Hands-On Learning Theory
Fall 2016, Lecture 9

Jean Honorio jhonorio@purdue.edu

1 Primal-Dual Witness: Some Motivation

For brevity, everywhere differentiable functions will be called *smooth*. Similarly, not everywhere differentiable functions will be called *nonsmooth*.

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 $\mathbf{w}^* \in \mathbb{R}^p$. In Theorem 8.1, our goal was to recover a vector $\widehat{\mathbf{w}}$ which is *close to* $\mathbf{w}^*$. Here, we also want to make sure that the vector $\widehat{\mathbf{w}}$ has the same set of non-zero elements and the correct signs of each element as of $\mathbf{w}^*$.

Next, we define the *support* of a vector.

Definition 9.1. *Given a vector $\mathbf{w} \in \mathbb{R}^p$, its support is defined as:*

$$\mathcal{S}(\mathbf{w}) = \{i \in \{1 \dots p\} \mid w_i \neq 0\}$$

For simplicity, we will use the shorthand notation $\mathcal{S} \equiv \mathcal{S}(\mathbf{w}^*)$.

Questions: We want to understand the sufficient conditions for which we have:

1. A unique solution for eq.(1).
2. Correct support recovery, i.e. $\mathcal{S}(\widehat{\mathbf{w}}) = \mathcal{S}$.
3. Correct sign recovery, i.e. $(\forall i \in \mathcal{S})\ \mathrm{sgn}(\widehat{w}_i) = \mathrm{sgn}(w_i^*)$.

In order to prove the above, we will make use of the *Karush-Kuhn-Tucker (KKT) conditions* for the specific problem in eq.(1). The KKT conditions that we will use are stationarity, complementary slackness and dual feasibility. In general, for any optimization problem, the optimal solution has to fulfill these KKT conditions.

We introduce some notation. Let $\mathcal{S}^c$ denote the complement of $\mathcal{S}$. For an arbitrary vector $\mathbf{a}$, let $\mathbf{a}_{\mathcal{S}}$ denote the vector that contains only the entries in the

set $\mathcal{S}$. For an arbitrary matrix $\mathbf{A}$, let $\mathbf{A}_{\mathcal{S}}$ denote the matrix that contains all rows of $\mathbf{A}$ but only the columns in the set $\mathcal{S}$. For an square matrix $\mathbf{A}$, let $\mathbf{A}_{\mathcal{S},\mathcal{S}}$ denote the matrix that contains only the rows and columns in the set $\mathcal{S}$.

First, note that $\|\mathbf{w}\|_1$ in eq.(1) is nonsmooth. The *stationarity condition* for eq.(1) is:

$$(\exists \widehat{\mathbf{z}} \in \partial \|\widehat{\mathbf{w}}\|_1)\ \nabla \ell(\widehat{\mathbf{w}}) + \lambda \widehat{\mathbf{z}} = \mathbf{0}$$

where $\partial \|\widehat{\mathbf{w}}\|_1$ is the subdifferential set of the function $\|\cdot\|_1$ at $\widehat{\mathbf{w}}$. Recall that by norm duality, we have:

$$\|\widehat{\mathbf{w}}\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \mathbf{z}, \widehat{\mathbf{w}} \rangle$$

Then, in order to maximize the above linear function, we arrive to the *complementary slackness condition*:

$$\widehat{\mathbf{z}} \in \partial \|\widehat{\mathbf{w}}\|_1 \ \Leftrightarrow\ \begin{cases} \widehat{z}_i = +1 & \wedge\ \widehat{w}_i > 0 \\ \widehat{z}_i = -1 & \wedge\ \widehat{w}_i < 0 \\ \widehat{z}_i \in [-1, +1] & \wedge\ \widehat{w}_i = 0 \end{cases}$$

From the above, first note that:

$$\|\widehat{\mathbf{z}}\|_\infty \leq 1$$

Second, note that:

$$|\widehat{z}_i| < 1 \Rightarrow \widehat{w}_i = 0$$

Recall that $\mathcal{S}^c$ is the complement of $\mathcal{S}$ and that $\mathcal{S} \equiv \mathcal{S}(\mathbf{w}^*)$. If we assume *strict dual feasibility*:

$$(\forall i \notin \mathcal{S})\ |\widehat{z}_i| < 1 \quad \text{or equivalently} \quad \|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1$$

then we can conclude that:

$$(\forall i \notin \mathcal{S})\ \widehat{w}_i = 0 \quad \text{or equivalently} \quad \widehat{\mathbf{w}}_{\mathcal{S}^c} = \mathbf{0}$$

The above would imply that there are no *false positives*. That is, every entry in $\widehat{\mathbf{w}}$ that is not in the support are zero. But some entries in $\widehat{\mathbf{w}}$ that are in the support could potentially be zero as well. Formally speaking we conclude that $\mathcal{S}(\widehat{\mathbf{w}}) \subseteq \mathcal{S}$.

Second, assume that we knew the support $\mathcal{S}$. We could solve a restricted problem of eq.(1) as follows:

$$\widetilde{\mathbf{w}}_{\mathcal{S}} = \operatorname*{arg\,min}_{\mathbf{w}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}} \ell((\mathbf{w}_{\mathcal{S}}, \mathbf{0})) + \lambda \|\mathbf{w}_{\mathcal{S}}\|_1$$

Clearly, if the Hessian is positive definite then ℓ is strictly convex and thus $\widetilde{\mathbf{w}}_{\mathcal{S}}$ would be the unique solution of the above problem. Furthermore, $(\widetilde{\mathbf{w}}, \mathbf{0})$ would be the unique solution of eq.(1). The positive definiteness requirement of the Hessian would be:

$$(\forall \mathbf{w}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}) \ \left[\nabla^2 \ell((\mathbf{w}_{\mathcal{S}}, \mathbf{0}))\right]_{\mathcal{S},\mathcal{S}} \succ \mathbf{0}$$

2 Primal-Dual Witness: The Method

The primal-dual witness method is not a practical algorithm for solving eq.(1). It is a proof technique that allows to answer the questions that we posed at the beginning. The idea is to assume the true support $\mathcal{S}$ as given and then finding out which conditions allow for the KKT conditions to hold.

Assume that we knew the support $\mathcal{S}$. We construct a primal-dual witness solution $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{z}})$ as follows:

Step 1: Assume that:

$$(\forall \mathbf{w}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}) \ \left[\nabla^2 \ell((\mathbf{w}_{\mathcal{S}}, \mathbf{0}))\right]_{\mathcal{S},\mathcal{S}} \succ \mathbf{0}$$

Step 2: Construct the primal variable $\widetilde{\mathbf{w}}$ by making $\widetilde{\mathbf{w}}_{\mathcal{S}^c} = \mathbf{0}$ and by solving the restricted problem:

$$\widetilde{\mathbf{w}}_{\mathcal{S}} = \underset{\mathbf{w}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}}{\arg\min}\, \ell((\mathbf{w}_{\mathcal{S}}, \mathbf{0})) + \lambda \|\mathbf{w}_{\mathcal{S}}\|_1 \tag{2}$$

Step 3: Choose the dual variable $\widetilde{\mathbf{z}}_{\mathcal{S}}$ in order to fufill the *complementary slackness condition*:

$$(\forall i \in \mathcal{S}) \ \begin{cases} \widetilde{z}_i = +1 & \text{if } \ \widetilde{w}_i > 0 \\ \widetilde{z}_i = -1 & \text{if } \ \widetilde{w}_i < 0 \\ \widetilde{z}_i \in [-1, +1] & \text{if } \ \widetilde{w}_i = 0 \end{cases}$$

Note that:

$$\|\widetilde{\mathbf{z}}_{\mathcal{S}}\|_\infty \leq 1$$

Step 4: Solve for the dual variable $\widetilde{\mathbf{z}}_{\mathcal{S}^c}$ in order that $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{z}})$ fulfills the *stationarity condition*:

$$\begin{aligned} [\nabla \ell((\widetilde{\mathbf{w}}_{\mathcal{S}}, \mathbf{0}))]_{\mathcal{S}} + \lambda \widetilde{\mathbf{z}}_{\mathcal{S}} &= \mathbf{0} \\ [\nabla \ell((\widetilde{\mathbf{w}}_{\mathcal{S}}, \mathbf{0}))]_{\mathcal{S}^c} + \lambda \widetilde{\mathbf{z}}_{\mathcal{S}^c} &= \mathbf{0} \end{aligned}$$

By solving the above:

$$\widetilde{\mathbf{z}}_{\mathcal{S}^c} = \frac{-1}{\lambda}\left[\nabla \ell((\widetilde{\mathbf{w}}_{\mathcal{S}}, \mathbf{0}))\right]_{\mathcal{S}^c}$$

Step 5: Verify that the *strict dual feasibility condition* is fulfilled. That is:

$$\|\widetilde{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1 \quad \text{or equivalently} \quad \frac{1}{\lambda}\|[\nabla \ell((\widetilde{\mathbf{w}}_{\mathcal{S}}, \mathbf{0}))]_{\mathcal{S}^c}\|_\infty < 1$$

If the primal-dual witness construction succeeds, then it acts as a witness to the fact that the solution $\widetilde{\mathbf{w}}$ to the restricted problem in eq.(2) is equal to the solution $\widehat{\mathbf{w}}$ to the problem in eq.(1). Note that all steps hold by construction, with the exception of Step (1) and Step (5). Thus, when applying this technique we should guarantee that both steps hold.

Step 6: Up until now, we would have shown that there are no *false positives* $\mathcal{S}(\widetilde{\mathbf{w}}) \subseteq \mathcal{S}$. For proving correct sign recovery, we need to show that $(\forall i \in \mathcal{S})\ \mathrm{sgn}(\widetilde{w}_i) = \mathrm{sgn}(w_i^*)$. By our derivations so far, we could alternatively show that $(\forall i \in \mathcal{S})\ \widetilde{z}_i = \mathrm{sgn}(w_i^*)$.

Very often, the analysis of Step (5) allows to arrive to an expression of the form $\|\widetilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}_{\mathcal{S}}^*\|_\infty \leq \varepsilon$. From here, it is trivial to show correct sign recovery by using the following lemma.

Lemma 9.1. *Let $a, b \in \mathbb{R}$ and fix $\varepsilon > 0$. We have that:*

$$|a - b| \leq \varepsilon \ \wedge\ |b| > 2\varepsilon \ \Rightarrow\ \mathrm{sgn}(a) = \mathrm{sgn}(b)$$

Proof. Assume $b > 2\varepsilon$, since $|a - b| \leq \varepsilon$ we have $a \geq b - \varepsilon > 2\varepsilon - \varepsilon = \varepsilon > 0$.
Assume $b < -2\varepsilon$, since $|a - b| \leq \varepsilon$ we have $a \leq b + \varepsilon < -2\varepsilon + \varepsilon = -\varepsilon < 0$.
In the two cases above, we have $\mathrm{sgn}(a) = \mathrm{sgn}(b)$. □

Then, we can invoke the above lemma for our purposes.

Lemma 9.2. *We have that:*

$$\|\widetilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}_{\mathcal{S}}^*\|_\infty \leq \varepsilon \ \wedge\ \min_{i \in \mathcal{S}} |w_i^*| > 2\varepsilon \ \Rightarrow\ (\forall i \in \mathcal{S})\ \mathrm{sgn}(\widetilde{w}_i) = \mathrm{sgn}(w_i^*)$$

Proof. Note that:

$$\|\widetilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}_{\mathcal{S}}^*\|_\infty \leq \varepsilon \ \wedge\ \min_{i \in \mathcal{S}} |w_i^*| > 2\varepsilon \ \Leftrightarrow\ (\forall i \in \mathcal{S})\ |\widetilde{w}_i - w_i^*| \leq \varepsilon \ \wedge\ |w_i^*| > 2\varepsilon$$

By Lemma 9.1, we prove our claim. □

3 Application: Noiseless Compressed Sensing

Let $\mathbf{A} \in \mathbb{R}^{m \times k}$ be an arbitrary matrix. Let $\mathbf{a}_i$ be the i-th row of $\mathbf{A}$. Denote by $|||\cdot|||_\infty$ the induced norm:

$$|||\mathbf{A}|||_\infty \equiv \max_{i=1}^{m} \|\mathbf{a}_i\|_1$$

Let $\|\cdot\|_2$ be the spectral norm. If $\mathbf{A} \in \mathbb{R}^{k \times k}$, we have that:

$$|||\mathbf{A}|||_\infty \leq \sqrt{k}\|\mathbf{A}\|_2$$

We will use $\phi_{\min}(\mathbf{A})$ and $\phi_{\max}(\mathbf{A})$ to denote the minimum and maximum eigenvalues of $\mathbf{A}$ respectively. The following property follows from the Cauchy-Schwarz inequality:

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_\infty &= \max_{i=1}^{m} |\langle \mathbf{a}_i, \mathbf{x} \rangle| \\
&\leq \max_{i=1}^{m} \|\mathbf{a}_i\|_1 \|\mathbf{x}\|_\infty \\
&= |||\mathbf{A}|||_\infty \|\mathbf{x}\|_\infty
\end{aligned}$$

Let $\mathbf{A} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{k \times k}$ be two matrices, from the above we also have:

$$\begin{aligned}
|||\mathbf{A}\mathbf{B}|||_\infty &= \max_{i=1}^{m} \|\mathbf{a}_i \mathbf{B}\|_1 \\
&\leq k \max_{i=1}^{m} \|\mathbf{a}_i \mathbf{B}\|_\infty \\
&= k \max_{i=1}^{m} \|\mathbf{B}^{\mathrm{T}} \mathbf{a}_i^{\mathrm{T}}\|_\infty \\
&\leq k |||\mathbf{B}^{\mathrm{T}}|||_\infty \max_{i=1}^{m} \|\mathbf{a}_i^{\mathrm{T}}\|_\infty \\
&= k \|\mathbf{A}\|_\infty |||\mathbf{B}^{\mathrm{T}}|||_\infty
\end{aligned}$$

The above properties of the induced $|||\cdot|||_\infty$ will be very useful in our proofs.

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 $\mathbf{x}_i \in \mathbb{R}^p$ and the black box returns an *output*:

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

In the above, we know that y_i is equal to $\langle \mathbf{x}_i, \mathbf{w}^* \rangle$, but we do not have access to $\mathbf{w}^*$. 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 $\widehat{\mathbf{w}}$ with the same support and signs of $\mathbf{w}^*$. Assume we obtain n pairs. Let $\mathbf{X} \in \{-1, +1\}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$. Note that:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* \tag{3}$$

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

$$\ell(\mathbf{w}) = \frac{1}{2n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \tag{4}$$

For answering our questions, we will have to show that the loss function ℓ fulfills Step (1) and Step (5). From eq.(3) and eq.(4), we have:

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

By the above, we can conclude that:

$$\begin{aligned}\nabla\ell(\mathbf{w}) &= \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{w} - \mathbf{w}^*) \\ \nabla\ell((\mathbf{w}_{\mathcal{S}}, \mathbf{0})) &= \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}} - \mathbf{w}^*_{\mathcal{S}}) \qquad (5) \\ \nabla^2\ell(\mathbf{w}) &= \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} \qquad (6)\end{aligned}$$

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

Assumption. We will assume that $n \geq 16|\mathcal{S}|^3 \log p$.

Proof for Step (1). Note that from eq.(6), we have:

$$(\forall \mathbf{w}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}) \;\; \left[\nabla^2\ell((\mathbf{w}_{\mathcal{S}}, \mathbf{0}))\right]_{\mathcal{S},\mathcal{S}} \succ \mathbf{0} \;\Leftrightarrow\; \frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}} \succ \mathbf{0}$$

The above is equivalent to assuming that there exists a $\beta > 0$ for which:

$$\phi_{\min}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right) > \beta \tag{7}$$

If such a bound exists, then the problem in eq.(2) has a unique solution. Note that $\mathbf{X}_{\mathcal{S}} \in \{-1, +1\}^{n \times |\mathcal{S}|}$ and thus $\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$. Therefore:

$$\begin{aligned}\left\|\left\|\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\right\|\right\|_{\infty} &\leq |\mathcal{S}|^{1/2}\left\|\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\right\|_2 \\ &< \frac{|\mathcal{S}|^{1/2}}{\beta} \qquad (8)\end{aligned}$$

The above inequality will be useful for proving Step (5) and Step (6).

We now prove that eq.(7) holds with high probability. Let $\mathbf{x}_{\mathcal{S}} \in \{-1, +1\}^{|\mathcal{S}|}$ denote one of the n rows of $\mathbf{X}_{\mathcal{S}}$. Note that $\mathbf{x}_{\mathcal{S}}$ is a random vector of Rademacher

variables, where each entry is independent and thus uncorrelated. More formally for all $i, j \in \mathcal{S}$ such that $i \neq j$ we have $\mathbb{E}[x_i x_j] = \mathbb{E}[x_i]\ \mathbb{E}[x_j] = 0$. Moreover, for all $i \in \mathcal{S}$ we have $\mathbb{E}[x_i^2] = 1$. Therefore, $\mathbb{E}[\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}] = \mathbf{I}$ and we conclude that:

$$\phi_{\min}\left(\mathbb{E}[\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}]\right) = 1$$

Since $\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}$ is a rank-1 matrix, we conclude that:

$$\phi_{\min}\left(\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}\right) = 0$$

Since $\mathbf{x}_{\mathcal{S}} \in \{-1, +1\}^{|\mathcal{S}|}$, by the *variational characterization* of eigenvalues:

$$\begin{aligned}
\phi_{\max}\left(\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}\right) &= \max_{\mathbf{a}^{\mathrm{T}}\mathbf{a}=1} \mathbf{a}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}\mathbf{a} \\
&= \max_{\mathbf{a}^{\mathrm{T}}\mathbf{a}=1} \left(\mathbf{a}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}\right)^2 \\
&= |\mathcal{S}|
\end{aligned}$$

the above follows from picking $(\forall i \in \mathcal{S})\ a_i = x_i/|\mathcal{S}|^{1/2}$. From the above three conclusions, we can invoke Theorem 1.1 in [1], we have by conveniently making $t = 1/2$:

$$\begin{aligned}
\mathbb{P}\left[\phi_{\min}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right) \leq 1 - t\right] &= \mathbb{P}\left[\phi_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{\mathcal{S}}^{(i)}{\mathbf{x}_{\mathcal{S}}^{(i)}}^{\mathrm{T}}\right) \leq (1-t)\phi_{\min}\big(\mathbb{E}[\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}]\big)\right] \\
&\leq |\mathcal{S}|\left(\frac{e^{-t}}{(1-t)^{1-t}}\right)^{\frac{n\phi_{\min}\left(\mathbb{E}[\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathrm{T}}]\right)}{|\mathcal{S}|}} \\
&= |\mathcal{S}|\left(\sqrt{\frac{2}{e}}\right)^{\frac{n}{|\mathcal{S}|}} \\
&\leq |\mathcal{S}|\left(e^{-1/8}\right)^{\frac{n}{|\mathcal{S}|}} \\
&= |\mathcal{S}|\ e^{\frac{-n}{8|\mathcal{S}|}} \\
&\leq |\mathcal{S}|\ e^{-2|\mathcal{S}|^2\log p} \\
&= |\mathcal{S}|/p^{2|\mathcal{S}|^2}
\end{aligned}$$

Therefore, with probability at least $1 - |\mathcal{S}|/p^{2|\mathcal{S}|^2}$ over the choice of $\mathbf{X}$, we have:

$$\phi_{\min}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right) > 1/2 = \beta$$

Proof for Step (5). The main goal of this step is to state the sufficient condition for which the *strict dual feasibility condition* is fulfilled (i.e., $\|\widetilde{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1$). From Step (4), we have:

$$\begin{aligned}
\left[\nabla\ell((\widetilde{\mathbf{w}}_{\mathcal{S}}, \mathbf{0}))\right]_{\mathcal{S}} + \lambda\widetilde{\mathbf{z}}_{\mathcal{S}} &= \mathbf{0} \\
\left[\nabla\ell((\widetilde{\mathbf{w}}_{\mathcal{S}}, \mathbf{0}))\right]_{\mathcal{S}^c} + \lambda\widetilde{\mathbf{z}}_{\mathcal{S}^c} &= \mathbf{0}
\end{aligned}$$

or equivalently, by eq.(5) we have:

$$\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}-\mathbf{w}_{\mathcal{S}}^{*})+\lambda\widetilde{\mathbf{z}}_{\mathcal{S}}=\mathbf{0}$$
$$\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}-\mathbf{w}_{\mathcal{S}}^{*})+\lambda\widetilde{\mathbf{z}}_{\mathcal{S}^c}=\mathbf{0}$$

We can reorganize the above equations and get:

$$\mathbf{w}_{\mathcal{S}}-\mathbf{w}_{\mathcal{S}}^{*}=-\lambda\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\widetilde{\mathbf{z}}_{\mathcal{S}} \tag{9}$$
$$\widetilde{\mathbf{z}}_{\mathcal{S}^c}=\frac{-1}{\lambda n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}-\mathbf{w}_{\mathcal{S}}^{*})=\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\widetilde{\mathbf{z}}_{\mathcal{S}}$$

Since $\|\widetilde{\mathbf{z}}_{\mathcal{S}}\|_\infty \leq 1$, from the above and eq.(8) we have:

$$\begin{aligned}
\|\widetilde{\mathbf{z}}_{\mathcal{S}^c}\|_\infty &= \left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\widetilde{\mathbf{z}}_{\mathcal{S}}\right\|_\infty \\
&\leq \left\|\left|\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\right|\right\|_\infty \|\widetilde{\mathbf{z}}_{\mathcal{S}}\|_\infty \\
&\leq |\mathcal{S}|\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right\|_\infty \left\|\left|\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\right|\right\|_\infty \|\widetilde{\mathbf{z}}_{\mathcal{S}}\|_\infty \\
&\leq \frac{|\mathcal{S}|^{3/2}}{\beta}\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right\|_\infty
\end{aligned}$$

From the above, it becomes clear that:

$$\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right\|_\infty < \frac{\beta}{|\mathcal{S}|^{3/2}} \;\Rightarrow\; \|\widetilde{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1 \tag{10}$$

Thus, the left-hand side of eq.(10) is the sufficient condition for which the *strict dual feasibility condition* is fulfilled (i.e., $\|\widetilde{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1$).

We now prove that the left-hand side of eq.(10) *holds with high probability.* 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, we know that $z_i \in [-1, +1]$ and $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}] = 0$ since the entries of $\mathbf{X}$ are independent and zero-mean. Thus, by Hoeffding's inequality (Corollary 2.2), the union bound and by conveniently making $t = \beta/|\mathcal{S}|^{3/2}$:

$$\begin{aligned}
\mathbb{P}\left[\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right\|_\infty \geq t\right] &= \mathbb{P}\left[(\exists j \in \mathcal{S}^c, k \in \mathcal{S})\;\left|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}\right| \geq t\right] \\
&\leq 2|\mathcal{S}^c||\mathcal{S}|\; e^{\frac{-2n^2t^2}{n2^2}}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2p|\mathcal{S}|\ e^{\frac{-nt^2}{2}} \\
&= 2p|\mathcal{S}|\ e^{\frac{-n\beta^2}{2|\mathcal{S}|^3}} \\
&\leq 2p|\mathcal{S}|\ e^{-2\log p} \\
&= 2|\mathcal{S}|/p
\end{aligned}
$$

Therefore, with probability at least $1 - 2|\mathcal{S}|/p$ over the choice of $\mathbf{X}$, we have:

$$
\left\| \frac{1}{n} \mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}} \mathbf{X}_{\mathcal{S}} \right\|_{\infty} < \frac{\beta}{|\mathcal{S}|^{3/2}}
$$

and thus, the *strict dual feasibility condition* is fulfilled (i.e., $\|\widetilde{\mathbf{z}}_{\mathcal{S}^c}\|_{\infty} < 1$).

Proof for Step (6). Since $\|\widetilde{\mathbf{z}}_{\mathcal{S}}\|_{\infty} \leq 1$, from eq.(8) and eq.(9):

$$
\begin{aligned}
\|\mathbf{w}_{\mathcal{S}} - \mathbf{w}_{\mathcal{S}}^*\|_{\infty} &= \lambda \left\| \left(\frac{1}{n} \mathbf{X}_{\mathcal{S}}^{\mathrm{T}} \mathbf{X}_{\mathcal{S}} \right)^{-1} \widetilde{\mathbf{z}}_{\mathcal{S}} \right\|_{\infty} \\
&\leq \lambda \left\| \left(\frac{1}{n} \mathbf{X}_{\mathcal{S}}^{\mathrm{T}} \mathbf{X}_{\mathcal{S}} \right)^{-1} \right\|_{\infty} \|\widetilde{\mathbf{z}}_{\mathcal{S}}\|_{\infty} \\
&< \frac{\lambda |\mathcal{S}|^{3/2}}{\beta}
\end{aligned}
$$

References

[1] J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[2] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.