ThRaSH Workshop 2011

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

Sebastian U. Stich[a,b], Christian L. Müller[a,b], Bernd Gärtner[a]

Institute of Theoretical Computer Science
Department of Computer Science[a], Swiss Institute of Bioinformatics[b]
ETH Zürich

Juli 9, 2011

Table of contents

Black-box scenario

Given: $f : E \to \mathbb{R}$
Goal: $f^* := f(x^*) := \min_{x \in E} f(x)$

$x \to$ [redacted] $\to f(x)$

- 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: $f : E \to \mathbb{R}$
Goal: $f^* := f(x^*) := \min_{x \in E} f(x)$

$x \to$  $\to f(x)$

- 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

Convex functions

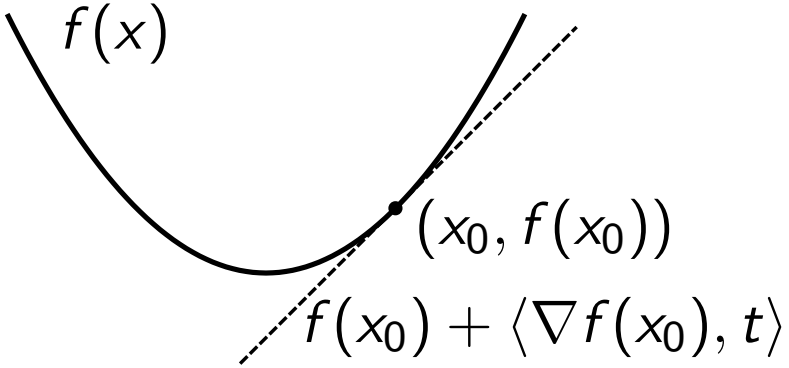
Definition

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

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in E$$



Convex vs. Black-box scenario

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

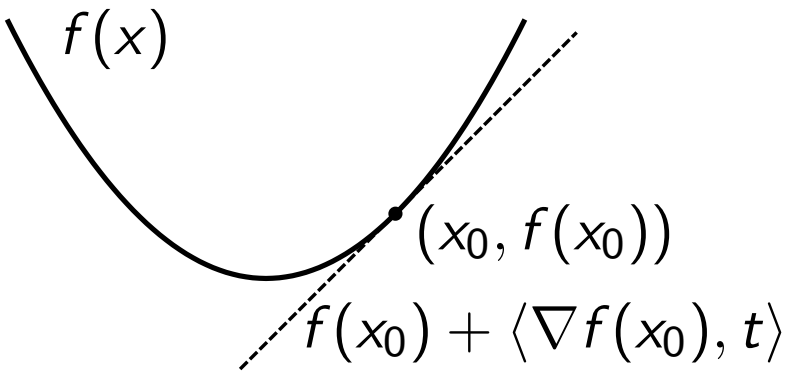
We want to focus on derivative-free optimization.

Strongly convex functions - Quadratic lower bound

Definition

$f \in C^1(E)$ 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

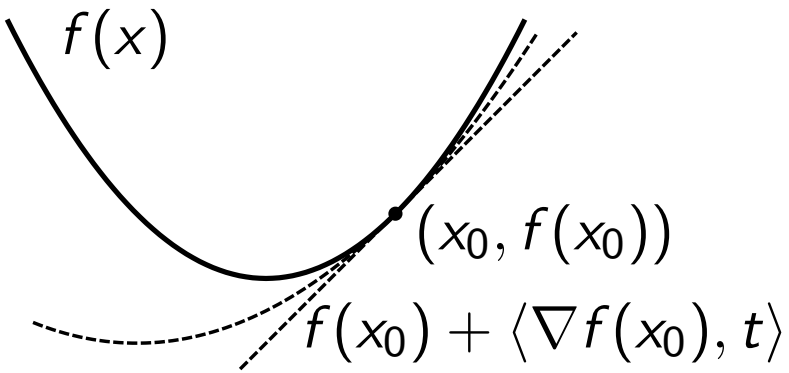
$$\frac{m}{2} \|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^n$$

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

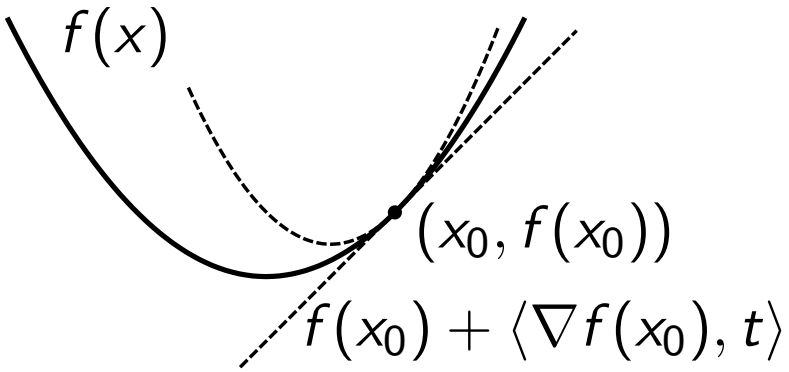
$$\frac{m}{2} \|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^n$$

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

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

$$g_0(x) = f'(x, u) \cdot u \qquad g_\mu(x) = \frac{f(x + \mu u) - f(x)}{\mu} \cdot u$$

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

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

$$g_0(x) = f'(x, u) \cdot u \qquad g_\mu(x) = \frac{f(x + \mu u) - f(x)}{\mu} \cdot u$$

Random gradient free optimization

	$\mathcal{RG}_0$	$\mathcal{RG}_\mu$	$\mathcal{FRG}_\mu$
smooth & strongly convex	$O\left(n\frac{L_1}{m}\ln\frac{L_1R^2}{\epsilon}\right)$		$O\left(n\sqrt{\frac{L_1}{m}}\ln\frac{f_0+R^2}{\epsilon}\right)$
smooth & convex	$O\left(n\frac{L_1R^2}{\epsilon}\right)$		$O\left(n\sqrt{\frac{L_1(f_0+R^2)}{\epsilon}}\right)$
convex	$O\left(n\frac{L_0^2R^2}{\epsilon^2}\right)$	$O\left(n^2\frac{L_0^2R^2}{\epsilon^2}\right)$	-

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

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

$$\begin{aligned} \mathcal{L}_\epsilon \colon \mathbb{R} \times S^{n-1} &\to \mathbb{R} \\ (x, u) &\mapsto h \end{aligned}$$

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

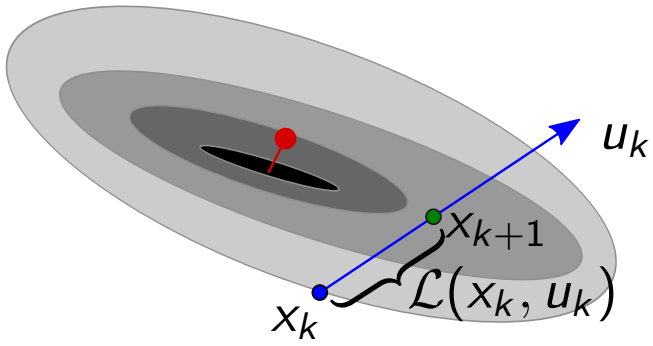
Definition (Line search oracle)

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

$$\mathcal{L}_\epsilon \colon \mathbb{R} \times S^{n-1} \to \mathbb{R}$$
$$(x, u) \mapsto h$$

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).

The setting

Assumptions:

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

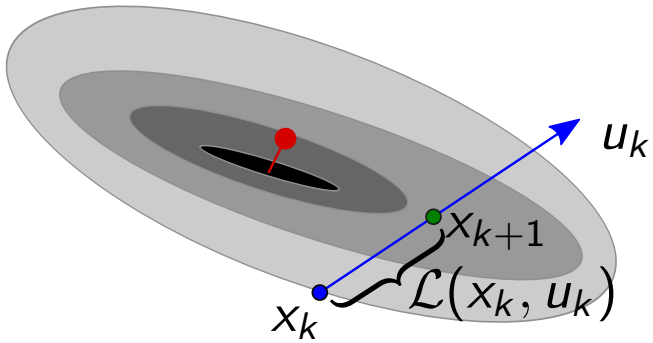
Definition (Line search oracle)

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

$$\mathcal{L}_\epsilon \colon \mathbb{R} \times S^{n-1} \to \mathbb{R}$$
$$(x, u) \mapsto h$$

with

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



Remark

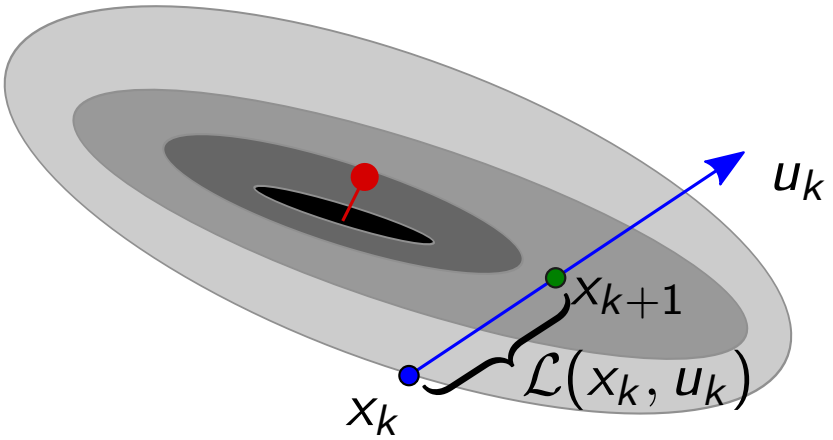
Relative accuracy of $\mathcal{L}$ yields similar results (with different analysis).

Random Pursuit for unconstrained minimization

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

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

No input parameters needed, except x_0.



The analysis

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

$$\begin{aligned} f(x_{k+1}) &= f(x_k + \mathcal{L}_\epsilon(x,u)u) \\ &\leq f(x_k + h^* u) + \epsilon \\ &\overset{\text{optimality of } h^*}{\leq} f(x_k + \frac{1}{L_1 n}\alpha(u)u) + \epsilon \end{aligned}$$

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]

The analysis

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

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

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]

How to choose α

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 $\alpha \colon \mathbb{R}^n \to \mathbb{R}$, s.t.

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

Proof.

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

How to choose α

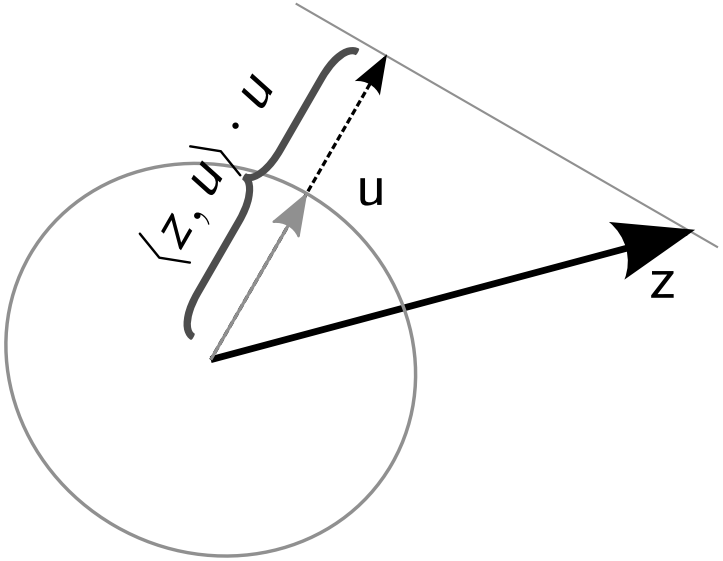
Lemma

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

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

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 $z = x^* - x_k$:

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

For strongly convex functions:

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

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 $z = x^* - x_k$:

$$\begin{aligned}\mathbb{E}\left[f(x_{k+1}) \mid x_t\right] &\leq f(x_k) + \frac{1}{L_1 n} \langle \nabla f(x_k), x^* - x_k \rangle + \frac{1}{L_1 n} \|x^* - x_k\|^2 + \epsilon \\ &\overset{\text{convex}}{\leq} \left(1 - \frac{1}{L_1 n}\right)(f(x_k) - f(x^*)) + \frac{1}{L_1 n} \|x^* - x_k\|^2 + \epsilon\end{aligned}$$

For strongly convex functions:

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

Accelerated method

Algorithm 4 Accelerated Random Pursuit: ($\mathcal{FRP}$)

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 $\lambda_k = \frac{\beta_k}{\gamma_{k+1}} m$, $\delta_k = \frac{\beta_k \gamma_k}{\gamma_k + \beta_k m}$, 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

	$\mathcal{RP}_0$	$\mathcal{RP}_\mu$	$\mathcal{FRP}$
smooth & strongly convex	$O\left(n\frac{L_1}{m}\ln\frac{f_0}{\epsilon}\right)$		-
smooth & convex	$O\left(n\frac{L_1R^2+f_0}{\epsilon}\right)$		-

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

Test function

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

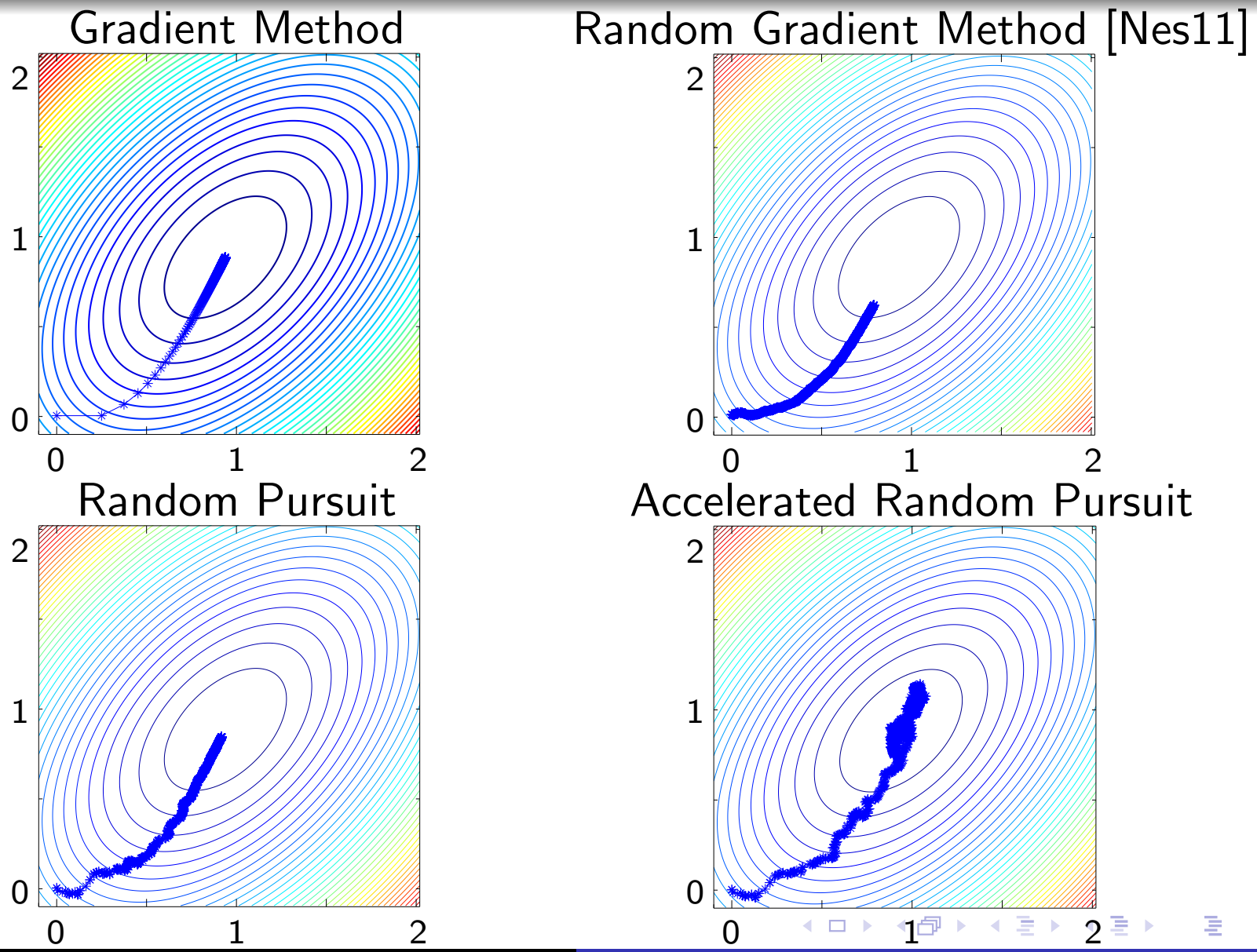
With parameters:

$$dim = 64$$

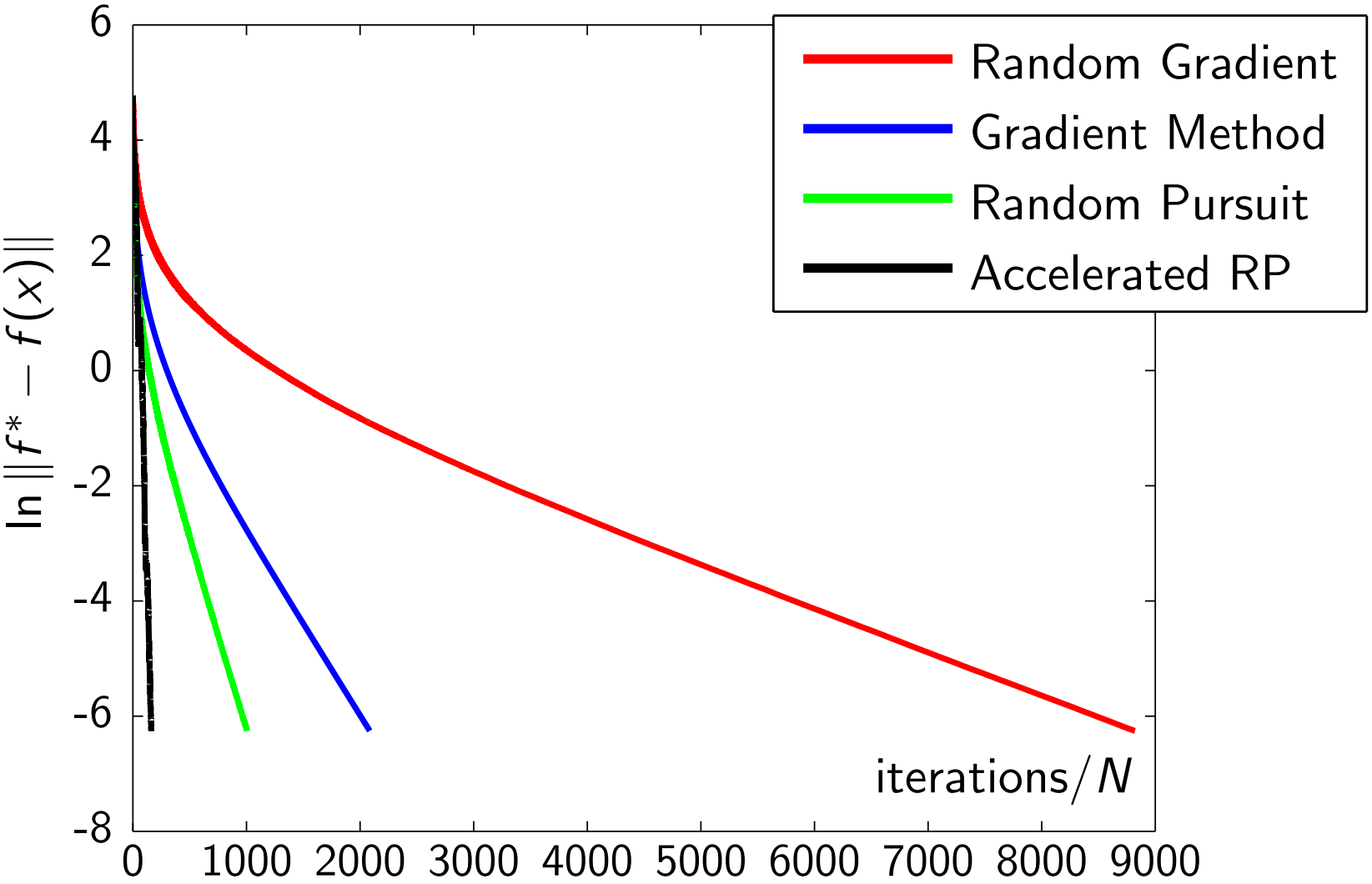
$$L_1 = 1000 \qquad m = 1$$

$$x_0 = 0 \qquad (x^*) = \begin{bmatrix} 0.939 \\ 0.881 \\ \vdots \end{bmatrix}$$

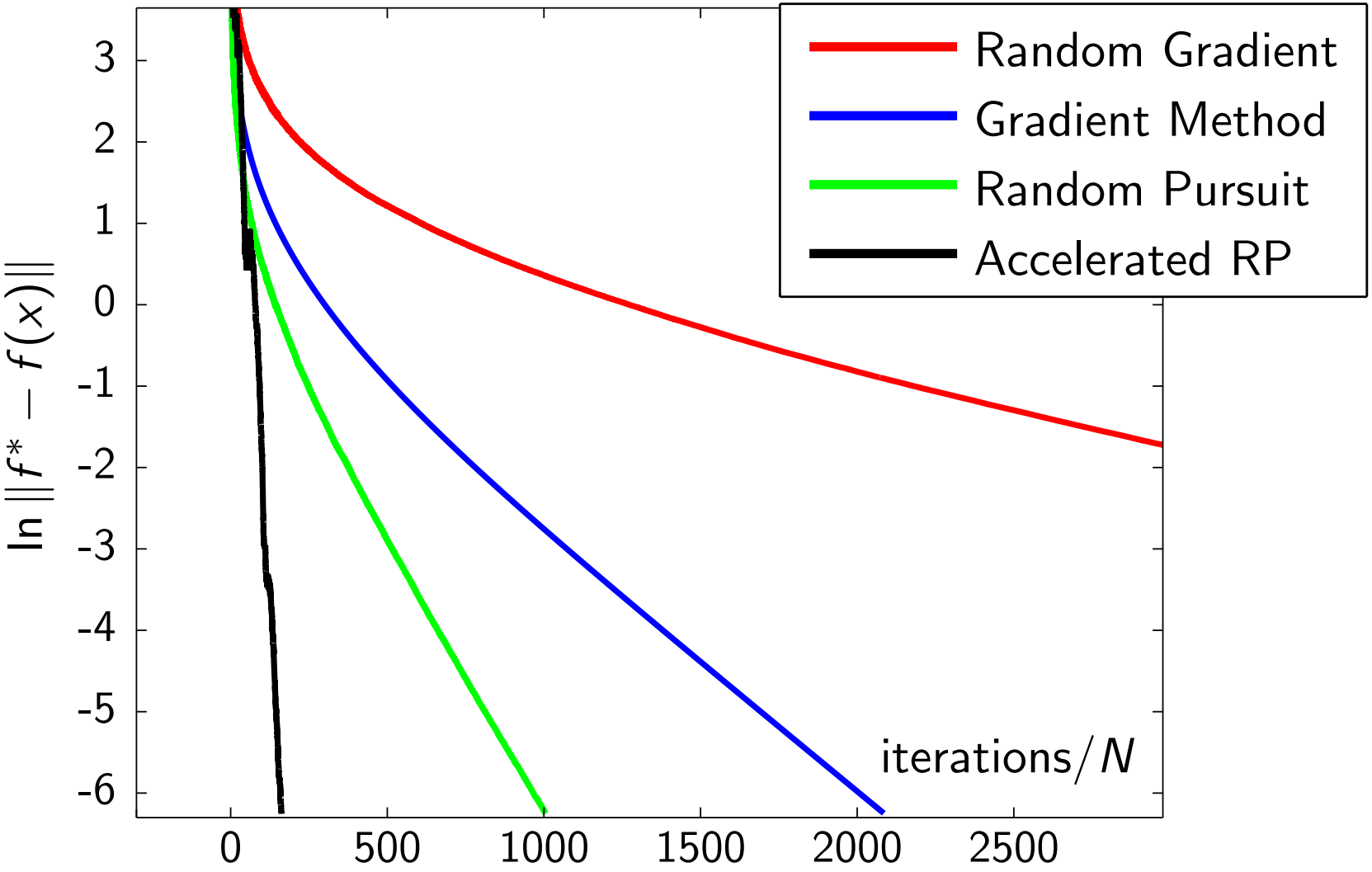
Results - Evolution paths



Results - Convergence plots



Results - Convergence plots



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

References

- Yu. Nesterov, Random Gradient-Free Minimization of Convex Functions, *Ecore Discussion Paper*, 2011.
- A. Kleiner, A. Rahimi, and M. I. Jordan. Random conic pursuit for semidefinite programming. In *Neural Information Processing Systems*, 2010.