Advanced Optimization

(10-801: CMU)

Lecture 28
Derivative free optimization

28 Apr 2014

Suvrit Sra

Introduction

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

Introduction

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

Optimizing without derivatives

Introduction

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

Optimizing without derivatives

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

Introduction

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

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

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

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?

$\boxed{\text{Derivative free optimization (DFO)}}$

DFO

WARNING!

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

Remarks

Not discussed today

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

Remarks

Not discussed today

- ♣ Automatic differentiation (http://www.autodiff.org)
- ♣ Fast Differentiation — $T(\nabla f) \leq 4T(f)$

  Baur, Strassen (1983) showed how to construct from a circuit computing f a circuit that computes both f and ∇f with at most 4-times increase in complexity.
- ♣ More general such result $T(\nabla f) \leq QT(f)$ by Kim, Nesterov, Cherkasskii (Sov. Math. Dokl., **29**, 384–387, (1984))

Remarks

Not discussed today

- ♣ Automatic differentiation (http://www.autodiff.org)
- ♣ Fast Differentiation — $T(\nabla f) \leq 4T(f)$

  Baur, Strassen (1983) showed how to construct from a circuit computing f a circuit that computes both f and ∇f with at most 4-times increase in complexity.
- ♣ More general such result $T(\nabla f) \leq QT(f)$ by Kim, Nesterov, Cherkasskii (Sov. Math. Dokl., **29**, 384–387, (1984))
- ♣ Various finite differencing techniques

Remarks

Not discussed today

- ♣ Automatic differentiation (http://www.autodiff.org)
- ♣ Fast Differentiation — $T(\nabla f) \leq 4T(f)$

  Baur, Strassen (1983) showed how to construct from a circuit computing f a circuit that computes both f and ∇f with at most 4-times increase in complexity.
- ♣ More general such result $T(\nabla f) \leq QT(f)$ by Kim, Nesterov, Cherkasskii (Sov. Math. Dokl., **29**, 384–387, (1984))
- ♣ Various finite differencing techniques
- ♣ Nonconvex DFO
- ♣ Recent book: "*Introduction to Derivative-Free Optimization*" by A. Conn, K. Scheinberg, and L. N. Vicente (MPS-SIAM, 2009).

DFO – brute force

$$\boxed{\min \quad f(x)}$$

DFO – brute force

$$\boxed{\min \quad f(x)}$$

Brute force method

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

DFO – brute force

$$\boxed{\min \quad f(x)}$$

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!

More cleverly: Bayesian / 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

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

(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 $f'(x; u) = \langle \nabla f(x), u \rangle$)

- ▶ If $\mathbb{E}_u(f'(x; u)u) \in \partial f(x)$ we are in good shape!

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 $f'(x;u) = \langle \nabla f(x), u \rangle$)

- If $\mathbb{E}_u(f'(x;u)u) \in \partial f(x)$ we are in good shape!
- 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 $f'(x; u) = \langle \nabla f(x), u \rangle$)

- If $\mathbb{E}_u(f'(x; u)u) \in \partial f(x)$ we are in good shape!
- 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 $f'(x; u) = \langle \nabla f(x), u \rangle$)

- If $\mathbb{E}_u(f'(x; u)u) \in \partial f(x)$ we are in good shape!
- Directional derivatives much simpler than gradient
- Can be reasonably approximated by finite differences
- Even for nonconvex functions

DFO – simulated gradients

$$x_{k+1} = x_k - h_k g_k, \qquad g_k \equiv f'(x_k; u)u.$$

- Above process may be viewed as stochastic subgradient method with random oracle

DFO – simulated gradients

$$x_{k+1} = x_k - h_k g_k, \qquad g_k \equiv f'(x_k; u)u.$$

- Above process may be viewed as stochastic subgradient method with random oracle
- Optimization problem: $\min f(x) := \mathbb{E}_u[F(x; u)]$

DFO – simulated gradients

$$x_{k+1} = x_k - h_k g_k, \qquad g_k \equiv f'(x_k; u)u.$$

- Above process may be viewed as stochastic subgradient method with random oracle
- Optimization problem: $\min f(x) := \mathbb{E}_u[F(x; u)]$
- Typical assumption here is boundedness of 2nd moment

$$\mathbb{E}_u(\|\nabla_x F(x, u)\|^2) \leq G^2 \quad x \in \mathbb{R}^n.$$

DFO – simulated gradients

$$x_{k+1} = x_k - h_k g_k, \qquad g_k \equiv f'(x_k; u)u.$$

- Above process may be viewed as stochastic subgradient method with random oracle
- Optimization problem: $\min f(x) := \mathbb{E}_u[F(x; u)]$
- Typical assumption here is boundedness of 2nd moment

$$\mathbb{E}_u(\|\nabla_x F(x, u)\|^2) \leq G^2 \quad x \in \mathbb{R}^n.$$

- In our case, if f differentiable at x

$$\mathbb{E}_u(\|f'(x; u)u\|^2) \leq (n + 4)\|\nabla f(x)\|^2$$

makes analysis simpler — but dimension dependent convergence rates.

DFO – smoothing idea

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 – smoothing idea

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 – smoothing idea

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 ($\mu_k > 0$)

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

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

DFO – smoothing idea

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 ($\mu_k > 0$)

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

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

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

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)
- ☞ $f(x) \le f_\mu(x)$.

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

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

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

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

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

Basic properties of f_μ

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

☞ $f(x) \le f_\mu(x)$. *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 $\frac{1}{\kappa}\int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (mean-zero Gaussian)

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

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 $\frac{1}{\kappa}\int_E ue^{-\frac{1}{2}\|u\|^2}du = 0$ (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:*

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

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

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

  last line follows as $\frac{1}{\kappa}\int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (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:*

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

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 $\frac{1}{\kappa}\int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (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:*

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

Basic properties of f_μ

- If f is convex, then f_μ is also convex (nonneg weighted sum)
- $f(x) \le f_\mu(x)$. *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 $\frac{1}{\kappa}\int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$ (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:*

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

- Similarly, prove that

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

Bounding moments

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

Bounding moments

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

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

Bounding moments

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

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

Bounding moments

We saw: $f(x) \le f_\mu(x)$. What about $f_\mu(x) \le f(x) + \text{something}$

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

Need to bound moments

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

Bounding moments

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

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

Need to bound moments

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

Two easy cases: $p = 0$ and $p = 2$

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

Bounding moments

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

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

Need to bound moments

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

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 (notice κ comes due to deriv. of log, and $\mathrm{Tr}(Buu^*) = \|u\|^2$)

Bounding moments

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

Proof: Simple exercise.

Bounding moments

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

Proof: Simple exercise.

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

Proof: Simple exercise.

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

Proof: Simple exercise.

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

Proof: Simple exercise.

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

Proof: Simple exercise.

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

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

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

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

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

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

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

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

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

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

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 $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- This lemma justifies the name "smoothing"

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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 $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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 $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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

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

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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 $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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

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

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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

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

Getting gradients, gradient bounds

Lemma If $f \in C^0_{L_0}$, then $f_\mu \in C^1_{L_1}$. (f_μ is **differentiable**)

- 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 $dP(u) = e^{-\frac{1}{2}\|u\|^2} u \, du$):

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 $dP(u) = e^{-\frac{1}{2}\|u\|^2} u\, du$):

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

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 $dP(u) = e^{-\frac{1}{2}\|u\|^2} u \, du$):

$$\begin{aligned} \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* &= \left| \tfrac{1}{\kappa} \int_E [\tfrac{f(x+\mu u) - f(x) + f(y) - f(y+\mu u)}{\mu}] dP(u) \right| \\ \leq \quad & \tfrac{1}{\mu\kappa} \int_E |f(x+\mu u) - f(x) + f(y) - f(y+\mu u)| \|u\| e^{-\frac{1}{2}\|u\|^2} du \end{aligned}$$

Lipschitz constant of ∇f_μ

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

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

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

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

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

Note: We got rid B in the $Bu\, du$ part because of $\|\cdot\|_*$

Simulated gradients

- Note

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

Simulated gradients

- Note

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

- **Exercise:** If f is differentiable at x, then $\nabla f_0(x) = \nabla f(x)$

Simulated gradients

- Note

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

- **Exercise:** If f is differentiable at x, then $\nabla f_0(x) = \nabla f(x)$
- More generally, if f is convex and Lipschitz continuous, then for any $x \in E$ and $\mu \geq 0$, we have

$$\nabla f_\mu(x) \in \partial_\epsilon f(x), \quad \epsilon = \mu L_0(f) \sqrt{n}$$

Gradient-free oracles

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: $g_0(x) = f'(x,u) \cdot Bu$
- ☞ Oracles g_μ and $\hat{g}_\mu$ more suitable for smooth functions

DFO Algorithm

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

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

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

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 analysis – key inequality

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

DFO analysis – key inequality

- 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 analysis – key inequality

- 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 analysis – key inequality

- 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 $\mathcal{R}_0$

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

DFO Algorithm – analysis $\mathcal{R}_0$

☞ Define $S_T := \sum_{k=0}^{T} h_k$.

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

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 $\mathcal{R}_0$

☞ Define $S_T := \sum_{k=0}^{T} h_k$.

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

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 $\mathcal{R}_0$

- ☞ 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 $\mathcal{R}_0$

☞ Define $S_T := \sum_{k=0}^{T} h_k$.

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

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.

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

Now, minimize over h_k (assuming fixed T)

DFO Algorithm – analysis $\mathcal{R}_0$

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 $\mathcal{R}_0$

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** $D(x) := \text{diam}\partial f(x)$
- If f is differentiable at x then $\mathbb{E}_u[\|g_0(x)\|_*^2] \leq (n+4)\|\nabla f_0(x)\|_*^2$

DFO Algorithm – analysis $\mathcal{R}_\mu$

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

DFO Algorithm – analysis $\mathcal{R}_\mu$

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

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

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

DFO – stochastic optimization

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

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

DFO – stochastic optimization

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

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

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

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

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

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

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

- Assume $f \in C^0_{L_0}$ is convex (weaker than all $F(x,\xi)$ 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
    $s_0(x) = F'_x(x,\xi;u) \cdot Bu$

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

Interesting directions

1. Can the dimension dependence be improved in special cases?
2. Nonconvex DFO
3. Parallel DFO
4. Distributed DFO
5. DFO for machine learning problems

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).