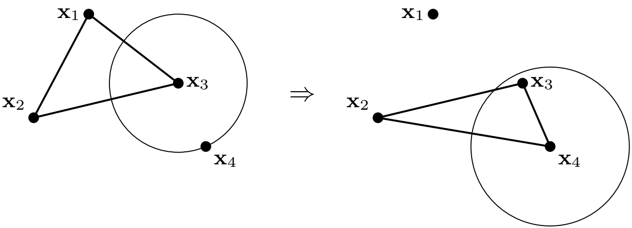
The Error of Multivariate Linear Extrapolation with Applications to Derivative-Free Optimization

Liyuan Cao, Zaiwen Wen

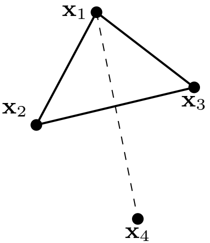
Peking University

2nd Derivative-Free Optimization Symposium
June 28, 2024

Motivation



(a) linear interpolation + trust region method



(b) simplex method

Problem Definition

objective function	$f : \mathbb{R}^n \to \mathbb{R}$
interpolation set	$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}\} \subset \mathbb{R}^n$ affinely independent
linear interpolation model	$\hat{f}(\mathbf{x}) = c + \mathbf{g} \cdot \mathbf{x}$ such that

$$\begin{bmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ & \vdots \\ 1 & \mathbf{x}_{n+1}^T \end{bmatrix} \begin{bmatrix} c \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_{n+1}) \end{bmatrix}.$$

Question: Assume $f \in C_\nu^{1,1}(\mathbb{R}^n)$, i.e.,

$$\|Df(\mathbf{u}) - Df(\mathbf{v})\| \leq \nu \|\mathbf{u} - \mathbf{v}\| \text{ for all } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Given $\{\mathbf{x}_i\}_{i=1}^{n+1}$ and $\mathbf{x}$, what is the (sharp) upper bound on the function approximation error $|\hat{f}(\mathbf{x}) - f(\mathbf{x})|$, particularly when $\mathbf{x} \notin \operatorname{conv}\left(\{\mathbf{x}_i\}_{i=1}^{n+1}\right)$?

Existing Results

1. **seminal work on interpolation error:** Philippe G Ciarlet and Pierre-Arnaud Raviart. "General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods". In: *Archive for Rational Mechanics and Analysis* 46.3 (1972), pp. 177–199

Theorem (error of general Lagrange interpolation)

Let $\hat{f}$ be a polynomial of degree d that interpolates a $d+1$ times continuous differentiable f on a poised set.

$$D^m \hat{f}(\mathbf{x}) - D^m f(\mathbf{x}) = \frac{1}{(d+1)!} \sum_{i=1}^{\binom{n+d}{d}} \left\{ D^{d+1} f(\xi_i) \cdot (\mathbf{x}_i - \mathbf{x})^{d+1} \right\} D^m \ell_i(\mathbf{x}),$$

where $\xi_i = \alpha_i \mathbf{x}_i + (1-\alpha_i)\mathbf{x}$ for some α_i.

2. **sharp bound on LI error:** Shayne Waldron. "The error in linear interpolation at the vertices of a simplex". In: *SIAM Journal on Numerical Analysis* 35.3 (1998), pp. 1191–1200

Theorem (sharp bound on linear interpolation)

Let $\mathbf{c}$ be the center and R the radius of the unique sphere containing $\Theta = \{\mathbf{x}_i\}_{i=1}^{n+1}$. Then, for each $\mathbf{x} \in conv(\Theta)$, there is the sharp inequality

$$|\hat{f}(\mathbf{x}) - f(\mathbf{x})| \leq \frac{1}{2} \left(R^2 - \|\mathbf{x} - \mathbf{c}\|^2 \right) |||D^2 f|||_{L_\infty(\mathrm{conv}(\Theta))}.$$

Definition (Lagrange Polynomial)

Given an affinely independent set $\{\mathbf{x}_i\}_{i=1}^{n+1} \subset \mathbb{R}^n$, a set of $n+1$ linear functions $\{\ell_j\}_{j=1}^{n+1}$ is called a basis of Lagrange polynomials if

$$\ell_j(\mathbf{x}_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Additionally, we define

$$\mathbf{x}_0 = \mathbf{x} \quad \text{and} \quad \ell_0 : \mathbb{R}^n \to -1.$$

They have the following properties:

$$\sum_{i=1}^{n+1} \ell_i(\mathbf{x}) f(\mathbf{x}_i) = \hat{f}(\mathbf{x}),$$

$$\sum_{i=0}^{n+1} \ell_i(\mathbf{x}) = 0,$$

$$\text{and} \quad \sum_{i=0}^{n+1} \ell_i(\mathbf{x}) \mathbf{x}_i = \mathbf{0}.$$

Define

$$\mathcal{I}_+ = \{i \in \{0, \ldots, n+1\} : \ell_i(\mathbf{x}) > 0\}$$
$$\mathcal{I}_- = \{i \in \{0, \ldots, n+1\} : \ell_i(\mathbf{x}) < 0\}.$$

Error Estimation Problem

Because **the sharp upper bound on error = the largest possible error**, the question can be formulated as

$$\max_f |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \quad \text{s.t.} \;\; f \in C^{1,1}_{\nu}(\mathbb{R}^n).$$

Because **the sharp upper bound on error = the largest possible error**, the question can be formulated as

$$\max_f |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \quad \text{s.t. } f \in C_\nu^{1,1}(\mathbb{R}^n).$$

This infinite dimensional problem has a finite dimensional equivalent

$$\begin{aligned} \max_{\mathbf{g}_i, y_i} \quad & \sum_{i=0}^{n+1} \ell_i(\mathbf{x}) y_i \\ \text{s.t.} \quad & y_j \leq y_i + \frac{1}{2}(\mathbf{g}_i + \mathbf{g}_j) \cdot (\mathbf{x}_j - \mathbf{x}_i) + \frac{\nu}{4}\|\mathbf{x}_j - \mathbf{x}_i\|^2 \\ & \qquad - \frac{1}{4\nu}\|\mathbf{g}_j - \mathbf{g}_i\|^2 \;\forall i, j = 0, 1, \ldots, n+1. \end{aligned}$$

Error Estimation Problem

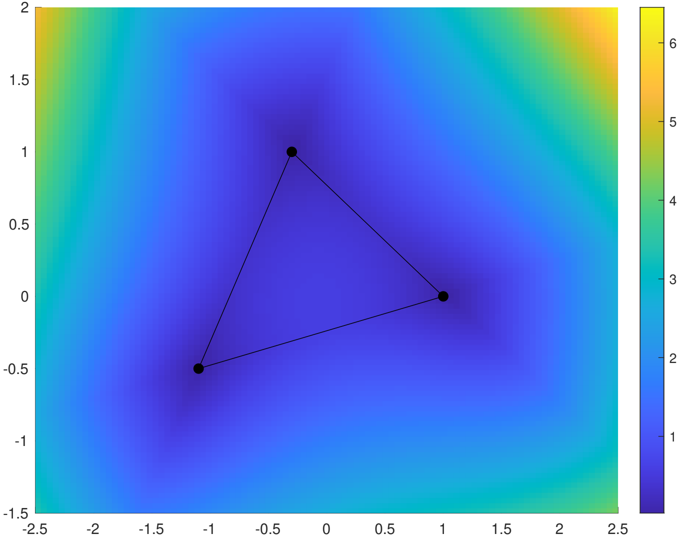


Figure: The sharp error bound on $|\hat{f}(\mathbf{x}) - f(\mathbf{x})|$ for each $\mathbf{x}$ on the 100×100 grid covering $[-2.5, 2.5] \times [-1.5, 2.5]$, where $\Theta = \{(-0.3, 1), (-1.1, -0.5), (1, 0)\}$ and $\nu = 1$.

An Improved Upper Bound

Theorem (An Improved Upper Bound)

Assume $f \in C^{1,1}_{\nu}(\mathbb{R}^n)$. *Let linear* $\hat{f}$ *interpolate* f *at* $\{\mathbf{x}_i\}_{i=1}^{n+1} \subset \mathbb{R}^n$. *Then*

$$\hat{f}(\mathbf{x}) - f(\mathbf{x}) \leq \frac{\nu}{2} \sum_{i=0}^{n+1} |\ell_i(\mathbf{x})| \|\mathbf{x}_i - \mathbf{u}\|^2 \text{ for any } \mathbf{u} \in \mathbb{R}^n.$$

Proof.

The bound is the weighted sum of the following inequalities

$$\begin{array}{lll} \ell_i(\mathbf{x}) & f(\mathbf{x}_i) - f(\mathbf{u}) - Df(\mathbf{u}) \cdot (\mathbf{x}_i - \mathbf{u}) \leq \frac{\nu}{2}\|\mathbf{x}_i - \mathbf{u}\|^2 & \text{for all } i \in \mathcal{I}_+, \\ -\ell_j(\mathbf{x}) & -f(\mathbf{x}_j) + f(\mathbf{u}) + Df(\mathbf{u}) \cdot (\mathbf{x}_j - \mathbf{u}) \leq \frac{\nu}{2}\|\mathbf{x}_j - \mathbf{u}\|^2 & \text{for all } j \in \mathcal{I}_-. \end{array}$$

- In existing results from the literature, the function f needs to be twice continuously differentiable and $\mathbf{u} = \mathbf{x}$.
- The point $\mathbf{u}$ can be set to the center of a trust region.
- Minimize the R.H.S. w.r.t. $\mathbf{u}$ to yield

$$\mathbf{u}^\star = \mathbf{w} \overset{\text{def}}{=} \frac{\sum_{i=0}^{n+1} |\ell_i(\mathbf{x})| \mathbf{x}_i}{\sum_{i=0}^{n+1} |\ell_i(\mathbf{x})|}$$

An Improved Upper Bound: Sharpness

Theorem

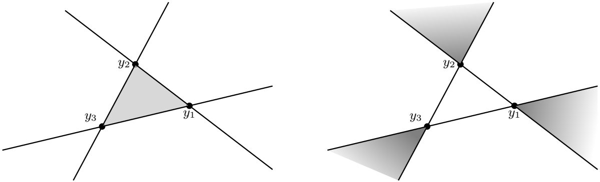
The bound $\hat{f}(\mathbf{x}) - f(\mathbf{x}) \leq \frac{\nu}{2} \sum_{i=0}^{n+1} |\ell_i(\mathbf{x})| \|\mathbf{x}_i - \mathbf{w}\|^2$ *is sharp under either of the two following conditions*

1. $\mathbf{x} \in conv(\Theta)$;
2. *there is only one positive term in* $\{\ell_i(\mathbf{x})\}_{i=1}^{n+1}$.

Proof.

This error can be achieved by the function

1. $f(\mathbf{x}) = \frac{\nu}{2} \|\mathbf{x}\|^2$ for the first case;
2. $f(\mathbf{x}) = -\frac{\nu}{2} \|\mathbf{x}\|^2$ for the second case.



(a) $\mathbf{x} \in \text{conv}(\Theta)$ **(b)** one positive ℓ

Worst Quadratic Function

Let f be a quadratic function of the form

$$f(\mathbf{u}) = c + \mathbf{g} \cdot \mathbf{u} + H\mathbf{u} \cdot \mathbf{u}/2 \text{ with } c \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^n, \text{ and symmetric } H \in \mathbb{R}^{n \times n}.$$

The error estimation problem can be formulated as

$$\begin{array}{ll} \max\limits_{H} & \hat{f}(\mathbf{x}) - f(\mathbf{x}) = G \cdot H/2 \\ \text{s.t.} & -\nu I \preceq H \preceq \nu I, \end{array}$$

where

$$G = \sum_{i=0}^{n+1} \ell_i(\mathbf{x}) \mathbf{x}_i \mathbf{x}_i^T.$$

Worst Quadratic Function

Let f be a quadratic function of the form

$$f(\mathbf{u}) = c + \mathbf{g} \cdot \mathbf{u} + H\mathbf{u} \cdot \mathbf{u}/2 \text{ with } c \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^n, \text{ and symmetric } H \in \mathbb{R}^{n \times n}.$$

The error estimation problem can be formulated as

$$\begin{array}{ll} \max\limits_{H} & \hat{f}(\mathbf{x}) - f(\mathbf{x}) = G \cdot H/2 \\ \text{s.t.} & -\nu I \preceq H \preceq \nu I, \end{array}$$

where

$$G = \sum_{i=0}^{n+1} \ell_i(\mathbf{x}) \mathbf{x}_i \mathbf{x}_i^T.$$

Analytical solution:

$$G \cdot H^\star/2 = \frac{\nu}{2} \sum_{i=1}^{n} |\lambda_i(G)|, \text{ where } \lambda_i\text{'s are the eigenvalues of } G.$$

Worst Quadratic Function

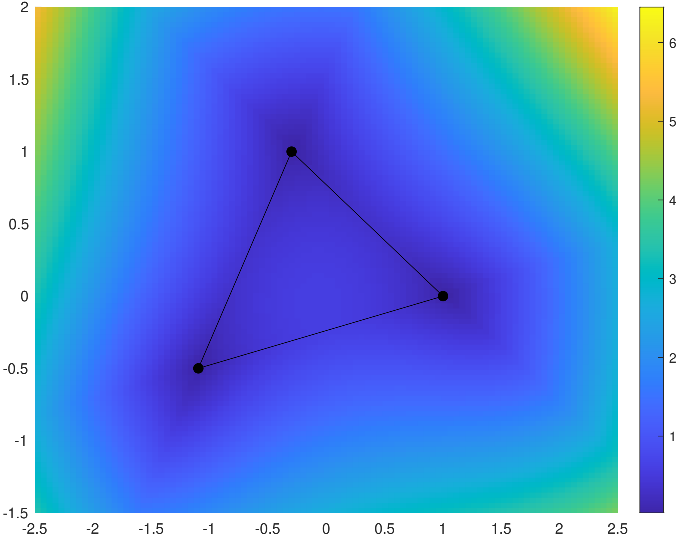
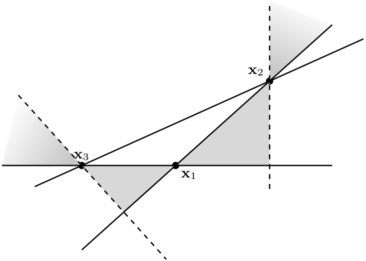


Figure: The sharp error bound on $|\hat{f}(\mathbf{x}) - f(\mathbf{x})|$ for each $\mathbf{x}$ on the 100×100 grid covering $[-2.5, 2.5] \times [-1.5, 2.5]$, where $\Theta = \{(-0.3, 1), (-1.1, -0.5), (1, 0)\}$ and $\nu = 1$.

Worst Quadratic Function: Not Bad Enough



Areas where

$$\max_f |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \quad \text{s.t. } f \in C^{1,1}_\nu(\mathbb{R}^n) \qquad \geq \qquad \max_f |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \quad \text{s.t. } f \in C^{1,1}_\nu(\mathbb{R}^n) \text{ and is quadratic..}$$

- At least for the bivariate case, the maximum error can be achieved by piecewise quadratic functions.
- There are up to 4 such open sets for bivariate extrapolation, but this number can be as large as 20 for trivariate extrapolation.
- The sufficient condition for $\nu/2 \sum_{i=1}^{n} |\lambda_i(G)|$ is an upper bound is complicated.

Maximizing Error over Quadratic Functions

Theorem (upper bound achieved by quadratic functions)

Assume $f \in C^{1,1}_{\nu}(\mathbb{R}^n)$*. For any* $\mathbf{x} \in \mathbb{R}^n$*, if* $\mu_{ij} \geq 0$ *for all* $(i,j) \in \mathcal{I}_+ \times \mathcal{I}_-$*, then*

$$|\hat{f}(\mathbf{x}) - f(\mathbf{x})| \leq \frac{1}{2} G \cdot H^\star = \frac{\nu}{2} \sum_{i=1}^{n} |\lambda_i(G)|.$$

Computation of $\{\mu_{ij}\}$:

1.
$$Y_+ = \begin{bmatrix} \text{---}(\mathbf{x}_i - \mathbf{x})^T\text{---} \\ \vdots \\ \text{---}(\qquad)^T\text{---} \end{bmatrix}_{i \in \mathcal{I}_+} \qquad Y_- = \begin{bmatrix} \text{---}(\mathbf{x}_j - \mathbf{x})^T\text{---} \\ \vdots \\ \text{---}(\qquad)^T\text{---} \end{bmatrix}_{j \in \mathcal{I}_-}$$

$$\mathrm{diag}(\ell_+) = \begin{bmatrix} \ell_i(\mathbf{x}) & \\ & \ddots \end{bmatrix}_{i \in \mathcal{I}_+} \qquad P_- = \begin{bmatrix} \cdots & \mathbf{p}_i & \cdots \end{bmatrix}_{i:\lambda_i < 0}$$

2. $$M = \mathrm{diag}(\ell_+) Y_+ P_- (Y_- P_-)^{-1} = \begin{bmatrix} & \vdots & \\ \cdots & \mu_{ij} & \cdots \\ & \vdots & \end{bmatrix}_{i \in \mathcal{I}_+, j \in \mathcal{I}_- \setminus \{0\}} \in \mathbb{R}^{|\mathcal{I}_+| \times (|\mathcal{I}_-| - 1)}$$

3. $\mu_{i0} = \ell_i(\mathbf{x}) - \sum_{j \in \mathcal{I}_- \setminus \{0\}} \mu_{ij}$ for all $i \in \mathcal{I}_+$.

Summary of Error Analysis Results

1. An improved upper bound:

$$\hat{f}(\mathbf{x}) - f(\mathbf{x}) \leq \frac{\nu}{2} \sum_{i=0}^{n+1} |\ell_i(\mathbf{x})| \|\mathbf{x}_i - \mathbf{u}\|^2 \text{ for any } \mathbf{u} \in \mathbb{R}^n,$$

   which is sometimes tight after $\mathbf{u}$ is optimized.

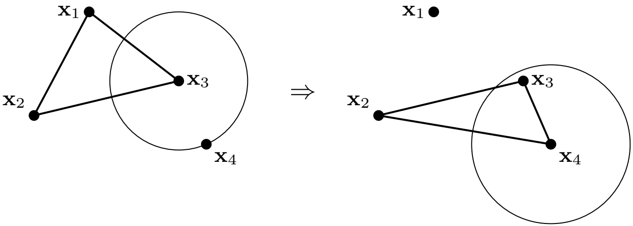
2. Error obtained by the worst quadratic function:

$$G \cdot H^\star/2 = \frac{\nu}{2} \sum_{i=1}^{n} |\lambda_i(G)|, \text{ where } G = \sum_{i=0}^{n+1} \ell_i(\mathbf{x}) \mathbf{x}_i \mathbf{x}_i^T,$$

   which is an upper error bound when $\{\mu_{ij}\}_{i \in \mathcal{I}_+, j \in \mathcal{I}_-}$ are all non-negative.

3. Piecewise quadratic functions can achieve the largest error in the remaining cases of bivariate linear interpolation. (For curiosity, not for any applications. Details not included in the talk.)

Application 1: Preventing Wasteful Evaluation in TR Methods



(a) linear interpolation + trust region method

Idea/Plan:

1. In TR DFO methods, $\hat{f}(\mathbf{x}_4)$ might be wildly inaccurate.
2. If $\texttt{error}(\mathbf{x}_4) \gg f(\mathbf{x}_3) - \hat{f}(\mathbf{x}_4)$, opt for a model step.

Results:

1. Preliminary results show some success, but occasional (depends on other parts of the algorithm and hyperparameters) and limited (up to 12% save).
2. Will not necessarily work because: bad approximation $\neq$ bad step.

Application 2: Tracking the Poisedness in TR Methods

Algorithm 0: Self-Correcting DFO-TR based on Linear Interpolation

Inputs: initial TR $B(\mathbf{c},\delta)$ and sample Θ; $\Lambda > 1$, $\eta \in (0,1)$, and $0 < \gamma_2 < 1 \leq \gamma_1$.

while *termination condition not met,* **do**

Linear interpolation: $\hat{f}(\mathbf{u}) = f(\mathbf{u})$ for all $\mathbf{u} \in \Theta$

Trust region method: Let $\mathbf{x} = \mathbf{c} - \delta/\|D\hat{f}\|D\hat{f}$ be the trial point. Compute

$$\rho = \frac{f(\mathbf{c}) - f(\mathbf{x})}{\hat{f}(\mathbf{c}) - \hat{f}(\mathbf{x})} \text{ and } \tau = \frac{1}{n}\sum_{\mathbf{u}\in\Theta} |\ell_{\mathbf{u}}(\mathbf{x})| \frac{\|\mathbf{u}-\mathbf{c}\|^2}{\delta^2}.$$

Then update the trust region as

$$(\mathbf{c},\delta) \leftarrow \begin{cases} (\mathbf{x}, \gamma_1\delta) & \text{if } \rho \geq \eta, & \text{(descent iteration)} \\ (\mathbf{c}, \delta) & \text{if } \rho < \eta \text{ and } \tau > \Lambda, & \\ & \text{or } \|D\hat{f}\| \text{ is too small,} & \text{(model improvement iteration)} \\ (\mathbf{x}, \gamma_2\delta) & \text{otherwise.} & \text{(trust region adjustment iteration)} \end{cases}$$

Sample set management: Let

$$\mathbf{r} = \arg\max_{\mathbf{u}\in\Theta} |\ell_{\mathbf{u}}(\mathbf{x})| \|\mathbf{u}-\mathbf{c}\|^2,$$

and replace $\mathbf{r}$ with $\mathbf{x}$ in Θ.

Application 2: Tracking the Poisedness in TR Methods

$$\tau = \frac{1}{n} \sum_{\mathbf{u} \in \Theta} |\ell_{\mathbf{u}}(\mathbf{x})| \frac{\|\mathbf{u}-\mathbf{c}\|^2}{\delta^2}$$

With our improved bound:

$$\hat{f}(\mathbf{x}) - f(\mathbf{x}) \leq \frac{\nu}{2}\left(|\ell_0(\mathbf{x})| \|\mathbf{x}-\mathbf{c}\|^2 + \sum_{\mathbf{u} \in \Theta} |\ell_{\mathbf{u}}(\mathbf{x})| \|\mathbf{u}-\mathbf{c}\|^2\right) = \frac{\nu}{2}(1+n\tau)\delta^2.$$

Lemma (small τ and small $\delta \Rightarrow$ descent iteration)

If $\delta \leq \frac{2(1-\eta)}{\nu(1+n\tau)} \|D\hat{f}\|$, *then* $\rho \geq \eta$.

$$\tau = \frac{1}{n} \sum_{\mathbf{u} \in \Theta} |\ell_{\mathbf{u}}(\mathbf{x})| \frac{\|\mathbf{u} - \mathbf{c}\|^2}{\delta^2}$$

With our improved bound:

$$\hat{f}(\mathbf{x}) - f(\mathbf{x}) \le \frac{\nu}{2} \left(|\ell_0(\mathbf{x})| \|\mathbf{x} - \mathbf{c}\|^2 + \sum_{\mathbf{u} \in \Theta} |\ell_{\mathbf{u}}(\mathbf{x})| \|\mathbf{u} - \mathbf{c}\|^2 \right) = \frac{\nu}{2}(1 + n\tau)\delta^2.$$

Lemma (small τ and small $\delta \Rightarrow$ descent iteration)

If $\delta \le \frac{2(1-\eta)}{\nu(1+n\tau)} \|D\hat{f}\|$, then $\rho \ge \eta$.

Lemma (model improvement iteration $\Rightarrow \psi$ decreases)

If the trust region does not change, then $\psi(\Theta, \mathbf{c}, \delta) - \psi(\Theta^+, \mathbf{c}, \delta) \ge \log \tau$.

Lemma (small $\psi \Rightarrow$ small τ)

If $\psi(\Theta, \mathbf{c}, \delta) \le \frac{1}{3} \log \Lambda$, then $\tau \le \Lambda$.

Application 3: Proving the Convergence Rate of Simplex Methods

Algorithm 1: A Baisc Simplex DFO Method

Start with a regular simplex with center $\mathbf{c}_0$ and radius δ.

for $k = 0, 1, 2, \ldots$ **do**

 Sort and label the points in Θ_k as $\{\mathbf{x}_i\}_{i=1}^{n+1}$ such that $f(\mathbf{x}_1) \leq \cdots \leq f(\mathbf{x}_{n+1})$.

 Let $\mathbf{x} = -\mathbf{x}_{n+1} + \frac{2}{n}\sum_{i=1}^{n} \mathbf{x}_i$, and evaluate $f(\mathbf{x})$.

 $\Theta_{k+1} \leftarrow \Theta_k \setminus \{\mathbf{x}_{n+1}\} \cup \{\mathbf{x}\}$.

Because

1. The simplex remains regular,
2. The size of the simplex does not change,

we always have

1. $\mu_{ij} = 1/n$ for all $i \in \mathcal{I}_+ = \{1, 2, \ldots, n\}$ and $j \in \mathcal{I}_- = \{0, n+1\}$,
2. $$G = \frac{2(n+1)}{n^2} \begin{bmatrix} -(n+1) & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{bmatrix} \quad \Rightarrow \quad \frac{\nu}{2}\sum_{i=1}^{n} |\lambda_i(G)| = \frac{2n+2}{n}\nu\delta^2$$

Application 3: Proving the Convergence Rate of Simplex Methods

Lemma (Range of the Reflection Point's Function Value)

Assume $f \in C_\nu^{1,1}(\mathbb{R}^n)$. In any iteration, the function value at the reflection point $\mathbf{x}$ is always bounded as

$$-f(\mathbf{x}_{n+1}) + \frac{2}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{2n+2}{n}\nu\delta^2 \leq f(\mathbf{x}) \leq -f(\mathbf{x}_{n+1}) + \frac{2}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{2n+2}{n}\nu\delta^2.$$

Application 3: Proving the Convergence Rate of Simplex Methods

Lemma (Range of the Reflection Point's Function Value)

Assume $f \in C_{\nu}^{1,1}(\mathbb{R}^n)$. In any iteration, the function value at the reflection point $\mathbf{x}$ is always bounded as

$$-f(\mathbf{x}_{n+1}) + \frac{2}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{2n+2}{n}\nu\delta^2 \leq f(\mathbf{x}) \leq -f(\mathbf{x}_{n+1}) + \frac{2}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{2n+2}{n}\nu\delta^2.$$

Then, let $\{\mathbf{x}_i^{(t)}\}_{i=1}^{n+1}$ and $\mathbf{x}^{(t)}$ be the simplex points and the reflection point in iteration t, respectively. We have,

$$\begin{aligned}
\sum_{\mathbf{u}\in\Theta_{k+1}} f(\mathbf{u}) &= \sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) - f(\mathbf{x}_{n+1}^{(k)}) + f(\mathbf{x}^{(k)}) \\
&\leq \sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) - f(\mathbf{x}_{n+1}) + \Big[- f(\mathbf{x}_{n+1}^{(k)}) + \frac{2}{n}\sum_{i=1}^{n} f(\mathbf{x}_i^{(k)}) + \frac{2n+2}{n}\nu\delta^2 \Big] \\
&= \sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) - \frac{2n+2}{n}\Big[f(\mathbf{x}_{n+1}^{(k)}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i^{(k)}) \Big] + \frac{2n+2}{n}\nu\delta^2.
\end{aligned}$$

After telescoping, we have

$$\sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) \leq \sum_{\mathbf{u}\in\Theta_0} f(\mathbf{u}) - \frac{2n+2}{n}\sum_{t=0}^{k-1}\left[f(\mathbf{x}_{n+1}^{(t)}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i^{(t)})\right] + k\frac{2n+2}{n}\nu\delta^2.$$

Use the fact that $\sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) \geq (n+1)f^\star$ and rearrange the terms to get

$$\frac{1}{k}\sum_{t=0}^{k-1}\left[f(\mathbf{x}_{n+1}^{(t)}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i^{(t)})\right] \leq \frac{n}{2k}\cdot\left[\frac{1}{n+1}\sum_{\mathbf{u}\in\Theta_0} f(\mathbf{u}) - f^\star\right] + \nu\delta^2.$$

Application 3: Proving the Convergence Rate of Simplex Methods

After telescoping, we have

$$\sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) \le \sum_{\mathbf{u}\in\Theta_0} f(\mathbf{u}) - \frac{2n+2}{n}\sum_{t=0}^{k-1}\left[f(\mathbf{x}_{n+1}^{(t)}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i^{(t)})\right] + k\frac{2n+2}{n}\nu\delta^2.$$

Use the fact that $\sum_{\mathbf{u}\in\Theta_k} f(\mathbf{u}) \ge (n+1)f^\star$ and rearrange the terms to get

$$\frac{1}{k}\sum_{t=0}^{k-1}\left[f(\mathbf{x}_{n+1}^{(t)}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i^{(t)})\right] \le \frac{n}{2k}\cdot\left[\frac{1}{n+1}\sum_{\mathbf{u}\in\Theta_0} f(\mathbf{u}) - f^\star\right] + \nu\delta^2.$$

Lemma (Low Function Value Difference ⇒ Small Model Gradient)

Assume $f \in C_\nu^{1,1}(\mathbb{R}^n)$. For any iteration k, let $\hat{f}$ be the linear function that interpolates f on Θ_k, and $\mathbf{c}_k$ the centroid of Θ_k. Then

$$\|D\hat{f}(\mathbf{c}_k)\| \le \frac{n}{\delta}\left[f(\mathbf{x}_{n+1}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i)\right].$$

Lemma (Model Gradient vs True Gradient)

Assume $f \in C_\nu^{1,1}(\mathbb{R}^n)$. For any iteration k, let $\hat{f}$ be the linear function that interpolates f on Θ_k, and $\mathbf{c}_k$ the centroid of Θ_k. Then

$$\|Df(\mathbf{c}_k) - D\hat{f}(\mathbf{c}_k)\|^2 \le \frac{n}{4}\nu^2\delta^2.$$

Application 3: Proving the Convergence Rate of Simplex Methods

Theorem (Convergence Rate with an Arbitrary δ)

Assume $f \in C^{1,1}_{\nu}(\mathbb{R}^n)$ and $f(\mathbf{u}) \geq f^\star$ for all $\mathbf{u} \in \mathbb{R}^n$. Let $\mathbf{c}_k$ be the centroid of Θ_k for each iteration $k = 0, 1, \ldots$. We have for any $k \geq 1$

$$\frac{1}{k}\sum_{t=0}^{k-1}\|Df(\mathbf{c}_t)\| \leq \frac{n^2}{2\delta k}\cdot\left[\frac{1}{n+1}\sum_{\mathbf{u}\in\Theta_0} f(\mathbf{u}) - f^\star\right] + \left(n + \frac{\sqrt{n}}{2}\right)\nu\delta.$$

If the Lipschitz constant ν is known, we can select the size of the simplex and a stopping criterion to obtain a solution of desired accuracy.

Theorem (Complexity for an ϵ-Stationary Solution)

Assume $f \in C^{1,1}_{\nu}(\mathbb{R}^n)$ and $f(\mathbf{u}) \geq f^\star$ for all $\mathbf{u} \in \mathbb{R}^n$. Given a desired accuracy $\epsilon > 0$, if $\delta = \frac{2\epsilon}{5n\nu}$ and the loop breaks after $\left[f(\mathbf{x}_{n+1}) - \frac{1}{n+1}\sum_{i=1}^{n+1} f(\mathbf{x}_i)\right] \leq 2\nu\delta^2$ is detected before the reflection step in some iteration k, then the algorithm would terminate in at most

$$\frac{25n^3\nu}{8\epsilon^2}\left[\frac{1}{n+1}\sum_{\mathbf{u}\in\Theta_0} f(\mathbf{u}) - f^\star\right]$$

iterations with $\|Df(\mathbf{c}_k)\| \leq \epsilon$.

Thank you! Grazie!