New Gradient and Hessian Approximation Methods for Derivative-free Optimisation

Chayne Planiden

School of Mathematics and Statistics
University of Wollongong

October 8, 2020

The Moreau envelope and the proximal mapping

The Moreau envelope of a proper, lsc function f is defined:

$$e_r f(x) = \inf_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{r}{2}\|y - x\|^2 \right\}.$$

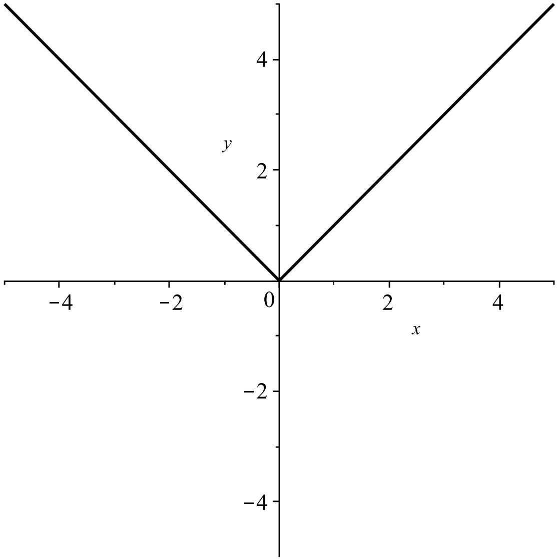
The proximal mapping is the set of points that yield the infimum:

$$P_r f(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{r}{2}\|y - x\|^2 \right\}.$$

Why are the Moreau envelope and proximal mapping useful?

- The Moreau envelope is a smoothing function, and for convex functions it maintains the same minimum value and minimizers.
- For convex functions, the gradient of the Moreau envelope has closed form.
- As the parameter r is increased, the Moreau envelope of f converges to f.
- The proximal mapping is a key component of many Optimization algorithms.

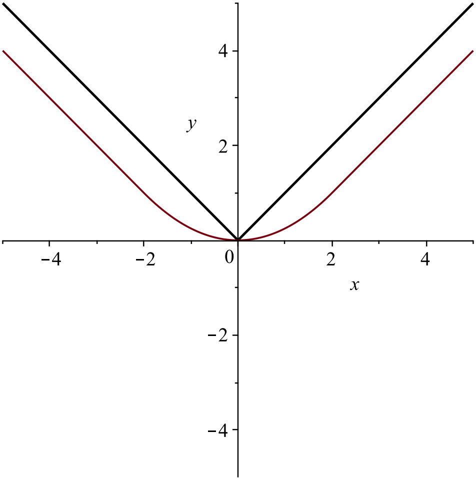
The Moreau envelope and the proximal mapping



The Moreau envelope and the proximal mapping

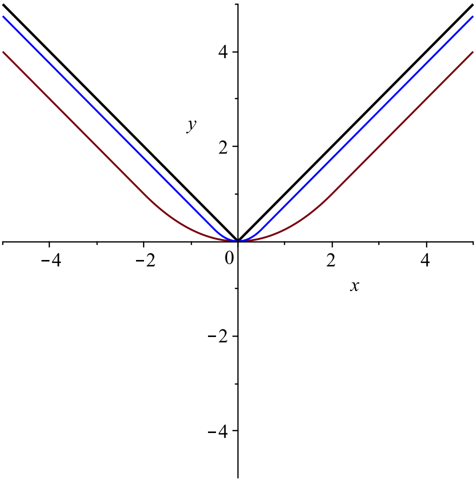
$$
\begin{aligned}
e_r f(x) &= \inf_y \left\{ |y| + \frac{r}{2}(y-x)^2 \right\} \\
&= \min \left[\inf_{y<0} \left\{ -y + \frac{r}{2}(y-x)^2 \right\}, \frac{r}{2}x^2, \inf_{y>0} \left\{ y + \frac{r}{2}(y-x)^2 \right\} \right] \\
&= \begin{cases} -x - \frac{1}{2r}, & \text{if } x < -\frac{1}{r}, \\ \frac{r}{2}x^2, & \text{if } -\frac{1}{r} \leq x \leq \frac{1}{r}, \\ x - \frac{1}{2r}, & \text{if } x > \frac{1}{r} \end{cases}
\end{aligned}
$$

The Moreau envelope and the proximal mapping

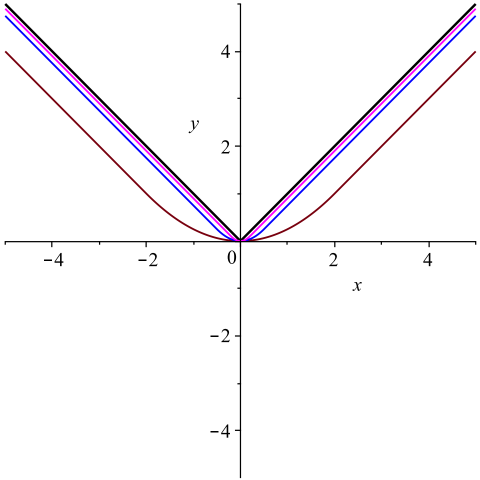


$r = 1$

The Moreau envelope and the proximal mapping



$r = 2$



$r = 5$

The gradient of $e_r f$

- For convex functions, the gradient of $e_r f$ is defined by:

$$\nabla e_r f(x) = r(x - p),$$

  where p is the proximal point of f at x.

- Since $\min f = \min e_r f$, the problem is converted into a smooth one .

Application to norm functions

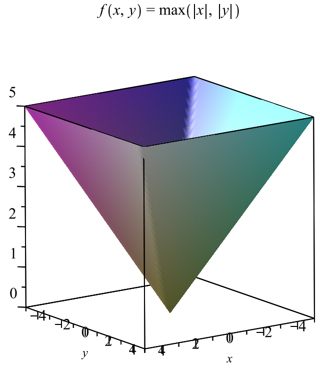
Theorem

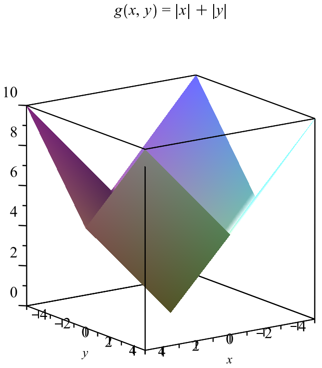
Let f be any norm function on a Hilbert space. Then the function

$$\sqrt{e_r(f^2)}$$

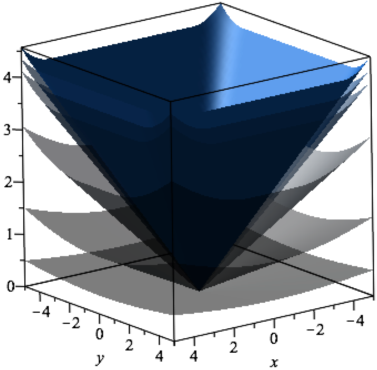
is also a norm function, and it is differentiable everywhere except at the origin.

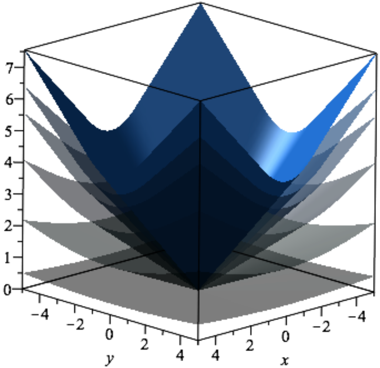
Application to norm functions





Application to norm functions

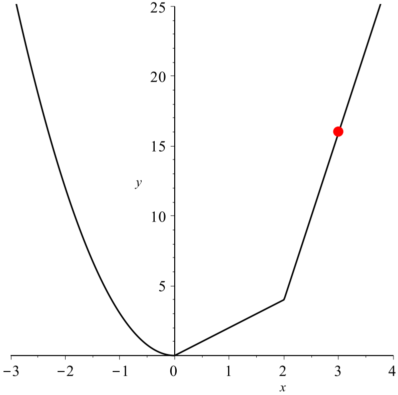




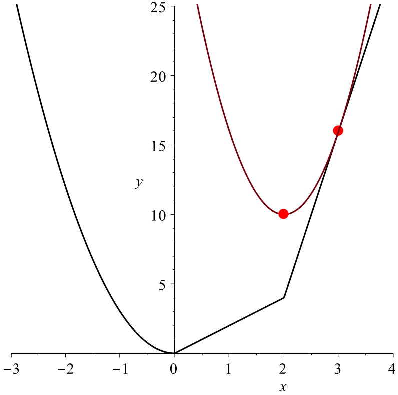
The proximal point algorithm is a minimization algorithm for convex nonsmooth functions developed by Martinet [1970], simple and beautiful:

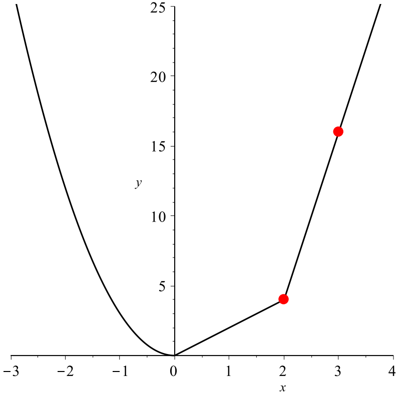
$$x_{k+1} = P_r f(x_k).$$

The proximal point algorithm

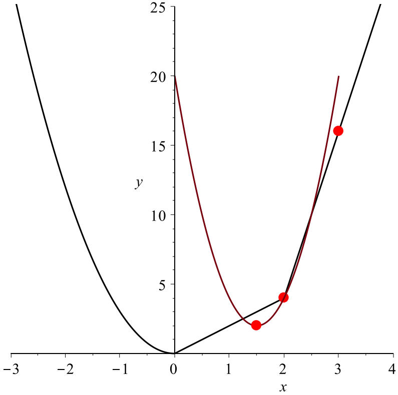


The proximal point algorithm

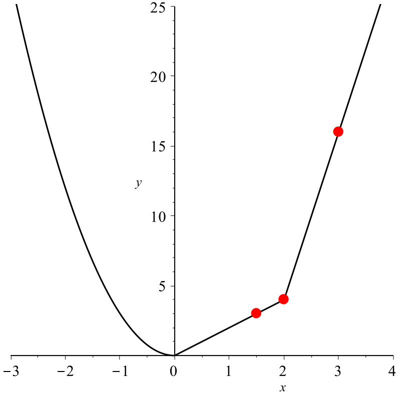



25
20
15
10
5
y
−3
−2
−1
0
1
2
3
4
x

The proximal point algorithm



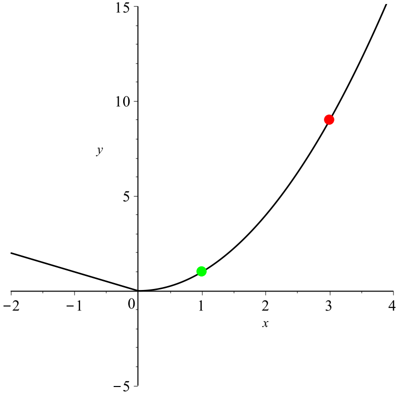
The proximal point algorithm



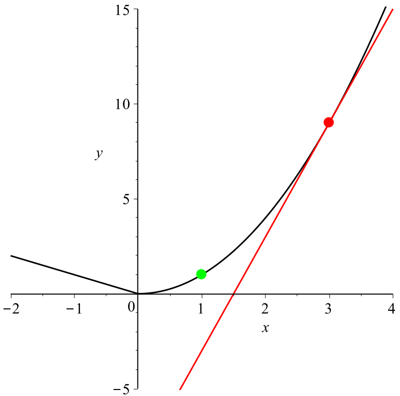
The proximal bundle method

- It can be difficult to find the proximal point.
- A proximal bundle method approximates f with a piecewise-linear function and finds the prox-point of the model function [Kiwiel 1995, Bonnans et al. 1997].
- The bundle is a collection of information recorded at each iteration to improve the model function at the next iteration.

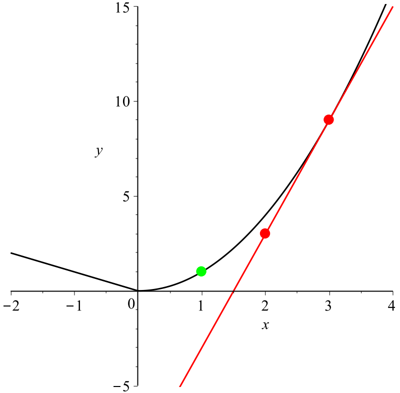
The proximal bundle method



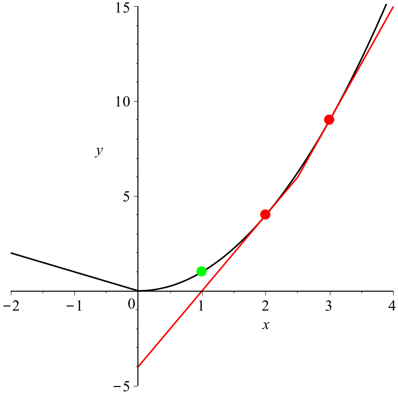
The proximal bundle method



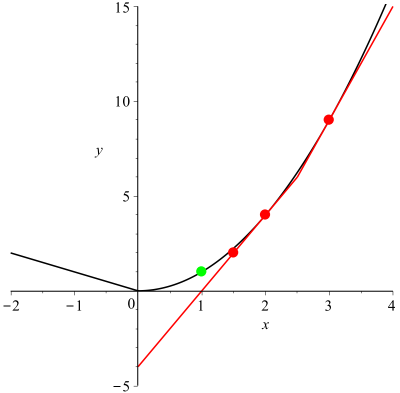
The proximal bundle method



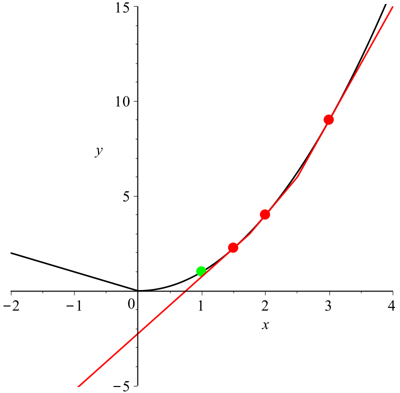
The proximal bundle method



The proximal bundle method



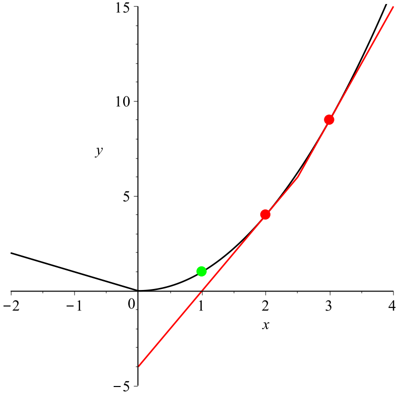
The proximal bundle method



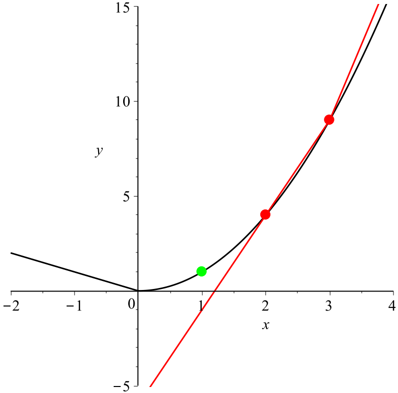
The proximal bundle method

- Proximal bundle methods are used as subroutines in many optimization algorithms.
- Derivative-free (DFO) methods are useful when finding gradients is either impossible or too expensive to do [Conn et al. 2009].
- We created a derivative-free proximal bundle method and used it in the DFO $\mathcal{VU}$-algorithm.

The proximal bundle method



The proximal bundle method



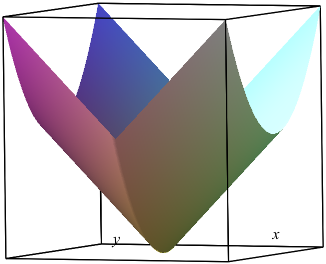
Our purpose in creating the DFO proximal bundle algorithm is to develop a DFO $\mathcal{VU}$-algorithm.

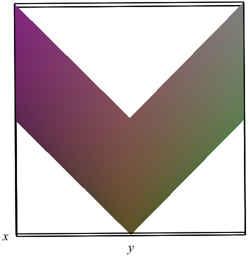
- Prox-point algorithms are slow, but necessary for optimizing nonsmooth functions.
- The $\mathcal{VU}$-algorithm speeds up the process by requiring a proximal step parallel to a subspace of $\mathbb{R}^n$, and then a quasi-Newton step parallel to the remaining subspace.

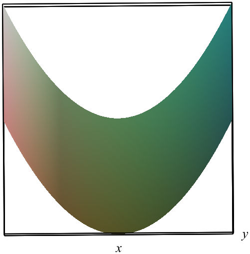
The $\mathcal{VU}$-algorithm

- The idea is to take advantage of the structure of the function, the fact that the nonsmoothness is due to a subspace only.
- We decompose the space into a $\mathcal{V}$-space where the nonsmooth structure is, and the orthogonal $\mathcal{U}$-space where the function behaves smoothly.

$$f(x,y) = x^2 + |y|$$




x
y


y
x

Algorithm:

(0) Initialize.

(1) Decompose: Compute the $\mathcal{VU}$-decomposition of the space at the current point.

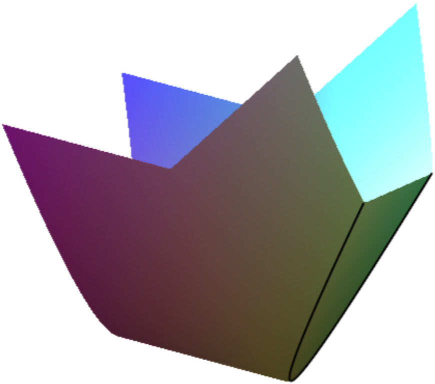
(2) $\mathcal{V}$-step: Run the proximal point method parallel to the $\mathcal{V}$-space.

(3) Stop check: If subgradient norm $\|s_k\|$ is small, then stop.

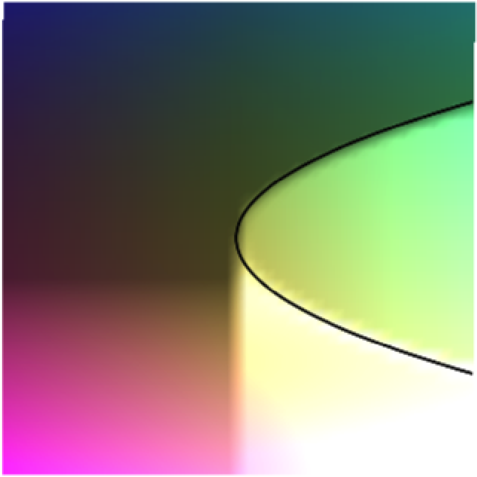
(4) $\mathcal{U}$-step: Find the $\mathcal{U}$-gradient ∇L and $\mathcal{U}$-Hessian $\nabla^2 L$. Take a quasi-Newton step parallel to the $\mathcal{U}$-space by solving

$$\nabla^2 L \Delta u = -\nabla L$$

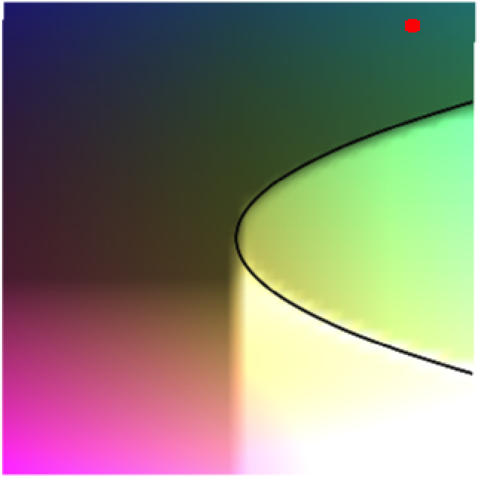
for Δu and setting $x_{k+1} = x_k + \Delta u$. Go to (1).



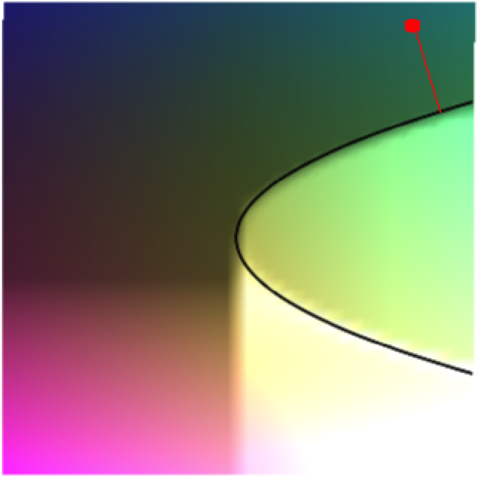
The $\mathcal{VU}$-algorithm



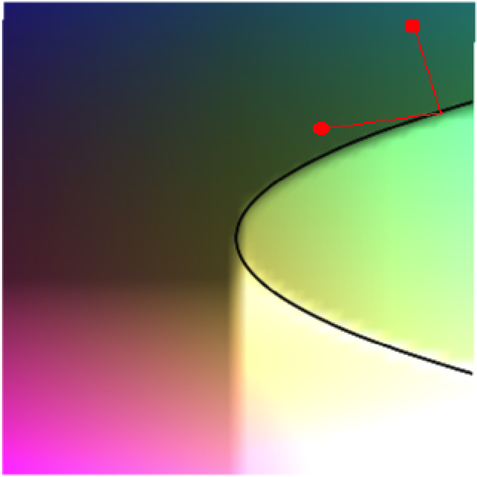
The $\mathcal{VU}$-algorithm

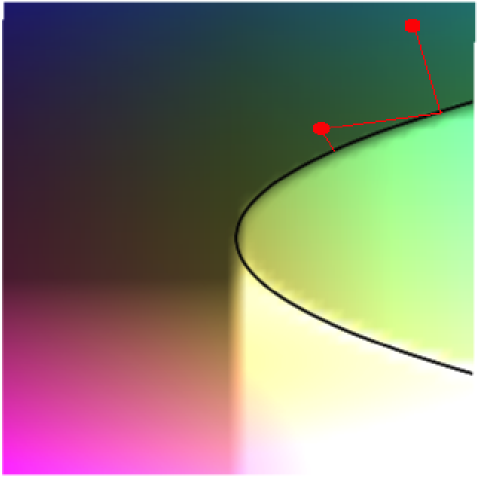


The $\mathcal{VU}$-algorithm

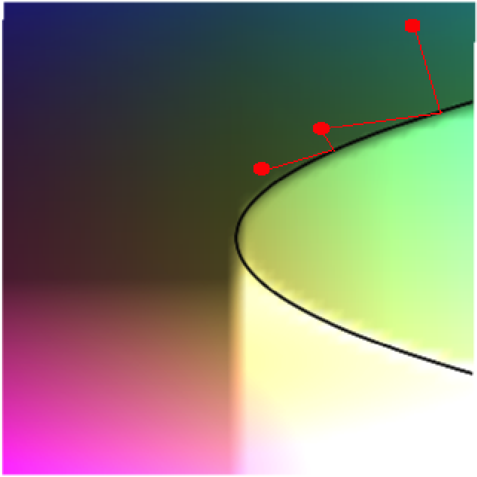


The $\mathcal{VU}$-algorithm

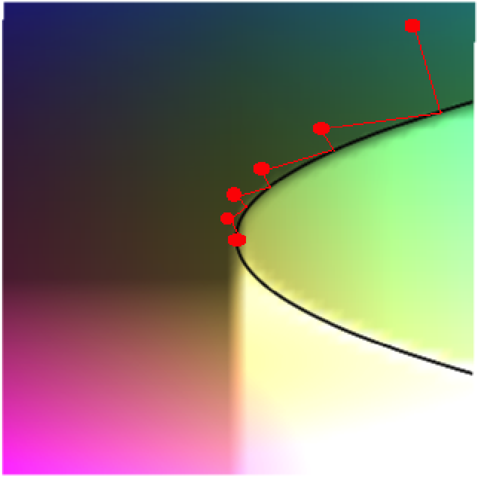




The $\mathcal{VU}$-algorithm



The $\mathcal{VU}$-algorithm

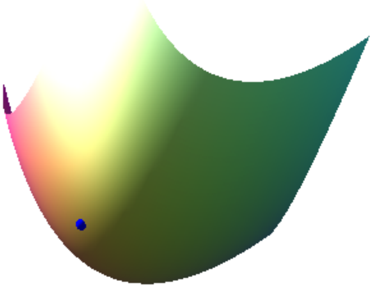


A derivative-free $\mathcal{VU}$-algorithm

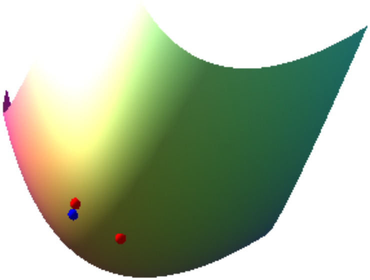
- The $\mathcal{V}$-step requires a proximal point, which can be approximated by our proximal bundle method.
- The $\mathcal{U}$-step requires the gradient and Hessian of f in the $\mathcal{U}$-space. In the DFO version, these are approximated via the simplex gradient and the minimum Frobenius norm, respectively.

The simplex gradient (SG) of f at x is the gradient of the linear interpolation function of f over a set of $n+1$ points close to x on $\mathbb{R}^n$.

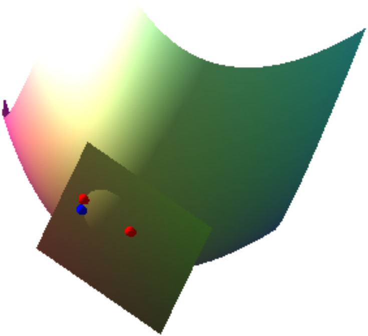
The approximate $\mathcal{U}$-gradient



The approximate $\mathcal{U}$-gradient



The approximate $\mathcal{U}$-gradient



The approximate $\mathcal{U}$-gradient

Definition

Let $\mathcal{X} = [x^0, x^1, \ldots, x^n]$ be affinely independent on $\mathbb{R}^n$. Then $\mathcal{X}$ forms a simplex, and the simplex gradient of f over $\mathcal{X}$ is given by

$$\nabla_s f(\mathcal{X}) = S^{-1}\delta_f(\mathcal{X}),$$

where

$$S = [x^0 - x^1 \quad \cdots \quad x^0 - x^n]^\top \text{ and } \delta_f(\mathcal{X}) = \begin{bmatrix} f(x^0) - f(x^1) \\ \vdots \\ f(x^0) - f(x^n) \end{bmatrix}.$$

For example, the matrix

$$\mathcal{X} = [x^0 \;\; x^0 + \Delta e_1 \;\; x^0 + \Delta e_2 \;\; \cdots \;\; x^0 + \Delta e_n]$$

forms a simplex. The condition number of $\mathcal{X}$ is given by $\|\widehat{S}^{-1}\|$, where

$$\widehat{S} = \frac{1}{\Delta}[x^0 - x^1 \;\; \cdots \;\; x^0 - x^n]^\top \text{ and } \Delta = \max_i \|x^0 - x^i\|.$$

An important feature of the condition number is that it is always possible to keep it from degrading, while making Δ arbitrarily close to zero.

The approximate $\mathcal{U}$-gradient

There is an error bound for the distance between the simplex gradient and the exact gradient:

Theorem

Let $\mathcal{X} = [x^0, x^1, \ldots, x^n]$ *form a simplex. Then there exists* $\mu = \mu(x^0) > 0$ *such that*

$$\|\nabla_s f(\mathcal{X}) - \nabla f(x^0)\| \leq \mu \|\widehat{S}^{-1}\| \Delta.$$

So by controlling Δ, we can approximate our $\mathcal{U}$-gradient as closely as we want.

The approximate $\mathcal{U}$-Hessian

Now to approximate a Hessian, we solve the minimum Frobenius norm problem.

Definition

The Frobenius norm of a matrix $H \in \mathbb{R}^{p\times q}$ with elements a_{ij} is defined by

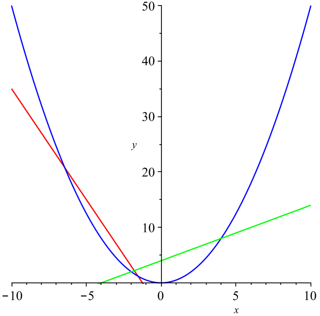
$$\|H_F\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}^2}.$$

The approximate $\mathcal{U}$-Hessian

Note: the DFO $\mathcal{VU}$-algorithm is for finite-max objective functions, i.e. they can be expressed as a max of a finite number of convex functions.

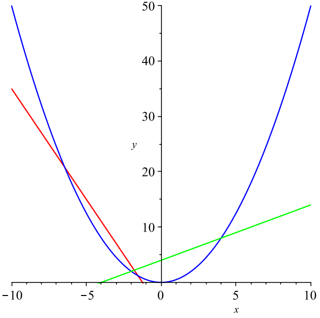
The approximate $\mathcal{U}$-Hessian

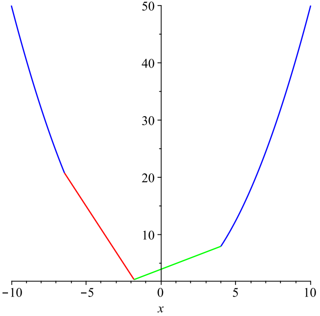
Note: the DFO $\mathcal{VU}$-algorithm is for finite-max objective functions, i.e. they can be expressed as a max of a finite number of convex functions.



The approximate $\mathcal{U}$-Hessian

Note: the DFO $\mathcal{VU}$-algorithm is for finite-max objective functions, i.e. they can be expressed as a max of a finite number of convex functions.





The approximate $\mathcal{U}$-Hessian

We use the matrix

$$Z = [x \;\; x + \Delta e_1 \;\; x - \Delta e_1 \;\; \cdots \;\; x + \Delta e_n \;\; x - \Delta e_n].$$

For each active function f_i at the current point x (i.e. $f_i(x) = f(x)$), and for $j = 1, \ldots, 2n+1$, we solve

$$\nabla_F^2 f_i(Z) = \operatorname{argmin} \|H_i\|_F \text{ such that } \frac{1}{2} Z_j^\top H_i Z_j + B_i^\top Z_j + C_i = f_i(Z_j),$$

where Z_j is column j of Z. With variables H_i, B_i, C_i, this is a quadratic programming problem.

The approximate $\mathcal{U}$-Hessian

Then, denoting

$$H = \frac{1}{|A(x)|} \sum_{i \in A(x)} \nabla_F^2 f_i(Z),$$

we define the approximate $\mathcal{U}$-Hessian:

$$\nabla_U^2 L = U^\top H U,$$

where U is a basis for the $\mathcal{U}$-space and $A(x)$ is the active set of functions at x.

The approximate $\mathcal{U}$-Hessian

Theorem

There exists $\mu = \mu(x)$ such that

$$\|\nabla^2_U L - \nabla^2 L\| \leq \left[2\sqrt{2}\sqrt{|A(x) - 1|}\|\|V^\dagger\|\|\|H\|(2\mu + \mu^2\Delta) + \mu\right]\Delta,$$

where V is a basis for the $\mathcal{V}$-space.

So once again, by controlling Δ we get as close an approximation to the $\mathcal{U}$-Hessian as necessary.

Algorithm:

(0) Initialize.

(1) Decompose: Compute the $\mathcal{VU}$-decomposition of the space at the current point.

(2) $\mathcal{V}$-step: Run the DFO proximal bundle method to find the prox-point within ε_k.

(3) Stop check: If ε_k and subgradient norm $\|s_k\|$ are small, stop.

(4) $\mathcal{U}$-step: Approximate the $\mathcal{U}$-gradient ∇L with the simplex gradient $\nabla_s L$, and the $\mathcal{U}$-Hessian $\nabla^2 L$ with the argmin of the minimum Frobenius norm $\nabla^2_{\varepsilon_k} L$. Solve

$$\nabla^2_{\varepsilon_k} L \Delta u = -\nabla_s L$$

for Δu and setting $x_{k+1} = x_k + \Delta u$. Go to (1).

Now, we want to improve the performance by using better gradient and Hessian approximations, and to generalise by relaxing the requirement on the number of points needed.

New Approximations

Now, we want to improve the performance by using better gradient and Hessian approximations, and to generalise by relaxing the requirement on the number of points needed.

- The generalised simplex gradient (GSG) does not necessarily use $n+1$ points in $\mathbb{R}^n$. It can be more (overdetermined case) or fewer (underdetermined case).

Now, we want to improve the performance by using better gradient and Hessian approximations, and to generalise by relaxing the requirement on the number of points needed.

- The generalised simplex gradient (GSG) does not necessarily use $n+1$ points in $\mathbb{R}^n$. It can be more (overdetermined case) or fewer (underdetermined case).
- Using k points, the GSG is defined

$$\nabla_s f(\mathcal{X}) = (S^\top)^\dagger \delta_f(\mathcal{X}),$$

where $S \in \mathbb{R}^{n \times k}$ and $S^\dagger$ is the Moore–Penrose pseudoinverse of S.

New Approximations

We have shown that

$$\|\nabla_s f(\mathcal{X}) - \nabla f(x^0)\| \leq \frac{\sqrt{k}}{2} L_{\nabla f} \left\| \left(\widehat{S}(\mathcal{X})^\top \right)^\dagger \right\| \Delta,$$

where $\widehat{S} = S/\Delta$ and $L_{\nabla f}$ is the Lipschitz constant of ∇f. We have also developed calculus rules for ∇_s (similar to the CSG coming up).

New Approximations

- The centred simplex gradient (CSG) uses $2n+1$ points in $\mathbb{R}^n$ rather than $n+1$, but it offers an error bound on the order of Δ^2.

New Approximations

- The centred simplex gradient (CSG) uses $2n+1$ points in $\mathbb{R}^n$ rather than $n+1$, but it offers an error bound on the order of Δ^2.
- Using the simplex $\mathcal{X} = \{x^0, x^1, \ldots, x^n\} = \{x^0, x^0 + d^1, \ldots, x^0 + d^n\}$ and the reflection $\mathcal{X}^- = \{x^0, x^0 - d^1, \ldots, x^0 - d^n\}$, we define

$$\delta_f^c(\mathcal{X}) = \begin{bmatrix} f(x^0 + d^1) - f(x^0 - d^1) \\ f(x^0 + d^2) - f(x^0 - d^2) \\ \vdots \\ f(x^0 + d^n) - f(x^0 - d^n) \end{bmatrix}$$

- The centred simplex gradient (CSG) uses $2n+1$ points in $\mathbb{R}^n$ rather than $n+1$, but it offers an error bound on the order of Δ^2.
- Using the simplex $\mathcal{X} = \{x^0, x^1, \ldots, x^n\} = \{x^0, x^0 + d^1, \ldots, x^0 + d^n\}$ and the reflection $\mathcal{X}^- = \{x^0, x^0 - d^1, \ldots, x^0 - d^n\}$, we define

$$\delta_f^c(\mathcal{X}) = \begin{bmatrix} f(x^0 + d^1) - f(x^0 - d^1) \\ f(x^0 + d^2) - f(x^0 - d^2) \\ \vdots \\ f(x^0 + d^n) - f(x^0 - d^n) \end{bmatrix}$$

- Then the CSG is defined

$$\nabla_c f(\mathcal{X}) = (S^\top)^{-1} \delta_f^c(\mathcal{X}).$$

New Approximations

It can be proved that the CSG is the average of two SGs:

$$\nabla_c f(\mathcal{X}) = \frac{1}{2}(\nabla_s f(\mathcal{X}) + \nabla_s f(\mathcal{X}^-)).$$

New Approximations

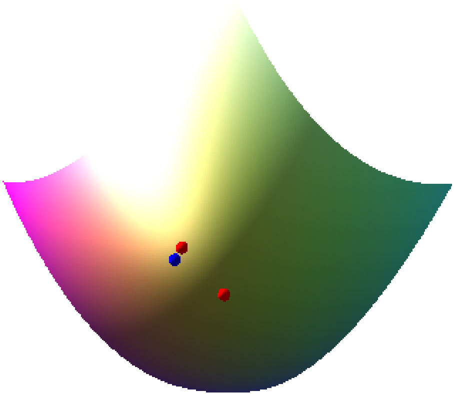
It can be proved that the CSG is the average of two SGs:

$$\nabla_c f(\mathcal{X}) = \frac{1}{2}(\nabla_s f(\mathcal{X}) + \nabla_s f(\mathcal{X}^-)).$$

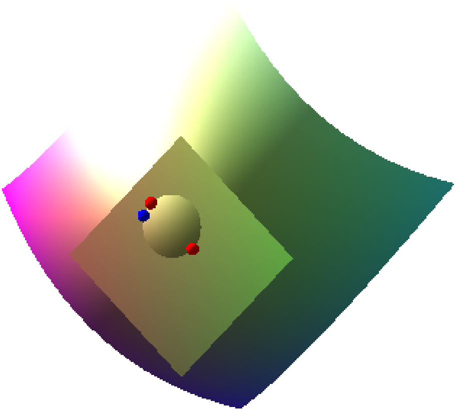
Now we generalise to any k points rather than $n+1$ and define the generalised centred simplex gradient (GCSG):

$$\nabla_c f(\mathcal{X}) = \left(S(\mathcal{X})^\top\right)^\dagger \delta_f^c(\mathcal{X}).$$

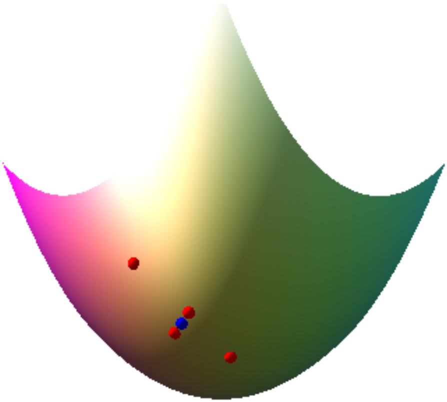
New Approximations



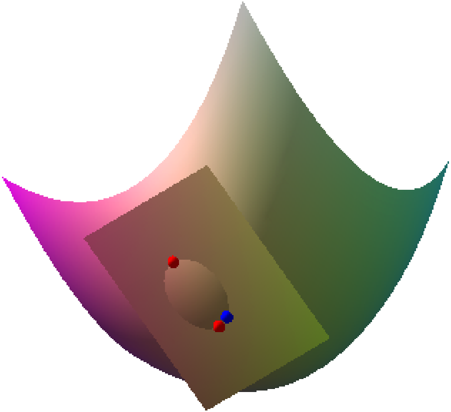
New Approximations



New Approximations



New Approximations



Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^{2+}$ on $B_{x^0}(\Delta)$ with $\nabla^2 f$ having Lipschitz constant L. Let $\mathcal{X} = [x^0 \ \cdots \ x^k]$ be well-poised. Then

$$\|\nabla_c f(\mathcal{X}) - \nabla f(x^0)\| \leq \frac{L\sqrt{k}}{6} \left\| \left(\widehat{S}(\mathcal{X})^\top \right)^\dagger \right\| \Delta^2, \qquad \text{(overdet.)}$$

$$\|\nabla_c f(\mathcal{X}) - \nabla f_U(x^0)\| \leq \frac{L\sqrt{k}}{6} \left\| \left(\widehat{S}(\mathcal{X})^\top \right)^\dagger \right\| \Delta^2, \qquad \text{(underdet.)}$$

where ∇f_U is the orthogonal projection of ∇f onto k-dimensional subspace U.

New Approximations

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^{2+}$ on $B_{x^0}(\Delta)$ with $\nabla^2 f$ having Lipschitz constant L. Let $\mathcal{X} = [x^0 \ \cdots \ x^k]$ be well-poised. Then

$$\|\nabla_c f(\mathcal{X}) - \nabla f(x^0)\| \leq \frac{L\sqrt{k}}{6} \left\| \left(\widehat{S}(\mathcal{X})^\top\right)^\dagger \right\| \Delta^2, \qquad \text{(overdet.)}$$

$$\|\nabla_c f(\mathcal{X}) - \nabla f_U(x^0)\| \leq \frac{L\sqrt{k}}{6} \left\| \left(\widehat{S}(\mathcal{X})^\top\right)^\dagger \right\| \Delta^2, \qquad \text{(underdet.)}$$

where ∇f_U is the orthogonal projection of ∇f onto k-dimensional subspace U.

We get order Δ^2 because in the Taylor-expansion proof, the first-order terms of $\mathcal{X}$ and $\mathcal{X}^-$ cancel out.

We created calculus rules as follows.

New Approximations

We created calculus rules as follows.

$\nabla(fg) = f\nabla g + g\nabla f$, so define

$\nabla_c(fg)(\mathcal{X}) = f(x^0)\nabla_c g(\mathcal{X}) + g(x^0)\nabla_c f(\mathcal{X})$.

We created calculus rules as follows.

$\nabla(fg) = f\nabla g + g\nabla f$, so define

$\nabla_c(fg)(\mathcal{X}) = f(x^0)\nabla_c g(\mathcal{X}) + g(x^0)\nabla_c f(\mathcal{X})$.

Then

$$\|\nabla_c(fg)(\mathcal{X}) - \nabla(fg)(x^0)\| \leq \frac{\sqrt{k}}{6}(L_g|f(x^0)| + L_f|g(x^0)|)\left\|\left(\widehat{S}^\top\right)^\dagger\right\|\Delta^2$$

where $|\mathcal{X}| = k+1$ and L_f, L_g are Lipschitz constants.

New Approximations

$$\|\nabla_c(fg)(\mathcal{X}) - \nabla(fg)(x^0)\| \leq \frac{\sqrt{k}}{6}(L_g|f(x^0)| + L_f|g(x^0)|) \left\| \left(\widehat{S}^\top\right)^\dagger \right\| \Delta^2$$

$$\|\nabla_c(f^p)(\mathcal{X}) - \nabla(f^p)(x^0)\| \leq \frac{L\sqrt{k}}{6} p|f(x^0)|^{p-1} \left\| \left(\widehat{S}^\top\right)^\dagger \right\| \Delta^2$$

$$\left\|\nabla_c\left(\frac{f}{g}\right)(\mathcal{X}) - \nabla\left(\frac{f}{g}\right)(x^0)\right\| \leq \frac{\sqrt{k}}{6}\left(L_f\left|\frac{1}{g(x^0)}\right| + L_g\left|\frac{f(x^0)}{g^2(x^0)}\right|\right) \left\| \left(\widehat{S}^\top\right)^\dagger \right\| \Delta^2$$

$$\|\nabla_c(f\circ g)(\mathcal{X}) - \nabla(f\circ g)(x^0)\| \leq \frac{\sqrt{k}p}{6}\left(\sqrt{k}L_{g_*}L_f\|(\widehat{S}^\top)^\dagger\| + \|\nabla f(g(x^0))\|L_{g_*^2}\right)\|(\widehat{S}^\top)^\dagger\|\Delta_*^2$$

New Approximations

$$\|\nabla_c(fg)(\mathcal{X}) - \nabla(fg)(x^0)\| \leq \frac{\sqrt{k}}{6}(L_g|f(x^0)| + L_f|g(x^0)|)\left\|\left(\widehat{S}^\top\right)^\dagger\right\|\Delta^2$$

$$\|\nabla_c(f^p)(\mathcal{X}) - \nabla(f^p)(x^0)\| \leq \frac{L\sqrt{k}}{6}p|f(x^0)|^{p-1}\left\|\left(\widehat{S}^\top\right)^\dagger\right\|\Delta^2$$

$$\left\|\nabla_c\left(\frac{f}{g}\right)(\mathcal{X}) - \nabla\left(\frac{f}{g}\right)(x^0)\right\| \leq \frac{\sqrt{k}}{6}\left(L_f\left|\frac{1}{g(x^0)}\right| + L_g\left|\frac{f(x^0)}{g^2(x^0)}\right|\right)\left\|\left(\widehat{S}^\top\right)^\dagger\right\|\Delta^2$$

$$\|\nabla_c(f\circ g)(\mathcal{X}) - \nabla(f\circ g)(x^0)\| \leq \frac{\sqrt{k}p}{6}\left(\sqrt{k}L_{g_*}L_f\|(\widehat{S}^\top)^\dagger\| + \|\nabla f(g(x^0))\|L_{g_*^2}\right)\|(\widehat{S}^\top)^\dagger\|\Delta_*^2$$

So the order Δ^2 gives much closer approximations as $\Delta \searrow 0$ and will certainly improve algorithm convergence rates.

New Approximations

- In much the same way, we made a new approximation for the Hessian, called the nested-set Hessian.

New Approximations

- In much the same way, we made a new approximation for the Hessian, called the nested-set Hessian.
- For second-order information, we need two sets of directions rather than one. Both are generalised to any finite number of points, and can be different numbers.

New Approximations

- In much the same way, we made a new approximation for the Hessian, called the nested-set Hessian.
- For second-order information, we need two sets of directions rather than one. Both are generalised to any finite number of points, and can be different numbers.

Let $f : \mathbb{R}^n \to \mathbb{R}$ and define

$$S = [s^1 \;\; s^2 \;\; \cdots \;\; s^m] \in \mathbb{R}^{n \times m}$$
$$T = [t^1 \;\; t^2 \;\; \cdots \;\; t^k] \in \mathbb{R}^{n \times k}$$

such that $x^0, x^0 + s^i, x^0 + t^j, x^0 + s^i + t^j \in \operatorname{dom} f$.

New Approximations

Define

$$\delta_f(x^0; T) = \begin{bmatrix} f(x^0 + t^1) - f(x^0) \\ f(x^0 + t^2) - f(x^0) \\ \vdots \\ f(x^0 + t^k) - f(x^0) \end{bmatrix}.$$

Define

$$\delta_f(x^0; T) = \begin{bmatrix} f(x^0 + t^1) - f(x^0) \\ f(x^0 + t^2) - f(x^0) \\ \vdots \\ f(x^0 + t^k) - f(x^0) \end{bmatrix}.$$

Using the notation

$$\nabla_s f(x^0; T) = (T^\top)^\dagger \delta_f(x^0; T)$$

for the GSG, we define the nested-set Hessian

$$\nabla_s^2 f(x^0; S, T) = (S^\top)^\dagger \delta_{\nabla_s f}(x^0; S, T),$$

where

$$\delta_{\nabla_s f}(x^0; S, T) = \begin{bmatrix} (\nabla_s f(x^0 + s^1; T) - \nabla_s f(x^0; T))^\top \\ (\nabla_s f(x^0 + s^2; T) - \nabla_s f(x^0; T))^\top \\ \vdots \\ (\nabla_s f(x^0 + s^m; T) - \nabla_s f(x^0; T))^\top \end{bmatrix}.$$

- With careful choices of S and T, we can guarantee at most $(n+1)(n+2)/2$ function evaluations.

New Approximations

- With careful choices of S and T, we can guarantee at most $(n+1)(n+2)/2$ function evaluations.
- The error bound between the nested-set Hessian and the true Hessian is order Δ.

New Approximations

- With careful choices of S and T, we can guarantee at most $(n+1)(n+2)/2$ function evaluations.
- The error bound between the nested-set Hessian and the true Hessian is order Δ.

$$\|\nabla_s^2 f(x^0; S, T) - \nabla^2 f(x^0)\| \leq \frac{m\sqrt{k}}{3} L_{\nabla^2 f} \left(2\frac{\Delta_u}{\Delta_l} + 3\right) \left\| \left(\widehat{S}^\top\right)^\dagger \right\| \left\| \widehat{T}^\dagger \right\| \Delta_u$$

New Approximations

- With careful choices of S and T, we can guarantee at most $(n+1)(n+2)/2$ function evaluations.
- The error bound between the nested-set Hessian and the true Hessian is order Δ.

$$\|\nabla_s^2 f(x^0; S, T) - \nabla^2 f(x^0)\| \leq \frac{m\sqrt{k}}{3} L_{\nabla^2 f} \left(2\frac{\Delta_u}{\Delta_l} + 3\right) \left\| \left(\widehat{S}^\top\right)^\dagger \right\| \left\| \widehat{T}^\dagger \right\| \Delta_u$$

- With further care (for instance $m = k = n$ and unit canonical directions), the bound can be further improved (for instance $\frac{11}{2} n^2 L_{\nabla^2 f}$).

New Approximations

Calculus rules:

$$\nabla^2(fg) = (\nabla^2 f)g + \nabla f(\nabla g)^\top + \nabla g(\nabla f)^\top + (\nabla^2 g)f,$$

so we define

$$\nabla_s^2(fg) = (\nabla_s^2 f)g + \nabla_s f(\nabla_s g)^\top + \nabla_s g(\nabla_s f)^\top + (\nabla_s^2 g)f.$$

New Approximations

Then

$$\|\nabla_s^2(fg)(x^0;S,T)-\nabla^2(fg)(x^0)\| \leq (E_{\nabla^2 sf}|g(x^0)|+E_{\nabla_s^2 g}|f(x^0)|+2M_{fg}^s)\Delta_u,$$

where

$$M_{fg}^s = \min\left\{\begin{array}{l} E_{\nabla_s f}E_{\nabla_s g}\Delta_u + E_{\nabla_s g}\|\nabla f(x^0)\| + E_{\nabla_s f}\|\nabla g(x^0)\|, \\ E_{\nabla_s f}\|\nabla_s g(x^0;T)\| + E_{\nabla_s g}\|\nabla f(x^0)\|, \\ E_{\nabla_s g}\|\nabla_s f(x^0;T)\| + E_{\nabla_s f}\|\nabla g(x^0)\| \end{array}\right\}.$$

New Approximations

Then

$$\|\nabla_s^2(fg)(x^0;S,T)-\nabla^2(fg)(x^0)\| \leq (E_{\nabla^2 sf}|g(x^0)|+E_{\nabla_s^2 g}|f(x^0)|+2M_{fg}^s)\Delta_u,$$

where

$$M_{fg}^s = \min\left\{\begin{array}{l} E_{\nabla_s f}E_{\nabla_s g}\Delta_u + E_{\nabla_s g}\|\nabla f(x^0)\| + E_{\nabla_s f}\|\nabla g(x^0)\|, \\ E_{\nabla_s f}\|\nabla_s g(x^0;T)\| + E_{\nabla_s g}\|\nabla f(x^0)\|, \\ E_{\nabla_s g}\|\nabla_s f(x^0;T)\| + E_{\nabla_s f}\|\nabla g(x^0)\| \end{array}\right\}.$$

Similar for quotient rule and power rule.

Summary.

- The generalised centred simplex gradient $\nabla_s f$ provides an improvement from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta^2)$ on the error bound with ∇f.

Summary.

- The generalised centred simplex gradient $\nabla_s f$ provides an improvement from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta^2)$ on the error bound with ∇f.
- It also provides greater flexibility on the number of points needed in the "simplex".

New Approximations

Summary.

- The generalised centred simplex gradient $\nabla_s f$ provides an improvement from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta^2)$ on the error bound with ∇f.
- It also provides greater flexibility on the number of points needed in the "simplex".
- The nested-set Hessian $\nabla_s^2 f$ provides $\mathcal{O}(\Delta)$ error bound with $\nabla^2 f$, as does the minimum Frobenius norm.

New Approximations

Summary.

- The generalised centred simplex gradient $\nabla_s f$ provides an improvement from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta^2)$ on the error bound with ∇f.
- It also provides greater flexibility on the number of points needed in the "simplex".
- The nested-set Hessian $\nabla_s^2 f$ provides $\mathcal{O}(\Delta)$ error bound with $\nabla^2 f$, as does the minimum Frobenius norm.
- However, Frobenius works for finite-max functions and the nested-set Hessian is not restricted to any particular class of functions.

Numerical experiments are forthcoming...

New Approximations

PhD Scholarship

Convergence Speed of Optimisation Algorithms

SCHOOL OF MATHEMATICS AND APPLIED STATISTICS (SMAS), UNIVERSITY OF WOLLONGONG, AUSTRALIA

An exciting PhD scholarship is available in the School of Mathematics and Applied Statistics (SMAS) at the University of Wollongong, South Western Sydney campus, in the area of Optimisation. The title of the project is Determining the Convergence Speed of Derivative-free Optimisation Algorithms. The UOW scholarship is $28,092AUD tax-free per year for three years full-time. Tuition fees (for up to 4 years) will be waived. The successful applicant will have the opportunity to work with both Australian and international collaborators, and extra funding may be available for conference travel.

Applications are invited from domestic and international students who are able to commence PhD studies at the University of Wollongong in 2021. Applicants should hold, or be close to completing, an Honours 1 undergraduate degree or a Master's degree in Applied Mathematics, Computational Mathematics or a closely related field. The ideal candidate will have an interest in optimisation algorithms and be happy to work with derivative-free methods. Excellent mathematical and programming skills and an interest in using them to further current research are essential. Self-motivation, a strong research potential and good oral and written communication skills are indispensable qualities the winning candidate will have.

The successful applicant will learn about Nonsmooth Optimisation techniques that do not use gradient information. Instead, they use approximations of gradients and are termed Derivative-free methods. In particular, the recently developed Derivative-free VU-algorithm (DFOVU) is of interest. The candidate will investigate, posit and prove the convergence speed and properties of the DFOVU method, comparing to the classical VU-algorithm and to other derivative-free methods. Extensive numerical testing and validation of convergence rates against similar methods is part of the research project.

HOW TO APPLY

If you are interested in applying for this scholarship, please contact Dr Chayne Planiden via email: chayne@uow.edu.au. Applications must include CV detailing previous education experience and academic transcripts. It is expected that the successful applicant will be available to commence this scholarship by 31 October 2021.

Applications close 30 November, 2020.

MORE INFORMATION

Dr Chayne Planiden, Lecturer
School of Mathematics & Applied Statistics, University of Wollongong, NSW, Australia
Email: chayne@uow.edu.au

UOW.EDU.AU #THISISUOW

U O W

Thank you!

1. Planiden and Wang, Strongly convex functions, Moreau envelopes and the generic nature of functions with strong minimizers, ***SIAM Journal on Optimization*** 26(2), 2016.

2. Planiden and Wang, Epiconvergence, the Moreau envelope and generalized linear-quadratic functions, ***Journal of Optimization Theory and Applications***, minor revisions 2017.

3. Hare and Planiden, Computing proximal points of convex functions with inexact subgradients, ***Set-valued and Variational Analysis***, accepted 2016.

4. Hare and Planiden and Sagastizábal, A derivative-free $\mathcal{VU}$-algorithm for convex finite-max problems, ***Mathematical Programming***, submitted 2017.

5. Hare and Jarry–Bolduc and Planiden, Error bounds for overdetermined and underdetermined generalized centred simplex gradients, ***IMA Journal of Numerical Analysis***, to appear.

6. Hare and Jarry–Bolduc and Planiden, Hessian Approximations, ***ArXiV preprint***.