CARL VON OSSIETZKY universität OLDENBURG

IfM INSTITUT FÜR MATHEMATIK

Linear convergence of descent methods for the unconstrained minimization of restricted strongly convex functions*

Frank Schöpfer[1]

IfM Preprint No. 2016-03
September 2016

Institut für Mathematik
Carl von Ossietzky Universität Oldenburg
D-26111 Oldenburg, Germany

*This version is available at `https://www.uni-oldenburg.de/mathematik/forschung/preprints/`
Published in *SIAM J. Optim.* 26 (2016), no. 3, 1883–1911, DOI: 10.1137/140992990

[1] Carl von Ossietzky Universitt Oldenburg, Germany; `frank.schoepfer@uni-oldenburg.de`

LINEAR CONVERGENCE OF DESCENT METHODS FOR THE UNCONSTRAINED MINIMIZATION OF RESTRICTED STRONGLY CONVEX FUNCTIONS

FRANK SCHÖPFER[*]

Abstract. Linear convergence rates of descent methods for unconstrained minimization are usually proven under the assumption that the objective function is strongly convex. Recently it was shown that the weaker assumption of restricted strong convexity suffices for linear convergence of the ordinary gradient descent method. A decisive difference to strong convexity is that the set of minimizers of a restricted strongly convex function need be neither a singleton nor bounded. In this paper we extend the linear convergence results under this weaker assumption to a larger class of descent methods including restarted nonlinear CG, BFGS and its damped limited memory variants L-D-BFGS. For twice continuously differentiable objective functions we even obtain r-step superlinear convergence for the CG_DESCENT conjugate gradient method of Hager and Zhang, where r is greater than or equal to the rank of the Hessian at a minimizer. This is remarkable since the Hessian of a restricted strongly convex function need not have full rank. Furthermore we show that convex quadratic splines and objective functions of the unconstrained duals to some linearly constrained optimization problems are restricted strongly convex. In particular this holds for the regularized Basis Pursuit problem and its analogues for nuclear norm minimization and Principal Component Pursuit.

Key words. linear convergence, restricted strong convexity, error bound, quadratic splines, BFGS, conjugate gradient

AMS subject classifications. 65K, 90C

1. Introduction.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex with a Lipschitz-continuous gradient and nonempty set of minimizers X_f. We consider the unconstrained minimization

$$f_{min} := \min_{x \in \mathbb{R}^n} f(x) \tag{1.1}$$

by iteration methods of the form

$$x_{k+1} = x_k + t_k \cdot d_k \quad \text{for} \quad k = 0, 1, 2, \dots \tag{1.2}$$

with step length $t_k > 0$ and search directions d_k of the conjugate gradient (CG) type

$$d_k = -g_k + \beta_k \cdot d_{k-1} \,,$$

where $g_k := \nabla f(x_k)$ and $\beta_k \in \mathbb{R}$ is a parameter, or Quasi-Newton methods,

$$d_k = -B_k^{-1} g_k \,, \tag{1.3}$$

with symmetric positive definite matrices $B_k \in \mathbb{R}^{n \times n}$. This comprises ordinary gradient descent for $d_k = -g_k$ as well as BFGS and its limited memory variant L-BFGS for large-scale problems, see eg. [45]. While the above assumptions on f are generally sufficient to prove global convergence results under mild restrictions on t_k and B_k, linear (or superlinear) convergence rates are usually obtained only under the additional assumption that f is strongly convex. In the recent papers [65, 66] it was shown that, for the ordinary gradient descent, linear convergence rates can be proven under the

[*] INSTITUT FÜR MATHEMATIK, CARL VON OSSIETZKY UNIVERSITÄT OLDENBURG (FRANK.SCHOEPFER@UNI-OLDENBURG.DE)

weaker assumption that f is only *restricted strongly convex* on $\mathbb{R}^n$, i.e. there is a constant $\nu > 0$ such that f satisfies for all $x \in \mathbb{R}^n$ the *restricted secant inequality*

$$\left\langle \nabla f(x) - \nabla f\left(P_{X_f}(x)\right), x - P_{X_f}(x) \right\rangle \geq \nu \cdot \left\| x - P_{X_f}(x) \right\|_2^2 , \tag{1.4}$$

where $P_{X_f}(x)$ denotes the orthogonal projection of x onto X_f. A decisive difference to strong convexity is that X_f need be neither a singleton nor bounded.

Here we analyse to what extend restricted strong convexity is also sufficient to guarantee linear convergence rates for the general methods above. In particular, for search directions of the form (1.3) and several line search strategies including Wolfe conditions and backtracking, we prove linear convergence rates for the decrease of both the function values $f(x_k) - f_{min}$ and the distance $\|x_k - \hat{x}\|_2$ of the iterates to some minimizer $\hat{x} \in X_f$ under the assumption that f is restricted strongly convex on the level set $\mathcal{L}_{f(x_0)} := \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ and that all matrices B_k, B_k^{-1} are uniformly bounded. In fact the assumption of restricted strong convexity allows for a very simple proof. A similar result was shown in [32] for the special case of f being a *convex quadratic spline*, i.e. a differentiable convex piecewise quadratic function. There the convergence analysis is based on error bounds proven in [31]. Especially it was shown in [31] that for any convex piecewise quadratic function f (not necessarily differentiable) and any $\delta > 0$ there exists a constant $\gamma > 0$ such that for all $x \in \mathbb{R}^n$ with $f(x) - f_{min} \leq \delta$ we have

$$\operatorname{dist}(x, X_f) \leq \gamma \cdot \sqrt{f(x) - f_{min}} . \tag{1.5}$$

In [38] convex, not necessarily differentiable, functions with property (1.5) were called *optimally strongly convex*, and for the minimization of such functions linear convergence of the so called *asynchronous stochastic coordinate descent* algorithm was proven. We show that (1.5) and (1.4) are actually equivalent on any level set of a differentiable convex function f. This immediately implies that a convex quadratic spline is also restricted strongly convex on any level set. Hence we extend the result of [32] to a larger class of functions, while giving an even simpler proof. Furthermore we show that some of the limited memory damped BFGS methods (L-D-BFGS) considered in [2] produce uniformly bounded matrices B_k, B_k^{-1}, and thus retain linear convergence rates under the weaker assumption of restricted strong convexity. For the standard BFGS method with a line search satisfying the Wolfe conditions we obtain linear convergence rates for the decrease of both the function values $f(x_k) - f_{min}$ and the distance $\operatorname{dist}(x_k, X_f)$ of the iterates to the set of minimizers. We also prove linear convergence rates for CG methods with restarts, several types of line searches and many choices of the parameter β_k found in the literature. For twice continuously differentiable objective functions we even obtain r-step superlinear convergence for the CG_DESCENT conjugate gradient method of Hager and Zhang [23], where r is greater than or equal to the rank of the Hessian at a minimizer. This is remarkable since the Hessian of a restricted strongly convex function need not have full rank.

An important class of problems that can be cast in the form (1.1) are linearly constrained convex optimization problems

$$\min_{y \in \mathbb{R}^m} g(y) \quad s.t. \quad Ay = b , \tag{1.6}$$

where $A \in \mathbb{R}^{n \times m}$, $b \in \mathcal{R}(A)$, and $g : \mathbb{R}^m \to \mathbb{R}$ is strongly convex, which implies that (1.6) has a unique solution $\hat{y} \in \mathbb{R}^m$. The objective function of the unconstrained dual (1.1) to (1.6) is then given by $f(x) := g^*(A^T x) - \langle b\,, x\rangle$, where g^* is the convex conjugate of g. It can be shown that linear convergence of $\operatorname{dist}(x_k, X_f)$ implies linear convergence of the iterates $y_k := \nabla g^*(A^T x_k)$ to $\hat{y}$, even if the iterates x_k do not converge. Hence all of the discussed methods can be applied to solve (1.6) with linear convergence rates in case f is restricted strongly convex on $\mathcal{L}_{f(x_0)}$. The question arises whether restricted strong convexity of f is automatically implied by some properties of g which are relatively simple to check. We can immediately give a positive answer for piecewise quadratic (not necessarily differentiable) functions g, because then g^* and f are actually convex quadratic splines, see [55]. One of the most prominent examples for this case is given by the regularized *Basis Pursuit problem*, which arises in the vast area of sparse optimization, see e.g. [7, 11, 19],

$$\min_{y \in \mathbb{R}^m} \tfrac{1}{2} \cdot \|y\|_2^2 + \tau \cdot \|y\|_1 \quad s.t. \quad Ay = b\,. \tag{1.7}$$

See also [56] for explicit values of $\tau > 0$ that guarantee exact recovery of sparse solutions. The dual objective function f_τ to (1.7) can be written as

$$f_\tau(x) = \tfrac{1}{2} \cdot \left\|\operatorname{shrink}_\tau(A^T x)\right\|_2^2 - \langle b\,, x\rangle \ ,$$

where $\operatorname{shrink}_\tau(y) := \operatorname{sign}(y) \cdot \max\{|y| - \tau, 0\}$ is the componentwise soft shrinkage operator. In [63] it was observed that the well known *linearized Bregman method* for the solution of (1.7) can be interpreted as an ordinary gradient descent method applied to the dual with f_τ, see also [40]. It follows from the results in [27] that the linearized Bregman method converges linearly for constant step length t_k, *kicking* [46] and a *BB-line search* [63]. Here we extend this result to more line search strategies and efficient descent methods like CG. It also follows from the results in [27] that f_τ is indeed restricted strongly convex on all of $\mathbb{R}^n$. The proof given there yields an explicit value for the constant ν in (1.4) but relies on the special structure of f_τ. Hence we may ask whether there is a simple characterization of convex quadratic splines that are restricted strongly convex on all of $\mathbb{R}^n$. We can give a partial answer by showing that a convex quadratic spline f is restricted strongly convex on all of $\mathbb{R}^n$ and has a bounded set of minimizers, if and only if f is coercive. It follows that all dual objective functions f to problems of the form (1.6) are restricted strongly convex on all of $\mathbb{R}^n$, if g is strongly convex piecewise quadratic (which is the case in (1.7)), and if A has full row rank, because then f is coercive. We point out that in [27] restricted strong convexity of f_τ was proven even without this additional assumption on A.

More generally we prove that the objective function f to the unconstrained dual of (1.6) is restricted strongly convex on $\mathcal{L}_{f(x_0)}$ if the subdifferential mapping ∂g is *calm* at the optimal solution $\hat{y}$ and if the constraint qualification $\operatorname{rint}\big(\partial g(\hat{y})\big) \cap \mathcal{R}(A^T) \neq \emptyset$ holds. We show that ∂g is always calm at $\hat{y}$ for functions of the form $g(Y) = h\big(\sigma(Y)\big)$ for matrices $Y \in \mathbb{R}^{m\times n}$ and an absolutely symmetric convex piecewise quadratic function h, where $\sigma(Y)$ is the vector of all singular values of Y. In particular this holds for the objective function of the *regularized nuclear norm* optimization problem,

$$\min_{Y \in \mathbb{R}^{m\times n}} \tfrac{1}{2} \cdot \|Y\|_F^2 + \tau \cdot \|Y\|_* \quad s.t. \quad \mathcal{A}(Y) = b\,, \tag{1.8}$$

and the objective function of the regularized *Principal Component Pursuit problem*,

$$\min_{Y_1, Y_2 \in \mathbb{R}^{m\times n}} \tfrac{1}{2} \cdot \|Y_1\|_F^2 + \tau_1 \cdot \|Y_1\|_* + \tfrac{1}{2} \cdot \|Y_2\|_F^2 + \tau_2 \cdot \|Y_2\|_1 \quad s.t. \quad Y_1 + Y_2 = B\,, \tag{1.9}$$

which are analogues to (1.7) in the area of low rank matrix solutions like *low rank matrix completion* and *robust principal component analysis*, see eg. [8, 9, 27, 36, 53] and [64], where also explicit parameter values are given that guarantee exact recovery of low rank solutions. To the best of our knowledge this is the first time that linear convergence rates can be proven for the singular value thresholding algorithm of [6] and its generalization for the solution of (1.8) and (1.9). But see also [26] for linear convergence of the proximal gradient method for nuclear norm regularized problems.

In the next section we review the most important properties of restricted strongly convex functions that are used for the convergence analysis in section 3, and establish the restricted strong convexity of convex quadratic splines and the objective functions of the unconstrained duals to some linearly constrained optimization problems.

2. Restricted strong convexity.

We slightly extend the definition of restricted strong convexity given in [27, 65] to the case that a function f can be restricted strongly convex only on a convex subset $C \subset \mathbb{R}^n$, and not necessarily on all of $\mathbb{R}^n$.

Definition 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex differentiable with a nonempty set of minimizers X_f, and let $C \subset \mathbb{R}^n$ be a closed convex subset with $X_f \cap C \neq \emptyset$. Then f is* restricted strongly convex *on C with constant $\nu > 0$ if it satisfies for all $x \in C$ the* restricted secant inequality

$$\left\langle \nabla f(x) - \nabla f\big(P_{X_f \cap C}(x)\big)\, , x - P_{X_f \cap C}(x) \right\rangle \geq \nu \cdot \left\| x - P_{X_f \cap C}(x) \right\|_2^2 . \tag{2.1}$$

Note that $\nabla f\big(P_{X_f \cap C}(x)\big) = 0$ and that strongly convex functions satisfy (2.1). We need the following well-known properties of a convex function with Lipschitz-continuous gradient (cf. Proposition 12.60 in [55]).

Lemma 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex with a Lipschitz-continuous gradient with constant $L > 0$. Then for all $x, y \in \mathbb{R}^n$ we have*

$$f(y) - f(x) \leq \langle \nabla f(x)\, , y - x \rangle + \frac{L}{2} \cdot \|y - x\|_2^2 , \tag{2.2}$$

$$\langle \nabla f(y) - \nabla f(x)\, , y - x \rangle \geq \frac{1}{L} \cdot \|\nabla f(y) - \nabla f(x)\|_2^2 . \tag{2.3}$$

Restricted strongly convex functions have nice properties and satisfy error bounds that are important for the convergence analysis of gradient based iteration methods. See also [41, 42, 47] for the use of error bounds in mathematical programming.

Lemma 2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be restricted strongly convex on C with constant $\nu > 0$. Then for all $x \in C$ we have*

$$\nu \cdot \left\| x - P_{X_f \cap C}(x) \right\|_2 \ \leq \ \|\nabla f(x)\|_2 , \tag{2.4}$$

$$\frac{\nu}{2} \cdot \left\| x - P_{X_f \cap C}(x) \right\|_2^2 \ \leq \ f(x) - f_{min} , \tag{2.5}$$

$$f(x) - f_{min} \ \leq \ \frac{1}{\nu} \cdot \|\nabla f(x)\|_2^2 . \tag{2.6}$$

If in addition f has a Lipschitz-continuous gradient with constant $L \geq 0$ then we also have

$$\|\nabla f(x)\|_2 \leq L \cdot \sqrt{\frac{2}{\nu}} \cdot \sqrt{f(x) - f_{min}} . \tag{2.7}$$

Proof. By applying the Cauchy-Schwartz inequality to (2.1) we get (2.4). To show (2.5) we adopt the proof of Lemma 1 in [65]. For $x \in C$ we set $\hat{x} := P_{X_f \cap C}(x)$ and $x_t := \hat{x} + t(x - \hat{x})$ for $t \in [0,1]$. Since C is convex we have $x_t \in C$ for all $t \in [0,1]$ and $\hat{x}$ is also the orthogonal projection of x_t onto $X_f \cap C$. By (2.1), and since $f(\hat{x}) = f_{min}$, we get

$$\begin{aligned} f(x) - f_{min} &= \int_0^1 \langle \nabla f(x_t)\,, x - \hat{x} \rangle \; dt = \int_0^1 \frac{1}{t} \langle \nabla f(x_t)\,, x_t - \hat{x} \rangle \; dt \\ &\geq \int_0^1 \frac{\nu}{t} \cdot \|x_t - \hat{x}\|_2^2 \; dt = \frac{\nu}{2} \cdot \|x - \hat{x}\|_2^2 \,. \end{aligned}$$

Furthermore, convexity of f implies

$$f(x) - f_{min} \leq \langle \nabla f(x)\,, x - \hat{x} \rangle \leq \|\nabla f(x)\|_2 \cdot \|x - \hat{x}\|_2 \,.$$

Inequality (2.6) then follows from (2.4), and (2.7) follows from (2.5). □

An immediate consequence of (2.5) is the following.

COROLLARY 2.3. *If $f : \mathbb{R}^n \to \mathbb{R}$ is restricted strongly convex on all of $\mathbb{R}^n$ and has a bounded set of minimizers then f is coercive, i.e. $\lim_{\|x\|_2 \to \infty} \frac{f(x)}{\|x\|_2} = \infty$.*

The following lemma shows that (2.1) and (2.5) are actually equivalent for convex differentiable functions.

LEMMA 2.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex differentiable with a nonempty set of minimizers X_f, and let $C \subset \mathbb{R}^n$ be a closed convex subset with $X_f \cap C \neq \emptyset$. If there is a constant $\gamma > 0$ such that for all $x \in C$ we have*

$$\operatorname{dist}(x, X_f \cap C) \leq \gamma \cdot \sqrt{f(x) - f_{min}}\,,$$

then f is restricted strongly convex on C with constant $\nu := \frac{1}{\gamma^2}$.

Proof. By convexity of f we have $f(x) - f_{min} \leq \left\langle \nabla f(x)\,, x - P_{X_f \cap C}(x) \right\rangle$, from which the assertion follows. □

In general it is not easy to check whether a function f is restricted strongly convex, since projections onto the set of mimimizers may not be easy to analyse. But there are some interesting classes of functions with this property. The following example is a slight variation of Theorem 5 in [65] and Theorem 2 in [66], and its proof is analoguous. It shows that the composition of a linear mapping with a strongly convex mapping is restricted strongly convex, even if the linear mapping has a nonempty nullspace, so that the composition cannot be strongly convex. Let $\mathcal{L}_\delta^f := \{x \in \mathbb{R}^n \,|\, f(x) \leq \delta\}$ denote a level set of a function $f : \mathbb{R}^n \to \mathbb{R}$.

EXAMPLE 2.5. *Let $A \in \mathbb{R}^{m \times n}$, $\delta > 0$, and $g : \mathbb{R}^m \to \mathbb{R}$ be differentiable such that the function $f(x) := g(Ax)$ has a nonempty set of minimizers. If g is strongly convex on $\mathcal{L}^g_{f_{min}+\delta}$ then f is restricted strongly convex on $\mathcal{L}^f_{f_{min}+\delta}$. If g is strongly convex on all of $\mathbb{R}^n$ then f is restricted strongly convex on all of $\mathbb{R}^n$. Obviously, if g is twice* *continuously differentiable then so is f.*

See also [60] for related results. Further examples are discussed in the next two subsections.

2.1. Convex quadratic splines. Several types of optimization problems can be formulated as unconstrained minimization problems with convex quadratic splines, e.g. the Basis Pursuit problem, least distance problems or convex quadratic programs, see [32, 35] and the references therein. Restricted strong convexity of convex quadratic splines is a consequence of Lemma 2.4 and the following error bound result.

Theorem 2.6 (Theorem 2.7 in [31]). *For any convex piecewise quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ with a nonempty set of minimizers and any $\delta > 0$ there exists a constant $\gamma > 0$ such that (1.5) holds for all $x \in \mathcal{L}^f_{f_{min}+\delta}$.*

Example 2.7. *Any convex quadratic spline $f : \mathbb{R}^n \to \mathbb{R}$ with a nonempty set of minimizers is restricted strongly convex on $\mathcal{L}^f_{f_{min}+\delta}$ for all $\delta > 0$.*

We point out that any convex quadratic spline also has a Lipschitz-continuous gradient. For further interesting properties we refer to [10]. Now we aim to characterize convex quadratic splines that are restricted strongly convex on all of $\mathbb{R}^n$.

Theorem 2.8. *A convex quadratic spline $f : \mathbb{R}^n \to \mathbb{R}$ is restricted strongly convex on all of $\mathbb{R}^n$ and has a bounded set of minimizers if and only if f is coercive.*

Proof. Because of Corollary 2.3 it remains to prove the "if"-part. Coercivity of f obviously implies a nonempty bounded set of minimizers X_f. Since f is piecewise quadratic there are finitely many polyhedral sets $C_j \subset \mathbb{R}^n$, $j = 1, \dots, p$, whose union equals $\mathbb{R}^n$ and relative to each of which $f(x)$ is given by a convex linear-quadratic function

$$f(x) = \tfrac{1}{2} \langle x \, , A_j x\rangle + \langle a_j \, , x\rangle + \alpha_j \quad , \quad x \in C_j \, ,$$

with symmetric positive semidefinite matrices $A_j \in \mathbb{R}^{n \times n}$, vectors $a_j \in \mathbb{R}^n$ and scalars $\alpha_j \in \mathbb{R}$. To show that f is restricted strongly convex on all of $\mathbb{R}^n$ we use Example 2.7 on a level set and analyse the behaviour of f on unbounded regions C_j. By Corollary 3.53 in [55] each polyhedral region C_j can be written in the form $C_j = D_j + K_j$ with a bounded polyhedral set D_j and a closed convex cone K_j, $j = 1, \dots, p$. Then we have

$$d_b := \max\big\{ \max_{j=1,\dots,p} \max_{y \in D_j} \|y\|_2 \, , \max_{y \in X_f} \|y\|_2 \big\} < \infty \, .$$

Let J_∞ be the set of all indices j such that C_j is unbounded. Especially we have $K_j \neq \{0\}$ for all $j \in J_\infty$. Fix some $y \in D_j$. Then for any $0 \neq z \in K_j$ and $t > 0$ we have $y + t \cdot z \in C_j$ and thus can write

$$f(y + t \cdot z) = f(y) + t \cdot \langle \nabla f(y) \, , z\rangle + \frac{t^2}{2} \langle z \, , A_j z\rangle \ .$$

By letting $t \to \infty$ it follows that $\langle z \, , A_j z\rangle > 0$, because f is coercive. Hence we also have

$$\lambda := \min_{j \in J_\infty} \min_{z \in K_j, \|z\|_2 = 1} \langle z \, , A_j z\rangle > 0 \, .$$

Choose $R > d_b$ large enough such that

$$\mu_R := \sqrt{\lambda} - (\sqrt{\lambda} + \sqrt{L}) \cdot \frac{2d_b}{R - d_b} > 0$$

and set

$$f_R := \max_{x \in \mathbb{R}^n, \|x\|_2 \le R} f(x) \, .$$

By Example 2.7 there is a constant $\nu_R > 0$ such that the restricted secant inequality (1.4) holds for all $x \in \mathcal{L}_{f_R}$. Now let $x \notin \mathcal{L}_{f_R}$ be arbitrary. Then we have $\|x\|_2 > R > d_b$ and hence x must lie in some unbounded region, i.e. $x = y + z \in C_j$ for some $j \in J_\infty$ and $y \in D_j$, $z \in K_j$. We set $\hat{x} := P_{X_f}(x)$. Then we have $\|x - \hat{x}\|_2 \ge R - d_b$ and hence

$$\|y - \hat{x}\|_2 \le 2d_b \le \frac{2d_b}{R - d_b} \cdot \|x - \hat{x}\|_2 \ .$$

With this we can estimate

$$\begin{aligned}
\sqrt{\langle x - \hat{x} \, , A_j(x - \hat{x}) \rangle} &\ge \sqrt{\langle z \, , A_j z \rangle} - \sqrt{\langle y - \hat{x} \, , A_j(y - \hat{x}) \rangle} \\
&\ge \sqrt{\lambda} \cdot \|z\|_2 - \sqrt{L} \cdot \|y - \hat{x}\|_2 \\
&\ge \sqrt{\lambda} \cdot \|x - \hat{x}\|_2 - (\sqrt{\lambda} + \sqrt{L}) \cdot \|y - \hat{x}\|_2 \\
&\ge \left(\sqrt{\lambda} - (\sqrt{\lambda} + \sqrt{L}) \cdot \frac{2d_b}{R - d_b} \right) \cdot \|x - \hat{x}\|_2 \\
&= \mu_R \cdot \|x - \hat{x}\|_2 \ ,
\end{aligned}$$

which yields

$$\langle x - \hat{x} \, , A_j(x - \hat{x}) \rangle \ge \mu_R^2 \cdot \|x - \hat{x}\|_2^2 \ . \tag{2.8}$$

Consider the line segment $x_t := \hat{x} + t(x - \hat{x})$ for $t \in [0, 1]$. Since $\|\hat{x}\|_2 < R < \|x\|_2$ there is a unique point $t_1 \in (0, 1)$ such that $\|x_{t_1}\|_2 = R < \|x_t\|_2$ for all $t \in (t_1, 1]$. Hence $x_{t_1} \in \mathcal{L}_{f_R}$. Furthermore we can find finitely many points $t_1 < t_2 < \ldots < t_N = 1$ and corresponding indices j_k such that $x_t \in C_{j_k}$ for all $t \in [t_k, t_{k+1}]$ and $k = 1, \ldots, N-1$. It follows that all C_{j_k} with $k > 1$ are unbounded and (2.8) also holds with x and A_j replaced by x_{t_k} and A_{j_k} respectively (note that we also have $\hat{x} = P_{X_f}(x_{t_k})$). Then we can write

$$\begin{aligned}
&\langle \nabla f(x) \, , x - \hat{x} \rangle \\
&= \sum_{k=1}^{N-1} \left\langle \nabla f(x_{t_{k+1}}) - \nabla f(x_{t_k}) \, , x - \hat{x} \right\rangle + \langle \nabla f(x_{t_1}) \, , x - \hat{x} \rangle \\
&= \sum_{k=1}^{N-1} \left\langle x_{t_{k+1}} - x_{t_k} \, , A_{j_k}(x - \hat{x}) \right\rangle + \langle \nabla f(x_{t_1}) \, , x - \hat{x} \rangle \\
&= \sum_{k=1}^{N-1} \frac{t_{k+1} - t_k}{t_{k+1}^2} \cdot \left\langle x_{t_{k+1}} - \hat{x} \, , A_{j_k}(x_{t_{k+1}} - \hat{x}) \right\rangle + \frac{1}{t_1} \langle \nabla f(x_{t_1}) \, , x_{t_1} - \hat{x} \rangle
\end{aligned}$$

Finally we estimate the first summand by (2.8) and the second summand by (1.4) to get

$$\begin{aligned}\langle \nabla f(x)\,, x-\hat{x}\rangle &\geq \sum_{k=1}^{N-1} \frac{t_{k+1}-t_k}{t_{k+1}^2}\cdot \mu_R^2 \cdot \left\|x_{t_{k+1}}-\hat{x}\right\|_2^2 + \frac{\nu_R}{t_1}\cdot \left\|x_{t_1}-\hat{x}\right\|_2^2 \\ &\geq \min\{\mu_R^2, \nu_R\}\cdot \left\|x-\hat{x}\right\|_2^2\ ,\end{aligned}$$

from which the assertion follows. □

Remark 2.9. *Let the polyhedral sets $C_j \subset \mathbb{R}^n$ be the defining regions of f as in the preceding proof. Since X_f is polyhedral too, we have*

$$d_{min} := \min\{\operatorname{dist}(C_j, X_f)\,|\, C_j \cap X_f = \emptyset\} > 0\,,$$

and hence $\operatorname{dist}(x, X_f) \geq d_{min}$ *for all $x \in C_j$ with $C_j \cap X_f = \emptyset$. As a consequence the iterates of the descent methods in section 3 will always stay in a* solution region C_j *(i.e. $C_j \cap X_f \neq \emptyset$) after finitely many iterations.*

2.2. Unconstrained duals to linearly constrained optimization problems. To establish the restricted strong convexity result we need the concepts of calmness of a set-valued mapping [55] and linear regularity of a collection of convex sets [3]. Let $B_\epsilon(x)$ denote the 2-norm ball with radius $\epsilon > 0$ and center $x \in \mathbb{R}^n$.

Definition 2.2. *A set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is* calm *at $\hat{x} \in \mathbb{R}^n$ if $S(\hat{x}) \neq \emptyset$ and there are constants $\epsilon, L > 0$ such that for all $x \in B_\epsilon(\hat{x})$*

$$S(x) \subset S(\hat{x}) + L\cdot \left\|x-\hat{x}\right\|_2 \cdot B_1(0)\,.$$

Such mappings are also called *locally upper Lipschitz-continuous* in [54].

Example 2.10.

(a) *Any* polyhedral multifunction, *i.e. a set-valued mapping whose graph is the union of finitely many polyhedral convex sets, is calm at each $\hat{x} \in \mathbb{R}^n$. In particular this holds for the subdifferential mapping*

$$\partial f(x) := \{x^* \in \mathbb{R}^n \,|\, f(y) \geq f(x) + \langle x^*\,, y-x\rangle \quad \textit{for all } y \in \mathbb{R}^n\}$$

of a convex piecewise quadratic function $f : \mathbb{R}^n \to \mathbb{R}$.

(b) *Let $h : \mathbb{R}^m \to \mathbb{R}$ be a convex piecewise quadratic function which is* absolutely symmetric, *i.e. $h(x_1, \ldots, x_m) = h\big(|x_{\pi(1)}|, \ldots, |x_{\pi(m)}|\big)$ for any permutation π of the indices. Then the subdifferential mapping of the convex function $g(X) := h\big(\sigma(X)\big)$ is calm at each $\hat{X} \in \mathbb{R}^{m\times n}$. In particular this holds for the* nuclear norm $\|X\|_* := \|\sigma(X)\|_1$, *the* spectral norm $\|X\|_2 := \|\sigma(X)\|_\infty$ *and the objective function in* (1.8), $g(X) = \frac{1}{2}\cdot \|X\|_F^2 + \tau \cdot \|X\|_*$, *where $\|X\|_F := \|\sigma(X)\|_2$ denotes the* Frobenius norm.

(c) *The subdifferential mapping of the objective function in* (1.9), $g(X_1, X_2) = \frac{1}{2}\cdot \|X_1\|_F^2 + \tau_1 \cdot \|X_1\|_* + \frac{1}{2}\cdot \|X_2\|_F^2 + \tau_2 \cdot \|X_2\|_1$, *is calm at each $(\hat{X_1}, \hat{X_2}) \in \mathbb{R}^{m\times n}\times \mathbb{R}^{m\times n}$. Here $\|X\|_1$ denotes the 1-norm of all entries of a matrix X.*

Proof. For (a) see Proposition 1 in [54], and (c) follows from (a) and (b). To show (b) we may without loss of generality assume that $m \leq n$. For $d \in \mathbb{R}^m$ we denote by $\operatorname{diag}(d) \in \mathbb{R}^{m\times n}$ a (rectangular) matrix which is zero except for the vector d on its main diagonal. Let $X = U \operatorname{diag}\big(\sigma(X)\big)V^T$ be a singular value decomposition (SVD) of a given matrix $X \in \mathbb{R}^{m\times n}$, i.e. $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal and $\sigma(X)$ is the vector of the singular values $\sigma_1(X) \geq \ldots \geq \sigma_m(X) \geq 0$ of X. It follows from Theorem 7.1 in [29] that the subdifferential of g at X is given by (see also [62])

$$\partial g(X) = \{U \operatorname{diag}(d)V^T \mid X = U \operatorname{diag}\big(\sigma(X)\big)V^T \text{ is a SVD of } X \text{ and } d \in \partial h\big(\sigma(X)\big)\}\,.$$

Fix some $\hat{X} \in \mathbb{R}^{m\times n}$. It follows from Lemma 4.3 in [58] that for any X near $\hat{X}$ and any SVD of $X = U \operatorname{diag}\big(\sigma(X)\big)V^T$ there are orthogonal matrices $\hat{U} \in \mathbb{R}^{m\times m}$ and $\hat{V} \in \mathbb{R}^{n\times n}$ such that $U - \hat{U} = \mathcal{O}(\|X - \hat{X}\|_2)$, $V - \hat{V} = \mathcal{O}(\|X - \hat{X}\|_2)$, and both $\hat{U}^T\hat{X}\hat{X}^T\hat{U}$ and $\hat{V}^T\hat{X}^T\hat{X}\hat{V}$ are diagonal where the first m diagonal elements are $\sigma_1(\hat{X})^2, \ldots, \sigma_m(\hat{X})^2$ and the remaining ones are zero. For the positive singular values $\sigma_i(\hat{X}) > 0$ we redefine the column vectors $\hat{U}_i := \frac{1}{\sigma_i(\hat{X})}\hat{X}\hat{V}_i$. Then $\hat{U}$ remains orthogonal and $\hat{X} = \hat{U} \operatorname{diag}\big(\sigma(\hat{X})\big)\hat{V}^T$ is a SVD of $\hat{X}$. Since the singular value functions $\sigma_i(X)$ are Lipschitz-continuous (cf. [30, 58]) we still have $U - \hat{U} = \mathcal{O}(\|X - \hat{X}\|_2)$. By (a) for X near $\hat{X}$ and any $d \in \partial h\big(\sigma(X)\big)$ there exists some $\hat{d} \in \partial h\big(\sigma(\hat{X})\big)$ with $d - \hat{d} = \mathcal{O}(\|X - \hat{X}\|_2)$. Hence for any $X^* \in \partial g(X)$ there exists some $\hat{X}^* \in \partial g(\hat{X})$ with $X^* - \hat{X}^* = \mathcal{O}(\|X - \hat{X}\|_2)$. □

The next theorem shows that the standard constraint qualification together with a boundedness assumption implies linear regularity. For a closed convex set $C \subset \mathbb{R}^n$ we denote by $\operatorname{rint}(C)$ the relative interior of C.

Theorem 2.11 (Corollary 6 in [3]). *Suppose $C_1, \ldots C_m \subset \mathbb{R}^n$ are closed convex sets, where $C_{r+1}, \ldots, C_m$ are polyhedral for some $r \in \{0, \ldots, m\}$. If $C := \bigcap_{i=1}^m C_i$ is bounded and $\bigcap_{i=1}^r \operatorname{rint}(C_i) \cap \bigcap_{i=r+1}^m C_i \neq \emptyset$, then the collection $\{C_1, \ldots C_m\}$ is* linearly regular, *i.e. there exists $\gamma > 0$ such that for all $x \in \mathbb{R}^n$ we have*

$$\operatorname{dist}(x, C) \leq \gamma \cdot \sum_{i=1}^m \operatorname{dist}(x, C_i)\,.$$

For a matrix $A \in \mathbb{R}^{n\times m}$ we denote by $\mathcal{R}(A)$ the range of A and by $\mathcal{N}(A)$ the nullspace of A. The convex conjugate g^* of a convex function $g : \mathbb{R}^m \to \mathbb{R}$ is defined as $g^*(y^*) := \sup_{y\in\mathbb{R}^m} \langle y^*, y\rangle - g(y)$, see [55]. Note that the optimization problems (1.8) and (1.9), or more generally problems of the form

$$\min_{Y\in\mathbb{R}^{m_1\times m_2}} \tilde{g}(Y) \quad s.t. \quad \mathcal{A}(Y) = b\,,$$

with matrix variables $Y \in \mathbb{R}^{m_1\times m_2}$, objective function $\tilde{g} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}$ and a linear operator $\mathcal{A} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}^n$ can equivalently be written in the form (1.6) by the usual way of identifying the matrix $Y = \operatorname{mat}(y)$ with the vector $y := \operatorname{vec}(Y) \in \mathbb{R}^{m_1\cdot m_2}$ of all of its columns, the operator $\mathcal{A}$ with a matrix $A \in \mathbb{R}^{n\times(m_1\cdot m_2)}$ such that $\mathcal{A}(Y) = A \cdot y$, and the objective function with $g(y) := \tilde{g}\big(\operatorname{mat}(y)\big) = \tilde{g}(Y)$. The subdifferentials are then related by $\partial g(y) = \operatorname{vec}\Big(\partial\tilde{g}\big(\operatorname{mat}(y)\big)\Big)$.

Theorem 2.12. *Consider the linearly constrained optimization problem* (1.6) *with* $A \in \mathbb{R}^{n\times m}$, $b \in \mathcal{R}(A)$, *and strongly convex* $g : \mathbb{R}^m \to \mathbb{R}$. *Then* (1.6) *has a unique solution* $\hat{y}$ *and the objective function* $f(x) := g^*(A^T x) - \langle b\,, x\rangle$ *of the unconstrained dual* (1.1) *to* (1.6) *is convex with a Lipschitz-continuous gradient* $\nabla f(x) = A\nabla g^*(A^T x) - b$ *and nonempty set of minimizers* $X_f = \{x \in \mathbb{R}^n \mid A\nabla g^*(A^T x) = b\}$. *If the subdifferential mapping of g is calm at* $\hat{y}$ *and if the collection* $\{\partial g(\hat{y}), \mathcal{R}(A^T)\}$ *is linearly regular, then f is restricted strongly convex on* $\mathcal{L}^f_{f_{min}+\delta}$ *for any* $\delta > 0$. *Especially this holds in any of the following cases:*

(a) g is a convex piecewise quadratic function.

(b) ∂g *is calm at* $\hat{y}$ *and the constraint qualification* $\operatorname{rint}\big(\partial g(\hat{y})\big) \cap \mathcal{R}(A^T) \neq \emptyset$ *holds.*

(c) $g(y) = \tilde{g}\big(\operatorname{mat}(y)\big)$ *where* $\tilde{g} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}$ *is of the form* $\tilde{g}(Y) = h\big(\sigma(Y)\big)$ *for matrices Y, and with an absolutely symmetric convex piecewise quadratic function h, and the constraint qualification* $\operatorname{rint}\big(\partial g(\hat{y})\big) \cap \mathcal{R}(A^T) \neq \emptyset$ *holds.*

(d) g has a Lipschitz-continuous gradient. If g is twice continuously differentiable then so is f.

Proof. At first we note that in cases (a) and (c) calmness of ∂g at $\hat{y}$ is automatically fulfilled by Example 2.10. By Proposition 12.60 in [55] it holds that if g is strongly convex with constant $\nu_g > 0$ then ∇g^* is Lipschitz-continuous with constant $L_{g^*} := \frac{1}{\nu_g}$, and therefore ∇f is Lipschitz-continuous with constant $L := L_{g^*} \cdot \|A\|_2^2$. Furthermore, if g has a Lipschitz-continuous gradient, then g^* is strongly convex, and hence in case (d) restricted strong convexity of f follows from Example 2.5. The assertion about twice continuous differentiability of g^*, and hence f, is an immediate consequence of the classical Legendre transform, cf. Example 11.9 in [55]. Note that the optimal solution $\hat{y}$ of (1.6) fulfills $\partial g(\hat{y}) \cap \mathcal{R}(A^T) \neq \emptyset$ and $\hat{y} = \nabla g^*(A^T x)$ for all $x \in X_f$. Furthermore, $\partial g(\hat{y}) \cap \mathcal{R}(A^T)$ is bounded because the closed convex set $\partial g(\hat{y})$ is bounded. Hence by Theorem 2.11 all cases (a), (b) and (c) imply linear regularity of $\{\partial g(\hat{y}), \mathcal{R}(A^T)\}$, which in turn implies that there exists $\gamma > 0$ such that for all $y^* \in \mathcal{R}(A^T)$ we have

$$\operatorname{dist}\big(y^*, \partial g(\hat{y}) \cap \mathcal{R}(A^T)) \leq \gamma \cdot \operatorname{dist}\big(y^*, \partial g(\hat{y})\big)\,. \tag{2.9}$$

We prove restricted strong convexity of f by contradiction. Assume that f is not restricted strongly convex on $\mathcal{L}^f_{f_{min}+\delta}$ for some $\delta > 0$. Then there exists a sequence $x_n \in \mathcal{L}^f_{f_{min}+\delta}$ such that $x_n \notin X_f$ and

$$\lim_{n\to\infty} \frac{\big\langle \nabla f(x_n)\,, x_n - P_{X_f}(x_n)\big\rangle}{\big\|x_n - P_{X_f}(x_n)\big\|_2^2} = 0\,. \tag{2.10}$$

We set $y_n^* := A^T x_n \in \mathcal{R}(A^T)$ and $y_n := \nabla g^*(y_n^*)$. Hence we have $y_n^* \in \partial g(y_n)$ and $A^T P_{X_f}(x_n) \in \partial g(\hat{y})$, but $y_n^* \notin \partial g(\hat{y})$. Since g is strongly convex there is a constant $c > 0$ such that we can estimate the denominator in (2.10) by

$$\big\langle \nabla f(x_n)\,, x_n - P_{X_f}(x_n)\big\rangle = \big\langle y_n - \hat{y}\,, y_n^* - A^T P_{X_f}(x_n)\big\rangle \geq c \cdot \|y_n - \hat{y}\|_2^2\,.$$

To y_n^* we can find some $\hat{x}_n \in X_f$ such that $\hat{y}_n^* := P_{\partial g(\hat{y})\cap\mathcal{R}(A^T)}(y_n^*) = A^T \hat{x}_n \in \partial g(\hat{y})$. By (2.9) we have

$$\|y_n^* - \hat{y}_n^*\|_2 = \operatorname{dist}\big(y_n^*, \partial g(\hat{y}) \cap \mathcal{R}(A^T)\big) \leq \gamma \cdot \operatorname{dist}\big(y_n^*, \partial g(\hat{y})\big)\,.$$

Let $\sigma_{min}(A) > 0$ denote the minimal positive singular value of A. Since we have $X_f + \mathcal{N}(A^T) \subset X_f$ we can estimate the nominator in (2.10) by

$$\begin{aligned}\left\|x_n - P_{X_f}(x_n)\right\|_2 &\leq \left\|x_n - \left((\hat{x}_n + P_{\mathcal{N}(A^T)}(x_n - \hat{x}_n)\right)\right\|_2 \\ &\leq \tfrac{1}{\sigma_{min}(A)} \cdot \left\|A^T x_n - A^T \hat{x}_n\right\|_2 = \tfrac{1}{\sigma_{min}(A)} \cdot \|y_n^* - \hat{y}_n^*\|_2 \\ &\leq \tfrac{\gamma}{\sigma_{min}(A)} \cdot \operatorname{dist}\left(y_n^*, \partial g(\hat{y})\right).\end{aligned}$$

Hence it follows from (2.10) that

$$\lim_{n\to\infty} \frac{\|y_n - \hat{y}\|_2}{\operatorname{dist}\left(y_n^*, \partial g(\hat{y})\right)} = 0 . \tag{2.11}$$

Furthermore, g^* is coercive since g is finite everywhere, see Theorem 11.8 in [55]. This implies boundedness of $A^T \mathcal{L}^f_{f_{min}+\delta}$. Hence $\operatorname{dist}\left(y_n^*, \partial g(\hat{y})\right)$ remains bounded and it follows from (2.11) that the sequence y_n converges to $\hat{y}$. But then calmness of ∂g at $\hat{y}$ implies that $\operatorname{dist}\left(y_n^*, \partial g(\hat{y})\right) = \mathcal{O}\big(\|y_n - \hat{y}\|_2\big)$ for all n large enough, which together with (2.11) leads to a contradiction. □

3. Linearly convergent descent methods. Here we analyse the convergence behaviour of iteration (1.2) to solve the unconstrained optimization problem (1.1). Throughout this section we make the following assumption about the function f.

Assumption 3.1. *$f : \mathbb{R}^n \to \mathbb{R}$ is restricted strongly convex on the level set $\mathcal{L}_{f(x_0)}$ with constant $\nu > 0$ and has a Lipschitz-continuous gradient with constant $L > 0$.*

Furthermore we assume that at each iteration we have $g_k := \nabla f(x_k) \neq 0$, because otherwise we stop iterating after finitely many iterations with x_k being a minimizer of f, and the estimates for the rate of decrease remain valid as long as $g_k \neq 0$. We set $f_k := f(x_k)$ and consider the following line search strategies to choose suitable step lengths $t_k > 0$ in (1.2) for *descent directions* d_k, i.e. $\langle d_k\,, g_k\rangle < 0$:

(LS1) **Explicit value**: Set $t_k := -\dfrac{c \cdot \langle g_k\,, d_k\rangle}{L \cdot \|d_k\|_2^2}$ for some constant $c \in (0, 2)$.

(LS2) **Wolfe conditions**: Let $0 < \alpha < \beta < 1$ and choose t_k such that

$$f_{k+1} \leq f_k + \alpha \cdot t_k \cdot \langle g_k\,, d_k\rangle \tag{3.1}$$

$$\langle g_{k+1}\,, d_k\rangle \geq \beta \cdot \langle g_k\,, d_k\rangle\ . \tag{3.2}$$

(LS3) **Backtracking**: Let $\eta \in (0, 1)$ and $0 < \tau_1 < \tau_2 < 1$.
If $t = 1$ satisfies $f(x_k + t \cdot d_k) \leq f_k + \eta \cdot t \cdot \langle g_k\,, d_k\rangle$ then set $t_k := t$. Otherwise choose a new $t \in [\tau_1 \cdot t, \tau_2 \cdot t]$ and repeat the test.
(This backtracking includes the Goldstein-Armijo rule of multiplying the old value t by a constant factor $\tau \in (0, 1)$, as well as polynomial interpolation methods.)

(LS4) **Exact line search**: Let $t_{max} > 1$ and $\hat{t}_k$ fulfill $\big\langle \nabla f(x_k + \hat{t}_k \cdot d_k)\,, d_k\big\rangle = 0$.
(LS4a) (without safeguard) Set $t_k := \hat{t}_k$.
(LS4b) (with safeguard) Set $t_k := \min\{\hat{t}_k, t_{max}\}$.

The explicit value in (LS1) requires knowledge of the Lipschitz-constant L. But since it requires no additional function or gradient evaluations it may be especially useful for large-scale applications. The Wolfe conditions (LS2) guarantee positive definite matrix-updates in Quasi-Newton methods. Backtracking (LS3) is often preferred in Newton-like methods when the approximate Hessian matrices B_k are close to the true Hessian. Assumption (3.1) ensures that a step length $t_k > 0$ satisfying (LS2) or (LS3) can always be found, cf. [45]. In general $\hat{t}_k$ in (LS4) need not exist if the set of minimizers is unbounded, as the following counterexample demonstrates.

Counterexample 3.2. *Let $C \subset \mathbb{R}^2$ be the closed convex and unbounded set $C := \{(u,v) \in \mathbb{R}^2 \,|\, u > 0\,,\, v \geq \frac{1}{u}\}$. Then the function $f(x) := \frac{1}{2}\left\|x - P_C(x)\right\|_2^2$ has a Lipschitz-continuous gradient $\nabla f(x) = x - P_C(x)$ and is restricted strongly convex on all of $\mathbb{R}^2$ with an unbounded set of minimizers $X_f = C$. For $x_0 := \begin{pmatrix}0\\0\end{pmatrix}$ and $B_0^{-1} := \begin{pmatrix}2 & -1\\ -1 & 1\end{pmatrix}$ we have $P_C(x_0) = \begin{pmatrix}1\\1\end{pmatrix}$, $g_0 = -\begin{pmatrix}1\\1\end{pmatrix}$, $d_0 := -B_0^{-1}\cdot g_0 = \begin{pmatrix}1\\0\end{pmatrix}$. Hence for all $t > 0$ we have $x_0 + t\cdot d_0 = \begin{pmatrix}t\\0\end{pmatrix} \notin C$ and therefore*

$$0 < f(x_0 + t\cdot d_0) \leq \frac{1}{2}\left\|\begin{pmatrix}t\\0\end{pmatrix} - \begin{pmatrix}t\\ \frac{1}{t}\end{pmatrix}\right\|_2^2 = \frac{1}{2t^2} \longrightarrow 0 \quad \textit{for} \quad t \to \infty\,,$$

i.e. the infimum is not attained.

But if f has a bounded set of mimimizers X_f then it follows from (2.5) that the level set $\mathcal{L}_{f(x_0)}$ is bounded, and hence a minimizer $\hat{t}_k$ as in (LS4) exists. For convex quadratic splines $\hat{t}_k$ in (LS4) always exists, even if X_f is unbounded, and an exact line search can often be performed cheaply, see [31, 34, 40, 48]. Under Assumption (3.1) all these line search methods guarantee sufficient decrease of the function value at each iteration, and provide some control on the step length.

Lemma 3.3. *All line searches (LS1)–(LS4) imply one of the following decrease conditions: There are constants $c_1, c_2 > 0$ such that at each iteration we have*

$$f_{k+1} \leq f_k - c_1 \cdot \frac{\left\langle g_k\,, d_k\right\rangle^2}{\left\|d_k\right\|_2^2} \tag{3.3}$$

$$\textit{or} \quad f_{k+1} \leq f_k + c_2 \cdot \left\langle g_k\,, d_k\right\rangle\,. \tag{3.4}$$

Furthermore, for search directions of the form (1.3) *we have*

(LS1) $\frac{c}{L\cdot\left\|B_k^{-1}\right\|_2} \leq t_k \leq \frac{c\cdot\|B_k\|_2}{L}$,

(LS2) $\frac{1-\beta}{L\cdot\left\|B_k^{-1}\right\|_2} \leq t_k \leq \frac{\|B_k\|_2}{\alpha\cdot\nu}$,

(LS3) $t_k = 1$ *or* $\frac{\tau_1\cdot(1-\eta)}{L\cdot\left\|B_k^{-1}\right\|_2} \leq t_k \leq 1$,

(LS4a) $\frac{1}{L\cdot\left\|B_k^{-1}\right\|_2} \leq t_k$,

(LS4b) $t_k = t_{max}$ *or* $\frac{1}{L\cdot\left\|B_k^{-1}\right\|_2} \leq t_k \leq t_{max}$.

Proof. (LS1) Since ∇f is Lipschitz-continuous we can estimate by (2.2)

$$f_{k+1} \leq f_k + t_k\cdot\left\langle g_k\,, d_k\right\rangle + t_k^2\cdot\frac{L}{2}\cdot\left\|d_k\right\|_2^2 \leq f_k - \frac{c\cdot(2-c)}{2L}\cdot\frac{\left\langle g_k\,, d_k\right\rangle^2}{\left\|d_k\right\|_2^2}\,,$$

which implies (3.3) for $c_1 := \frac{c\cdot(2-c)}{2L}$. The estimates for t_k follow from

$$\frac{-\langle g_k\,,d_k\rangle}{\|B_k\|_2} \leq \|d_k\|_2^2 \leq -\left\|B_k^{-1}\right\|_2 \cdot \langle g_k\,,d_k\rangle\ .$$

(LS2) It is well known that the Wolfe conditions imply (3.3). But since we need some intermediate result we repeat the proof here. By the second Wolfe condition (3.2) and Lipschitz-continuity of ∇f we have

$$-t_k\cdot(1-\beta)\cdot\langle g_k\,,d_k\rangle \leq \langle g_{k+1}-g_k\,,x_{k+1}-x_k\rangle \leq L\cdot t_k^2\cdot\|d_k\|_2^2\ , \tag{3.5}$$

which implies

$$t_k \geq -\frac{(1-\beta)\cdot\langle g_k\,,d_k\rangle}{L\cdot\|d_k\|_2^2} \geq \frac{1-\beta}{L\cdot\left\|B_k^{-1}\right\|_2}\ . \tag{3.6}$$

Inserting this into the first Wolfe condition (3.1) yields (3.3) for $c_1 := \frac{\alpha\cdot(1-\beta)}{L}$. Furthermore it follows from (3.1) and (2.6) (restricted strong convexity) that

$$t_k \leq \frac{f_k - f_{k+1}}{-\alpha\cdot\langle g_k\,,d_k\rangle} \leq \frac{\|g_k\|_2^2}{-\alpha\cdot\nu\cdot\langle g_k\,,d_k\rangle} \leq \frac{\|B_k\|_2}{\alpha\cdot\nu}\ . \tag{3.7}$$

(LS3) This is proven in Lemma 4.1 of [4] and the proof given there is valid under Assumption (3.1).
(LS4a) Since $f(x_k + \hat{t}_k\cdot d_k) \leq f(x_k + \tilde{t}_k\cdot d_k)$ for $\tilde{t}_k$ as in (LS1) (and $c := 1$), we have (3.3) for $c_1 := \frac{1}{2L}$. The lower estimate for t_k holds, because the second Wolfe condition (3.2) is fulfilled with equality for $\beta := 0$.
(LS4b) In case $\langle g_{k+1}\,,d_k\rangle \geq \frac{1}{t_{max}}\cdot\langle g_k\,,d_k\rangle$ the second Wolfe condition (3.2) is satisfied with $\beta := \frac{1}{t_{max}}$. Hence we have $t_k \geq -\frac{(1-\beta)\cdot\langle g_k\,,d_k\rangle}{L\cdot\|d_k\|_2^2} =: \tilde{t}_k$, where $\tilde{t}_k$ is a step length of the form (LS1). Since $f(x_k + t\cdot d_k)$ is monotonically decreasing for $t\in[0,\hat{t}_k]$ we have $f_{k+1} \leq f(x_k + \tilde{t}_k\cdot d_k)$ and (3.3) follows as for (LS1). In case $\langle g_{k+1}\,,d_k\rangle < \frac{1}{t_{max}}\cdot\langle g_k\,,d_k\rangle < 0$ we must have $t_k = t_{max}$. By convexity of f we then estimate

$$f_k - f_{k+1} \geq \langle g_{k+1}\,,x_k - x_{k+1}\rangle = -t_{max}\cdot\langle g_{k+1}\,,d_k\rangle > -\langle g_k\,,d_k\rangle\ ,$$

which yields (3.4) for $c_2 := 1$. □

We do not know whether the exact minimizer in (LS4a) can be bounded from above, even if we always choose the first minimizer.

3.1. Quasi-Newton methods. Consider iteration (1.2) with any of the line searches (LS1)–(LS4) and search directions of the form (1.3).

THEOREM 3.4. *If all matrices B_k and B_k^{-1} in* (1.3) *are uniformly bounded, i.e. for some constant $M > 0$ we have $\|B_k\|_2\,,\left\|B_k^{-1}\right\|_2 \leq M$, then the following linear convergence results hold: There exist constants $q\in(0,1)$ and $\gamma_1 > 0$ such that*

$$f_{k+1} - f_{min} \leq q\cdot(f_k - f_{min}) \leq (f_0 - f_{min})\cdot q^{k+1}\,, \tag{3.8}$$

$$\operatorname{dist}(x_k, X_f) \leq \gamma_1\cdot q^{\frac{k}{2}}\ . \tag{3.9}$$

Furthermore any but the exact linesearch (LS4a) without safeguard guarantees that the iterates x_k converge linearly to some minimizer $\hat{x} \in X_f$, i.e. there is a constant $\gamma_2 > 0$ such that

$$\|x_k - \hat{x}\|_2 \leq \gamma_2 \cdot q^{\frac{k}{2}} . \tag{3.10}$$

Proof. By Lemma 3.3 all line searches imply at each iteration (3.3) or (3.4), where (3.3) in turn implies

$$f_{k+1} \leq f_k - \frac{c_1}{(\|B_k\|_2 \cdot \|B_k^{-1}\|_2)^2} \cdot \|g_k\|_2^2 \leq f_k - \frac{c_1}{M^2} \cdot \|g_k\|_2^2 ,$$

and (3.4) in turn implies

$$f_{k+1} \leq f_k - \frac{c_2}{\|B_k\|_2} \cdot \|g_k\|_2^2 \leq f_k - \frac{c_2}{M} \cdot \|g_k\|_2^2 .$$

Hence in any case there is a constant $c > 0$ such that

$$f_{k+1} - f_{min} \leq f_k - f_{min} - c \cdot \|g_k\|_2^2 . \tag{3.11}$$

It follows inductively that all iterates x_k remain in $\mathcal{L}_{f(x_0)}$. Since f is assumed to be restricted strongly convex on $\mathcal{L}_{f(x_0)}$ we can use inequality (2.6) to estimate

$$-\|g_k\|_2^2 \leq -\nu \cdot (f_k - f_{min}) ,$$

from which (3.8) follows with $q := 1 - c \cdot \nu$. Together with (2.5) this immediately implies (3.9). It remains to prove convergence of the iterates x_k. By (2.7) we have

$$\begin{aligned} \|x_{k+1} - x_k\|_2 = t_k \cdot \|d_k\|_2 &\leq t_k \cdot \|B_k^{-1}\|_2 \cdot \|g_k\|_2 \\ &\leq t_k \cdot \|B_k^{-1}\|_2 \cdot L \cdot \sqrt{\frac{2}{\nu}} \cdot \sqrt{f_k - f_{min}} . \end{aligned}$$

Since $\|B_k\|_2 \leq M$ we know by Lemma 3.3 that all but the exact linesearch (LS4a) guarantee that the step length t_k remains bounded. Hence together with (3.8) it follows that for some constant $\gamma > 0$ we have $\|x_{k+1} - x_k\|_2 \leq \gamma \cdot q^{\frac{k}{2}}$. From this we conclude that x_k is a Cauchy-sequence and hence convergent to some $\hat{x} \in \mathbb{R}^n$ with the rate (3.10) for $\gamma_2 := \frac{\gamma}{1-\sqrt{q}}$. Finally (3.9) shows that we indeed have $\hat{x} \in X_f$. □

REMARK 3.5. *For (LS1), (LS2) and (LS4a) it actually suffices that the condition numbers $\|B_k\|_2 \cdot \|B_k^{-1}\|_2$ of B_k are uniformly bounded.*

One of the referees pointed out that a part of Theorem 3.4 follows from a more general result of [61]. There the convergence of a class of projected gradient methods for the solution of constrained optimization problems was analyzed for functions f fulfilling some error bounds, which in the unconstrained case are fulfilled under our Assumption 3.1. For unconstrained problems Theorem 8 in [61] reads as follows.

THEOREM 3.6. *Let $\beta, \gamma > 0$ such that the iterates x_k satisfy*

$$\begin{aligned} x_{k+1} &= x_k - t_k \cdot g_k + e_k , \\ \|e_k\|_2 &\leq \beta \cdot \|x_k - x_{k+1}\|_2 , \\ f_k - f_{k+1} &\geq \gamma \cdot \|x_k - x_{k+1}\|_2^2 , \end{aligned} \tag{3.12}$$

with step lengths t_k such that $\inf_k t_k > 0$. Then the function values f_k decrease linearly, i.e. (3.8) *holds.*

Indeed by setting $e_k := t_k \cdot (g_k - B_k^{-1} g_k)$ iteration (1.2), (1.3) can be written in the form (3.12). For uniformly bounded $\|B_k\|_2 , \left\|B_k^{-1}\right\|_2 \leq M$ we have

$$\|e_k\|_2 \leq (1+M) \cdot t_k \cdot \|g_k\|_2 \leq (1+M) \cdot M \cdot \|x_k - x_{k+1}\|_2 \ ,$$

and $\|x_k - x_{k+1}\|_2 \leq t_k \cdot M \cdot \|g_k\|_2$. By Lemma 3.3 all line searches imply (3.11) and all but (LS4a) guarantee that the step length remains bounded and away from zero i.e. $0 < c_1 \leq t_k \leq c_2$. Hence we also have $\inf_k t_k > 0$ and

$$f_k - f_{k+1} \geq c \cdot \|g_k\|_2^2 \geq \frac{c}{(M \cdot c_2)^2} \cdot \|x_k - x_{k+1}\|_2^2 \ .$$

Nevertheless we decided to present our proof of Theorem 3.4 because of its simplicity and because we do not see how to infer linear convergence from the result of [61] for the exact line search (LS4a), which is meaningful for convex quadratic splines.

Now we analyse the convergence behaviour when the matrices B_k in (1.3) are obtained by the BFGS update formula

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{\langle s_k \, , B_k s_k \rangle} + \frac{y_k y_k^T}{\langle y_k \, , s_k \rangle} \ , \tag{3.13}$$

where $y_k := g_{k+1} - g_k$ and $s_k := x_{k+1} - x_k$, see eg. [4, 5]. Matrix-updates of the form (3.13) remain positive definite if the initial matrix B_0 is positive definite and if $\langle y_k \, , s_k \rangle > 0$, which is guaranteed by the second Wolfe condition (3.2), see Theorem 7.8. in [18].

Theorem 3.7. *Consider iteration* (1.2) *together with the Wolfe conditions (LS2). Choose a symmetric positive definite matrix $B_0 \in \mathbb{R}^{n \times n}$ and for $k = 1, 2, \dots$ let the matrices B_k in* (1.3) *be obtained by the BFGS update formula* (3.13). *Then there exist constants $q \in (0,1)$ and $\gamma > 0$ such that*

$$\begin{aligned} f_k - f_{min} &\leq (f_0 - f_{min}) \cdot q^k \ , \\ \operatorname{dist}(x_k, X_f) &\leq \gamma \cdot q^{\frac{k}{2}} \ . \end{aligned}$$

Proof. We define $p_k := \frac{\langle s_k \, , B_k s_k \rangle}{\|B_k s_k\|_2^2} = -\frac{\langle g_k \, , d_k \rangle}{\|g_k\|_2^2}$. From the first Wolfe condition (3.1) and (2.6) it follows that

$$\begin{aligned} f_{k+1} - f_{min} &\leq f_k - f_{min} - \alpha \cdot t_k \cdot p_k \cdot \|g_k\|_2^2 \\ &\leq (f_k - f_{min}) \cdot (1 - \nu \cdot \alpha \cdot t_k \cdot p_k) \\ &\leq (f_0 - f_{min}) \cdot \prod_{j=0}^{k} (1 - \nu \cdot \alpha \cdot t_j \cdot p_j) \ . \end{aligned}$$

Applying the geometric/arithmetic mean inequality twice we obtain

$$\begin{aligned} f_{k+1} - f_{min} &\leq (f_0 - f_{min}) \cdot \left(\frac{1}{k+1} \cdot \sum_{j=0}^{k} (1 - \nu \cdot \alpha \cdot t_j \cdot p_j) \right)^{k+1} \\ &\leq (f_0 - f_{min}) \cdot \left(1 - \nu \cdot \alpha \cdot \left(\prod_{j=0}^{k} t_j \cdot p_j \right)^{\frac{1}{k+1}} \right) . \end{aligned}$$

Hence to prove the assertion it suffices to show that there exists $c > 0$ such that

$$\prod_{j=0}^{k} t_j \cdot p_j \geq c^{k+1} .$$

To this end we adopt the common strategy to analyse the trace and the determinant of the matrices B_k. By (3.13) the trace obeys the recursion

$$\mathrm{tr}(B_{k+1}) = \mathrm{tr}(B_k) - \frac{\|B_k s_k\|_2^2}{\langle s_k \, , B_k s_k \rangle} + \frac{\|y_k\|_2^2}{\langle y_k \, , s_k \rangle} .$$

With (2.3) we can estimate $\frac{\|y_k\|_2^2}{\langle y_k \, , s_k \rangle} \leq L$ and thus

$$\mathrm{tr}(B_{k+1}) \leq \mathrm{tr}(B_k) - \frac{1}{p_k} + L \leq \mathrm{tr}(B_0) - \sum_{j=0}^{k} \frac{1}{p_j} + (k+1) \cdot L .$$

It follows that $\mathrm{tr}(B_k) \leq \mathrm{tr}(B_0) + k \cdot L$ and

$$\sum_{j=0}^{k} \frac{1}{p_j} \leq (k+1) \cdot (\mathrm{tr}(B_0) + L) .$$

Again applying the geometric/arithmetic mean inequality we get

$$\prod_{j=0}^{k} \frac{1}{p_j} \leq \left(\frac{1}{k+1} \sum_{j=0}^{k} \frac{1}{p_j} \right)^{k+1} \leq (\mathrm{tr}(B_0) + L)^{k+1} ,$$

which yields

$$\prod_{j=0}^{k} p_j \geq \left(\frac{1}{\mathrm{tr}(B_0) + L} \right)^{k+1} .$$

To obtain a lower estimate for $\prod_{j=0}^{k} t_j$ we use the determinant

$$\det(B_{k+1}) = \det(B_k) \cdot \frac{\langle y_k \, , s_k \rangle}{\langle s_k \, , B_k s_k \rangle} .$$

The nominator equals $-t_k^2 \cdot \langle g_k \, , d_k \rangle$ and, as a consequence of the second Wolfe condition (3.2), the denominator can be estimated as in (3.5) by

$$\langle y_k \, , s_k \rangle \geq -t_k \cdot (1 - \beta) \cdot \langle g_k \, , d_k \rangle .$$

Hence we have

$$\det(B_{k+1}) \geq \det(B_k) \cdot \frac{1-\beta}{t_k} \geq \det(B_0) \cdot \prod_{j=0}^{k} \frac{1-\beta}{t_j}\,,$$

which yields

$$\prod_{j=0}^{k} t_j \geq \frac{\det(B_0) \cdot (1-\beta)^{k+1}}{\det(B_{k+1})}\,.$$

Together with the following chain of inequalities,

$$\det(B_k) \leq \|B_k\|_2^n \leq \operatorname{tr}(B_k)^n \leq \big(\operatorname{tr}(B_0) + k \cdot L\big)^n\,,$$

we finally conclude that there is a constant $c > 0$ such that $\prod_{j=0}^{k} t_j \cdot p_j \geq c^{k+1}$. □

For large-scale problems it is preferable to use the limited memory BFGS method (L-BFGS), see [37]. At each iteration k only the m most recent vector pairs s_j, y_j for $j = k-m, \ldots, k-1$ are used to build up an approximation to the inverse Hessian $H_k = B_k^{-1}$. For the convergence analysis it is more convenient to use B_k itself. Then the updates can be described as follows:

At each iteration k an initial symmetric positive definite matrix B_k^0 (which is allowed to be different for each k) is updated only m times according to (3.13), i.e. for $j = k-m, \ldots, k-1$,

$$B_k^{m-k+j+1} = B_k^{m-k+j} - \frac{B_k^{m-k+j} s_j s_j^T B_k^{m-k+j}}{\left\langle s_j\,, B_k^{m-k+j} s_j \right\rangle} + \frac{y_j y_j^T}{\langle y_j\,, s_j\rangle}\,, \tag{3.14}$$

and $B_k := B_k^m$. In the first $k < m$ iterations we use $m = k$.

In the recent paper [2] it was reported that in several cases it can be advantageous to use a damped version of L-BFGS, which is called the limited memory damped BFGS method (L-D-BFGS). In L-D-BFGS the vectors y_j in the update formula (3.14) are replaced by

$$y_j^k = \phi_j^k \cdot y_j + (1 - \phi_j^k) \cdot B_k^{m-k+j} s_j\,, \tag{3.15}$$

where $\phi_j^k \in (0,1]$ is a damping parameter. The following lemma shows that the matrices B_k, B_k^{-1} remain symmetric positive definite and uniformly bounded for a suitable choice of the damping parameter. Hence by Theorem 3.4 we can extend the convergence result obtained in [2] for strongly convex functions to the case of restricted strongly convex functions.

Lemma 3.8. *Consider the L-D-BFGS update formula* (3.14) *with y_j replaced by y_j^k as in* (3.15) *for $j = k-m, \ldots, k-1$ and define*

$$\tau_j^k := \frac{\langle y_j\,, s_j\rangle}{\left\langle s_j\,, B_k^{m-k+j} s_j \right\rangle} \geq 0\,.$$

If all initial matrices $B_k^0, (B_k^0)^{-1}$ are symmetric positive definite and uniformly bounded and the parameters $\phi_j^k \in (0,1]$ are chosen such that for some constant $c > 0$

$$\phi_j^k \cdot \tau_j^k + (1 - \phi_j^k) \geq c \tag{3.16}$$

then all B_k, B_k^{-1} are symmetric positive definite and uniformly bounded as well.

Proof. At first we note that all line searches (LS1)–(LS4) ensure that $s_j \neq 0$ as long as $g_j \neq 0$, and by (2.3) we always have $L \cdot \langle y_j \, , s_j \rangle \geq \|y_j\|_2^2 \geq 0$. This implies $\tau_j^k \geq 0$ and that $\langle y_j \, , s_j \rangle = 0$ if and only if $y_j = 0$. By (3.15) and (3.16) we get

$$\frac{\langle y_j^k \, , s_j \rangle}{\left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle} = \phi_j^k \cdot \tau_j^k + (1 - \phi_j^k) \geq c > 0 \, .$$

From this we inductively infer that $\left\langle y_j^k \, , s_j \right\rangle > 0$ and hence all matrix updates are symmetric positive definite. It remains to show the boundedness of B_k, B_k^{-1}. Since the function $h(y,t) = \frac{\|y\|_2^2}{t}$ is convex for $(y,t) \in \mathbb{R}^n \times \mathbb{R}_+$, we consider the points

$$\left(y_j^k, \langle y_j^k \, , s_j \rangle \right) = \phi_j^k \cdot \left(y_j, \langle y_j \, , s_j \rangle \right) + (1 - \phi_j^k) \cdot \left(B_k^{m-k+j} s_j, \left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle \right)$$

on the line segment between $\left(y_j, \langle y_j \, , s_j \rangle \right)$ and $\left(B_k^{m-k+j} s_j, \left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle \right)$ to get

$$\frac{\|y_j^k\|_2^2}{\langle y_j^k \, , s_j \rangle} \leq \phi_j^k \cdot \frac{\|y_j\|_2^2}{\langle y_j \, , s_j \rangle} + (1 - \phi_j^k) \cdot \frac{\left\| B_k^{m-k+j} s_j \right\|_2^2}{\left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle} \, .$$

As in the proof of Theorem 3.7 we can now estimate the trace from above by

$$\begin{aligned} \operatorname{tr}(B_k^{m-k+j+1}) &= \operatorname{tr}(B_k^{m-k+j}) - \frac{\left\| B_k^{m-k+j} s_j \right\|_2^2}{\left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle} + \frac{\|y_j^k\|_2^2}{\langle y_j^k \, , s_j \rangle} \\ &\leq \operatorname{tr}(B_k^{m-k+j}) - \phi_j^k \cdot \frac{\left\| B_k^{m-k+j} s_j \right\|_2^2}{\left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle} + \phi_j^k \cdot \frac{\|y_j\|_2^2}{\langle y_j \, , s_j \rangle} \\ &\leq \operatorname{tr}(B_k^{m-k+j}) + L \, , \end{aligned}$$

which implies $\operatorname{tr}(B_k) \leq \operatorname{tr}(B_k^0) + m \cdot L$. Hence all B_k are uniformly bounded. For the determinant we get

$$\begin{aligned} \det(B_k^{m-k+j+1}) &= \det(B_k^{m-k+j}) \cdot \frac{\langle y_j^k \, , s_j \rangle}{\left\langle s_j \, , B_k^{m-k+j} s_j \right\rangle} \\ &= \det(B_k^{m-k+j}) \cdot \left(\phi_j^k \cdot \tau_j^k + (1 - \phi_j^k) \right) , \end{aligned}$$

which by (3.16) implies that $\det(B_k) \geq \det(B_k^0) \cdot c^m$. Hence all B_k^{-1} are uniformly bounded as well. □

Corollary 3.9. *Under the assumptions of Lemma 3.8 the assertions of Theorem 3.4 hold for the L-D-BFGS method.*

The simplest admissible choice for the damping parameter is $\phi_j^k \in (0, 1-c)$. The following choice was considered in [2] and switches to the undamped L-BFGS for certain values of τ_j^k:

$$\phi_j^k := \begin{cases} \frac{\sigma_2}{1-\tau_j^k} & , \tau_j^k < 1-\sigma_2 \\ \frac{\sigma_3}{\tau_j^k - 1} & , \tau_j^k > 1+\sigma_3 \\ 1 & , \text{otherwise} \end{cases}$$

with positive constants $\sigma_2 < 1$ and σ_3. This choice fullfills (3.16) since we have

$$\phi_j^k \cdot \tau_j^k + (1-\phi_j^k) = \begin{cases} 1-\sigma_2 & , \tau_j^k < 1-\sigma_2 \\ 1+\sigma_3 & , \tau_j^k > 1+\sigma_3 \\ \tau_j^k \geq 1-\sigma_2 & , \text{otherwise} \end{cases} .$$

For $\sigma_3 = \infty$ and $\sigma_2 = 0.8$ it reduces to the one given in [52]. Several other choices were also compared numerically in [2], but as far as we can see some of them guarantee (3.16) only under the stronger assumption that f is strongly convex.

Finally we remark that in [35] for a large class of convex quadratic splines a regularized Newton method called QSPLINE was shown to find a solution of (1.1) after finitely many iterations.

3.2. Conjugate Gradient methods. Nonlinear CG methods differ by the choice of the parameter β_k. Several well-known examples are listed in Table 3.1. For quadratic objective functions all those choices coincide, but for general nonlinear functions the corresponding CG methods behave quite differently.

Most results about CG methods in the literature are concerned with proving *global convergence* for general nonlinear functions in the sense that $\liminf_{k\to\infty} g_k = 0$, see e.g. [16, 24] and the many references therein. Here we consider restricted strongly convex functions under Assumption 3.1 and descent methods with $\langle d_k\,, g_k\rangle < 0$ and decreasing function values f_k. Together with Lemma 2.2 this immediately implies the following.

Corollary 3.10. *For any globally convergent CG method we have*

$$\lim_{k\to\infty} f_k = f_{min} \quad , \quad \lim_{k\to\infty} g_k = 0 \quad , \quad \lim_{k\to\infty} \operatorname{dist}(x_k, X_f) = 0\,.$$

Results about convergence rates are scarce and mostly concentrate on quadratic objective functions, or n-step quadratic or n-step superlinear convergence, see eg. [13, 14, 28, 43, 57]. To obtain linear convergence rates we consider CG methods with restarts after every r-th iteration, i.e. the search directions are chosen as

$$d_k := \begin{cases} -g_k & , k = 0 \text{ or } k \text{ is a multiple of } r \\ -g_k + \beta_k \cdot d_{k-1} & , \text{otherwise} \end{cases} . \tag{3.17}$$

We do not necessarily assume $r \geq n$. The following Lemma shows that restarts ensure that the norm of the search direction is bounded by the norm of the gradient.

Table 3.1
Several choices for the conjugate gradient parameter β_k (with $y_{k-1} := g_k - g_{k-1}$).

β_k^{HS}	$= \dfrac{\langle g_k \,, y_{k-1}\rangle}{\langle d_{k-1} \,, y_{k-1}\rangle}$	(Hestenes and Stiefel [25])
β_k^{FR}	$= \dfrac{\lVert g_k\rVert_2^2}{\lVert g_{k-1}\rVert_2^2} \ \geq 0$	(Fletcher and Reeves [21])
β_k^{PRP}	$= \dfrac{\langle g_k \,, y_{k-1}\rangle}{\lVert g_{k-1}\rVert_2^2}$	(Polak, Ribière and Polyak [49, 50])
β_k^{CD}	$= \dfrac{\lVert g_k\rVert_2^2}{-\langle d_{k-1} \,, g_{k-1}\rangle} \ \geq 0$	(“Conjugate Descent” [20])
β_k^{LS}	$= \dfrac{\langle g_k \,, y_{k-1}\rangle}{-\langle d_{k-1} \,, g_{k-1}\rangle}$	(Liu and Storey [39])
β_k^{DY}	$= \dfrac{\lVert g_k\rVert_2^2}{\langle d_{k-1} \,, y_{k-1}\rangle} \ \geq 0$	(Dai and Yuan [17])
β_k^{PRP+}	$= \max\{\beta_k^{PRP}, 0\} \ \geq 0$	(Powell [51])
β_k^{TS}	$= \max\{0, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\} \ \geq 0$	(Touati-Ahmed and Storey [59])
β_k^{GN}	$= \max\{-\beta_k^{FR}, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}$	(Gilbert and Nocedal [22])
β_k^{DYCD}	$= \min\{\beta_k^{DY}, \beta_k^{CD}\} \ \geq 0$	(Dai [15])
$\beta_k(\lambda_k, \mu_k)$	$= \frac{\lambda_k \cdot \lVert g_k\rVert_2^2 + (1-\lambda_k)\cdot\langle g_k \,, y_{k-1}\rangle}{\mu_k \cdot \lVert g_{k-1}\rVert_2^2 + (1-\mu_k)\cdot\langle d_{k-1} \,, y_{k-1}\rangle}$	$, \lambda_k, \mu_k \in [0,1]$ (Nazareth [44])
β_k^{N}	$= \beta_k^{HS} - \lambda \cdot \dfrac{\lVert y_{k-1}\rVert_2^2 \cdot \langle g_k \,, d_{k-1}\rangle}{(\langle d_{k-1} \,, y_{k-1}\rangle)^2}$	$, \lambda > 1/4$ (Hager and Zhang [24])

Lemma 3.11. *Consider a CG method with restarts* (3.17)*. If there is a constant $c > 0$ such that the parameter β_k fulfills*

$$|\beta_k| \leq c \cdot \frac{\|g_k\|_2}{\|g_{k-1}\|_2}\,, \tag{3.18}$$

then there exists some constant $\eta > 0$ such that for all $k \geq 0$ we have

$$\|d_k\|_2 \leq \eta \cdot \|g_k\|_2\ . \tag{3.19}$$

Especially this holds for all β_k with

$$|\beta_k| \leq c \cdot \max\{\beta_k^{FR}, |\beta_k^{PRP}|\}\,, \tag{3.20}$$

and hence for $\beta_k^{PRP+}, \beta_k^{TS}, \beta_k^{GN}$ and $\beta_k(\lambda_k, \mu_k)$ with $\lambda_k \in [0,1]$ and $0 < \mu \leq \mu_k \leq 1$.

Proof. By (3.17) and (3.18) we get

$$\|d_k\|_2 \leq \|g_k\|_2 + |\beta_k| \cdot \|d_{k-1}\|_2 \leq \|g_k\|_2 \cdot \left(1 + c \cdot \frac{\|d_{k-1}\|_2}{\|g_{k-1}\|_2}\right)\,,$$

which yields

$$\frac{\|d_k\|_2}{\|g_k\|_2} \leq 1 + c \cdot \frac{\|d_{k-1}\|_2}{\|g_{k-1}\|_2}\,.$$

Since $d_{l\cdot r} = -g_{l\cdot r}$ for all $l \geq 0$ we infer that for all $0 \leq i \leq r-1$ we have

$$\frac{\|d_{l\cdot r+i}\|_2}{\|g_{l\cdot r+i}\|_2} \leq \sum_{j=0}^{i} c^j \leq \sum_{j=0}^{r-1} c^j =: \eta\,,$$

from which (3.19) follows. To prove that the assertion holds for all β_k with (3.20), it suffices to show that β_k^{FR} and β_k^{PRP} fulfill (3.18). Since the function values f_k are assumed to be decreasing, we can estimate by Lemma 2.2

$$\|g_k\|_2 \leq L\cdot\sqrt{\frac{2}{\nu}}\cdot\sqrt{f_k - f_{min}} \leq L\cdot\sqrt{\frac{2}{\nu}}\cdot\sqrt{f_{k-1} - f_{min}} \leq \frac{\sqrt{2}\cdot L}{\nu}\cdot\|g_{k-1}\|_2\,,$$

i.e. $\|g_k\|_2 \leq c\cdot\|g_{k-1}\|_2$ with $c := \frac{\sqrt{2}\cdot L}{\nu}$. Hence we have $\beta_k^{FR} \leq c\cdot\frac{\|g_k\|_2}{\|g_{k-1}\|_2}$ and

$$|\beta_k^{PRP}| = \frac{|\langle g_k\,, g_k - g_{k-1}\rangle|}{\|g_{k-1}\|_2^2} \leq \frac{\|g_k\|_2}{\|g_{k-1}\|_2}\cdot\frac{\|g_k\|_2 + \|g_{k-1}\|_2}{\|g_{k-1}\|_2} \leq (1+c)\cdot\frac{\|g_k\|_2}{\|g_{k-1}\|_2}\,.$$

Finally by (2.3) we get $\langle d_{k-1}\,, y_{k-1}\rangle = \frac{1}{t_{k-1}}\cdot\langle x_k - x_{k-1}\,, g_k - g_{k-1}\rangle \geq 0$, which implies that for all $\lambda_k \in [0,1]$ and $0 < \mu \leq \mu_k \leq 1$ we have $|\beta_k(\lambda_k,\mu_k)| \leq \frac{1}{\mu}\cdot(\beta_k^{FR} + |\beta_k^{PRP}|)$. □

Not all line searches guarantee descent directions for all choices of β_k. Notable exceptions are β_k^{DYCD} [15] and β_k^{N} [24]. At first we prove linear convergence with exact line searches. Note that in this case we have $\langle d_k\,, g_k\rangle = -\|g_k\|_2^2$, i.e. d_k is always a descent direction. Furthermore several of the parameters β_k in Table 3.1 coincide, i.e. $\beta_k^{FR} = \beta_k^{CD} = \beta_k^{DY}$ and $\beta_k^{PRP} = \beta_k^{HS} = \beta_k^{LS} = \beta_k^{N}$ and $\beta_k(\lambda_k,\mu_k) = \lambda_k\cdot\beta_k^{FR} + (1-\lambda_k)\cdot\beta_k^{PRP}$ for all $\lambda_k,\mu_k \in [0,1]$.

Theorem 3.12. *If a CG method with restarts* (3.17) *and exact line search (LS4a) fulfills* (3.18) *then it is linearly convergent in the sense of* (3.8) *and* (3.9). *In particular this holds for all parameters β_k in Table 3.1.*

Proof. An exact line search (LS4a) implies (3.3) with $c_1 = \frac{1}{2L}$. Together with (2.6) and (3.19) we get

$$f_{k+1} - f_{min} \leq f_k - f_{min} - \frac{1}{2L}\cdot\frac{\|g_k\|_2^2}{\|d_k\|_2^2}\cdot\|g_k\|_2^2 \leq \left(1 - \frac{\nu}{2L\cdot\eta}\right)\cdot(f_k - f_{min})\,.$$

□

Now we turn to the CG methods of Dai and Yuan, and Hager and Zhang, for which a Wolfe line search suffices to ensure that d_k is a descent direction.

Theorem 3.13. *The CG methods with β_k^{DY}, β_k^{DYCD} or β_k^{N} with restarts* (3.17) *and a Wolfe line search (LS2) produce descent directions d_k and are linearly convergent in the sense of* (3.8) *and* (3.10), *i.e. the iterates x_k converge to some minimizer $\hat{x} \in X_f$.*

Proof. In the proofs of Theorem 4.1 in [16], formula (4.12), and Theorem 3.4 in [15], formula (3.17), it was inductively proven that for β_k^{DY} and β_k^{DYCD}, with initial search direction $d_0 = -g_0$ and without restarts, d_k is a descent direction and we have

$$\frac{\|d_k\|_2^2}{\langle g_k\,, d_k\rangle^2} \le \frac{\|d_{k-1}\|_2^2}{\langle g_{k-1}\,, d_{k-1}\rangle^2} + \frac{1}{\|g_k\|_2^2}\,, \tag{3.21}$$

from which follows that $\frac{\|d_k\|_2^2}{\langle g_k\,,d_k\rangle^2} \le \sum_{j=0}^{k} \frac{1}{\|g_j\|_2^2}$. With restarts we can use the same method of proof to conclude that d_k is a descent direction and that for all $l \ge 0$ and $0 \le i \le r-1$ we have

$$\frac{\|d_{l\cdot r+i}\|_2^2}{\langle g_{l\cdot r+i}\,, d_{l\cdot r+i}\rangle^2} \le \sum_{j=0}^{i} \frac{1}{\|g_{l\cdot r+j}\|_2^2}\,.$$

Since the function values f_k are then decreasing, we can estimate as in the proof of Lemma 3.11 to get $\|g_{l\cdot r+i}\|_2 \le c\cdot\|g_{l\cdot r+j}\|_2$ for all $0 \le j \le i$. Hence we have

$$\frac{\|d_{l\cdot r+i}\|_2^2}{\langle g_{l\cdot r+i}\,, d_{l\cdot r+i}\rangle^2} \le \frac{c\cdot(i+1)}{\|g_{l\cdot r+i}\|_2^2} \le \frac{c\cdot r}{\|g_{l\cdot r+i}\|_2^2}\,, \tag{3.22}$$

which implies $-\frac{\langle g_k\,,d_k\rangle^2}{\|d_k\|_2^2} \le -\frac{\|g_k\|_2^2}{c\cdot r}$. Linear convergence of the function values then follows by (3.3) and (2.6). Finally by (3.22) and (3.7) we get

$$\|x_{k+1}-x_k\|_2 = t_k\cdot\|d_k\|_2 \le \frac{\|g_k\|_2\cdot\|d_k\|_2}{-\alpha\cdot\nu\cdot\langle g_k\,, d_k\rangle}\cdot\|g_k\|_2 \le \frac{\sqrt{c\cdot r}}{\alpha\cdot\nu}\cdot\|g_k\|_2\,,$$

from which (3.10) follows similarly as in the proof of Theorem 3.4.

The search directions with β_k^N fulfill the *sufficient descent condition* [24]

$$\langle g_k\,, d_k\rangle \le -\big(1-\tfrac{1}{4\lambda}\big)\cdot\|g_k\|_2^2\,. \tag{3.23}$$

For notational simplicity in the following we use a generic constant $c > 0$. By (3.23), the second Wolfe condition (3.2) and Lemma 3.11 we can estimate the first summand in β_k^N by

$$|\beta_k^{HS}| \le c\cdot|\beta_k^{PRP}| \le c\cdot\frac{\|g_k\|_2}{\|g_{k-1}\|_2}\,,$$

and, since $\|y_{k-1}\|_2 \le \|g_k\|_2 + \|g_{k-1}\|_2 \le c\cdot\|g_{k-1}\|_2$, we can estimate the second summand by

$$\frac{\|y_{k-1}\|_2^2\cdot|\,\langle g_k\,, d_{k-1}\rangle\,|}{(\langle d_{k-1}\,, y_{k-1}\rangle)^2} \le c\cdot\frac{\|g_{k-1}\|_2^2\cdot\|g_k\|_2\cdot\|d_{k-1}\|_2}{\|g_{k-1}\|_2^4} = c\cdot\frac{\|g_k\|_2\cdot\|d_{k-1}\|_2}{\|g_{k-1}\|_2^2}\,.$$

Hence we have $|\beta_k^N| \le c\cdot\frac{\|g_k\|_2}{\|d_{k-1}\|_2}\cdot\left(\frac{\|d_{k-1}\|_2}{\|g_{k-1}\|_2} + \frac{\|d_{k-1}\|_2^2}{\|g_{k-1}\|_2^2}\right)$ and arrive at the recursion

$$\frac{\|d_k\|_2}{\|g_k\|_2} \le 1 + c\cdot\frac{\|d_{k-1}\|_2}{\|g_{k-1}\|_2} + c\cdot\frac{\|d_{k-1}\|_2^2}{\|g_{k-1}\|_2^2}\,,$$

which for restarts yields $\|d_k\|_2 \le c\cdot\|g_k\|_2$. Together with (3.6), (3.7) and (3.23) we conclude that t_k remains bounded and away from zero, i.e. $0 < t_{\min} \le t_k \le t_{\max}$.

Hence (3.23) and the first Wolfe condition (3.1) imply $f_{k+1} \leq f_k - c \cdot \|g_k\|_2^2$. Furthermore we get $\|x_{k+1} - x_k\|_2 \leq c \cdot \|g_k\|_2$, from which the linear convergence results follow. Note that β_k^N then also remains bounded. □

For very large scale problems it may be advantageous to dispense with a line search procedure and instead rely on the explicit step length (LS1), see also [12].

THEOREM 3.14. *A CG method with restarts* (3.17) *and* $t_k := -\frac{c \cdot \langle g_k\,, d_k\rangle}{L \cdot \|d_k\|_2^2}$ *produces descent directions* d_k *and is linearly convergent in the sense of* (3.8) *and* (3.10)*, i.e. the iterates* x_k *converge to some minimizer* $\hat{x} \in X_f$*, in any of the following cases*

(a) $0 < c \leq 1$ *and* $\beta_k \geq 0$ *satisfies* (3.18)*,*

(b) $0 < c < 2$ *and* $\beta_k = \beta_k^{CD}$ *or* $\beta_k = \beta_k^{DYCD}$*.*

Proof. By the Lipschitz-continuity of ∇f we have

$$\|y_{k-1}\|_2 = \|g_k - g_{k-1}\|_2 \leq L \cdot t_{k-1} \cdot \|d_{k-1}\|_2 = -c \cdot \frac{\langle g_{k-1}\,, d_{k-1}\rangle}{\|d_{k-1}\|_2}\,.$$

Hence for all $\beta_k \geq 0$ we get

$$\begin{aligned}\langle g_k\,, d_k\rangle &= -\|g_k\|_2^2 + \beta_k \cdot \big(\langle g_{k-1}\,, d_{k-1}\rangle + \langle y_{k-1}\,, d_{k-1}\rangle\big)\\ &\leq -\|g_k\|_2^2 + \beta_k \cdot (1-c) \cdot \langle g_{k-1}\,, d_{k-1}\rangle\ .\end{aligned}$$

From this we inductively infer that $\langle g_k\,, d_k\rangle \leq -\|g_k\|_2^2 < 0$ for all $0 < c \leq 1$. For β_k^{CD} we (inductively) get

$$\langle g_k\,, d_k\rangle = -\left(2 + \frac{\langle y_{k-1}\,, d_{k-1}\rangle}{\langle g_{k-1}\,, d_{k-1}\rangle}\right) \cdot \|g_k\|_2^2 \leq -(2-c) \cdot \|g_k\|_2^2 < 0\,,$$

and hence also $\beta_k^{CD} \leq \frac{1}{2-c} \cdot \frac{\|g_k\|_2}{\|g_{k-1}\|_2}$, i.e. (3.18) holds. For β_k^{DYCD} inequality (3.21) was shown to hold independent of the line search in [15]. Hence in all cases it follows together with (3.19) that we have $-\frac{\langle g_k\,, d_k\rangle^2}{\|d_k\|_2^2} \leq -c_1 \cdot \|g_k\|_2^2$ and $\|x_{k+1} - x_k\|_2 \leq c_2 \cdot \|g_k\|_2$ for some constants $c_1, c_2 > 0$. The assertion now follows from (3.3) and (2.6) similarly as in the proof of Theorem 3.4. □

The established estimates for the constant q in the convergence rates of the CG methods are rather pessimistic, and indeed much worse than those for ordinary gradient descent. But this holds also for nonlinear strongly convex functions. For instance for CG_DESCENT with β_k^N we have $\|d_k\|_2 \leq c \cdot \|g_k\|_2$ with

$$c = 1 + \frac{L}{\mu} + \lambda \cdot \frac{L^2}{\mu^2}\,,$$

where μ is the modulus of strong convexity, cf. proof of Theorem 2.2 in [23] (this also holds without restarts). Since c determines the lower estimate (3.6) for t_k, a Wolfe line search (3.1) then leads to the estimate

$$q \leq 1 - \frac{\alpha \cdot (1-\beta)}{c^2} \cdot \left(1 - \tfrac{1}{4\lambda}\right)^2 \cdot \frac{\nu}{L}\,,$$

which is worse than the corresponding one for ordinary gradient descent, since c may be considerably larger than 1. This is in sharp contrast to the good numerical performance observed in applications. CG_DESCENT often even outperforms L-BFGS, see [23]. Nevertheless in the initial iterations CG methods can sometimes be observed to be actually worse than ordinary gradient descent, which conforms with the theoretical analysis. But for a sufficiently accurate line search and twice continuously differentiable strongly convex functions n-step superlinear or even n-step quadratic convergence can be established, which explains the superior numerical performance near the solution, see eg. [13, 28, 43].

Based on the idea of proof in [43] we extend this result to restricted strongly convex functions for CG_DESCENT of Hager and Zhang. In the following we again use a generic constant $c > 0$. For an *increasingly accurate line search* we replace the second Wolfe condition (3.2) by the *strong Wolfe condition*

$$|\left\langle g_{k+1}\,, d_k \right\rangle| \leq -\min\{c, \epsilon_k\} \cdot \left\langle g_k\,, d_k \right\rangle\ . \tag{3.24}$$

with a null sequence $\epsilon_k > 0$. The assertions and estimates of Theorem 3.13 then still hold true, i.e. the iterates x_k converge linearly to some minimizer $\hat{x} \in X_f$ with $\|x_k - \hat{x}\|_2 = \mathcal{O}(q^{\frac{k}{2}})$ for some $0 < q < 1$, and we get $|\left\langle g_{k+1}\,, d_k \right\rangle| \leq c \cdot \epsilon_k \cdot \|g_k\|_2^2$. Furthermore we infer from the sufficient descent condition (3.23) that $\|g_k\|_2 \leq c{\cdot}\|d_k\|_2$. We will make repeated use of the following facts established so far

$$\begin{gathered}\|g_k\|_2 = \mathcal{O}(\|d_k\|_2) \quad , \quad \|d_k\|_2 = \mathcal{O}(\|g_k\|_2) \quad , \quad \|g_{k+1}\|_2 = \mathcal{O}(\|g_k\|_2) \\ \left\langle g_{k+1}\,, d_k \right\rangle = \mathcal{O}(\epsilon_k \cdot \|g_k\|_2^2) \quad , \quad \|g_k\|_2 = \mathcal{O}(q^{\frac{k}{2}}) \quad , \quad \|x_k - \hat{x}\|_2 = \mathcal{O}(q^{\frac{k}{2}}) \\ 0 < t_{min} \leq t_k \leq t_{max} \quad , \quad |\beta_k^N| \leq c\,. \end{gathered} \tag{3.25}$$

We assume that f is twice continuously differentiable and denote by $H(x)$ its Hessian at x. We set $\hat{H} := H(\hat{x})$ and define

$$E_k := \int_0^1 \left(H(x_k + \tau \cdot t_k \cdot d_k) - \hat{H}\right) d\tau\,. \tag{3.26}$$

Note that $\lim_{k\to\infty} \|E_k\|_2 = 0$. If $H(x)$ is even Lipschitz-continuous then we have $\|E_k\|_2 = \mathcal{O}(\|x_k - \hat{x}\|_2)$.

At first we analyse the asymptotic behaviour of the step length t_k, the CG parameter β_k^N and the search directions d_k with respect to the Hessian $\hat{H}$ at the minimizer.

Lemma 3.15. *If f is twice continuously differentiable then for CG_DESCENT with increasingly accurate line search there is $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$ we have*

(a) $\frac{\left\langle d_k\,, \hat{H} d_k \right\rangle}{\|g_k\|_2^2} \geq c$ *and* $\frac{\left\langle d_{k+1}\,, \hat{H} d_k \right\rangle}{\|g_k\|_2^2} = \mathcal{O}(\epsilon_k + \|E_k\|_2)$,

(b) $t_k = \frac{-\left\langle g_k\,, d_k \right\rangle}{\left\langle d_k\,, \hat{H} d_k \right\rangle} + \mathcal{O}(\epsilon_k + \|E_k\|_2)$,

(c) $\beta_k^N = \beta_k^{HS} + \mathcal{O}(\epsilon_{k-1})$.

Proof. By the sufficient descent condition (3.23) and the strong Wolfe condition (3.24) we get

$$c \cdot \|g_k\|_2^2 \leq \left\langle g_{k+1} - g_k\,, d_k \right\rangle = t_k \cdot \left\langle d_k\,, \hat{H} d_k \right\rangle + t_k \cdot \left\langle d_k\,, E_k d_k \right\rangle,$$

from which the first part of (a) and (b) follow together with the facts stated in (3.25). By the same arguments and $\|y_{k-1}\|_2 = \|g_k - g_{k-1}\|_2 = \mathcal{O}(\|g_{k-1}\|_2)$ we can estimate the second summand in β_k^N to obtain (c). The second part of (a) then follows from

$$\begin{aligned} t_k \cdot \left\langle d_{k+1}\,, \hat{H} d_k \right\rangle &= \langle d_{k+1}\,, g_{k+1} - g_k \rangle - t_k \cdot \langle d_{k+1}\,, E_k d_k \rangle \\ &= - \langle g_{k+1}\,, g_{k+1} - g_k \rangle + \beta_{k+1}^N \cdot \langle d_k\,, g_{k+1} - g_k \rangle + \mathcal{O}(\|g_k\|_2^2 \cdot \|E_k\|_2) \\ &= - \langle g_{k+1}\,, y_k \rangle + \beta_{k+1}^{HS} \cdot \langle d_k\,, y_k \rangle + \mathcal{O}\big(\|g_k\|_2^2 \cdot (\epsilon_k + \|E_k\|_2)\big) \\ &= \mathcal{O}\big(\|g_k\|_2^2 \cdot (\epsilon_k + \|E_k\|_2)\big)\,, \end{aligned}$$

where the last equality is due to the definition of β_{k+1}^{HS}. □

Now we can proof r-step superlinear convergence.

Theorem 3.16. *Let f be twice continuously differentiable. Consider CG_DESCENT with increasingly accurate line search (3.24) and restarts with $r \geq \hat{r}$, where $\hat{r}$ is the rank of the Hessian $\hat{H}$ at the minimizer $\hat{x} = \lim_{k\to\infty} x_k$. Then it holds*

(a) We have r-step superlinear convergence in the following sense

$$\lim_{m\to\infty} \frac{\|g_{m\cdot r+\hat{r}}\|_2}{\|g_{m\cdot r}\|_2} = 0 \quad \textit{and} \quad \lim_{m\to\infty} \frac{\operatorname{dist}(x_{m\cdot r+\hat{r}}, X_f)}{\operatorname{dist}(x_{m\cdot r}, X_f)} = 0\,.$$

(b) If $H(x)$ is Lipschitz-continuous and $\epsilon_k = \mathcal{O}(\|g_k\|_2)$ then we have

$$\frac{\|g_{m\cdot r+\hat{r}}\|_2}{\|g_{m\cdot r}\|_2} = \mathcal{O}(q^{\frac{m\cdot r}{2}}) \quad \textit{and} \quad \frac{\operatorname{dist}(x_{m\cdot r+\hat{r}}, X_f)}{\operatorname{dist}(x_{m\cdot r}, X_f)} = \mathcal{O}(q^{\frac{m\cdot r}{2}})\,.$$

(c) If there is a unique minimizer $X_f = \{\hat{x}\}$ then under the assumptions in (b) we have r-step quadratic convergence in the following sense

$$\frac{\|g_{m\cdot r+\hat{r}}\|_2}{\|g_{m\cdot r}\|_2} = \mathcal{O}(\|g_{m\cdot r}\|_2^2) \quad \textit{and} \quad \frac{\|x_{m\cdot r+\hat{r}} - \hat{x}\|_2}{\|x_{m\cdot r} - \hat{x}\|_2} = \mathcal{O}(\|x_{m\cdot r} - \hat{x}\|_2^2)\,.$$

Proof. For simplicity we write $k := m \cdot r$ for all $m \in \mathbb{N}$. For all $0 \leq j \leq r$ we have

$$M_{k,j} := \max_{0\leq i\leq j} \{\|E_{k+i}\|_2\,, \epsilon_{k+i}\} \to 0 \quad , \quad k \to \infty\,.$$

At first we inductively show that for all $0 \leq i \leq j < r$ we have

$$\begin{aligned} \langle g_{k+j+1}\,, d_{k+i} \rangle &= \mathcal{O}(\|g_{k+i}\|_2 \cdot \|g_k\|_2 \cdot M_{k,j})\,, \\ \langle g_{k+j+1}\,, g_{k+i} \rangle &= \mathcal{O}(\|g_{k+i}\|_2 \cdot \|g_k\|_2 \cdot M_{k,j})\,, \\ \left\langle d_{k+j+1}\,, \hat{H} d_{k+i} \right\rangle &= \mathcal{O}(\|g_{k+i}\|_2 \cdot \|g_k\|_2 \cdot M_{k,j})\,. \end{aligned} \tag{3.27}$$

For $j = 0$ this follows from the strong Wolfe condition (3.24) together with $d_k = -g_k$ for $k = m \cdot r$, and Lemma 3.15 (a). Now assume that the assertion holds for some $j < r - 1$ and all $0 \leq i \leq j$. Then we have for $i \leq j$

$$\begin{aligned} \langle g_{k+j+2}\,, d_{k+i} \rangle &= \langle g_{k+j+1}\,, d_{k+i} \rangle + \langle g_{k+j+2} - g_{k+j+1}\,, d_{k+i} \rangle \\ &= \langle g_{k+j+1}\,, d_{k+i} \rangle + t_{k+j+1} \cdot \left\langle (\hat{H} + E_{k+j+1}) d_{k+j+1}\,, d_{k+i} \right\rangle \\ &= 2 \cdot \mathcal{O}(\|g_{k+i}\|_2 \cdot \|g_k\|_2 \cdot M_{k,j}) + \mathcal{O}(\|g_{k+i}\|_2 \cdot \|g_k\|_2 \cdot \|E_{k+j+1}\|_2) \\ &= \mathcal{O}(\|g_{k+i}\|_2 \cdot \|g_k\|_2 \cdot M_{k,j+1}) \end{aligned}$$

And for $i = j+1$ this follows again from the strong Wolfe condition (3.24). Then we also get for $0 < i \leq j+1$

$$\begin{aligned}\langle g_{k+j+2}\,, g_{k+i}\rangle &= -\langle g_{k+j+2}\,, d_{k+i}\rangle + \beta^N_{k+i}\cdot\langle g_{k+j+2}\,, d_{k+i-1}\rangle\\ &= \mathcal{O}(\|g_{k+i}\|_2\cdot\|g_k\|_2\cdot M_{k,j+1}) + \beta^N_{k+i}\cdot\mathcal{O}(\|g_{k+i-1}\|_2\cdot\|g_k\|_2\cdot M_{k,j+1})\\ &= \mathcal{O}(\|g_{k+i}\|_2\cdot\|g_k\|_2\cdot M_{k,j+1})\,,\end{aligned}$$

where the last estimate follows from $\left\|\beta^N_{k+i}\cdot d_{k+i-1}\right\|_2 = \|d_{k+i}+g_{k+i}\|_2 = \mathcal{O}(\|g_{k+i}\|_2)$ and $\|g_{k+i-1}\|_2 = \mathcal{O}(\|d_{k+i-1}\|_2)$. For $i=0$ this holds since $g_k = -d_k$. Finally for $i \leq j$ we get

$$\begin{aligned}\left\langle d_{k+j+2}\,, \hat{H}d_{k+i}\right\rangle &= -\left\langle g_{k+j+2}\,, \hat{H}d_{k+i}\right\rangle + \beta^N_{k+j+2}\cdot\left\langle d_{k+j+1}\,, \hat{H}d_{k+i}\right\rangle\\ &= -\left\langle g_{k+j+2}\,, \hat{H}d_{k+i}\right\rangle + \mathcal{O}(\|g_{k+i}\|_2\cdot\|g_k\|_2\cdot M_{k,j})\,,\end{aligned}$$

where we have

$$\begin{aligned}t_{k+i}\left\langle g_{k+j+2}\,, \hat{H}d_{k+i}\right\rangle &= \langle g_{k+j+2}\,, g_{k+i+1}-g_{k+i}\rangle - t_{k+i}\cdot\langle g_{k+j+2}\,, E_{k+i}d_{k+i}\rangle\\ &= \mathcal{O}(\|g_{k+i}\|_2\cdot\|g_k\|_2\cdot M_{k,j+1})\,,\end{aligned}$$

which yields $\left\langle d_{k+j+2}\,, \hat{H}d_{k+i}\right\rangle = \mathcal{O}(\|g_{k+i}\|_2\cdot\|g_k\|_2\cdot M_{k,j+1})$ for $i \leq j$. For $i = j+1$ this follows from Lemma 3.15 (a). Hence (3.27) is proven.

Now we define $u_{k+i} := \frac{\hat{H}^{1/2}d_{k+i}}{\left\|\hat{H}^{1/2}d_{k+i}\right\|_2} \in \mathcal{R}(\hat{H}^{1/2}) = \mathcal{R}(\hat{H})$ for $0 \leq i \leq r$, where $\hat{H}^{1/2}$ is the unique symmetric positive semidefinite square root of $\hat{H}$. Note that by Lemma 3.15 (a) we have $\left\|\hat{H}^{1/2}d_{k+i}\right\|_2 \geq c\cdot\|g_{k+i}\|_2$. Then $\|u_{k+i}\|_2 = 1$ and from (3.27) we infer that for all $0 \leq i \leq j < r$ we have

$$\langle u_{k+j+1}\,, u_{k+i}\rangle = \mathcal{O}\left(\tfrac{\|g_k\|_2}{\|g_{k+j+1}\|_2}\cdot M_{k,j}\right)\,. \tag{3.28}$$

We will prove assertions (a) and (b) of this theorem by contradiction. Assume that (a) is not true. Then there exists a subsequence of $k = m\cdot r$, again denoted by k, such that $\frac{\|g_{k+r}\|_2}{\|g_k\|_2} \geq c$. It follows that $\|g_k\|_2 = \mathcal{O}(\|g_{k+r}\|_2) = \mathcal{O}(\|g_{k+j+1}\|_2)$. Hence by (3.28) we get $\lim_{k\to\infty}\langle u_{k+j+1}\,, u_{k+i}\rangle = 0$, i.e. we may assume without loss of generality that the vectors $u_k,\ldots,u_{k+\hat{r}-1}$ converge to an orthonormal basis $\hat{u}_0,\ldots,\hat{u}_{\hat{r}-1}$ of $\mathcal{R}(\hat{H}^{1/2})$. But then, using (3.28) again for $j = \hat{r}-1$, we arrive at a contradiction, because

$$1 = \sum_{i=0}^{\hat{r}-1}|\langle u_{k+\hat{r}}\,, \hat{u}_i\rangle|^2 = \sum_{i=0}^{\hat{r}-1}|\langle u_{k+\hat{r}}\,, u_{k+i}\rangle + \langle u_{k+\hat{r}}\,, \hat{u}_i - u_{k+i}\rangle|^2 \to 0 \quad , \quad k\to\infty$$

Now we turn to (b). Since $\|x_k - \hat{x}\|_2 = \mathcal{O}(q^{\frac{k}{2}})$ and $\|g_k\|_2 = \mathcal{O}(q^{\frac{k}{2}})$, a Lipschitz-continuous Hessian $H(x)$ and $\epsilon_k = \mathcal{O}(\|g_k\|_2)$ imply $\|E_k\|_2 = \mathcal{O}(q^{\frac{k}{2}})$ and $\epsilon_k = \mathcal{O}(q^{\frac{k}{2}})$. Hence $M_{k,j} = \mathcal{O}(q^{\frac{k}{2}})$. Assume that the assertion in (b) is not true. Then there exists a subsequence of $k = m\cdot r$, again denoted by k, such that $\frac{\|g_{k+r}\|_2}{\|g_k\|_2\cdot q^{\frac{k}{2}}} \to \infty$ for $k\to\infty$. But then we also get $\frac{\|g_{k+j+1}\|_2}{\|g_k\|_2\cdot q^{\frac{k}{2}}} \geq c\cdot\frac{\|g_{k+r}\|_2}{\|g_k\|_2\cdot q^{\frac{k}{2}}} \to \infty$. Hence, again by (3.28), we get $\lim_{k\to\infty}\langle u_{k+j+1}\,, u_{k+i}\rangle = 0$, which leads to the same contradiction as in case (a).

Finally (c) follows similarly, because for a unique minimizer $X_f = \{\hat{x}\}$ we have $\|x_k - \hat{x}\|_2 = \mathrm{dist}(x_k, X_f) = \mathcal{O}(\|g_k\|_2)$, and we can use $\|g_k\|_2$ instead of $q^{\frac{k}{2}}$. □

As demonstrated in Counterexample 3.2 an exact minimizer t_k for the line search need not exist for restricted strongly convex functions. And even if it exists, e.g. in case of a bounded set of minimizers X_f, convergence of the whole sequence x_k to a minimizer $\hat{x} \in X_f$ cannot be guaranteed, which is crucial for the proof above. Hence we actually have to use inexact line searches, both from theoretical as well as computational considerations. Note that, since the rank $\hat{r}$ of the Hessian $\hat{H}$ at $\hat{x}$ may be considerably smaller than n, Theorem 3.16 suggests that in this case faster convergence can be achieved by making restarts with r smaller than n. This may then also be the case for strongly convex functions, if the Hessian $\hat{H}$ has several eigenvalues close to zero. Preliminary numerical experiments seem to support this conjecture, but more experiments are needed for decisive conclusions. To the best of our knowledge this is the first kind of such a convergence result for CG methods applied to functions with a rank deficient Hessian. In this regard Theorem 3.16 gives further strong theoretical support for the superior numerical performance of CG_DESCENT compared to several other CG methods. Indeed a close look at the proof of Theorem 3.16 reveals that we needed the following properties of CG_DESCENT:

- sufficient descent (3.23) with a (strong) Wolfe line search,
- the (asymptotic) conjugacy condition $\langle y_{k-1}, d_k \rangle = 0$ inherited from the CG method with β_k^{HS} of Hestenes and Stiefel, cf. Lemma 3.15 (c).

So far we do not see how to extend this result to other CG methods, or even (L)-BFGS.

Finally we consider the case of strongly convex quadratic splines f. For such functions it was shown in [33] that the iterates of a special CG method reach the unique minimizer $\hat{x}$ of f after finitely many iterations if f is twice continuously differentiable at $\hat{x}$. This method uses exact line searches and, instead of cyclic restarts, automatically restarts if an iterate enters a new polyhedral region, where the representing linear quadratic function changes. Here we can extend this result to the CG methods with any of the parameters β_k in Table 3.1.

THEOREM 3.17. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex quadratic spline with unique minimizer $\hat{x}$. Consider a CG method with an exact line search (LS4a) and any of the parameters β_k in Table 3.1. If we restart only if an iterate enters a new polyhedral region, where the linear quadratic function representing f changes, then the iterates converge linearly to $\hat{x}$ in the sense of* (3.8) *and* (3.10). *If f is twice continuously differentiable at $\hat{x}$ then the iterates reach $\hat{x}$ after finitely many iterations.*

Proof. For strongly convex f and exact line searches all CG methods with any of the parameters β_k in Table 3.1 are known to be globally convergent without restarts. It follows from Corollary 3.10 and Remark 2.9 that after finitely many iterations k_0 all iterates will always be in a solution region of f. If never more than n consecutive iterates stay in the same solution region, then restarts always occur after at most $\max\{n, k_0\}$ iterations, in which case the assertions of Theorem 3.12 remain valid. Otherwise the linear quadratic function representing f remains the same for at least n consecutive iterations, and, since a restart occured at entering the region, the method reduces to the standard linear CG method which terminates after at most n iterations with the exact solution $\hat{x}$. Finally, if f is twice continuously differentiable at $\hat{x}$ then $\hat{x}$ lies in the interior of some solution region and hence the iterates will eventually stay in this same region, cf. proof of Theorem 3 in [33]. □

3.3. Linearly constrained optimization problems. Together with Theorem 2.12, applying the descent methods to the unconstrained dual of (1.6), we immediately get linear convergence of the primal iterates.

COROLLARY 3.18. *Consider the linearly constrained convex optimization problem* (1.6) *with a strongly convex objective function* $g : \mathbb{R}^m \to \mathbb{R}$ *and unique solution* $\hat{y} \in \mathbb{R}^m$*. Apply any of the descent methods discussed in the previous subsections to the unconstrained dual* (1.1) *with objective function* $f(x) = g^*(A^T x) - \langle b\,, x\rangle$*, and define the primal iterates as* $y_k := \nabla g^*(A^T x_k)$*. If* f *satisfies Assumption 3.1, then there exist constants* $q \in (0,1)$ *and* $\gamma > 0$ *such that*

$$\|y_k - \hat{y}\|_2 \leq \gamma \cdot q^{\frac{k}{2}}\,.$$

Especially this holds under the assumptions of Theorem 2.12. Furthermore, if g *is twice continuously differentiable then the CG method with* β_k^N *of Hager and Zhang guarantees* r*-step superlinear convergence in the sense of Theorem 3.16, and also with the corresponding expressions for* $\|y_k - \hat{y}\|_2$ *replacing* $\operatorname{dist}(x_k, X_f)$*.*

Proof. The assertion follows from the estimates

$$\|y_k - \hat{y}\|_2 = \left\|\nabla g^*(A^T x_k) - \nabla g^*\big(A^T P_{X_f}(x_k)\big)\right\|_2 \leq L_{g^*} \cdot \|A\|_2 \cdot \operatorname{dist}(x_k, X_f)$$

and $\operatorname{dist}(x_k, X_f) = \mathcal{O}(\left\|\nabla f(x_k) - \nabla f\big(P_{X_f}(x_k)\big)\right\|_2) = \mathcal{O}(\|y_k - \hat{y}\|_2)$. □

We remark that here the explicit value (LS1) may be replaced by $t_k := -\frac{c\cdot\langle g_k\,, d_k\rangle}{L_{g^*}\cdot\|A^T d_k\|_2^2}$, which only involves L_{g^*} and needs no estimate for $\|A\|_2$. And instead of the iterates x_k it suffices to store the iterates $y_k^* := A^T x_k$, because we have $y_{k+1}^* = y_k^* + t_k \cdot A^T d_k$, $y_k = \nabla g^*(y_k^*)$ and $g_k = \nabla f^*(x_k) = A y_k - b$.

4. Conclusions. We have proven that, for a large class of descent methods for unconstrained minimization including nonlinear CG and BFGS, linear convergence rates can still be guaranteed if one replaces the assumption of strong convexity by the weaker assumption of restricted strong convexity. For CG_DESCENT we could obtain r-step superlinear convergence, even if the Hessian at a minimizer is rank deficient. Somewhat surprisingly there remains a little gap. So far we succeeded for the standard BFGS method und its damped limited memory variant L-D-BFGS, but neither for the damped BFGS [1] nor the undamped L-BFGS method. It would be interesting to know whether this gap can be filled. Furthermore we have shown that convex quadratic splines and objective functions of the unconstrained duals to some linearly constrained optimization problems are restricted strongly convex. Future research should establish restricted strong convexity for more classes of functions.

References.

[1] M. Al-Baali and L. Grandinetti, *On practical modifications of the quasi-Newton BFGS method*, Adv. Model. Optim, 11 (2009), pp. 63–76.

[2] M. Al-Baali, L. Grandinetti, and O. Pisacane, *Damped techniques for the limited memory BFGS method for large-scale optimization*, Journal of Optimization Theory and Applications, 161 (2014), pp. 688–699.

[3] H. H. Bauschke, J. M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, jamesons property (G), and error bounds in convex optimization*, Mathematical Programming, 86 (1999), pp. 135–160.

[4] R .H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with applications to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.

[5] R .H. Byrd, J. Nocedal, and Y.-X. Yuan, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM Journal on Numerical Analysis, 24 (1987), pp. 1171–1190.

[6] J.-F. Cai, E. J. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 20 (2010), pp. 1956–1982.

[7] J.-F. Cai, S. Osher, and Z. Shen, *Convergence of the linearized Bregman iteration for ℓ_1-norm minimization*, Math. Comp., 78 (2009), pp. 2127–2136.

[8] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, Journal of the ACM, 58 (2011), p. 11.

[9] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM Journal on Optimization, 21 (2011), pp. 572–596.

[10] B. T. Chen, K. Madsen, and S. Zhang, *On the characterization of quadratic splines*, Journal of Optimization Theory and Applications, 124 (2005), pp. 93–111.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.

[12] X. Chen and J. Sun, *Global convergence of a two-parameter family of conjugate gradient methods without line search*, Journal of Computational and Applied Mathematics, 146 (2002), pp. 37–45.

[13] A. I. Cohen, *Rate of convergence of several conjugate gradient algorithms*, SIAM Journal on Numerical Analysis, 9 (1972), pp. 248–259.

[14] H. Crowder and P. Wolfe, *Linear convergence of the conjugate gradient method*, IBM Journal of Research and Development, 16 (1972), pp. 431–433.

[15] Y. H. Dai, *A nonmonotone conjugate gradient algorithm for unconstrained optimization*, Journal of Systems Science and Complexity, 15 (2002), pp. 139–145.

[16] ——, *Nonlinear conjugate gradient methods*, Wiley Encyclopedia of Operations Research and Management Science, (2011).

[17] Y. H. Dai and Y. Yuan, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM Journal on Optimization, 10 (1999), pp. 177–182.

[18] J. E. Dennis and J. J. Moré, *Quasi-newton methods, motivation and theory*, SIAM review, 19 (1977), pp. 46–89.

[19] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*, Springer, 2010.

[20] R. Fletcher, *Practical methods of optimization*, John Wiley & Sons, 2013.

[21] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, The computer journal, 7 (1964), pp. 149–154.

[22] J. C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM Journal on optimization, 2 (1992), pp. 21–42.

[23] W. W. Hager and H. Zhang, *A new conjugate gradient method with guaranteed descent and an efficient line search*, SIAM Journal on Optimization, 16 (2005), pp. 170–192.

[24] ——, *A survey of nonlinear conjugate gradient methods*, Pacific journal of Optimization, 2 (2006), pp. 35–58.

[25] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, vol. 49, NBS, 1952.

[26] K. Hou, Z. Zhou, A. M.-C. So, and Z.-Q. Luo, *On the linear convergence of the proximal gradient method for trace norm regularization*, in Advances in Neural Information Processing Systems, 2013, pp. 710–718.

[27] M. J. Lai and W. Yin, *Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm*, SIAM J. Imaging Sci., 6 (2013), pp. 1059–1091.

[28] M. L. Lenard, *Convergence conditions for restarted conjugate gradient methods with inaccurate line searches*, Mathematical Programming, 10 (1976), pp. 32–51.

[29] A. S. Lewis and H. S. Sendov, *Nonsmooth analysis of singular values. part i: Theory*, Set-Valued Analysis, 13 (2005), pp. 213–241.

[30] ———, *Nonsmooth analysis of singular values. part ii: Applications*, Set-Valued Analysis, 13 (2005), pp. 243–264.

[31] W. Li, *Error bounds for piecewise convex quadratic programs and applications*, SIAM Journal on Control and Optimization, 33 (1995), pp. 1510–1529.

[32] ———, *Linearly convergent descent methods for the unconstrained minimization of convex quadratic splines*, Journal of Optimization Theory and Applications, 86 (1995), pp. 145–172.

[33] ———, *A conjugate gradient method for the unconstrained minimization of strictly convex quadratic splines*, Mathematical Programming, 72 (1996), pp. 17–32.

[34] W. Li and J. Swetits, *A newton method for convex regression, data smoothing, and quadratic programming with bounded constraints*, SIAM Journal on Optimization, 3 (1993), pp. 466–488.

[35] ———, *Regularized newton methods for minimization of convex quadratic splines with singular hessians*, in Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods, Springer, 1999, pp. 235–257.

[36] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, *Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix*, Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 61 (2009).

[37] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical Programming, 45 (1989), pp. 503–528.

[38] J. Liu and S. J. Wright, *Asynchronous stochastic coordinate descent: Parallelism and convergence properties*, SIAM Journal on Optimization, 25 (2015), pp. 351–376.

[39] Y. Liu and C. Storey, *Efficient generalized conjugate gradient algorithms, part 1: theory*, Journal of Optimization Theory and Applications, 69 (1991), pp. 129–137.

[40] D. A. Lorenz, F. Schöpfer, and S. Wenger, *The linearized Bregman method via split feasibility problems: Analysis and generalizations*, SIAM J. Imaging Sciences, 7 (2014), pp. 1237–1262.

[41] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization, 30 (1992), pp. 408–425.

[42] ———, *Error bounds and convergence analysis of feasible descent methods: a general approach*, Annals of Operations Research, 46 (1993), pp. 157–178.

[43] G. P. McCormick and K. Ritter, *Alternative proofs of the convergence properties of the conjugate-gradient method*, Journal of Optimization Theory and Applications, 13 (1974), pp. 497–518.

[44] J. L. Nazareth, *Conjugate-gradient methods*, in Encyclopedia of Optimization, Springer, 2001, pp. 319–323.

[45] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 1999.

[46] S. Osher, Y. Mao, B. Dong, and W. Yin, *Fast linearized Bregman iteration for compressive sensing and sparse denoising*, Communications in Mathematical Sciences, 8 (2010), pp. 93–111.

[47] J. S. Pang, *Error bounds in mathematical programming*, Mathematical Programming, 79 (1997), pp. 299–332.

[48] P. M. Pardalos and N. Kovoor, *An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds*, Mathematical Programming, 46 (1990), pp. 321–328.

[49] E. Polak and G. Ribiere, *Note sur la convergence de méthodes de directions conjuguées*, Revue française d'informatique et de recherche opérationnelle, série rouge, 3 (1969), pp. 35–43.

[50] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 94–112.

[51] M. JD. Powell, *Nonconvex minimization calculations and the conjugate gradient method*, Springer, 1984.

[52] M. J. D. Powell, *Algorithms for nonlinear constraints that use Lagrangian functions*, Mathematical programming, 14 (1978), pp. 224–248.

[53] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review, 52 (2010), pp. 471–501.

[54] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Mathematical Programming Study, 14 (1981), pp. 206–214.

[55] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 2009.

[56] F. Schöpfer, *Exact regularization of polyhedral norms*, SIAM J. Optim., 22 (2012), pp. 1206–1223.

[57] Z.-J. Shi and J. Guo, *A new algorithm of nonlinear conjugate gradient method with strong convergence*, Computational & Applied Mathematics, 27 (2008), pp. 93–106.

[58] D. Sun J. Sun, *Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems*, SIAM Journal on Numerical Analysis, 40 (2002), pp. 2352–2367.

[59] D. Touati-Ahmed and C. Storey, *Efficient hybrid conjugate gradient techniques*, Journal of Optimization Theory and Applications, 64 (1990), pp. 379–397.

[60] P. Tseng, *Approximation accuracy, gradient methods, and error bound for structured convex optimization*, Mathematical Programming, 125 (2010), pp. 263–295.

[61] P.-W. Wang and C.-J. Lin, *Iteration complexity of feasible descent methods for convex optimization*, The Journal of Machine Learning Research, 15 (2014), pp. 1523–1548.

[62] G. A. Watson, *Characterization of the subdifferential of some matrix norms*, Linear Algebra and its Applications, 170 (1992), pp. 33–45.

[63] W. Yin, *Analysis and generalizations of the linearized Bregman method*, SIAM J. Imaging Sci., 3 (2010), pp. 856–877.

[64] H. Zhang, J. F. Hui Cai, L. Cheng, and J. Zhu, *Strongly convex programming for exact matrix completion and robust principal component analysis*, Inverse Problems and Imaging, 6 (2012), pp. 357–372.

[65] H. Zhang and L. Cheng, *Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization*, Optimization Letters, (2014), pp. 1–19.

[66] H. Zhang and W. Yin, *Gradient methods for convex minimization: better rates under weaker conditions*, arXiv, (2013).