Noisy Optimization Complexity

Under Locality Assumption

Jeremie Decock

Olivier Teytaud

Inria

March 17, 2014

Introduction

The aim of this presentation

- Study convergence rate for numerical optimization with noisy objective function
- Reduce the gap between upper and lower bounds
  - for a given family of noisy objective function
  - under some assumptions

Plan

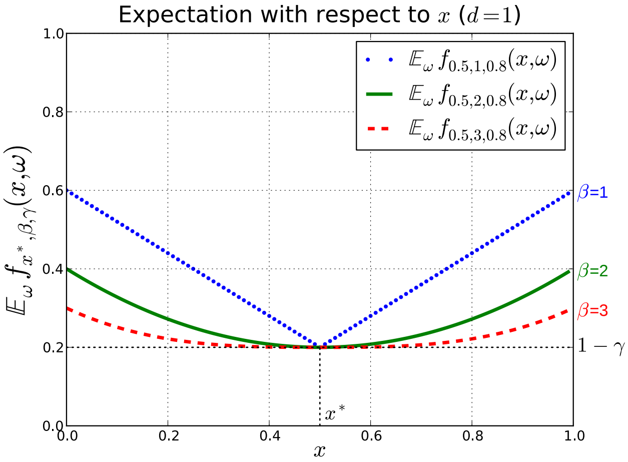
Framework

Considered family F of objective functions

$$f_{\mathbf{x}^*,\beta,\gamma} : \quad [0,1]^d \to \{0,1\}$$
$$x \mapsto \mathcal{B}\left(\underbrace{\gamma\left(\frac{\|\mathbf{x}-\mathbf{x}^*\|}{\sqrt{d}}\right)^{\beta} + (1-\gamma)}_{p}\right)$$

- $f_{\mathbf{x}^*,\beta,\gamma} \in F$: a stochastic objective function (a random variable)
- $\mathcal{B}(p)$: the Bernoulli distribution with parameter p
- $d \in \mathbb{N}^*$: the dimension of the domain of $f_{\mathbf{x}^*,\beta,\gamma}$
- $\mathbf{x}^* \in [0,1]^d$: the optimum
- $\beta \in \mathbb{R}_+^*$: the "flatness" of the expectation $\mathbb{E}f$ around $\mathbf{x}^*$
- $\gamma \in [0,1]$: a noise parameter (variance at x^*)

Considered objective functions



Considered objective functions

- we assume that the optimum is unique
- we consider noisy optimization in the case of local convergence

All families including the family of function F are also concerned by this presentation

The simulation protocol for our noisy optimization setting

Require: ω, ω', $\mathbf{x}^*$, β, γ and (unknown).
for all $n = 1, 2, 3, \ldots$ **do**
$\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'} = \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\omega,\omega'}, \ \ldots \ , \mathbf{x}_{\mathbf{x}^*,n-1,\omega,\omega'}, \ y_1, \ \ldots \ , y_{n-1}, \ \omega')$
if $\omega_n \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'})$ **then**
$y_n = 1$
else
$y_n = 0$
end if
end for
return $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'}$

The framework requires:

- ω uniform random variable for simulating the Bernoulli in f
- ω' a random seed of the algorithm (optimizer's internal randomness)
- $\mathbf{x}^*$ the optimum (that minimizes $\mathbb{E}_\omega f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}, \omega)$)
- β and γ two fixed parameters of f

The simulation protocol for our noisy optimization setting

Require: ω, ω', $\mathbf{x}^*$, β, γ and (unknown).
for all $n = 1, 2, 3, \ldots$ **do**
$\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'} = \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\omega,\omega'}, \ \ldots \ , \mathbf{x}_{\mathbf{x}^*,n-1,\omega,\omega'}, \ y_1, \ \ldots \ , \ y_{n-1}, \ \omega')$
if $\omega_n \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'})$ **then**
$y_n = 1$
else
$y_n = 0$
end if
end for
return $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'}$

It's an iterative process. For each iteration:

- *Optimize* (the optimization algorithm) returns the next point to visit
  - looking for $\mathbf{x}^*$
  - according to some inputs
- This point is evaluated by the *fitness function* (it-then-else statement)

The simulation protocol for our noisy optimization setting

Require: ω, ω', $\mathbf{x}^*$, β, γ and (unknown).
  for all $n = 1, 2, 3, \dots$ **do**
    $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'} = \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\omega,\omega'}, \ \dots \ , \mathbf{x}_{\mathbf{x}^*,n-1,\omega,\omega'}, \ y_1, \ \dots \ , y_{n-1}, \ \omega')$
    if $\omega_n \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'})$ **then**
      $y_n = 1$
    else
      $y_n = 0$
    end if
  end for
  return $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'}$

Optimize makes its recommendation according to:

- the sequence of former visited points $\mathbf{x}_n$
- their binary noisy fitness values y_n
- the optimizer's internal randomness ω'

The simulation protocol for our noisy optimization setting

Require: ω, ω', $\mathbf{x}^*$, β, γ and (unknown).
for all $n = 1, 2, 3, \ldots$ **do**
$\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'} = \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\omega,\omega'}, \ \ldots \ , \mathbf{x}_{\mathbf{x}^*,n-1,\omega,\omega'}, \ y_1, \ \ldots \ , \ y_{n-1}, \ \omega')$
if $\omega_n \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'})$ **then**
$y_n = 1$
else
$y_n = 0$
end if
end for
return $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'}$

The fitness function $f_{\mathbf{x}^*,\beta,\gamma}$ outputs

- 1 if random $(= \omega_n)$ is less than $\mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'})$
- 0 otherwise

The simulation protocol for our noisy optimization setting

Require: ω, ω', $\mathbf{x}^*$, β, γ and (unknown).
for all $n = 1, 2, 3, \ldots$ **do**
$\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'} = \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\omega,\omega'}, \ \ldots \ , \mathbf{x}_{\mathbf{x}^*,n-1,\omega,\omega'}, \ y_1, \ \ldots \ , \ y_{n-1}, \ \omega')$
if $\omega_n \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'})$ **then**
$y_n = 1$
else
$y_n = 0$
end if
end for
return $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'}$

Eventually, the estimation of the optimum is returned

Sampling strategies

Sampling close to the current estimation of the optimum

- most evolution strategies do that

Sampling far from the current estimation of the optimum

- when f is learnable
- the optimizer can uses a model of f to sample far from x^*
- *optimize*'s outputs can have different meanings
  - be the most informative points to sample (exploration)
  - provide an estimate of $\arg\min \mathbb{E}f$ (recommendation)

These strategies lead to different convergence rates.

Our assumptions

We assume that the optimization algorithm

- doesn't require a model of the objective function
- samples close to the optimum (first strategy)

The *locality* assumption

$$\forall f \in F, \quad P\left(\forall i \leq n, \quad \|\mathbf{x}_i - \mathbf{x}^*\| \leq \frac{C(d)}{i^\alpha}\right) \geq 1 - \frac{\delta}{2}$$

- for some $0 < \delta < 1/2$
- $C(d) > 0$: a constant depending on d only
- $\alpha > 0$: convergence speed (large α implies a fast convergence)

Complexity

Noisy optimization complexity

What is the best theoretical convergence rate of an optimization algorithm on $f \in F$ assuming this *locality assumption* ?

$$\forall f \in F, \quad P\left(\forall i \leq n, \quad \|\mathbf{x}_i - \mathbf{x}^*\| \leq \frac{C(d)}{i^\alpha}\right) \geq 1 - \frac{\delta}{2}$$

That is to say...

What is the supremum of possible α ?

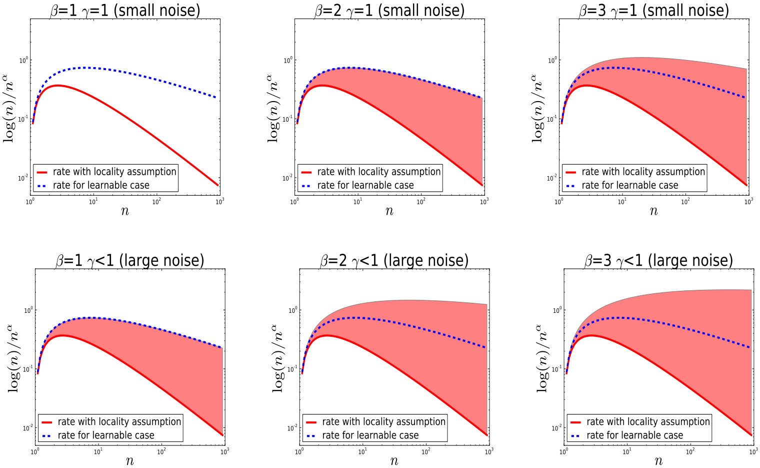
State of art

Complexity of the optimization algorithm on f (subject to probability):

$$||\mathbf{x}_n - \mathbf{x}^*|| = O\left(\frac{\log(n)}{n^\alpha}\right)$$

	$\gamma = 1$ (small noise)	$\gamma < 1$ (large noise)
Proved rate for R-EDA	$\frac{1}{\beta} \leq \alpha$	$\frac{1}{2\beta} \leq \alpha$
Former lower bounds	$\alpha \leq 1$	$\alpha \leq 1$
R-EDA experimental rates	$\alpha = \frac{1}{\beta}$	$\alpha = \frac{1}{2\beta}$
Rate by active learning	$\alpha = \frac{1}{2}$	$\alpha = \frac{1}{2}$

State of art


β=1 γ=1 (small noise)
β=2 γ=1 (small noise)
β=3 γ=1 (small noise)
β=1 γ<1 (large noise)
β=2 γ<1 (large noise)
β=3 γ<1 (large noise)
log(n)/n^α
n
rate with locality assumption
rate for learnable case

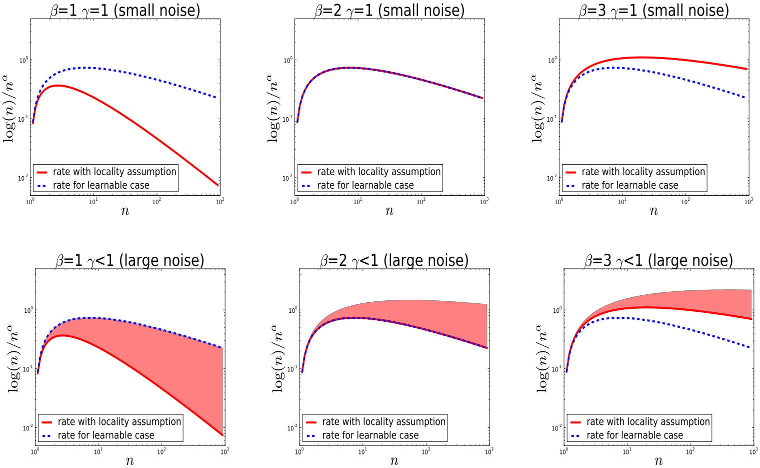
The aim of this presentation (that we prove)

Complexity of the optimization algorithm on f (subject to probability):

$$||\mathbf{x}_n - \mathbf{x}^*|| = O\left(\frac{\log(n)}{n^\alpha}\right)$$

	$\gamma = 1$ (small noise)	$\gamma < 1$ (large noise)
Proved rate for R-EDA	$\frac{1}{\beta} \leq \alpha$	$\frac{1}{2\beta} \leq \alpha$
Former lower bounds	$\alpha \leq 1$	$\alpha \leq 1$
This paper (lower bounds)	$\alpha \leq \frac{1}{\beta}$	$\alpha \leq \frac{1}{\beta}$

New possible values of α whith our theorem

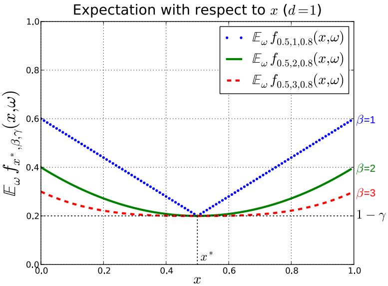


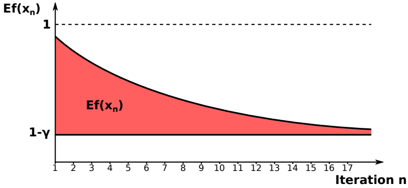
Lemmas

Introduction

With probability at least $1-\delta/2$:

$$\underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} \leq \mathbb{E}f(\mathbf{x}_n) \leq \underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{n^{\alpha\beta}}$$





Lemma 1

Lemma 1

Definition

$X_{n,\Omega}$ = the set of all the $\mathbf{x}_{\mathbf{x}^*,n,\omega,\omega'}$ for all $\mathbf{x}^*$ (with $\Omega = (\omega, \omega'))$

Combinatorial lemma (number of possible outcomes)

The cardinality of $X_{n,\Omega}$ is at most 2^N, where N is the cardinality of

$$\left\{ 1 \leq i \leq n \;\; ; \;\; \underbrace{\mathbb{E} f(\mathbf{x}^*)}_{1-\gamma} \;\; \leq \;\; \omega_i \;\; \leq \;\; \underbrace{\mathbb{E} f(\mathbf{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}} \frac{C(d)^\beta}{i^{\alpha\beta}} \right\}$$

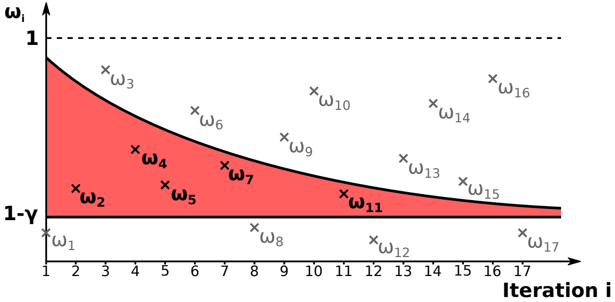
Lemma 1 - Proof

Require: $\boldsymbol{\omega}$, $\boldsymbol{\omega}'$, $\mathbf{x}^*$, β and γ.
  for all $n = 1, 2, 3, \ldots$ **do**
    $\mathbf{x}_{\mathbf{x}^*,n,\boldsymbol{\omega},\boldsymbol{\omega}'} = \text{Optimize}(\underbrace{\mathbf{x}_{\mathbf{x}^*,1,\boldsymbol{\omega},\boldsymbol{\omega}'}, \ \ldots \ , \ \mathbf{x}_{\mathbf{x}^*,n-1,\boldsymbol{\omega},\boldsymbol{\omega}'}}, \ \underbrace{y_1, \ \ldots \ , \ y_{n-1}}, \ \underbrace{\boldsymbol{\omega}'})$
    if $\omega_n \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,n,\boldsymbol{\omega},\boldsymbol{\omega}'})$ **then**
      $y_n = 1$
    else
      $y_n = 0$
    end if
  end for
  return $\mathbf{x}_{\mathbf{x}^*,n,\boldsymbol{\omega},\boldsymbol{\omega}'}$

Lemma 1 - Proof

$$
\begin{aligned}
\mathbf{x}_{\mathbf{x}^*,1,\boldsymbol{\omega},\boldsymbol{\omega}'} &= \text{Optimize}(\boldsymbol{\omega}') \\
\forall\Omega, |X_{1,\Omega}| &= 1 \\
y_1 &= \begin{cases} 1 & \text{if } \omega_1 \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,1,\boldsymbol{\omega},\boldsymbol{\omega}'}) \\ 0 & \text{otherwise} \end{cases} \\
\mathbf{x}_{\mathbf{x}^*,2,\boldsymbol{\omega},\boldsymbol{\omega}'} &= \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\boldsymbol{\omega},\boldsymbol{\omega}'},\ y_1,\ \boldsymbol{\omega}') \\
\forall\Omega, |X_{2,\Omega}| &\leq 2 \\
y_2 &= \begin{cases} 1 & \text{if } \omega_2 \leq \mathbb{E} f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}_{\mathbf{x}^*,2,\boldsymbol{\omega},\boldsymbol{\omega}'}) \\ 0 & \text{otherwise} \end{cases} \\
\mathbf{x}_{\mathbf{x}^*,3,\boldsymbol{\omega},\boldsymbol{\omega}'} &= \text{Optimize}(\mathbf{x}_{\mathbf{x}^*,1,\boldsymbol{\omega},\boldsymbol{\omega}'},\ \mathbf{x}_{\mathbf{x}^*,2,\boldsymbol{\omega},\boldsymbol{\omega}'},\ y_1,\ y_2,\ \boldsymbol{\omega}') \\
\forall\Omega, |X_{3,\Omega}| &\leq 4 \\
\dots & \\
\forall\Omega, |X_{n,\Omega}| &\leq 2^{n-1}
\end{aligned}
$$

Lemma 1 - Proof



The cardinality of $X_{n,\Omega}$ is at most 2^N

$$N = \left| \left\{ 1 \leq i \leq n \;\; ; \;\; \underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} \;\; \leq \;\; \omega_i \;\; \leq \;\; \underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}} \frac{C(d)^\beta}{i^{\alpha\beta}} \right\} \right|$$

Lemma 2

Lemma 2

Definition (reminder)

N is the cardinality of

$$\left\{ 1 \leq i \leq n \;\; ; \;\; \underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} \;\leq\; \omega_i \;\leq\; \underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}} \frac{C(d)^\beta}{i^{\alpha\beta}} \right\}$$

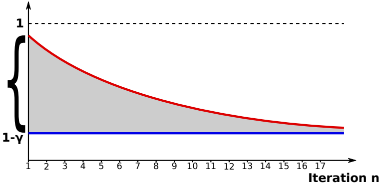
Lemma (number of ones)

N has expectation and variance at most

$$z = \frac{\gamma}{d^{\beta/2}} \, C(d)^\beta \sum_{i=1}^{n} i^{-\alpha\beta}$$

Lemma 2

$$\mathbb{E}f(\mathbf{x}^*) + \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}} - \mathbb{E}f(\mathbf{x}^*) = \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}}$$



$$Z_i \sim \mathcal{B}\left(\frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}}\right)$$

$$\mathbb{E}(Z_i) = \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}} \qquad V(Z_i) = \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}}\left(1 - \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}}\right)$$

$$\text{Expectation of N} = \sum_{i=1}^{n} \mathbb{E}(Z_i) = \sum_{i=1}^{n} \frac{\gamma}{d^{\beta/2}} \frac{C(d)^{\beta}}{i^{\alpha\beta}} = \frac{\gamma}{d^{\beta/2}} C(d)^{\beta} \sum_{i=1}^{n} \frac{1}{i^{\alpha\beta}} = z$$

$$\text{Variance of N} = \sum_{i=1}^{n} V(Z_i) \leq \sum_{i=1}^{n} \mathbb{E}(Z_i) \quad (\text{as } V(Z_i) \leq \mathbb{E}(Z_i)) \leq z$$

Lemma 3

Lemma 3

Lemma (Lemma 2 + Chebyshev's inequality)

Consider $\delta \in [0,1]$

$$N \leq z + \sqrt{z}\,(\delta/2)^{-1/2}$$

with probability at least $1 - \delta/2$

The cardinality of $X_{n,\Omega}$

Lemmas 1 and 3 together imply that the cardinality of $X_{n,\Omega}$ is at most

$$2^{z+\frac{\sqrt{z}}{\sqrt{\delta/2}}}$$

with probability at least $1-\delta/2$ and with

$$z = \frac{\gamma}{d^{\beta/2}}\ C(d)^{\beta} \sum_{i=1}^{n} i^{-\alpha\beta}$$

Main result

Theorem

Theorem

Assume the objective function $f \in F$

Assume the *locality assumption*

$$\forall f \in F, \quad P\left(\forall i \leq n, \quad \|\mathbf{x}_i - \mathbf{x}^*\| \leq \frac{C(d)}{i^\alpha}\right) \geq 1 - \frac{\delta}{2}$$

Then $\alpha \leq 1/\beta$.

Proof

Let us show that $\alpha\beta \leq 1$

In order to do so, let us assume, in order to get a contradiction, that $\alpha\beta > 1$

Proof

Knowing convergence of Riemann series for $\alpha\beta > 1$

$$\sum_{i=1}^{n} \frac{1}{i^{\alpha\beta}} < \frac{\alpha\beta}{\alpha\beta - 1}$$

lemma 2 leads to

$$z \leq \frac{\gamma C(d)^{\beta}}{d^{\beta/2}} \frac{\alpha\beta}{\alpha\beta - 1} \text{ if } \alpha\beta > 1$$

That is to say...

When $\alpha\beta > 1$, $X_{n,\Omega}$ is a finite set

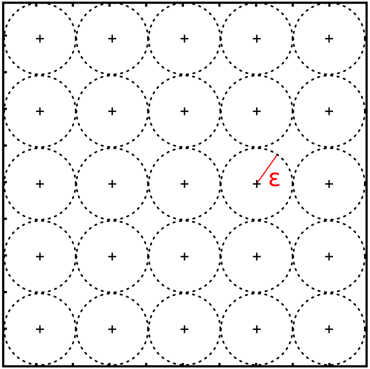
Proof

Definition

Consider R a set of points with lower bounded distance to each other:

- two distinct elements of R are at distance greater than 2ϵ from each other with

$$\epsilon = \frac{C(d)}{i^{\alpha}}$$



Reminder

The *locality* assumption

$$\forall f \in F, \quad P\left(\forall i \leq n, \quad \|\mathbf{x}_i - \mathbf{x}^*\| \leq \frac{C(d)}{i^{\alpha}}\right) \geq 1 - \frac{\delta}{2}$$

Proof

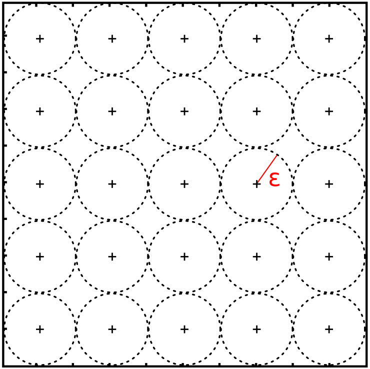
If x^* is uniformly drawn in R...

Then to respect the locality assumption:

- $X_{n,\Omega}$ should have points at maximum distance ϵ of a certain percentage $(1 - \delta/2)$ of R's elements

That is to say:

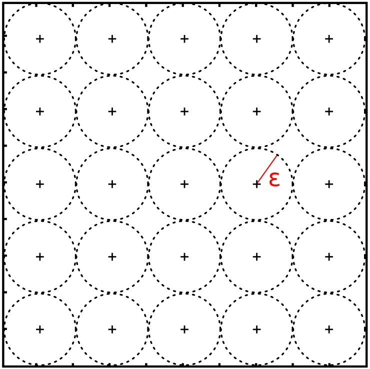
- *optimize* should have the opportunity to choose points which are at most at distance ϵ to (almost) each possible x^*



Proof

But...

- if ϵ decreases (x_i approach x^*) then $|R|$ increases
- and $X_{n,\Omega}$ is finite (with great probability)



Up to a certain timestep, the locality assumption can't be respected.

We have a contradiction on our assumption.

$\alpha\beta > 1$ is wrong

Conclusion

Conclusion

We proved that $\alpha \leq \frac{1}{\beta}$ for $f \in F$ assuming...

$$\forall f \in F, \quad P\left(\forall i \leq n, \quad \|\mathbf{x}_i - \mathbf{x}^*\| \leq \frac{C(d)}{i^\alpha}\right) \geq 1 - \frac{\delta}{2}$$

	$\gamma = 1$ (small noise)	$\gamma < 1$ (large noise)
Proved rate for R-EDA	$\frac{1}{\beta} \leq \alpha$	$\frac{1}{2\beta} \leq \alpha$
Former lower bounds	$\alpha \leq 1$	$\alpha \leq 1$
This paper (lower bounds)	$\alpha \leq \frac{1}{\beta}$	$\alpha \leq \frac{1}{\beta}$

Conclusion

	$\gamma = 1$ (small noise)	$\gamma < 1$ (large noise)
Rate with locality assumption	$\alpha = \frac{1}{\beta}$	$\frac{1}{2\beta} \leq \alpha \leq \frac{1}{\beta}$
Rate by active learning	$\alpha = \frac{1}{2}$	$\alpha = \frac{1}{2}$

Second conclusion

faster rates can only be obtained by sampling far from the optimum (it requires an appropriate objective function and an accurate model of it)

Future work

- reduce the remaining gap between the upper and the lower bound (in the large noise case $\gamma < 1$)
- investigate on intermediate models (sampling both close and far from the optimum)
- consider the case of global convergence
- find a dependency in d ?

Thank you for your attention

Questions ?

Lemmas - Introduction (1)

The objective function

$$f_{\mathbf{x}^*,\beta,\gamma}(\mathbf{x}) = \mathcal{B}\left(\gamma\left(\frac{\|\mathbf{x}-\mathbf{x}^*\|}{\sqrt{d}}\right)^{\beta} + (1-\gamma)\right)$$

and the *locality assumption*

$$\forall f \in F, \quad P\left(\forall i \leq n, \quad \underbrace{\|\mathbf{x}_i - \mathbf{x}^*\| \leq \frac{C(d)}{i^{\alpha}}}\right) \geq 1 - \frac{\delta}{2}$$

imply

$$\underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} \quad \leq \quad \overbrace{\mathbb{E}f(\mathbf{x}_n) \quad \leq \quad \underbrace{\mathbb{E}f(\mathbf{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}}\frac{C(d)^{\beta}}{n^{\alpha\beta}}}$$

with probability at least $1 - \delta/2$

(f and $\mathbb{E}f(\mathbf{x})$ are short notations for $f_{\mathbf{x}^*,\beta,\gamma}$ and $\mathbb{E}_{\boldsymbol{\omega}} f(\mathbf{x},\boldsymbol{\omega})$)

Lemmas - Introduction (2)

$$
\begin{aligned}
& \|\mathbf{x}_i - \mathbf{x}^*\| \quad \leq \quad \frac{C(d)}{i^\alpha} \quad \text{(the locality assumption)} \\
\Leftrightarrow \quad & \underbrace{\gamma \left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|}{\sqrt{d}} \right)^\beta + (1-\gamma)} \quad \leq \quad \underbrace{(1-\gamma)} + \gamma \left(\frac{C(d)}{i^\alpha \sqrt{d}} \right)^\beta \\
\Leftrightarrow \quad & \mathbb{E} f(\mathbf{x}_i) \quad \leq \quad \mathbb{E} f(\mathbf{x}^*) + \gamma \left(\frac{C(d)}{i^\alpha \sqrt{d}} \right)^\beta \\
\Leftrightarrow \quad & \mathbb{E} f(\mathbf{x}_i) \quad \leq \quad \mathbb{E} f(\mathbf{x}^*) + \frac{\gamma C(d)^\beta}{i^{\alpha\beta} \sqrt{d}^\beta}
\end{aligned}
$$

Proof

Consider $f_{\mathbf{x}^*} = f_{\mathbf{x}^*,\beta,\gamma}$ with $\mathbf{x}^*$ uniformly distributed in R. Then:

$$\begin{aligned}
& P(\|\mathbf{x}_n - \mathbf{x}^*\| \leq \epsilon) \\
\leq \quad & \mathbb{E}_\Omega P_{\mathbf{x}^*}(\mathbf{x}^* \in \mathrm{Enl}(X_{n,\Omega}, \epsilon)) \\
\leq \quad & P(\#X_{n,\Omega} \leq C) P_{\mathbf{x}^*}(\mathbf{x}^* \in \mathrm{Enl}(X_{n,\Omega}, \epsilon) | \#X_{n,\Omega} \leq C) \\
& \quad + P(\#X_{n,\Omega} > C) \\
\leq \quad & (1 - \frac{\delta}{2}) \frac{C}{C'} + \frac{\delta}{2} \\
< \quad & 1 - \frac{\delta}{2}
\end{aligned}$$

where $\mathrm{Enl}(U, \epsilon)$ is the ϵ-enlargement of U defined as:

$$\mathrm{Enl}(U, \epsilon) = \left\{ \mathbf{x}; \exists \mathbf{x}' \in U, \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon \right\}.$$

This contradicts the locality assumption.

This concludes the proof of $\alpha\beta \leq 1$.

R-EDA (1)

Algorithm 2 R-EDA: algorithm for optimizing noisy fitness functions. *Bernstein* denotes a Bernstein race, as defined in Algorithm 3. The initial domain is $[x_0^-, x_0^+] \in \mathbb{R}^D$, δ is the confidence parameter. This algorithm goes back to [15, 16].

$n \leftarrow 0$
while True **do**
  // Pick the coordinate with highest uncertainty
  $c_n = \arg\max_i (x_n^+)_i - (x_n^-)_i$
  $\delta_n^{\max} = (x_n^+)_{c_n} - (x_n^-)_{c_n}$
  for $i \in [[1, 3]]$ **do**
    // Consider the middle point
    $x'^i_n \leftarrow \frac{1}{2}(x_n^- + x_n^+)$
    // The c_n^{th} coordinate may take 3 $\neq$ values
    $(x'^i_n)_{c_n} \leftarrow (x_n^-)_{c_n} + \frac{i-1}{2}(x_n^+ - x_n^-)_{c_n}$
  end for
  $(good_n, bad_n) = Bernstein(x'^1_n, x'^2_n, x'^3_n, \frac{6\delta}{\pi^2(n+1)^2})$.
  // A good and a bad point
  Let H_n be the halfspace
  $\{x \in \mathbb{R}^D; ||x - good_n|| \leq ||x - bad_n||\}$
  Split the domain: $[x_{n+1}^-, x_{n+1}^+] = H_n \cap [x_n^-, x_n^+]$
  $n \leftarrow n + 1$
end while

R-EDA (2)

Algorithm 3 Bernstein race between 3 points. Eq. 3 is Bernstein's inequality to compute the precision for empirical estimates (see e.g. [18, p124]); $\hat{\sigma}_i$ is the empirical estimate of the standard deviation of point x_i's associated random variable $F_t(x_i)$ (it is 0 in the first iteration, which does not alter the algorithm's correctness); $\hat{f}(x)$ is the average of the fitness measurements at x.

$Bernstein(x_1, x_2, x_3, \delta')$
$T = 0$
repeat
  $T \leftarrow T + 1$
  Evaluate the fitness of points x_1, x_2, x_3 once, *i.e.* evaluate the noisy fitness at each of these points.
  Evaluate the precision:

$$\epsilon_{(T)} = 3\log\left(\frac{3\pi^2T^2}{6\delta'}\right)/T + \max_i \hat{\sigma}_i \sqrt{2\log\left(\frac{3\pi^2T^2}{6\delta'}\right)/T}. \quad (3)$$

until Two points $(good, bad)$ satisfy $\hat{f}(bad) - \hat{f}(good) \geq 2\epsilon$
— **return** $(good, bad)$

R-EDA (3)

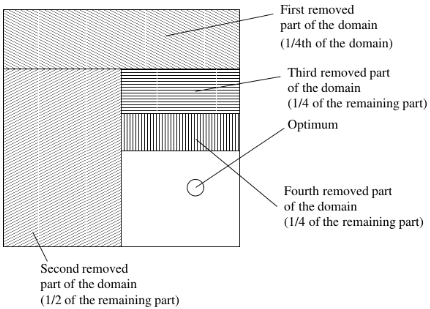


Fig. 1. Noisy optimization algorithm (cf Algorithm 2). At each iteration, a main axis is selected (the one on which the domain has maximum range). Three equally spaced points are generated in the domain on this axis (this is the offspring). Then, a Bernstein race is applied for choosing a "good" and a "bad" arm among these points. The domain is reduced thanks to this knowledge, removing one fourth or one half of the domain (depending on the position of the good arm and of the bad arm - the best case is when the good and the bad arm are diametrically opposed: see Fig. 2).