**MID-TERM EXAM**

**YangyangLi**
li002252@umn.edu

# 1 Problem 1

## 1.1 a

**Answers:**

$g(z) = |z|$ that is convex function but non-differentiable at point $0$. So using subdifferential to generalize gradient of differential functions at $z = 0$.

$$\partial g(z_0) = \{x \in \mathbb{R} : g(z) \geq g(z_0) + x(z - z_0)\}$$

Where $z_0 = 0 \quad g(z_0) = 0$.

So,

$$\begin{aligned}\partial g(z_0) &= \{x \in \mathbb{R} : g(z) \geq g(z_0) + x(z - z_0)\} \\ &= \{x \in \mathbb{R} : g(z) \geq g(z_0) + xz\} \\ &= \{x \in \mathbb{R} : |z| \geq xz\} \\ &= \{x \in \mathbb{R} : -1 \leq x \leq 1\}\end{aligned}$$

Also, $\partial_z |z| = 1$ when $z > 0$. $\partial_z |z| = -1$ when $z < 0$. Hence,

$$\partial_z |z| = \begin{cases} 1 & z > 0 \\ -1 & z < 0 \\ [-1, 1] & z = 0 \end{cases}.$$

## 1.2 b

**Answers:**

As,

$$f(\boldsymbol{x}) \doteq \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$

Let $g(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$, its gradient is $\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$, which is shown in previous HWs. Concretely,

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ &\to \|\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{\delta}) - \boldsymbol{y}\|_2^2 \\ &= \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + 2\left\langle \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}), \boldsymbol{\delta} \right\rangle + o(\|\boldsymbol{\delta}\|_2)\end{aligned}$$

So, $\partial_x g(\boldsymbol{x}) = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$. As $\boldsymbol{A} \in \mathbb{R}^{N \times d}$, $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{x} \in \mathbb{R}^{d \times 1}$

$$\partial_{\boldsymbol{x}} g(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \\ \vdots \\ \frac{\partial g}{\partial x_d} \end{bmatrix}$$

Let $h(x) = \|\boldsymbol{x}\|_1$,

$$\partial_{\boldsymbol{x}} h(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial h}{\partial x_1} \\ \frac{\partial h}{\partial x_2} \\ \vdots \\ \frac{\partial h}{\partial x_d} \end{bmatrix}$$

As shown in 1.1, for every scalar $x_i$, $\frac{\partial h}{\partial x_i} = \operatorname{sign}(x_i)$. So,

$$\partial_{\boldsymbol{x}} h(\boldsymbol{x}) = \begin{bmatrix} \operatorname{sign}(x_1) \\ \operatorname{sign}(x_2) \\ \vdots \\ \operatorname{sign}(x_d) \end{bmatrix}$$

Consequently,

$$\begin{aligned} \partial_{\boldsymbol{x}} f(\boldsymbol{x}) &= \partial_{\boldsymbol{x}} g(\boldsymbol{x}) + \partial_{\boldsymbol{x}} h(\boldsymbol{x}) \\ &= \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) + \lambda \operatorname{sign}(\boldsymbol{x}) \end{aligned}$$

Where $\operatorname{sign}(\boldsymbol{x})$ means applying the sign function elementwise.

## 1.3 c

**Answers:**

Let $f(\boldsymbol{z}) = \|\boldsymbol{z}\|_2$. It is not deifferential when $\boldsymbol{z} = \boldsymbol{0}$. So,

$$\begin{aligned} \partial_{\boldsymbol{z}} f(\boldsymbol{z}_0) &= \left\{\boldsymbol{x} \in \mathbb{R}^N : f(\boldsymbol{z}) \geq f(\boldsymbol{z}_0) + \langle \boldsymbol{x}, \boldsymbol{z} - \boldsymbol{z}_0 \rangle \right\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{z}\| \geq \boldsymbol{x}^T \boldsymbol{z} \right\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{z}\| \leq 1 \right\} \end{aligned}$$

Where $\boldsymbol{z}_0 = \boldsymbol{0}$.
When $\boldsymbol{z}$ does not equal to $\boldsymbol{0}$,

$$\partial_{\boldsymbol{z}} f(\boldsymbol{z}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{z})}{\partial z_1} \\ \frac{\partial f(\boldsymbol{z})}{\partial z_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{z})}{\partial z_n} \end{bmatrix}$$

Where $\frac{\partial f(z)}{\partial z_i} = \frac{z_i}{\|z\|}$.
Hence,

$$\partial_{\boldsymbol{z}} \|\boldsymbol{z}\|_2 = \begin{cases} \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} & \boldsymbol{z} \neq \mathbf{0} \\ \left\{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\| \leq 1\right\} & \boldsymbol{z} = \mathbf{0} \end{cases}$$

## 1.4 d

### 1.4.1 i

**Answers:**

I implement Lasso to estimate $\boldsymbol{x}$ on the Colab Notebook that is uploaded as well. Please feel free to check that.

### 1.4.2 ii

**Answers:**

Let $g(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$. As proved in the questions 1.3, its subgradient is shown below in terms of chain rule.

$$g(\boldsymbol{x}) = \begin{cases} \frac{\left(\mathbf{1}^T(\boldsymbol{A} \otimes (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}))\right)^T}{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2} & \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \neq \mathbf{0} \\ \{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq 1\} & \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} = \mathbf{0} \end{cases}$$

Where $\boldsymbol{A} \otimes (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$ is broadcasting operation.

I implement Lasso to estimate $\boldsymbol{x}$ on the Colab Notebook that is uploaded as well. Please feel free to check that.

## 1.5 iii

**Answers:**

I fix the $\lambda = 1e - 2$, and try $\sigma^2 = 0.7, 2, 4$. I find that both Lasso and square-root Lasso have similar recovery performance of groundtruth $x_0$. Also, The higher the variance is, the worse the performance will be.
I finish this part on the Colab Notebook that is uploaded as well. Please feel free to check that.

# 2 Problem 2

## 2.1 a

**Answers:**

$$\min_{\boldsymbol{w} \in \mathbb{H}} \sum_{i=1}^{N} \left(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle - y_i\right)^2 + \lambda \|\boldsymbol{w}\|_{\mathbb{H}}^2 = \min_{\boldsymbol{w} \in \mathbb{H}} \|\Phi(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{w}\|_{\mathbb{H}}^2$$

Where $\Phi(\boldsymbol{X}) = \begin{bmatrix} \Phi\left(\boldsymbol{x}_1^T\right) \\ \Phi\left(\boldsymbol{x}_2^T\right) \\ \vdots \\ \Phi\left(\boldsymbol{x}_N^T\right) \end{bmatrix}$. the gradient is,

$$\nabla_{\boldsymbol{w}} = 2\Phi(\boldsymbol{X})^T (\Phi(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}) + 2\lambda \boldsymbol{w}$$

the global minimizer is $\boldsymbol{w}_*$,

$$\left(\Phi(\boldsymbol{X})^T \Phi(\boldsymbol{X}) + \lambda \boldsymbol{I}\right) \boldsymbol{w}_* = \Phi(\boldsymbol{X})^T \boldsymbol{y}$$

.

For any nonzero vector $\boldsymbol{u}$,

$$\boldsymbol{u}^T \left(\Phi(\boldsymbol{X})^T \Phi(\boldsymbol{X}) + \lambda \boldsymbol{I}\right) \boldsymbol{u} = \|\Phi(\boldsymbol{X})\boldsymbol{u}\|^2 + \lambda \|\boldsymbol{u}\|^2 > 0$$

.

So $\Phi(\boldsymbol{X})^T \Phi(\boldsymbol{X}) + \lambda \boldsymbol{I}$ is positive definite, which implies that it is invertible.

$$\boldsymbol{w}_* = \left(\Phi(\boldsymbol{X})^T \Phi(\boldsymbol{X}) + \lambda \boldsymbol{I}\right)^{-1} \Phi(\boldsymbol{X})^T \boldsymbol{y} = \Phi(\boldsymbol{X})^T \left(\Phi(\boldsymbol{X}) \Phi(\boldsymbol{X})^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y}$$

.

Rewriting $\boldsymbol{w}_* = \sum_{i=1}^N \alpha_i \Phi(\boldsymbol{x}_i)$, $\quad \boldsymbol{\alpha} = \left(\Phi(\boldsymbol{X}) \Phi(\boldsymbol{X})^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y} = (\boldsymbol{G} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}$
where $G$ is the Gram matrix generated from $K$ on our training data points $\{x_i\}_{i=1}^N$.

Hence, the regularized regression problem can be solved by find the optimal $\boldsymbol{\alpha}$. Also, $\boldsymbol{\alpha}$ only relates to $\boldsymbol{G}$ and $\boldsymbol{y}$

## 2.2 b

**Answers:**

The new optimization problem need to find the optimal $\boldsymbol{\alpha}$. As proved in question 2.1, $\boldsymbol{\alpha} = \left(\Phi(\boldsymbol{X}) \Phi(\boldsymbol{X})^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y} = (\boldsymbol{G} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}$. Obviously, it is unique global minimizer.

## 2.3 c

**Answers:**

the find predictor is $\langle \boldsymbol{w}_*, \Phi(\boldsymbol{x}) \rangle$. The matrix form is,

$$\Phi\left(\hat{X}\right) \boldsymbol{w}_* = \Phi\left(\hat{X}\right) \Phi(\boldsymbol{X})^T \left(\Phi(\boldsymbol{X}) \Phi(\boldsymbol{X})^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y} = \hat{\boldsymbol{G}} (\boldsymbol{G} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}$$

Where train data $\Phi(\boldsymbol{X}) = \begin{bmatrix} \Phi\left(\boldsymbol{x}_1^T\right) \\ \Phi\left(\boldsymbol{x}_2^T\right) \\ \vdots \\ \Phi\left(\boldsymbol{x}_N^T\right) \end{bmatrix}$. test data $\Phi\left(\hat{X}\right) = \begin{bmatrix} \Phi\left(\boldsymbol{x}_1^T\right) \\ \Phi\left(\boldsymbol{x}_2^T\right) \\ \vdots \end{bmatrix}$. $G$ is the Gram matrix generated from $K$ on our training data points $\{x_i\}_{i=1}^N$. $\hat{G}$ is the Gram matrix generated from $K$ on our testing data points and training data points.

# 3 Problem 3

## 3.1 a

**Answers:**

The problem is convex because $\frac{1}{2}\|\boldsymbol{w}\|_2^2$, $-v\rho$ and $\sum_{i=1}^{N}\xi_i$ are all convex. Their Hessian matrix $\nabla^2 \geq 0$. The positive combination $\frac{1}{2}\|\boldsymbol{w}\|_2^2 - v\rho + \sum_{i=1}^{N}\epsilon_i$ is convex. Also, the constraints is linear. So, the problem is convex.

## 3.2 b

**Answers:**

For the soft-margin SVM, there exist $\boldsymbol{w}_0$, $b_0$ and $\xi_i$ satisfying,

$$y_i\left(\langle \boldsymbol{w}_0, \boldsymbol{x}_i\rangle + b_0\right) \geq 1 - \xi_i \quad \forall i$$

So we can find $\lambda > 0$ so that,

$$y_i\left(\langle \lambda\boldsymbol{w}_0, \boldsymbol{x}_i\rangle + \lambda b_0\right) > \rho - \xi_i \quad \forall i$$

Where $\xi_i > 0$, $\rho > 0$.

So strict feasibility can be verified and we can invoke the KKT optimality condition. The Lagrangian is,

$$\mathcal{L}\left(\boldsymbol{w}, b, \rho, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{u}, \pi\right) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - v\rho + \sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\lambda_i\left(\rho - \xi_i - y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right)\right) - \sum_{i=1}^{N}u_i\xi_i - \pi\rho$$

The KKT condition is,

stationarity:

$$\boldsymbol{w} = \sum_{i=1}^{N}\lambda_i y_i \boldsymbol{x}_i \quad \sum_{i=1}^{N}\lambda_i y_i = 0 \quad \mathbf{1} = \boldsymbol{\lambda} + \boldsymbol{u} \quad \sum_{i=1}^{N}\lambda_i = \pi + v$$

feasibility:

$$y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right) \geq \rho - \xi_i,\ \xi_i \geq 0,\ \lambda_i \geq 0.\ u_i \geq 0\ \forall i \quad \rho \geq 0,\ \pi \geq 0$$

complementary slackness:

$$\lambda_i\left(\rho - \xi_i - y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\right) + b\right) = 0, \quad u_i\xi_i = 0\ \forall i \quad \pi\rho = 0$$

## 3.3 c

**Answers:**

the support vectors satisfy $y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right) = \rho - \xi_i$. So, $\lambda_i \neq 0$ for the support vectors. As $\rho > 0$, $\pi = 0 \Rightarrow \sum_{i=1}^{N}\lambda_i = v$ in terms of the KKT conditions.

For the outliers, $\xi_i \geq 0 \Rightarrow u_i = 0 \Rightarrow \lambda_i = 1$. $\lambda_i = 1$ for all the outliers. Let $N_o$ denote the number of outliers. $N_o \leq N$

$$N_o = \sum_{i=1}^{N_o} 1 \leq \sum_{i=1}^{N} \lambda_i = v$$

Where $\lambda_i$ denote value of $\lambda$ of all data points.

The support vectors contain points on the hyperplanes and outliers. For the points on the hyerplanes, $\xi_i = 0 \Rightarrow u_i > 0 \Rightarrow \lambda_i < 1$ in terms of KKT conditions.

Hence, $\lambda_i \leq 1$ for the support vectors. $\lambda_i = 0$ for other points. Let $N_s$ denote the number of support vectors.

$$N_s = \sum_{i=1}^{N_s} 1 \geq \sum_{i=1}^{N} \lambda_i = v$$

Where $\lambda_i$ denote value of $\lambda$ of all data points.

### 3.4 d

**Answers:**

As the training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ is linearly separable, there exists $\boldsymbol{w}_0$ and $b_0$ so that

$$y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b\right) \geq 0 \quad \forall i$$

So $\rho \neq 0$ in that $\rho - \xi_i \leq 0$ if $\rho = 0$. The constraint will be loose so that we cannot get right and optimal hyperplane.

So $\rho > 0$. Also, $v$ is an upper bound on the number of outliers bounded as proved above 3.3. If $v = \frac{1}{2}$, there exists no outliers. Hence, $\xi_i = 0$ for all data points.

The objective function will be,

$$\min_{\boldsymbol{w}, b, \rho} \frac{1}{2} \|\boldsymbol{w}\|_2^2 - v\rho \;\text{ s. t. } y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b\right) \geq \rho$$

Rewriting that,

$$\min_{\boldsymbol{w}, b, \rho} \frac{1}{2} \left\|\frac{\boldsymbol{w}}{\rho}\right\|_2^2 - \frac{v}{\rho} \;\text{ s. t. } y_i\left(\left\langle \frac{\boldsymbol{w}}{\rho}, \boldsymbol{x}_i \right\rangle + \frac{b}{\rho}\right) \geq 1$$

Let $\boldsymbol{w}' = \frac{\boldsymbol{w}}{\rho}$, $v' = \frac{v}{\rho}$, $b' = \frac{b}{\rho}$.

$$\min_{\boldsymbol{w}', b'} \frac{1}{2} \|\boldsymbol{w}'\|_2^2 - v' \;\text{ s. t. } y_i\left(\langle \boldsymbol{w}', \boldsymbol{x}_i \rangle + b'\right) \geq 1$$

We can remove $v'$, which does not influence minimization,

$$\min_{\boldsymbol{w}', b'} \frac{1}{2} \|\boldsymbol{w}'\|_2^2 \;\text{ s. t. } y_i\left(\langle \boldsymbol{w}', \boldsymbol{x}_i \rangle + b'\right) \geq 1 \quad \forall i$$

Now the objective function is consistent with hard-margin SVM.
Consequently,

If $\rho > 0$, this problem yields the same binary classifier as that of hard-margin SVM.

### 3.5 e

**Answers:**
Assume $(\boldsymbol{w}_*, b_*, \xi_*, \rho_*)$ is a global minimizer of problem,

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\rho} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \nu\rho + \sum_{i=1}^{N}\xi_i \quad \text{s.t. } y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right) \geq \rho - \xi_i,\ \xi_i \geq 0\ \forall\, i \quad \rho \geq 0$$

$$y_i\left(\langle \boldsymbol{w}_*, x_i\rangle + b_*\right) \geq \rho_* - \xi_i^* \quad \forall i$$

So,

$$y_i\left(\left\langle \frac{\boldsymbol{w}}{\rho}, x_i\right\rangle + \frac{b}{\rho}\right) \geq 1 - \frac{\xi_i}{\rho} \quad \forall i$$

Let $\boldsymbol{w}' = \frac{\boldsymbol{w}}{\rho}$, $b' = \frac{b}{\rho}$, $\xi' = \frac{\xi_i}{\rho}$

$$y_i\left(\langle \boldsymbol{w}', x_i\rangle + b'\right) \geq 1 - \xi' \quad \forall i$$

As $\rho_* > 0$, The objective function is,

$$\min_{\boldsymbol{w},b,\xi,\rho} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - v\rho + \sum_{i=1}^{N}\xi_i = \min_{\boldsymbol{w},b,\xi,\rho} \frac{1}{2}\left\|\frac{\boldsymbol{w}}{\rho}\right\|_2^2 + \rho^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\rho} - \frac{v}{\rho}$$

Let $v' = v/\rho$, the objective function is,

$$min_{\boldsymbol{w},b,\xi,\rho}\frac{1}{2}\|\boldsymbol{w}'\|_2^2 + \rho^{-1}\sum_{i=1}^{N}\xi' - v'$$

Assume $\rho$ is constant. Removing $v'$, which does not influence minimization,

$$min_{\boldsymbol{w},b,\xi,}\frac{1}{2}\|\boldsymbol{w}'\|_2^2 + \rho^{-1}\sum_{i=1}^{N}\xi' \quad y_i\left(\langle \boldsymbol{w}', x_i\rangle + b'\right) \geq 1 - \xi' \quad \forall i$$

This is soft-margin SVM with $C = \rho^{-1}$. As $(\boldsymbol{w}_*, b_*, \xi_*, \rho_*)$ is a global minimizer of original problem, the global minimizer $\{\boldsymbol{w}, b, \boldsymbol{\xi}\}$ of soft-margin SVM with $C = \rho_*^{-1}$ are $\frac{\boldsymbol{w}_*}{\rho_*}$,$\frac{b_*}{\rho_*}$ and $\frac{\boldsymbol{\xi}_*}{\rho_*}$ respectively.

## 4 Problem 4

### 4.1 a

The $\ell_2$ norm case is,

$$\mathcal{H} \doteq \{\boldsymbol{x} \to \langle \boldsymbol{w}, \boldsymbol{x}\rangle : \|\boldsymbol{w}\|_2 \leq 1\}$$

$$
\begin{aligned}
N * G(\boldsymbol{X}) &= \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_2 \leq 1} \langle \boldsymbol{X}\boldsymbol{w}, \boldsymbol{g} \rangle \\
&= \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_2 \leq 1} \sum_{i=1}^{N} g_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \\
&= \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_2 \leq 1} \left\langle \boldsymbol{w}, \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\rangle \\
&\leq \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \left\| \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\|_2
\end{aligned}
$$

using Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \left\| \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\|_2 &= \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \left( \left\| \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\|_2^2 \right)^{1/2} \\
&\leq \left( \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \left\| \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\|_2^2 \right)^{1/2}
\end{aligned}
$$

$g_1, g_2, \ldots, g_n$ are sampled from normal Gaussian distribution and independent.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \left\| \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\|_2^2 &= \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} \sum_{i,j} g_i g_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\
&= \sum_{i \neq j} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} g_i g_j + \sum_{i=1}^{N} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \mathbb{E}_{\boldsymbol{g} \sim_{iid} \mathsf{N}(0,1)} g_i^2 \\
&= \sum_{i=1}^{N} \|\boldsymbol{x}_i\|_2^2 \leq N \max_i \|\boldsymbol{x}_i\|_2^2
\end{aligned}
$$

Hence,

$$
G(\boldsymbol{X}) \leq \frac{\max_i \|\boldsymbol{x}_i\|_2}{\sqrt{N}}
$$

## 4.2 b

The $\ell_\infty$ norm case is,

$$
\mathcal{H} \doteq \{\boldsymbol{x} \to \langle \boldsymbol{w}, \boldsymbol{x} \rangle : \|\boldsymbol{w}\|_\infty \leq 1\}
$$

using Holder's inequality,

$$
\begin{aligned}
N * G(\boldsymbol{X}) &= \mathbb{E}_{\boldsymbol{g}\sim_{iid}\mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_\infty \le 1} \langle \boldsymbol{X}\boldsymbol{w}, \boldsymbol{g} \rangle \\
&= \mathbb{E}_{\boldsymbol{g}\sim_{iid}\mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_\infty \le 1} \sum_{i=1}^{N} g_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \\
&= \mathbb{E}_{\boldsymbol{g}\sim_{iid}\mathsf{N}(0,1)} \sup_{\|\boldsymbol{w}\|_2 \le 1} \left\langle \boldsymbol{w}, \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\rangle \\
&\le \mathbb{E}_{\boldsymbol{g}\sim_{iid}\mathsf{N}(0,1)} \left\| \sum_{i=1}^{N} g_i \boldsymbol{x}_i \right\|_1
\end{aligned}
$$

For each $j \in [d]$, let $v_j = (x_1, j \ldots x_N, j) \in \mathbb{R}^N$. Note that $\|\boldsymbol{v}_j\|_2 \le \sqrt{N} \max_i \|\boldsymbol{x}_i\|_1$. Let $V = \{\boldsymbol{v}_1, \ldots, -\boldsymbol{v}_1, \cdots - \boldsymbol{v}_n\}$. The right-hand side is $N\ G(V)$. Using Massart lemma,

$$
G(V) \le \max_i \|\boldsymbol{x}_i\|_1 \sqrt{2\log(2d)/N}
$$