Random derivative-free optimization of convex functions using a line search oracle

Sebastian U. Stich\textsuperscript{a,b}, Christian L. Müller\textsuperscript{a,b}, Bernd Gärtner\textsuperscript{a}

Institute of Theoretical Computer Science
Department of Computer Science\textsuperscript{a}, Swiss Institute of Bioinformatics\textsuperscript{b}
ETH Zürich

Juli 9, 2011
1 Introduction
   - Black-box scenario
   - Convex functions
   - Existing results

2 Random Pursuit
   - The setting
   - The algorithm
   - Exp. Results

3 Outlook
   - Summary
   - Open problems
Given: \( f: E \rightarrow \mathbb{R} \)

Goal: \( f^* := f(x^*) := \min_{x \in E} f(x) \)

- \( E \) some (continuous) domain (e.g. \( \mathbb{R}^n \))
- oracle access only to the value \( f(x) \) for \( x \in E \)
- especially no access to gradients (may not even exist)
- no knowledge of the structure of \( f \)

Assumption
- \( f \) convex
Black-box scenario

**Given:** \( f : E \rightarrow \mathbb{R} \)

**Goal:** \( f^* := f(x^*) := \min_{x \in E} f(x) \)

- \( E \) some (continuous) domain (e.g. \( \mathbb{R}^n \))
- oracle access only to the value \( f(x) \) for \( x \in E \)
- especially no access to gradients (may not even exist)
- no knowledge of the structure of \( f \)

**Assumption**

- \( f \) convex
**Definition**

\( f \) is convex on \( E \) if \( E \) is a convex set and

\[
f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y), \quad \forall x, y \in E, \theta \in [0, 1]
\]

**first-order condition:** for \( f \in C^1(E) \), equivalent to

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in E
\]
If one knows that $f$ is convex, one knows probably much more . . .

- $f$ composed of simple convex functions, e.g.

$$f(x) = \ln \sum_{i=1}^{n} e^{x_i}$$

**Remark**

- $f$ given as a series of differentiable operations:
  
  *fast differentiation technique:* $C(\nabla f) \leq 4C(f)$

- if the gradients are known, very efficient methods are available

We want to focus on derivative-free optimization.
**Strongly convex functions - Quadratic lower bound**

**Definition**

\[ f \in C^1(E) \text{ is strongly convex on } E \text{ if } E \text{ is a convex set and} \]

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2, \quad \forall x, y \in E \]

if \( E = \mathbb{R}^n \), this implies \( f \) has a unique minimizer \( x^* \) and

\[ \frac{m}{2} \|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^n \]
Strongly convex functions - Quadratic lower bound

Definition

\( f \in C^1(E) \) is strongly convex on \( E \) if \( E \) is a convex set and

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \| y - x \|^2, \quad \forall x, y \in E
\]

if \( E = \mathbb{R}^n \), this implies \( f \) has a unique minimizer \( x^* \) and

\[
\frac{m}{2} \| x - x^* \|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m} \| \nabla f(x) \|^2, \quad \forall x \in \mathbb{R}^n
\]
Smooth functions - Quadratic upper bound

Definition

\( f \in C^{1,1}(E) \) if gradient is Lipschitz-continuous:

\[
\|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\|
\]

if \( E = \mathbb{R}^n \) and \( f \) has a minimizer \( x^* \), implies

\[
\frac{1}{2L_1} \|\nabla f(x)\|^2 \leq f(x) - f(x^*) \leq \frac{L_1}{2} \|x - x^*\|^2, \quad \forall x \in \mathbb{R}^n
\]
Random gradient free optimization

- Random Conic Pursuit [Kleiner et al. 2010]
- Random Gradient Method [Nesterov 2011]

Algorithm 1 Random Gradient method, [Nesterov 2011]: $(\mathcal{RG}_\mu)$

1. for $k = 0$ to $N$ do
2. \[ u_k \sim \mathcal{N}(0, I_n) \]
3. \[ x_{k+1} = x_k - h_k g_\mu(x_k) \]
4. end for

with $h_k = h_k(L_1, N)$ stepsizes, $g$ directional derivative/finite difference in random direction:

\[ g_0(x) = f'(x, u) \cdot u \quad g_\mu(x) = \frac{f(x + \mu u) - f(x)}{\mu} \cdot u \]
Random gradient free optimization

- Random Conic Pursuit [Kleiner et al. 2010]
- Random Gradient Method [Nesterov 2011]

**Algorithm 2** Random Gradient method, [Nesterov 2011]: ($\mathcal{RG}_\mu$)

1: for $k = 0$ to $N$ do
2: \[ u_k \sim \mathcal{N}(0, I_n) \]
3: \[ x_{k+1} = x_k - h_k g_\mu(x_k) \]
4: end for

with $h_k = h_k(L_1, N)$ stepsizes, $g$ directional derivative/finite difference in random direction:

\[ g_0(x) = f'(x, u) \cdot u \quad g_\mu(x) = \frac{f(x + \mu u) - f(x)}{\mu} \cdot u \]
Random gradient free optimization

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{RG}_0$</th>
<th>$\mathcal{RG}_\mu$</th>
<th>$\mathcal{FRG}_\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>smooth &amp; strongly convex</td>
<td>$O \left( n \frac{L_1}{m} \ln \frac{L_1 R^2}{\epsilon} \right)$</td>
<td>$O \left( n \sqrt{\frac{L_1}{m}} \ln \frac{f_0 + R^2}{\epsilon} \right)$</td>
<td>$O \left( n \sqrt{\frac{L_1}{m}} (f_0 + R^2) \right)$</td>
</tr>
<tr>
<td>smooth &amp; convex</td>
<td>$O \left( n \frac{L_1 R^2}{\epsilon} \right)$</td>
<td>$O \left( n \sqrt{\frac{L_1}{m} (f_0 + R^2)} \right)$</td>
<td>$O \left( n \sqrt{\frac{L_1}{m} (f_0 + R^2)} \right)$</td>
</tr>
<tr>
<td>convex</td>
<td>$O \left( n \frac{L_0^2 R^2}{\epsilon^2} \right)$</td>
<td>$O \left( n^2 \frac{L_0^2 R^2}{\epsilon^2} \right)$</td>
<td>-</td>
</tr>
</tbody>
</table>

Where $R = \| x_0 - x^* \|$, $f_0 = f(x_0) - f(x^*)$. Usually $n$ times more iterations than standard gradient methods.
Assumptions:

- $f$ convex
- $f \in C^{1,1}(\mathbb{R}^n)$, $L_1$-Lipschitz continuous gradients
- efficient 1-dim optimization

**Definition (Line search oracle)**

For $x, u \in \mathbb{R}^n$, and $\epsilon \geq 0$ a function

$$L_\epsilon : \mathbb{R} \times S^{n-1} \rightarrow \mathbb{R}$$

$$(x, u) \mapsto h$$

with

$$f(x + hu) \leq \min_{h \in \mathbb{R}^n} f(x + hu) + \epsilon.$$
The setting

Assumptions:

- \( f \) convex
- \( f \in C^{1,1}(\mathbb{R}^n) \), \( L_1 \)-Lipschitz continuous gradients
- efficient 1-dim optimization

Definition (Line search oracle)

For \( x, u \in \mathbb{R}^n \), and \( \epsilon \geq 0 \) a function

\[
\mathcal{L}_\epsilon : \mathbb{R} \times S^{n-1} \rightarrow \mathbb{R}
\]

\[(x, u) \mapsto h\]

with

\[
f(x + hu) \leq \min_{h \in \mathbb{R}^n} f(x + hu) + \epsilon.
\]

Remark

Relative accuracy of \( \mathcal{L} \) yields similar results (with different analysis).
The setting

Assumptions:

- $f$ convex
- $f \in C^{1,1}(\mathbb{R}^n)$, $L_1$-Lipschitz continuous gradients
- efficient 1-dim optimization

**Definition (Line search oracle)**

For $x, u \in \mathbb{R}^n$, and $\epsilon \geq 0$ a function

$$ \mathcal{L}_\epsilon : \mathbb{R} \times S^{n-1} \to \mathbb{R} $$

$$(x, u) \mapsto h$$

with

$$ f(x + hu) \leq \min_{h \in \mathbb{R}^n} f(x + hu) + \epsilon. $$

**Remark**

Relative accuracy of $\mathcal{L}$ yields similar results (with different analysis).
Random Pursuit for unconstrained minimization

**Algorithm 3** Random Pursuit: ($\mathcal{RP}_\epsilon$)

1: for $k = 0$ to $N$ do
2: $u_k \sim \mathcal{N}(0, I_n)$
3: $\hat{h} = \mathcal{L}_\epsilon(x_k, u_k)$
4: $x_{k+1} = x_k + \hat{h}u_k$
5: end for

No input parameters needed, except $x_0$. 
The analysis

Let $x_k$ be the current iterate, $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}$ arbitrarily.

$$f(x_{k+1}) = f(x_k + \mathcal{L}_\epsilon(x, u)u)$$
$$\leq f(x_k + h^* u) + \epsilon$$
$$\leq f(x_k + \frac{1}{L_1 n} \alpha(u)u) + \epsilon$$

Quadratic upper bound:

$$f(x_{k+1}) \leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), \alpha(u)u \rangle + \frac{1}{L_1 n^2} \|\alpha(u)u\|^2 + \epsilon$$

Similar trick was used by [Kleiner et al. 2010]
Let $x_k$ be the current iterate, $\alpha: \mathbb{R}^n \to \mathbb{R}$ arbitrarily.

\[
\begin{align*}
f(x_{k+1}) &= f(x_k + \mathcal{L}_\epsilon(x, u)u) \\
&\leq f(x_k + h^* u) + \epsilon \\
\text{optimality of } h^*
&\leq f(x_k + \frac{1}{L_1 n} \alpha(u)u) + \epsilon \\
\text{want: } &\approx \frac{1}{L_1 n} (x^* - x_k)
\end{align*}
\]

Quadratic upper bound:

\[
f(x_{k+1}) \leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), \alpha(u)u \rangle + \frac{1}{L_1 n^2} \|\alpha(u)u\|^2 + \epsilon
\]

Similar trick was used by [Kleiner et al. 2010]
Lemma

Let $U$ be the uniform distribution on all unit length vectors $u \in S^{n-1}$. For $z \in \mathbb{R}^n$, there exists $\alpha: \mathbb{R}^n \rightarrow \mathbb{R}$, s.t.

\[
\mathbb{E}_U [\alpha(u)u] = z \quad \mathbb{E}_U [\|\alpha(u)u\|^2] = n \|z\|^2.
\]

Proof.

Choose

$\alpha(u) = n \langle u, z \rangle$. \hfill \Box
How to choose $\alpha$

**Lemma**

Let $U$ be the uniform distribution on all unit length vectors $u \in S^{n-1}$. For $z \in \mathbb{R}^n$, there exists $\alpha : \mathbb{R}^n \to \mathbb{R}$, s.t.

$$
\mathbb{E}_U [\alpha(u)u] = z \quad \mathbb{E}_U [\|\alpha(u)u\|^2] = n \|z\|^2.
$$

**Proof.**

Choose

$$
\alpha(u) = n \langle u, z \rangle.
$$

□
Progress rate

\[
f(x_{k+1}) \leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), \alpha(u)u \rangle + \frac{1}{L_1 n^2} \|\alpha(u)u\|^2 + \epsilon
\]

Using Lemma with \( z = x^* - x_k \):

\[
\mathbb{E}[f(x_{k+1}) \mid x_t] \leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), x^* - x_k \rangle + \frac{1}{L_1 n} \|x^* - x_k\|^2 + \epsilon
\]

For strongly convex functions:

\[
\mathbb{E}[f(x_{k+1}) \mid x_t] \leq \left(1 - \frac{m}{L_1 n}\right) \cdot (f(x_k) - f(x^*)) + \epsilon
\]
Progress rate

\[
f(x_{k+1}) \leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), \alpha(u)u \rangle + \frac{1}{L_1 n^2} \|\alpha(u)u\|^2 + \epsilon
\]

Using Lemma with \( z = x^* - x_k \):

\[
\mathbb{E}[f(x_{k+1}) \mid x_t] \leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), x^* - x_k \rangle + \frac{1}{L_1 n} \|x^* - x_k\|^2 + \epsilon
\]

\[
\overset{\text{convex}}{\leq} \left(1 - \frac{1}{L_1 n}\right)(f(x_k) - f(x^*)) + \frac{1}{L_1 n} \|x^* - x_k\|^2 + \epsilon
\]

For strongly convex functions:

\[
\mathbb{E}[f(x_{k+1}) \mid x_t] \leq \left(1 - \frac{m}{L_1 n}\right) \cdot (f(x_k) - f(x^*)) + \epsilon
\]
Accelerated method

Algorithm 4 Accelerated Random Pursuit: (FRP)

1: $\theta = \frac{1}{L_1 n}$, $\gamma_0 \geq m$
2: for $k = 0$ to $N$ do
3: Compute $\beta_k > 0$ satisfying $\theta^{-1} \beta_k^2 = (1 - \beta_k) \gamma_k + \beta_k m =: \gamma_{k+1}$
4: Set $\lambda_k = \frac{\beta_k m}{\gamma_{k+1}}$, $\delta_k = \frac{\beta_k \gamma_k}{\gamma_k + \beta_k m}$, and $y_k = (1 - \beta_k) x_k + \beta_k v_k$.
5: $u_k \sim \mathcal{N}(0, I_n)$
6: $\hat{h} = \mathcal{L}_\epsilon(x_k, u_k)$
7: Set $x_{k+1} = y_k + \hat{h} u_k$, $v_{k+1} = (1 - \lambda_k) v_k + \lambda_k y_k + \frac{h^*}{\beta_k n} u_k$
8: end for

- maintain two iterates $x_k, y_k$
- adaptation of $\mathcal{FGM}$, [Nesterov 2011]
- no proof . . .
Theoretical results

\[
\begin{array}{|c|c|c|}
\hline
& \mathcal{RP}_0 & \mathcal{RP}_\mu & \mathcal{FRP} \\
\hline
\text{smooth & strongly convex} & O \left( n \frac{L_1}{m} \ln \frac{f_0}{\epsilon} \right) & - \\
\text{smooth & convex} & O \left( n \frac{L_1 R^2 + f_0}{\epsilon} \right) & - \\
\hline
\end{array}
\]

Where \( \|x_k - x^*\| \leq R, \ k = 0, \ldots, N, \ f_0 = f(x_0) - f(x^*). \)
Test function

\[ f_{m,L_1}(x) = \frac{(L_1 - m)}{8} \left[ (x_1)^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + (x_n)^2 - 2x_1 \right] + \frac{m}{2} \|x\|^2 \]

With parameters:

- \( dim = 64 \)
- \( L_1 = 1000 \)
- \( m = 1 \)
- \( x_0 = 0 \)
- \( (x^*) = \begin{bmatrix} 0.939 \\ 0.881 \\ \vdots \end{bmatrix} \)
Results - Evolution paths

Gradient Method

Random Gradient Method [Nes11]

Random Pursuit

Accelerated Random Pursuit

S. Stich, Ch. Müller, B. Gärtner

Random Pursuit
Results - Convergence plots

\[ \ln \|f^* - f(x)\| \]

- Red: Random Gradient
- Blue: Gradient Method
- Green: Random Pursuit
- Black: Accelerated RP

Iterations / \( N \)
Results - Convergence plots

\[
\ln \left\| f^* - f(x) \right\|
\]

\begin{align*}
&\text{Random Gradient} \\
&\text{Gradient Method} \\
&\text{Random Pursuit} \\
&\text{Accelerated RP}
\end{align*}

iterations/N
Summary

- Random Conic Pursuit [Kleiner et al. 2010]
  - explicit bounds
  - linear convergence for strongly convex functions
- Random Gradient [Nesterov 2011]
  - no input parameters needed ($L_1$)
Open problems

- Constraint handling: \( \min_{x \in \mathcal{K}} f(x) \)
- Accelerated methods
- Non-differentiable functions
  - \( \rightarrow \) random noise
- Variable metric
Open problems

- Constraint handling: \( \min_{x \in \mathcal{K}} f(x) \)
- Accelerated methods
- Non-differentiable functions
  - \( \rightarrow \) random noise
- Variable metric

Thank you