Convex Optimization

(EE227A: UC Berkeley)

Lecture 27
(Derivative free optimization)

30 Apr, 2013

Suvrit Sra

Introduction

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

Introduction

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

Optimizing without derivatives

Introduction

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

Optimizing without derivatives

(CD): $x_j^{k+1} \leftarrow \operatorname{argmin}_{x_j} f(\ldots, x_j, \ldots)$

Introduction

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

Optimizing without derivatives

(CD): $x_j^{k+1} \leftarrow \operatorname{argmin}_{x_j} f(\ldots, x_j, \ldots)$

- Requires **subroutine** to solve for each coordinate, or
- explicit access to f, or
- ability to restrict computation to jth coordinate

Introduction

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

Optimizing without derivatives

(CD): $x_j^{k+1} \leftarrow \operatorname{argmin}_{x_j} f(\ldots, x_j, \ldots)$

- Requires **subroutine** to solve for each coordinate, or
- explicit access to f, or
- ability to restrict computation to jth coordinate

Sometimes may not be possible / practical!

Optimizing without derivatives

Why care?

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...
- Burden of mathematical modelling

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...
- Burden of mathematical modelling
- Programmer time vs computer time

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...
- Burden of mathematical modelling
- Programmer time vs computer time
- Extra storage needed by *Fast Differentiation*

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...
- Burden of mathematical modelling
- Programmer time vs computer time
- Extra storage needed by *Fast Differentiation*
- Dealing with nonsmooth, nonconvex functions

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...
- Burden of mathematical modelling
- Programmer time vs computer time
- Extra storage needed by *Fast Differentiation*
- Dealing with nonsmooth, nonconvex functions
- Ease of use, laziness?

Optimizing without derivatives

Why care?

- Legacy code, access to executables only, ...
- Burden of mathematical modelling
- Programmer time vs computer time
- Extra storage needed by *Fast Differentiation*
- Dealing with nonsmooth, nonconvex functions
- Ease of use, laziness?

Derivative free optimization (DFO)

DFO

WARNING!

If you can somehow obtain derivatives, use them. Turn to DFO if derivatives too expensive or impossible to get!

DFO

WARNING!

If you can somehow obtain derivatives, use them. Turn to DFO if derivatives too expensive or impossible to get!

Not discussing today

♣ Automatic differentiation (http://www.autodiff.org)

DFO

WARNING!

If you can somehow obtain derivatives, use them. Turn to DFO if derivatives too expensive or impossible to get!

Not discussing today

- ♣ Automatic differentiation (http://www.autodiff.org)
- ♣ Fast Differentiation — cost(∇f) ≤ 4cost(f).

DFO

WARNING!

If you can somehow obtain derivatives, use them. Turn to DFO if derivatives too expensive or impossible to get!

Not discussing today

- ♣ Automatic differentiation (http://www.autodiff.org)
- ♣ Fast Differentiation — cost$(\nabla f) \leq 4$cost(f).
- ♣ Various finite differencing techniques

DFO

WARNING!

If you can somehow obtain derivatives, use them. Turn to DFO if derivatives too expensive or impossible to get!

Not discussing today

- ♣ Automatic differentiation (http://www.autodiff.org)
- ♣ Fast Differentiation — cost$(\nabla f) \leq 4$cost(f).
- ♣ Various finite differencing techniques
- ♣ Nonconvex DFO
- ♣ See recent book: "*Introduction to Derivative-Free Optimization*" by A. Conn, K. Scheinberg, and L. N. Vicente (MPS-SIAM, 2009).

DFO – brute force

$$\min \quad f(x)$$

DFO – brute force

$$\min \quad f(x)$$

Brute force method

- Start at $x_0 \in \mathbb{R}^n$

DFO – brute force

$$\min \quad f(x)$$

Brute force method

- Start at $x_0 \in \mathbb{R}^n$
- At iteration $k \geq 0$:
  - ✈ Sample a point y from $\mathcal{N}(x_k, \Sigma_k)$

DFO – brute force

$$\min \quad f(x)$$

Brute force method

- Start at $x_0 \in \mathbb{R}^n$
- At iteration $k \geq 0$:
  - ✈ Sample a point y from $\mathcal{N}(x_k, \Sigma_k)$
  - ✈ **If** $f(y) < f(x_k)$, **then** $x_{k+1} \leftarrow y$

DFO – brute force

$$\min \quad f(x)$$

Brute force method

- Start at $x_0 \in \mathbb{R}^n$
- At iteration $k \geq 0$:
  - ✈ Sample a point y from $\mathcal{N}(x_k, \Sigma_k)$
  - ✈ **If** $f(y) < f(x_k)$, **then** $x_{k+1} \leftarrow y$
  - ✈ **otherwise** $x_{k+1} \leftarrow x_k$

DFO – brute force

$$\min \quad f(x)$$

Brute force method

- Start at $x_0 \in \mathbb{R}^n$
- At iteration $k \geq 0$:
  - ✈ Sample a point y from $\mathcal{N}(x_k, \Sigma_k)$
  - ✈ **If** $f(y) < f(x_k)$, **then** $x_{k+1} \leftarrow y$
  - ✈ **otherwise** $x_{k+1} \leftarrow x_k$
- repeat above procedure until tired

DFO – brute force

$$\min \quad f(x)$$

Brute force method

- Start at $x_0 \in \mathbb{R}^n$
- At iteration $k \geq 0$:
  - ✈ Sample a point y from $\mathcal{N}(x_k, \Sigma_k)$
  - ✈ **If** $f(y) < f(x_k)$, **then** $x_{k+1} \leftarrow y$
  - ✈ **otherwise** $x_{k+1} \leftarrow x_k$
- repeat above procedure until tired

Nothing but completely random search!

Can be done more cleverly: see e.g. probabilistic optimization

DFO – simulating gradients

- At iteration k pick $u \in \mathbb{S}^{n-1}$ at random

DFO – simulating gradients

- At iteration k pick $u \in \mathbb{S}^{n-1}$ at random
- Update the guess as

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} \right] u$$

Scheme might "work" as $\mu_k \to 0$; it becomes

DFO – simulating gradients

- At iteration k pick $u \in \mathbb{S}^{n-1}$ at random
- Update the guess as

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} \right] u$$

Scheme might "work" as $\mu_k \to 0$; it becomes

$$x_{k+1} = x_k - h_k \underbrace{f'(x_k; u)}_{\text{directional deriv}} u,$$

(notice that if f is differentiable, then $f'(x; u) = \langle \nabla f(x), u \rangle$)

DFO – simulating gradients

- At iteration k pick $u \in \mathbb{S}^{n-1}$ at random
- Update the guess as

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} \right] u$$

Scheme might "work" as $\mu_k \to 0$; it becomes

$$x_{k+1} = x_k - h_k \underbrace{f'(x_k; u)}_{\text{directional deriv}} u,$$

(notice that if f is differentiable, then $f'(x; u) = \langle \nabla f(x), u \rangle$)

- ▶ Directional derivatives much simpler than gradient

DFO – simulating gradients

- At iteration k pick $u \in \mathbb{S}^{n-1}$ at random
- Update the guess as

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} \right] u$$

Scheme might "work" as $\mu_k \to 0$; it becomes

$$x_{k+1} = x_k - h_k \underbrace{f'(x_k; u)}_{\text{directional deriv}} u,$$

(notice that if f is differentiable, then $f'(x; u) = \langle \nabla f(x), u \rangle$)

- ▶ Directional derivatives much simpler than gradient
- ▶ Can be reasonably approximated by finite differences

DFO – simulating gradients

- At iteration k pick $u \in \mathbb{S}^{n-1}$ at random
- Update the guess as

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} \right] u$$

Scheme might "work" as $\mu_k \to 0$; it becomes

$$x_{k+1} = x_k - h_k \underbrace{f'(x_k; u)}_{\text{directional deriv}} u,$$

(notice that if f is differentiable, then $f'(x; u) = \langle \nabla f(x), u \rangle$)

- Directional derivatives much simpler than gradient
- Can be reasonably approximated by finite differences
- Even for nonconvex functions

DFO via randomization

Def. (Smoothing). Let $\mu > 0$, and $u \sim P$ with density p, then

$$f_\mu(x) := \int f(x + \mu u)p(u)du.$$

DFO via randomization

Def. (Smoothing). Let $\mu > 0$, and $u \sim P$ with density p, then

$$f_\mu(x) := \int f(x + \mu u) p(u) du.$$

Main ideas today:

♠ For deterministic $f(x)$,

$$x_{k+1} = x_k - h_k f'(x_k; u) u,$$

at worst $O(n)$ slower than usual subgradient method

DFO via randomization

Def. (Smoothing). Let $\mu > 0$, and $u \sim P$ with density p, then

$$f_\mu(x) := \int f(x + \mu u)p(u)du.$$

Main ideas today:

- For deterministic $f(x)$,

$$x_{k+1} = x_k - h_k f'(x_k; u)u,$$

  at worst $O(n)$ slower than usual subgradient method

- Finite-differencing version

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k}\right] u,$$

  at worst $O(n^2)$ slower.

DFO via randomization

Def. (Smoothing). Let $\mu > 0$, and $u \sim P$ with density p, then

$$f_\mu(x) := \int f(x + \mu u) p(u) du.$$

Main ideas today:

- For deterministic $f(x)$,

$$x_{k+1} = x_k - h_k f'(x_k; u) u,$$

  at worst $O(n)$ slower than usual subgradient method

- Finite-differencing version

$$x_{k+1} = x_k - h_k \left[\frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} \right] u,$$

  at worst $O(n^2)$ slower.

- For **stochastic optimization**, i.e., $f(x) = E_z[F(x, z)]$, both iterations above extend naturally.

DFO – setup

- ☞ We'll work in some Euclidean space E; let its dual be E^*
- ☞ (If E is column-vectors in $\mathbb{R}^n$, then E^* are row vectors in $\mathbb{R}^n$)
- ☞ Let $B = B^* \succ 0$ be a linear operator from $E^* \to E$

DFO – setup

- ☞ We'll work in some Euclidean space E; let its dual be E^*
- ☞ (If E is column-vectors in $\mathbb{R}^n$, then E^* are row vectors in $\mathbb{R}^n$)
- ☞ Let $B = B^* \succ 0$ be a linear operator from $E^* \to E$

We'll use the following pair of norms (dual to each other)

$$\|x\| = \langle Bx, x\rangle^{1/2}, \quad x \in E,$$
$$\|g\|_* = \langle g, B^{-1}g\rangle^{1/2}, \quad g \in E^*.$$

DFO – setup

- ☞ We'll work in some Euclidean space E; let its dual be E^*
- ☞ (If E is column-vectors in $\mathbb{R}^n$, then E^* are row vectors in $\mathbb{R}^n$)
- ☞ Let $B = B^* \succ 0$ be a linear operator from $E^* \to E$

We'll use the following pair of norms (dual to each other)

$$\|x\| = \langle Bx,\, x\rangle^{1/2}, \quad x \in E,$$
$$\|g\|_* = \langle g,\, B^{-1}g\rangle^{1/2}, \quad g \in E^*.$$

Function classes

- ▶ $f \in C^0_{L_0}(E)$: $|f(x) - f(y)| \leq L_0(f)\|x - y\|$, $x, y \in E$

DFO – setup

- ☞ We'll work in some Euclidean space E; let its dual be E^*
- ☞ (If E is column-vectors in $\mathbb{R}^n$, then E^* are row vectors in $\mathbb{R}^n$)
- ☞ Let $B = B^* \succ 0$ be a linear operator from $E^* \to E$

We'll use the following pair of norms (dual to each other)

$$\|x\| = \langle Bx, x \rangle^{1/2}, \quad x \in E,$$
$$\|g\|_* = \langle g, B^{-1}g \rangle^{1/2}, \quad g \in E^*.$$

Function classes

- $f \in C^0_{L_0}(E)$: $|f(x) - f(y)| \leq L_0(f)\|x - y\|$, $x, y \in E$
- $f \in C^1_{L_1}(E)$: $\|\nabla f(x) - \nabla f(y)\|_* \leq L_1(f)\|x - y\|$, $x, y \in E$
  Equivalently:
  $|f(y) - f(y) - \langle \nabla f(x), y - x \rangle| \leq \frac{1}{2}L_1(f)\|x - y\|^2$

DFO – Gaussian smoothing

Assumption: Let $f : E \to \mathbb{R}$. Assume at each $x \in E$, directional derivative of f exists in every direction.

DFO – Gaussian smoothing

Assumption: Let $f : E \to \mathbb{R}$. Assume at each $x \in E$, directional derivative of f exists in every direction.

Def. (Gaussian approximation.) Let $\mu \geq 0$, we define

$$f_\mu(x) := \frac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du.$$

DFO – Gaussian smoothing

Assumption: Let $f : E \to \mathbb{R}$. Assume at each $x \in E$, directional derivative of f exists in every direction.

Def. (Gaussian approximation.) Let $\mu \geq 0$, we define

$$f_\mu(x) := \frac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du.$$

Notes:

- Remember, we are using: $\|u\|^2 = \langle Bu, u \rangle$
- κ is the normalization constant $\kappa := \int_E e^{-\frac{1}{2}\|u\|^2} du$

DFO – Gaussian smoothing

Assumption: Let $f : E \to \mathbb{R}$. Assume at each $x \in E$, directional derivative of f exists in every direction.

Def. (Gaussian approximation.) Let $\mu \geq 0$, we define

$$f_\mu(x) := \frac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du.$$

Notes:

- Remember, we are using: $\|u\|^2 = \langle Bu, u \rangle$
- κ is the normalization constant $\kappa := \int_E e^{-\frac{1}{2}\|u\|^2} du$

Key point: Smoothed function f_μ nicer than $f(x)$

Basic properties of f_μ

☞ If f is convex, then f_μ is also convex (nonneg weighted sum)

Basic properties of f_μ

☞ If f is convex, then f_μ is also convex (nonneg weighted sum)

☞ $f(x) \leq f_\mu(x)$.

Basic properties of f_μ

- If f is convex, then f_μ is also convex (nonneg weighted sum)
- $f(x) \leq f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

Basic properties of f_μ

☞ If f is convex, then f_μ is also convex (nonneg weighted sum)

☞ $f(x) \le f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$f_\mu(x) \;=\; \frac{1}{\kappa}\int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du$$

Basic properties of f_μ

☞ If f is convex, then f_μ is also convex (nonneg weighted sum)

☞ $f(x) \le f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \frac{1}{\kappa}\int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &\ge \frac{1}{\kappa}\int_E \left[f(x) + \mu\langle g,\, u\rangle\right] e^{-\frac{1}{2}\|u\|^2} du \end{aligned}$$

Basic properties of f_μ

- If f is convex, then f_μ is also convex (nonneg weighted sum)
- $f(x) \leq f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \tfrac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &\geq \tfrac{1}{\kappa} \int_E \left[f(x) + \mu \langle g, u \rangle \right] e^{-\frac{1}{2}\|u\|^2} du \\ &= f(x), \end{aligned}$$

  last line follows as $\frac{1}{\kappa} \int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (mean-zero Gaussian)

Basic properties of f_μ

- ☞ If f is convex, then f_μ is also convex (nonneg weighted sum)
- ☞ $f(x) \leq f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \tfrac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &\geq \tfrac{1}{\kappa} \int_E \left[f(x) + \mu \langle g, u \rangle \right] e^{-\frac{1}{2}\|u\|^2} du \\ &= f(x), \end{aligned}$$

  last line follows as $\frac{1}{\kappa} \int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (mean-zero Gaussian)
- ☞ If $f \in C^0_{L_0}$, then $f_\mu \in C^0_{L_0}$ with $L_0(f_\mu) \leq L_0(f)$.

Basic properties of f_μ

☞ If f is convex, then f_μ is also convex (nonneg weighted sum)

☞ $f(x) \leq f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \tfrac{1}{\kappa} \int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &\geq \tfrac{1}{\kappa} \int_E \left[f(x) + \mu \langle g,\, u \rangle \right] e^{-\frac{1}{2}\|u\|^2} du \\ &= f(x), \end{aligned}$$

last line follows as $\frac{1}{\kappa} \int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (mean-zero Gaussian)

☞ If $f \in C^0_{L_0}$, then $f_\mu \in C^0_{L_0}$ with $L_0(f_\mu) \leq L_0(f)$. *Proof:*

$$|f_\mu(x) - f_\mu(y)| \quad \leq \quad \tfrac{1}{\kappa} \int_E |f(x+\mu u) - f(y+\mu u)| e^{-\frac{1}{2}\|u\|^2} du$$

Basic properties of f_μ

- If f is convex, then f_μ is also convex (nonneg weighted sum)
- $f(x) \le f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \frac{1}{\kappa}\int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &\ge \frac{1}{\kappa}\int_E \left[f(x) + \mu\langle g,\, u\rangle\right] e^{-\frac{1}{2}\|u\|^2} du \\ &= f(x), \end{aligned}$$

  last line follows as $\frac{1}{\kappa}\int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (mean-zero Gaussian)
- If $f \in C^0_{L_0}$, then $f_\mu \in C^0_{L_0}$ with $L_0(f_\mu) \le L_0(f)$. *Proof:*

$$\begin{aligned} |f_\mu(x) - f_\mu(y)| &\le \frac{1}{\kappa}\int_E |f(x+\mu u) - f(y+\mu u)| e^{-\frac{1}{2}\|u\|^2} du \\ &\le L_0(f)\|x-y\| \frac{1}{\kappa}\int_E e^{-\frac{1}{2}\|u\|^2} du \end{aligned}$$

Basic properties of f_μ

- If f is convex, then f_μ is also convex (nonneg weighted sum)
- $f(x) \le f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \tfrac{1}{\kappa}\int_E f(x+\mu u)e^{-\frac{1}{2}\|u\|^2}du \\ &\ge \tfrac{1}{\kappa}\int_E \left[f(x)+\mu\langle g,\, u\rangle\right] e^{-\frac{1}{2}\|u\|^2}du \\ &= f(x), \end{aligned}$$

  last line follows as $\frac{1}{\kappa}\int_E ue^{-\frac{1}{2}\|u\|^2}du = 0$ (mean-zero Gaussian)
- If $f \in C^0_{L_0}$, then $f_\mu \in C^0_{L_0}$ with $L_0(f_\mu) \le L_0(f)$. *Proof:*

$$\begin{aligned} |f_\mu(x) - f_\mu(y)| &\le \tfrac{1}{\kappa}\int_E |f(x+\mu u) - f(y+\mu u)|e^{-\frac{1}{2}\|u\|^2}du \\ &\le L_0(f)\|x-y\|\tfrac{1}{\kappa}\int_E e^{-\frac{1}{2}\|u\|^2}du \\ &= L_0(f)\|x-y\|. \end{aligned}$$

Basic properties of f_μ

- If f is convex, then f_μ is also convex (nonneg weighted sum)
- $f(x) \leq f_\mu(x)$. *Proof:* Let $g \in \partial f(x)$, then

$$\begin{aligned} f_\mu(x) &= \frac{1}{\kappa}\int_E f(x+\mu u)e^{-\frac{1}{2}\|u\|^2}du \\ &\geq \frac{1}{\kappa}\int_E \left[f(x)+\mu\langle g,\, u\rangle\right] e^{-\frac{1}{2}\|u\|^2}du \\ &= f(x), \end{aligned}$$

  last line follows as $\frac{1}{\kappa}\int_E ue^{-\frac{1}{2}\|u\|^2}du = 0$ (mean-zero Gaussian)
- If $f \in C^0_{L_0}$, then $f_\mu \in C^0_{L_0}$ with $L_0(f_\mu) \leq L_0(f)$. *Proof:*

$$\begin{aligned} |f_\mu(x)-f_\mu(y)| &\leq \frac{1}{\kappa}\int_E |f(x+\mu u)-f(y+\mu u)|e^{-\frac{1}{2}\|u\|^2}du \\ &\leq L_0(f)\|x-y\|\frac{1}{\kappa}\int_E e^{-\frac{1}{2}\|u\|^2}du \\ &= L_0(f)\|x-y\|. \end{aligned}$$

- Similarly, prove that

$$\|\nabla f_\mu(x)-\nabla f_\mu(y)\|_* \leq L_1(f)\|x-y\|, \quad x,y\in E.$$

Bounding moments

We saw: $f(x) \leq f_\mu(x)$. What about $f_\mu(x) \leq f(x) + ?$

Bounding moments

We saw: $f(x) \le f_\mu(x)$. What about $f_\mu(x) \le f(x) + ?$

Seek bounds on

$$\theta(p) := \frac{1}{\kappa} \int_E \|u\|^p e^{-\frac{1}{2}\|u\|^2} du.$$

Bounding moments

We saw: $f(x) \leq f_\mu(x)$. What about $f_\mu(x) \leq f(x) + ?$

Seek bounds on

$$\theta(p) := \frac{1}{\kappa} \int_E \|u\|^p e^{-\frac{1}{2}\|u\|^2} du.$$

Lemma Let $p \geq 0$. The function $\log \theta(p)$ is convex.

Proof: Simple exercise.

Bounding moments

We saw: $f(x) \leq f_\mu(x)$. What about $f_\mu(x) \leq f(x)+?$

Seek bounds on

$$\theta(p) := \frac{1}{\kappa}\int_E \|u\|^p e^{-\frac{1}{2}\|u\|^2} du.$$

Lemma Let $p \geq 0$. The function $\log\theta(p)$ is convex.

Proof: Simple exercise.
Two easy cases: $p = 0$ and $p = 2$

$$\begin{aligned} p = 0, \qquad & \theta(0) = \frac{1}{\kappa}\int_E e^{-\frac{1}{2}\|u\|^2} du = 1 \\ p = 2, \qquad & \theta(2) = \frac{1}{\kappa}\int_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du = n. \end{aligned}$$

Proof: $\log\int e^{-\frac{1}{2}\|u\|^2} du = \log\int e^{-\frac{1}{2}\langle Bu, u\rangle} du = \frac{1}{2}(n\log(2\pi) - \log\det(B))$.
Differentiate both sides wrt B to obtain, $\frac{1}{\kappa}\int_E uu^* e^{-\frac{1}{2}\|u\|^2} du = B^{-1}$.
Now multiply by B and take trace.

Bounding moments

Lemma For $p \in [0, 2]$, we have

$$\theta(p) \leq n^{p/2}.$$

For $p \geq 2$ we have two-sided bounds

$$n^{p/2} \leq \theta(p) \leq (p + n)^{p/2}.$$

Bounding moments

Lemma For $p \in [0, 2]$, we have

$$\theta(p) \leq n^{p/2}.$$

For $p \geq 2$ we have two-sided bounds

$$n^{p/2} \leq \theta(p) \leq (p + n)^{p/2}.$$

Proof:

- Say, $p \in [0, 2]$. Since $\log \theta(p)$ is convex, write $p = (1 - \alpha) \cdot 0 + \alpha \cdot 2$

Bounding moments

Lemma For $p \in [0,2]$, we have

$$\theta(p) \leq n^{p/2}.$$

For $p \geq 2$ we have two-sided bounds

$$n^{p/2} \leq \theta(p) \leq (p+n)^{p/2}.$$

Proof:

- Say, $p \in [0,2]$. Since $\log \theta(p)$ is convex, write $p = (1-\alpha) \cdot 0 + \alpha \cdot 2$
- Thus, $\log \theta(p) \leq (1-\alpha) \log \theta(0) + \alpha \log \theta(2)$

Bounding moments

Lemma For $p \in [0,2]$, we have

$$\theta(p) \leq n^{p/2}.$$

For $p \geq 2$ we have two-sided bounds

$$n^{p/2} \leq \theta(p) \leq (p+n)^{p/2}.$$

Proof:

- Say, $p \in [0,2]$. Since $\log \theta(p)$ is convex, write $p = (1-\alpha) \cdot 0 + \alpha \cdot 2$
- Thus, $\log \theta(p) \leq (1-\alpha) \log \theta(0) + \alpha \log \theta(2)$
- So we get: $\log \theta(p) \leq \frac{p}{2} \log n$

Bounding moments

Lemma For $p \in [0, 2]$, we have

$$\theta(p) \leq n^{p/2}.$$

For $p \geq 2$ we have two-sided bounds

$$n^{p/2} \leq \theta(p) \leq (p+n)^{p/2}.$$

Proof:

- Say, $p \in [0, 2]$. Since $\log \theta(p)$ is convex, write $p = (1-\alpha) \cdot 0 + \alpha \cdot 2$
- Thus, $\log \theta(p) \leq (1-\alpha) \log \theta(0) + \alpha \log \theta(2)$
- So we get: $\log \theta(p) \leq \frac{p}{2} \log n$
- The other case, $p \geq 2$ requires some more work.

Lipschitz properties of f_μ

Theorem A. If $f \in C^0_{L_0}$ then

$$|f_\mu(x) - f(x)| \leq \mu L_0(f)\sqrt{n}, \quad x \in E$$

Lipschitz properties of f_μ

Theorem A. If $f \in C^0_{L_0}$ then

$$|f_\mu(x) - f(x)| \le \mu L_0(f)\sqrt{n}, \quad x \in E$$

Proof: We have $f_\mu(x) - f(x) = \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du$

Lipschitz properties of f_μ

Theorem A. If $f \in C^0_{L_0}$ then

$$|f_\mu(x) - f(x)| \leq \mu L_0(f)\sqrt{n}, \quad x \in E$$

Proof: We have $f_\mu(x) - f(x) = \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du$

$$|f_\mu(x) - f(x)| \quad \leq \quad \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du$$

Lipschitz properties of f_μ

Theorem A. If $f \in C^0_{L_0}$ then

$$|f_\mu(x) - f(x)| \leq \mu L_0(f)\sqrt{n}, \quad x \in E$$

Proof: We have $f_\mu(x) - f(x) = \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du$

$$\begin{array}{rcl} |f_\mu(x) - f(x)| & \leq & \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du \\ & \leq & \frac{\mu L_0(f)}{\kappa}\int_E \|u\| e^{-\frac{1}{2}\|u\|^2} du \end{array}$$

Lipschitz properties of f_μ

Theorem A. If $f \in C^0_{L_0}$ then

$$|f_\mu(x) - f(x)| \leq \mu L_0(f)\sqrt{n}, \quad x \in E$$

Proof: We have $f_\mu(x) - f(x) = \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du$

$$\begin{aligned} |f_\mu(x) - f(x)| &\leq \frac{1}{\kappa}\int_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2} du \\ &\leq \frac{\mu L_0(f)}{\kappa}\int_E \|u\| e^{-\frac{1}{2}\|u\|^2} du \\ &\leq \mu L_0(f)\sqrt{n}. \end{aligned}$$

Lipschitz properties of f_μ

Theorem B. If $f \in C^1_{L_1}$ then

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2}{2} L_1(f) n, \quad x \in E.$$

Lipschitz properties of f_μ

Theorem B. If $f \in C^1_{L_1}$ then

$$|f_\mu(x) - f(x)| \le \frac{\mu^2}{2} L_1(f) n, \quad x \in E.$$

Proof: If $f \in C^1_{L_1}$, then

$$f_\mu(x) - f(x) \;=\; \frac{1}{\kappa} \int_E [f(x+\mu u) - f(x) - \mu \langle \nabla f(x), u \rangle] e^{-\frac{1}{2}\|u\|^2} du$$

Lipschitz properties of f_μ

Theorem B. If $f \in C^1_{L_1}$ then

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2}{2} L_1(f) n, \quad x \in E.$$

Proof: If $f \in C^1_{L_1}$, then

$$\begin{aligned} f_\mu(x) - f(x) &= \frac{1}{\kappa} \int_E [f(x+\mu u) - f(x) - \mu \langle \nabla f(x), u \rangle] e^{-\frac{1}{2}\|u\|^2} du \\ |f_\mu(x) - f(x)| &\leq \frac{\mu^2 L_1(f)}{2\kappa} \int_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du \end{aligned}$$

Lipschitz properties of f_μ

Theorem B. If $f \in C^1_{L_1}$ then

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2}{2} L_1(f) n, \quad x \in E.$$

Proof: If $f \in C^1_{L_1}$, then

$$\begin{aligned} f_\mu(x) - f(x) &= \frac{1}{\kappa} \int_E [f(x+\mu u) - f(x) - \mu \langle \nabla f(x), u \rangle] e^{-\frac{1}{2}\|u\|^2} du \\ |f_\mu(x) - f(x)| &\leq \frac{\mu^2 L_1(f)}{2\kappa} \int_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du \\ &= \frac{\mu^2 L_1(f)}{2} n. \end{aligned}$$

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

First, let's get the gradient

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

First, let's get the gradient

$$f_\mu(x) = \frac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du,$$

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

First, let's get the gradient

$$\begin{aligned} f_\mu(x) &= \tfrac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du, \\ f_\mu(x) &= \tfrac{1}{\kappa \mu^n} \int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} dy, \quad (y = x + (\mu I) u) \end{aligned}$$

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

First, let's get the gradient

$$\begin{aligned}
f_\mu(x) &= \tfrac{1}{\kappa} \textstyle\int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du, \\
f_\mu(x) &= \tfrac{1}{\kappa\mu^n} \textstyle\int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} dy, \quad (y = x + (\mu I)u) \\
\nabla f_\mu(x) &= \tfrac{1}{\mu^n \kappa} \textstyle\int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} \tfrac{1}{\mu^2} B(y-x) dy
\end{aligned}$$

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

First, let's get the gradient

$$\begin{aligned}
f_\mu(x) &= \tfrac{1}{\kappa} \int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du, \\
f_\mu(x) &= \tfrac{1}{\kappa \mu^n} \int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} dy, \quad (y = x + (\mu I)u) \\
\nabla f_\mu(x) &= \tfrac{1}{\mu^n \kappa} \int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} \tfrac{1}{\mu^2} B(y-x) dy \\
&= \tfrac{1}{\mu \kappa} \int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} Bu \, du
\end{aligned}$$

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$.

- This lemma justifies the name "smoothing"

Proof: We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

First, let's get the gradient

$$\begin{aligned}
f_\mu(x) &= \tfrac{1}{\kappa} \int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du, \\
f_\mu(x) &= \tfrac{1}{\kappa \mu^n} \int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} dy, \quad (y = x + (\mu I)u) \\
\nabla f_\mu(x) &= \tfrac{1}{\mu^n \kappa} \int_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} \tfrac{1}{\mu^2} B(y-x) dy \\
&= \tfrac{1}{\mu \kappa} \int_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} Bu\, du \\
&= \tfrac{1}{\kappa} \int_E \tfrac{f(x+\mu u) - f(x)}{\mu} e^{-\frac{1}{2}\|u\|^2} Bu\, du.
\end{aligned}$$

Lipschitz constant of ∇f_μ

We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

Lipschitz constant of ∇f_μ

We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

Now, let's get $L_1(f_\mu)$ (write $dP(u) = e^{-\frac{1}{2}\|u\|^2} Bu\, du$):

Lipschitz constant of ∇f_μ

We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

Now, let's get $L_1(f_\mu)$ (write $dP(u) = e^{-\frac{1}{2}\|u\|^2} Bu\, du$):

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* = \tfrac{1}{\kappa} \int_E [\tfrac{f(x+\mu u) - f(x) + f(y) - f(y+\mu u)}{\mu}] dP(u)$$

Lipschitz constant of ∇f_μ

We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

Now, let's get $L_1(f_\mu)$ (write $dP(u) = e^{-\frac{1}{2}\|u\|^2} Bu\, du$):

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* = \tfrac{1}{\kappa} \int_E [\tfrac{f(x+\mu u) - f(x) + f(y) - f(y+\mu u)}{\mu}] dP(u)$$

$$\leq \tfrac{1}{\mu\kappa} \int_E |f(x+\mu u) - f(x) + f(y) - f(y+\mu u)| \|Bu\|_2 e^{-\frac{1}{2}\|u\|^2} du$$

Lipschitz constant of ∇f_μ

We show that $f_\mu \in C^1_{L_1}$ with

$$L_1(f_\mu) = \frac{2\sqrt{n}}{\mu} L_0(f).$$

Now, let's get $L_1(f_\mu)$ (write $dP(u) = e^{-\frac{1}{2}\|u\|^2} Bu\, du$):

$$\begin{aligned}
& \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* = \tfrac{1}{\kappa} \int_E \big[\tfrac{f(x+\mu u) - f(x) + f(y) - f(y+\mu u)}{\mu}\big] dP(u) \\
\leq \; & \tfrac{1}{\mu\kappa} \int_E |f(x+\mu u) - f(x) + f(y) - f(y+\mu u)| \|Bu\|_2 e^{-\frac{1}{2}\|u\|^2} du \\
\leq \; & \tfrac{2L_0(f)}{\kappa\mu} \int_E \|u\| e^{-\frac{1}{2}\|u\|^2} du \\
\leq \; & \tfrac{2L_0(f)}{\mu} \sqrt{n}.
\end{aligned}$$

DFO gradient oracles

Let $u \sim \mathcal{N}(0, B^{-1})$. For $\mu \geq 0$, we define **gradient-free oracles**

☞ Sample $u \in E$ and return $g_\mu(x) = \left[\frac{f(x+\mu u) - f(x)}{\mu}\right] Bu$

DFO gradient oracles

Let $u \sim \mathcal{N}(0, B^{-1})$. For $\mu \geq 0$, we define **gradient-free oracles**

- ☞ Sample $u \in E$ and return $g_\mu(x) = \left[\frac{f(x+\mu u)-f(x)}{\mu}\right] Bu$
- ☞ $\hat{g}_\mu(x) = \left[\frac{f(x+\mu u)-f(x-\mu u)}{2\mu}\right] Bu$

DFO gradient oracles

Let $u \sim \mathcal{N}(0, B^{-1})$. For $\mu \geq 0$, we define **gradient-free oracles**

- ☞ Sample $u \in E$ and return $g_\mu(x) = \left[\frac{f(x+\mu u) - f(x)}{\mu}\right] Bu$
- ☞ $\hat{g}_\mu(x) = \left[\frac{f(x+\mu u) - f(x-\mu u)}{2\mu}\right] Bu$
- ☞ More generally: $g_0(x) = f'(x, u) \cdot Bu$

Note:

$$f'(x, u) = \lim_{\mu \downarrow 0} \frac{f(x+\mu u) - f(x)}{\mu}$$

DFO gradient oracles

Let $u \sim \mathcal{N}(0, B^{-1})$. For $\mu \geq 0$, we define **gradient-free oracles**

- ☞ Sample $u \in E$ and return $g_\mu(x) = \left[\frac{f(x+\mu u)-f(x)}{\mu}\right] Bu$
- ☞ $\hat{g}_\mu(x) = \left[\frac{f(x+\mu u)-f(x-\mu u)}{2\mu}\right] Bu$
- ☞ More generally: $g_0(x) = f'(x,u) \cdot Bu$

Note:

$$\begin{aligned} f'(x,u) &= \lim_{\mu \downarrow 0} \frac{f(x+\mu u)-f(x)}{\mu} \\ \nabla f_0(x) &= \frac{1}{\kappa} \int_E f'(x,u) e^{-\frac{1}{2}\|u\|^2} Bu \, du. \end{aligned}$$

DFO gradient oracles

Let $u \sim \mathcal{N}(0, B^{-1})$. For $\mu \geq 0$, we define **gradient-free oracles**

- ☞ Sample $u \in E$ and return $g_\mu(x) = \left[\frac{f(x+\mu u)-f(x)}{\mu}\right] Bu$
- ☞ $\hat{g}_\mu(x) = \left[\frac{f(x+\mu u)-f(x-\mu u)}{2\mu}\right] Bu$
- ☞ More generally: $g_0(x) = f'(x,u) \cdot Bu$

Note:

$$\begin{aligned} f'(x,u) &= \lim_{\mu \downarrow 0} \frac{f(x+\mu u)-f(x)}{\mu} \\ \nabla f_0(x) &= \frac{1}{\kappa} \int_E f'(x,u) e^{-\frac{1}{2}\|u\|^2} Bu \, du. \end{aligned}$$

Exercise: If f is differentiable at x, show that $\nabla f_0(x) = \nabla f(x)$

DFO Algorithm

$$\min_{x \in \mathcal{X}} f(x)$$

DFO Algorithm

$$\min_{x \in \mathcal{X}} f(x)$$

Method: $\mathcal{R}_\mu$

- Choose $x_0 \in \mathcal{X}$ (If $\mu = 0$, x_0 must be unconstrained min!)

DFO Algorithm

$$\min_{x \in \mathcal{X}} f(x)$$

Method: $\mathcal{R}_\mu$

- Choose $x_0 \in \mathcal{X}$ (If $\mu = 0$, x_0 must be unconstrained min!)
- At iteration $k \geq 0$:
  - ✈ Generate $u_k \in E$ and compute $g_\mu(x_k)$

DFO Algorithm

$$\min_{x \in \mathcal{X}} f(x)$$

Method: $\mathcal{R}_\mu$

- Choose $x_0 \in \mathcal{X}$ (If $\mu = 0$, x_0 must be unconstrained min!)
- At iteration $k \geq 0$:
  - Generate $u_k \in E$ and compute $g_\mu(x_k)$
  - Update $x_{k+1} = P_{\mathcal{X}}(x_k - h_k B^{-1} g_\mu(x_k))$

DFO Algorithm – analysis

- Method generates a random sequence $\{x_k\}$.

DFO Algorithm – analysis

- Method generates a random sequence $\{x_k\}$.
- Denote collection of random variables up to iteration k as

$$\mathcal{U}_k := (u_0, u_1, \ldots, u_k),$$

where u_k are i.i.d.

DFO Algorithm – analysis

- Method generates a random sequence $\{x_k\}$.
- Denote collection of random variables up to iteration k as

$$\mathcal{U}_k := (u_0, u_1, \ldots, u_k),$$

  where u_k are i.i.d.
- Let $\phi_0 := f(x_0)$ and $\phi_k := E_{\mathcal{U}_{k-1}}[f(x_k)]$, for $k \geq 1$

DFO Algorithm – analysis

- Method generates a random sequence $\{x_k\}$.
- Denote collection of random variables up to iteration k as

$$\mathcal{U}_k := (u_0, u_1, \ldots, u_k),$$

  where u_k are i.i.d.
- Let $\phi_0 := f(x_0)$ and $\phi_k := E_{\mathcal{U}_{k-1}}[f(x_k)]$, for $k \geq 1$

Theorem Let $\{x_k\}$ be generated by $\mathcal{R}_0$. Then, for $T \geq 0$

$$\sum_{k=0}^{T} h_k(\phi_k - f^*) \leq \tfrac{1}{2}\|x_0 - x^*\|^2 + \tfrac{(n+4)L_0^2(f)}{2} \sum_{k=0}^{T} h_k^2.$$

Now a subgradient type stepsize selection

DFO Algorithm – analysis

- ☞ Define $S_T := \sum_{k=0}^{T} h_k$.
- ☞ Set $\hat{x}_T := \operatorname{argmin}_{0 \leq k \leq T} f(x_k)$

DFO Algorithm – analysis

- Define $S_T := \sum_{k=0}^{T} h_k$.
- Set $\hat{x}_T := \operatorname{argmin}_{0 \le k \le T} f(x_k)$

Theorem With above choice, and assuming $\|x_0 - x^*\| \le R$, we have

$$E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* \le L_0(f) R (n+4)^{1/2} \frac{1}{\sqrt{T+1}}$$

DFO Algorithm – analysis

- ☞ Define $S_T := \sum_{k=0}^{T} h_k$.
- ☞ Set $\hat{x}_T := \operatorname{argmin}_{0 \le k \le T} f(x_k)$

Theorem With above choice, and assuming $\|x_0 - x^*\| \le R$, we have

$$E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* \le L_0(f) R (n+4)^{1/2} \frac{1}{\sqrt{T+1}}$$

Proof: Let us show this $O(1/\sqrt{T})$ result.

DFO Algorithm – analysis

☞ Define $S_T := \sum_{k=0}^{T} h_k$.

☞ Set $\hat{x}_T := \operatorname{argmin}_{0 \le k \le T} f(x_k)$

Theorem With above choice, and assuming $\|x_0 - x^*\| \le R$, we have

$$E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* \le L_0(f) R (n+4)^{1/2} \frac{1}{\sqrt{T+1}}$$

Proof: Let us show this $O(1/\sqrt{T})$ result.

$$f(\hat{x}_T) - f^* \quad \le \quad \frac{1}{S_T} \sum_{k=0}^{T} h_k (f(x_k) - f^*)$$

DFO Algorithm – analysis

- ☞ Define $S_T := \sum_{k=0}^{T} h_k$.
- ☞ Set $\hat{x}_T := \operatorname{argmin}_{0 \le k \le T} f(x_k)$

Theorem With above choice, and assuming $\|x_0 - x^*\| \le R$, we have

$$E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* \le L_0(f) R (n+4)^{1/2} \frac{1}{\sqrt{T+1}}$$

Proof: Let us show this $O(1/\sqrt{T})$ result.

$$\begin{aligned} f(\hat{x}_T) - f^* &\le \frac{1}{S_T} \sum_{k=0}^{T} h_k (f(x_k) - f^*) \\ E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* &\le E_{\mathcal{U}_{T-1}}\left[\frac{1}{S_T} \sum_{k=0}^{T} h_k (f(x_k) - f^*)\right] \\ &\le \tfrac{1}{S_T}\left[\tfrac{1}{2}\|x_0 - x^*\|^2 + \tfrac{n+4}{2} L_0^2(f) \sum_{k=0}^{T} h_k^2\right] \end{aligned}$$

Now, minimize over h_k (assuming fixed T)

DFO Algorithm – analysis

Fixed step-size

$$h_k = \frac{R}{\sqrt{n+4}L_0(f)\sqrt{T+1}}, \quad k = 0, \ldots, T.$$

Which yields the desired bound.

DFO Algorithm – analysis

Fixed step-size

$$h_k = \frac{R}{\sqrt{n+4}L_0(f)\sqrt{T+1}}, \quad k = 0, \ldots, T.$$

Which yields the desired bound.

Corollary. $\mathcal{R}_0$ yields $E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* \leq \epsilon$ in

$$\frac{(n+4)L_0^2(f)R^2}{\epsilon^2} = O(1/\epsilon^2),$$

iterations.

- Theorem relies on being able to bound $E_u[\|g_0(x)\|_*^2]$. For convex f, this can be shown to be bounded by $(n+4)[\|\nabla f_0(x)\|_*^2 + nD^2(x)]$, where **diameter** $D(x) :=$ diam$\partial f(x)$

DFO Algorithm – analysis

For $\mu > 0$, we run method $\mathcal{R}_\mu$ for which we have

DFO Algorithm – analysis

For $\mu > 0$, we run method $\mathcal{R}_\mu$ for which we have

Theorem Select μ and h_k as follows

$$\mu = \frac{\epsilon}{2L_0(f)\sqrt{n}}, \quad h_k = \frac{R}{(n+4)L_0(f)\sqrt{T+1}}, \quad k = 0, \ldots, T.$$

Then, we have $E_{\mathcal{U}_{T-1}}[f(\hat{x}_T)] - f^* \leq \epsilon$, with

$$T = \frac{4(n+4)^2 L_0^2(f) R^2}{\epsilon^2}.$$

☞ Note: Dependency on dimension n is now quadratic.

DFO – stochastic optimization

$$f(x) = E_\xi[F(x,\xi)] = \int_\Xi F(x,\xi)dP(\xi)$$

- Assume $f \in C^0_{L_0}$ is convex (weaker than all: $F(x,\xi)$ being convex)

DFO – stochastic optimization

$$f(x) = E_\xi[F(x,\xi)] = \int_\Xi F(x,\xi)dP(\xi)$$

- Assume $f \in C^0_{L_0}$ is convex (weaker than all: $F(x,\xi)$ being convex)
- Replace our DF oracles by *DF-stochastic oracles*:

DFO – stochastic optimization

$$f(x) = E_\xi[F(x,\xi)] = \int_\Xi F(x,\xi) dP(\xi)$$

- Assume $f \in C^0_{L_0}$ is convex (weaker than all: $F(x,\xi)$ being convex)
- Replace our DF oracles by *DF-stochastic oracles*:
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $s_\mu(x) = \left[\frac{F(x+\mu u,\xi) - F(x,\xi)}{\mu}\right] Bu$

DFO – stochastic optimization

$$f(x) = E_\xi[F(x,\xi)] = \int_\Xi F(x,\xi)dP(\xi)$$

- Assume $f \in C^0_{L_0}$ is convex (weaker than all: $F(x,\xi)$ being convex)
- Replace our DF oracles by *DF-stochastic oracles*:
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $s_\mu(x) = \left[\frac{F(x+\mu u,\xi)-F(x,\xi)}{\mu}\right] Bu$
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $\hat{s}_\mu(x) = \left[\frac{F(x+\mu u,\xi)-F(x-\mu u,\xi)}{2\mu}\right] Bu$

DFO – stochastic optimization

$$f(x) = E_\xi[F(x,\xi)] = \int_\Xi F(x,\xi)dP(\xi)$$

- Assume $f \in C^0_{L_0}$ is convex (weaker than all: $F(x,\xi)$ being convex)
- Replace our DF oracles by *DF-stochastic oracles*:
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $s_\mu(x) = \left[\frac{F(x+\mu u,\xi)-F(x,\xi)}{\mu}\right] Bu$
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $\hat{s}_\mu(x) = \left[\frac{F(x+\mu u,\xi)-F(x-\mu u,\xi)}{2\mu}\right] Bu$
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $s_0(x) = F'_x(x,\xi;u) \cdot Bu$

DFO – stochastic optimization

$$f(x) = E_\xi[F(x,\xi)] = \int_\Xi F(x,\xi)dP(\xi)$$

- Assume $f \in C^0_{L_0}$ is convex (weaker than all: $F(x,\xi)$ being convex)
- Replace our DF oracles by *DF-stochastic oracles*:
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $s_\mu(x) = \left[\frac{F(x+\mu u,\xi)-F(x,\xi)}{\mu}\right] Bu$
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $\hat{s}_\mu(x) = \left[\frac{F(x+\mu u,\xi)-F(x-\mu u,\xi)}{2\mu}\right] Bu$
  - ☞ Sample $u \in E$, $\xi \in \Xi$, return $s_0(x) = F'_x(x,\xi;u) \cdot Bu$

Here also one gets $O(n^2/\epsilon^2)$ for $\mu > 0$

References

♡ D. P. Bertsekas. *Stochastic Optimization Problems with Nondifferentiable Cost Functionals*, (1973)

♡ Yu. Nesterov. *Random gradient-free minimization of convex functions*. (2011). (all proofs are from this reference).