 1.7: On the left Wolfe condition (which chooses derivatives that are substantially zero). On the right part the intervals selected by Wolfe.

Another issue of this conditions is that the derivative in the interval is quite big on the right side.

Definition 1.5 (Strong Wolfe condition)**.**

$$|\varphi'(\alpha)| \leq m_3 |\varphi'(0)| = -m_3 \varphi'(0)$$

Fact 1.3. *If* $\varphi'(\alpha) \not\gg 0$ *and (A)* $\cap$ *(W) / (W') then all local minima (& maxima) are captured unless* m_1 *too close to* 1 *(that's why usually* $m_1 \approx 0.0001$*)*

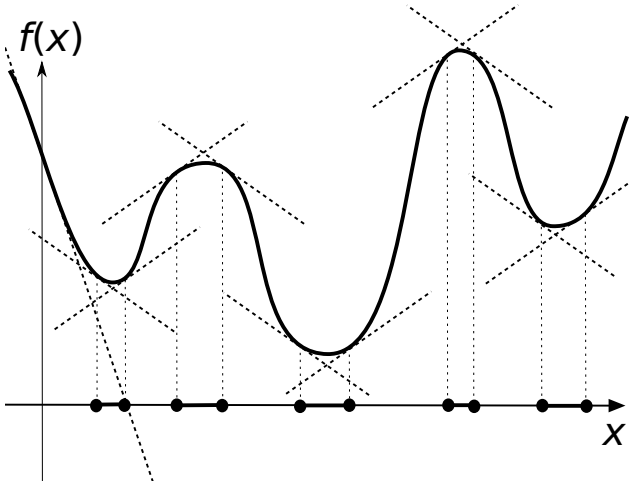

Figure 1.8: Strong Wolfe condition.

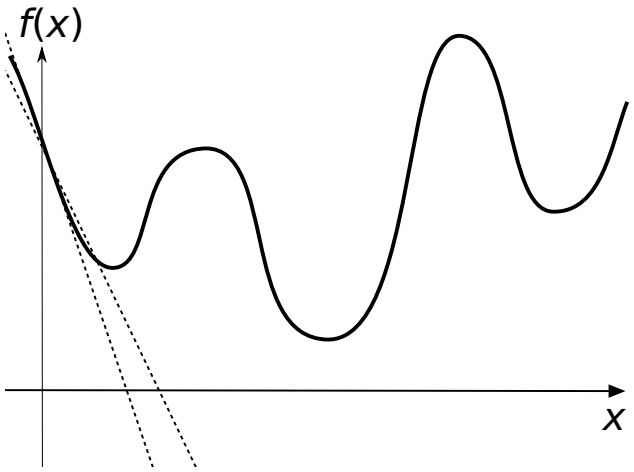

Figure 1.9: If the line is too close to the original one only a few points will satisfy Armijo condition which means that the intersection between Armijo and Wolfe will almost be empty.

The m_i are like the hyperparameters of machine learning. Less formally, if we choose an m_1 far enough from 1 everything works fine.

Theorem 1.4. *Let $\varphi \in C^1$ and $\varphi(\alpha)$ bounded below for $\alpha \geq 0$ then $\exists\ \alpha$ s.t. (A) $\cap$ (W') holds.*

Proof. $l(\alpha) = \varphi(0) + m_1\alpha\varphi'(0)$, $d(\alpha) = l(\alpha) - \varphi(\alpha) \Longrightarrow$
$d(0) = 0$, $d'(0) = (m_1 - 1)\varphi'(0) > 0$ $(m_1 < 1)$ □

0 and $\bar{\alpha}$ are the two roots of the function d, so we can use Rolle's theorem, in order to prove that the function d has a stationary point in the interval $[0, \bar{\alpha}]$.

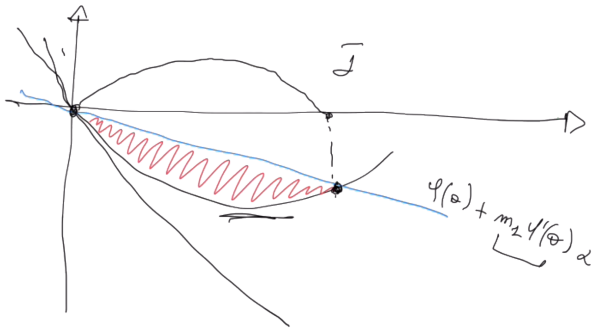

Figure 1.10: If the function isn't going to $-\infty$ the blue line and the function will meet again and we denote the value along the x axis $\bar{\alpha}$.

$$d(\alpha) = \varphi(0) + \alpha(m_1\varphi'(0)) - \varphi(\alpha)$$

$$d'(\alpha) = m_1\varphi'(0) + \varphi'(\alpha)$$

So $d'(\alpha^*)$ iif $\varphi'(\alpha^*) = m_1\varphi'(0)$. Then strong Wolfe requests that $|\varphi'(\alpha^*)| \leq m_1|\varphi'(0)|$.

How can we find such a point?

Algorithm 1.3 Pseudocode for backtracking line search.

```
1: procedure BLS(φ, φ', α, m_1, τ)
2:     while (φ(α) > φ(0) + m_1 α φ'(0)) do
3:         α ← τ α; τ < 1;
4:     end while
5: end procedure
```

- Fundamental assumption: ∇f Lipschitz $\implies$ φ' Lipschitz and L does not depend on x^i (**check**)
- Recall: $\exists\bar{\alpha}$ s.t. (A) holds $\forall\alpha \in]0, \bar{\alpha}]$ and $\varphi'(\bar{\alpha}) > m_1\varphi'(0) > \varphi'(0)$;
- φ' Lipschitz $\implies$ $\bar{\alpha}$ is "large" if $\|\nabla f(x^i)\|$ is:

$$L(\bar{\alpha} - 0) \geq \varphi'(\bar{\alpha}) - \varphi'(0) > (1 - m_1)(-\varphi'(0)) \implies \bar{\alpha} > (1 - m_1)\frac{\|\nabla f(x^i)\|\,|}{L}$$

  (recall $-\varphi'(0) = \|\nabla f(x^i)\|$);
- Fundamental trick: $\bar{\alpha}$ can $\searrow 0$, but only as fast as $\|\nabla f(x^i)\|$ does;
- Enough to prove that $\alpha^i \geq \bar{\alpha}$, or "not too smaller".

Now we can prove the following

Theorem 1.5. *If (A) ∩ (W) holds ∀i then either* $\{f(x^i)\} \to -\infty$ *or* $\{\|\nabla f(x^i)\|\} \to 0$.

Proof. **By contraddiction**, we assume $-\varphi'(0) = \|\nabla f(x^i)\| \geq \varepsilon > 0 \forall i$. Then

1. (W) $\Longrightarrow \alpha^i \geq \bar{\alpha} > (1 - m_1)\frac{\|\nabla f(x^i)\|}{L} \Longrightarrow \alpha^i \geq \delta > 0$;
2. (A) $\Longrightarrow f(x^{i+1}) \leq f(x^i) - m_1 \alpha^i \|\nabla f(x^i)\| \leq f(x^i) - m_1 \delta \varepsilon$;
3. So $\{f(x^i)\} \to -\infty$ (or $\{\|\nabla f(x^i)\|\} \to 0$).

□

Backtracking is similar: for simplicity, $\alpha = 1$ (input)
$\|\nabla f(x^i)\| > \varepsilon \forall i \Longrightarrow \bar{\alpha} > \delta > 0 \forall i$
$h = \min\{k : \tau^{-k} \leq \delta\} \Longrightarrow \alpha^i \geq \tau^{-h} > 0 \forall i \Longrightarrow f(x^{i+1}) \leq f(x^i) - m_1 \tau^{-h} \varepsilon \Longrightarrow \{f(x^i)\} \to -\infty$ or ↯.