Hands-On Learning Theory
Fall 2016, Lecture 8

Jean Honorio jhonorio@purdue.edu

1 Restricted Strong Convexity

Let $\mathbf{w}$ be a vector and ℓ be a loss function. In general, ℓ_1-norm regularized loss minimization can be written as follows for some $\lambda > 0$:

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

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 $\widehat{\mathbf{w}}$ 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 (α, τ, 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 (α, τ, g) as in Definition 8.1. Let k be the number of nonzero elements in $\mathbf{w}^*$. For a regularization weight $\lambda \geq 2\|\nabla\ell(\mathbf{w}^*)\|_\infty$ and a sufficient number of samples $17\frac{\tau}{\alpha}kg(n,p) \leq 1$, we have that:*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq 102\frac{\sqrt{k}}{\alpha}\lambda$$
$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \leq 408\frac{k}{\alpha}\lambda$$

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* $\ell : \mathbb{R}^p \to \mathbb{R}$ *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:*

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

Proof. Let $\Delta \equiv \widehat{\mathbf{w}} - \mathbf{w}^*$. 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:

$$\begin{aligned}
\|\widehat{\mathbf{w}}\|_1 &= \|\mathbf{w}^* + \Delta\|_1 \\
&= \|\mathbf{w}^*_\mathcal{K} + \Delta_\mathcal{K} + \Delta_{\mathcal{K}^c}\|_1 \\
&= \|\mathbf{w}^*_\mathcal{K} + \Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1 \\
&\geq \|\mathbf{w}^*_\mathcal{K}\|_1 - \|\Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1 \\
&= \|\mathbf{w}^*\|_1 - \|\Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1 \qquad (2)
\end{aligned}$$

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

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

and therefore:

$$\ell(\widehat{\mathbf{w}}) - \ell(\mathbf{w}^*) \leq -\lambda\|\widehat{\mathbf{w}}\|_1 + \lambda\|\mathbf{w}^*\|_1 \qquad (3)$$

By convexity of ℓ, 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:

$$\begin{aligned}
\ell(\widehat{\mathbf{w}}) - \ell(\mathbf{w}^*) &\geq \langle \nabla\ell(\mathbf{w}^*), \Delta\rangle \\
&\geq -\|\nabla\ell(\mathbf{w}^*)\|_\infty\|\Delta\|_1 \\
&\geq -\tfrac{1}{2}\lambda\|\Delta\|_1 \qquad (4)
\end{aligned}$$

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

$$\begin{aligned}
0 &\geq -\|\Delta\|_1 + 2\|\widehat{\mathbf{w}}\|_1 - 2\|\mathbf{w}^*\|_1 \\
&\geq -\|\Delta\|_1 + 2\|\mathbf{w}^*\|_1 - 2\|\Delta_\mathcal{K}\|_1 + 2\|\Delta_{\mathcal{K}^c}\|_1 - 2\|\mathbf{w}^*\|_1 \\
&= -\|\Delta\|_1 - 2\|\Delta_\mathcal{K}\|_1 + 2\|\Delta_{\mathcal{K}^c}\|_1 \\
&= -\|\Delta_\mathcal{K}\|_1 - \|\Delta_{\mathcal{K}^c}\|_1 - 2\|\Delta_\mathcal{K}\|_1 + 2\|\Delta_{\mathcal{K}^c}\|_1 \\
&= -3\|\Delta_\mathcal{K}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1
\end{aligned}$$

where the second line follows from eq.(2). Given the above, we have:

$$\begin{aligned}\|\Delta\|_1 &= \|\Delta_{\mathcal{K}}\|_1 + \|\Delta_{\mathcal{K}^c}\|_1 \\ &\leq \|\Delta_{\mathcal{K}}\|_1 + 3\|\Delta_{\mathcal{K}}\|_1 \\ &= 4\|\Delta_{\mathcal{K}}\|_1 \\ &\leq 4\sqrt{k}\|\Delta_{\mathcal{K}}\|_2 \\ &\leq 4\sqrt{k}\|\Delta\|_2 \end{aligned} \tag{5}$$

which proves our claim. □

Next, we provide the final proof.

Proof of Theorem 8.1. Let $\Delta \equiv \widehat{\mathbf{w}} - \mathbf{w}^*$. First, since we assume that $\lambda \geq 2\|\nabla\ell(\mathbf{w}^*)\|_\infty$ we can invoke Lemma 8.1 and therefore:

$$\|\Delta\|_1 \leq 4\sqrt{k}\|\Delta\|_2 \tag{6}$$

For $\mathbf{w} = \widehat{\mathbf{w}}$, by restricted strong convexity of ℓ around $\mathbf{w}^*$ with parameters (α, τ, g) as in Definition 8.1, by eq.(6) and since $17\frac{\tau}{\alpha}kg(n,p) \leq 1$, we have:

$$\begin{aligned}\ell(\widehat{\mathbf{w}}) - \ell(\mathbf{w}^*) - \langle \nabla\ell(\mathbf{w}^*), \Delta\rangle &\geq \alpha\|\Delta\|_2^2 - \tau g(n,p)\|\Delta\|_1^2 \\ &\geq (\alpha - 16k\tau g(n,p))\|\Delta\|_2^2 \\ &\geq (\alpha - \tfrac{16}{17}\alpha)\|\Delta\|_2^2 \\ &= \tfrac{1}{17}\alpha\|\Delta\|_2^2 \end{aligned} \tag{7}$$

By eq.(6) and eq.(7), 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:

$$\begin{aligned}\tfrac{1}{17}\alpha\|\Delta\|_2^2 &\leq \ell(\widehat{\mathbf{w}}) - \ell(\mathbf{w}^*) - \langle \nabla\ell(\mathbf{w}^*), \Delta\rangle \\ &\leq -\lambda\|\widehat{\mathbf{w}}\|_1 + \lambda\|\mathbf{w}^*\|_1 + \|\nabla\ell(\mathbf{w}^*)\|_\infty\|\Delta\|_1 \\ &\leq -\lambda\|\widehat{\mathbf{w}}\|_1 + \lambda\|\mathbf{w}^*\|_1 + \tfrac{1}{2}\lambda\|\Delta\|_1 \\ &\leq -\lambda\|\widehat{\mathbf{w}}\|_1 + \lambda\|\widehat{\mathbf{w}}\|_1 + \lambda\|\Delta\|_1 + \tfrac{1}{2}\lambda\|\Delta\|_1 \\ &= \tfrac{3}{2}\lambda\|\Delta\|_1 \\ &\leq 6\sqrt{k}\lambda\|\Delta\|_2 \end{aligned}$$

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). □

2 Application: Compressed Sensing

Assume that there is an unknown but fixed $\mathbf{w}^* \in \mathbb{R}^p$. The only way to access $\mathbf{w}^*$ is through a *black box* that works as follows. We (somehow) generate an *input* vector $\mathbf{x}_i \in \{-1, +1\}^p$ and the black box returns an *output*:

$$y_i = \langle \mathbf{x}_i, \mathbf{w}^*\rangle + \varepsilon_i$$

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 $\langle \mathbf{x}_i, \mathbf{w}^* \rangle + \varepsilon_i$, but we do not have access to $\mathbf{w}^*$ or ε_i. We only have access to the output y_i, and of course the input $\mathbf{x}_i$.

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

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon} \tag{8}$$

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

$$\ell(\mathbf{w}) = \frac{1}{2n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \tag{9}$$

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

$$\begin{aligned}
\ell(\mathbf{w}) &= \frac{1}{2n}\|\mathbf{X}(\mathbf{w} - \mathbf{w}^*) - \boldsymbol{\epsilon}\|_2^2 \\
&= \frac{1}{2n}(\mathbf{w} - \mathbf{w}^*)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{w} - \mathbf{w}^*) - \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}(\mathbf{w} - \mathbf{w}^*) + \frac{1}{2n}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} \qquad (10)
\end{aligned}$$

By the above, we can conclude that:

$$\begin{aligned}
\ell(\mathbf{w}^*) &= \frac{1}{2n}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} \qquad (11) \\
\nabla\ell(\mathbf{w}) &= \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{w} - \mathbf{w}^*) - \frac{1}{n}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon} \\
\nabla\ell(\mathbf{w}^*) &= -\frac{1}{n}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon} \qquad (12)
\end{aligned}$$

In what follows we will assume that each entry of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ is independent and Rademacher distributed.

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

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\epsilon}}\left[(\exists j = 1 \ldots p) \ \left|\frac{1}{n}\langle \mathbf{x}^j, \boldsymbol{\epsilon} \rangle\right| \geq \frac{\lambda}{2} \middle| \mathbf{X}\right] &\leq 2p \ e^{\frac{-2n^2(\lambda/2)^2}{n 2^2}} \\
&= 2p \ e^{-\frac{n\lambda^2}{8}} \\
&= 2p \ e^{-2\log p} \\
&= 2/p
\end{aligned}$$

By eq.(12) and the above, we have:

$$\begin{aligned}\mathbb{P}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\nabla\ell(\mathbf{w}^*)\|_\infty \geq \frac{\lambda}{2}\right] &= \mathbb{P}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\left\|\frac{1}{n}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_\infty \geq \frac{\lambda}{2}\right] \\ &= \mathbb{P}_{\mathbf{X},\boldsymbol{\epsilon}}\left[(\exists j = 1\dots p)\ \left|\frac{1}{n}\left\langle \mathbf{x}^j, \boldsymbol{\epsilon}\right\rangle\right| \geq \frac{\lambda}{2}\right] \\ &= \mathbb{E}_{\mathbf{X}}\left[\mathbb{P}_{\boldsymbol{\epsilon}}\left[(\exists j = 1\dots p)\ \left|\frac{1}{n}\left\langle \mathbf{x}^j, \boldsymbol{\epsilon}\right\rangle\right| \geq \frac{\lambda}{2}\middle|\mathbf{X}\right]\right] \\ &\leq \mathbb{E}_{\mathbf{X}}\left[2/p\right] \\ &= 2/p\end{aligned}$$

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

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

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

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

$$\begin{aligned}\ell(\mathbf{w}) - \ell(\mathbf{w}^*) - \langle\nabla\ell(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^*\rangle &= \frac{1}{2n}(\mathbf{w} - \mathbf{w}^*)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{w} - \mathbf{w}^*) \\ &= \frac{1}{2n}\|\mathbf{X}(\mathbf{w} - \mathbf{w}^*)\|_2^2\end{aligned}$$

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

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

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

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

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 $i = 1\dots n$ are independent random variables. Moreover, $z_i \in [-1, +1]$ and we also know that $\mathbb{E}_{\mathbf{X}}[\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}] = 0$ since the entries of $\mathbf{X}$ are independent and zero-mean. Thus, by Hoeffding's inequality (Corollary 2.2), the union bound and by assuming $t = \sqrt{6\frac{\log p}{n}}$:

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}}\left[\max_{j\neq k}\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\right| \geq t\right] &= \mathbb{P}_{\mathbf{X}}\left[(\exists j\neq k)\ \left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\right| \geq t\right] \\
&\leq 2\binom{p}{2}\ e^{\frac{-2n^2t^2}{n2^2}} \\
&\leq p^2\ e^{\frac{-nt^2}{2}} \\
&= p^2 e^{-3\log p} \\
&= 1/p
\end{aligned}$$

Therefore, with probability at least $1 - 1/p$ over the choice of $\mathbf{X}$, we have that:

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

Note that since $\|\mathbf{v}\|_1 = 1$, then:

$$\begin{aligned}
\sum_{j\neq k}|v_j v_k| &\leq \sum_{j=1}^{p}\sum_{k=1}^{p}|v_j v_k| \\
&= \sum_{j=1}^{p}\sum_{k=1}^{p}|v_j|\ |v_k| \\
&= \|\mathbf{v}\|_1^2 \\
&= 1
\end{aligned} \tag{14}$$

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

$$\begin{aligned}
\frac{1}{2n}\|\mathbf{X}\mathbf{v}\|_2^2 &= \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{X}\mathbf{v})_i^2 \\
&= \frac{1}{2n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij}v_j\right)^2 \\
&= \frac{1}{2n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij}^2 v_j^2 + \sum_{j\neq k} x_{ij}x_{ik}v_j v_k\right) \\
&= \frac{1}{2n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} v_j^2 + \sum_{j\neq k} x_{ij}x_{ik}v_j v_k\right) \\
&= \frac{1}{2}\|\mathbf{v}\|_2^2 + \frac{1}{2}\sum_{j\neq k}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\right) v_j v_k
\end{aligned}$$

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

where the previous-to-the-last step follows from the Cauchy-Schwarz inequality $(\forall \mathbf{a}, \mathbf{b})\ |\langle \mathbf{a}, \mathbf{b}\rangle| \leq \|\mathbf{a}\|_1 \|\mathbf{b}\|_\infty$. 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:

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

Therefore, we conclude that the loss function ℓ 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 $17\frac{\tau}{\alpha}kg(n,p) \leq 1$. That is:

$$17\frac{\tau}{\alpha}kg(n,p) = 17\sqrt{6}\ k\sqrt{\frac{\log p}{n}} \leq 1$$

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

[1] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 2012.

[2] G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.