Project number IST-25582

CGL

Computational Geometric Learning

Variable Metric Random Pursuit

STREP

Information Society Technologies

Period covered:	November 1, 2011–October 31, 2012
Date of preparation:	October 22, 2012
Date of revision:	October 22, 2012
Start date of project:	November 1, 2010
Duration:	3 years
Project coordinator name:	Joachim Giesen (FSU)
Project coordinator organisation:	Friedrich-Schiller-Universität Jena Jena, Germany

Noname manuscript No.
(will be inserted by the editor)

Variable Metric Random Pursuit

S. U. Stich · C. L. Müller · B. Gärtner

Received: date / Accepted: date

Abstract We consider unconstrained randomized optimization of smooth convex objective functions in the gradient-free setting. We analyze Random Pursuit (RP) algorithms with fixed (F-RP) and variable metric (V-RP). The algorithms only use zeroth-order information about the objective function and compute an approximate solution by repeated optimization over randomly chosen one-dimensional subspaces. The distribution of search directions is dictated by the chosen metric. Variable Metric RP uses novel variants of a randomized zeroth-order Hessian approximation scheme recently introduced by Leventhal and Lewis (D. Leventhal and A. S. Lewis., Optimization 60(3), 329–245, 2011). We here present (i) a refined analysis of the expected single step progress of RP algorithms and their global convergence on (strictly) convex functions and (ii) novel convergence bounds for V-RP on convex quadratic functions. We also quantify how well the employed metric needs to match the local geometry of the function in order for the RP algorithms to converge with the best possible rate on strongly convex functions. Our theoretical results are accompanied by illustrative experiments on Nesterov's worst case function and quadratic functions with (inverse) sigmoidal eigenvalue distributions.

Keywords gradient-free optimization · convex optimization · variable metric · line search

1 Introduction

Since its inception by Davidon in the late 1950's [2] variable metric methods have become a cornerstone in first-order (non-)convex continuous optimization. Among the many instances of variable metric schemes Quasi-Newton methods such as the BFGS scheme [1,3,4,17] are ubiquitous in all areas of science and engineering. In zeroth-order (or gradient-free) optimization, the idea of using a variable metric guiding the search for local or global optima has surprisingly been used to a far less extent. Although "directional adaptation" has been conjectured to be useful for randomized gradient-free schemes in the late 1960's [16] the literature on this topic is scarce and scattered across different communities ranging from electrical

The project CG Learning acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 255827.

S. U. Stich, C. L. Müller, B. Gärtner
Institute of Theoretical Computer Science, ETH Zürich,
E-mail: {stich,christian.mueller,gaertner}@inf.ethz.ch

engineering, optimal control, bio-inspired optimization to mathematical programming. Important examples include the Gaussian Adaptation algorithm developed by Kjellström and Taxen [11,14] in the context of analog circuit design, Marti's controlled random search schemes using concepts from optimal control [13], and the arguably most popular scheme, Hansen's Evolution Strategy with Covariance Matrix Adaptation [5] that emerged in the bio-inspired optimization community.

Despite their great appeal in practice many randomized gradient-free variable metric schemes lack a thorough theoretical convergence analysis. A marked exception is Leventhal and Lewis' recent work on Randomized Hessian approximation [12]. We here adopt some of their ideas and extend our framework of Random Pursuit (RP) [19], eventually leading to Variable Metric Random Pursuit (V-RP) schemes. We solely consider optimization problems of the kind:

$$\min f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \mathbb{R}^n\,, \tag{1}$$

where f is a smooth convex function. We assume that there is a global minimum and that the curvature of the function f is bounded. Moreover, we assume that we have only access to function values of f. No analytic gradient or higher order information about f is available.

To motivate Variable Metric Random Pursuit, let us first sketch the working mechanism of standard Random Pursuit on an illustrative example. Each iteration of standard Random Pursuit consists of two steps: (i) a random direction is sampled from an isotropic probability distribution; (ii) the next iterate is chosen such as to (approximately) minimize the objective function along this direction (using an approximate line search procedure). In [19] we have shown that the expected error in function value decreases by a factor of

$$\left(1 - \frac{m}{n\ell_1}\right)$$

in every step, if $m > 0$ and $\ell_1 > 0$ are parameters of quadratic functions that bound the difference between f and any of its linear approximations from below and above; more precisely,

$$\frac{m}{2}\left\|\mathbf{y}-\mathbf{x}\right\|^2 \leq \ell_{\mathbf{x}}(\mathbf{y}) := f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \frac{\ell_1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|^2 \tag{2}$$

is assumed to hold for all $\mathbf{x}, \mathbf{y}$. As an example, let us consider the function

$$f_0(x_1, x_2) = 100x_1^2 + x_2^2,$$

for which $\ell_{\mathbf{x}}(\mathbf{y}) = 100(x_1 - y_1)^2 + (x_2 - y_2)^2$. This means that $m = 2$ and $\ell_1 = 200$ are the best possible parameters in (2), and the progress rate in every step is no better than $(1 - 1/200)$. This also matches our intuition: every level set of f_0 is a long and skinny ellipse, stretching out along the x_2-axis; if we start from a point close to the x_2 axis, the progress in a step will be small, unless we almost sample in x_2-direction.

For this particular function f_0, it would be better to sample from an anisotropic distribution that favors the x_2-direction. Once we fix such an anisotropic sampling distribution, however, other functions become "bad"; in fact, without prior knowledge about f, anisotropic sampling makes no sense at all. Here is where the "variable metric" approach comes in. The idea is to gradually *adapt* the sampling distribution to the function f while we run the algorithm. Suppose that we can somehow estimate the Hessians at the various iterates. Under the assumption that f is wedged between two quadratic functions—whose Hessians are not necessarily multiples of the identity, as in (2)—these estimates will allow us to learn a suitable metric that guides the sampling distribution. In case of f_0, we would start with the isotropic

one and then converge to a distribution that indeed favors the x_2-direction with the right proportion.

In this contribution we present a framework for analyzing the convergence behavior of Random Pursuit algorithms on convex functions. In a first step we analyze the Fixed Metric Random Pursuit (F-RP) algorithm, a natural extension of Random Pursuit with an arbitrary but fixed anisotropic sampling distribution. We obtain a progress rate that depends on the chosen distribution, as well as on lower and upper quadratic approximations to f. Our progress rate bound improves over the one we would get from the standard Random Pursuit analysis [19], after applying an affine transformation that maps the sampling distribution back to the isotropic one. The improvement is due to the fact that the new analysis takes the whole spectrum of the quadratic upper bound into account rather than just a scalar value ℓ_1.

In a second step we equip Random Pursuit with a randomized scheme to update the metric that defines the sampling distribution in every step: the Variable Metric Random Pursuit. We present two novel variants of an update scheme recently proposed by Leventhal and Lewis [12]. These learning schemes are generic in the sense that they work for all convex functions and do not require any prior knowledge of the function's shape. We then concentrate on the class of convex quadratic functions and rigorously prove in this case that the sampling distribution converges to a distribution that yields asymptotically optimal (and function-independent) progress rates. We also provide bounds on the number of iterations that the algorithm requires in order to attain optimal progress rates with high probability.

The remainder of the paper is structured as follows. In Section 2 we give a generic description of the different Random Pursuit algorithms and their essential building blocks. We introduce all relevant mathematical definitions such as matrix upper and lower bounds of convex functions and expressions for certain scalar and matrix expectations in Section 3. We derive the expected single-step progress and global convergence of F-RP in Section 4. Section 5 is dedicated to Variable Metric Random Pursuit. We derive theoretical convergence results and show a number of illustrative numerical examples. We discuss the key results of the paper and outline future research goals in Section 6.

2 Fixed and Variable Metric Random Pursuit

All Random Pursuit algorithms are designed for problems as defined in (1) where f is assumed to be a differentiable convex function with the property that it has a minimum along every line in $\mathbb{R}^n$. Before stating the formal definition of the considered RP algorithms we need to define two primitives.

Definition 1 (Multivariate normal distribution) The multivariate normal distribution arises from independent and identically distributed (i.i.d.) standard normals. The vector $\mathbf{v} \in \mathbb{R}^n$ is standard multivariate normally distributed, i.e., $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, I_n)$ if $v_i \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, n$ and I_n the identity matrix. For $\boldsymbol{\mu} \in \mathbb{R}^n, \Sigma = CC^T \in \mathrm{PD}_n$, where PD_n is the set of symmetric positive definite matrices, $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance Σ if $\mathbf{u} = \boldsymbol{\mu} + C\mathbf{v}$ and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, I_n)$.

Definition 2 (Line search oracle) For $\mathbf{x} \in \mathbb{R}^n$, a direction $\mathbf{u} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ (not necessarily of unit length), and a convex function f, a function $\mathtt{LS}_f \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with

$$\mathtt{LS}_f(\mathbf{x}, \mathbf{u}) = \arg\min_{h \in \mathbb{R}} f(\mathbf{x} + h\mathbf{u}) \tag{3}$$

is called an *exact line search oracle*. The analysis in [19] shows that an *approximate line search oracle* is sufficient to establish the same asymptotic convergence bounds which is important in practical applications. For simplicity we assume here that we have access to an exact line search oracle.

The two RP schemes considered here are summarized in Fig. 1.

F-RP$(f, \mathbf{x}_0, \Sigma, N)$

Output: Approximate solution x_N to (1)
1 **for** $k = 1$ *to* N **do**
2 $\quad \mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \Sigma)$
3 $\quad \mathbf{x}_k \leftarrow \mathtt{LS}_f(\mathbf{x}_{k-1}, \mathbf{u}_k)$
4 **return** $\mathbf{x}_N$

V-RP$(f, \mathbf{x}_0, B_0, N)$

Output: Approximate solution x_N to (1)
1 **for** $k = 1$ *to* N **do**
2 $\quad B_k \leftarrow \mathtt{updateHess}(f, \mathbf{x}, B_{k-1})$
3 $\quad \mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, B_k^{-1})$
4 $\quad \mathbf{x}_k \leftarrow \mathtt{LS}_f(\mathbf{x}_{k-1}, \mathbf{u}_k)$
5 **return** $\mathbf{x}_N$

Fig. 1 Fixed Metric Random Pursuit (left panel) and the Variable Metric version (right panel). The generic sub-routine `updateHess` on line 2 exemplifies any function that generates the metric B_k in step k. Two specific instantiations are discussed in Section 5 (cf. Fig. 2).

In Fixed Metric Random Pursuit (F-RP) a direction $\mathbf{u} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ is sampled from a multivariate normal distribution with fixed covariance Σ at iteration k of the algorithm. The next iterate $\mathbf{x}_k$ is calculated from the current iterate $\mathbf{x}_{k-1}$ as

$$\mathbf{x}_k := \mathbf{x}_{k-1} + \mathtt{LS}_f(\mathbf{x}_{k-1}, \mathbf{u}) \cdot \mathbf{u}. \tag{4}$$

This algorithm only requires function evaluations on top of the exact line search oracle. Note that an approximate line search oracle can efficiently be implemented with function evaluations and binary search. No additional first or second-order information about the objective is needed. Besides the starting point no further input parameters describing function properties (such as curvature constant etc.) are necessary. The actual run time will, however, depend on the specific properties of the objective function.

Variable Metric Random Pursuit (V-RP) comprises an independent process that gives an approximation of the Hessian at each iteration. The inverse of the Hessian is then used as covariance matrix in the multivariate normal distribution to generate the current search direction. In principle, any deterministic or randomized gradient-free estimator can be used for this purpose. In Section 5 we present two variants of a Randomized Hessian approximation scheme recently proposed in [12] for which theoretical guarantees are known.

3 Definitions and Notations

3.1 Quadratic norms

Let PD_n denote the set of symmetric positive definite $n \times n$ matrices. With respect to $A \in \mathrm{PD}_n$, we can define an 'anisotropic' inner product and a corresponding norm by

$$\langle \mathbf{x}, \mathbf{y} \rangle_A := \mathbf{y}^T A \mathbf{x}\,, \qquad \text{and} \qquad \|\mathbf{x}\|_A^2 := \langle \mathbf{x}, \mathbf{x} \rangle_A\,,$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. In statistics the induced metric is also known as the Mahalanobis metric. We observe that

$$\lambda_{\min}(A)\,\|\mathbf{x}\|^2 \leq \|\mathbf{x}\|_A^2 \leq \lambda_{\max}(A)\,\|\mathbf{x}\|^2\,, \tag{5}$$

due to $\lambda_{\min}(A) = \min\{\mathbf{x}^T A \mathbf{x} : \|x\| = 1\}$ and $\lambda_{\max}(A) = \max\{\mathbf{x}^T A \mathbf{x} : \|x\| = 1\}$. We will frequently need the following lemma.

Lemma 1 *Let $A, B \in \mathrm{PD}_n, \mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} \neq \mathbf{0}$. Then*

$$\lambda_{\min}(B^{-1}A) \leq \frac{\|\mathbf{x}\|_A}{\|\mathbf{x}\|_B} \leq \lambda_{\max}(B^{-1}A)\,.$$

Proof Suppose that $B = CC^T$ and set $\mathbf{y} = C^T\mathbf{x}$. Then we have

$$\frac{\|\mathbf{x}\|_A}{\|\mathbf{x}\|_B} = \frac{\|\mathbf{y}\|_{C^{-1}AC^{-T}}}{\|\mathbf{y}\|}\,,$$

where C^{-T} is a shorthand for $(C^T)^{-1}$. Using (5), we can argue that

$$\lambda_{\min}(C^{-1}AC^{-T}) \leq \frac{\|\mathbf{y}\|_{C^{-1}AC^{-T}}}{\|\mathbf{y}\|} \leq \lambda_{\max}(C^{-1}AC^{-T})\,.$$

It remains to show that the matrices $C^{-1}AC^{-T}$ and $C^{-T}C^{-1}A = B^{-1}A$ have the same eigenvalues. This follows from the "Rotation" Lemma 2. □

Lemma 2 (Rotation) *Let $L \in \mathbb{R}^{n\times n}$, and let $C \in \mathbb{R}^{n\times n}$ be an invertible matrix. Then the two matrices*

$$P := LCC^T \text{ and } Q := C^TLC$$

have the same eigenvalues.

Proof We show that P and Q have the same characteristic polynomial. For this, we first observe that

$$P - tI_n = C^{-T}(Q - tI_n)C^T.$$

It follows that

$$\begin{aligned}\det(P - tI_n) &= \det(C^{-T}(Q - tI_n)C^T) = \det(C^{-T})\det(C^T)\det(Q - tI_n)\\ &= \det(Q - tI_n)\,. \qquad \square\end{aligned}$$

3.2 Quadratic bounds

We now introduce some important inequalities that are useful for the subsequent analysis. We always assume that the objective function f is differentiable and convex. The latter property is equivalent to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle\,, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n\,. \tag{6}$$

We also require that the curvature of f is bounded. However, we allow for different curvatures depending on the direction. By this we mean that for some fixed symmetric and positive definite matrix $L_1 \in \mathrm{PD}_n$,

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \leq \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_{L_1}^2\,, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n\,. \tag{7}$$

We will also refer to this inequality as the *(matrix) quadratic upper bound*. It means that the deviation of f from any of its linear approximations can be bounded by a quadratic function.

A differentiable function is *strongly convex* with parameter $M \in \mathrm{PD}_n$ if the *(matrix) quadratic lower bound*

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \geq \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_M^2\,, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n\,, \tag{8}$$

holds. Let $\mathbf{x}^*$ be the unique minimizer of a strongly convex function f with parameter M. Then equation (8) implies this useful relation:

$$\frac{1}{2}\left\|\mathbf{x}-\mathbf{x}^*\right\|_M^2 \leq f(\mathbf{x})-f(\mathbf{x}^*) \leq \frac{1}{2}\|\nabla f(\mathbf{x})\|_{M^{-1}}^2\,, \quad \forall \mathbf{x}\in\mathbb{R}^n\,. \tag{9}$$

The former inequality uses $\nabla f(\mathbf{x}^*)=0$, and the latter one follows from (8) via

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^*-\mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{x}^*-\mathbf{x}\right\|_M^2 \\ &\geq f(\mathbf{x}) + \min_{\mathbf{y}\in\mathbb{R}^n}\left(\langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_M^2\right) = f(\mathbf{x}) - \frac{1}{2}\|\nabla f(\mathbf{x})\|_{M^{-1}}^2 \end{aligned}$$

by standard calculus.

3.3 Expectations involving normally distributed random variables

We here review and derive certain facts about the moments of the standard normal distribution for the later convergence analysis.

Lemma 3 *Let $\mathbf{u}\in\mathcal{N}(\mathbf{0},\Sigma)$ be drawn from the multivariate normal distribution over $\mathbb{R}^n$ with covariance $\Sigma\in\mathrm{PD}_n$.*

(i) Then for all indices i,j,k,l,

$$\mathbb{E}[u_i]=0\,, \qquad \mathbb{E}[u_iu_j]=\Sigma_{ij}\,, \qquad \mathbb{E}\left[\mathbf{u}\mathbf{u}^T\right]=\Sigma\,,$$

and

$$\mathbb{E}[u_iu_ju_ku_l]=\Sigma_{ij}\Sigma_{kl}+\Sigma_{ik}\Sigma_{jl}+\Sigma_{il}\Sigma_{jk}\,.$$

(ii) Let $A\in\mathrm{SYM}_n$ be a symmetric $n\times n$ matrix. Then

$$\mathbb{E}\left[\mathbf{u}^TA\mathbf{u}\right]=\mathrm{Tr}[A\Sigma]\,, \qquad \mathbb{E}\left[\mathbf{u}^TA\mathbf{u}\cdot\mathbf{u}\mathbf{u}^T\right]=\mathrm{Tr}[A\Sigma]\Sigma+2\Sigma A\Sigma\,.$$

Proof The first three equalities in part (i) are easy consequences of the above definition. The last one is known as *Isserlis' Theorem* [9].

The matrix equations in (ii) easily follow from (i) via

$$\mathbb{E}\left[\mathbf{u}^TA\mathbf{u}\right]=\sum_{i,j=1}^n\mathbb{E}\left[u_iu_jA_{ij}\right]=\sum_{i,j=1}^n\Sigma_{ij}A_{ij}=\mathrm{Tr}[A^T\Sigma]\,,$$

and

$$\begin{aligned} \left(\mathbb{E}\left[\mathbf{u}^TA\mathbf{u}\cdot\mathbf{u}\mathbf{u}^T\right]\right)_{ij} &= \sum_{k,l=1}^n\mathbb{E}\left[u_iu_ju_ku_lA_{kl}\right] \\ &= \sum_{k,l=1}^n A_{kl}\left(\Sigma_{ij}\Sigma_{kl}+\Sigma_{ik}\Sigma_{jl}+\Sigma_{il}\Sigma_{jk}\right) \\ &= \mathrm{Tr}[A\Sigma]\Sigma_{ij}+\left(\Sigma A\Sigma\right)_{ij}+\left(\Sigma A\Sigma\right)_{ji}\,, \end{aligned}$$

using symmetry of $\Sigma A\Sigma$. □

Lemma 4 *Let $\mathbf{v}\sim S^{n-1}$ a random unit vector and let $A\in\mathrm{SYM}_n$. Then*

$$\mathbb{E}\left[\mathbf{v}^TA\mathbf{v}\cdot\mathbf{v}\mathbf{v}^T\right]=\frac{\mathrm{Tr}[A]I_n+2A}{n(n+2)}\,.$$

Proof Let $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, I_n)$. We observe that the random vector $\mathbf{w} = \mathbf{u}/\left\|\mathbf{u}\right\|$ has the same distribution as $\mathbf{v}$. In particular

$$\mathbb{E}_{\mathbf{v}}\left[\mathbf{v}^T A\mathbf{v} \cdot \mathbf{v}\mathbf{v}^T\right] = \mathbb{E}_{\mathbf{u}}\left[\frac{\mathbf{u}^T A\mathbf{u} \cdot \mathbf{u}\mathbf{u}^T}{\left\|\mathbf{u}\right\|^4}\right] = \frac{\mathbb{E}_{\mathbf{u}}\left[\mathbf{u}^T A\mathbf{u} \cdot \mathbf{u}\mathbf{u}^T\right]}{\mathbb{E}_{\mathbf{u}}\left[\left\|\mathbf{u}\right\|^4\right]} = \frac{\mathrm{Tr}[A] I_n + 2A}{n(n+2)}\,,$$

where the second equality follows from independence of $\frac{\mathbf{u}^T A\mathbf{u} \cdot \mathbf{u}\mathbf{u}^T}{\left\|\mathbf{u}\right\|^4}$ and $\left\|\mathbf{u}\right\|^4$ [6]. Using Lemma 3 (ii) with $A = \Sigma = I_n$, we get

$$\mathbb{E}\left[\left\|\mathbf{u}\right\|^4\right] = \mathrm{Tr}\left[\mathbb{E}\left[\mathbf{u}^T\mathbf{u} \cdot \mathbf{u}\mathbf{u}^T\right]\right] = n(n+2)\,. \qquad \square$$

We finally need the following lemma on the expectation of certain scalar products:

Lemma 5 *Let $\mathbf{u} \in \mathcal{N}(\mathbf{0}, \Sigma)$ be drawn from the multivariate normal distribution over $\mathbb{R}^n$ with covariance $\Sigma \in \mathrm{PD}_n$, let $A \in \mathrm{PD}_n$ and $\mathbf{x} \in \mathbb{R}^n$. Then*

$$\mathbb{E}\left[\langle \mathbf{x}, \mathbf{u}\rangle\, \mathbf{u}\right] = \Sigma\mathbf{x}\,, \quad \textit{and} \quad \mathbb{E}\left[\left\|\langle \mathbf{x}, \mathbf{u}\rangle\, \mathbf{u}\right\|_A^2\right] = \mathrm{Tr}[A\Sigma]\left\|\mathbf{x}\right\|_\Sigma^2 + 2\left\|\mathbf{x}\right\|_{\Sigma A\Sigma}^2\,.$$

Proof We calculate

$$\mathbb{E}\left[\langle \mathbf{x}, \mathbf{u}\rangle\, \mathbf{u}\right] = \mathbb{E}\left[\mathbf{u}\mathbf{u}^T\mathbf{x}\right] = \mathbb{E}\left[\mathbf{u}\mathbf{u}^T\right]\mathbf{x} = \Sigma\mathbf{x}\,,$$

by Lemma 3 (i). For the second moment we get

$$\mathbb{E}\left[\left\|\langle \mathbf{x}, \mathbf{u}\rangle\, \mathbf{u}\right\|_A^2\right] = \mathbb{E}[\mathbf{x}^T\mathbf{u}\mathbf{u}^T A\mathbf{u}\mathbf{u}^T\mathbf{x}] = \mathbf{x}^T\,\mathbb{E}[\mathbf{u}\mathbf{u}^T A\mathbf{u}\mathbf{u}^T]\mathbf{x}\,,$$

and the claim follows by Lemma 3 (ii), observing that $\mathbf{u} \cdot \mathbf{u}^T A\mathbf{u} \cdot \mathbf{u}^T = \mathbf{u}^T A\mathbf{u} \cdot \mathbf{u}\mathbf{u}^T$. $\square$

4 Convergence of Fixed Metric Random Pursuit

To prepare the convergence proof of Algorithm F-RP, we study the expected progress in a single step, which is the quantity

$$f(\mathbf{x}_k) - \mathbb{E}\left[f(\mathbf{x}_{k+1}) \mid \mathbf{x}_k\right]\,.$$

We will now derive the global convergence rates for convex and strongly convex functions.

4.1 Single step progress

Once a search direction is determined, the subsequent iterate is chosen according to Equation 4. Because the step size is determined by the line search oracle, we can derive the following lower bound on the single step progress.

Lemma 6 *Let f be convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}_+$ be the next iterate after one step of Algorithm F-RP in direction $\mathbf{u} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, cf. Equation 4. Then*

$$f(\mathbf{x}_+) \leq f(\mathbf{x}) + t\,\langle \nabla f(\mathbf{x}), \mathbf{u}\rangle + \frac{t^2}{2}\left\|\mathbf{u}\right\|_{L_1}^2\,, \tag{10}$$

for every $t \in \mathbb{R}$.

Proof This is a simple consequence of Definition 2. For every t it holds

$$\begin{aligned} f(\mathbf{x}_+) &= f(\mathbf{x} + \mathtt{LS}_f(\mathbf{x}, \mathbf{u}) \cdot \mathbf{u}) \\ &\leq f(\mathbf{x} + t\mathbf{u}) \,, \end{aligned}$$

and the claim follows from the quadratic upper bound (7) with $\mathbf{y} = \mathbf{x} + t\mathbf{u}$. □

In order to get a bound on the expected single step progress, let us fix a (possibly suboptimal) step size $t = -h \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle$ with (yet to determine) parameter $h \in \mathbb{R}$. With this step size the single step progress can be calculated.

Lemma 7 *Let f we convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}_+$ be the next iterate after one step of Algorithm F-RP with direction $\mathbf{u} \sim \mathcal{N}(0, \Sigma)$, eg. Equation 4. Then*

$$\mathbb{E}\left[f(\mathbf{x}_+) \mid \mathbf{x}\right] \leq f(\mathbf{x}) - h \left\|\nabla f(\mathbf{x})\right\|_\Sigma^2 \tag{11}$$

$$+ \frac{h^2}{2} \left(\mathrm{Tr}\left[L_1 \Sigma\right] \left\|\nabla f(\mathbf{x})\right\|_\Sigma^2 + 2 \left\|\nabla f(\mathbf{x})\right\|_{\Sigma L_1 \Sigma}^2 \right) , \tag{12}$$

for every $h \in \mathbb{R}$.

Proof Let $t = -h \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle$ in (10). Then

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{x}_+) \mid \mathbf{x}\right] &\leq f(\mathbf{x}) - h\mathbb{E}\left[\langle \nabla f(\mathbf{x}), \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle \cdot \mathbf{u} \rangle\right] \\ &\quad + \frac{h^2}{2} \mathbb{E}\left[\left\| \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle \cdot \mathbf{u} \right\|_{L_1}^2 \right] \\ &= f(\mathbf{x}) - h \left\langle \nabla f(\mathbf{x}), \Sigma \nabla f(\mathbf{x}) \right\rangle \\ &\quad + \frac{h^2}{2} \left(\mathrm{Tr}\left[L_1 \Sigma\right] \left\|\nabla f(\mathbf{x})\right\|_\Sigma^2 + 2 \left\|\nabla f(\mathbf{x})\right\|_{\Sigma L_1 \Sigma}^2 \right) , \end{aligned}$$

with Lemma 5. □

Now we are ready to eliminate the last free parameter h in our bound on the single step progress. The best choice of the parameter is obviously the one that minimizes the right hand side of (11). By taking the derivative with respect to h, we see that h must satisfy

$$- \left\|\nabla f(\mathbf{x})\right\|_\Sigma^2 + h \left(\mathrm{Tr}\left[L_1 \Sigma\right] \left\|\nabla f(\mathbf{x})\right\|_\Sigma^2 + 2 \left\|\nabla f(\mathbf{x})\right\|_{\Sigma L_1 \Sigma}^2 \right) = 0 \,.$$

With this choice of h, we readily obtain our final bound on the single step progress.

Lemma 8 *Let f be convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, $\mathbf{x} \in \mathbb{R}^n$ such that $\nabla f(\mathbf{x}) \neq \mathbf{0}$, and let $\mathbf{x}_+$ be the next iterate after one step of Algorithm (V)RP with direction $\mathbf{u} \sim \mathcal{N}(0, \Sigma)$, cf. Equation 4. Then*

$$\mathbb{E}\left[f(\mathbf{x}_+) \mid \mathbf{x}\right] \leq f(\mathbf{x}) - \frac{1}{2\mathrm{Tr}\left[L_1 \Sigma\right] + 4\sigma(\mathbf{x})} \left\|\nabla f(\mathbf{x})\right\|_\Sigma^2 , \tag{13}$$

where $\sigma(\mathbf{x}) := \frac{\left\|\nabla f(\mathbf{x})\right\|_{\Sigma L_1 \Sigma}^2}{\left\|\nabla f(\mathbf{x})\right\|_\Sigma^2}$.

This lemma shows that, on average, there is progress in every single step if $\left\|\nabla f(\mathbf{x})\right\|_\Sigma$ is bounded away from zero.[1] This can be assured for all strongly convex functions, but not for convex functions in general. For this reason, we also derive a slightly different bound on the one-step progress.

[1] Here we use that $\mathrm{Tr}\left[L_1 \Sigma\right] > 0$; see the remark after Theorem 2.

Lemma 9 *Let f be convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}_+$ be the next iterate after one step of Algorithm F-RP with direction $\mathbf{u} \sim \mathcal{N}(0, \Sigma)$ (see Eq. 4). In addition, let $\mathbf{x}^* \in \mathbb{R}^n$ be one of the minimizers of f. Then, for every positive $h \geq 0$ it holds that*

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_+) - f(\mathbf{x}^*) \mid \mathbf{x}\right] \leq (1-h)\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) \\ + \frac{h^2}{2}\left(\frac{\mathrm{Tr}\left[L_1\Sigma\right]}{\lambda_{\min}(\Sigma)} + 2\lambda_{\max}(L_1)\right)\left\|\mathbf{x} - \mathbf{x}^*\right\|^2 .\end{aligned} \tag{14}$$

Proof Let $t = -h\left\langle \Sigma^{-1}(\mathbf{x} - \mathbf{x}^*), \mathbf{u}\right\rangle$ in (10). Then, similarly to the proof of Lemma 7 we derive

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_+) \mid \mathbf{x}\right] \leq f(\mathbf{x}) - h\left\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^*\right\rangle \\ + \frac{h^2}{2}\left(\mathrm{Tr}\left[L_1\Sigma\right]\left\|\mathbf{x} - \mathbf{x}^*\right\|_{\Sigma^{-1}}^2 + 2\left\|\mathbf{x} - \mathbf{x}^*\right\|_{L_1}^2\right) ,\end{aligned} \tag{15}$$

with Lemma 5. Using the definition of convexity (6) we can bound the term $\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^*\rangle$ from below by $f(\mathbf{x}) - f(\mathbf{x}^*)$. By subtracting $f(\mathbf{x}^*)$ on both sides of (15) we thus arrive at

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_+) - f(\mathbf{x}^*) \mid \mathbf{x}\right] \leq (1-h)\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) \\ + \frac{h^2}{2}\left(\mathrm{Tr}\left[L_1\Sigma\right]\left\|\mathbf{x} - \mathbf{x}^*\right\|_{\Sigma^{-1}}^2 + 2\left\|\mathbf{x} - \mathbf{x}^*\right\|_{L_1}^2\right) .\end{aligned} \tag{16}$$

By bounding the last two terms from above with (5), we obtain (14), using $\lambda_{\max}(\Sigma^{-1}) = 1/\lambda_{\min}(\Sigma)$. □

4.2 Global convergence

We now use the previously derived bounds on the expected single step progress (Lemma 8 and Lemma 9) to show convergence of F-RP in expectation. We first show convergence on smooth but not necessarily strongly convex functions.

Theorem 1 *Let f be convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, let $\mathbf{x}^* \in \mathbb{R}^n$ be a minimizer of f and let the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ be generated by Algorithm F-RP with covariance $\Sigma \in \mathrm{PD}_n$. Assume there exists $R \in \mathbb{R}$, s.t. $\|\mathbf{y} - \mathbf{x}_0\| \leq R$ for all $\mathbf{y} \in \mathbb{R}^n$ with $f(\mathbf{y}) \leq f(\mathbf{x}_0)$. Then, for any $N \geq 0$, we have*

$$\mathbb{E}\left[f(\mathbf{x}_N) - f(\mathbf{x}^*)\right] \leq \frac{Q}{N+1} , \tag{17}$$

where

$$Q := \max\left\{2\omega(L_1, \Sigma), f(\mathbf{x}_0) - f(\mathbf{x}^*)\right\} \quad , \quad \omega(L_1, \Sigma) := \frac{\mathrm{Tr}\left[L_1\Sigma\right]}{\lambda_{\min}(\Sigma)} + 2\lambda_{\max}(L_1) .$$

Proof By assumption, $\|\mathbf{x}_k - \mathbf{x}^*\| \leq R$ for all $k = 0, 1, \ldots, N$. With Lemma 9 it follows for any step size $h_k \geq 0$:

$$\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \mid \mathbf{x}_k\right] \leq (1-h_k)\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right) + h_k^2\frac{R^2\omega(L_1, \Sigma)}{2} . \tag{18}$$

Taking expectations over $\mathbf{x}_k$, we obtain

$$\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right] \leq (1-h_k)\,\mathbb{E}\left[f(\mathbf{x}_k) - f(\mathbf{x}^*)\right] + h_k^2\frac{R^2\omega(L_1, \Sigma)}{2} . \tag{19}$$

As in [19, Theorem 5.3], the choice $h_k := \frac{2}{k+1}$ for $k = 0, 1, \ldots, (N-1)$ yields

$$\mathbb{E}\left[f(\mathbf{x}_N) - f(\mathbf{x}^*)\right] \leq \frac{4}{N+1} \cdot \frac{R^2\omega(L_1, \Sigma)}{2} . \qquad \square \tag{20}$$

On strongly convex function the convergence of F-RP is linear.

Theorem 2 *Let f be convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, and let f in addition be strongly convex with parameter $M \in \mathrm{PD}_n$. Let $\mathbf{x}^* \in \mathbb{R}^n$ denote the unique minimizer of f and let the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ be generated by Algorithm F-RP with covariance $\Sigma \in \mathrm{PD}_n$. Then*

$$\mathbb{E}\left[f(\mathbf{x}_N) - f(\mathbf{x}^*)\right] \leq \left(1 - \frac{\lambda_{\min}(M\Sigma)}{\mathrm{Tr}[L_1\Sigma] + 2\lambda_{\max}(L_1\Sigma)}\right)^N \cdot \left(f(\mathbf{x}_0) - f(\mathbf{x}^*)\right) . \quad (21)$$

Proof We use Lemma 1 to bound

$$\sigma(\mathbf{x}_k) = \frac{\|\nabla f(\mathbf{x}_k)\|^2_{\Sigma L_1 \Sigma}}{\|\nabla f(\mathbf{x}_k)\|^2_{\Sigma}} \leq \lambda_{\max}(\Sigma^{-1}\Sigma L_1 \Sigma) = \lambda_{\max}(L_1\Sigma)$$

for $k = 0, 1, \ldots, N-1$. Thus Lemma 8 yields

$$\mathbb{E}\left[f(\mathbf{x}_{k+1}) \mid \mathbf{x}_k\right] \leq f(\mathbf{x}_k) - \left(\frac{1}{2\mathrm{Tr}[L_1\Sigma] + 4\lambda_{\max}(L_1\Sigma)}\right) \|\nabla f(\mathbf{x}_k)\|^2_{\Sigma} , \quad (22)$$

for $k = 0, \ldots, N-1$. Using Lemma 1 again, we get

$$\|\nabla f(\mathbf{x}_k)\|^2_{\Sigma} \geq \lambda_{\min}(M\Sigma) \|\nabla f(\mathbf{x}_k)\|^2_{M^{-1}} .$$

Applying the quadratic lower bound (9) to further bound the latter term from below yields

$$\|\nabla f(\mathbf{x}_k)\|^2_{\Sigma} \geq 2\lambda_{\min}(M\Sigma)\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right) .$$

Now, plugging this into (22) and subtracting $f(\mathbf{x}^*)$ on both sides yields

$$\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \mid \mathbf{x}_k\right] \leq \left(1 - \frac{\lambda_{\min}(M\Sigma)}{\mathrm{Tr}[L_1\Sigma] + 2\lambda_{\max}(L_1\Sigma)}\right) \cdot \left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right) ,$$

for $k = 0, \ldots, N-1$. Finally, taking expectation (over $\mathbf{x}_k$) yields the recurrence

$$\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right] \leq \left(1 - \frac{\lambda_{\min}(M\Sigma)}{\mathrm{Tr}[L_1\Sigma] + 2\lambda_{\max}(L_1\Sigma)}\right) \cdot \mathbb{E}\left[f(\mathbf{x}_k) - f(\mathbf{x}^*)\right] ,$$

and the theorem follows. □

We remark that the progress is strict: With $\Sigma = CC^T$, "Rotation" Lemma 2 in Section 3 along with Sylvester's law of inertia yields that all the three terms $\lambda_{\min}(M\Sigma)$,$\mathrm{Tr}[L_1\Sigma]$ and $\lambda_{\max}(L_1\Sigma)$ are positive.

It is not necessary that the function f is strongly convex everywhere for linear convergence to hold. Let us recall that strong convexity with parameter M implies that

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|^2_M , \forall \mathbf{x} \in \mathbb{R}^n . \quad (23)$$

It turns out that, instead of strong convexity (8), the weaker condition (23) is enough for linear convergence.

Theorem 3 *Let f be convex with quadratic upper bound $L_1 \in \mathrm{PD}_n$, and let f have unique minimizer $\mathbf{x}^* \in \mathbb{R}^n$ satisfying (23) with $M \in \mathrm{PD}_n$. Let the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ be generated by Algorithm F-RP with covariance $\Sigma \in \mathrm{PD}_n$. Then*

$$\mathbb{E}\left[f(\mathbf{x}_N) - f(\mathbf{x}^*)\right] \leq \left(1 - \frac{1}{4\theta}\right)^N \cdot \left(f(\mathbf{x}_0) - f(\mathbf{x}^*)\right) , \quad (24)$$

where

$$\theta = \frac{\mathrm{Tr}\left[L_1\Sigma\right]}{\lambda_{\min}(M\Sigma)} + \frac{2}{\lambda_{\min}(ML_1^{-1})} .$$

Proof To see this, we use (16) to get

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_+) - f(\mathbf{x}^*) \mid \mathbf{x}\right] &\leq (1-h)\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) \\ &\quad + \frac{h^2}{2}\left(\operatorname{Tr}\left[L_1\Sigma\right]\|\mathbf{x}-\mathbf{x}^*\|^2_{\Sigma^{-1}} + 2\,\|\mathbf{x}-\mathbf{x}^*\|^2_{L_1}\right) \\ &\leq \left(1 - h + h^2\theta\right)\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right)\,,\end{aligned}$$

where we used Lemma 1 to bound

$$\|\mathbf{x}-\mathbf{x}^*\|^2_{\Sigma^{-1}} \leq \|\mathbf{x}-\mathbf{x}^*\|^2_M \cdot \frac{1}{\lambda_{\min}(M\Sigma)}$$

and

$$\|\mathbf{x}-\mathbf{x}^*\|^2_{L_1} \leq \|\mathbf{x}-\mathbf{x}^*\|^2_M \cdot \frac{1}{\lambda_{\min}(ML_1^{-1})}\,.$$

followed by (23).

Setting h to $\frac{1}{2\theta}$ the term in the left bracket becomes $\left(1 - \frac{1}{4\theta}\right)$ and the proof continues as the proof of Theorem 2. □

5 Metric Learning in Random Pursuit

Our bounds on the progress rate of F-RP with fixed covariance matrix Σ concisely describe the influence of the matrix upper and lower bounds and the chosen covariance matrix on the convergence rate of RP algorithms. For instance, for the special case of quadratic functions of the form $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x}$ with $H \in \mathrm{PD}_n$ and trivial quadratic upper and lower bound $L_1 = M = H$, the choice $\Sigma = H^{-1}$ leads to the expected progress rate $\left(1 - \frac{1}{n+2}\right)$ (cf. Thm 2). This rate is (i) near-optimal from a theoretical point of view [10] and (ii) independent of the function f (i.e., the spectrum of H). For general strongly convex functions $f\colon \mathbb{R}^n \to \mathbb{R}$, the Hessian $\nabla^2 f(\mathbf{x})$ is not constant for all $\mathbf{x} \in \mathbb{R}^n$. However, if we assume that the Hessian is only mildly changing (for instance, Lipschitz continuous) then for all $\mathbf{x} \in \mathbb{R}^n$ that are close to the unique minimizer $\mathbf{x}^* \in \mathbb{R}^n$ the corresponding Hessians will also be close, meaning that $\nabla^2 f(\mathbf{x}) \approx \nabla^2 f(\mathbf{x}^*)$ in some norm. The choice $\Sigma^{-1} = \nabla^2 f(\mathbf{x}^*)$ is thus likely to also yield a good convergence rate on this function class.

We are now left with the challenge of how to efficiently learn a suitable covariance matrix (that induces the right metric) on smooth convex functions in the present gradient-free setting. Iterative stochastic covariance matrix adaptation schemes are well-established in gradient-free continuous optimization [11,5,14] but notoriously difficult to study theoretically. A welcome alternative has recently been introduced by Leventhal and Lewis [12] in form of a Randomized Hessian approximation scheme. We here review and extend their scheme which we refer to as Variable Metric (VM) update. For twice differentiable functions $f\colon \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$ and initial iterate $B_0 \in \mathrm{PD}_n$, Leventhal and Lewis already showed that the VM scheme generates a random sequence $\{B_k\}_{k\in\mathbb{N}}$ of iterates that converge to $\nabla^2 f(\mathbf{x})$. For quadratic functions, their analysis also reveals exact rates for the convergence in expectation of the sequence of estimates $\{B_k\}_{k\in\mathbb{N}}$ to the true Hessian H. We are, however, interested in the inverses of these matrices, since we want to understand convergence of the covariances $\Sigma = B_k^{-1}, k \in \mathbb{N}$ to the optimal covariance H^{-1}. We know that for a sufficiently large number K of steps, $B_K^{-1} \approx H^{-1}$ almost surely, but explicit bounds for K do not directly follow from existing results.

We here address this question and provide, in addition, two novel, theoretically sound, and easy-to-implement VM updates. We also study their numerical performance on Nesterov's worst case function and quadratic functions with (inverse) sigmoidal spectral distribution. For simplicity, we analyze VM updates and their interplay with Random Pursuit solely on convex quadratic functions. The general convex case is subject of future research.

5.1 Variable Metric update schemes

The generic VM update [12] comprises direct updates of a Hessian estimate. Given a symmetric matrix $B \in \mathrm{PD}_n$ as current Hessian estimate, a direction $\mathbf{u}$ is selected uniformly at random from the n-dimensional hypersphere S^{n-1} (i.e., $\mathbf{u} \sim S^{n-1}$). The next iterate B_+ is determined according to:

$$B_+ = B + \mathbf{u}^T (H - B)\, \mathbf{u} \cdot \mathbf{u}\mathbf{u}^T . \tag{25}$$

This formula requires the evaluation of $\mathbf{u}^T H \mathbf{u}$ with unknown H. For twice differentiable functions f the second derivative of f at $\mathbf{x}$ in direction $\mathbf{u}$ can be well approximated by finite differences:

$$\mathbf{u}^T H \mathbf{u} \approx \frac{f(\mathbf{x} + \epsilon\mathbf{u}) - 2f(\mathbf{x}) + f(\mathbf{x} - \epsilon\mathbf{u})}{\epsilon^2} \tag{26}$$

for some small $\epsilon > 0$. In the convex quadratic case, the above formula is exact for arbitrary $\epsilon > 0$. Note that this formula only requires two additional function evaluations at $\mathbf{x} \pm \epsilon\mathbf{u}$. In addition, the formula implies that the estimate B_+ behaves at $\mathbf{x}$ like the unknown Hessian along direction $\mathbf{u}$, that is, $\mathbf{u}^T B_+ \mathbf{u} = \mathbf{u}^T H \mathbf{u}$. This can be seen directly from (25) by noting that $\mathbf{u}^T \mathbf{u}\mathbf{u}^T \mathbf{u} = 1$. Unfortunately, the update does not guarantee that the matrix B_+ stays positive definite. In order to be useful in Variable Metric Random Pursuit, an additional correction step is thus required. Leventhal and Lewis suggest an *ad hoc* projection of B_+ onto the cone of PD_n matrices. They numerically show that this yields a practicable algorithm [12]. We here introduce two alternatives, `updateHess` and `updateHessCorr`, as outlined in Fig. 2.

`updateHess`$(f, \mathbf{x}, B, \epsilon)$

Requires: Global variable T, initialized at first invocation to $T = B$
Output : Hessian estimate $B_+ \in \mathrm{PD}_n$
$\mathbf{u} \sim S^{n-1}$
$\Delta_{\mathbf{u}} \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{u}) - 2f(\mathbf{x}) + f(\mathbf{x}-\epsilon\mathbf{u})}{\epsilon^2} - \mathbf{u}^T B \mathbf{u}$
if $T \leftarrow T + \Delta_{\mathbf{u}} \cdot \mathbf{u}\mathbf{u}^T \in \mathrm{PD}_n$ **then**
  $B_+ \leftarrow T$
else
  $B_+ \leftarrow B$
return B_+

`updateHessCorr`$(f, \mathbf{x}, B, \epsilon)$

Output: Hessian estimate $B_+ \in \mathrm{PD}_n$
$\mathbf{u} \sim S^{n-1}$
$\Delta_{\mathbf{u}} \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{u}) - 2f(\mathbf{x}) + f(\mathbf{x}-\epsilon\mathbf{u})}{\epsilon^2} - \mathbf{u}^T B \mathbf{u}$
if $T \leftarrow B + \Delta_{\mathbf{u}} \cdot \mathbf{u}\mathbf{u}^T \in \mathrm{PD}_n$ **then**
  $B_+ \leftarrow T$
else
  $\mathbf{v} \leftarrow$ `smallestEVec`(T)
  $\Delta_{\mathbf{v}} \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{v}) - 2f(\mathbf{x}) + f(\mathbf{x}-\epsilon\mathbf{v})}{\epsilon^2} - \mathbf{v}^T T \mathbf{v}$
  $B_+ \leftarrow \left(B + \Delta_{\mathbf{v}} \cdot \mathbf{v}\mathbf{v}^T\right) + \Delta_{\mathbf{u}} \cdot \mathbf{u}\mathbf{u}^T$
return B_+

Fig. 2 Two implementations of the VM update scheme (25). Left panel: The update (25) is applied to a temporary matrix T and the matrix B is only updated if T is positive definite. Right panel: The matrix B is updated in every step. Positive definiteness is established by an additional correction step (see main text for further information).

In sub-routine `updateHess` the current matrix T is returned if it is positive definite. Otherwise, the *last known* positive definite matrix is used. As long as the iterates are positive definite, no additional computational effort is needed. The

update can be implemented in $O(n^2)$ by using a rank-one update on the Cholesky decomposition of T. However, if T is not positive semidefinite this approach fails, and the condition on line 3 of the algorithm must be checked by computation of the smallest eigenvalue.

In sub-routine `updateHessCorr` we ensure that the generated iterates are always positive definite. In case B_+ is not positive definite, we apply a second VM update step in direction $\mathbf{v}$, where $\mathbf{v}$ is an eigenvector of B_+ corresponding to the smallest (hence negative) eigenvalue of B_+. By standard matrix perturbation theory, as detailed in Lemma 10 below, the twice updated matrix will be positive semidefinite definite again (as H is). This scheme comes at the expense of two additional function evaluations at $\mathbf{x} \pm \epsilon \mathbf{v}$. This version of the VM update has already been successfully used in a recent numerical study [18].

Lemma 10 *Let $A \in \mathrm{PD}_n$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z}_1 \in \mathbb{R}^n$ an eigenvector corresponding to the smallest eigenvalue of $(A - \mathbf{x}\mathbf{x}^T)$. Then*

$$B := A - \mathbf{x}\mathbf{x}^T + \left|\lambda_{\min}(A - \mathbf{x}\mathbf{x}^T)\right| \mathbf{z}_1\mathbf{z}_1^T \in \mathrm{PD}_n\,.$$

Proof The matrix $(A - \mathbf{x}\mathbf{x}^T)$ is symmetric. Let $(A - \mathbf{x}\mathbf{x}^T) = \sum_{i=1}^n \lambda_i \mathbf{z}_i\mathbf{z}_i^T$ denote its spectral decomposition with $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_n$ in increasing order. If $\lambda_1 \geq 0$, then there is nothing to show. Otherwise, we observe that by a variant of Weyl's theorem (cf. [7, Theorem 4.3.4]), $0 \leq \lambda_i(A) \leq \lambda_{i+1}(A - \mathbf{x}\mathbf{x}^T) = \lambda_{i+1}$ for $i = 1, \ldots, n-1$. Thus at most λ_1 can be negative. We conclude

$$\mathbf{y}^T B \mathbf{y} = \mathbf{y}^T \left(\sum_{i=1}^n \lambda_i \mathbf{z}_i\mathbf{z}_i^T + |\lambda_1|\, \mathbf{z}_1\mathbf{z}_1^T \right) \mathbf{y} \geq \mathbf{y}^T \left(\lambda_1 \mathbf{z}_1\mathbf{z}_1^T + |\lambda_1|\, \mathbf{z}_1\mathbf{z}_1^T \right) \mathbf{y} \geq 0\,,$$

for all $\mathbf{y} \in \mathbb{R}^n$. □

Remark 1 The present VM update schemes can be used in several ways in combination with Random Pursuit: (i) One can use the VM scheme at the initial iterate $\mathbf{x}_0 \in \mathbb{R}^n$ multiple times in order to get a good approximation B of the Hessian $\nabla^2 f(\mathbf{x}_0)$ and then employ $\Sigma = B^{-1}$ in F-RP; (ii) one could continuously toggle between VM update steps and line searches (as shown in Fig. 1, right panel), yielding a fully adaptive scheme. The analysis of the latter combination is naturally more evolved. On quadratic functions, however, the VM update is independent of the current position $\mathbf{x} \in \mathbb{R}^n$ and thus, V-RP is amenable to a theoretical analysis. Finally, note also that the update schemes are easily parallelizable due to the independence of the directions $\mathbf{u}$ and the calculation of $\mathbf{u}H\mathbf{u}$. This can be used on today's multi-core machines to speed-up the algorithm.

5.2 Convergence of Variable Metric Random Pursuit

We now show that the combination of RP with sequential VM updates indeed yields a convergent algorithm on convex quadratic functions. To simplify the notation, we write the VM update in terms of the *error matrix* $E = (B - H)$.

Definition 3 (VM update scheme) Let $E_0 \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $\{\mathbf{u}_k\}_{k \geq 0}$ a sequence of independent vectors sampled uniformly from S^n. Then the sequence $\{E_k\}_{k>0}$ is generated by the VM update if the iterates satisfy

$$E_{k+1} = E_k - \mathbf{u}_k^T E_k \mathbf{u}_k \cdot \mathbf{u}_k \mathbf{u}_k^T\,. \tag{27}$$

We observe that this definition matches with (25) for $E_k = B_k - H$.

The following Lemma summarizes the most important properties of the VM update scheme.

Lemma 11 *Let $\{E_k\}_{k\geq 0}$ be generated by the VM update (27) with E_0 symmetric and let $K = n(n+2)\ln(a\sqrt{b})$ for parameters $a \geq 1$, $b > 1$. Then*

(i) $\|E_k\|_F \leq \|E_{k-1}\|_F$, *for* $k \geq 1$,
(ii) $\mathbb{E}\left[\|E_k\|_F^2\right] \leq \left(1 - \frac{2}{n(n+2)}\right)^k \|E_0\|_F^2$ *for* $k \geq 0$, *and*
(iii) $\|E_k\|_F \leq \frac{\|E_0\|_F}{a}$ *for all* $k \geq K$, *with probability at least* $1 - \frac{1}{b}$.

Corollary 1 *Let $\{B_k\}_{k\geq 0}$ with B_0 symmetric a sequence of iterates generated either by the VM update as implemented in* `updateHessCorr` *or the sequence of internal matrices T in* `updateHess`. *Then the statement from Lemma 11 also holds for the sequence of error matrices $\{E'_k\}_{k\geq 0} := \{B_k - H\}_{k\geq 0}$.*

Proof The internal matrices T in `updateHess` are exactly updated according to (27), hence nothing is to show. The iterates of `updateHessCorr` are almost generated according to (27). The additional correction step (if necessary) can be viewed as one step of the VM update (25) in a special (not random) direction. However, by (i) of Lemma 11 the Frobenius norm of the error matrix will not increase by this step. □

Proof (of Lemma 11) We note that (i) and (ii) were already proven by Leventhal and Lewis [12, Theorem 2.1]. It remains to show that (iii) follows from Markov's inequality. Since $\|E_k\|_F$ decreases monotonically, it suffices to prove (iii) for $k = K$. We have

$$\left(1 - \frac{2}{n(n+2)}\right)^K \leq \frac{1}{(a\sqrt{b})^2},$$

hence, by (ii),

$$\mathbb{E}[\|E_K\|_F^2] \leq \frac{1}{b}\frac{\|E_0\|_F^2}{a^2}.$$

By the Markov inequality, the probability that $\|E_K\|_F^2$ exceeds its expectation by more than a factor of b is at most $1/b$, and this yields the lemma. □

In order to show the effectiveness of the V-RP algorithm we have to argue that the convergence factor in Theorem 2 describing the single-step progress of F-RP converges to $1 - 1/(n+2)$. To provide some intuition on the convergence behavior, let us consider a quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x}$, with quadratic upper and lower bound $M = L_1 = H$. Then the convergence factor in Theorem 2 depends only on the spectrum of $H\Sigma$, with $\Sigma = B_k^{-1} = (H + E_k)^{-1}$.

Lemma 12 (VM convergence factor) *Let $H \in \mathrm{PD}_n$, let $\{E_k\}_{k\geq 0}$ with E_0 symmetric be a sequence of iterates generated by the VM update (27) on the function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x}$ and let $c < 1$. If for $k' \geq 0$:*

$$\|E_{k'}\|_2 \leq \frac{c}{\|H^{-1}\|_2},$$

then the convergence factor from Theorem 2 can be upper bounded by

$$1 - \frac{\lambda_{\min}(H\Sigma)}{\mathrm{Tr}[H\Sigma] + 2\lambda_{\max}(H\Sigma)} \leq 1 - \frac{(1-c)^2}{n+2}, \tag{28}$$

for $\Sigma = (H + E_{k'})^{-1}$.

Proof To show the Lemma, we derive bounds on the smallest and largest eigenvalues of the matrix $H\Sigma$. First we observe that

$$\max\left\{\left|\lambda_{\min}(E_{k'}H^{-1})\right|, \left|\lambda_{\max}(E_{k'}H^{-1})\right|\right\} = \left\|E_{k'}H^{-1}\right\|_2 \leq \left\|E_{k'}\right\|_2 \left\|H^{-1}\right\|_2$$

by the definition of the 2-norm and submultiplicativity. With the assumption on the product of the two norms, the whole expression can be upper bounded by c:

$$\left\|E_{k'}\right\|_2 \left\|H^{-1}\right\|_2 \leq c < 1\,.$$

Therefore, the largest eigenvalue of $H\Sigma$ is well defined and finite:

$$\begin{aligned}\lambda_{\max}(H\Sigma) &= \frac{1}{\lambda_{\min}((H\Sigma)^{-1})} = \frac{1}{\lambda_{\min}(I_n + E_{k'}H^{-1})} \\ &= \frac{1}{1+\lambda_{\min}(E_{k'}H^{-1})} \leq \frac{1}{1-c}\,.\end{aligned}$$

With a similar argumentation we obtain a lower bound on the smallest eigenvalue of $H\Sigma$:

$$\lambda_{\min}(H\Sigma) = \frac{1}{1+\lambda_{\max}(E_{k'}H^{-1})} \geq 1 - \lambda_{\max}(E_{k'}H^{-1}) \geq 1 - c\,,$$

where the first inequality follows from $\frac{1}{1+x} \geq 1 - x$ for $x > -1$. These two bounds together with the trivial estimate $\mathrm{Tr}[H\Sigma] \leq n\lambda_{\max}(H\Sigma)$ yield:

$$\frac{\lambda_{\min}(H\Sigma)}{\mathrm{Tr}[H\Sigma] + 2\lambda_{\max}(H\Sigma)} \geq \frac{\lambda_{\min}(H\Sigma)}{(n+2)\lambda_{\max}(H\Sigma)} \geq \frac{(1-c)^2}{(n+2)}\,,$$

and the claim follows. □

Corollary 2 *Let $c < 1$ and $K = n(n+2)\ln(\left\|H^{-1}\right\|_2 \left\|E_0\right\|_F \sqrt{b}/c)$. Then* (28) *holds with probability at least $1 - \frac{1}{b}$.*

Proof We observe that $\left\|E_k\right\|_2 \leq \left\|E_k\right\|_F$. Now the result follows from Lemma 12 and Lemma 11 (iii) with $a = \left\|H^{-1}\right\|_2 \left\|E_0\right\|_F / c$. □

We can also interpret the statement of Corollary 2 as follows: By solving the equation for c and plugging the result into (28), we obtain an upper bound on the convergence factor depending only on b, the success parameter, and K, the number of iterations. With probability at least $1 - \frac{1}{b}$ it holds:

$$c \leq \left\|H^{-1}\right\|_2 \left\|E_0\right\|_F \sqrt{b} e^{-\frac{K}{n(n+2)}}\,. \tag{29}$$

In summary, these results guarantee the near-optimal linear convergence of V-RP after an initial learning phase of at most K steps with high probability.

5.3 Illustrative numerical examples

We now illustrate the typical convergence behavior of Variable Metric Random Pursuit on three instances of convex quadratic functions. The first instance is Nesterov's worst case function f_{Nes}:

$$f_{\mathrm{Nes}}(\mathbf{x}) = \frac{\ell - m}{4}\left(\frac{1}{2}\left[x_1^2 + \sum_{i=1}^{n-1}\left(x_{i+1} - x_i\right)^2 + x_n^2\right] - x_1\right) + \frac{m}{2}\left\|x\right\|^2\,.$$

We here use this strongly convex function with parameter $m = 1$ and $\ell = 10^7$. Further information about this function as well as extensive numerical results of

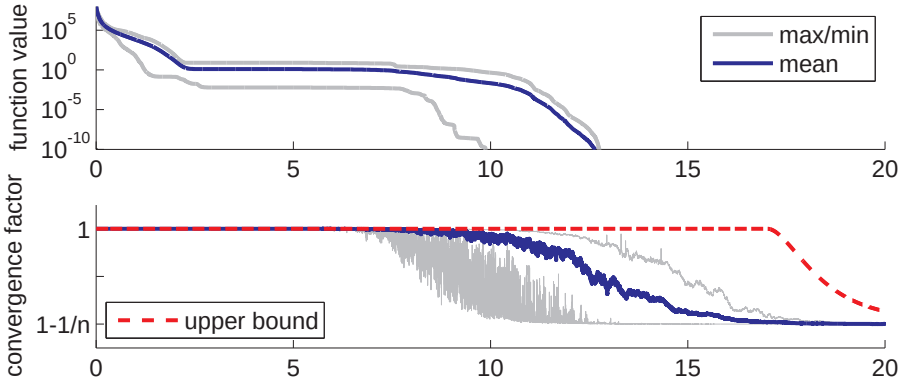

Fig. 3 Convergence of V-RP with `updateHessCorr` on f_{Nes}. Upper panel: Mean (blue) and max/min (grey) function values vs. # ITS over 111 runs. Lower Panel: Mean (blue) and max/min (grey) convergence factor (see Thm. 2) vs. # ITS over 111 runs. Theoretical upper bound (red) on the convergence factor. See main text for further information.

the convergence of standard RP on this function can be found in [19]. In addition, we test V-RP on two quadratic functions with sigmoidal and inverse sigmoidal (almost flat) distribution of eigenvalues, referred to as f_{Sigm} and f_{Flat}, respectively. These functions have been introduced in [18] and are defined as

$$f_{\text{Sigm}}(\mathbf{x}) = \sum_{i=1}^{n} \text{nm}_i \left(\left(1 + e^{15 - \frac{30(t-1)}{n-1}} \right)^{-1} + \frac{1}{2} \right) (x_i - 1)^2 \,.$$

and

$$f_{\text{Flat}}(\mathbf{x}) = \sum_{i=1}^{n} \text{nm}_i \left(-\log \left(\left(10^{-6} + \frac{(t-1)(1 - 2 \cdot 10^{-6})}{n-1} \right)^{-1} - 1 \right) \right) (x_i - 1)^2 \,.$$

with the normalizing function $\text{nm}_i(f(t)) = \frac{\ell - 1}{2} \frac{f(i)}{|f(1)|} + \frac{\ell + 1}{2}$ with $\ell = 10^7$.

For all considered quadratic functions the ratio of largest to smallest eigenvalue of the Hessians (i.e. the condition number) is 10^7, and their global minimum at $\mathbf{x}^* = \mathbf{1}_n$ (where $\mathbf{1}_n$ is the all-ones vector) with $f(\mathbf{x}^*) = 0$. On all functions we conduct 111 runs of V-RP in $n = 50$ dimensions. The initial conditions are $\mathbf{x}_0 = \mathbf{0}$, $B_0 = 0.5 \cdot 10^7 \cdot I_n$. For comparison both VM update schemes (see Fig. 2) are tested. We here report the evolution of the mean, maximum, and minimum function value vs. number of iterations (# ITS). We also calculate and report the derived convergence factor from Thm. 2. We observe a number of interesting features of the numerical optimization trajectories. Independent of the concrete function instance we see that V-RP shows three distinct phases of convergence in function values: (i) a first short phase of rapid improvement, (ii) a metric learning phase with only marginal progress in function decrease, and (iii) a final rapid decrease in function value. For the considered functions this last phase appears after 10-15 n^2 iterations. We also observe that, on average, the theoretically derived convergence factor indeed captures the on-set of the fast convergence of V-RP. Both VM update schemes have a similar effect on convergence, with `updateHess` producing a more discontinuous trajectory of function values. This is expected because in V-RP with `updateHess` the employed metric stays fixed over longer periods of iterations (until the Hessian estimate becomes positive definite again). On f_{Nes} the numerical data suggest that `updateHess` is preferred over `updateHessCorr` because the former strategy consistently needs less iterations (both on average and on the tails) to reach the global minimum (see Figs. 3 and 4, respectively). This can also be observed on

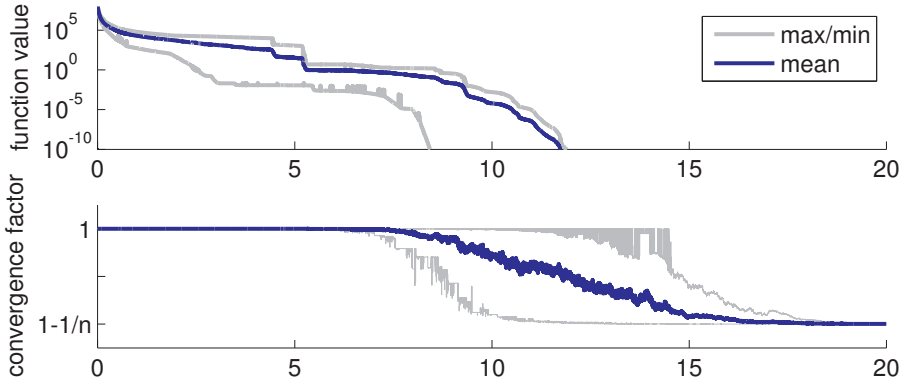

Fig. 4 Convergence of V-RP with `updateHess` on f_{Nes}. Upper panel: Mean (blue) and max/min (grey) function values vs. # ITS over 111 runs. Lower Panel: Mean (blue) and max/min (grey) convergence factor vs. # ITS over 111 runs. See main text for further information.

f_{Flat} (see Figs. 5 and 6). On f_{Sigm}, however, the V-RP scheme with `updateHessCorr` is superior (see Figs. 7 and 8, respectively).

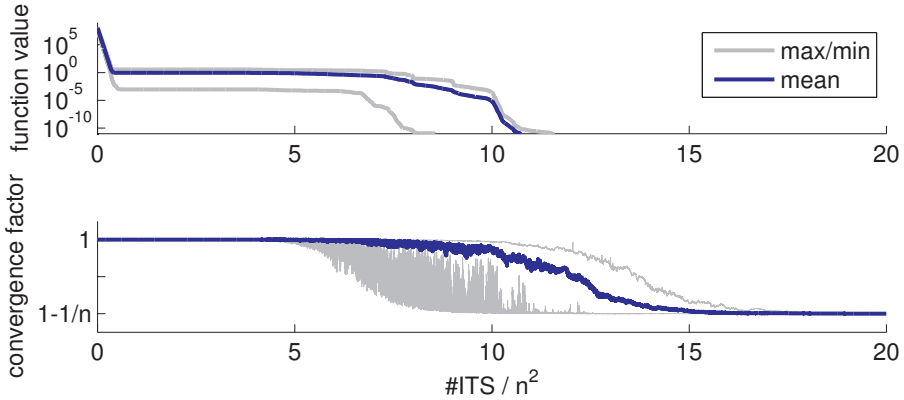

Fig. 5 Convergence of V-RP with `updateHessCorr` on f_{Flat}. Upper panel: Mean (blue) and max/min (grey) function values vs. # ITS over 111 runs. Lower Panel: Mean (blue) and max/min (grey) convergence factor (see Thm. 2) vs. # ITS over 111 runs. See main text for further information.

Overall, the numerical results show the clear advantage of learning a variable metric. Non-adaptive schemes such as standard RP or Nesterov's randomized gradient-free schemes would not even come close to the achieved solution accuracy within the considered function evaluation budget [15,19].

In Figs. 3 and 7 we show the upper bound on the convergence factor obtained in Lem. 12 and Cor. 2 with parameter $b = 2$ (see also the discussion at the end of Sec. 5.2). In general, we expect the bound to be loose because the derivation in both Lem. 12 and Cor. 2 are based on worst-case assumptions that do not reflect the average situation. We indeed observe that the bound is very conservative on Nesterov's function and does over-estimate the observed convergence factors. On function f_{Sigm}, however, the bound is much closer to the observed convergence factors. This is in full agreement with previous numerical studies [18] that identified f_{Sigm} as a challenging function for adaptive schemes. The presented experiments suggest that the convergence factor, derived in Thm. 2, is an interesting quantity that deserves further theoretical investigation.

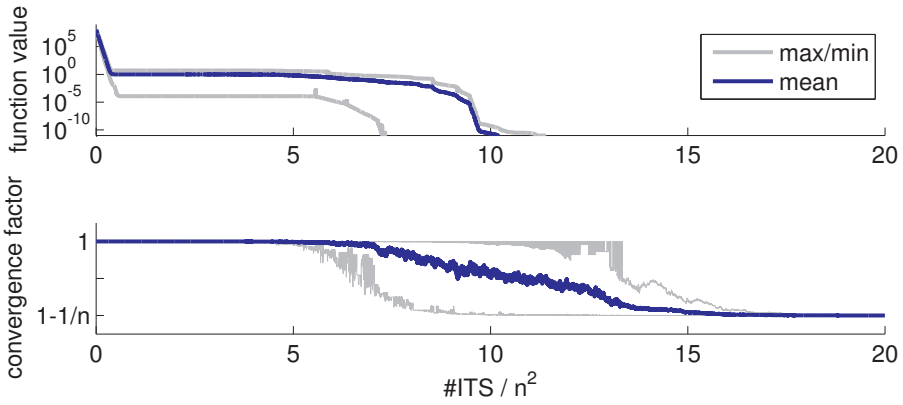

Fig. 6 Convergence of V-RP with `updateHess` on f_{Flat}. Upper panel: Mean (blue) and max/min (grey) function values vs. # ITS over 111 runs. Lower Panel: Mean (blue) and max/min (grey) convergence factor (see Thm. 2) vs. # ITS over 111 runs. See main text for further information.

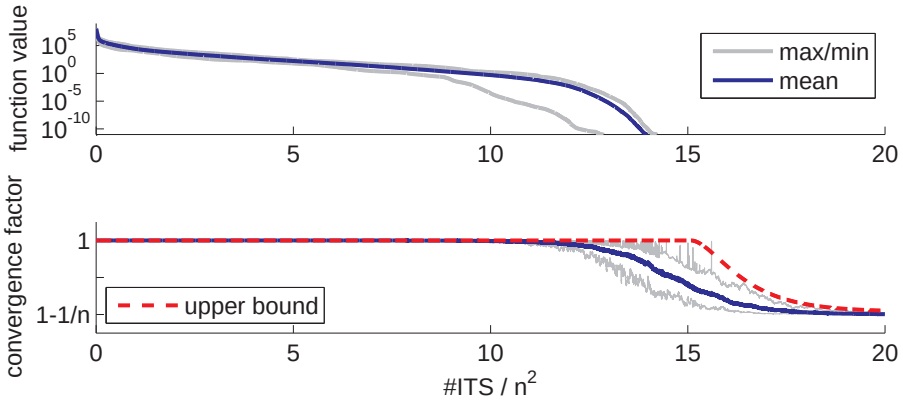

Fig. 7 Convergence of V-RP with `updateHessCorr` on f_{Sigm}. Upper panel: Mean (blue) and max/min (grey) function values vs. # ITS over 111 runs are reported. Lower Panel: Mean (blue) and max/min (grey) convergence factor (see Thm. 2) vs. # ITS over 111 runs are reported. Theoretical upper bound (red) on the convergence factor. See main text for further information.

6 Discussion and Conclusion

In this contribution we have analyzed Random Pursuit algorithms that employ (i) a fixed but arbitrary metric (Fixed Metric Random Pursuit) and (ii) a variable metric learning procedure (Variable Metric Random Pursuit). We have detailed convergence proofs and convergence rates for these Random Pursuit algorithms on convex functions. We have used an improved (matrix) quadratic upper bound technique to show expected single-step progress and global convergence of Fixed Metric Random Pursuit on (strictly) convex functions. We have also shown that Variable Metric Random Pursuit can achieve a near-optimal convergence rate on convex quadratic functions that, after a finite learning phase of length at most $O(n^2)$, does not depend on the underlying properties of the unknown Hessian of the function. Compared to standard Random Pursuit [19] we have thus removed one of our previously identified challenges toward the design of competitive gradient-free optimization methods that are easy to implement, possess theoretical convergence guarantees, and are useful in practice

Nonetheless, a number of theoretical challenges remain. Firstly, it would be very interesting to give tighter upper and lower bounds on the expected convergence

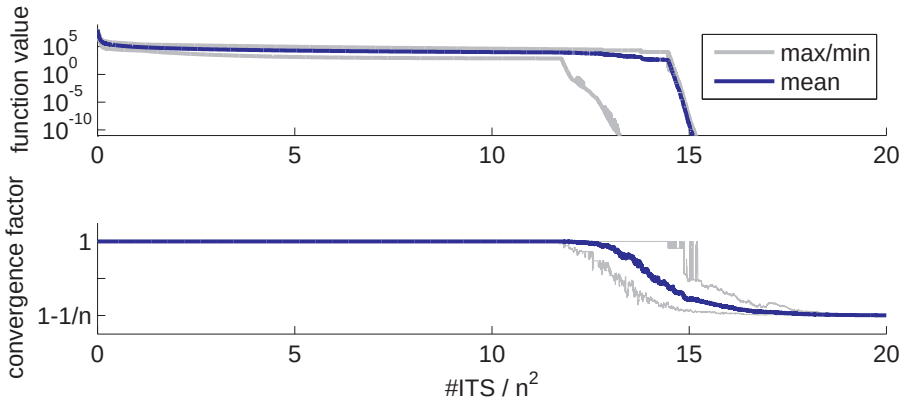

Fig. 8 Convergence of V-RP with `updateHess` on f_{Sigm}. Upper panel: Mean (blue) and max/min (grey) function values vs. # ITS over 111 runs are reported. Lower Panel: Mean (blue) and max/min (grey) convergence factor (see Thm. 2) vs. # ITS over 111 runs are reported. See main text for further information.

factor for Variable Metric Random Pursuit. This is subject to further research, although some first results have already been obtained in [20]. Secondly, it is still an open question how to analyze Random Pursuit schemes for constrained optimization problems of the form

$$\min f(x) \quad \text{subject to} \quad x \in \mathcal{K}\,, \tag{30}$$

where $\mathcal{K} \subset \mathbb{R}^n$ is a convex set. Furthermore, it is open how to derive convergence guarantees for Random Pursuit schemes on the class of globally convex (or δ-convex) functions [8], or on noisy functions with certain bounds on the variance of the noise. Finally, convergence on the important class of non-smooth convex functions is another fundamental challenge for gradient-free optimization that, most likely, needs novel tools and techniques to be developed by the mathematical programming community.

References

1. Broyden, C.G.: The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. IMA Journal of Applied Mathematics **6**(1), 76–90 (1970). DOI 10.1093/imamat/6.1.76. URL `http://imamat.oxfordjournals.org/content/6/1/76.abstract`
2. Davidon, W.C.: Variable Metric Method for Minimization. SIAM Journal on Optimization **1**(1), 1–17 (1991). DOI 10.1137/0801001. URL `http://link.aip.org/link/?SJE/1/1/1`
3. Fletcher, R.: A new approach to variable metric algorithms. The Computer Journal **13**(3), 317–322 (1970). DOI 10.1093/comjnl/13.3.317. URL `http://comjnl.oxfordjournals.org/content/13/3/317.abstract`
4. Goldfarb, D.: A Family of Variable-Metric Methods Derived by Variational Means. Mathematics of Computation **24**(109), 23–26 (1970). URL `http://www.jstor.org/stable/2004873`
5. Hansen, N., Ostermeier, A.: Completely Derandomized Self-Adaption in Evolution Strategies. Evolutionary Computation **9**(2), 159–195 (2001)
6. Heijmans, R.: When does the expectation of a ratio equal the ratio of expectations? Statistical Papers **40**, 107–115 (1999)
7. Horn, R.A., Johnson, C.R.: Matrix analysis, reprint 1990 edn. Cambridge University Press (1985)
8. Hu, T.C., Klee, V., Larman, D.: Optimization of globally convex functions. SIAM Journal on Control and Optimization **27**(5), 1026–1047 (1989). DOI 10.1137/0327055. URL `http://link.aip.org/link/?SJC/27/1026/1`
9. Isserlis, L.: On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika **12**, 134–139 (1918)
10. Jägerskküpper, J.: Rigorous runtime analysis of the (1+1) ES: 1/5-rule and ellipsoidal fitness landscapes. In: FOGA, *LNCS*, vol. 3469, pp. 356–361 (2005)

11. Kjellström, G., Taxen, L.: Stochastic Optimization in System Design. IEEE Trans. Circ. and Syst. **28**(7) (1981)
12. Leventhal, D., Lewis, A.S.: Randomized Hessian estimation and directional search. Optimization **60**(3), 329–345 (2011). DOI 10.1080/02331930903100141. URL `http://www.tandfonline.com/doi/abs/10.1080/02331930903100141`
13. Marti, K.: Controlled random search procedures for global optimization. In: V. Arkin, A. Shiraev, R. Wets (eds.) Stochastic Optimization, *Lecture Notes in Control and Information Sciences*, vol. 81, pp. 457–474. Springer (1986)
14. Müller, C.L., Sbalzarini, I.F.: Gaussian adaptation revisited - an entropic view on covariance matrix adaptation. In: C. Di Chio et al. (ed.) EvoApplications, no. 6024 in Lecture Notes in Computer Science, pp. 432–441. Springer (2010)
15. Nesterov, Y.: Random Gradient-Free Minimization of Convex Functions. Tech. rep., ECORE (2011)
16. Schumer, M., Steiglitz, K.: Adaptive step size random search. Automatic Control, IEEE Transactions on **13**(3), 270–276 (1968). DOI 10.1109/TAC.1968.1098903
17. Shanno, D.F.: Conditioning of Quasi-Newton Methods for Function Minimization. Mathematics of Computation **24**(111), 647–656 (1970). URL `http://www.jstor.org/stable/2004840`
18. Stich, S.U., Müller, C.L.: On spectral invariance of Randomized Hessian and Covariance Matrix Adaptation schemes. In: Parallel Problem Solving From Nature (PPSN). Springer (2012)
19. Stich, S.U., Müller, C.L., Gärtner, B.: Optimization of convex functions with Random Pursuit. http://arxiv.org/abs/1111.0194 (2011)
20. Stich, S.U., Müller, C.L., Gärtner, B.: Matrix-valued Iterative Random Projections. Tech. rep., ETH Zürich (2012). URL `http://people.inf.ethz.ch/sstich/mrp.pdf`. Manuscript