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
{jdl17,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

$$\begin{aligned} f(x^+) - f(x) &= g(x^+) - g(x) + h(x^+) - h(x) \\ &\le g(x^+) - g(x) + th(x + \Delta x) + (1-t)h(x) - h(x) \\ &= g(x^+) - g(x) + t(h(x + \Delta x) - h(x)) \\ &= \nabla g(x)^T(t\Delta x) + t(h(x + \Delta x) - h(x)) + O(t^2), \end{aligned}$$

which proves (8).

Δx steps to the minimizer of h plus our quadratic approximation to g so $t\Delta x$ satisfies

$$\begin{aligned} &\nabla g(x)^T \Delta x + \frac{1}{2}\Delta x^T H \Delta x + h(x + \Delta x) \\ &\le \nabla g(x)^T(t\Delta x) + \frac{t^2}{2}\Delta x^T H \Delta x + h(x^+) \\ &\le t\nabla g(x)^T \Delta x + \frac{t^2}{2}\Delta x^T H \Delta x + th(x + \Delta x) + (1-t)h(x). \end{aligned}$$

We can rearrange and then simplify to obtain

$$(1-t)\nabla g(x)^T \Delta x + \frac{1}{2}(1-t^2)\Delta x^T H \Delta x + (1-t)(h(x + \Delta x) - h(x)) \le 0$$

$$\nabla g(x)^T \Delta x + \frac{1}{2}(1+t)\Delta x^T H \Delta x + h(x + \Delta x) - h(x) \le 0$$

$$\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \le \frac{1}{2}(1+t)\Delta x^T H \Delta x.$$

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

$$\begin{aligned}
&f(x^+) - f(x) = g(x^+) - g(x) + h(x^+) - h(x) \\
&\le \int_0^1 \nabla g(x + s(t\Delta x))^T (t\Delta x) ds + t h(x+\Delta x) + (1-t)h(x) - h(x) \\
&= \nabla g(x)^T (t\Delta x) + t(h(x+\Delta x) - h(x)) \\
&\quad + \int_0^1 (\nabla g(x + s(t\Delta x)) - \nabla g(x))^T (t\Delta x) ds \\
&\le t \big(\nabla g(x)^T (t\Delta x) + h(x+\Delta x) - h(x) \\
&\quad + \int_0^1 \|\nabla g(x + s(\Delta x)) - \nabla g(x)\| \|\Delta x\| ds \bigg).
\end{aligned}$$

∇g is Lipschitz continuous so

$$\begin{aligned}
f(x^+) - f(x) &\le t\left(\nabla g(x)^T \Delta x + h(x+\Delta x) - h(x) + \frac{L_1 t^2}{2}\|\Delta x\|^2 \right) \\
&= t\left(\Delta + \frac{L_1 t}{2}\|\Delta x\|^2 \right).
\end{aligned} \tag{18}$$

If we choose $t \le \frac{2m}{L_1}(1-\alpha)$, then

$$\frac{L_1 t}{2}\|\Delta x\|^2 \le m(1-\alpha)\|\Delta x\|^2 \le (1-\alpha)\Delta x^T H \Delta x \le -(1-\alpha)\Delta. \tag{19}$$

We can substitute (19) into (18) to obtain

$$f(x^+) - f(x) \le t\left(\Delta - (1-\alpha)\Delta\right) = t(\alpha\Delta).$$

□

A.3 Proof of Theorem 3.2

Proof. $\{f(x_k)\}$ is a nonincreasing sequence because because Δx 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 Δ_k must decay to zero. Δ_k satisfies

$$\begin{aligned}
\Delta_k &= \nabla g(x_k)^T \Delta x_k + h(x_k + \Delta x_k) - h(x_k) \\
&\le -\Delta x_k^T H_k \Delta x_k \le -m\|\Delta x_k\|^2,
\end{aligned}$$

where first inequality follows from (9). We reverse this inequality to obtain

$$\|\Delta x_k\|^2 \le \frac{1}{m}\Delta x_k^T H_k \Delta x_k \le -\frac{1}{m}\Delta_k$$

so the search directions Δx_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 ∂h is monotone. H is a symmetric, positive definite matrix so we have

$$\begin{aligned}
(\partial h(x) - \partial h(y))^T (x-y) &\ge 0 \\
(x-y)^T H (x-y) &\ge m\|x-y\|^2.
\end{aligned}$$

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

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

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

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

□

A.5 Proof of Theorem 3.5

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

$$x_{k+1} = \operatorname{prox}_h^{H_k}\left(x_k - {H_k}^{-1}\nabla g(x_k)\right).$$

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

$$\begin{aligned} \|x_{k+1} - x^\star\| &\leq \|R_k \circ S_k(x_k) - R_k \circ S_k(x^\star)\| \\ &\leq \|S_k(x_k) - S_k(x^\star)\| \\ &\leq \|S_k(x_k) - S_k(x^\star) - \nabla S_k(x^\star)(x_k - x^\star)\| \\ &\quad + \|\nabla S_k(x^\star)(x_k - x^\star)\|. \end{aligned} \tag{20}$$

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

$$\begin{aligned} \|\nabla S_k(x^\star)(x_k - x^\star)\| &\leq \frac{1}{m}\|\nabla^2 g(x_k) - \nabla^2 g(x^\star)\|\|x_k - x^\star\| \\ &\leq \frac{L_2}{m}\|x_k - x^\star\|^2. \end{aligned} \tag{21}$$

$\{x_k\} \to x^\star$ and ∇g is continuous, so for k sufficiently large,

$$\begin{aligned} &\|S_k(x_k) - S_k(x^\star) - \nabla S_k(x^\star)(x_k - x^\star)\| \\ &= \left\|\int_0^1 \left(\nabla S_k(x^\star + t(x_k - x^\star)) - \nabla S_k(x^\star)\right)(x_k - x^\star)dt\right\| \\ &\leq \int_0^1 \|(\nabla S_k(x^\star + t(x_k - x^\star)) - \nabla S_k(x^\star))\|\, \|x_k - x^\star\| dt \\ &\leq \int_0^1 \frac{1}{m}\left\|\nabla^2 g(x^\star) - \nabla^2 g(x^\star + t(x_k - x^\star))\right\|\, \|x_k - x^\star\| dt \\ &\leq \int_0^1 \frac{L_2}{m} t \|x_k - x^\star\|^2 dt \leq \frac{L_2}{2m}\|x_k - x^\star\|^2. \end{aligned} \tag{22}$$

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

$$\|x_{k+1} - x^\star\| \leq \frac{3L_2}{2m}\|x_k - x^\star\|^2.$$

□

A.6 Proof of Lemma 3.6

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

$$g(x + t\Delta x) \leq g(x) + t\nabla g(x)^T\Delta x + \frac{1}{2}t^2\Delta x^T\nabla^2 g(x)\Delta x + \frac{1}{6}L_2 t^3\|\Delta x\|^3.$$

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

$$f(x + \Delta x) \leq f(x) + \Delta + \frac{1}{2}\Delta x^T(\nabla^2 g(x) - H)\Delta x + \frac{1}{2}\Delta x^T H \Delta x + \frac{1}{6} L_2 \|\Delta x\|^3. \tag{23}$$

$\frac{1}{2}\Delta x^T(\nabla^2 g(x) - H)\Delta x$ can be split into two terms that can be bounded using the Lipschitz continuity of $\nabla^2 g$ and the Dennis-Moré criterion:

$$\begin{aligned}
&\frac{1}{2}\Delta x^T(\nabla^2 g(x) - H)\Delta x \\
&= \frac{1}{2}\Delta x^T(\nabla^2 g(x) - \nabla^2 g(x^\star))\Delta x + \frac{1}{2}\Delta x^T(\nabla^2 g(x^\star) - H)\Delta x \\
&\leq L_2\|x - x^\star\|\|\Delta x\|^2 \; + \frac{1}{2}\|\Delta x\| \left\|\left(\nabla^2 g(x^\star) - H\right)\Delta x\right\| \\
&= o\left(\|\Delta x\|^2\right) + o\left(\|\Delta x\|^2\right).
\end{aligned}$$

$\Delta x^T H \Delta x$ can also be bounded using $\Delta x^T H \Delta x \leq -\Delta$. We substitute these expressions into (23) and rearrange to obtain

$$\begin{aligned}
f(x + \Delta x) - f(x) &\leq \Delta - \frac{1}{2}\Delta + \frac{1}{2}\Delta x^T(\nabla^2 g(x) - H)\Delta x + \frac{1}{6} L_2\|\Delta x\|^2\|\Delta x\| \\
&\leq \frac{1}{2}\Delta + \frac{1}{6}L_2\|\Delta x\|^2\Delta + o\left(\|\Delta x\|^2\right).
\end{aligned}$$

$\{\Delta x_k\} \to 0$ (see the proof of Theorem 3.2) so $f(x_k + \Delta x_k) - f(x_k) \leq \frac{1}{2}\Delta_k$ after sufficiently many iterations and thus the unit step length shall eventually satisfy the sufficient descent condition. □

A.7 Proof of Lemma 3.7

Proof. Δx and $\Delta \hat{x}$ are the solutions to their respective subproblems so they are also the solutions to

$$\begin{aligned}
\Delta x &= \arg\min_d \; \nabla g(x)^T d + \Delta x^T H d + h(x + d), \\
\Delta \hat{x} &= \arg\min_d \; \nabla g(x)^T d + \Delta \hat{x}^T \hat{H} d + h(x + d).
\end{aligned}$$

Hence Δx and $\Delta \hat{x}$ satisfy

$$\begin{aligned}
&\nabla g(x)^T \Delta x + \Delta x^T H \Delta x + h(x + \Delta x) \\
&\leq \nabla g(x)^T \Delta \hat{x} + \Delta \hat{x}^T H \Delta \hat{x} + h(x + \Delta \hat{x})
\end{aligned}$$

and

$$\begin{aligned}
&\nabla g(x)^T \Delta \hat{x} + \Delta \hat{x}^T \hat{H} \Delta \hat{x} + h(x + \Delta \hat{x}) \\
&\leq \nabla g(x)^T \Delta x + \Delta x^T \hat{H} \Delta x + h(x + \Delta x).
\end{aligned}$$

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

$$\begin{aligned}
&\Delta x^T H \Delta x - 2\Delta x^T H \Delta \hat{x} + \Delta \hat{x}^T H \Delta \hat{x} \\
&\leq \Delta x^T(\hat{H} - H)\Delta \hat{x} + \Delta \hat{x}^T(H - \hat{H})\Delta \hat{x}.
\end{aligned}$$

The left hand side is $\|\Delta x - \Delta\hat{x}\|_H^2$ and the eigenvalues of H are bounded so

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

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

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

We denote this constant using c and conclude that

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

We substitute this inequality into (24) to obtain

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

□

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

$$x_{k+1} = \operatorname{prox}_h^{H_k} \left(x_k - \nabla^2 g(x_k)^{-1} \nabla g(x_k) \right).$$

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

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

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

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

The second term $\|\Delta x_k - \Delta x_k^{nt}\| = O\left(\left\| (\nabla^2 g(x_k) - H_k)\Delta x_k \right\|^{1/2} \|\Delta x_k\|^{1/2} \right) \|$ (Lemma 3.7). We can show that $\left\| \left(\nabla^2 g(x_k) - H_k\right) \Delta x_k \right\| = o(\|\Delta x_k\|)$:

$$\begin{aligned} &\left\| \left(\nabla^2 g(x_k) - H_k\right) \Delta x_k \right\| \\ &\leq \left\| \left(\nabla^2 g(x_k) - \nabla^2 g(x^\star)\right) \Delta x_k \right\| + \left\| \left(\nabla^2 g(x^\star) - H_k\right) \Delta x_k \right\| \\ &\leq L_2 \|x_k - x^\star\| \|\Delta x_k\| + o(\|\Delta x_k\|). \end{aligned}$$

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

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

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