Non-smooth bundle trust-region method

Dominikus Noll

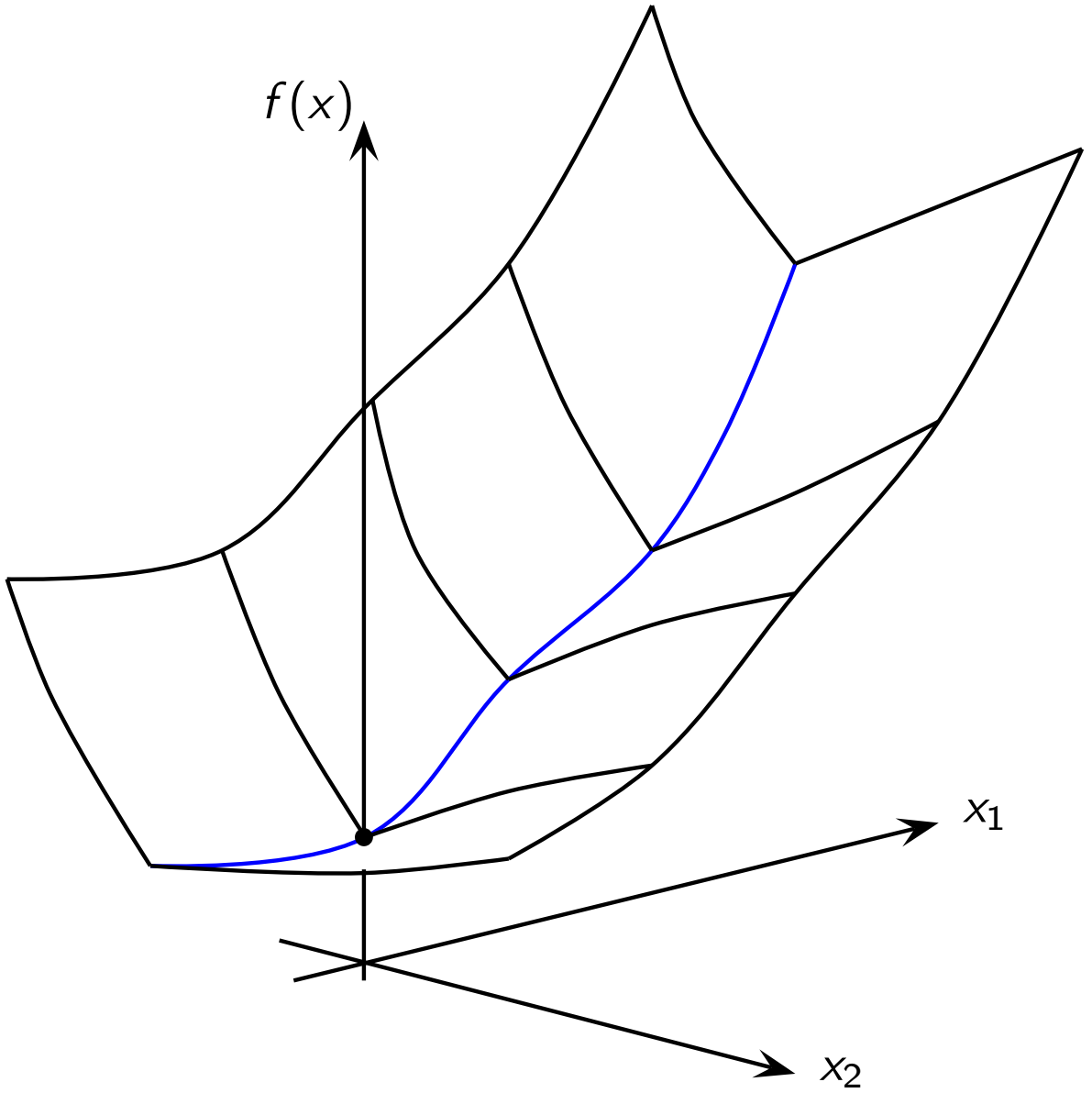
Université Paul Sabatier

Smooth optimization :

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

- Stopping test : $\dfrac{\|\nabla f(x^j)\|}{1 + |f(x^j)|} < \epsilon$
- Convergence to critical point $x^j \to x^*$, $\nabla f(x^*) = 0$

How about the non-smooth case ?


$f(x)$
x_1
x_2

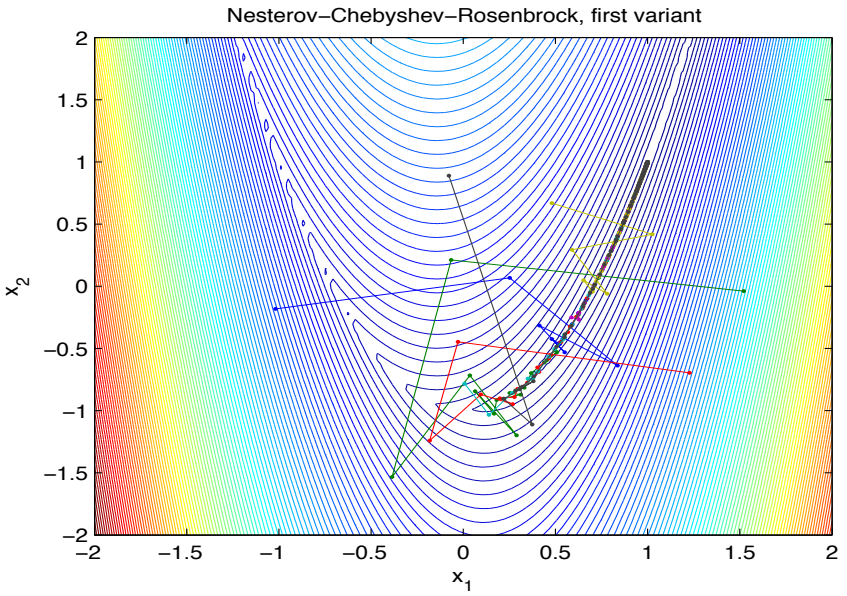
Arrange :

- f smooth outside blue Talweg.
- $\|\nabla f(x)\| \geq 1$ for x outside active manifold (= blue Talweg)
- $(0, 0)$ is minimum
- f may have concave features near minimum.
- Numerical method bound to never hit Talweg.

No subdifferential can give valid stopping test

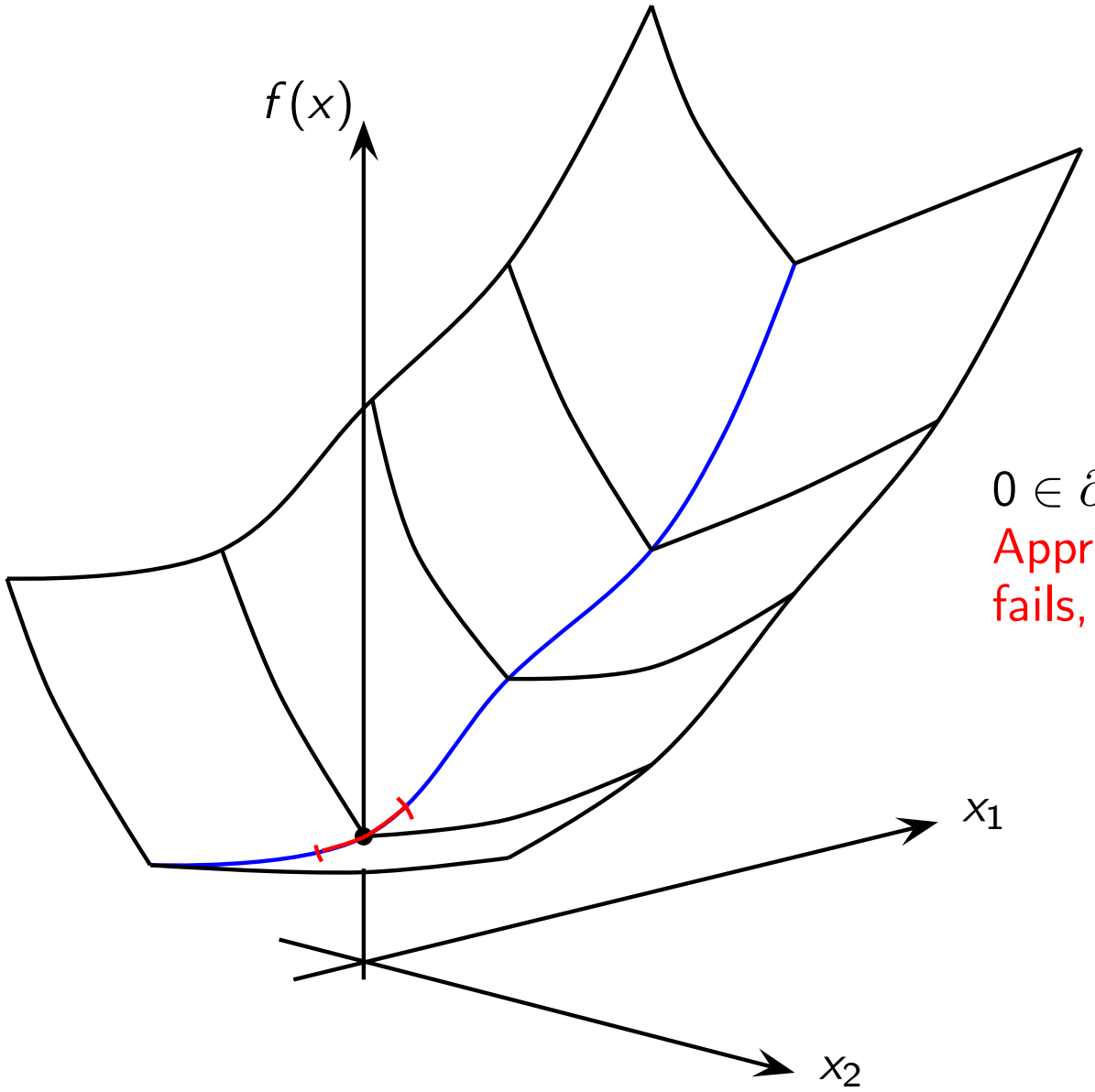
$$\|\partial f(x^j)\|_- = \inf\{\|g\| : g \in \partial f(x^j)\} \not\to 0$$

Nesterov's variant of Rosenbrock function :
$f(x) = \frac{1}{4}(x_1 - 2)^2 + |x_2 - 2x_1^2 + 1|$



Non-smooth BFGS

Gürbüzbalaban & Overton 2011, Lewis & Overton 2008-13



$0 \in \partial f(x^*) + \epsilon B$

Approximate optimality fails, too

Consequences : Fall back on other stopping criteria like

- Linesearch gives only marginal progress.
- Tangent program finds no trial steps which give sufficient progress.
- Progress tests like

$$\frac{\|x^{j+1} - x^j\|}{1 + \|x^j\|} < \mathrm{tol}, \qquad \frac{|f(x^{j+1}) - f(x^j)|}{1 + |f(x^j)|} < \mathrm{tol}$$

But are those justified ?

- Why not use stopping test of tangent program ?

Theorem

(Lewis 2003, Hare & Lewis 2003 – 2007)
The proximal point method detects the active manifold.

Theorem

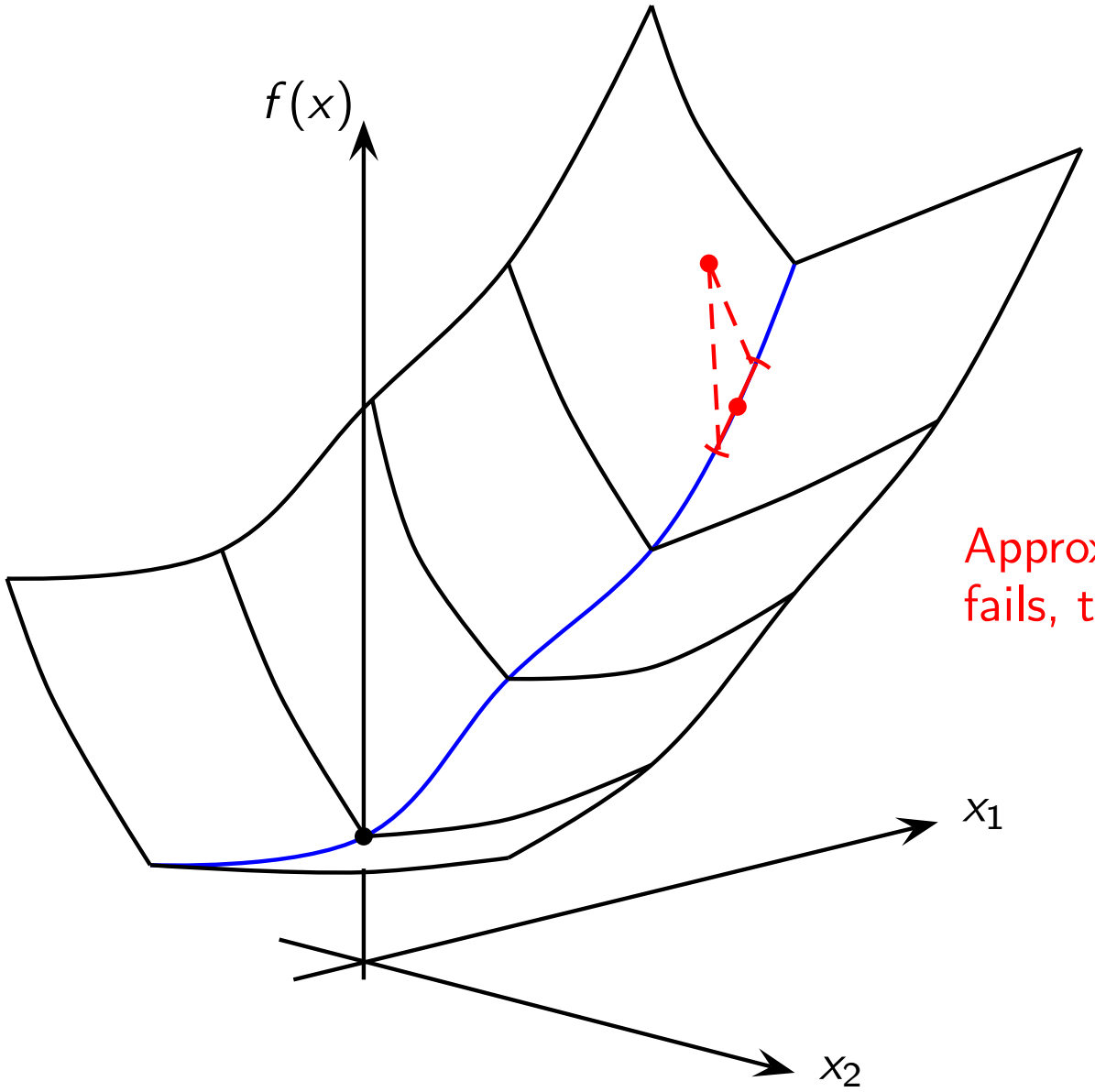
(Lewis 2003, Hare & Lewis 2003 – 2007)
The proximal point method detects the active manifold.

- Unfortunately, proximal point method is not a *method*.
- Tangent program

$$\min_{y \in \mathbb{R}^n} f(y) + \lambda_j \|y - x^j\|^2$$

  not practical
- But can compute points arbitrarily close to active manifold.
- Cannot use stopping test of tangent program :

$$0 \in \partial f(x^{j+1}) + \underbrace{\lambda_j (x^{j+1} - x^j)}_{\to 0} + \epsilon B$$



Approximate stopping fails, too

Observe :

- Principled difference between e.g. augmented Lagrangian method in smooth optimization, and proximal-point method in non-smooth optimization.
- Both run a NLP to optimality to compute a single iterate x^j, which is impossible.
- BUT, in the case of the AL we can stop early at an approximate solution based on an algorithmic criterion and still assure convergence.

  $\Longrightarrow$ *can make AL a practical method.*
- This is in general not possible with the non-smooth proximal point method.

Are there rigorous methods with convergence and stopping ?

- Convex bundle method (Lemaréchal).
- BT-methods (Zowe 1970s)
- Linesearch bundle method (Mifflin 1970s)
- Non-convex bundle method (since 2005)
- Bundle method for composite convex functions (Ruszczyński 2007, Sagastizábal-Hare 2009)

Today : *Bundle trust-region method*

Observe :

- Convex combination g_j^* of subgradients at trial points y^{kj} around serious iterates x^j converges to 0 :

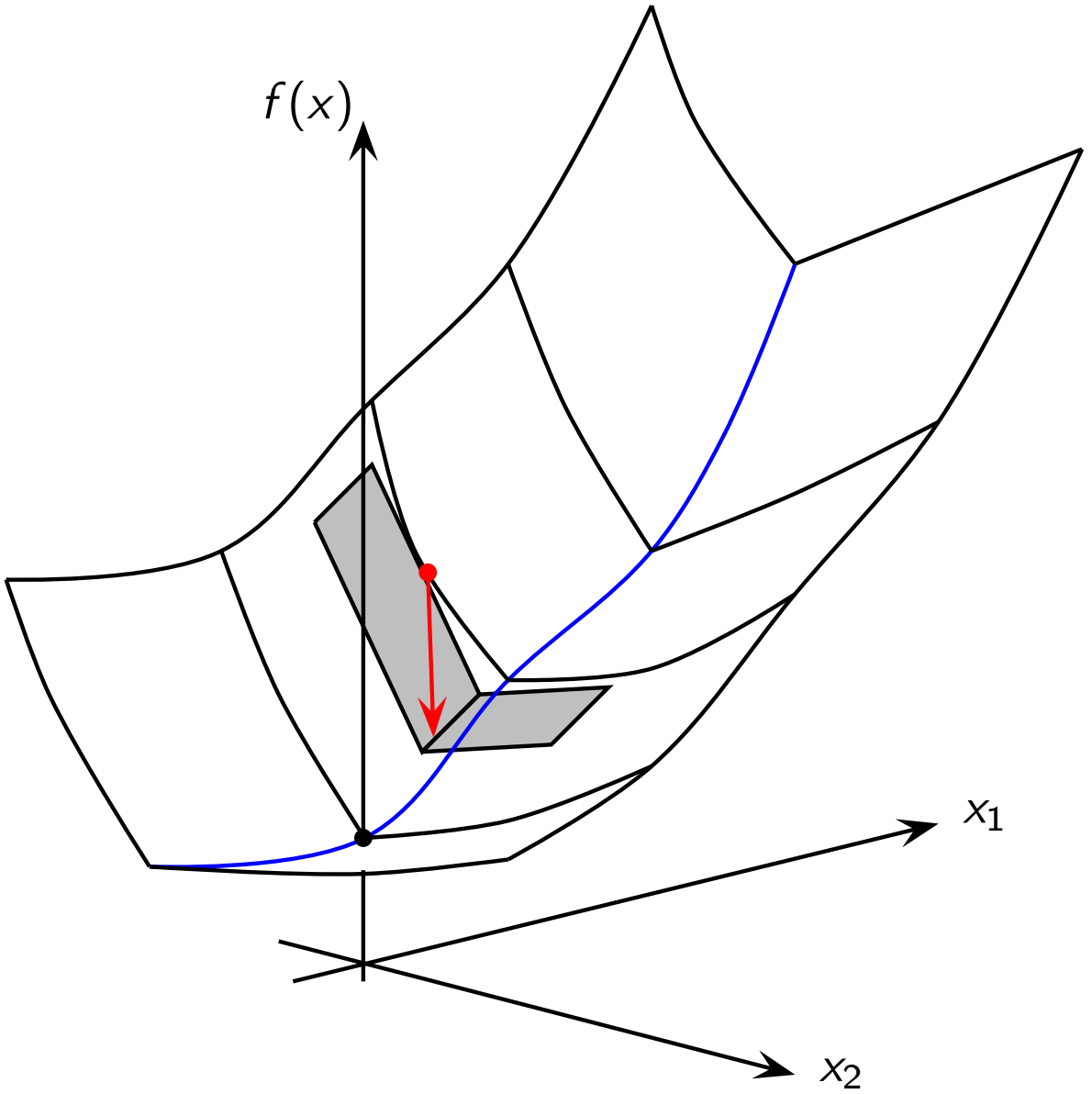
$$g_j^* = \sum_{j'=1, k \in I_{j'}}^{j} c \, \lambda_{kj'} g_{kj'} \to 0, \quad g_{kj'} \in \partial f(y^{kj'}) \quad \text{as } (j \to \infty),$$

$y^{kj'}$ trial points at serious iterates $x^{j'}$

(g_j^* = aggregate subgradient at x^j)

- Therefore need ∂f to be upper semi-continuous, closed and convex to be able to conclude $0 \in \partial f(x^*)$.
- Also need that minimum x^* satisfies $0 \in \partial f(x^*)$.
- Usually $g_{kj} = \nabla f(y^{kj})$. No need to compute ∂f nor $f'(x, d)$.

Can have a look what SQP does


$f(x)$
x_1
x_2

Observe :

- SQP does *not* try to reach the active manifold. It goes for its linearization, which is computable.
- SQP generates a convex combination of gradients at trial points which converge to 0.
- Minimize $\max_{i \in I} f_i(x)$:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f_i(x) \leq t, i \in I \end{array}$$

- KKT :

$$\nabla_{(t,x)} L(t, x, \lambda) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \sum_{i \in I}^{c} \lambda_i \begin{pmatrix} -1 \\ \nabla f_i(x) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$\Longrightarrow$ Clarke subdifferential

What is known about non-smooth trust-regions ?

Bundling included

- Ruszczyński 2007. Convex bundle trust-region method.
- Apkarian, Noll, Ravanbod 2015. Non-convex bundle trust-regions.

Without bundling

- Y. Yuan 1983. Trust-region method for $f = g \circ F$, g convex. Tangent program

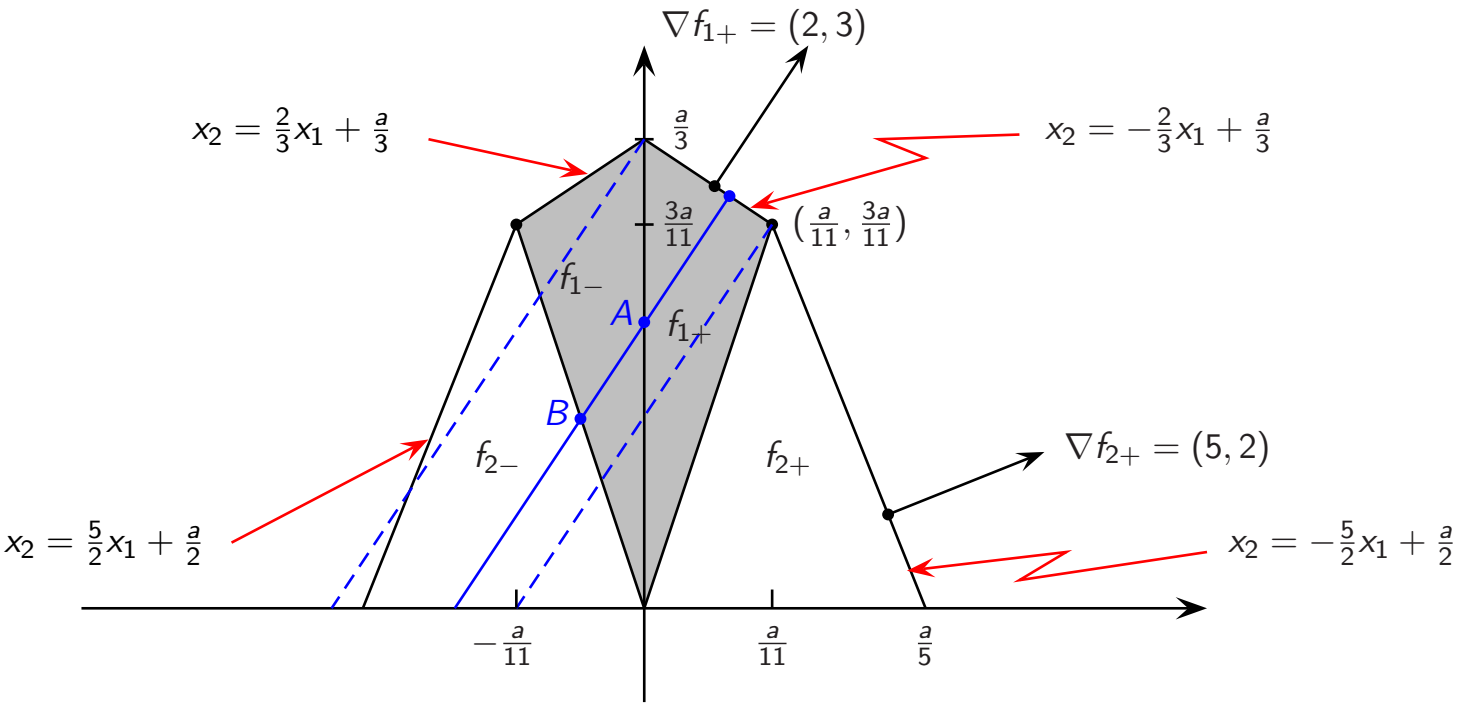
$$\min_{\|y - x^j\| \leq R} g(F(x^j) + F'(x^j)(y - x^j))$$

- 95% either re-discover Yuan, or blunder because they believe that Cauchy point works.
- Some work uses smooth approximations (à la Nesterov).

Hiriart-Urruty, Lemaréchal : Failure of steepest descent

$$f(x) = \max\{f_0(x), f_{\pm 1}(x), f_{\pm 2}(x)\}$$

$$f_0(x) = -100, f_{\pm 1}(x) = \pm 2x_1 + 3x_2, f_{\pm 2}(x) = \pm 5x_1 + 2x_2.$$



Consequences :

- Steepest descent with linesearch fails.
- Other gradient-oriented descent methods fail, too.
- Methods which vary the necessary optimality conditions also fail.
- Cauchy step in trust-regions fails, hence the usual strategy to prove convergence fails.

There are many temptations to go astray :

- Steepest descent with fixed or pre-defined steplengths works for convex f.
- Even without convexity : steepest descent convenient to define. Leads to nice Fenchel duality.
- Complexity theory harps on methods which work in the convex case, but fail when combined with globalization techniques such as linesearch or trust-regions.

Consequences for trust-regions :

- Dennis, Li, Tapia 1995. Axiomatic approach. Little scope due to regularity assumptions.
- Conn, Gould, Toint 2000. Believe that Cauchy point works.

Another consequence :

- Kurdyka-Łojasiewicz convergence theory not really effective as yet for non-smooth f.

Non-smooth bundle trust-region method

Want :

- Approximate f locally by a simple (polyhedral) model like in bundle method.
  $\Longrightarrow$ Need cutting planes.
- Tangent program should be CQP or even a LP.
- Use standard trust-region management to control stepsize.
- Prove convergence in the sense of subsequences.
  $\Longrightarrow$ Need some *ersatz* for the Cauchy point
- Get a rigorous stopping test.

Model approach :

Definition

(Noll, Prot, Rondepierre 2008).
A function $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a first-order model of $f : \mathbb{R}^n \to \mathbb{R}$ if the following axioms are satisfied :

(M_1) $\phi(\cdot, x)$ is convex, $\phi(x, x) = f(x)$, $\partial_1 \phi(x, x) \subset \partial f(x)$

(M_2) $f(y) \leq \phi(y, x) + o(\|y - x\|)$ as $y \to x$

(M_3) $\limsup_{y' \to y, x' \to x} \phi(y', x') \leq \phi(y, x)$.

Definition

The model ϕ is called strict if

$(\widehat{M}_2)$ $f(y) \leq \phi(y, x) + o(\|y - x\|)$ as $y - x \to 0$
uniformly on bounded sets

The model ϕ is called strong if

$(\widetilde{M}_2)$ $f(y) \leq \phi(y, x) + O(\|y - x\|^2)$ as $y - x \to 0$
uniformly on bounded sets

1st-order Taylor expansion

$$\phi(y,x) = f(x) + \nabla f(x)^T (y-x)$$

Standard model

$$\phi^\sharp(y,x) = f(x) + f^\circ(x, y-x)$$

f *convex.* Is its own strong model :

$$\phi(y,x) = f(y)$$

$f = g \circ F$ *composite convex.* Natural strict model :

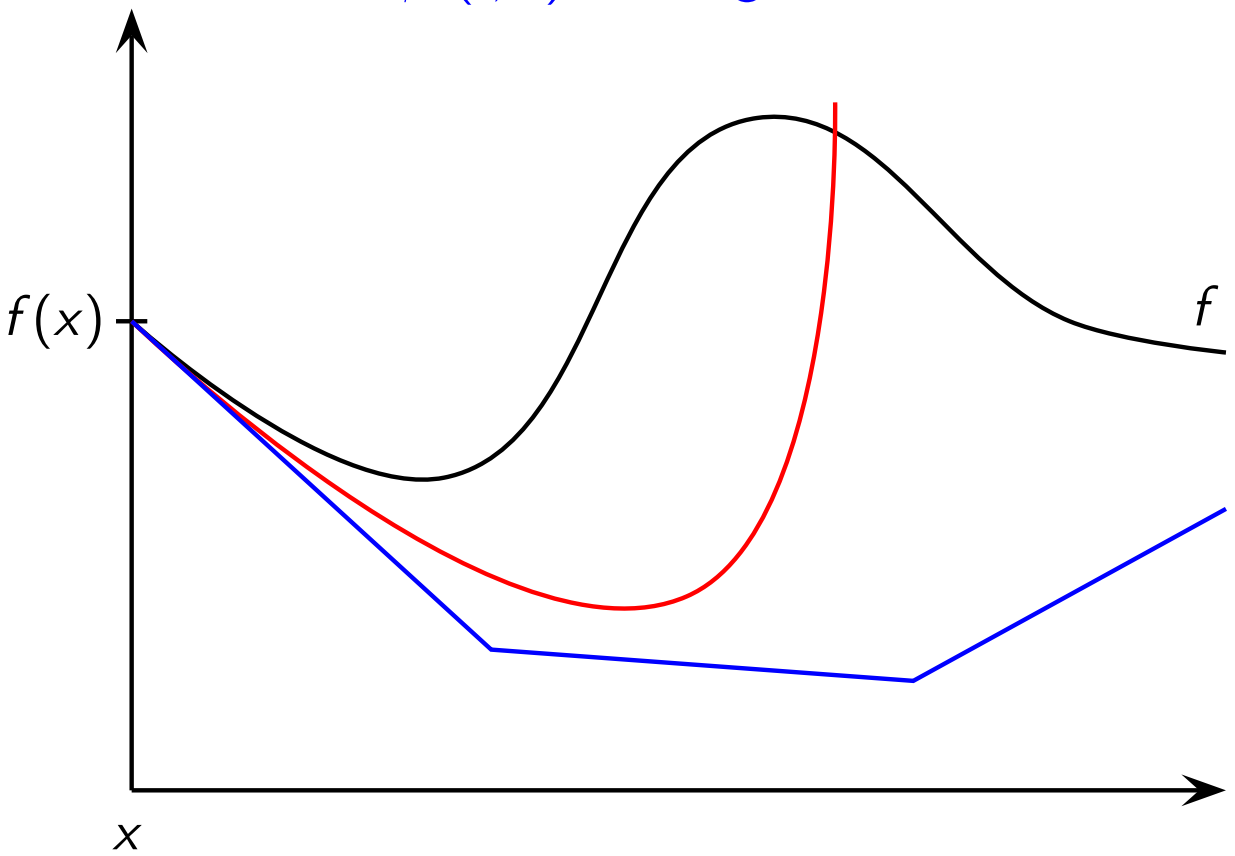
$$\phi^n(y,x) = g\left(F(x) + F'(x)(y-x)\right)$$

f *prox-regular* :

$$\phi(y,x) = f(y) + \mu\|y-x\|^2$$

$\phi(\cdot, x)$ ideal model

$\phi_k(\cdot, x)$ working model



Observe :

- Every model ϕ of f gives rise to one bundle method.

  $\Longrightarrow$ The more models, the more methods

- Standard model $\phi^\sharp$ not always strict.
- $f = g \circ F$, g convex, F class C^1 :

$$\phi^n(y, x) = g(F(x) + F'(x)(y - x))$$

- Every lower C^1-function has a strict model.

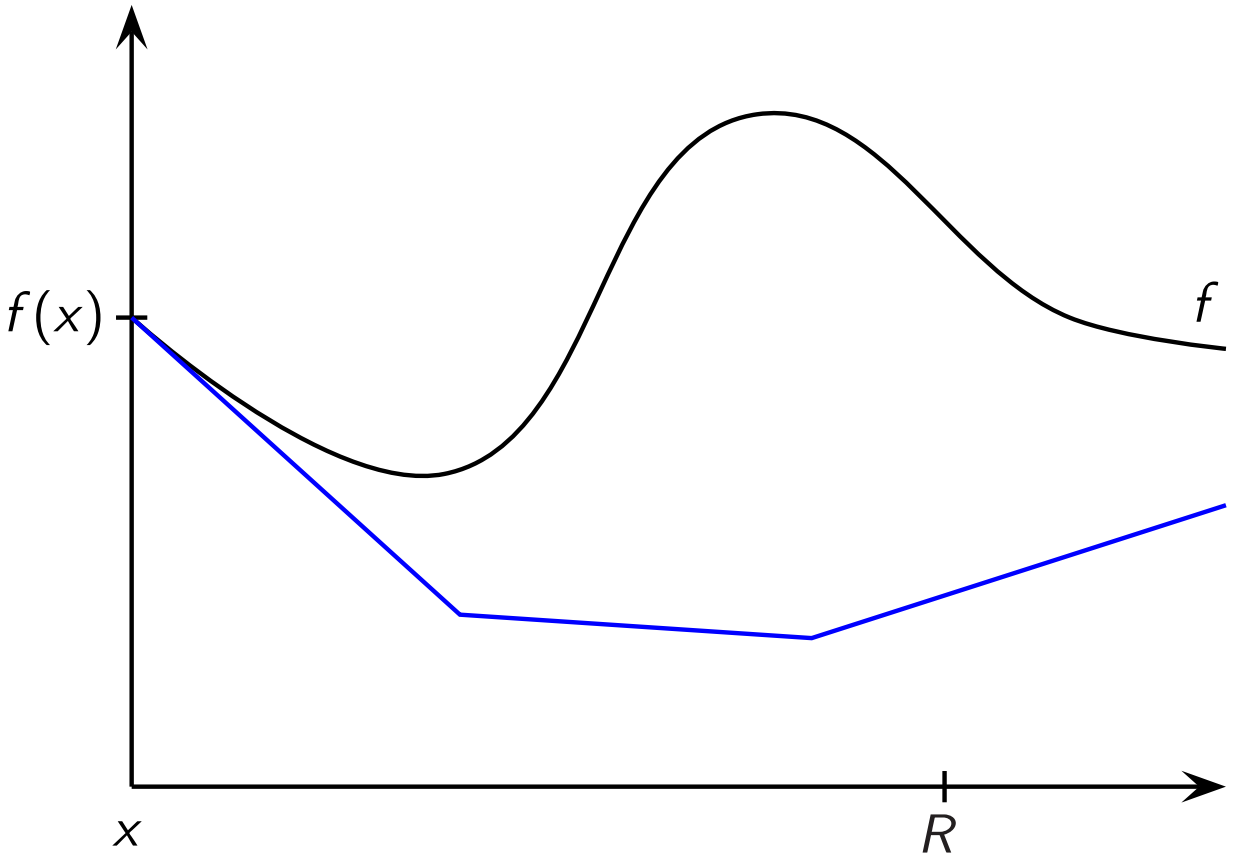
  $\Longrightarrow$ Upper envelope of downshifted tangents

- Every upper C^1-function has a strict model.

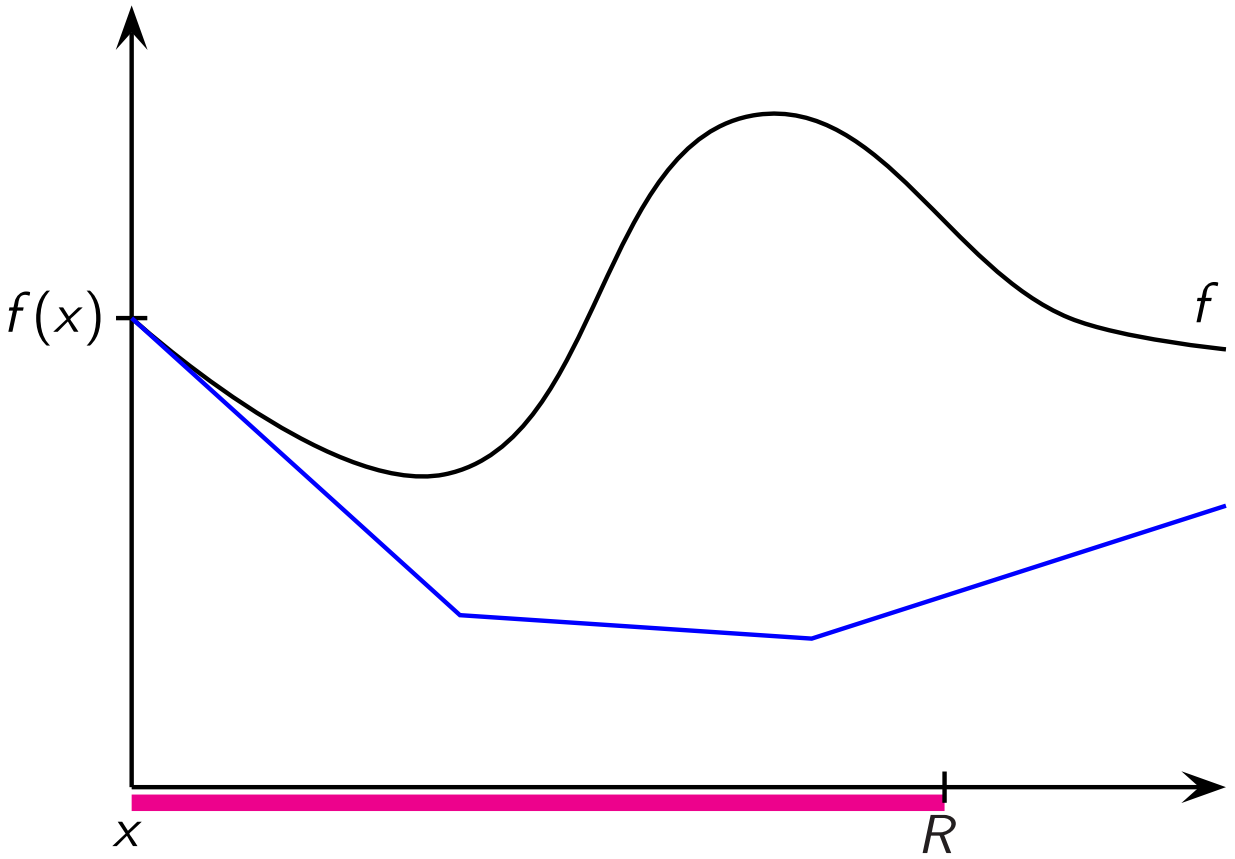
  $\Longrightarrow$ To wit, its standard model $\phi^\sharp$ is strict.

Bundle trust-region method

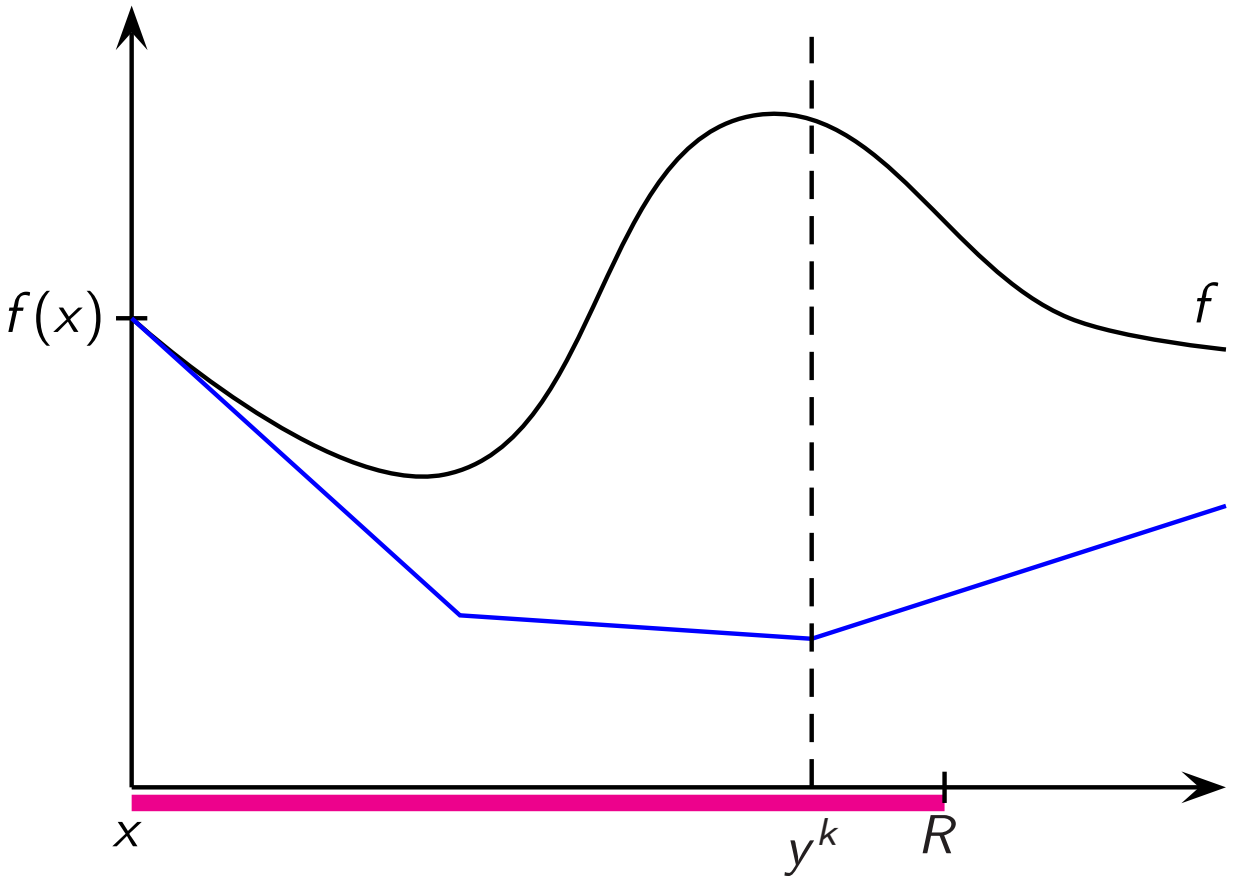
$\phi_k(\cdot, x)$ working model

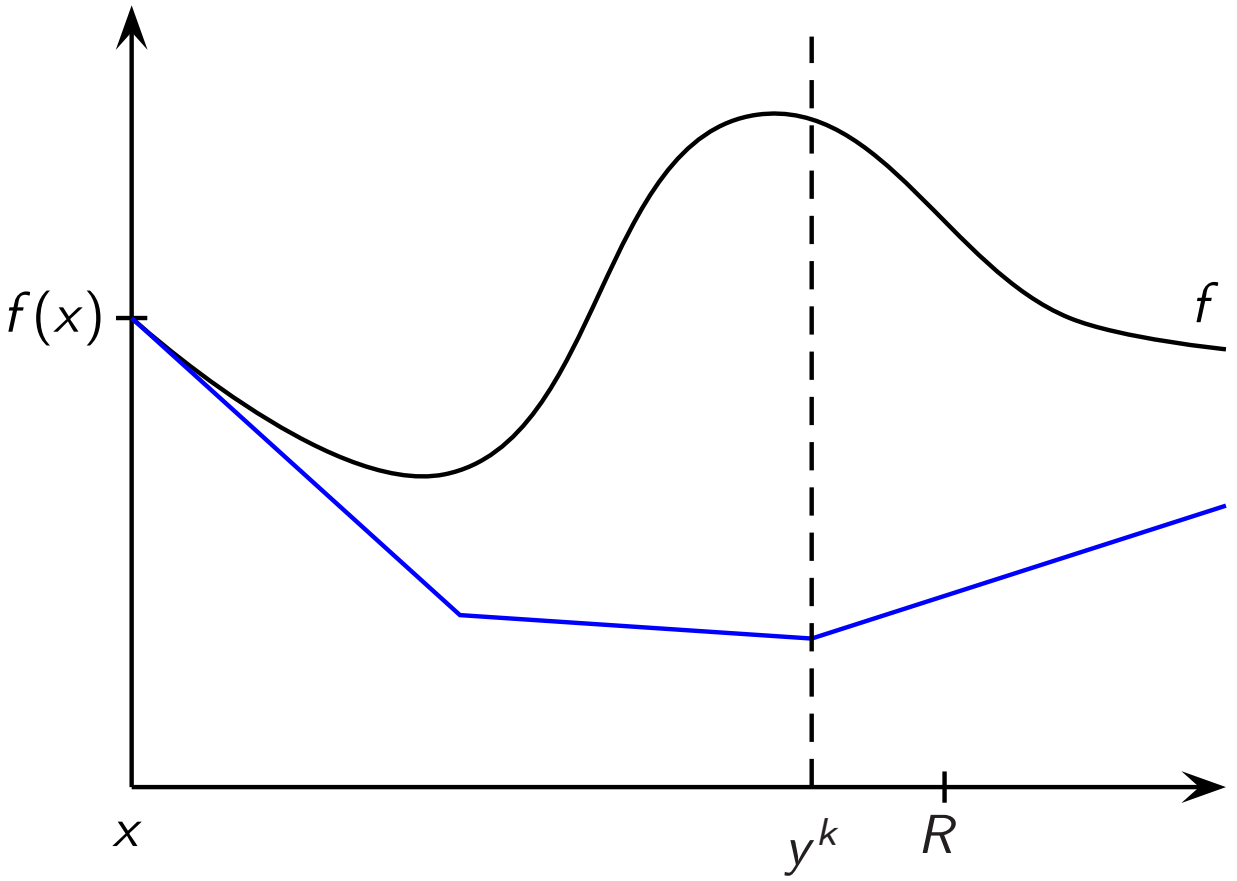


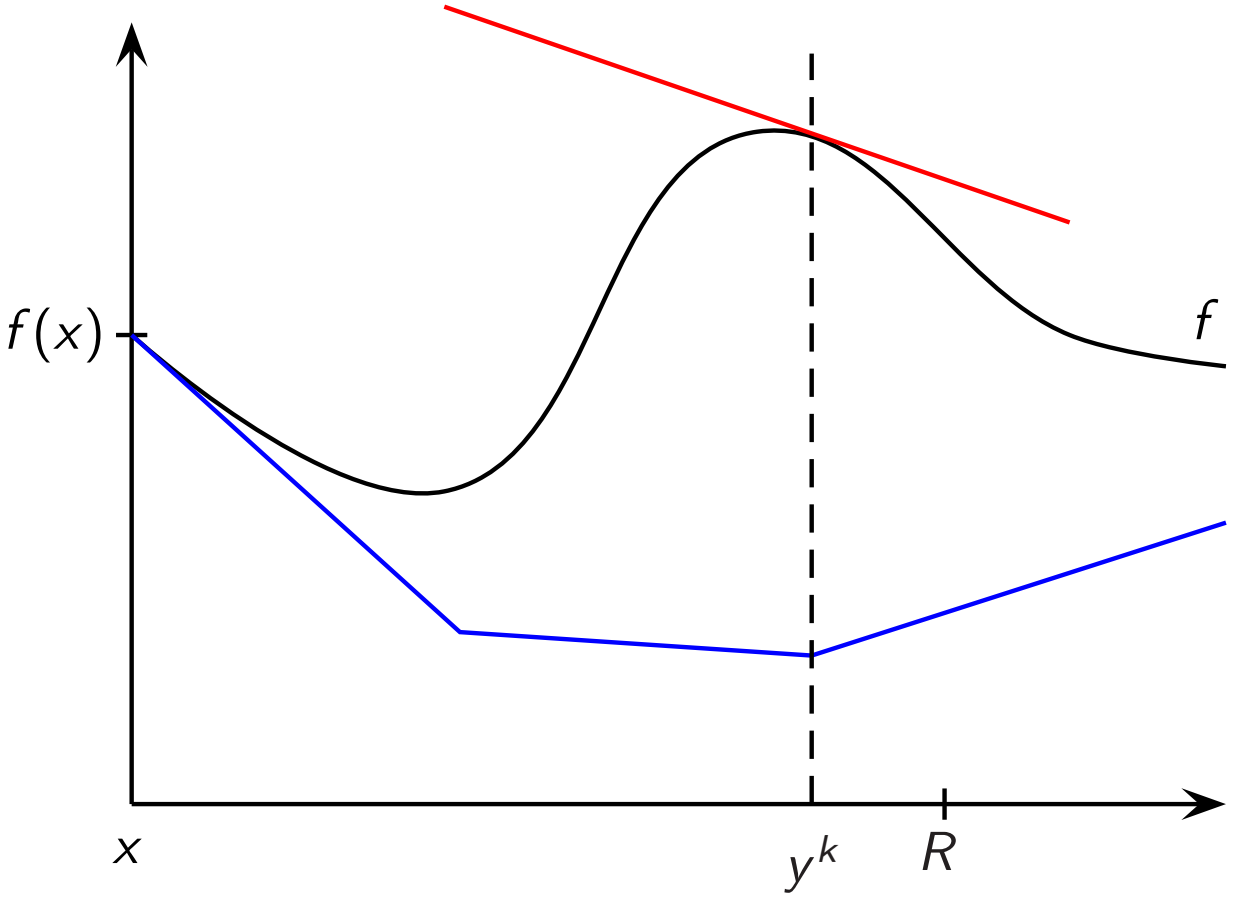
$$\min_{\|y-x\| \le R} \phi_k(y, x)$$

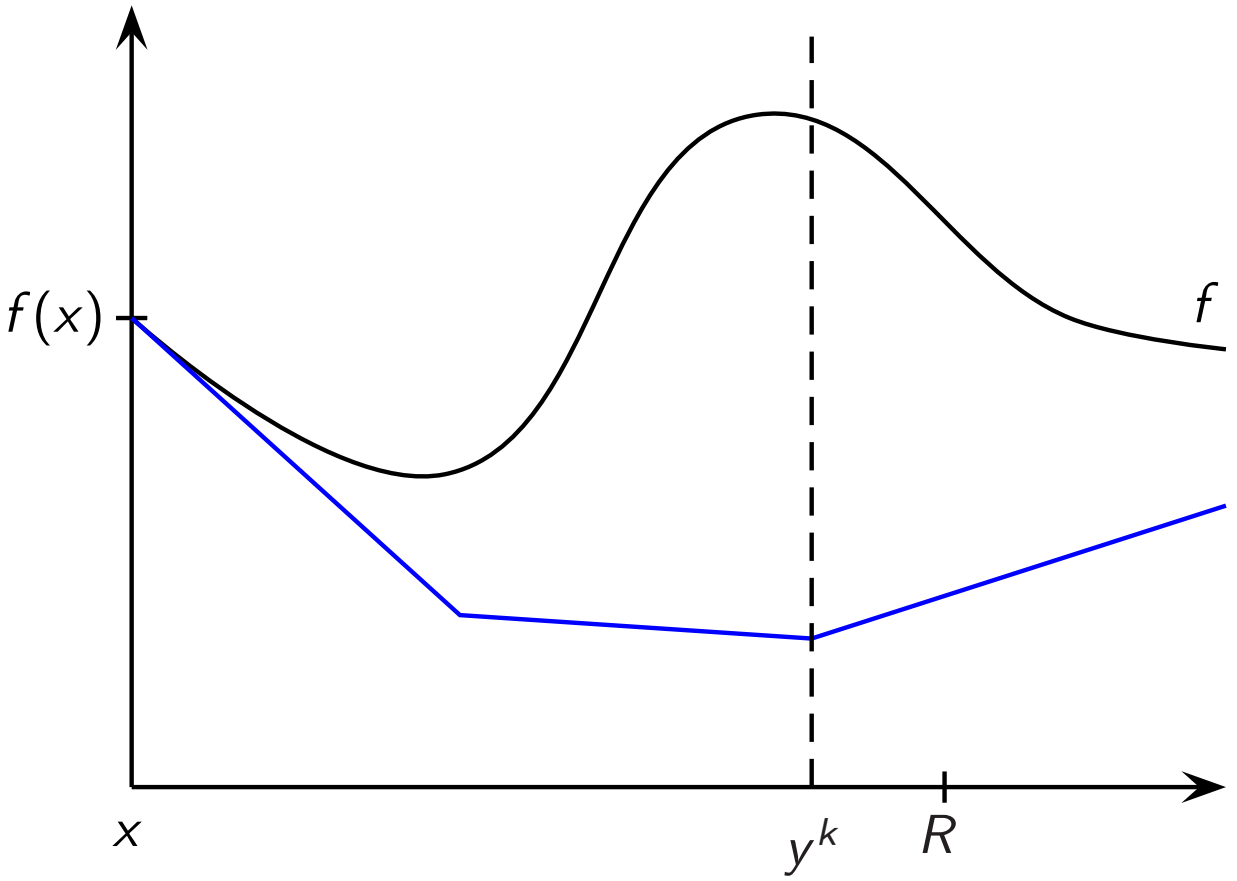


$$\min_{\|y-x\| \le R} \phi_k(y, x)$$

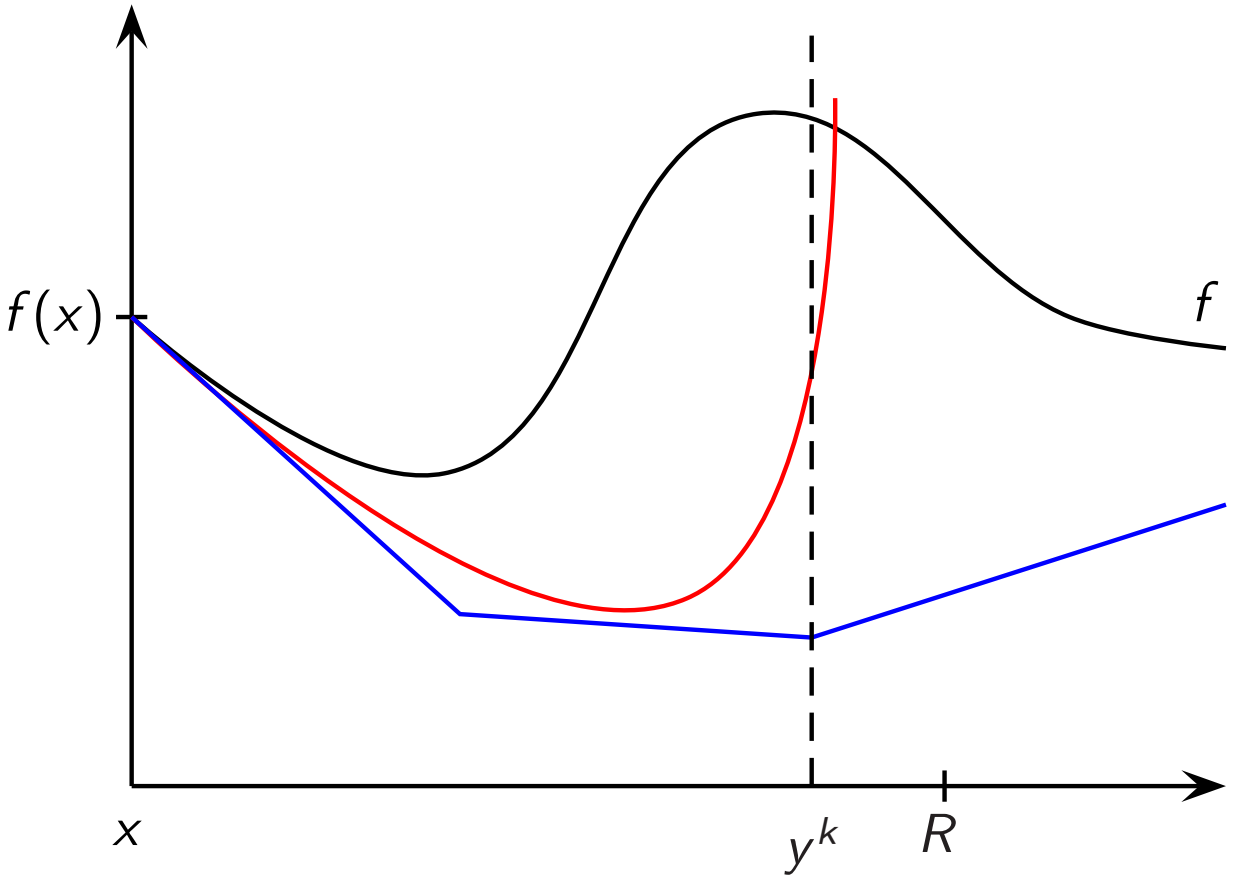



$f(x)$
f
x
y^k
R

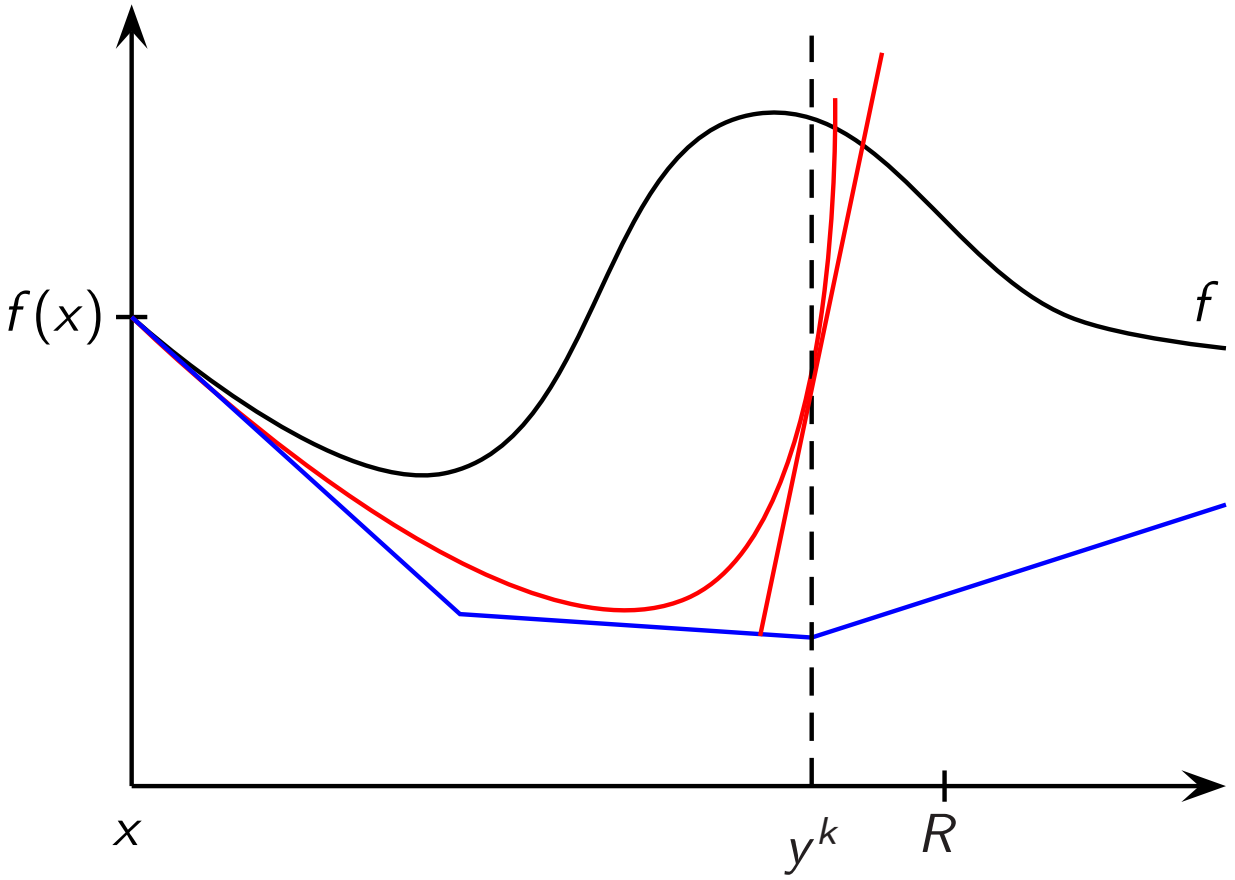

$f(x)$
f
x
y^k
R


$f(x)$
f
x
y^k
R

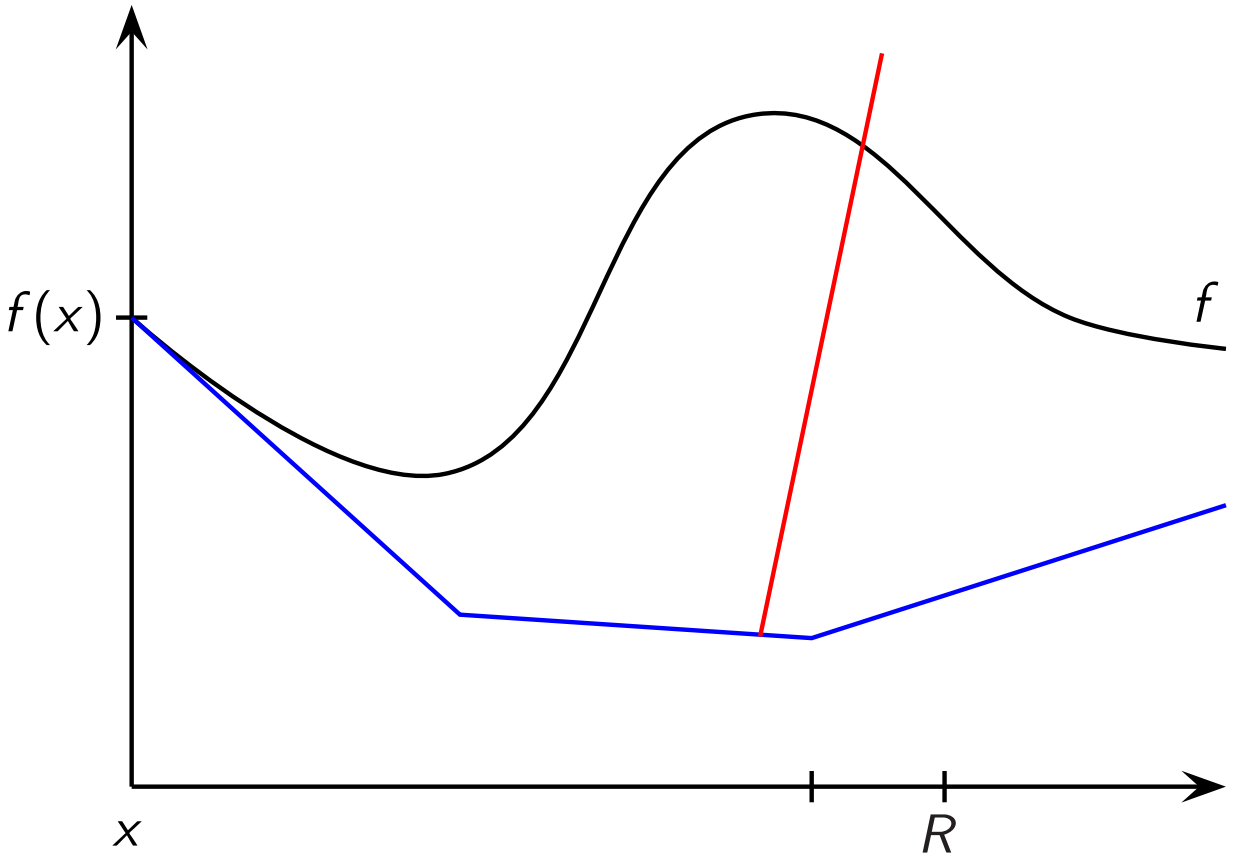
$\phi(\cdot, x)$ ideal model



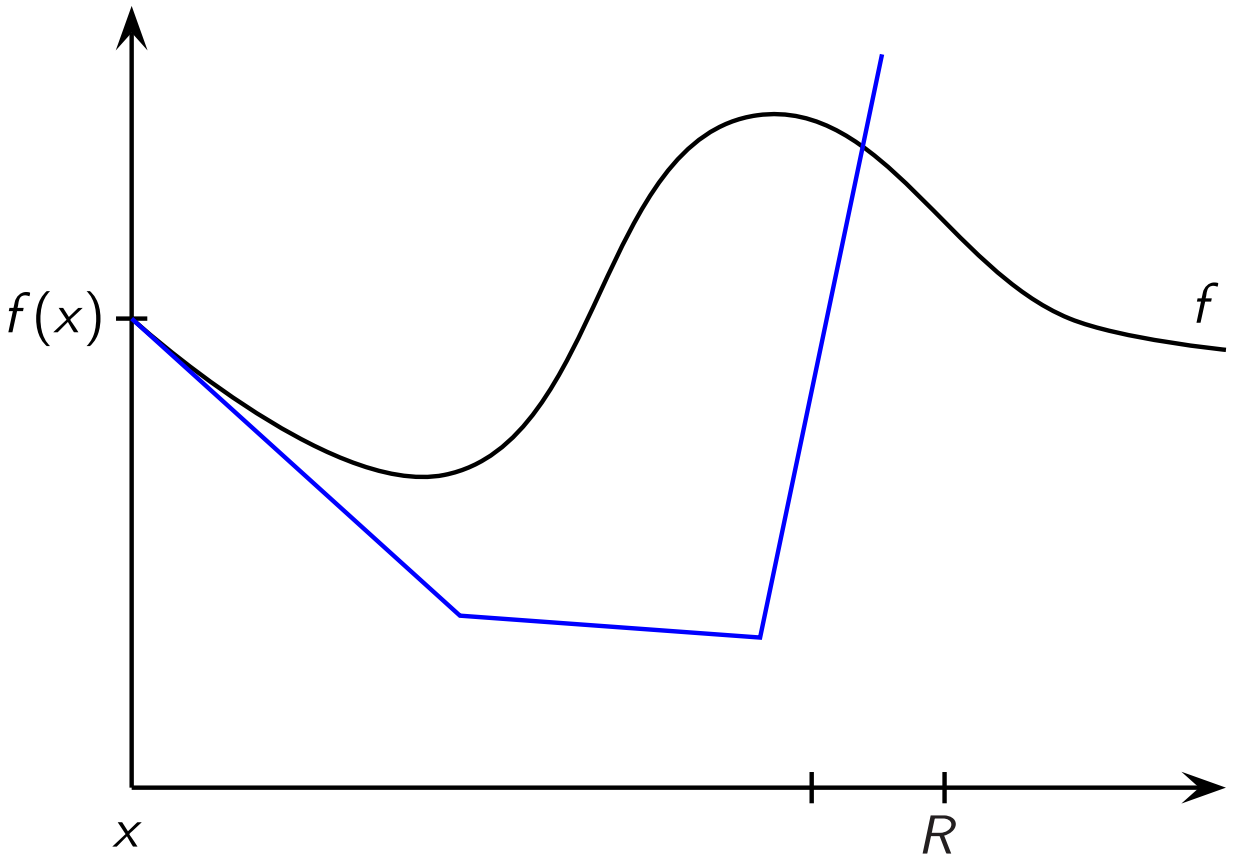
$\phi(\cdot, x)$ ideal model



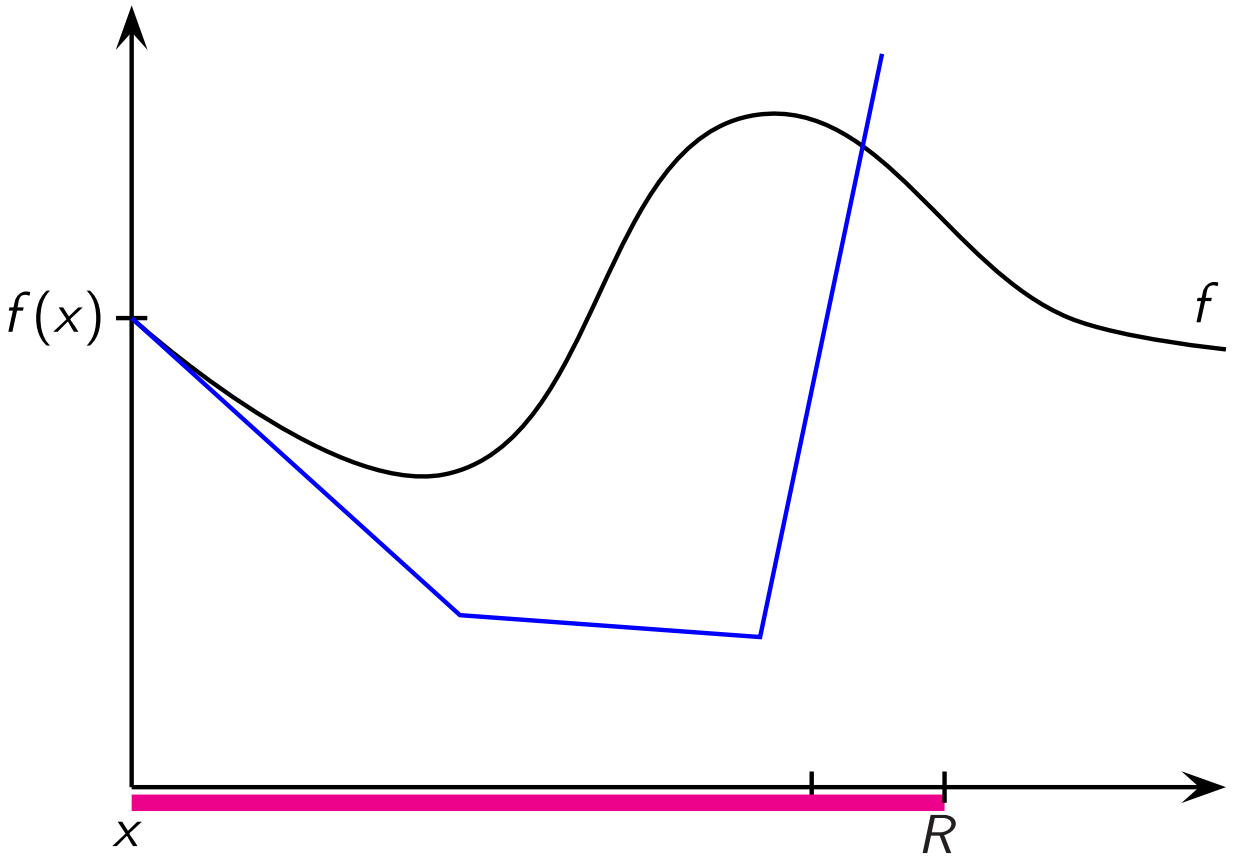
$\phi_k(\cdot, x)$ working model



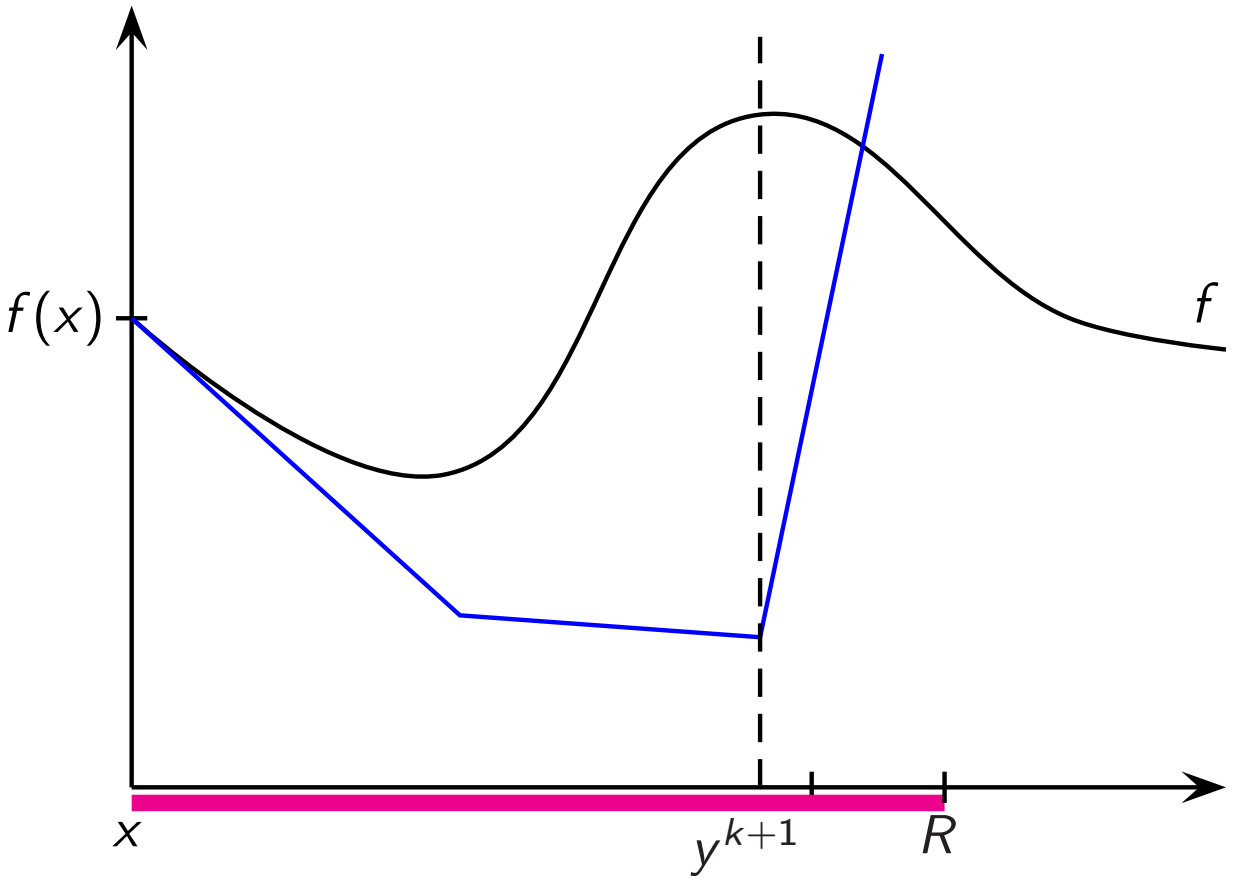
$\phi_{k+1}(\cdot, x)$ working model

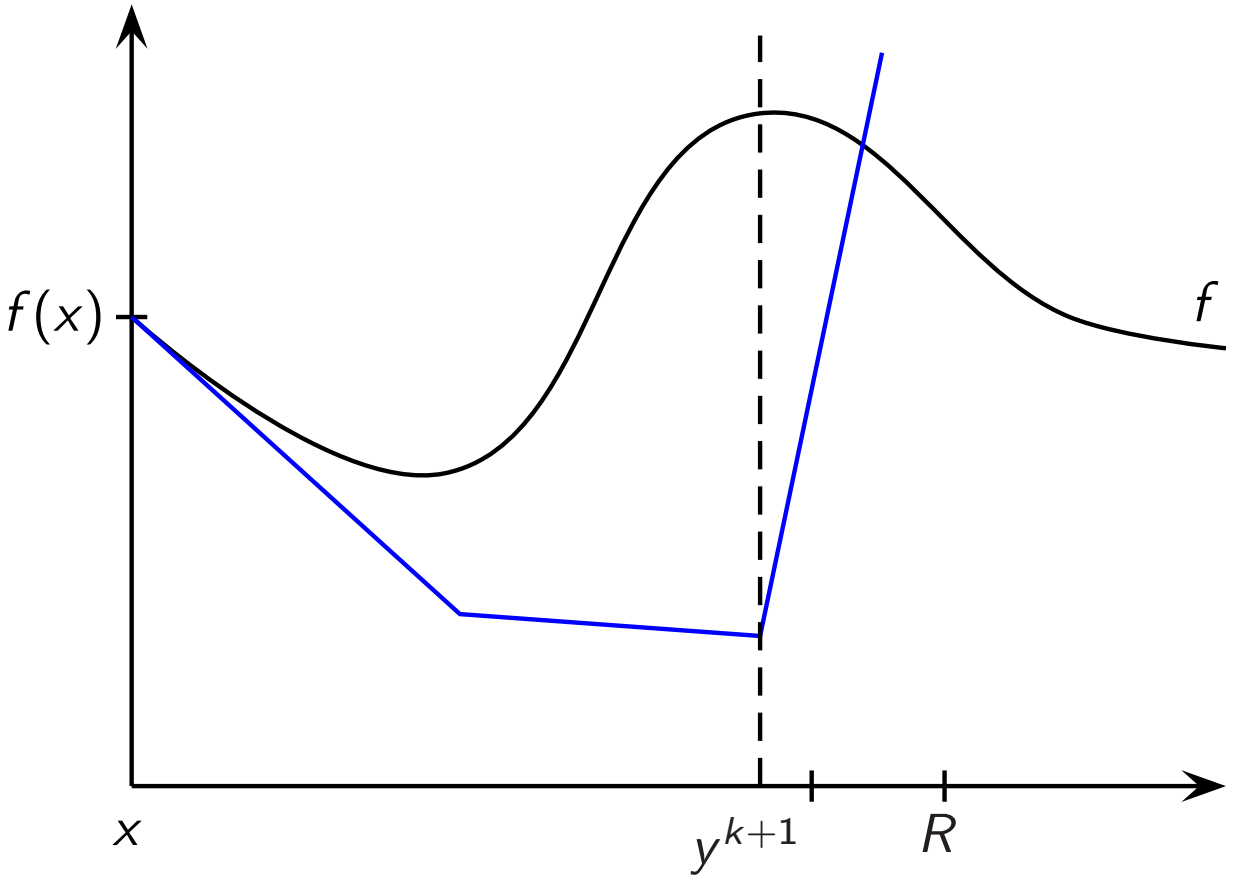


$$\min_{\|y-x\| \le R} \phi_{k+1}(y, x)$$

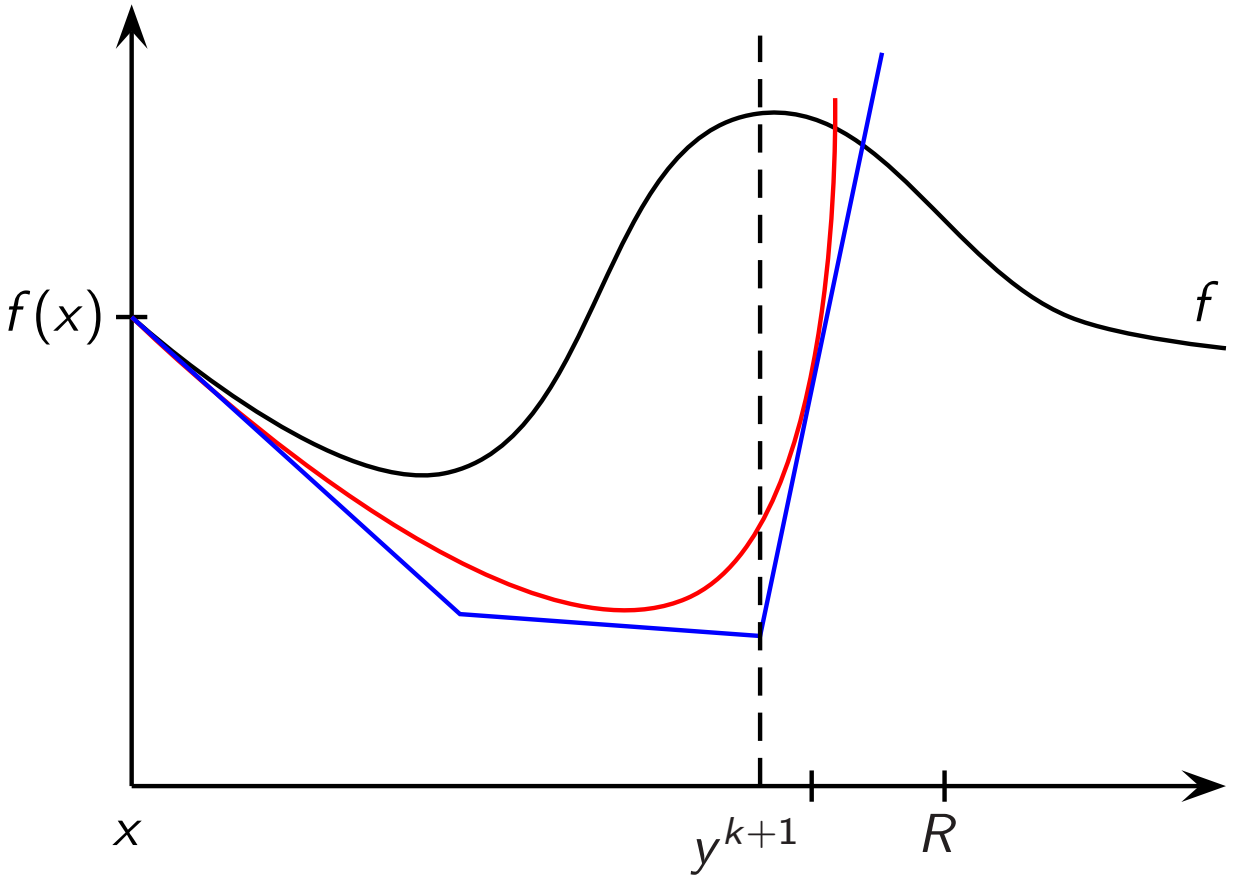


$$\min_{\|y-x\| \le R} \phi_{k+1}(y, x)$$

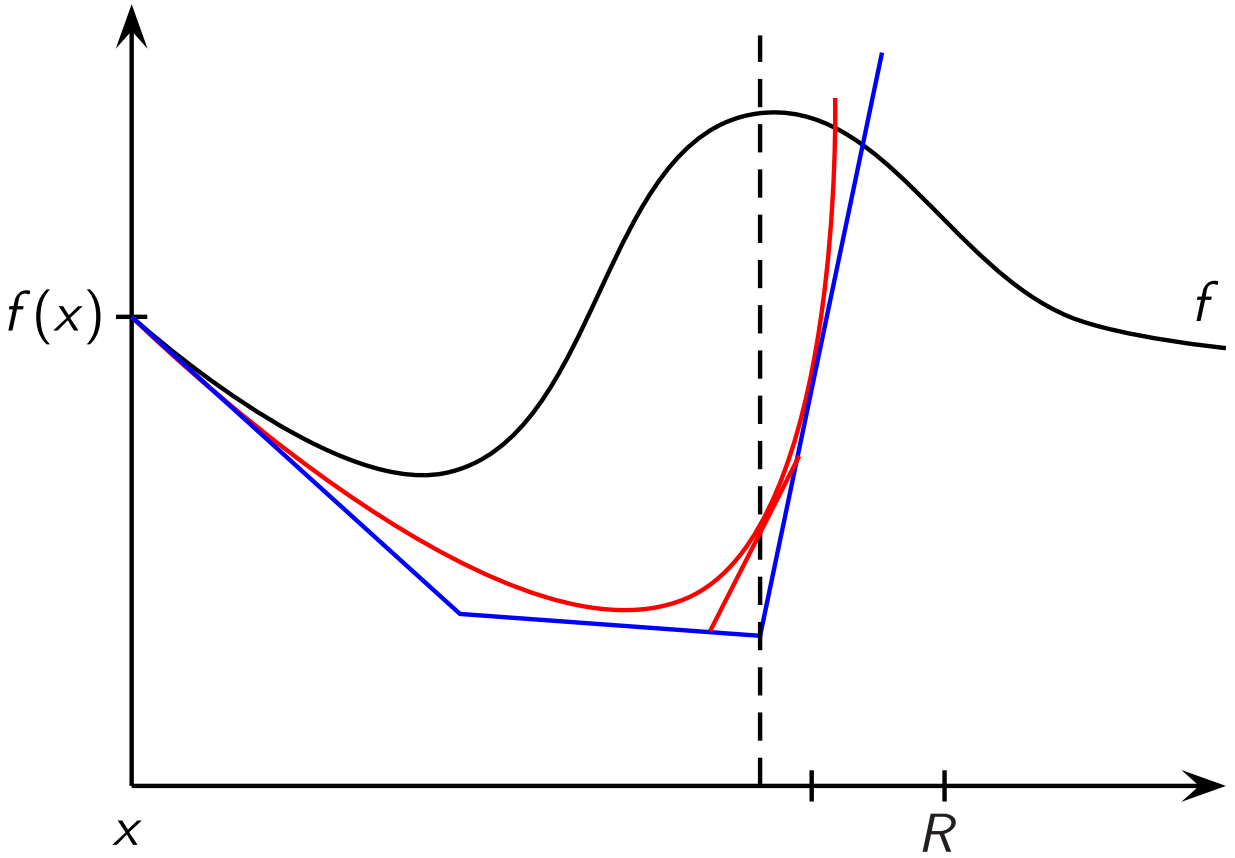



$f(x)$
f
x
y^{k+1}
R

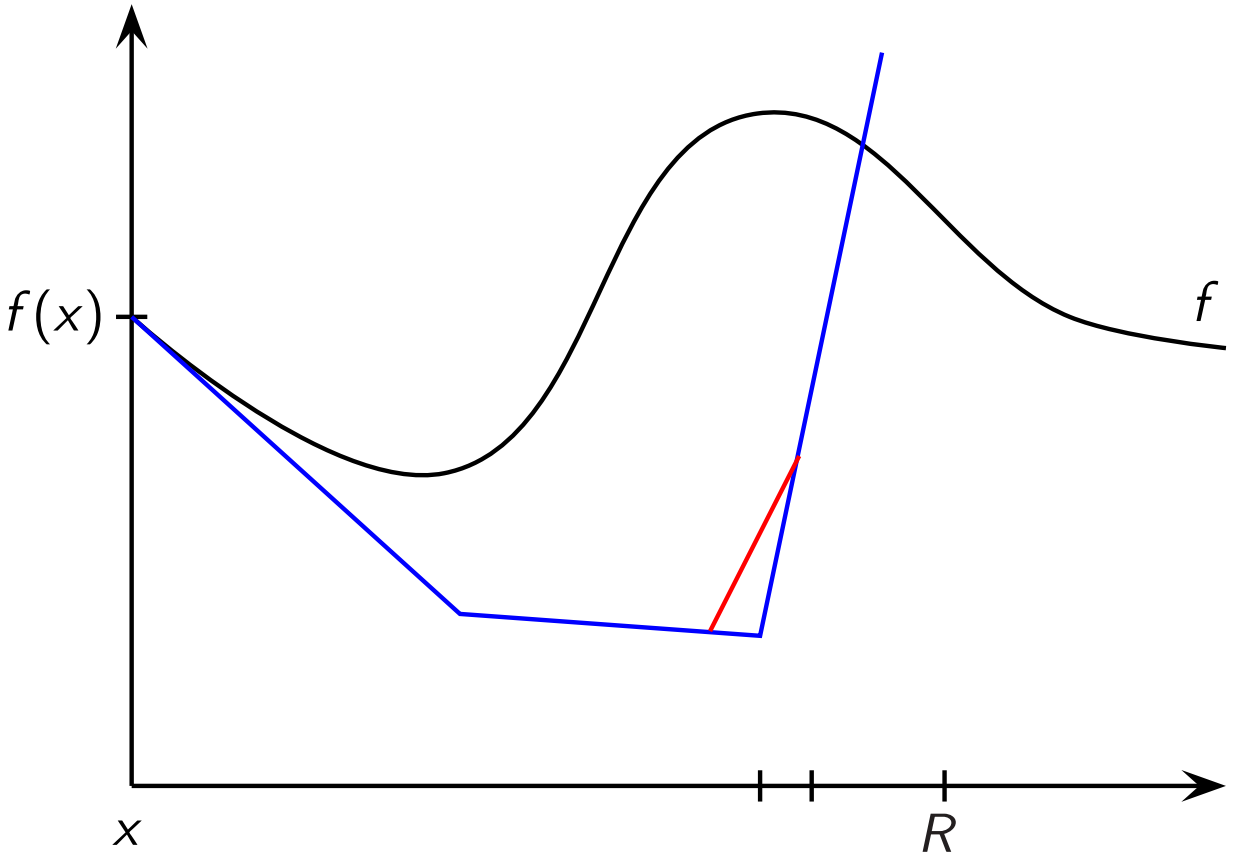
$\phi(\cdot, x)$ ideal model



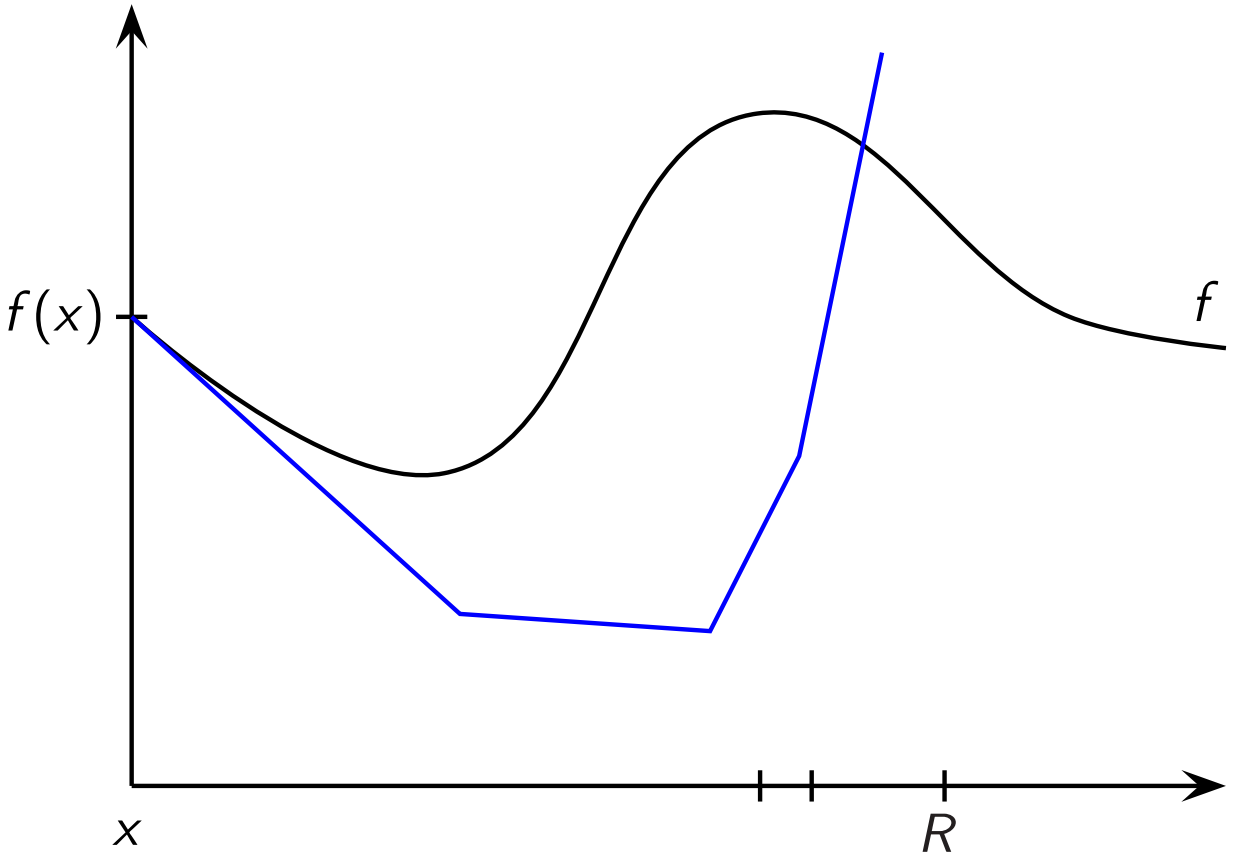
$\phi(\cdot, x)$ ideal model



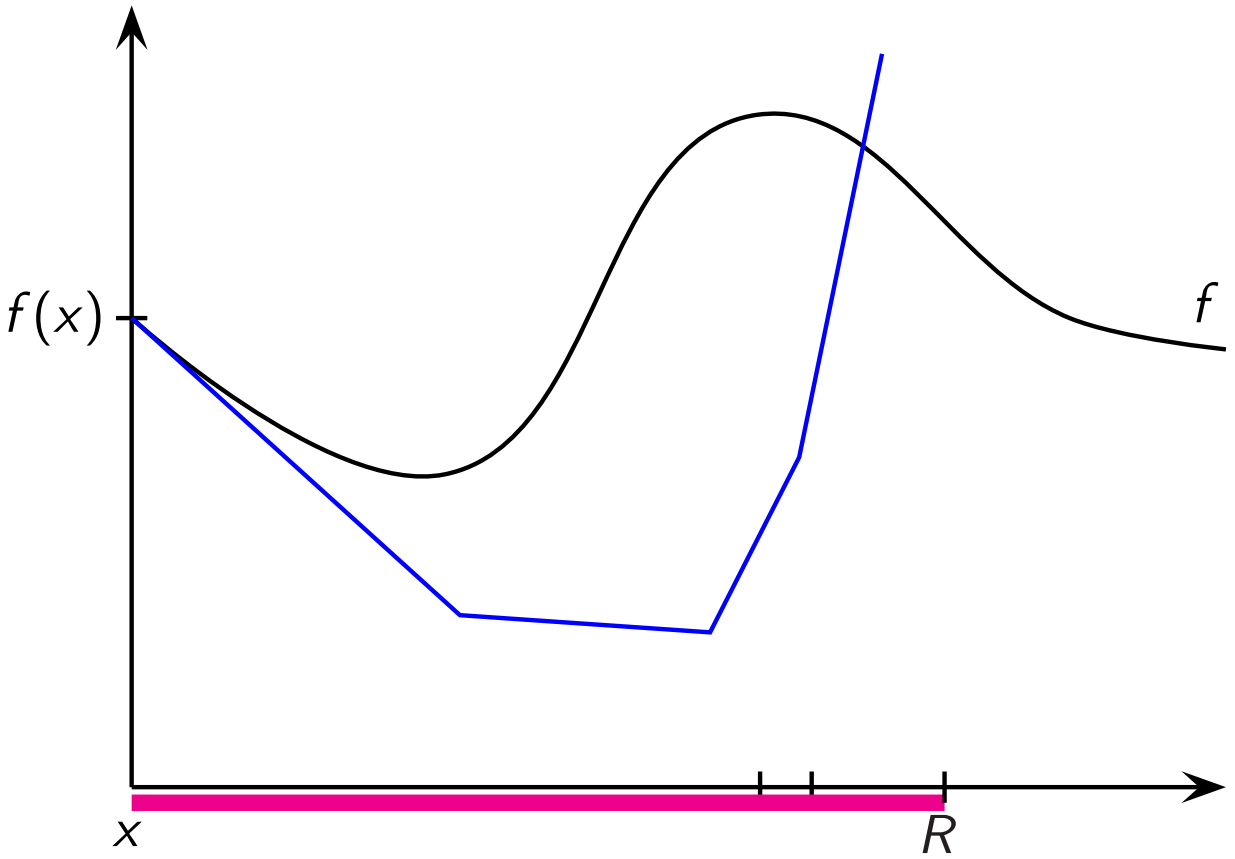
$\phi_{k+1}(\cdot, x)$ working model



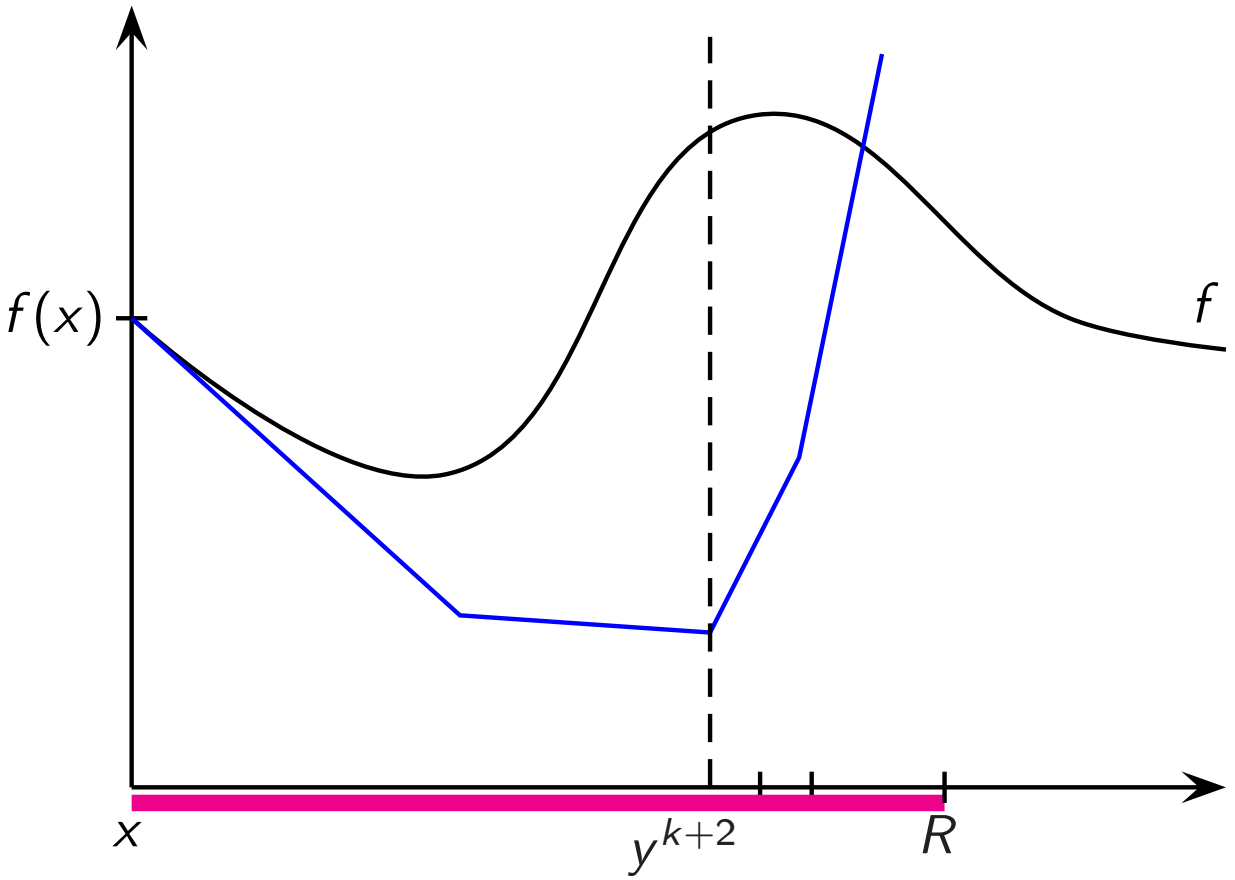
$\phi_{k+2}(\cdot, x)$ working model

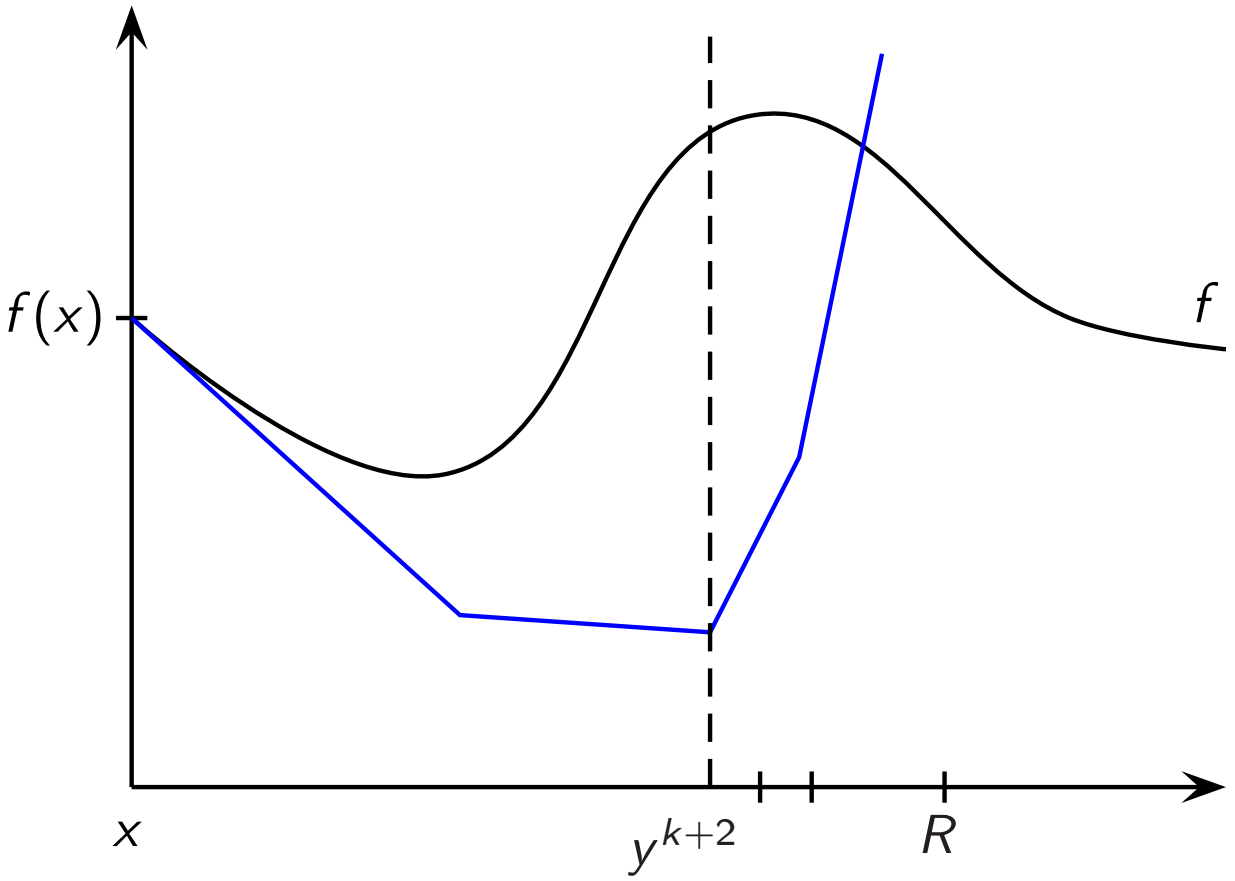


$$\min_{\|y-x\| \le R} \phi_{k+2}(y, x)$$

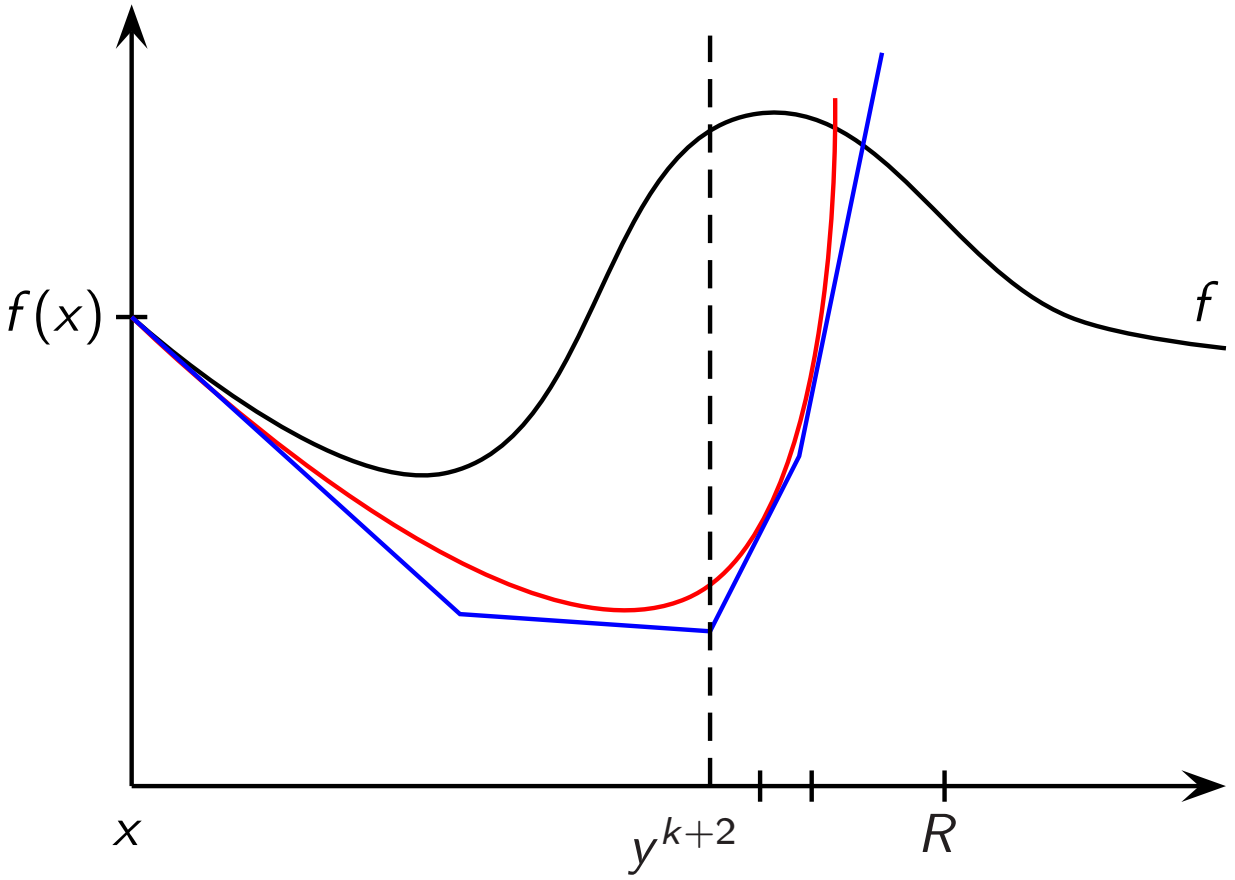


$$\min_{\|y-x\| \le R} \phi_{k+2}(y, x)$$

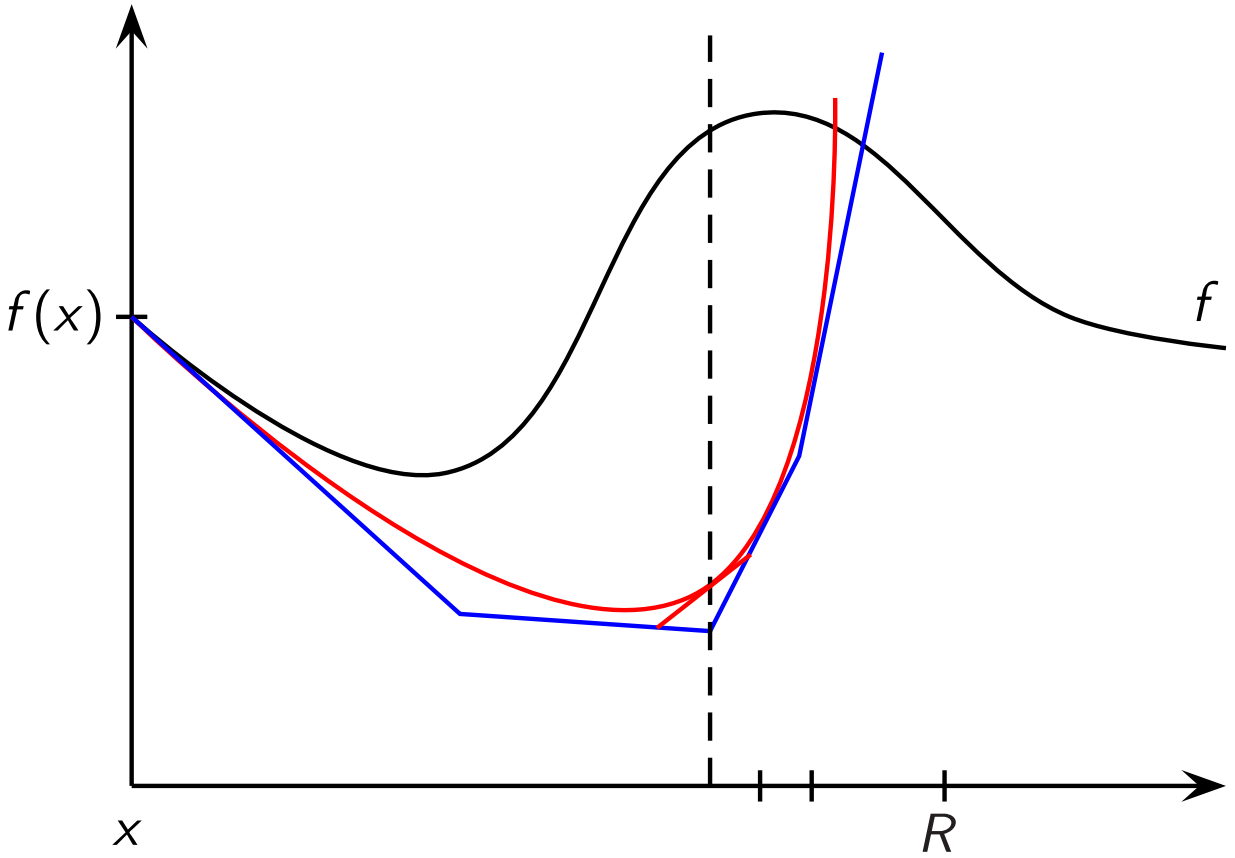



f(x)
f
x
y^{k+2}
R

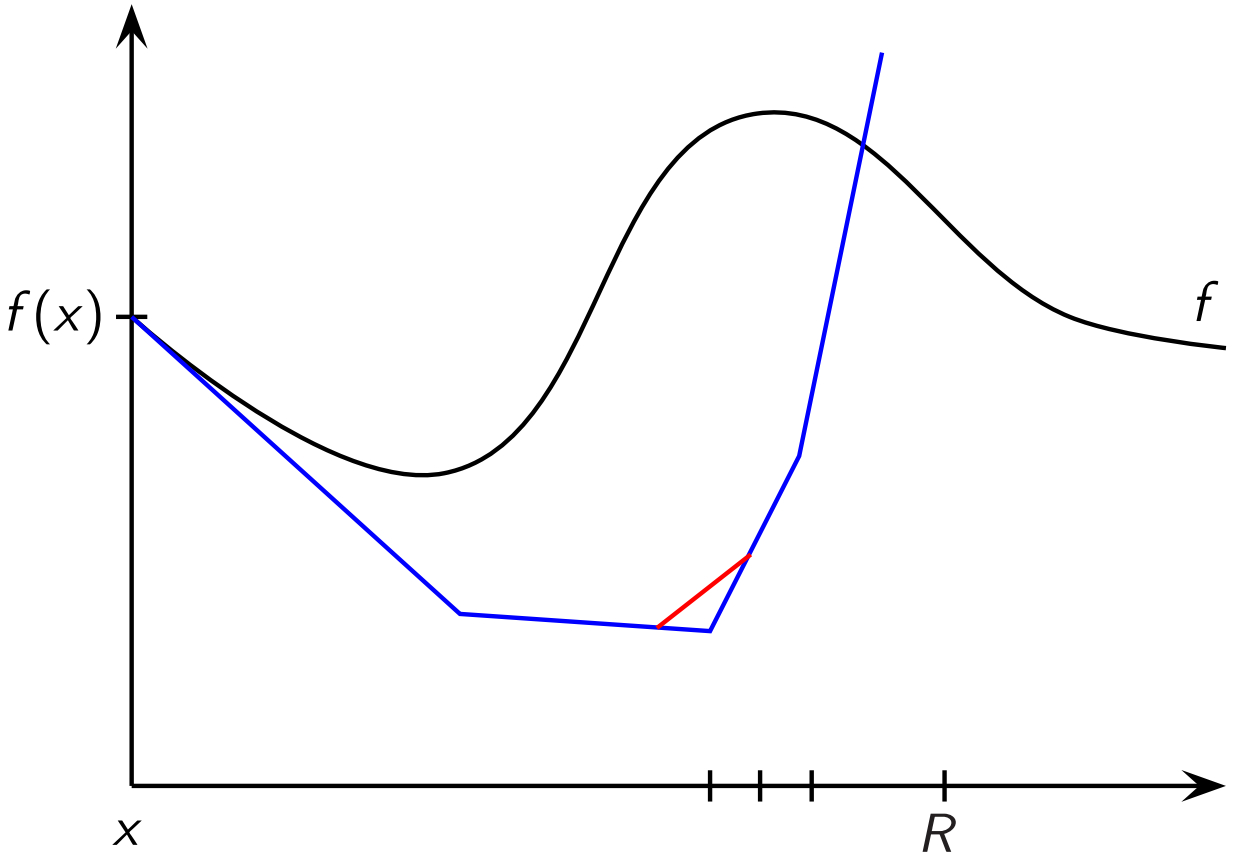
$\phi(\cdot, x)$ ideal model



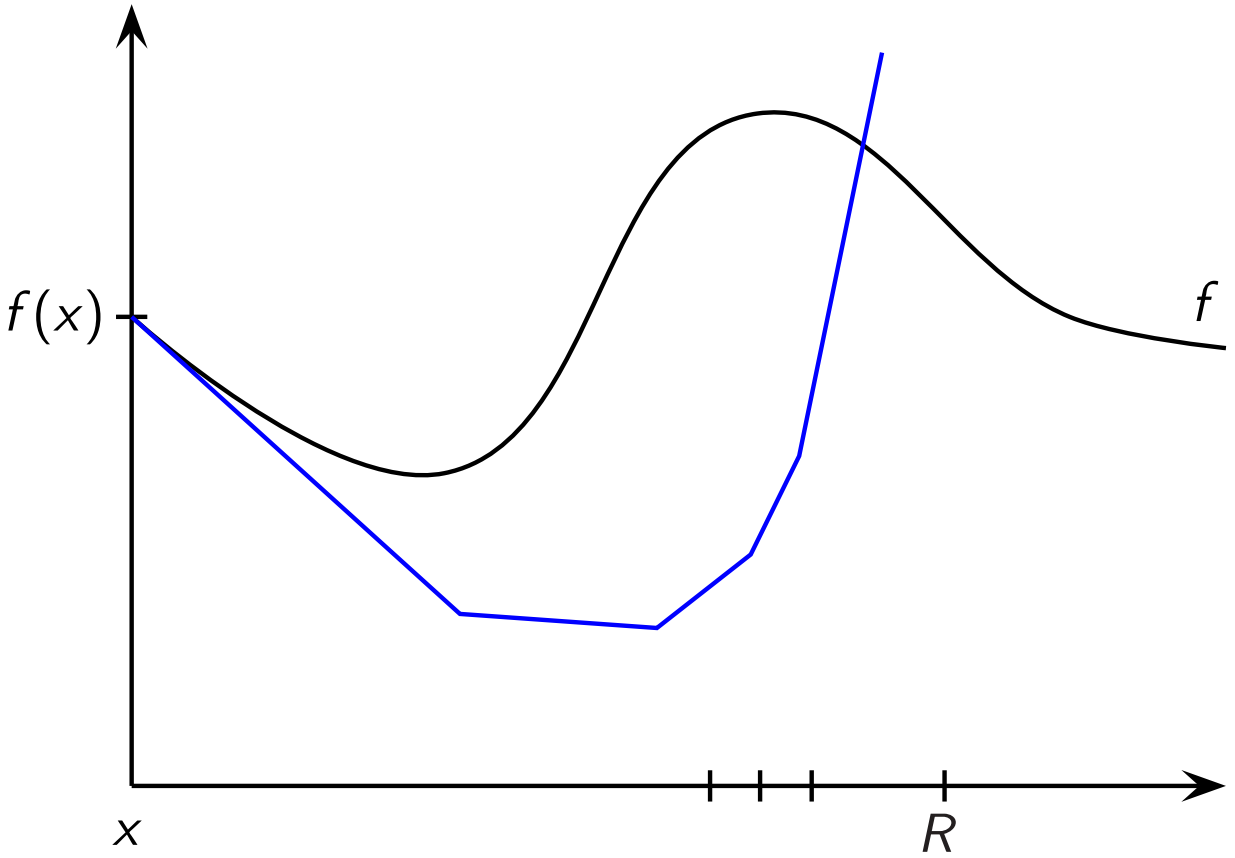
$\phi(\cdot, x)$ ideal model



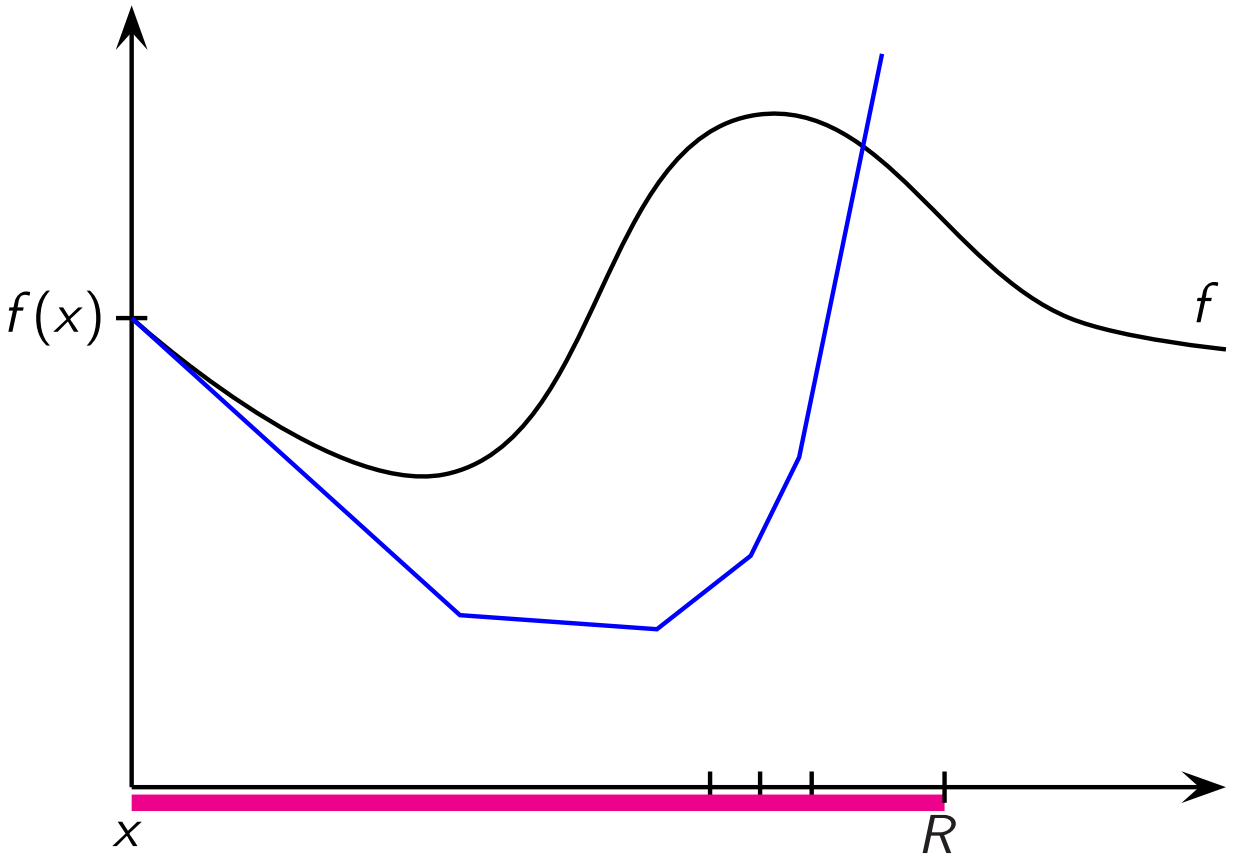
$\phi_{k+2}(\cdot, x)$ working model



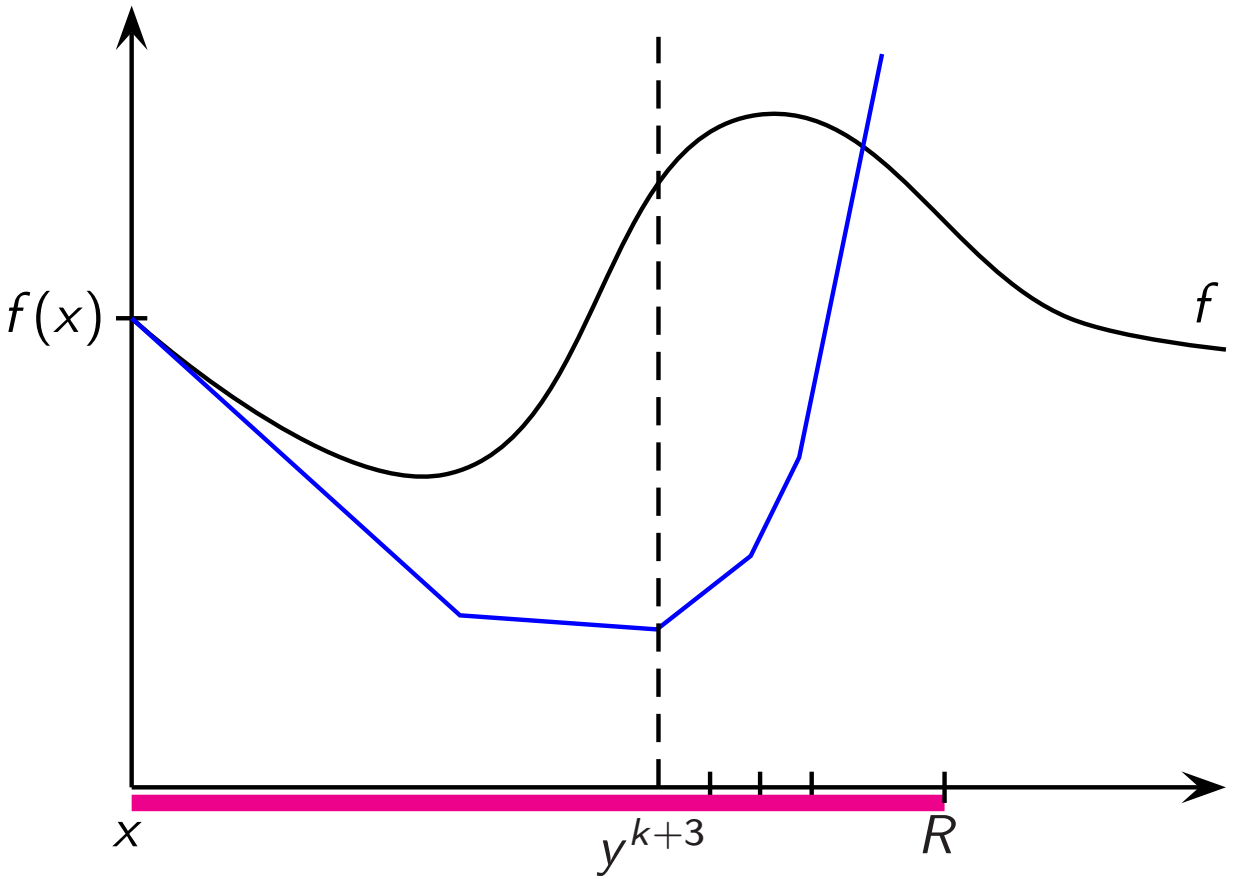
$\phi_{k+3}(\cdot, x)$ working model

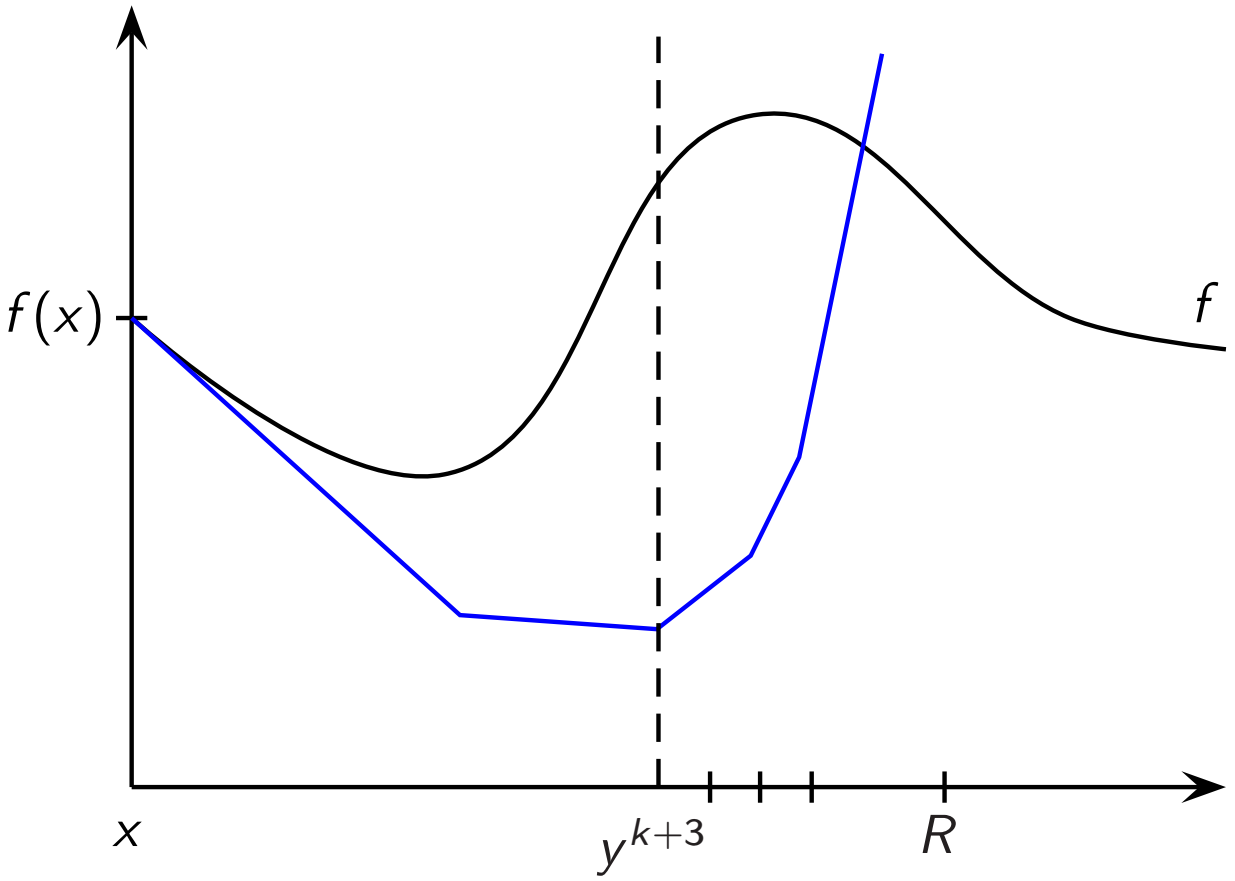


$$\min_{\|y-x\| \le R} \phi_{k+3}(y, x)$$

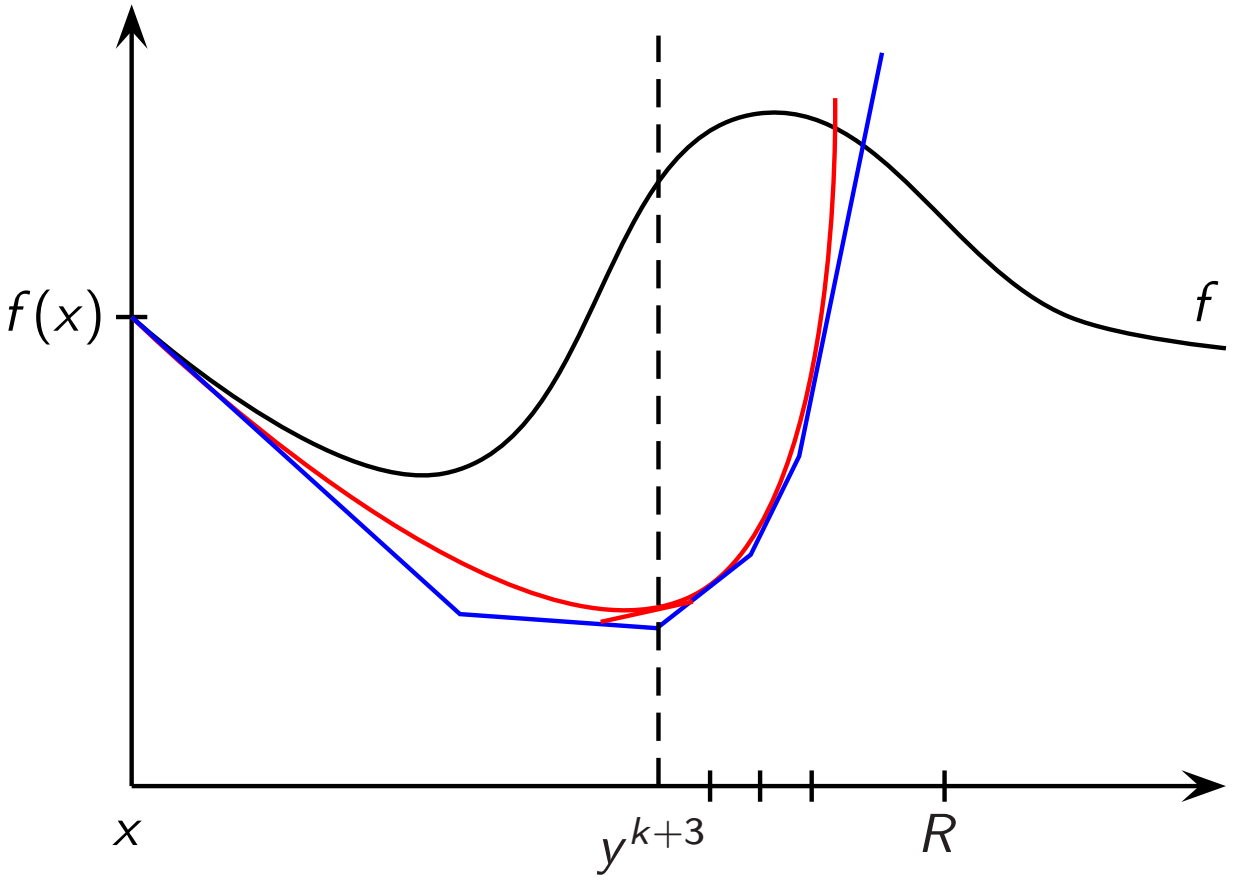


$$\min_{\|y-x\| \le R} \phi_{k+3}(y, x)$$

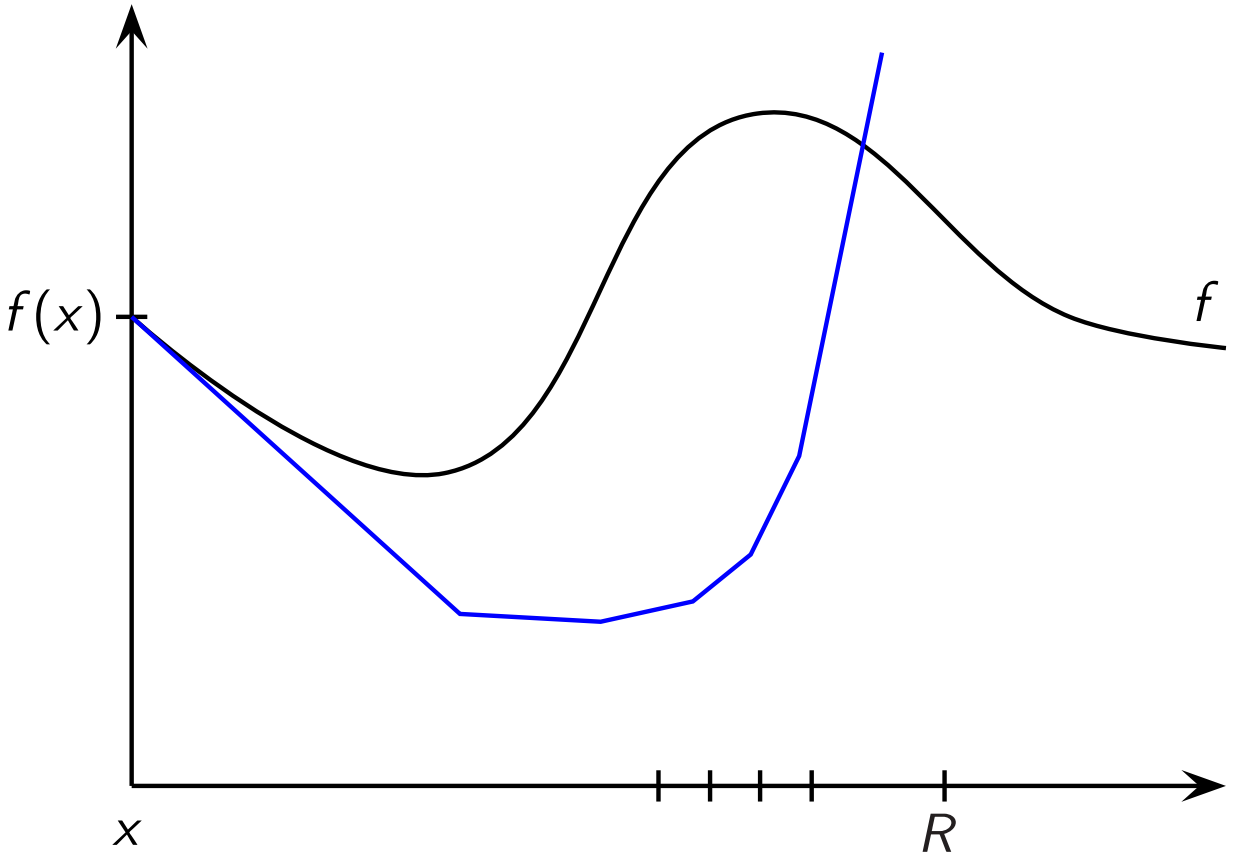



$f(x)$
f
x
y^{k+3}
R

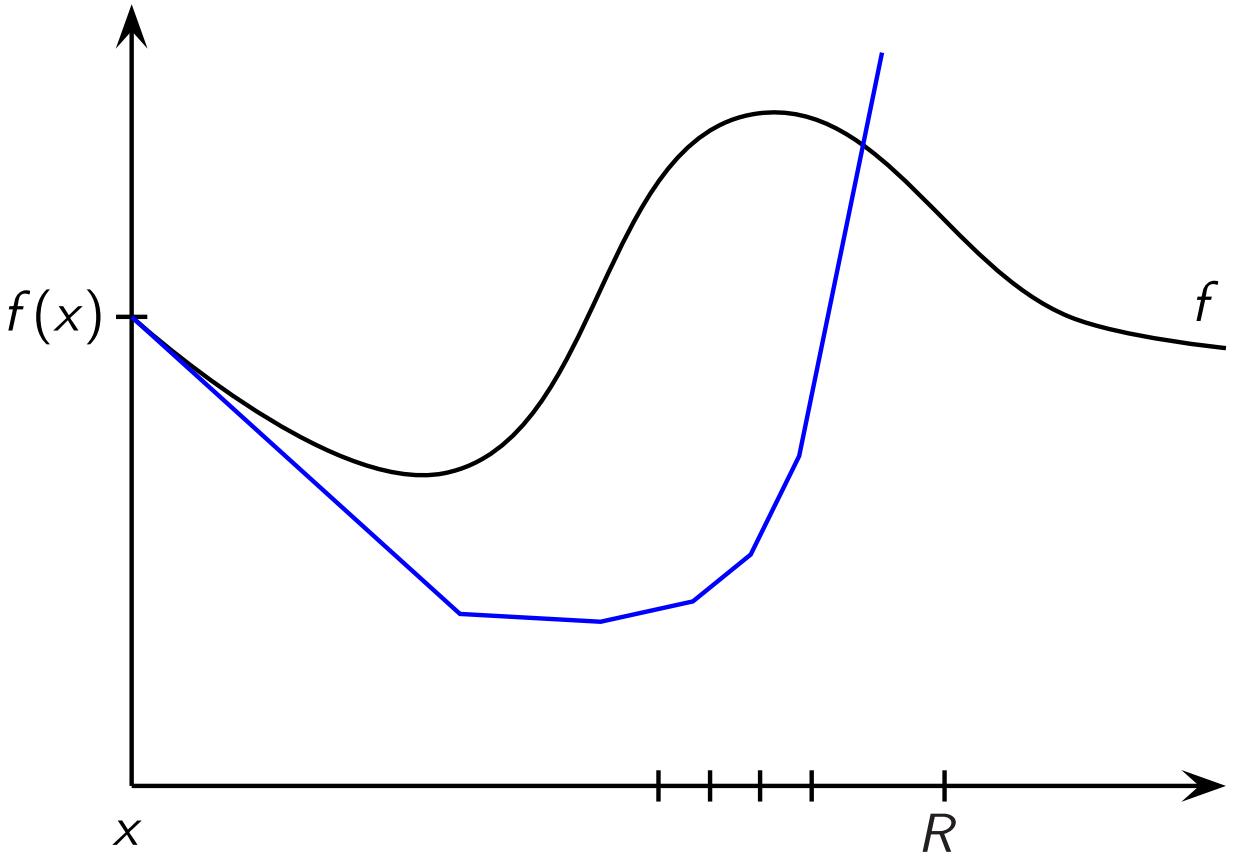
$\phi(\cdot, x)$ ideal model



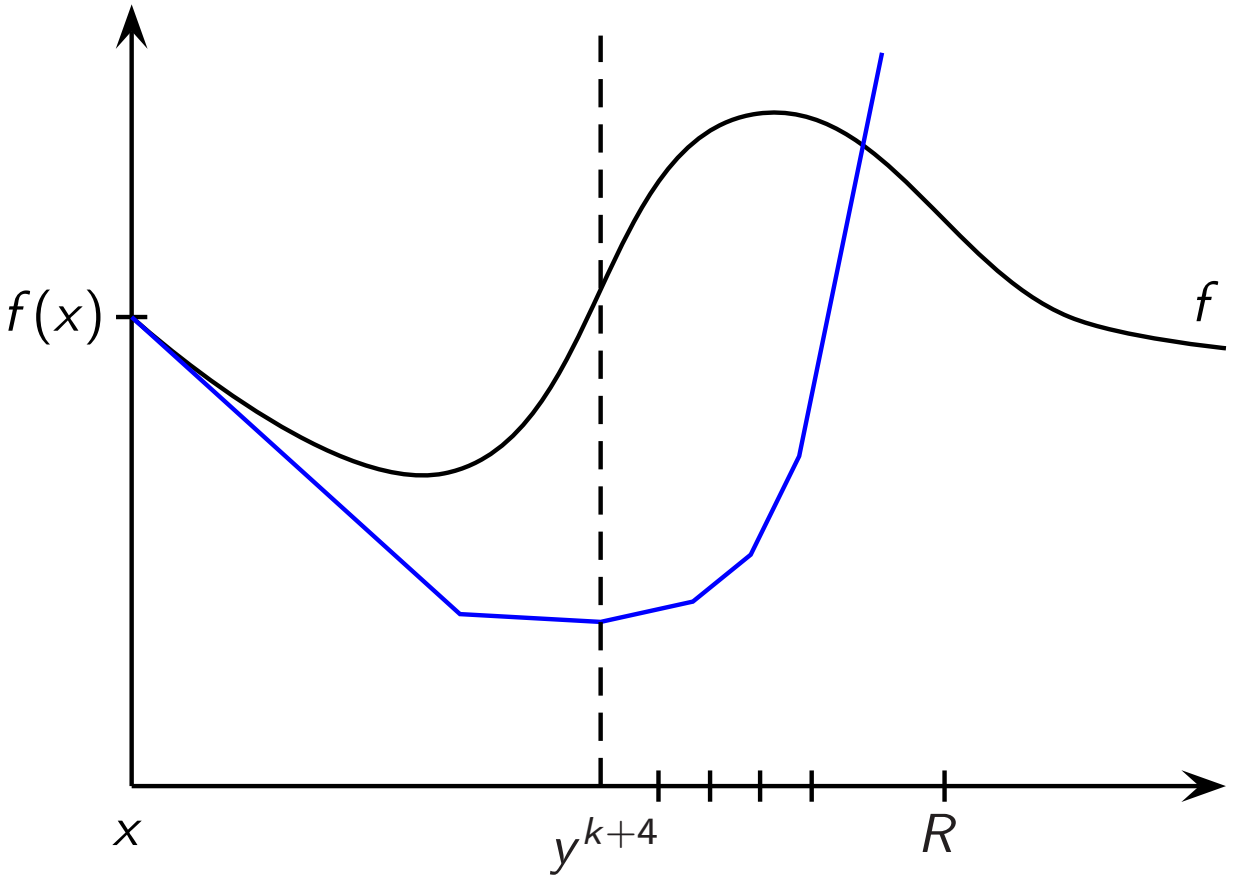
$\phi_{k+4}(\cdot, x)$ working model



$\phi_{k+4}(\cdot, x)$ working model

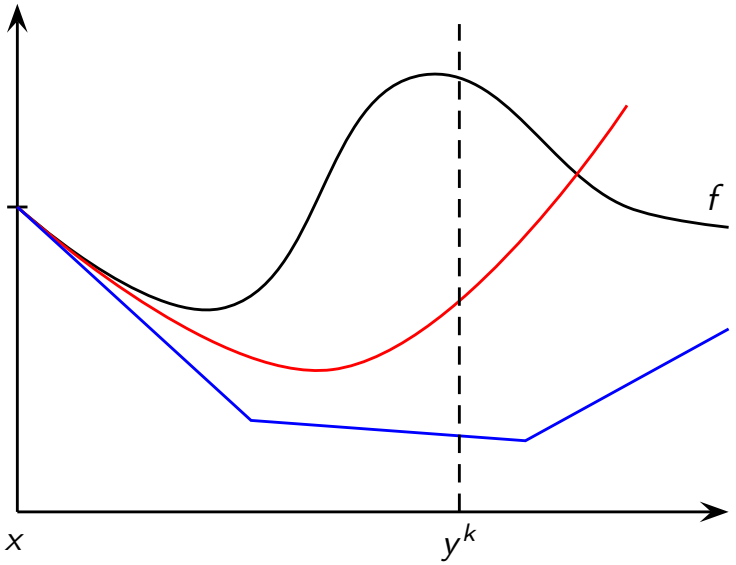


$\phi_{k+4}(\cdot, x)$ working model

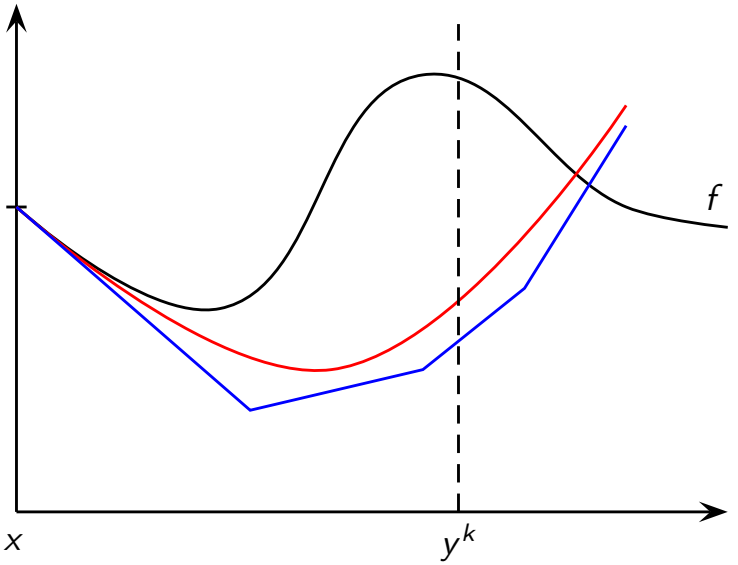


Observe :

- Trust-region radius R can be fixed once and for all in convex trust-region method (Ruszczyński 2007).
- No longer possible without convexity, because ϕ_k approximates ϕ, and not necessarily f.
- Need the option to force smaller steps to get better agreement between ϕ and f :
  - reduce $R^+ = R/2$
- But at what stage should we reduce trust-region radius ?
  - should we *always* backtrack when null step ? ?
- Don't want to consent to backtracking too willingly, as this leads to exceedingly small steps.



$R^+ = R$



$R^+ = \frac{1}{2}R$

$$\widetilde{\rho}_k = \frac{f(x) - \phi(y^k, x)}{f(x) - \phi_k(y^k, x)} \overset{??}{\approx} 1$$

Secondary test :

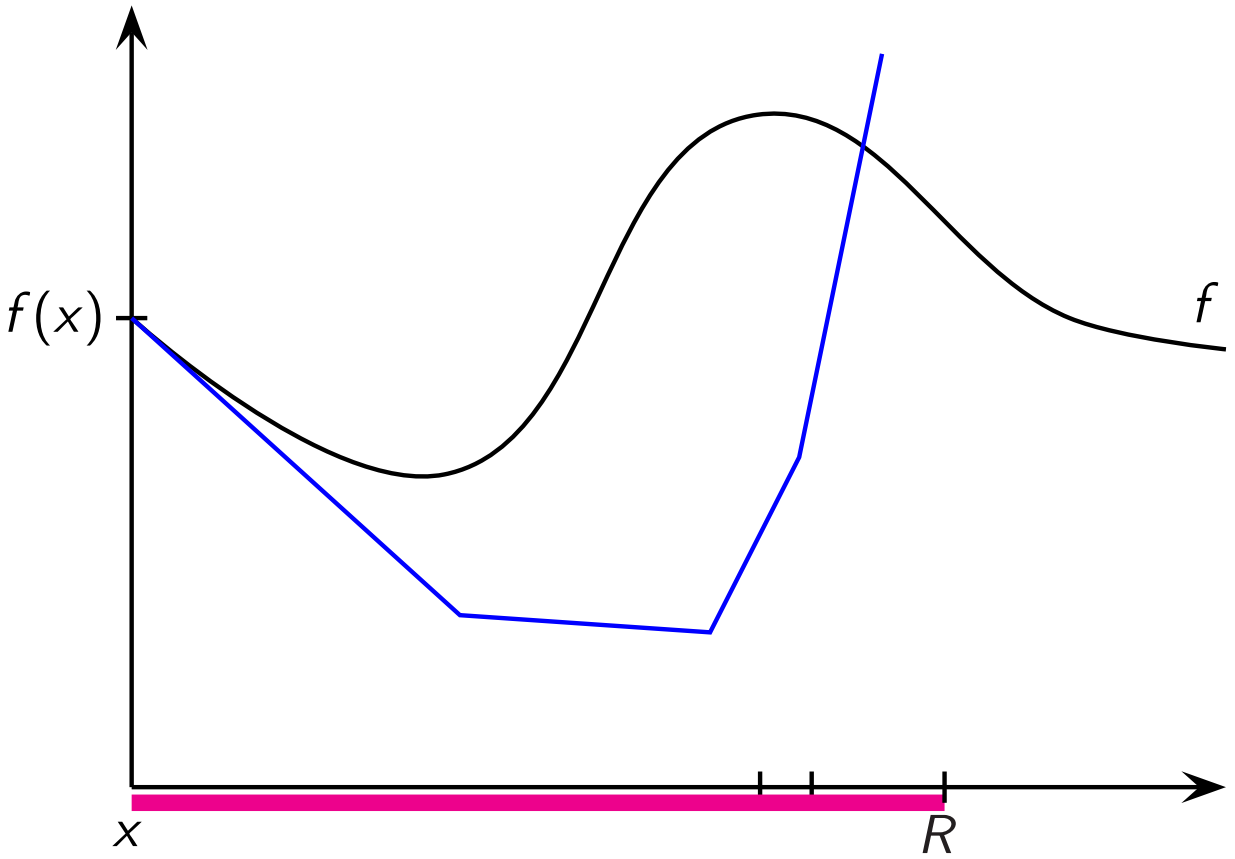
$$\widetilde{\rho}_k = \frac{f(x) - \phi(y^k, x)}{f(x) - \phi_k(y^k, x)}$$

$$= \frac{\text{model progress at } y^k \text{ had we already known the cutting plane}}{\text{progress predicted by working model } \phi_k(\cdot, x) \text{ at } y^k}$$

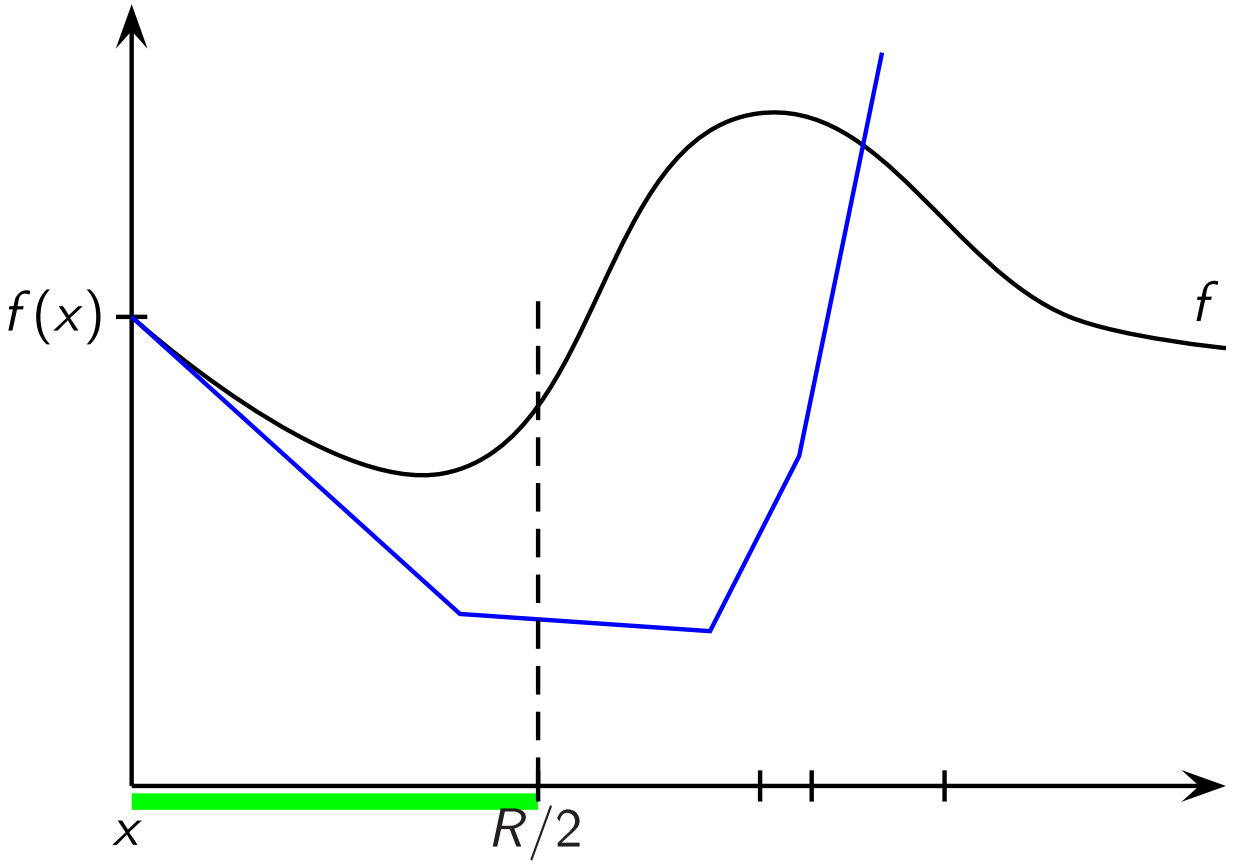
Decision :

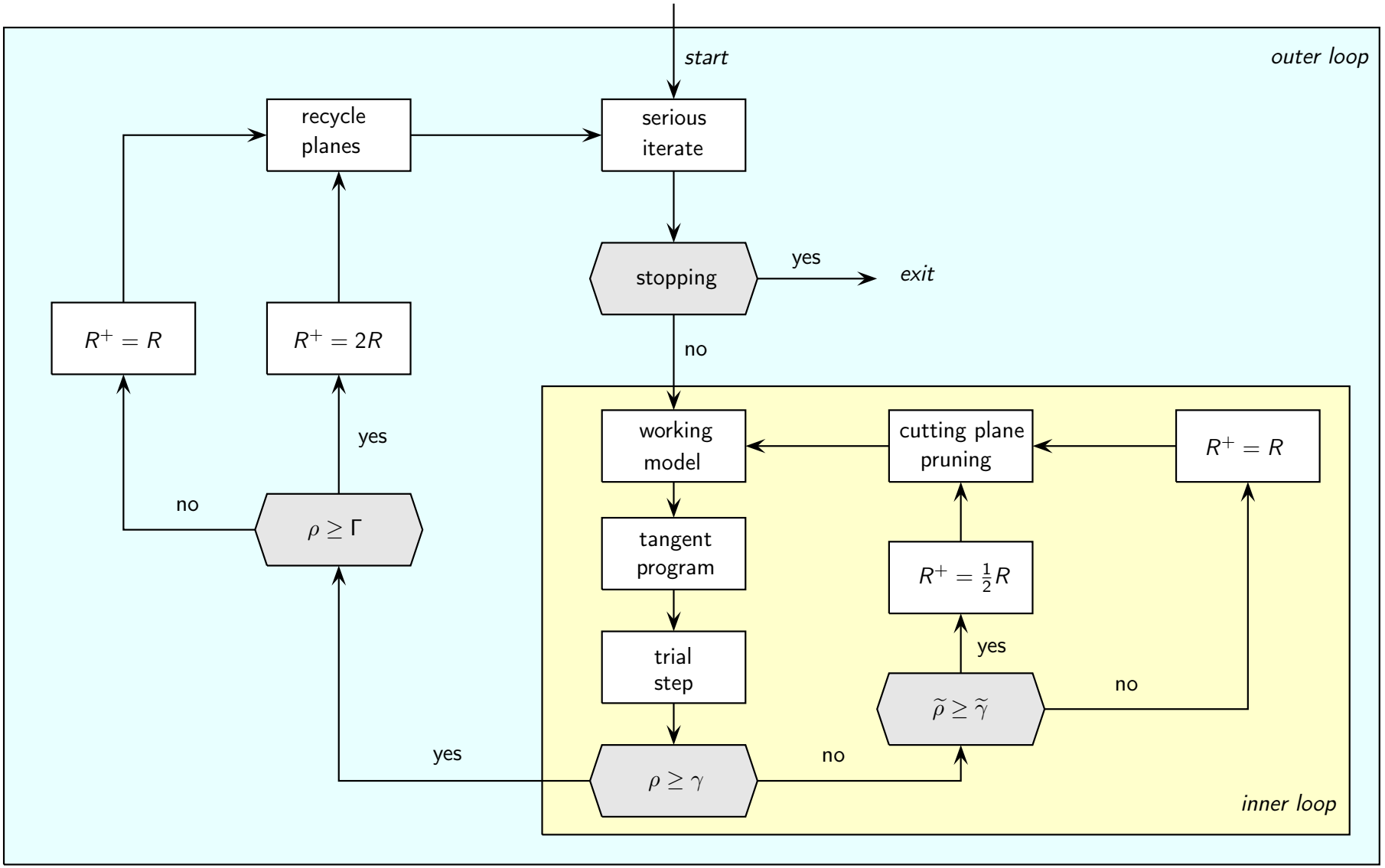
$$R_{k+1} = \begin{cases} R_k & \text{if } \widetilde{\rho}_k < \widetilde{\gamma} \text{ and } \rho_k < \gamma \\ \frac{1}{2} R_k & \text{if } \widetilde{\rho}_k \geq \widetilde{\gamma} \text{ and } \rho_k < \gamma \end{cases}$$

$$\min_{\|y-x\| \le R} \phi_{k+2}(y, x)$$



$$\min_{\|y-x\| \le R/2} \phi_{k+2}(y, x)$$




start
outer loop
recycle planes
serious iterate
stopping
yes
exit
no
$R^+ = R$
$R^+ = 2R$
yes
no
$\rho \geq \Gamma$
working model
cutting plane pruning
$R^+ = R$
tangent program
$R^+ = \frac{1}{2}R$
trial step
yes
$\widetilde{\rho} \geq \widetilde{\gamma}$
no
yes
$\rho \geq \gamma$
no
inner loop

Tangent program :

$$\min_{\|y-x^j\|\leq R} \phi_k(y, x^j)$$

Acceptance test :

$$\rho_k = \frac{f(x^j) - f(y^k)}{f(x^j) - \phi_k(y^k, x^j)} = \frac{\text{actual progress}}{\text{model predicted progress}} \stackrel{?}{\geq} \gamma$$

Pruning :

Can limit number of cuts to $n+1$ by Carathéodory's theorem.

Aggregation à la Kiwiel is open problem.

Trial step : Accept neighbouring points z^k

$$f(x^j) - \phi_k(z^k, x^j) \geq \theta \left(f(x^j) - \phi_k(y^k, x^j) \right)$$

Theorem

Suppose $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ *is bounded and* f *has a strict ideal model* ϕ*. Let* x^j *be the sequence of serious iterates generated by the bundle trust-region method. Then every accumulation point* x^* *of* x^j *is critical.*

Stopping test :

$$\min\{\|g_j^*\|, \|g_{j'}^*\|\} \to 0$$

Here :

If R was never reduced during *jth* inner loop, then $j' < j$ is the largest index of an inner loop, where such a reduction occurred for the last time. Otherwise $j' = j$.

Observe :

- For strong model ϕ just take aggregate at acceptance : $\|g_j^*\| < \epsilon$. (Applies to composite convex functions).
- Rigorous stopping test justifies stopping when tangent program gives only slight progress.
- As compared to bundle method proof gives new challenges.
- Even for convex f convergence to single limit not assured.
- y^k ersatz for Cauchy point.

Can we save the Cauchy point ?

Observe :

- If f is almost everywhere strictly differentiable, then can choose z^k as point of strict differentiability near y^k.
- If in addition $\phi^\sharp$ is used, bundle trust-region method coincides with its classical alter ego based on first-order model :

$$\min_{\|y-x^j\|\leq R} f(x^j) + \nabla f(x^j)^T(y - x^j)$$

- Then y^k = Cauchy point.
- But only justified if $\phi^\sharp$ strict.

Proposition

f upper C^1 $\implies$ f has strict standard model $\phi^\sharp$.

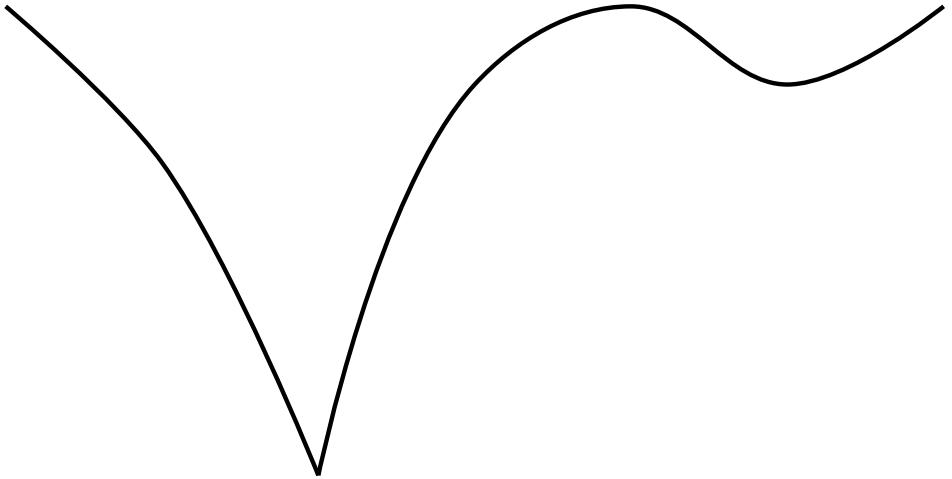
Proof. Upper C^1 at $\bar{x}$ $\implies$ $\forall \epsilon > 0\ \exists \delta > 0\ \forall x, x + td \in B(\bar{x}, \delta)$, $\|d\| = 1$

$$\frac{f(x + td) - f(x)}{t} \leq -f^0(x, -d) + \epsilon.$$

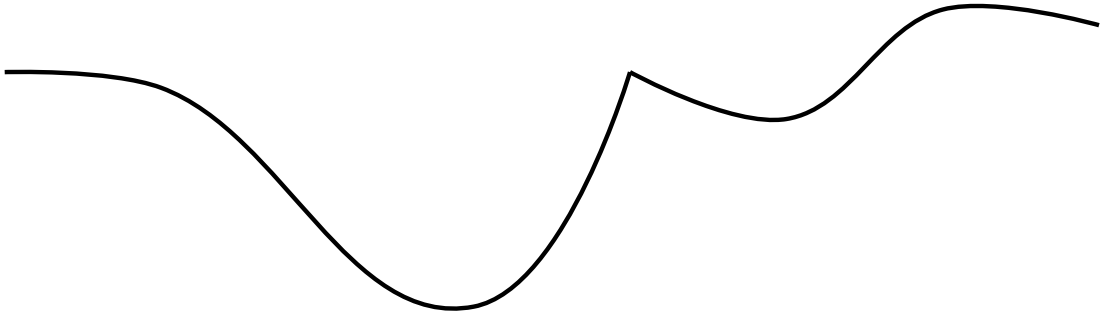
(Daniilidis & Georgiev 2004), (Ngai, Luc, Thera 2000).
Strictness at $\bar{x}$ means the formally weaker :

$$\frac{f(x + td) - f(x)}{t} \leq f^0(x, d) + \epsilon.$$

□

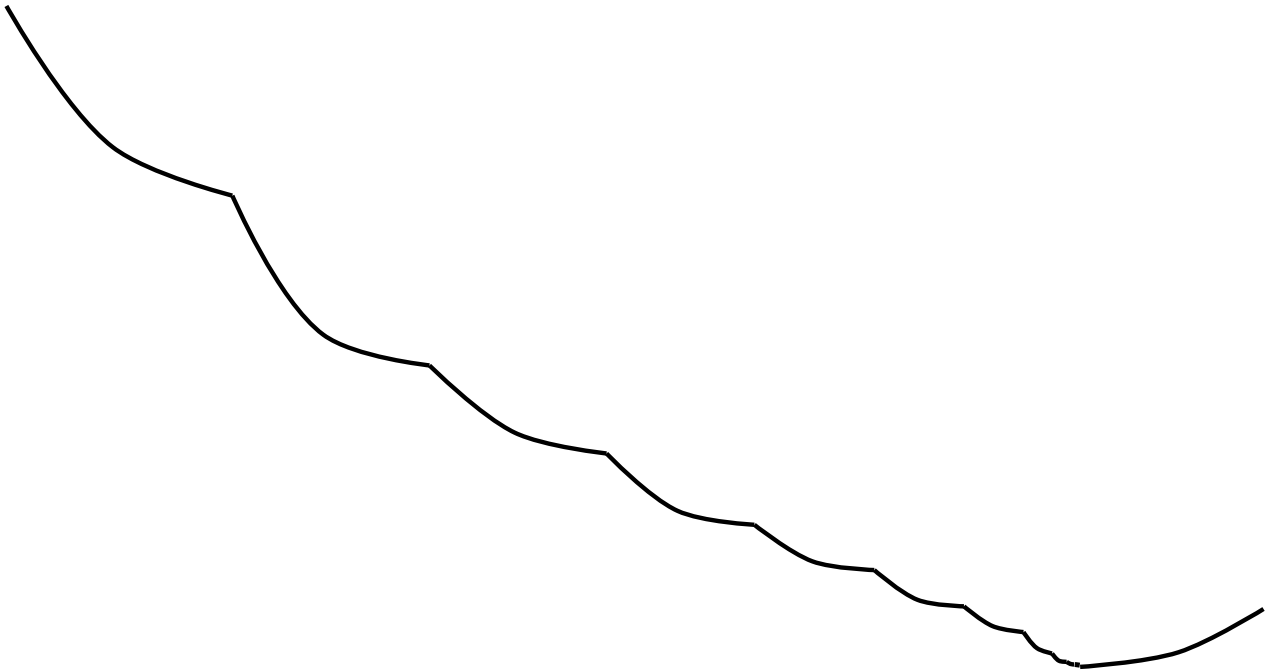


lower C^1



upper C^1

Example. Upper C^1-function with non smoothness near minimum.



Upward kinks accumulate at minimum
No local minima on curve

Example of upper-C^1 function :

$$\begin{aligned} f(x) &= f_s(x) + \int_0^1 \min_{i \in I} f_i(x,t)\, dt \\ &= f_s(x) + \min_{\sigma \in I^{[0,1]}} \int_0^1 f_{\sigma(t)}(x,t)\, dt \end{aligned}$$

is upper C^1 if first-order partial derivatives with respect to x are continuous.

M.N. Dao, J. Gwinner, D. Noll, N. Ovcharova. Nonconvex bundle method with application to a delamination problem. `arXiv`

Observe :

- Lightning function (Klatte & Kummer 2002) has strict $\phi^\sharp$, but is not upper-C^1.
- Suppose $x =$ serious, $z^k =$ trial step. Then $\phi^\sharp$ cutting plane has subgradient $g \in \partial f(x)$ which attains maximum

$$f^\circ(x, z^k - x) = g^T(z^k - x).$$

- f strictly differentiable at $x \implies$ only one $\phi^\sharp$ cutting plane $\implies$ method becomes classical trust-region method.
- Functions with strict standard model $\phi^\sharp$ can be optimized as if they were smooth. Cauchy point can be saved.
- Not possible for convex non-smooth optimization.

Applications with upper-C^1 :

- Delamination of composite materials (Dao, Gwinner, Noll, Ovcharova 2014). `arXiv`.
- Parametric robust feedback control design (Apkarian, Dao, Noll 2014, IEEE TAC).
- Branch and bound algorithm for the robustness analysis of uncertain systems (Apkarian, Noll, Ravanbod 2015).

Favorable use of $\|\cdot\|_\infty$ instead of $\|\cdot\|_2$.

Concluding remarks :

- Investigate possibilities for trial step z^k.
  $\Longrightarrow$ include second-order information if available
  $\Longrightarrow$ non-convex tangent QP
  $\Longrightarrow$ non-smooth BFGS.
- For f only locally Lipschitz use ϵ-model, where $\partial_1\phi(x,x) \subset \partial f(x) + \epsilon B$.
  $\Longrightarrow$ strict ϵ-model always exists
  $\Longrightarrow$ convergence to x^* with $0 \in \partial f(x^*) + \epsilon B$.
- Aggregation à la Kiwiel.