A. Appendix

A.1. Concentration Inequalities and Sampling Schemes

For the sake of simplicity we shall drop the iteration subscript k in the following results of this section.

A.1.1. Gradient Sampling

First, we extend the Vector Bernstein inequality as it can be found in (Gross, 2011) to the *average* of independent, zero-mean vector-valued random variables.

Lemma 18 (Vector Bernstein Inequality)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent vector-valued random variables with common dimension d and assume that each one is centered, uniformly bounded and also the variance is bounded above:*

$$\mathbb{E}\left[\mathbf{x}_i\right] = 0 \textit{ and } \|\mathbf{x}_i\|_2 \leq \mu \textit{ as well as } \mathbb{E}\left[\,\|\mathbf{x}_i\|^2\right] \leq \sigma^2$$

Let

$$\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i.$$

Then we have for $0 < \epsilon < \sigma^2/\mu$

$$P\left(\,\|\mathbf{z}\| \geq \epsilon\right) \leq \exp\left(-n \cdot \frac{\epsilon^2}{8\sigma^2} + \frac{1}{4}\right) \tag{35}$$

Proof: Theorem 6 in (Gross, 2011) gives the following Vector Bernstein inequality for independent, zero-mean vector-valued random variables

$$P\left(\left\|\sum_{n=1}^{n}\mathbf{x}_i\right\| \geq t + \sqrt{V}\right) \leq \exp\left(-\frac{t^2}{4V}\right), \tag{36}$$

where $V = \sum_{i=1}^{n}\mathbb{E}\left[\,\|\mathbf{x}_i\|^2\right]$ is the sum of the variances of the centered vectors $\mathbf{x}_i$.

First, we shall define $\epsilon = t + \sqrt{V}$, which allows us to rewrite the above equation as

$$P\left(\left\|\sum_{i=1}^{n}\mathbf{x}_i\right\| \geq \epsilon\right) \leq \exp\left(-\frac{1}{4}\left(\frac{\epsilon - \sqrt{V}}{\sqrt{V}}\right)^2\right) = \exp\left(-\frac{1}{4}\left(\frac{\epsilon}{\sqrt{V}} - 1\right)^2\right). \tag{37}$$

Based on the observation that

$$\begin{aligned}
& -\frac{1}{4}\left(\frac{\epsilon}{\sqrt{V}} - 1\right)^2 \leq -\frac{1}{4}\left(\frac{\epsilon^2}{2V}\right) + \frac{1}{4} \\
\Leftrightarrow & -\frac{\epsilon^2}{V} + 2\frac{\epsilon}{\sqrt{V}} - 1 \leq -\frac{\epsilon^2}{2V} + 1 \\
\Leftrightarrow & \; 0 \leq \frac{\epsilon^2}{2V} - 2\frac{\epsilon}{\sqrt{V}} + 2 \\
\Leftrightarrow & \; 0 \leq \left(\frac{\epsilon}{\sqrt{2V}} - \sqrt{2}\right)^2
\end{aligned} \tag{38}$$

always holds, we can formulate a slightly weaker Vector Bernstein version as follows

$$P\left(\left\|\sum_{i=1}^{n}\mathbf{x}_i\right\| \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{8V} + \frac{1}{4}\right). \tag{39}$$

Since the individual variance is assumed to be bounded above, we can write

$$V = \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{x}_i\|^2 \right] \leq n\sigma^2. \tag{40}$$

This term also constitutes an upper bound on the variance of $\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i$, because the $\mathbf{x}_i$ are independent and thus uncorrelated . However, $\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and we must account for the averaging term. Since the $\mathbf{x}_i$ are centered we have $\mathbb{E}[\mathbf{z}] = 0$, and thus

$$\begin{aligned} Var(\mathbf{z}) = \mathbb{E}\left[\|\mathbf{z} - \mathbb{E}[\mathbf{z}]\|^2 \right] = \mathbb{E}\left[\|\mathbf{z}\|^2 \right] = \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \right\|^2 \right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbf{x}_i \right)^{\intercal} \left(\sum_{j=1}^{n} \mathbf{x}_j \right) \right] \\ = \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j} \left(\mathbf{x}_j^{\intercal}\mathbf{x}_i \right) \right] = \frac{1}{n^2}\sum_{i,j}\mathbb{E}\left[\left(\mathbf{x}_j^{\intercal}\mathbf{x}_i \right) \right] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbf{x}_i^{\intercal}\mathbf{x}_i \right) \right] + \sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left[\left(\mathbf{x}_i^{\intercal}\mathbf{x}_j \right) \right] \right) \\ = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{x}_i\|^2 \right] \leq \frac{1}{n}\sigma^2, \end{aligned} \tag{41}$$

where we used the fact that the expectation of a sum equals the sum of the expectations and the cross-terms $\mathbb{E}\left[\mathbf{x}_j^{\intercal}\mathbf{x}_i \right] = 0, j \neq i$ because of the independence assumption. Hence, we can bound the term $V \leq \frac{1}{n}\sigma^2$ for the random vector sum $\mathbf{z}$.

Now, since $n > 1$ and $\epsilon > 0$, as well as $P(\mathbf{z} > a)$ is falling in a and $\exp(-x)$ falling in x, we can use this upper bound on the variance of $\mathbf{z}$ in (39), which gives the desired inequality

$$P\left(\|\mathbf{z}\| \geq \epsilon \right) \leq \exp\left(-n \cdot \frac{\epsilon^2}{8\sigma^2} + \frac{1}{4} \right) \tag{42}$$

□

This result was applied in order to find the probabilistic bound on the deviation of the sub-sampled gradient from the full gradient as stated in Lemma 6, for which we will give the proof next.

Proof of Lemma 6:

To apply vector Bernstein's inequality (35) we need to center the gradients. Thus we define

$$\mathbf{x}_i = g_i(\mathbf{x}) - \nabla f(\mathbf{x}),\ i = 1, \dots, |S_g| \tag{43}$$

and note that from the Lipschitz continuity of f (A3), we have

$$\|\mathbf{x}_i\| = \|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \|\mathbf{g}_i(\mathbf{x})\| + \|\nabla f(\mathbf{x})\| \leq 2\kappa_f \text{ and } \|\mathbf{x}_i\|^2 \leq 4\kappa_f^2,\ i = 1, \dots, |S_g|. \tag{44}$$

With $\sigma^2 := 4\kappa_f^2$ and

$$\mathbf{z} = \frac{1}{|S_g|}\sum_{i\in S_g} \mathbf{x}_i = \frac{1}{|S_g|}\sum_{i\in S_g} \mathbf{g}_i(\mathbf{x}) - \frac{1}{|S_g|}\sum_{i\in S_g} \nabla f(\mathbf{x}) = \mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x}) \tag{45}$$

in equation (35), we can require the probability of a deviation larger or equal to ϵ to be lower than some $\delta \in (0, 1]$

$$\begin{aligned} P\left(\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\| > \epsilon \right) \leq & 2d\exp\left(-|S_g| \cdot \frac{\epsilon^2}{32\kappa_f^2} + \frac{1}{4} \right) \overset{!}{\leq} \delta \\ \Leftrightarrow & |S_g| \cdot \frac{\epsilon^2}{32\kappa_f^2} - \frac{1}{4} \overset{!}{\geq} \log((2d)/\delta) \\ \Leftrightarrow & \epsilon \geq 4\sqrt{2}\kappa_f\sqrt{\frac{\log\left((2d)/\delta \right) + 1/4}{|S_g|}}. \end{aligned} \tag{46}$$

Conversely, the probability of a deviation of

$$\epsilon < 4\sqrt{2}\kappa_f \sqrt{\frac{\log\left((2d)/\delta\right) + 1/4}{|S_g|}} \tag{47}$$

is higher or equal to $1 - \delta$.

□

Of course, any sampling scheme that guarantees the right hand side of (16) to be smaller or equal to M times the squared step size, directly satisfies the sufficient gradient agreement condition (A5). Consequently, plugging the former into the latter and rearranging for the sample size gives Theorem 7 as we shall prove now.

Proof of Theorem 7:

By use of Lemma 6 we can write

$$\begin{aligned} \|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\| &\leq M \|\mathbf{s}\|^2 \\ \Leftrightarrow 4\sqrt{2}\kappa_f \sqrt{\frac{\log(1/\delta + 1/4)}{|S_g|}} &\leq M \|\mathbf{s}\|^2 \\ |S_g| &\geq \frac{32\kappa_f^2 \log\left(1/\delta + 1/4\right)}{M^2 \|\mathbf{s}\|^4} \end{aligned} \tag{48}$$

□

A.1.2. Hessian Sampling

Lemma 19 (Matrix Bernstein Inequality)**.** *Let $\mathbf{A}_1, .., \mathbf{A}_n$ be independent random Hermitian matrices with common dimension $d \times d$ and assume that each one is centered, uniformly bounded and also the variance is bounded above:*

$$\mathbb{E}\left[\mathbf{A}_i\right] = 0 \textit{ and } \left\|\mathbf{A}_i\right\|_2 \leq \mu \textit{ as well as } \left\|\mathbb{E}\left[\mathbf{A}_i^2\right]\right\|_2 \leq \sigma^2$$

Introduce the sum

$$\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i$$

Then we have

$$P(\left\|\mathbf{Z}\right\| \geq \epsilon) \leq 2d \cdot \exp\left(-n \cdot \min\{\frac{\epsilon^2}{4\sigma^2}, \frac{\epsilon}{2\mu}\}\right) \tag{49}$$

Proof: Theorem 12 in (Gross, 2011) gives the following Operator-Bernstein inequality

$$P\left(\left\|\sum_{i=1}^{n}\mathbf{A}_i\right\| \geq \epsilon\right) \leq 2d \cdot \exp\left(\min\{\frac{\epsilon^2}{4V}, \frac{\epsilon}{2\mu}\}\right), \tag{50}$$

where $V = n\sigma^2$. As well shall see, this is an upper bound on the variance of $\mathbf{Y} = \sum_{i=1}^{n}\mathbf{A}_i$ since the $\mathbf{A}_i$ are independent and have an expectation of zero ($\mathbb{E}\left[Y\right] = 0$).

$$\begin{aligned} Var(\mathbf{Y}) &= \left\|\mathbb{E}\left[\mathbf{Y}^2\right] - \mathbb{E}\left[\mathbf{Y}\right]^2\right\| = \left\|\mathbb{E}\left[(\sum_i \mathbf{A}_i)^2\right]\right\| = \left\|\mathbb{E}\left[\sum_{i,j}\mathbf{A}_i\mathbf{A}_j\right]\right\| = \left\|\sum_{i,j}\mathbb{E}\left[\mathbf{A}_i\mathbf{A}_j\right]\right\| \\ &= \left\|\sum_i \mathbb{E}\left[\mathbf{A}_i\mathbf{A}_i\right] + \sum_i\sum_{j\neq i}\mathbb{E}\left[\mathbf{A}_i\mathbf{A}_j\right]\right\| = \left\|\sum_i \mathbb{E}\left[\mathbf{A}_i^2\right]\right\| \leq \sum_i \left\|\mathbb{E}\left[\mathbf{A}_i^2\right]\right\| \leq n\sigma^2, \end{aligned} \tag{51}$$

where we used the fact that the expectation of a sum equals the sum of the expectations and the cross-terms $\mathbb{E}\left[\mathbf{A}_j\mathbf{A}_i\right] = 0, j \neq i$ because of the independence assumption.

However, $\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i$ and thus

$$Var(\mathbf{Z}) = \left\|\mathbb{E}\left[\mathbf{Z}^2\right]\right\| = \left\|\mathbb{E}\left[(\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i)^2\right]\right\| = \frac{1}{n^2}\left\|\mathbb{E}\left[(\sum_{i=1}^{n}\mathbf{A}_i)^2\right]\right\| \leq \frac{1}{n}\sigma^2. \tag{52}$$

Hence, we can bound $V \leq \frac{1}{n}\sigma^2$ for the *average* random matrix sum $\mathbf{Z}$. Furthermore, since $n > 1$ and $\epsilon, \mu > 0$ as well as $\exp(-\alpha)$ decreasing in $\alpha \in \mathbb{R}$ we have that

$$\exp\left(-\frac{\epsilon}{2\mu}\right) \leq \exp\left(-\frac{\epsilon}{n2\mu}\right). \tag{53}$$

Together with the Operator-Bernstein inequality, (52) and (53) give the desired inequality (49).

□

This result exhibits that sums of independent random matrices provide normal concentration near its mean in a range determined by the variance of the sum. We apply it in order to derive the bound on the deviation of the sub-sampled Hessian from the full Hessian as stated in Lemma 8, which we shall prove next.

Proof of Lemma 8: Bernstein's Inequality holds as $f \in C^2$ and thus the Hessian is symmetric by Schwarz's Theorem. Since the expectation of the random matrix needs to be zero, we center the individual Hessians,

$$\mathbf{X}_i = \mathbf{H}_i(\mathbf{x}) - \mathbf{H}(\mathbf{x}), i = 1, ..., |S_H|$$

and note that now from the Lipschitz continuity of $\mathbf{g}$ (A3):

$$\|\mathbf{X}_i\|_2 \leq 2\kappa_g, i = 1...|S_H| \text{ and } \left\|\mathbf{X}_i^2\right\|_2 \leq 4\kappa_g^2, i = 1...|S_H|.$$

Hence, for $\epsilon \leq 4\kappa_g$, we are in the *small deviation* regime of Bernstein's bound with a sub-gaussian tail. Then, we may plug

$$\frac{1}{|S_H|}\sum_{i=1}^{|S_H|} \mathbf{X}_i = \mathbf{B}(\mathbf{x}) - \mathbf{H}(\mathbf{x})$$

into (49), to get

$$P(\|\mathbf{B}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| \geq \epsilon) \leq 2d \cdot \exp\left(-\frac{\epsilon^2|S_H|}{16\kappa_g^2}\right). \tag{54}$$

Finally, we shall require the probability of a deviation of ϵ or higher to be lower than some $\delta \in (0, 1]$

$$\begin{aligned} & 2d \cdot \exp\left(-\frac{\epsilon^2|S_H|}{16\kappa_g^2}\right) \overset{!}{=} \delta \\ \Leftrightarrow & -\frac{\epsilon^2|S_H|}{16\kappa_g^2} = \log(\delta/2d) \\ \Leftrightarrow & \epsilon = 4\kappa_g\sqrt{\frac{\log(2d/\delta)}{|S_H|}}, \end{aligned} \tag{55}$$

which is equivalent to $\|\mathbf{B}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\|$ staying within this particular choice of ϵ with probability $(1-\delta)$, generally perceived as *high probability*.

□

Proof of Theorem 9: Since $\|\mathbf{Av}\| \leq \|\mathbf{A}\|_{op}\|\mathbf{v}\|$ for every $\mathbf{v} \in V$ we have for the choice of the spectral matrix norm and euclidean vector norm that any $\mathbf{B}$ that fulfils $\|(\mathbf{B}(\mathbf{x}) - \mathbf{H}(\mathbf{x}))\| \leq C\|\mathbf{s}\|$ also satisfies condition A4. Furthermore

$$\begin{aligned} & \|(\mathbf{B} - \mathbf{H}(\mathbf{x}))\| \leq C\|\mathbf{s}\| \\ \Leftrightarrow & 4\kappa_g\sqrt{\frac{\log(2d/\delta)}{|S_H|}} \leq C\|\mathbf{s}\| \\ \Leftrightarrow & |S_H| \geq \frac{16\kappa_g^2\log(2d/\delta)}{(C\|\mathbf{s}\|)^2}, \qquad C > 0. \end{aligned} \tag{56}$$

□

Note that there may be a less restrictive sampling conditions that satisfy A4 since condition (56) is based on the worst case bound $\|\mathbf{Av}\| \leq \|\mathbf{A}\|_{op}\|\mathbf{v}\|$ which indeed only holds with equality if $\mathbf{v}$ happens to be (exactly in the direction of) the largest eigenvector of A.

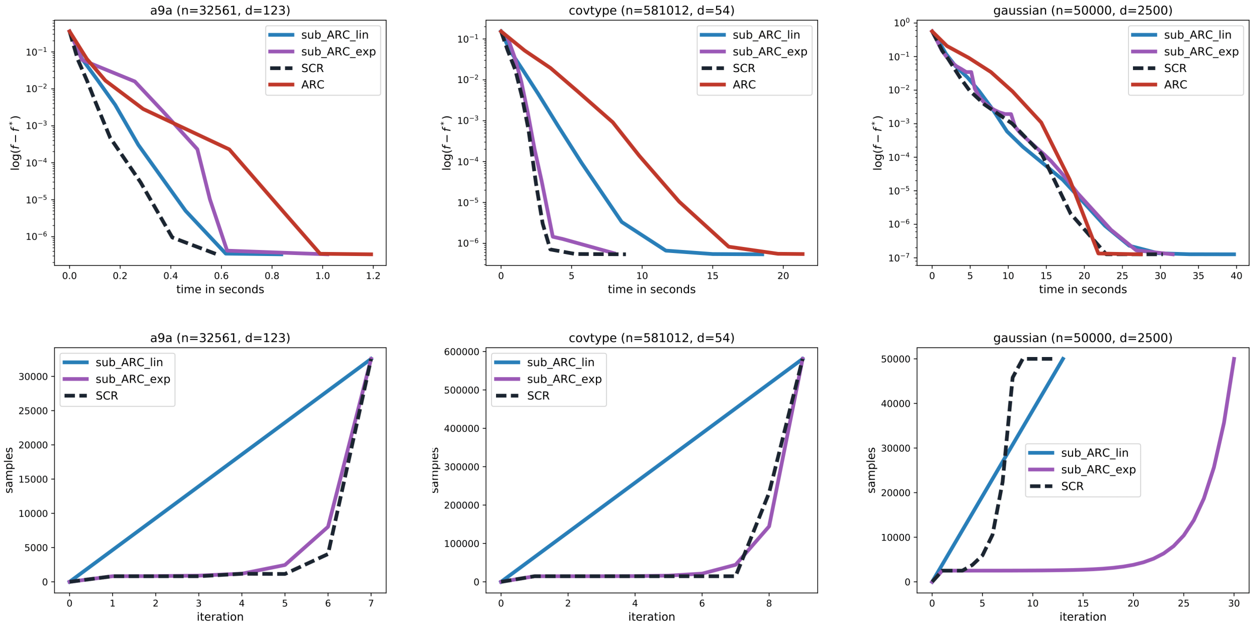
Finally, we shall state a Lemma which illustrates that the stepsize goes to zero as the algorithm converges. The proof can be found in Section 5 of (Cartis et al., 2011a).

Lemma 20. *Let $\{f(\mathbf{x}_k)\}$ be bounded below by some $f_{\inf} > -\infty$. Also, let $\mathbf{s}_k$ satisfy A1 and σ_k be bounded below by some $\sigma_{\inf} > 0$. Then we have for all successful iterations that*

$$\|\mathbf{s}_k\| \to 0, \text{ as } k \to \infty \tag{57}$$

A.1.3. Illustration

In the top row of Figure 2 we illustrate the Hessian sample sizes that result when applying SCR with a practical version of Theorem 9 to the datasets used in our experiments [4]. In the bottom row of Figure 2, we benchmark our algorithm to the deterministic as well as two naive stochastic versions of ARC with *linearly* and *exponentially* increasing sample sizes.



1. a9a

2. covtype

3. gaussian

Figure 2. Suboptimality (top row) and sample sizes (bottom row) for different cubic regularization methods on *a9a*, *covtype* and *gaussian*. Note that the automatic sampling scheme of SCR follows an exponential curve, which means that it can indeed save a lot of computation in the early stage of the optimization process.

Note that both the linear and the exponential sampling schemes do not quite reach the same performance as SCR even though they were carefully fine tuned to achieve the best possible performance. Furthermore, the sampling size was manually set to reach the full sample size at the very last iteration. This highlights another advantage of the automatic sampling scheme that does not require knowledge of the total number of iterations.

A.2. Convergence Analysis

A.2.1. Preliminary results

Proof of Lemma 10:

The lower bound σ_{inf} follows directly from Step 7 in the algorithm design (see Algorithm 1). Within the upper bound, the constant σ_0 accounts for the start value of the penalty parameter. Now, we show that as soon as some $\sigma_k > 3(\frac{2M+C+\kappa_g}{2})$, the iteration is very successful and $\sigma_{k+1} < \sigma_k$. Finally, γ_2 allows for σ_k being 'close to' the successful threshold, but increased 'one last time'.

Any iteration with $f(\mathbf{x}_k + \mathbf{s}_k) \leq m(\mathbf{s}_k)$ yields a $\rho_k \geq 1 \geq \eta_2$ and is thus very successful. From a 2nd-order Taylor

[4] see Section A.3 for details

approximation of $f(\mathbf{x}_k + \mathbf{s}_k)$ around $\mathbf{x}_k$ we have:

$$\begin{aligned}
f(\mathbf{x}_k + \mathbf{s}_k) - m_k(\mathbf{s}_k) &= (\nabla f(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_k))^\intercal \mathbf{s}_k + \frac{1}{2}\mathbf{s}_k^\intercal(\mathbf{H}(\mathbf{x}_k + t\mathbf{s}_k) - \mathbf{B}_k)\mathbf{s}_k - \frac{\sigma}{3}\|\mathbf{s}_k\|^3 \\
&\leq \mathbf{e}_k^\intercal \mathbf{s}_k + \frac{1}{2}\|\mathbf{s}_k\|^2 \|\mathbf{H}(\mathbf{x}_k + t\mathbf{s}_k) - \mathbf{H}(\mathbf{x})\| + \frac{1}{2}\|\mathbf{H}(\mathbf{x}_k) - \mathbf{B}_k\| \|\mathbf{s}_k\| - \frac{\sigma_k}{3}\|\mathbf{s}_k\|^3 \\
&\leq \|\mathbf{e}_{\mathbf{k}}\| \|\mathbf{s}_k\| + \left(\frac{C + \kappa_g}{2} - \frac{\sigma_k}{3}\right)\|\mathbf{s}_k\|^3 \\
&\leq M\|\mathbf{s}_k\|^3 + \left(\frac{C + \kappa_g}{2} - \frac{\sigma_k}{3}\right)\|\mathbf{s}_k\|^3 \\
&= \left(\frac{2M + C + \kappa_g}{2} - \frac{\sigma_k}{3}\right)\|\mathbf{s}_k\|^3
\end{aligned} \tag{58}$$

Requiring the right hand side to be non-positive and solving for σ_k gives the desired result.

□

Proof of Lemma 11 : By definition of the stochastic model $m_k(\mathbf{s}_k)$ we have

$$\begin{aligned}
f(\mathbf{x}_k) - m_k(\mathbf{s}_k) =& - \mathbf{s}_k^\intercal \mathbf{g}(\mathbf{x}_k) - \frac{1}{2}\mathbf{s}_k^\intercal \mathbf{B}_k \mathbf{s}_k - \frac{1}{3}\sigma_k \|\mathbf{s}_k\|^3 \\
=& \frac{1}{2}\mathbf{s}_k^\intercal \mathbf{B}_k \mathbf{s}_k + \frac{2}{3}\sigma_k \|\mathbf{s}_k\|^3 \\
\geq& \frac{1}{6}\sigma_k \|\mathbf{s}_k\|^3,
\end{aligned} \tag{59}$$

where we applied equation (11) first and equation (12) secondly.

□

Before proving the lower bound on the stepsize $\|\mathbf{s}_k\|$ we first transfer the rather technical result from Lemma 4.6 in (Cartis et al., 2011a) to our framework of stochastic gradients. For this purpose, let $\mathbf{e}_k$ be the gradient approximation error, i.e. $\mathbf{e}_k := \mathbf{g}_k - \nabla f(\mathbf{x}_k)$.

Lemma 21. *Let $f \in C^2$, Lipschitz continuous gradients (A3) and TC (A2) hold. Then, for each (very-)successful k, we have*

$$\begin{aligned}
&(1 - \kappa_\theta)\|\nabla f(\mathbf{x}_{k+1})\| \leq \sigma_k \|\mathbf{s}_k\|^2 + \\
&\underbrace{\left(\left\|\int_0^1 (\mathbf{H}(\mathbf{x}_k + t\mathbf{s}_k) - \mathbf{H}(\mathbf{x}_k))dt\right\| + \frac{\|(\mathbf{H}(\mathbf{x}_k) - \mathbf{B}_k)\mathbf{s}_k\|}{\|\mathbf{s}_k\|} + \kappa_\theta \kappa_g \|\mathbf{s}_k\| + (1 + \kappa_\theta \kappa_g)\frac{\|\mathbf{e}_k\|}{\|\mathbf{s}_k\|}\right)}_{=d_k} \cdot \|\mathbf{s}_k\|
\end{aligned} \tag{60}$$

with $\kappa_\theta \in (0, 1)$ as in TC (13).

Proof: We shall start by writing

$$\begin{aligned}
\|\nabla f(\mathbf{x}_k + \mathbf{s}_k)\| &\leq \|\nabla f(\mathbf{x}_k + \mathbf{s}_k) - \nabla m_k(\mathbf{s}_k)\| + \|\nabla m_k(\mathbf{s}_k)\| \\
&\leq \underbrace{\|\nabla f(\mathbf{x}_k + \mathbf{s}_k) - \nabla m_k(\mathbf{s}_k)\|}_{(a)} + \underbrace{\theta_k \|\mathbf{g}_k(\mathbf{x}_k)\|}_{(b)},
\end{aligned} \tag{61}$$

where the last inequality results from TC (Eq. (13)). Now, we can find the following bounds on the individual terms:

(a) By (5) we have

$$\|\nabla f(\mathbf{x}_k + \mathbf{s}_k) - \nabla m_k\| = \|\nabla f(\mathbf{x}_k + \mathbf{s}_k) - \mathbf{g}_k(\mathbf{x}_k) - \mathbf{B}_k \mathbf{s}_k - \sigma_k \mathbf{s}_k \|\mathbf{s}_k\|\|. \tag{62}$$

We can rewrite the right-hand side by a Taylor expansion of $\nabla f_{k+1}(\mathbf{x}_k + \mathbf{s}_k)$ around $\mathbf{x}_k$ to get

$$(62) = \left\| \nabla f(\mathbf{x}_k) + \int_0^1 \mathbf{H}(\mathbf{x}_k + t\mathbf{s}_k)\mathbf{s}_k dt - \mathbf{g}_k(\mathbf{x}_k) - \mathbf{B}_k\mathbf{s}_k - \sigma_k \mathbf{s}_k \|\mathbf{s}_k\| \right\|. \tag{63}$$

Contrary to the case of deterministic gradients, the first and third summand no longer cancel out. Applying the triangle inequality repeatedly, we thus get an error term in the final bound on (a):

$$\begin{aligned} \|\nabla f(\mathbf{x}_k + \mathbf{s}_k) - \nabla m_k\| &\leq \left\| \int_0^1 \mathbf{H}((\mathbf{x}_k + t\mathbf{s}_k) - \mathbf{B}_k)\mathbf{s}_k dt \right\| + \sigma_k \|\mathbf{s}_k\|^2 + \|\nabla f(\mathbf{x}_k) - \mathbf{g}_k(\mathbf{x}_k)\| \\ &\leq \left\| \int_0^1 \mathbf{H}((\mathbf{x}_k + t\mathbf{s}_k)dt - \mathbf{H}(\mathbf{x}_k) \right\| \cdot \|\mathbf{s}_k\| + \|(\mathbf{H}(\mathbf{x}_k) - \mathbf{B}_k)\mathbf{s}_k\| \\ &\quad + \sigma_k \|\mathbf{s}_k\|^2 + \|\mathbf{e}_k\|. \end{aligned} \tag{64}$$

(b) To bound the second summand, we can write

$$\begin{aligned} \|\mathbf{g}(\mathbf{x}_k)\| &\leq \|\nabla f(\mathbf{x}_k)\| + \|\mathbf{e}_k\| \\ &\leq \|\nabla f(\mathbf{x}_k + \mathbf{s}_k)\| + \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k + \mathbf{s}_k)\| + \|\mathbf{e}_k\| \\ &\leq \|\nabla f(\mathbf{x}_k + \mathbf{s}_k)\| + \kappa_g \|\mathbf{s}_k\| + \|\mathbf{e}_k\| \end{aligned} \tag{65}$$

Finally, using the definition of θ_k as in (13) (which also gives $\theta_k \leq \kappa_\theta$ and $\theta_k \leq \kappa_\theta h_k$) and combining (a) and (b) we get the above result.

□

Proof of Lemma 12: The conditions of Lemma 21 are satisfied. By multiplying $d_k \|\mathbf{s}_k\|$ out in equation (60), we get

$$\begin{aligned} &(1 - \kappa_\theta) \|\nabla f(\mathbf{x}_{k+1})\| \leq \\ &\left\| \int_0^1 (\mathbf{H}(\mathbf{x}_k + t\mathbf{s}_k) - \mathbf{H}(\mathbf{x}_k))dt \right\| \|\mathbf{s}_k\| + \|(\mathbf{H}(\mathbf{x}_k) - \mathbf{B}_k)\mathbf{s}_k\| + \kappa_\theta\kappa_g \|\mathbf{s}_k\|^2 + (1 + \kappa_\theta\kappa_g) \|\mathbf{e}_k\| + \sigma_k \|\mathbf{s}_k\|^2. \end{aligned} \tag{66}$$

Now, applying the strong agreement conditions (A4) and (A5), as well as the Lipschitz continuity of H, we can rewrite this as

$$(1 - \kappa_\theta) \|\nabla f(\mathbf{x}_{k+1})\| \leq (\frac{1}{2}\kappa_g + C + (1 + \kappa_\theta\kappa_g)M + \sigma_{\max} + \kappa_\theta\kappa_g) \|\mathbf{s}_k\|^2, \tag{67}$$

for all sufficiently large, successful k. Solving for the stepsize $\|\mathbf{s}_k\|$ give the desired result.

□

A.2.2. Local convergence

Before we can study the convergence rate of SCR in a locally convex neighbourhood of a local minimizer w_* we first need to establish three crucial properties:

1. a lower bound on $\|\mathbf{s}_k\|$ that depends on $\|\mathbf{g}_k\|$.
2. an upper bound on $\|\mathbf{s}_k\|$ that depends on $\|\mathbf{g}_{k+1}\|$.
3. an eventually full sample size
4. conditions under which all steps are eventually very successful.

With this at hand we will be able to relate $\|\mathbf{g}_{k+1}\|$ to $\|\mathbf{g}_k\|$, show that this ratio eventually goes to zero at a quadratic rate and conclude from a Taylor expansion around $\mathbf{g}_k$ that the iterates themselves converge as well.

Assumption 22 (Sampling Scheme)**.** *Let $\mathbf{g}_k$ and $\mathbf{B}_k$ be sampled such that 17 and 19 hold in each iteration k. Furthermore, for unsuccessful iterations, assume that the sample size is not decreasing.*

We have already established a lower stepsize bound in Lemma 12 so let us turn our attention directly to 2.:

Lemma 23 (Upper bound on stepsize)**.** *Suppose that $\mathbf{s}_k$ satisfies (11) and that the Rayleigh coefficient*

$$R_k(\mathbf{s}_k) := \frac{\mathbf{s}_k^\intercal \mathbf{B}_k \mathbf{s}_k}{\|\mathbf{s}_k\|^2} \tag{68}$$

is positive, then

$$\|\mathbf{s}_k\| \leq \frac{1}{R_k(\mathbf{s}_k)} \|\mathbf{g}_k\| = \frac{1}{R_k(\mathbf{s}_k)} \|\nabla f(\mathbf{w}_k) + \mathbf{e}_k\| \leq \frac{1}{R_k(\mathbf{s}_k)} (\|\nabla f(\mathbf{w}_k)\| + \|\mathbf{e}_k\|) \tag{69}$$

Proof: Given the above assumptions we can rewrite (11) as follows

$$R_k(\mathbf{s}_k) \|\mathbf{s}_k\|^2 = -\mathbf{s}_k^\intercal \mathbf{g}_k - \sigma_k \|\mathbf{s}_k\|^3 \leq \|\mathbf{s}_k\| \|\mathbf{g}_k\|, \tag{70}$$

where we used Cauchy-Schwarz inequality as well as the fact that $\sigma_k > 0, \ \forall k$. Solving (70) for $\|\mathbf{s}_k\|$ gives (69).

□

Lemma 24 (Eventually full sample size)**.** *Let $\{f(\mathbf{x}_k)\}$ be bounded below by some $f_{\text{inf}} > -\infty$. Also, let A1, A3 hold and let $\mathbf{g}_k$ and $\mathbf{B}_k$ be sampled according to A22. Then we have w.h.p. that*

$$|S_{g,k}| \to n \text{ and } |S_{B,k}| \to n \text{ as } k \to \infty \tag{71}$$

The sampling schemes from Theorem 7 and Theorem 9 imply that the sufficient agreement assumptions A5 and A4 hold with high probability. Thus, we can deduce from Lemma 10 that after a certain number of consecutive unsuccessful iterates the penalty parameter is so high ($\sigma_k \geq \sigma_{sup}$) that we are guaranteed to find a successful step. Consequently, the number of successful iterations must be infinite ($|\mathcal{S}| = \infty$) when we consider the asymptotic convergence properties of SCR. We are left with two possible scenarios:

(i) If the number of unsuccessful iterations is finite ($|\mathcal{U}| \leq \infty$ & $|\mathcal{S}| = \infty$) we have that $\exists\ \hat{k}$ after which all iterates are successful, i.e. $k \in \mathcal{S}, \forall\ k > \hat{k}$. From Lemma 20 we know that for all successful iterations $\|\mathbf{s}_k\| \to 0$ as $k \to \infty$. Consequently, due to the sampling scheme as specified in Theorem 7 and Theorem 9, $\exists\ \bar{k} \geq \hat{k}$ with $|S_{g,k}| = |S_{B,k}| = n, \ \forall\ k \geq \bar{k}$.

(ii) If the number of unsuccessful iterations is infinite ($|\mathcal{U}| = \infty$ & $|\mathcal{S}| = \infty$) we know for the same reasons that for the *sub*sequence of successful iterates $\{k = 0, 1, \dots \infty | k \in \mathcal{S}\}$ again $\|\mathbf{s}_k\| \to 0$, as $k \in \mathcal{S} \to \infty$ and hence $\exists\ \tilde{k}$ with $|S_{g,k}| = |S_{B,k}| = n, \ \forall\ k \geq \tilde{k} \in \mathcal{S}$. Given that we do specifically not decrease the sample size in unsuccessful iterations we have that $|S_{g,k}| = |S_{B,k}| = n, \ \forall\ k \geq \tilde{k}$.

As a result the sample sizes eventually equal n with high probability in all conceivable scenarios which proves the assertion[5].

□

Now that we have (asymptotic) stepsize bounds and gradient (Hessian) agreement we are going to establish that, when converging, all SCR iterations are indeed very successful asymptotically.

Lemma 25 (Eventually successful iterations)**.** *Let $f \in C^2$, ∇f uniformly continuous and $\mathbf{B}_k$ bounded above. Let $\mathbf{B}_k$ and $\mathbf{g}_k$ be sampled according to A22, as well as $\mathbf{s}_k$ satisfy (11). Furthermore, let*

$$\mathbf{w}_k \to \mathbf{w}_*, \text{ as } k \to \infty, \tag{72}$$

with $\nabla f(\mathbf{w}_*) = 0$ and $\mathbf{H}(\mathbf{w}_*)$ positive definite. Then there exists a constant $R_{min} > 0$ such that for all k sufficiently large

$$R_k(\mathbf{s}_k) \geq R_{\text{min}}. \tag{73}$$

Furthermore, all iterations are eventually very successful w.h.p.

[5] We shall see that, as a result of Lemma 25, the case of an infinite number of unsuccessful steps can actually not happen

Proof: Since f is continuous, the limit (72) implies that $\{f(\mathbf{w}_k)\}$ is bounded below. Since $\mathbf{H}(\mathbf{w}_*)$ is positive definite per assumption, so is $\mathbf{H}(\mathbf{w}_k)$ for all k sufficiently large. Therefore, there exists a constant $R_{\min}$ such that

$$\frac{\mathbf{s}_k^\intercal \mathbf{H}(\mathbf{w}_k)\mathbf{s}_k}{\|\mathbf{s}_k\|^2} > 2R_{\min} > 0, \text{ for all } k \text{ sufficiently large.} \tag{74}$$

As a result of Lemma 24 we have that $\|\mathbf{e}_k\| \to 0$ as $k \to \infty$. Hence, Lemma 23 yields $\|\mathbf{s}_k\| \leq 1/R_{\min} \|\nabla f_k\|$ which implies that the step size converges to zero as we approximate w^*. Consequently, we are able to show that eventually all iterations are indeed very successful. Towards this end we need to ensure that the following quantity r_k becomes negative for sufficiently large k:

$$r_k := \underbrace{f(\mathbf{w}_k + \mathbf{s}_k) - m(\mathbf{s}_k)}_{(i)} + (1 - \eta_2)\underbrace{(m(\mathbf{s}_k) - f(\mathbf{w}_k))}_{(ii)}, \tag{75}$$

where $\eta_2 \in (0, 1)$ is the ”very successful” threshold.

(i) By a (second-order) Taylor approximation around $f(\mathbf{w}_k)$ and applying the Cauchy-Schwarz inequality, we have:

$$\begin{aligned} f(\mathbf{w}_k + \mathbf{s}_k) - m(\mathbf{s}_k) =& (\nabla f(\mathbf{w}) - \mathbf{g}_k)^\intercal \mathbf{s}_k + \frac{1}{2}\mathbf{s}_k^\intercal((\mathbf{H}(\mathbf{w}_k + \tau\mathbf{s}_k) - \mathbf{B}_k)\mathbf{s}_k - \frac{\sigma_k}{3}\|\mathbf{s}\|^3 \\ \leq& \|\mathbf{e}_k\| \|\mathbf{s}_k\| + \frac{1}{2}\|((\mathbf{H}(\mathbf{w}_k + \tau\mathbf{s}_k) - \mathbf{B}_k)\mathbf{s}_k\| \|\mathbf{s}_k\|, \end{aligned} \tag{76}$$

where the term $\|\mathbf{e}_k\| \|\mathbf{s}_k\|$ is extra compared to the case of deterministic gradients.

(ii) Regarding the second part we note that if $\mathbf{s}_k$ satisfies (11), we have by the definition of R_k and equation (73) that

$$\begin{aligned} f(\mathbf{w}_k) - m_k(\mathbf{s}_k) =& \frac{1}{2}\mathbf{s}_k^\intercal B\mathbf{s}_k + \frac{2}{3}\sigma_k\|\mathbf{s}_k\|^3 \\ \geq& \frac{1}{2}R_{\min}\|\mathbf{s}_k\|^2, \end{aligned} \tag{77}$$

which negated gives the desired bound on (ii). All together, the upper bound on r_k is written as

$$r_k \leq \frac{1}{2}\|\mathbf{s}_k\|^2 \left(\frac{2\|\mathbf{e}_k\|}{\|\mathbf{s}_k\|} + \frac{\|((\mathbf{H}(\mathbf{w}_k + \tau\mathbf{s}_k) - \mathbf{B}_k)\mathbf{s}_k\|}{\|\mathbf{s}_k\|} - (1 - \eta_2)R_{\min}\right). \tag{78}$$

Let us add and subtract $\mathbf{H}(\mathbf{w}_k)$ to the second summand and apply the triangle inequality

$$r_k \leq \frac{1}{2}\|\mathbf{s}_k\|^2 \left(\frac{2\|\mathbf{e}_k\|}{\|\mathbf{s}_k\|} + \frac{\|(\mathbf{H}(\mathbf{w}_k + \tau\mathbf{s}_k) - \mathbf{H}_k)\mathbf{s}_k\| + \|(\mathbf{H}_k - \mathbf{B}_k)\mathbf{s}_k\|}{\|\mathbf{s}_k\|} - (1 - \eta_2)R_{\min}\right). \tag{79}$$

Now applying $\|\mathbf{A}\mathbf{v}\| \leq \|\mathbf{A}\| \|\mathbf{v}\|$ we get

$$r_k \leq \frac{1}{2}\|\mathbf{s}_k\|^2 \left(\frac{2\|\mathbf{e}_k\|}{\|\mathbf{s}_k\|} + \|\mathbf{H}(\mathbf{w}_k + \tau\mathbf{s}_k) - \mathbf{H}_k\| + \|(\mathbf{H}_k - \mathbf{B}_k)\| - (1 - \eta_2)R_{\min}\right). \tag{80}$$

We have already established in Lemma 24 that $\|\mathbf{e}_k\| \to 0$ and $\|(\mathbf{H}_k - \mathbf{B}_k)\| \to 0$. Together with Lemma 23 and the assumption $\|\nabla f_k\| \to 0$ this implies $\|\mathbf{s}_k\| \to 0$. Furthermore, since $\tau \in [0, 1]$ we have that $\|\mathbf{w}_k + \tau\mathbf{s}_k\| \leq \|\mathbf{w}_k + \mathbf{s}_k\| \leq \|\mathbf{s}_k\|$. Hence, $\mathbf{H}(\mathbf{w}_k + \tau\mathbf{s}_k)$ and $\mathbf{H}(\mathbf{w}_k)$ eventually agree. Finally, $\eta_2 < 1$ and $R_{\min} > 0$ such that r_k is negative for all k sufficiently large, which implies that every such iteration is very successful.

□

Proof of Theorem 13:

From Lemma 10 we have $\sigma_k \leq \sigma_{sup}$. Furthermore, all assumptions needed for the step size bounds of Lemma 12 and 23 hold. Finally, Lemma 25 gives that all iterations are eventually successful. Thus, we can combine the upper (69) and lower (24) bound on the stepsize for all k sufficiently large to obtain

$$\frac{1}{R_{\min}}(\|\nabla f(\mathbf{w}_k)\| + \|\mathbf{e}_k\|) \geq \|\mathbf{s}_k\| \geq \kappa_s\sqrt{\|\nabla f(\mathbf{w}_{k+1})\|} \tag{81}$$

which we can solve for the gradient norm ratio

$$\frac{\|\nabla f(\mathbf{w}_{k+1})\|}{\|\nabla f(\mathbf{w}_k)\|^2} \leq \left(\frac{1}{R_{\min}\kappa_s} \left(1 + \frac{\|\mathbf{e}_k\|}{\|\nabla f(\mathbf{w}_k)\|} \right) \right)^2 . \tag{82}$$

Consequently, as long as the right hand side of (82) stays below infinity, i.e. $\|\mathbf{e}_k\| \,/\, \|\nabla f(\mathbf{w}_k)\| \not\to \infty$, we have quadratic convergence of the gradient norms. From Lemma 24 we have that $\|\mathbf{e}_k\| \to 0$ as $k \to \infty$ w.h.p. and furthermore κ_s is bounded above by a constant and $R_{\min}$ is a positive constant itself which gives quadratic convergence of the gradient norm ratio with high probability. Finally, the convergence rate of the iterates follows from a Taylor expansion around $\mathbf{g}_k$.

□

A.2.3. First order global convergence

Note that the preliminary results Lemma 11 and 12 allow us to lower bound the function decrease of a successful step in terms of the *full* gradient ∇f_{k+1}. Combined with Lemma 10, this enables us to give a *deterministic* global convergence guarantee while using only *stochastic* first order information[6].

Proof of Theorem 14:

We will consider two cases regarding the number of successful steps for this proof.

Case (i): SCR takes only finitely many successful steps. Hence, we have some index k_0 which yields the very last successful iteration and all further iterates stay at the same point $\mathbf{x}_{k_0+1}$. That is $\mathbf{x}_{k_0+1} = \mathbf{x}_{k_0+i}, \; \forall \, i \geq 1$. Let us assume that $\|\nabla f(\mathbf{x}_{k_0+1})\| = \epsilon > 0$, then

$$\|\nabla f(\mathbf{x}_k)\| = \epsilon, \; \forall \, k \geq k_0 + 1. \tag{83}$$

Since, furthermore, all iterations $k \geq k_0 + 1$ are unsuccessful σ_k increases by γ, such that

$$\sigma_k \to \infty \text{ as } k \to \infty. \tag{84}$$

However, this is in contradiction with Lemma 10, which states that σ_k is bounded above. Hence, the above assumption cannot hold and we have $\|\nabla f(\mathbf{x}_{k_0+1})\| = \|\nabla f(\mathbf{x}^*)\| = 0$.

Case (ii): sARC takes infinitely many successful steps. While unsuccessful steps keep $f(\mathbf{x}_k)$ constant, (very) successful steps strictly decrease $f(\mathbf{x}_k)$ and thus the sequence $\{f(\mathbf{x}_k)\}$ is monotonically decreasing. Furthermore, it is bounded below per assumption and thus the objective values converge

$$f(\mathbf{x}_k) \to f_{\inf}, \text{ as } k \to \infty. \tag{85}$$

All requirements of Lemma 11 and Lemma 12 hold and we thus can use the sufficient function decrease equation (31) to write

$$f(\mathbf{x}_k) - f_{\inf} \geq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{6}\eta_1 \sigma_{\inf} \kappa_s^3 \, \|\nabla f(\mathbf{x}_{k+1})\|^{3/2} . \tag{86}$$

Since $(f(\mathbf{x}_k) - f_{\inf}) \to 0$ as $k \to \infty$, $\sigma_{\inf} > 0, \eta_1 > 0$ and $\kappa_s^3 \geq 0$ (as $\sigma_{\sup} < \infty$), we must have $\|\nabla f(\mathbf{x}_k)\| \to 0$, giving the result.

□

A.2.4. Second order global convergence and worst case iteration complexity

For the proofs of Theorem 15 and Theorem 17 we refer the reader to Theorem 5.4 in (Cartis et al., 2011a) and Corollary 5.3 in (Cartis et al., 2011b). Note that, as already laid out above in the proofs of Lemma 10 and Lemma 11, the constants involved in the convergence Theorems change due to the stochastic gradients used in our framework.

A.3. Details concerning experimental section

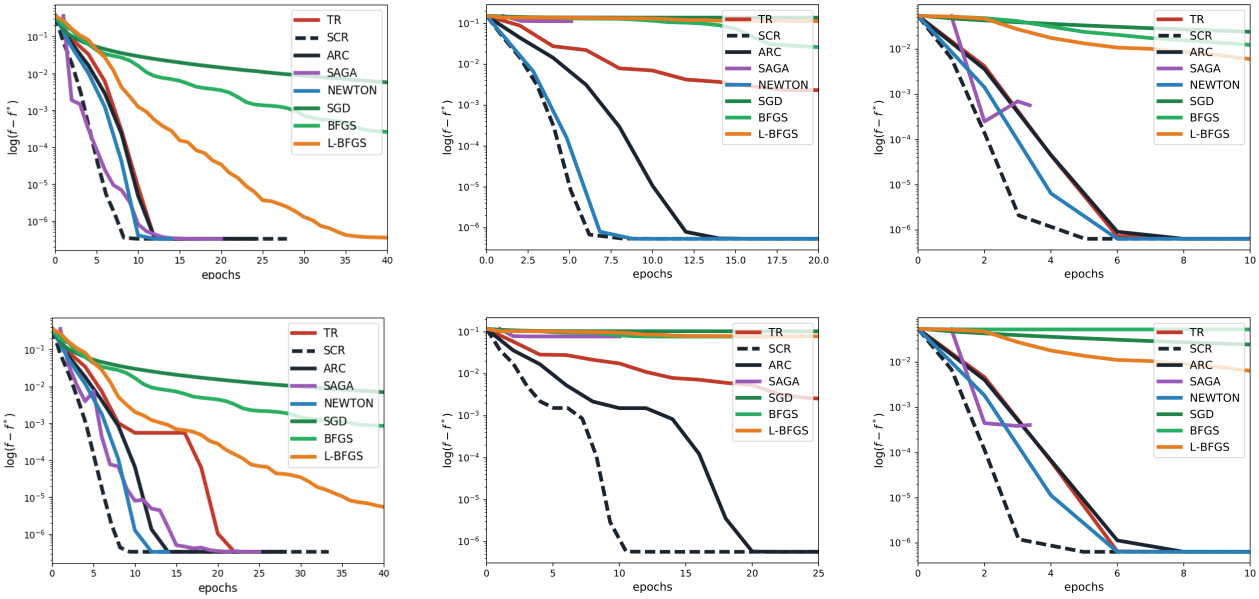
We here provide additional results and briefly describe the baseline algorithms used in the experiments as well as the choice of hyper-parameters. All experiments were run on a CPU with a 2.4 GHz nominal clock rate.

[6] Note that this result can also be proven without Lipschitz continuity of H and less strong agreement conditions as done in Corollary 2.6 in (Cartis et al., 2011a).

Datasets The real-world datasets we use represent very common instances of Machine Learning problems and are part of the libsvm library (Chang & Lin, 2011), except for *cifar* which is from Krizhevsky & Hinton (2009). A summary of their main characteristic can be found in table 1. The multiclass datasets are both instances of so-called image classification problems. The *mnist* images are greyscale and of size 28×28. The original *cifar* images are $32 \times 32 \times 3$ but we converted them to greyscale so that the problem dimensionality is comparable to *mnist*. Both datasets have 10 different classes, which multiplies the problem dimensionality of the multinomial regression by 10.

	type	n	d	$\kappa(H*)$	λ
a9a	Classification	$32,561$	123	761.8	$1e^{-3}$
a9a nc	Classification	$32,561$	123	$1,946.3$	$1e^{-3}$
covtype	Classification	$581,012$	54	$3 \cdot 10^9$	$1e^{-3}$
covtype nc	Classification	$581,012$	54	$25,572,903.1$	$1e^{-3}$
higgs	Classification	$11,000,000$	28	$1,412.0$	$1e^{-4}$
higgs nc	Classification	$11,000,000$	28	$2,667.7$	$1e^{-4}$
mnist	Multiclass	$60,000$	$7,840$	$10,281,848$	$1e^{-3}$
cifar	Multiclass	$50,000$	$10,240$	$1 \cdot 10^9$	$1e^{-3}$

Table 1. Overview over the real-world datasets used in our experiments with convex and non-convex (nc) regularizer. $\kappa(H*)$ refers to the condition number of the Hessian at the optimizer and λ is the regularization parameter applied in the loss function and its derivatives.



1. A9A 2. COVTYPE 3. HIGGS

Figure 3. Results from Section 5 over epochs. Top (bottom) row shows the log suboptimality of convex (non-convex) regularized logistic regressions over epochs (average of 10 independent runs).

Benchmark methods

- Stochastic Gradient Descent (SGD): To bring in some variation, we select a mini-batch of the size $\lceil n/10 \rceil$ on the real world classification- and $\lceil n/100 \rceil$ on the multiclass problems. On the artificial datasets we only sample 1 datapoint per iteration and update the parameters with respect to this point. We use a problem-dependent, constant step-size as this yields faster initial convergence (Hofmann et al., 2015),(Roux et al., 2012).
- SAGA: is a variance-reduced variant of SGD that only samples 1 datapoint per iteration and uses a constant step-size.

- Broyden-Fletcher-Goldfarb-Shanno (BFGS) is the most popular and stable Quasi-Newton method.
- Limited-memory BFGS is a variant of BFGS which uses only the recent K iterates and gradients to construct an approximate Hessian. We used $K = 20$ in our experiments. Both methods employs a line-search technique that satisfies the strong Wolfe condition to select the step size.
- NEWTON is the classic version of Newton's method which we apply with a backtracking line search.

For L-BFGS and BFGS we used the implementation available in the optimization library of scipy. All other methods are our own implementation. The code for our implementation of SCR is publicly available on the authors' webpage.

Initialization. All of our experiments were started from the initial weight vector $\mathbf{w}_0 := (0, \ldots, 0)$.

Choice of parameters for ARC and SCR. The regularization parameter updating is analog to the rule used in the reported experiments of (Cartis et al., 2011a), where $\gamma = 2$. Its goal is to reduce the penalty rapidly as soon as convergence sets in, while keeping some regularization in the non asymptotic regime. A more sophisticated approach can be found in (Gould et al., 2012). In our experiments we start with $\sigma_0 = 1, \eta_1 = 0.2,$ and $\eta_2 = 0.8$ as well as an initial sample size of 5%.

Influence of dimensionality To test the influence of the dimensionality on the progress of the above applied methods we created artificial datasets of three different sizes, labeled as *gaussian s*, *gaussian m* and *gaussian l*.

	type	n	d	$\kappa(H^*)$	λ
gaussian s	Classification	$50,000$	100	$2,083.3$	$1e^{-3}$
gaussian m	Classification	$50,000$	$1,000$	$98,298.9$	$1e^{-3}$
gaussian l	Classification	$50,000$	$10,000$	$1,167,211.3$	$1e^{-3}$

Table 2. Overview over the synthetic datasets used in our experiments with convex regularizer

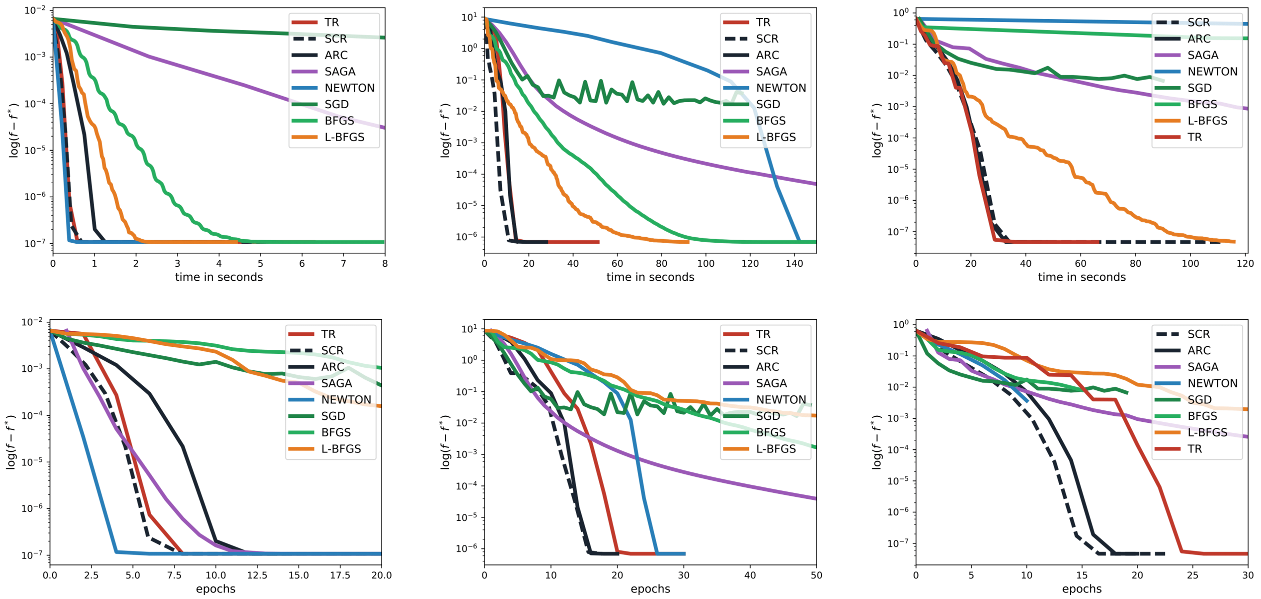
The feature vectors $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d), \mathbf{x}_i \in \mathbb{R}^n$ were drawn from a multivariate Gaussian distribution

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{87}$$

with a mean of zero $\mu = (0, \ldots, 0)$ and a covariance matrix that has reasonably uniformly distributed off-diagonal elements in the interval $(-1, 1)$.

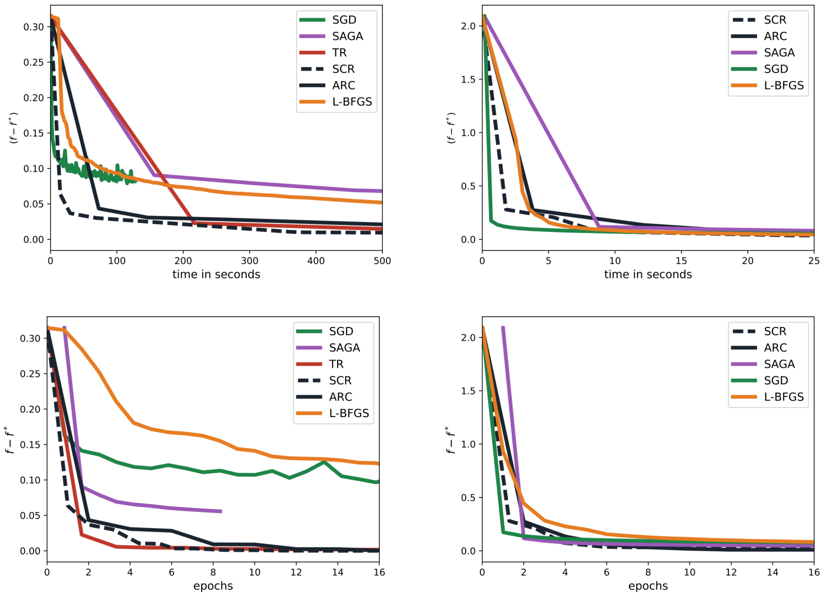
As expected, the classic Newton methods suffers heavily from an increase in the dimension. The regularized Newton methods on the other hand scale comparably very well since they only need indirect access to the Hessian via matrix-vector products. Evidently, these methods outperform the quasi-newton approaches even in high dimensions. Among these, the limited memory version of BFGS is significantly faster than its original variant.

Multiclass regression In this section we leave the trust region method out because our implementation is not optimized towards solving multi-class problems. We do not run Newton's method or BFGS either as the above results suggests that they are unlikely to be competitive. Furthermore, Figure 5 does not show logarithmic but linear suboptimality because optimizing these problems to high precision takes very long and yields few additional benefits. For example, the 25th SCR iteration drove the gradient norm from $3.8 \cdot 10^{-5}$ to $5.6 \cdot 10^{-8}$ after building up a Krylov space of dimensionality 7800. It took 9.47 hours and did *not* change any of the first 13 digits of the loss. As can be seen, SCR provides early progress at a comparable rate to other methods but gives the opportunity to solve the problem to high precision if needed.



1. GAUSSIAN S 2. GAUSSIAN M 3. GAUSSIAN L

Figure 4. Top (bottom) row shows the log suboptimality of convex regularized logistic regressions over time (epochs) (average of 10 independent runs).



1. CIFAR 2. MNIST

Figure 5. Top (bottom) row shows suboptimality of the empirical risk of convex regularized multinominal regressions over time (epochs)