Shashank Singh[1]

4 Convergence rate of generalized gradient descent [15 points] (Adona)

For convenience, we use the notation $\langle \cdot, \cdot \rangle$ to denote the dot product in several expressions.

(a) A second-order Taylor approximation gives, for some $\xi \in \mathbb{R}^n$,

$$g(y) \leq g(x) + (\nabla g(x))^T (y - x) + \frac{1}{2}(y - x)^T (\nabla^2 f(\xi))(y - x).$$

Since all directional derivatives of ∇g are bounded in magnitude by L (this is immediate from the definition of directional derivative) and $\|(\nabla^2 f(\xi))(y-x)\|$ is the magnitude of the derivative of ∇g at ξ in the direction $y - x$, $\|(\nabla^2 f(\xi))(y - x)\| \leq L\|y - x\|$. Thus, by Cauchy-Schwartz,

$$(y - x)^T (\nabla^2 f(\xi))(y - x) \leq \|(y - x)\|_2 \|(\nabla^2 f(\xi))(y - x)\|_2 \leq L\|(y - x)\|_2^2.$$

Then, since $f = g + h$,

$$f(y) \leq g(x) + (\nabla g(x)^T)(y - x) + \frac{L}{2}\|y - x\|_2^2 + h(y). \quad \blacksquare \tag{1}$$

(b) Substituting $y = x^+ = x - tG_t(x)$ into (1) gives

$$\begin{aligned} f(x^+) &\leq g(x) + (\nabla g(x)^T)(x - tG_t(x) - x) + \frac{L}{2}\|x - tG_t(x) - x\|_2^2 + h(x - tG_t(x)). \\ &= g(x) - t(\nabla g(x)^T)G_t(x) + \frac{Lt^2}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)). \\ &\leq g(x) - t\langle \nabla g(x), G_t(x)\rangle + \frac{t}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)), \end{aligned} \tag{2}$$

where the last inequality follows by bounding a factor of t by $1/L$. $\blacksquare$

(c) From the definitions of G_t and prox_t, we have

$$x - tG_t(x) = \text{prox}_t(x - t\nabla g(x)) = \operatorname*{argmin}_{z \in \mathbb{R}^n} \frac{1}{2t}\|x - t\nabla g(x) - z\|_2^2 + h(z).$$

The zero subgradient characterization of optimality and definition of argmin then imply

$$\begin{aligned} 0 &\in \partial \frac{1}{2t}\|x - t\nabla g(x) - (x - tG_t(x))\|_2^2 + h(x - tG_t(x)) \\ &= \partial \frac{1}{2}\|G_t(x) - \nabla g(x)\|_2^2 + \partial h(x - tG_t(x)) \\ &= \{-(G_t(x) - \nabla g(x)\} + \partial h(x - tG_t(x)), \end{aligned}$$

and hence $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$. $\blacksquare$

[1]sss1@andrew.cmu.edu

(d) Since $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$,

$$
\begin{aligned}
h(x - tG_x(x)) &\le h(z) - \langle G_t(x) - \nabla g(x), z - (x - tG_t(x)) \rangle \\
&= h(z) + \langle G_t(x), x - z \rangle - t\|G_t(x)\|_2^2 + \langle \nabla g(x), x - z \rangle + t\langle \nabla g(x), G_t(x) \rangle \\
&\le h(z) + \langle G_t(x), x - z \rangle - t\|G_t(x)\|_2^2 + g(z) - g(x) + t\langle \nabla g(x), G_t(x) \rangle,
\end{aligned}
$$

by bilinearity of the inner product and convexity of g (since $\langle \nabla g, x - z \rangle \le g(z) - g(x)$). Substituting this bound for $h(x - tG_x(x))$ in (2) and observing that several terms cancel gives

$$
\begin{aligned}
f(x^+) &\le h(z) + g(z) + \langle G_t(x), x - z \rangle - \frac{t}{2}\|G_t(x)\|_2^2 \\
&\le f(z) + \langle G_t(x), x - z \rangle - \frac{t}{2}\|G_t(x)\|_2^2 \qquad (3)
\end{aligned}
$$

since $f = g + h$. ■

(e) Plugging $z = x$ into (3) gives

$$
f(x^+) \le f(x) + \langle G_t(x), x - x \rangle - \frac{t}{2}\|G_t(x)\|_2^2 \le f(x)
$$

(and the latter inequality is strict if and only if $G_t(x) \neq 0$), so that generalized gradient descent does indeed decrease the criterion f in each iteration. Plugging $z = x^\star$ into (3) gives

$$
f(x^+) \le f(x^\star) + \langle G_t(x), x - x^\star \rangle - \frac{t}{2}\|G_t(x)\|_2^2.
$$

Substituting $G_t(x) = \frac{x - x^+}{t}$ and simplifying gives the desired result:

$$
f(x^+) \le f(x^\star) + \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2 \right). \quad ■
$$

(f) Since each $f(x^{(k)}) \le f(x^{(k-1)})$, by the result of part (e)

$$
\begin{aligned}
k(f(x^{(k)}) - f(x^\star)) = \sum_{i=1}^{k} f(x^{(k)}) - f(x^\star) &\le \sum_{i=1}^{k} f(x^{(i)}) - f(x^\star) \\
&\le \sum_{i=1}^{k} \frac{1}{2t}\left(\|x^{(k-1)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2 \right) \\
&\le \frac{\|x^{(0)} - x^\star\|_2^2}{2t},
\end{aligned}
$$

since the last sum telescopes. Dividing by k gives the desired result:

$$
f(x^{(k)}) - f(x^\star) \le \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}. \quad ■
$$