Mixed Optimization for Smooth Functions

Mehrdad Mahdavi  Lijun Zhang  Rong Jin
Department of Computer Science and Engineering, Michigan State University, MI, USA
{mahdavim,zhanglij,rongjin}@msu.edu

Before proving the lemmas we recall the definition of $F(w)$, $F'(w)$, $g$, and $\tilde{g}_i(w)$ as:

$$F(w) = \frac{\lambda}{2} \|w\|^2 + \lambda \langle w, \bar{w} \rangle + \frac{1}{n} \sum_{i=1}^{n} g_i(w + \bar{w}),$$

$$F'(w) = \frac{\lambda}{2\gamma} \|w\|^2 + \frac{\lambda}{\gamma} \langle w, \bar{w}' \rangle + \frac{1}{n} \sum_{i=1}^{n} g_i(w + \bar{w}'),$$

$$g = \lambda \bar{w} + \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\bar{w}),$$

$$\tilde{g}_i(w) = g_i(w + \bar{w}) - \langle w, \nabla g_i(\bar{w}) \rangle.$$  

We also recall that $\bar{w}_s$ and $\bar{w}_s'$ are the optimal solutions that minimize $F(w)$ and $F'(w)$ over the domain $W_k$ and $W_{k+1}$, respectively.

**Lemma 1.**

$$F(w_t) - F(\bar{w}_s) \leq \frac{\|w_t - \bar{w}_s\|^2}{2\eta} - \frac{\|w_{t+1} - \bar{w}_s\|^2}{2\eta} + \frac{\gamma}{2} \|\nabla \tilde{g}_i(w_t) + \lambda \bar{w}_t\|^2 + \langle g, w_t - w_{t+1} \rangle + \langle \nabla \tilde{g}_i(\bar{w}_s) - \nabla \tilde{F}(\bar{w}_s), w_t - \bar{w}_s \rangle + \langle -\nabla \tilde{g}_i(w_t) + \nabla \tilde{g}_i(\bar{w}_s) - \nabla \tilde{F}(\bar{w}_s) + \nabla \tilde{F}(w_t), w_t - \bar{w}_s \rangle$$

**Proof.** For each iteration $t$ in the $k$th epoch, from the strong convexity of $F(w)$ we have

$$F(w_t) - F(\bar{w}_s) \leq \langle \nabla F(w_t), w_t - \bar{w}_s \rangle - \frac{\lambda}{2} \|w_t - \bar{w}_s\|^2$$

$$= \langle g + \nabla \tilde{g}_i(w_t) + \lambda \bar{w}_t, w_t - \bar{w}_s \rangle + \langle -\nabla \tilde{g}_i(w_t) + \nabla \tilde{F}(w_t), w_t - \bar{w}_s \rangle - \frac{\lambda}{2} \|w_t - \bar{w}_s\|^2,$$

where $\tilde{F}(w) = \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_i(w)$. We now try to upper bound the first term in the right hand side. Since

$$\langle g + \nabla \tilde{g}_i(w_t) + \lambda \bar{w}_t, w_t - \bar{w}_s \rangle = \langle g + \nabla \tilde{g}_i(w_t) + \lambda \bar{w}_t, w_t - \bar{w}_s \rangle - \frac{\|w_t - \bar{w}_s\|^2}{2\eta} + \frac{\|w_t - \bar{w}_s\|^2}{2\eta}$$

$$\leq \langle g + \nabla \tilde{g}_i(w_t) + \lambda \bar{w}_t, w_t - w_{t+1} \rangle - \frac{\|w_t - w_{t+1}\|^2}{2\eta} - \frac{\|w_{t+1} - \bar{w}_s\|^2}{2\eta} + \frac{\|w_t - \bar{w}_s\|^2}{2\eta}$$

$$\leq \langle g, w_t - w_{t+1} \rangle - \frac{\|w_{t+1} - \bar{w}_s\|^2}{2\eta} + \frac{\|w_t - \bar{w}_s\|^2}{2\eta} + \frac{\|w_t - \bar{w}_s\|^2}{2\eta}$$

$$= \langle g, w_t - w_{t+1} \rangle - \frac{\|w_{t+1} - \bar{w}_s\|^2}{2\eta} + \frac{\|w_t - \bar{w}_s\|^2}{2\eta} + \frac{\|w_t - \bar{w}_s\|^2}{2\eta}$$
We bound (Bernstein’s inequality for martingales). Let

With a probability \( \Pr \left[ |w - \hat{w}| \leq \Delta \right] \),

The proof is based on the Berstein inequality for Martingales [1] which is restated here for completeness.

**Theorem 1.** (Bernstein’s inequality for martingales). Let \( X_1, \ldots, X_n \) be a bounded martingale difference sequence with respect to the filtration \( \mathcal{F} = (\mathcal{F}_t)_{t \leq n} \) and with \( \|X_t\| \leq K \). Let

\[
S_t = \sum_{j=1}^{i} X_j
\]

be the associated martingale. Denote the sum of the conditional variances by

\[
\Sigma_n^2 = \sum_{i=1}^{n} \mathbb{E} \left[ X_i^2 | \mathcal{F}_{t-1} \right],
\]

Then for all constants \( t, \nu > 0 \),

\[
\Pr \left[ \max_{i=1, \ldots, n} S_i > t \text{ and } \Sigma_n^2 \leq \nu \right] \leq \exp \left( -\frac{t^2}{2(\nu + Kt/3)} \right),
\]

and therefore,

\[
\Pr \left[ \max_{i=1, \ldots, n} S_i > \sqrt{2\nu t} + \frac{\sqrt{2} K t}{3} \text{ and } \Sigma_n^2 \leq \nu \right] \leq e^{-t}.
\]
Equipped with this theorem, we are now in a position to upper bound $B_T$ and $C_T$ as follows.

**Proof.** (of Lemma 3) Denote $X_t = \langle \nabla g_i(\w_t) - \nabla \hat{F}(\w_t), \w_t - \w_* \rangle$. We have that the conditional expectation of $X_t$, given randomness in previous rounds, is $E_{t-1}[X_t] = 0$. We now apply Theorem 1 to the sum of martingale differences. In particular, we have, with a probability $1 - e^{-t}$,

$$B_T \leq \frac{\sqrt{2}}{3} K t + \sqrt{2 \Sigma t}$$

where

$$K = \max_{1 \leq t \leq T} (\nabla g_i(\w_*) - \nabla \hat{F}(\w_*), \w_t - \w_*) \leq 2 \beta \Delta^2$$

$$\Sigma = \sum_{t=1}^T E_t \left[ (\nabla g_i(\w_*), \w_t - \w_*)^2 \right] \leq \beta^2 \Delta^4 T$$

Hence, with a probability $1 - \delta$, we have

$$B_T \leq \beta \Delta^2 \left( \ln \frac{1}{\delta} + \sqrt{2 T \ln \frac{1}{\delta}} \right)$$

Similar, for $C_T$, we have, with a probability $1 - \delta$,

$$C_T \leq 2 \beta \Delta^2 \left( \ln \frac{1}{\delta} + \sqrt{2 T \ln \frac{1}{\delta}} \right)$$

\[\square\]

**Lemma 4.** $\|\w_* - \w\| \leq \gamma \|\w - \w_*\|$.  

**Proof.** We rewrite $F(w)$ as

$$F(w) = \frac{\lambda}{2} \|w\|^2 + \lambda \langle w, \w \rangle + \frac{1}{n} \sum_{i=1}^n g_i(w + \w)$$

$$= \frac{\lambda}{2} \|w - \w + \w\|^2 + \lambda \langle w - \w + \w, \w \rangle + \frac{1}{n} \sum_{i=1}^n g_i(w - \w + \w')$$

Define $z = w - \w$. We have

$$F(w) = \frac{\lambda}{2} \|z + \w\|^2 + \lambda \langle z, \w \rangle + \frac{1}{n} \sum_{i=1}^n g_i(z + \w')$$

$$= \frac{\lambda}{2} \|z\|^2 + \lambda \langle z, \w' \rangle + \frac{1}{n} \sum_{i=1}^n g_i(z + \w') + \frac{\lambda}{2} \|\w\|^2 + \lambda \langle \w, \w \rangle$$

$$= \bar{F}(z) + \frac{\lambda}{2} \|\w\|^2 + \lambda \langle \w, \w \rangle$$

where

$$\bar{F}(z) = \frac{\lambda}{2} \|z\|^2 + \lambda \langle z, \w' \rangle + \frac{1}{n} \sum_{i=1}^n g_i(z + \w')$$

Define $\w_* = \w_* - \w$. Evidently, $\w_*$ minimizes $\bar{F}(w)$. The only difference between $\bar{F}(w)$ and $F'(w)$ is that they use different modulus of strong convexity $\lambda$. Thus, following [2], we have

$$\|\w_* - \w_*\| \leq \frac{1 - \gamma^{-1}}{\gamma^{-1}} \|\w_*\| \leq (\gamma - 1) \|\w_*\|$$

Hence,

$$\|\w_*\| \leq \gamma \|\w_*\| = \gamma \|\w_* - \w\|$$

which completes the proofs. \[\square\]
References
