1 Restricted Strong Convexity

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

$$\hat{w} = \arg \min_{w \in \mathbb{R}^p} \ell(w) + \lambda \|w\|_1 \quad (1)$$

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

Next, we define restricted strong convexity [1].

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

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

For many specific learning problems $g(n,p) = \sqrt{\log p/n}$ or $g(n,p) = \log p/n$. 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 $\ell : \mathbb{R}^p \rightarrow \mathbb{R}$ is restricted strongly convex around $w^*$ with parameters $(\alpha, \tau, g)$ as in Definition 8.1. Let $k$ be the number of nonzero elements in $w^*$. For a regularization weight $\lambda \geq 2\|\nabla \ell(w^*)\|_\infty$ and a sufficient number of samples $17\frac{\tau}{\alpha}kg(n,p) \leq 1$, we have that:

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

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

For many specific learning problems $\lambda \in O(\sqrt{\log p/n})$ 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 $\ell : \mathbb{R}^p \to \mathbb{R}$ is convex. Let $k$ be the number of nonzero elements in $\mathbf{w}^*$. For a regularization weight $\lambda \geq 2\|\nabla \ell(\mathbf{w}^*)\|_\infty$, we have:

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

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

Since by definition $\mathbf{w}^* = \mathbf{w}^*_\mathcal{K}$ and by the reverse triangle inequality, we have:

$$\|\hat{\mathbf{w}}\|_1 = \|\mathbf{w}^* + \Delta\|_1 = \|\mathbf{w}^*_\mathcal{K} + \Delta_\mathcal{K} + \Delta_{\mathcal{K}^c}\|_1 = \|\mathbf{w}^*_\mathcal{K} + \Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1 \geq \|\mathbf{w}^*_\mathcal{K}\|_1 - \|\Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1 = \|\mathbf{w}^*\|_1 - \|\Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1$$

(2)

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

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

and therefore:

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

(3)

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

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

(4)

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

$$0 \geq -\|\Delta\|_1 + 2\|\hat{\mathbf{w}}\|_1 - 2\|\mathbf{w}^*\|_1$$

$$\geq -\|\Delta\|_1 + 2\|\mathbf{w}^*\|_1 - 2\|\Delta_\mathcal{K}\|_1 + 2\|\Delta_{\mathcal{K}^c}\|_1 - 2\|\mathbf{w}^*\|_1$$

$$= -\|\Delta\|_1 - 2\|\Delta_\mathcal{K}\|_1 + 2\|\Delta_{\mathcal{K}^c}\|_1$$

$$= -\|\Delta_\mathcal{K}\|_1 - \|\Delta_{\mathcal{K}^c}\|_1 - 2\|\Delta_\mathcal{K}\|_1 - 2\|\Delta_{\mathcal{K}^c}\|_1$$

$$= -3\|\Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1$$
where the second line follows from eq.(2). Given the above, we have:

\[
\|\Delta\|_1 = \|\Delta_K\|_1 + \|\Delta_{K'}\|_1 \\
\leq \|\Delta_K\|_1 + 3\|\Delta_K\|_1 \\
= 4\|\Delta_K\|_1 \\
\leq 4\sqrt{k}\|\Delta\|_2 \\
\leq 4\sqrt{k}\|\Delta\|_2
\]

which proves our claim.

Next, we provide the final proof.

**Proof of Theorem 8.1.** Let \( \Delta = \hat{w} - w^* \). 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 \( 17\frac{\tau}{\alpha}kg(n, p) \leq 1 \), we have:

\[
\ell(\hat{w}) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \geq \alpha\|\Delta\|_2^2 - \tau g(n, p)\|\Delta\|_1^2 \\
\geq (\alpha - 16k\tau g(n, p))\|\Delta\|_2^2 \\
\geq (\alpha - \frac{16k\tau}{\alpha})\|\Delta\|_2^2 \\
= \frac{1}{17}\alpha\|\Delta\|_2^2
\]

(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 \( \|\Delta\|_2^2 \) 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 \( \epsilon \in \{-1, +1\}^n \). Note that:
\[
y = Xw^* + \epsilon \tag{8}
\]
We solve eq.(1) by using the loss function:
\[
\ell(w) = \frac{1}{2n} \| Xw - y \|^2 \tag{9}
\]
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:
\[
\ell(w) = \frac{1}{2n} \| X(w - w^*) - \epsilon \|^2 \\
= \frac{1}{2n} (w - w^*)^T X^T X (w - w^*) - \frac{1}{n} \epsilon^T X (w - w^*) + \frac{1}{2n} \epsilon^T \epsilon \tag{10}
\]
By the above, we can conclude that:
\[
\ell(w^*) = \frac{1}{2n} \epsilon^T \epsilon \tag{11}
\]
\[
\nabla \ell(w) = \frac{1}{n} X^T X (w - w^*) - \frac{1}{n} X^T \epsilon \tag{12}
\]
In what follows we will assume that each entry of \( X \) and \( \epsilon \) is independent and
Rademacher distributed.

**First Condition:** \( \lambda \geq 2 \| \nabla \ell(w^*) \|_{\infty} \). Assume that we set the regularization
weight as follows \( \lambda = \sqrt{\frac{\log p}{n}} \). Let \( x^j \in \{-1, +1\}^n \) be the \( j \)-th column of
\( X \). Fix \( j \). Note that \( \frac{1}{n} \langle x^j, \epsilon \rangle = \frac{1}{n} \sum_{i=1}^n x_{ij} \varepsilon_i = \frac{1}{n} \sum_{i=1}^n z_i \) where \( z_i \equiv x_{ij} \varepsilon_i \) for \( i = 1 \ldots n \) are independent random variables. Moreover, \( z_i \in [-1, +1] \) and
\( \mathbb{E}[\frac{1}{n} \langle x^j, \epsilon \rangle \mid X] = 0 \) since \( \mathbb{E}[\epsilon \mid X] = 0 \). Thus, by Hoeffding’s inequality (Corol-
mary 2.2) and the union bound:
\[
\mathbb{P}_\epsilon \left[ \exists j = 1 \ldots p \left| \frac{1}{n} \langle x^j, \epsilon \rangle \right| \geq \frac{\lambda}{2} \right] \leq 2 p \ e^{- \frac{2n^2 (\lambda/2)^2}{n^2}} \\
= 2 p \ e^{-2n\lambda^2} \\
= 2 p \ e^{-2\log p} \\
= \frac{2}{p}
\]
By eq.(12) and the above, we have:

$$
P_{X, \epsilon} \left[ \|\nabla \ell(w^*)\|_\infty \geq \frac{\lambda}{2} \right] = P_{X, \epsilon} \left[ \left\| \frac{1}{n} X^T \epsilon \right\|_\infty \geq \frac{\lambda}{2} \right]
$$

$$
= P_{X, \epsilon} \left[ \exists j = 1 \ldots p \mid \frac{1}{n} \langle x^j, \epsilon \rangle \geq \frac{\lambda}{2} \right]
$$

$$
= \mathbb{E}_X \left[ P_{\epsilon} \left[ \exists j = 1 \ldots p \mid \frac{1}{n} \langle x^j, \epsilon \rangle \geq \frac{\lambda}{2} \right] \right]
$$

$$
\leq \mathbb{E}_X \left[ \frac{2}{p} \right] = \frac{2}{p}
$$

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 $\lambda = \sqrt{\log p/n}$.

**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:

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

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

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

$$
= \frac{1}{2n} \|X(w - w^*)\|_2^2
$$

Let $v = w - w^*$. Our goal is to show that:

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

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:

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

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 \( t = \sqrt{\frac{6 \log p}{n}} \),

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

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

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

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

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

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:

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

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

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

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

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

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

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

**Third Condition: Sufficient Number of Samples.** Theorem 8.1 also requires that \(17 \frac{\tau}{\alpha} kg(n, p) \leq 1\). That is:
\[
17 \frac{\tau}{\alpha} kg(n, p) = 17 \sqrt{6} k \sqrt{\log \frac{p}{n}} \leq 1
\]

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

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**
