1 Restricted Strong Convexity

Let $\mathbf{w}$ be a vector and $\ell$ be a loss function. In general, $\ell_1$-norm regularized loss minimization can be written as follows for some $\lambda > 0$:

$$
\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^p} \ell(\mathbf{w}) + \lambda \|\mathbf{w}\|_1
$$

(1)

We will also assume that there is an unknown but fixed $\mathbf{w}^* \in \mathbb{R}^p$. Our goal will be to recover a vector $\hat{\mathbf{w}}$ which is close to $\mathbf{w}^*$.

Next, we define restricted strong convexity [1].

**Definition 8.1.** Let $\alpha > 0$, $\tau \geq 0$ and $g: \mathbb{N} \times \mathbb{N} \to \mathbb{R}^+$. A loss function $\ell : \mathbb{R}^p \to \mathbb{R}$ is restricted strongly convex around $\mathbf{w}^*$ with parameters $(\alpha, \tau, g)$ if and only if:

$$
(\forall \mathbf{w} \in \mathbb{R}^p) \, \ell(\mathbf{w}) - \ell(\mathbf{w}^*) - \langle \nabla \ell(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle \geq \alpha \|\mathbf{w} - \mathbf{w}^*\|_2^2 - \tau g(n,p) \|\mathbf{w} - \mathbf{w}^*\|_1^2
$$

For many specific learning problems $g(n,p) = \sqrt{\frac{\log p}{n}}$ or $g(n,p) = \frac{\log p}{n}$. In what follows, we analyze the sufficient number of samples for the problem in eq.(1).

**Theorem 8.1.** Assume that the convex loss function $\ell : \mathbb{R}^p \to \mathbb{R}$ is restricted strongly convex around $\mathbf{w}^*$ with parameters $(\alpha, \tau, g)$ as in Definition 8.1. Let $k$ be the number of nonzero elements in $\mathbf{w}^*$. For a regularization weight $\lambda \geq \frac{2 \|\nabla \ell(\mathbf{w}^*)\|_\infty}{\alpha}$ and a sufficient number of samples $\frac{17}{\alpha}kg(n,p) \leq 1$, we have that:

$$
\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq 102 \frac{\sqrt{k}}{\alpha} \lambda
$$

$$
\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 \leq 408 \frac{k}{\alpha} \lambda
$$

For many specific learning problems $\lambda \in \mathcal{O}(\sqrt{\frac{\log p}{n}})$ and thus, the above theorem establishes consistency as the number of samples $n$ grows.

First, we derive an intermediate lemma needed for the final proof.
Lemma 8.1. Assume that the loss function $\ell : \mathbb{R}^p \rightarrow \mathbb{R}$ is convex. Let $k$ be the number of nonzero elements in $w^*$. For a regularization weight $\lambda \geq 2\|\nabla \ell(w^*)\|_\infty$, we have:

$$\|\hat{w} - w^*\|_1 \leq 4\sqrt{k}\|\hat{w} - w^*\|_2$$

Proof. Let $\Delta \equiv \hat{w} - w^*$. Let $K$ be the set of nonzero elements of $w^*$ and let $K^c$ be the complement of $K$. Note that $k \equiv |K|$ is the number of nonzero elements in $w^*$. For an arbitrary vector $w$, let $w_K$ denote the original vector $w$ with zeros on the entries in $K$ and $w_K^c$ denote the original vector $w$ with zeros on the entries in $K$.

Since by definition $w^* = w_K^*$ and by the reverse triangle inequality, we have:

$$\|\hat{w}\|_1 = \|w^* + \Delta\|_1 = \|w_K^* + \Delta_K + \Delta_{K^c}\|_1 = \|w_K^* + \Delta_K\|_1 + \|\Delta_{K^c}\|_1 \geq \|w_K^*\|_1 - \|\Delta_K\|_1 + \|\Delta_{K^c}\|_1 = \|w^*\|_1 - \|\Delta_K\|_1 + \|\Delta_{K^c}\|_1$$  \hspace{1cm} (2)

By optimality of $\hat{w}$ in eq.(1), we have:

$$\ell(\hat{w}) + \lambda\|\hat{w}\|_1 \leq \ell(w^*) + \lambda\|w^*\|_1$$

and therefore:

$$\ell(\hat{w}) - \ell(w^*) \leq -\lambda\|\hat{w}\|_1 + \lambda\|w^*\|_1$$ \hspace{1cm} (3)

By convexity of $\ell$, the Cauchy-Schwarz inequality $(\forall a, b) |\langle a, b \rangle| \leq \|a\|_1 \|b\|_\infty$, and since we assume that $\lambda \geq 2\|\nabla \ell(w^*)\|_\infty$, we have:

$$\ell(\hat{w}) - \ell(w^*) \geq \langle \nabla \ell(w^*), \Delta \rangle \geq -\|\nabla \ell(w^*)\|_\infty \|\Delta\|_1 \geq -\frac{1}{2}\lambda\|\Delta\|_1$$ \hspace{1cm} (4)

By eq.(3) and eq.(4), it follows that $-\frac{1}{2}\lambda\|\Delta\|_1 \leq -\lambda\|\hat{w}\|_1 + \lambda\|w^*\|_1$ or equivalently since $\lambda > 0$:

$$0 \geq -\|\Delta\|_1 + 2\|\hat{w}\|_1 - 2\|w^*\|_1 \geq -\|\Delta\|_1 + 2\|w^*\|_1 - 2\|\Delta_K\|_1 + 2\|\Delta_{K^c}\|_1 - 2\|w^*\|_1 = -\|\Delta\|_1 + 2\|\Delta_K\|_1 + 2\|\Delta_{K^c}\|_1 = -\|\Delta_K\|_1 - \|\Delta_{K^c}\|_1 - 2\|\Delta_K\|_1 + 2\|\Delta_{K^c}\|_1 = -3\|\Delta_K\|_1 + \|\Delta_{K^c}\|_1$$

2
where the second line follows from eq.(2). Given the above, we have:
\[ \|\Delta\|_1 = \|\Delta_K\|_1 + \|\Delta_{K^c}\|_1 \]
\[ \leq \|\Delta_K\|_1 + 3\|\Delta_K\|_1 \]
\[ = 4\|\Delta_K\|_1 \]
\[ \leq 4\sqrt{k}\|\Delta\|_2 \]
\[ \leq 4\sqrt{k}\|\Delta\|_2 \]

which proves our claim. \(\square\)

Next, we provide the final proof.

**Proof of Theorem 8.1.** Let \(\Delta \equiv \hat{w} - w^*\). First, since we assume that \(\lambda \geq 2\|\nabla \ell(w^*)\|_\infty\) we can invoke Lemma 8.1 and therefore:
\[ \|\Delta\|_1 \leq 4\sqrt{k}\|\Delta\|_2 \]
(6)

For \(w = \hat{w}\), by restricted strong convexity of \(\ell\) around \(w^*\) with parameters \((\alpha, \tau, g)\) as in Definition 8.1, by eq.(6) and since 17, \[\frac{\alpha}{\tau^2}kg(n, p)\leq 1\], we have:
\[ \ell(\hat{w}) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \geq 0\|\Delta\|_2^2 - \tau g(n, p)\|\Delta\|_1^2 \]
\[ \geq (\alpha - 16k\tau g(n, p))\|\Delta\|_2^2 \]
\[ \geq (\alpha - \frac{16k}{\tau^2}\alpha)\|\Delta\|_2^2 \]
\[ = \frac{1}{17}\alpha\|\Delta\|_2^2 \]
(7)

By eq.(6) and eq.(7), the Cauchy-Schwarz inequality \((\forall a, b) |\langle a, b \rangle| \leq \|a\|_1 \|b\|_\infty\), and since we assume that \(\lambda \geq 2\|\nabla \ell(w^*)\|_\infty\), we have:
\[ \frac{1}{17}\alpha\|\Delta\|_2^2 \leq \ell(\hat{w}) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \]
\[ \leq -\lambda\|\hat{w}\|_1 + \lambda\|w^*\|_1 + \|\nabla \ell(w^*)\|_\infty\|\Delta\|_1 \]
\[ \leq -\lambda\|\hat{w}\|_1 + \lambda\|w^*\|_1 + \frac{1}{2}\lambda\|\Delta\|_1 \]
\[ \leq -\lambda\|\hat{w}\|_1 + \lambda\|w^*\|_1 + \lambda\|\Delta\|_1 + \frac{1}{2}\lambda\|\Delta\|_1 \]
\[ = \frac{3}{2}\lambda\|\Delta\|_1 \]
\[ \leq 6\sqrt{k}\lambda\|\Delta\|_2 \]
which proves the first claim after canceling \(\|\Delta\|_2\) on both sides of the inequality. The second claim can be proven by the above and eq.(6). \(\square\)

2 **Application: Compressed Sensing**

Assume that there is an unknown but fixed \(w^* \in \mathbb{R}^p\). The only way to access \(w^*\) is through a **black box** that works as follows. We (somehow) generate an **input** vector \(x_i \in \{-1, +1\}^p\) and the black box returns an **output**:
\[ y_i = \langle x_i, w^* \rangle + \varepsilon_i \]
where \( \varepsilon_i \in \{-1, +1\} \) is a Rademacher random variable (see Definition 6.1). In the above, we know that \( y_i \) is equal to \( \langle x_i, w^* \rangle + \varepsilon_i \), but we do not have access to \( w^* \) or \( \varepsilon_i \). We only have access to the output \( y_i \), and of course the input \( x_i \).

The question is how many pairs \((x_i, y_i)\) are sufficient in order to recover a vector \( \hat{w} \) which is close to \( w^* \). Assume we obtain \( n \) pairs. Let \( X \in \{-1, +1\}^{n \times p} \), \( y \in \mathbb{R}^n \) and \( \varepsilon \in \{-1, +1\}^n \). Note that:

\[
    y = Xw^* + \varepsilon
\]

We solve eq.(1) by using the loss function:

\[
    \ell(w) = \frac{1}{2n} \| Xw - y \|^2
\]

For answering our question, we will have to show that the loss function \( \ell \) fulfills the conditions of Theorem 8.1. From eq.(8) and eq.(9), we have:

\[
    \ell(w) = \frac{1}{2n} \| (w - w^*) + \varepsilon \|^2
    = \frac{1}{2n} (w - w^*)^T X^T X (w - w^*) - \frac{1}{n} \varepsilon^T X (w - w^*) + \frac{1}{2n} \varepsilon^T \varepsilon
\]

By the above, we can conclude that:

\[
    \ell(w^*) = \frac{1}{2n} \varepsilon^T \varepsilon
\]

\[
    \nabla \ell(w) = \frac{1}{n} X^T X (w - w^*) - \frac{1}{n} X^T \varepsilon
\]

\[
    \nabla \ell(w^*) = -\frac{1}{n} X^T \varepsilon
\]

In what follows we will assume that each entry of \( X \) and \( \varepsilon \) is independent and Rademacher distributed.

**First Condition:** \( \lambda \geq 2 \| \nabla \ell(w^*) \|_{\infty} \). Assume that we set the regularization weight as follows \( \lambda = 4 \sqrt{\frac{p \log p}{n}} \). Let \( x^j \in \{-1, +1\}^n \) be the \( j \)-th column of \( X \). Fix \( j \). Note that \( \frac{1}{n} \langle x^j, \varepsilon \rangle = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \varepsilon_i = \frac{1}{n} \sum_{i=1}^{n} z_i \) where \( z_i \equiv x_{ij} \varepsilon_i \) for \( i = 1 \ldots n \) are independent random variables. Moreover, \( z_i \in [-1, +1] \) and \( E_{\varepsilon} \left[ \frac{1}{n} \langle x^j, \varepsilon \rangle \mid X \right] = 0 \) since \( E_{\varepsilon} \left[ \varepsilon \mid X \right] = 0 \). Thus, by Hoeffding’s inequality (Corollary 2.2) and the union bound:

\[
    P_{\varepsilon} \left[ \exists j = 1 \ldots p \mid \left| \frac{1}{n} \langle x^j, \varepsilon \rangle \right| \geq \frac{\lambda}{2} \right] \leq 2p \ e^{-\frac{n^2 \lambda^2 \ (1/2)^2}{8n^2}}
    = 2p \ e^{-\frac{n^2 \lambda^2}{8n}}
    = 2p \ e^{-2 \log p}
    = 2/p
\]
By eq.(12) and the above, we have:

\[ P_{X,\epsilon} \left[ \left\| \nabla \ell(w^*) \right\|_\infty \geq \frac{\lambda}{2} \right] = P_{X,\epsilon} \left[ \left\| -\frac{1}{n} X^T \epsilon \right\|_\infty \geq \frac{\lambda}{2} \right] = P_{X,\epsilon} \left[ \exists j = 1 \ldots p \left| \frac{1}{n} \langle x^j, \epsilon \rangle \right| \geq \frac{\lambda}{2} \right] = P_{\epsilon} \left[ \exists j = 1 \ldots p \left| \frac{1}{n} \langle x^j, \epsilon \rangle \right| \geq \frac{\lambda}{2} \right] \]

Therefore, with probability at least \(1 - 2/p\) over the choice of \(X\) and \(\epsilon\), we have that \(\lambda \geq 2\|\nabla \ell(w^*)\|_\infty\) when we use the regularization weight \(\lambda = 4\sqrt{\log p/n}\).

Second Condition: Restricted Strong Convexity. Theorem 8.1 requires that the loss function \(\ell\) in eq.(9) fulfills Definition 8.1. Here, we will show that indeed this is fulfilled. That is:

\[ (\forall w \in \mathbb{R}^p) \ell(w) - \ell(w^*) - \langle \nabla \ell(w^*), w - w^* \rangle \geq \alpha \|w - w^*\|_2^2 - \tau g(n, p) \|w - w^*\|_1^2 \]

Note that by eq.(10), eq.(11) and eq.(12), we have:

\[ \ell(w) - \ell(w^*) - \langle \nabla \ell(w^*), w - w^* \rangle = \frac{1}{2n} (w - w^*)^T X^T X (w - w^*) \]

Let \(v = w - w^*\). Our goal is to show that:

\[ (\forall v \in \mathbb{R}^p) \frac{1}{2n} \|Xv\|_2^2 \geq \alpha \|v\|_2^2 - \tau g(n, p) \|v\|_1^2 \]

Since the above is trivially fulfilled for \(v = 0\) and since if the above holds for some \(v \in \mathbb{R}^p\) then it also holds for \(cv\) for all \(c \in \mathbb{R}\), we will equivalently show that:

\[ (\forall \|v\|_1 = 1) \frac{1}{2n} \|Xv\|_2^2 \geq \alpha \|v\|_2^2 - \tau g(n, p) \]

Fix \(j \neq k\). Note that \(\frac{1}{n} \sum_{i=1}^n x_{ij} x_{ik} = \frac{1}{n} \sum_{i=1}^n z_i\) where \(z_i \equiv x_{ij} x_{ik}\) for \(i = 1 \ldots n\) are independent random variables. Moreover, \(z_i \in [-1, 1]\) and we also know that \(\mathbb{E}_X [\frac{1}{n} \sum_{i=1}^n x_{ij} x_{ik}] = 0\) since the entries of \(X\) are independent and zero-mean. Thus, by Hoeffding’s inequality (Corollary 2.2), the union bound and by
assuming \( t = \sqrt{\frac{6 \log p}{n}} \):

\[
P_X \left[ \max_{j \neq k} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right| \geq t \right] = P_X \left[ (\exists j \neq k) \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right| \geq t \right]
\]

\[
\leq 2 \left( \frac{p}{2} \right) e^{-\frac{2n^2t^2}{n}}
\]

\[
\leq p^2 e^{-\frac{n^2t^2}{2}}
\]

\[
= p^2 e^{-3 \log p}
\]

\[
= \frac{1}{p}
\]

Therefore, with probability at least \( 1 - \frac{1}{p} \) over the choice of \( X \), we have that:

\[
\max_{j \neq k} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right| \leq \sqrt{\frac{6 \log p}{n}} \tag{13}
\]

Note that since \( \|v\|_1 = 1 \), then:

\[
\sum_{j \neq k} |v_j v_k| \leq \sum_{j=1}^{p} \sum_{k=1}^{p} |v_j v_k|
\]

\[
= \sum_{j=1}^{p} \sum_{k=1}^{p} |v_j| |v_k|
\]

\[
= \|v\|_1^2
\]

\[
= 1 \tag{14}
\]

Since \( (\forall ij) \ x_{ij}^2 = 1 \) and the above, we have:

\[
\frac{1}{2n} \|Xv\|_2^2 = \frac{1}{2n} \sum_{i=1}^{n} (Xv)^2_i
\]

\[
= \frac{1}{2n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} x_{ij} v_j \right)^2
\]

\[
= \frac{1}{2n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} x_{ij}^2 v_j^2 + \sum_{j \neq k} x_{ij} x_{ik} v_j v_k \right)
\]

\[
= \frac{1}{2n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} v_j^2 + \sum_{j \neq k} x_{ij} x_{ik} v_j v_k \right)
\]

\[
= \frac{1}{2} \|v\|_2^2 + \frac{1}{2} \sum_{j \neq k} \left( \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right) v_j v_k
\]
\[ \begin{align*}
&\geq \frac{1}{2} \|v\|_2^2 - \frac{1}{2} \max_{j \neq k} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right| \sum_{j \neq k} |v_j v_k| \\
&\geq \frac{1}{2} \|v\|_2^2 - \frac{1}{2} \sqrt{\frac{6 \log p}{n}}
\end{align*} \]

where the previous-to-the-last step follows from the Cauchy-Schwarz inequality \((\forall a, b) \ | \langle a, b \rangle \| \leq \|a\|_1 \|b\|_\infty\). The last step follows from eq.(13) and eq.(14).

Note that given our brief introduction at the beginning of the proof, we have shown that:

\[ (\forall w \in \mathbb{R}^p) \ell(w) - \ell(w^*) - \langle \nabla \ell(w^*), w - w^* \rangle \geq \frac{1}{2} \|w - w^*\|_2^2 - \frac{1}{2} \sqrt{\frac{6 \log p}{n}} \|w - w^*\|_1^2 \]

Therefore, we conclude that the loss function \(\ell\) in eq.(9) fulfills Definition 8.1 with \(\alpha = 1/2\), \(\tau = \sqrt{6}/2\) and \(g(n,p) = \sqrt{\frac{\log p}{n}}\).

**Third Condition: Sufficient Number of Samples.** Theorem 8.1 also requires that

\[ 17^2 kg(n,p) \leq 1 \]

That is:

\[ \frac{17^2 k \sqrt{6} \log p}{n} \leq 1 \]

Thus, we require \(n \geq \frac{17^2 6k^2 \log p}{n}\).

A proof for possibly correlated Gaussian random variables can be found in [2] where they obtained results with \(g(n,p) = \frac{\log p}{n}\), which is better for the required number of samples in the Third Condition.

**References**
