Supplementary material for “Proximal Newton-type methods for convex optimization”

Jason D. Lee† and Yuekai Sun∗
Institute for Computational and Mathematical Engineering
Stanford University, Stanford, CA
{jdll7, yuekai}@stanford.edu

Michael A. Saunders
Department of Management Science and Engineering
Stanford University, Stanford, CA
saunders@stanford.edu

November 9, 2012

A Proofs

A.1 Proof of Lemma 2.2

Proof. $h$ is convex so for $t \in (0, 1]$, we have
\[
\begin{align*}
f(x^+) - f(x) &= g(x^+) - g(x) + h(x^+) - h(x) \\
&\leq g(x^+) - g(x) + th(x + \Delta x) + (1-t)h(x) - h(x) \\
&= g(x^+) - g(x) + t(h(x + \Delta x) - h(x)) \\
&= \nabla g(x)^T (t\Delta x) + t(h(x + \Delta x) - h(x)) + O(t^2),
\end{align*}
\]
which proves (8).

$\Delta x$ steps to the minimizer of $h$ plus our quadratic approximation to $g$ so $t\Delta x$ satisfies
\[
\begin{align*}
\nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T H \Delta x + h(x + \Delta x) \\
&\leq \nabla g(x)^T (t\Delta x) + \frac{t^2}{2} \Delta x^T H \Delta x + h(x^+) \\
&\leq t \nabla g(x)^T \Delta x + \frac{t^2}{2} \Delta x^T H \Delta x + t(h(x + \Delta x) - h(x)) + (1-t)h(x).
\end{align*}
\]
We can rearrange and then simplify to obtain
\[
(1-t) \nabla g(x)^T \Delta x + \frac{1}{2}(1-t^2) \Delta x^T H \Delta x + (1-t)(h(x + \Delta x) - h(x)) \leq 0,
\]
\[
\nabla g(x)^T \Delta x + \frac{1}{2}(1+t) \Delta x^T H \Delta x + h(x + \Delta x) - h(x) \leq 0,
\]
\[
\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \leq \frac{1}{2}(1+t) \Delta x^T H \Delta x.
\]
Finally, we let $t \to 1$ and rearrange to obtain (9).

∗ Equal contributors
† Equal contributors
A.2 Proof of Lemma 2.3

Proof. We can bound the decrease at each iteration by
\[
\begin{align*}
    f(x^+) - f(x) &= g(x^+) - g(x) + h(x^+) - h(x) \\
    &\leq \int_0^1 \nabla g(x + s(t\Delta x))^T(t\Delta x)ds + th(x + \Delta x) + (1 - t)h(x) - h(x) \\
    &= \nabla g(x)^T(t\Delta x) + t(h(x + \Delta x) - h(x)) \\
    &\quad + \int_0^1 (\nabla g(x + s(t\Delta x)) - \nabla g(x))^T(t\Delta x)ds \\
    &\leq t \left( \nabla g(x)^T(t\Delta x) + h(x + \Delta x) - h(x) \\
    &\quad + \int_0^1 \|\nabla g(x + s(\Delta x)) - \nabla g(x)\|\Delta x\|ds \right).
\end{align*}
\]
\(\nabla g\) is Lipschitz continuous so
\[
    f(x^+) - f(x) \leq t \left( \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) + \frac{L_1t^2}{2} \|\Delta x\|^2 \right) \\
    = t \left( \Delta + \frac{L_1t}{2} \|\Delta x\|^2 \right). \tag{18}
\]
If we choose \(t \leq \frac{2m}{L_1}(1 - \alpha)\), then
\[
    \frac{L_1t}{2} \|\Delta x\|^2 \leq m(1 - \alpha)\|\Delta x\|^2 \leq (1 - \alpha)\Delta x^T H \Delta x \leq -(1 - \alpha)\Delta. \tag{19}
\]
We can substitute (19) into (18) to obtain
\[
    f(x^+) - f(x) \leq t (\Delta - (1 - \alpha)\Delta) = t(\alpha\Delta).
\]

A.3 Proof of Theorem 3.2

Proof. \{f(x_k)\} is a nonincreasing sequence because because \(\Delta x\) is a descent direction, and there exist step lengths that satisfy (10) (Lemma 2.3). \(f\) is also bounded below so \{f(x_k)\} must converge; i.e.
\[
f(x_k) - f(x_{k+1}) = \alpha t_k \Delta_k \to 0.
\]
The step lengths \(t_k\) are bounded away from zero because sufficiently small step lengths satisfy the sufficient descent condition so \(\Delta_k\) must decay to zero. \(\Delta_k\) satisfies
\[
    \Delta_k = \nabla g(x_k)^T \Delta x_k + h(x_k + \Delta x_k) - h(x_k) \\
    \leq -\Delta x_k^T H_k \Delta x_k \leq -m\|\Delta x_k\|^2,
\]
where first inequality follows from (9). We reverse this inequality to obtain
\[
    \|\Delta x_k\|^2 \leq \frac{1}{m} \Delta x_k^T H_k \Delta x_k \leq -\frac{1}{m} \Delta_k
\]
so the search directions \(\Delta x_k\) must also converge to zero. This is sufficient the sequence \{x_k\} converges to be a minimizer of \(f\) (Lemma 3.1).

A.4 Proof of Lemma 3.4

Proof. \(h\) is convex, so \(\partial h\) is monotone. \(H\) is a symmetric, positive definite matrix so we have
\[
    (\partial h(x) - \partial h(y))^T(x - y) \geq 0
    \]
\[
    (x - y)^T H(x - y) \geq m\|x - y\|^2.
\]

11
We add the two equations above and divide by $m$ to obtain

$$
\frac{1}{m}(Hx + \partial h(x) - Hy + \partial h(y))^T(x - y) \geq \|x - y\|^2
$$

$$
\left(\frac{1}{m}(H + \partial h) \right)(x) - \left(\frac{1}{m}(H + \partial h) \right)(y))^T(x - y) \geq \|x - y\|^2.
$$

Let $u$ and $v$ denote $\frac{1}{m}(H + \partial h)(x)$ and $\frac{1}{m}(H + \partial h)(y)$ respectively. Then, after simplifying,

$$(u - v)^T(R(u) - R(v)) \geq \|R(u) - R(v)\|^2.
$$

Proof. The assumptions of Lemma 3.3 are satisfied so step lengths of unity satisfy the sufficient descent condition after sufficiently many iterations. Hence, for $k$ sufficiently large, we have

$$
x_{k+1} = \text{prox}_{H_k}^1(x_k - H_k^{-1}\nabla g(x_k)).
$$

Let $\nabla S_k(x)$ denote $\frac{1}{m}(H_k - \nabla^2 g(x))$. $R$ is nonexpansive (Lemma 3.4) so

$$
\|x_{k+1} - x^*\| \leq \|R_k \circ S_k(x_k) - R_k \circ S_k(x^*)\|
$$

$$
\leq \|S_k(x_k) - S_k(x^*)\|
$$

$$
\leq \|S_k(x_k) - S_k(x^*) - \nabla S_k(x^*)(x_k - x^*)\|
$$

$$
+ \|\nabla S_k(x^*)(x_k - x^*)\|. \tag{20}
$$

We choose $H_k = \nabla^2 g(x_k)$ and $\nabla^2 g$ is Lipschitz continuous; hence

$$
\|\nabla S_k(x^*)(x_k - x^*)\| \leq \frac{1}{m}\|\nabla^2 g(x_k) - \nabla^2 g(x^*)\|\|x_k - x^*\|
$$

$$
\leq \frac{L}{m}\|x_k - x^*\|^2. \tag{21}
$$

$$
\{x_k\} \to x^* \text{ and } \nabla g \text{ is continuous, so for } k \text{ sufficiently large,}
$$

$$
\|S_k(x_k) - S_k(x^*) - \nabla S_k(x^*)(x_k - x^*)\|
$$

$$
= \left\|\int_0^1 (\nabla S_k(x^* + t(x_k - x^*)) - \nabla S_k(x^*))(x_k - x^*)dt\right\|
$$

$$
\leq \int_0^1 \|\nabla S_k(x^* + t(x_k - x^*)) - \nabla S_k(x^*)\|\|x_k - x^*\|dt
$$

$$
\leq \int_0^1 \frac{1}{m}\|\nabla^2 g(x^*) - \nabla^2 g(x^* + t(x_k - x^*))\|\|x_k - x^*\|dt
$$

$$
\leq \int_0^1 \frac{L^2}{m}\|x_k - x^*\|^2dt \leq \frac{L^2}{2m}\|x_k - x^*\|^2. \tag{22}
$$

Substituting (21) and (22) into (20), we have

$$
\|x_{k+1} - x^*\| \leq \frac{3L^2}{2m}\|x_k - x^*\|^2.
$$

Proof. The Lipschitz continuity of $\nabla^2 g$ imposes a cubic upper bound on $g$:

$$
g(x + t\Delta x) \leq g(x) + t\nabla g(x)^T\Delta x + \frac{1}{2}t^2 \Delta x^T\nabla^2 g(x)\Delta x + \frac{1}{6}L_2t^3\|\Delta x\|^3.
$$
We set $t = 1$ and add $h(x + \Delta x)$ to both sides to obtain
\[
f(x + \Delta x) \leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{1}{6} L_2 \|\Delta x\|^3 + h(x + \Delta x).
\]
We then add and subtract $h(x)$ and $\frac{1}{2} \Delta x^T H \Delta x$ from the right hand side and simplify to obtain
\[
f(x + \Delta x) \leq f(x) + \Delta + \frac{1}{2} \Delta x^T (\nabla^2 g(x) - H) \Delta x + \frac{1}{6} L_2 \|\Delta x\|^3 + h(x + \Delta).
\]
\[
\frac{1}{2} \Delta x^T (\nabla^2 g(x) - H) \Delta x \leq f(x) - f(x_k) - \frac{1}{2} \Delta x^T H \Delta x + \frac{1}{6} L_2 \|\Delta x\|^3
\]
Proof. $\Delta x$ and $\Delta \hat{x}$ are the solutions to their respective subproblems so they are also the solutions to
\[
\Delta x = \arg \min_d \nabla g(x)^T d + \Delta x^T H d + h(x + d),
\]
\[
\Delta \hat{x} = \arg \min_d \nabla g(x)^T d + \Delta \hat{x}^T H d + h(x + d).
\]
Hence $\Delta x$ and $\Delta \hat{x}$ satisfy
\[
\nabla g(x)^T \Delta x + \Delta x^T H \Delta x + h(x + \Delta x)
\]
\[
\leq \nabla g(x)^T \Delta \hat{x} + \Delta \hat{x}^T H \Delta \hat{x} + h(x + \Delta \hat{x})
\]
and
\[
\nabla g(x)^T \Delta \hat{x} + \Delta \hat{x}^T \hat{H} \Delta \hat{x} + h(x + \Delta \hat{x})
\]
\[
\leq \nabla g(x)^T \Delta x + \Delta x^T \hat{H} \Delta x + h(x + \Delta x).
\]
We sum these two inequalities and rearrange to obtain
\[
\Delta x^T H \Delta x - \Delta x^T (H + \hat{H}) \Delta \hat{x} + \Delta \hat{x}^T \hat{H} \Delta \hat{x} \leq 0.
\]
We can complete the square on the left hand side and rearrange to obtain
\[
\Delta x^T H \Delta x - 2 \Delta x^T H \Delta \hat{x} + \Delta \hat{x}^T H \Delta \hat{x}
\]
\[
\leq \Delta x^T (H - H) \Delta \hat{x} + \Delta \hat{x}^T (H - H) \Delta \hat{x}.
\]
The left hand side is $\|\Delta x - \Delta \hat{x}\|_H^2$ and the eigenvalues of $H$ are bounded so

$$\|\Delta x - \Delta \hat{x}\| \leq \frac{1}{\sqrt{m}} \left( \Delta x^T (\hat{H} - H) \Delta x + \Delta \hat{x}^T (H - \hat{H}) \Delta \hat{x} \right)^{1/2} \leq \frac{1}{\sqrt{m}} \left(\|\hat{H} - H\| \Delta \hat{x}\|^{1/2} \right) (\|\Delta x\| + \|\Delta \hat{x}\|)^{1/2}. \quad (24)$$

We use a result due to Tseng and Yun (Lemma 3 in [21]) to bound $(\|\Delta x\| + \|\Delta \hat{x}\|)$. Let $P = \hat{H}^{-1/2} H \hat{H}^{-1/2}$, then $\|\Delta x\|$ and $\|\Delta \hat{x}\|$ satisfy

$$\|\Delta x\| \leq \left( M \left(1 + \lambda_{\max}(P) + \sqrt{1 - 2\lambda_{\min}(P) + \lambda_{\max}(P)^2} \right) \right) \|\Delta \hat{x}\|. \quad (25)$$

We denote this constant using $c$ and conclude that

$$\|\Delta x\| + \|\Delta \hat{x}\| \leq (1 + c) \|\Delta \hat{x}\|. \quad (25)$$

We substitute this inequality into (24) to obtain

$$\|\Delta x - \Delta \hat{x}\|^2 \leq \frac{1 + c}{m} \left(\|\hat{H} - H\| \Delta \hat{x}\|^{1/2} \right) \|\Delta \hat{x}\|^{1/2}. \quad \square$$

### A.8 Proof of Theorem 3.8

**Proof.** We select unit step lengths after sufficiently many iterations (Lemma 3.6) so for large $k$, we have

$$x_{k+1} = \text{prox}_{H_k} (x_k - \nabla^2 g(x_k)^{-1} \nabla g(x_k)).$$

We can split $\|x_{k+1} - x^*\|$ into two terms:

$$\|x_{k+1} - x^*\| \leq \| x_k + \Delta x_k^{nt} - x^* \| + \| \Delta x_k - \Delta x_k^{nt} \|.$$

The first term decays to zero quadratically because the proximal Newton method converges to $x^*$ quadratically; i.e.

$$\| x_k + \Delta x_k^{nt} - x^* \| = O \left(\|x_k^{nt} - x^*\|^2 \right).$$

The second term $\|\Delta x_k - \Delta x_k^{nt}\| = O \left(\|
\left(\nabla^2 g(x_k) - H_k\right) \Delta x_k\| \right) = o(\|\Delta x_k\|)$.

We can show that $\left(\nabla^2 g(x_k) - H_k\right) \Delta x_k\| = o(\|\Delta x_k\|)$:

$$\|\nabla^2 g(x_k) - H_k\| \Delta x_k\| \leq \| (\nabla^2 g(x_k) - \nabla^2 g(x^*)) \Delta x_k\| + \| (\nabla^2 g(x^*) - H_k) \Delta x_k\| \leq L_2 \|x_k - x^*\| \|\Delta x_k\| + o(\|\Delta x_k\|).$$

thus $\|\Delta x_k^{nt}\| = o(\|\Delta x_k\|)$.

$\|\Delta x_k\|$ is within a factor $c_k$ of $\|\Delta x_k^{nt}\|$ (Lemma 3 in [21]) so

$$\|\Delta x_k\| \leq c_k \|\Delta x_k^{nt}\| = c_k \|x_{k+1}^{nt} - x_k\| \leq c_k \left(\|x_{k+1}^{nt} - x^*\| + \|x^* - x_k\|\right) \leq O \left(\|x_k - x^*\|^2 + O(\|x_k - x^*\|).$$

The second inequality follows from $c_k = O(1)$, due to the bounded eigenvalues of $H_k$ and $\nabla^2 g(x_k)$. Hence $\|\Delta x_k\| = O(\|x_k - x^*\|)$ and $\|x_{k+1} - x^*\| \leq o(\|x_k - x^*\|).$ \square