## Appendix B. Supplementary Material

In the supplementary material, we will provide the proofs for DRO representation and asymptotic result for logistic regression, which were discussed in Theorem 2 and Theorem 4, in Section B.1 and Section B.2. In addition, we will provide the results under the high dimension setting for linear regression, where the number of predictors growth with the sample size, as a generalization of Theorem 3 in Section B.3.

### B.1. Proof of DRO for Logistic Regression

**Proof** [Proof for Theorem 2]By applying strong duality results for semi-infinity linear programming problem in Blanchet et al. (2016), we can write the worst case expected loss function as,

$$\begin{aligned}
&\sup_{P:D_c(P,P_n)\le\delta} \mathbb{E}_P\left[\log\left(1+\exp\left(-Y\beta^TX\right)\right)\right] \\
&= \min_{\gamma\ge 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^n \sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^Tu\right)\right) - \gamma\left\|X_i-u\right\|_{\alpha^{-1}\text{-}(q,t)}\right\}\right\}.
\end{aligned}$$

For each $i$, we can apply Lemma 1 in Shafieezadeh-Abadeh et al. (2015) and the dual norm result in Proposition 5 to deal with the inner optimization problem. It gives us,

$$\begin{aligned}
&\sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^Tu\right)\right) - \gamma\left\|X_i-u\right\|_{\alpha^{-1}\text{-}(q,t)}\right\} \\
&= \begin{cases} \log\left(1+\exp\left(-Y_i\beta^TX_i\right)\right) & \text{if} \quad \|\beta\|_{\alpha\text{-}(p,s)} \le \gamma, \\ \infty & \text{if} \quad \|\beta\|_{\alpha\text{-}(p,s)} > \gamma. \end{cases}
\end{aligned}$$

Moreover, since the outer player wishes to minimize, $\gamma$ will be chosen to satisfy $\gamma \ge \|\beta\|_{\alpha\text{-}(p,s)}$. We then conclude

$$\begin{aligned}
&\min_{\gamma\ge 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^n \sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^Tu\right)\right) - \gamma\left\|X_i-u\right\|_{\alpha^{-1}\text{-}(q,t)}\right\}\right\} \\
&= \min_{\gamma\ge\|\beta\|_{\alpha\text{-}(p,s)}} \left\{\delta\gamma + \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^TX_i\right)\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^TX_i\right)\right) + \delta\left\|\beta\right\|_{\alpha\text{-}(p,s)},
\end{aligned}$$

where the last equality is obtained by noting that the objective function is continuous and monotone increasing in $\gamma$, thus $\gamma = \|\beta\|_{\alpha\text{-}(p,s)}$ is optimal. Hence, we conclude the DRO formulation for GR-Lasso logistic regression. ■

### B.2. Proof of Optimal Selection of Regularization for Logistic Regression

**Proof** [Proof of Theorem 4]We can apply strong duality result for semi-infinite linear programming problem in Section B of Blanchet et al. (2016), and write the scaled RWP function evaluated at $\beta^*$ in the dual form as,

$$\sqrt{n}R_n\left(\beta^*\right) = \max_{\zeta}\left\{\zeta^T Z_n - \mathbb{E}_{P_n}\phi\left(X,Y,\beta^*,\zeta\right)\right\},$$

where $Z_n = \frac{1}{n}\sum_i^n \frac{Y_i X_i}{1+\exp\left(Y_i X_i^T \beta^*\right)}$ and

$$\phi\left(X,Y,\beta^*,\zeta\right) = \max_u\left\{Y\zeta^T\left(\frac{X}{1+\exp\left(YX^T\beta^*\right)} - \frac{u}{1+\exp\left(Yu^T\beta^*\right)}\right) - \|X-u\|_{\alpha^{-1}\text{-}(q,t)}\right\}.$$

We proceed as in our proof of Theorem 3 in this paper and also adapting the case $\rho = 1$ for Theorem 1 in Blanchet et al. (2016). We can apply Lemma 2 in Blanchet et al. (2016) and conclude that the optimizer $\zeta$ can be taken to lie within a compact set with high probability as $n \to \infty$. We can combine the uniform law of large number estimate as in Lemma 3 of Blanchet et al. (2016) and obtain

$$\sqrt{n}R_n\left(\beta\right) = \max_{\zeta}\left\{\zeta^T Z_n - \mathbb{E}_P\phi\left(X,Y,\beta^*,\zeta\right)\right\} + o_P(1).$$

For the optimization problem defining $\phi\left(\cdot\right)$, we can apply results in Lemma 5 in Section A.3 of Blanchet et al. (2016), we know, for any choice of $\tilde{\zeta}$, if,

$$\operatorname*{ess\,sup}_{X,Y}\left\|\tilde{\zeta}^T\frac{y\left(1+\exp\left(YX^T\beta^*\right)\right)I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2}\right\|_{\alpha\text{-}(p,s)} > 1,$$

we have $\mathbb{E}\left[\phi\left(X,Y,\beta^*,\tilde{\zeta}\right)\right] = \infty$. Since the outer optimization problem is maximization over $\zeta$, the player will restrict $\zeta$ within the set $A$, where

$$A = \left\{\zeta \in \mathbb{R}^d : \operatorname*{ess\,sup}_{X,Y}\left\|\zeta^T\frac{y\left(1+\exp\left(YX^T\beta^*\right)\right)I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2}\right\|_{\alpha\text{-}(p,s)} \leq 1\right\}.$$

Moreover, it is easy to calculate, if $\zeta \in A$, we have $\mathbb{E}[\phi\left(X,Y,\beta^*,\zeta\right)] = 0$, thus we have the scaled RWP function has the following estimate, as $n \to \infty$

$$\sqrt{n}R_n\left(\beta\right) = \max_{\zeta\in A}\zeta^T Z_n + o_P(1).$$

Letting $n \to \infty$, we obtain the exact asymptotic result.

For the stochastic upper bound, let us recall for the definition of the set $A$ and consider the following estimate

$$\begin{aligned}
&\left\|\zeta^T\frac{y\left(1+\exp\left(YX^T\beta^*\right)\right)I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2}\right\|_{\alpha\text{-}(p,s)} \\
&\geq \left\|\frac{Y\zeta}{1+\exp\left(Y\beta^{*\,T}X\right)}\right\|_{\alpha\text{-}(p,s)} - \left\|\frac{\zeta^T X\beta^*}{\left(1+\exp\left(Y\beta^{*\,T}X\right)\right)^2}\right\|_{\alpha\text{-}(p,s)} \\
&\geq \left(\frac{1}{1+\exp\left(Y\beta^{*\,T}X\right)} - \frac{\|X\|_{\alpha^{-1}\text{-}(q,t)}\|\beta^*\|_{\alpha\text{-}(p,s)}}{\left(1+\exp\left(Y\beta^{*\,T}X\right)\right)\left(1+\exp\left(-Y\beta^{*\,T}X\right)\right)}\right)\|\zeta\|_{\alpha\text{-}(p,s)}.
\end{aligned}$$

The first inequality is due to application of triangle inequality in Proposition 5, while the second estimate follows from Hölder's inequality and $Y \in \{-1, +1\}$. Since we assume positive probability density for the predictor $X$, we can argue that, if $\|\zeta\|_{\alpha\text{-}(p,s)} = (1-\epsilon)^{-2} > 1$ and $\epsilon > 0$ is chosen arbitrarily small, we can conclude from the above estimate that, we have

$$\left\| \zeta^T \frac{y\left(1+\exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} > 1.$$

Thus, we proved the claim that $A \subset \left\{\zeta, \|\zeta\|_{\alpha\text{-}(p,s)} \leq 1\right\}$. The stochastic upper bound is derived by replacing $A$ by $\left\{\zeta, \|\zeta\|_{\alpha\text{-}(p,s)} \leq 1\right\}$, i.e.

$$L_3 = \sup_{\zeta\in A} \zeta^T Z \leq \sup_{\|\zeta\|_{\alpha\text{-}(p,s)}\leq 1} \zeta^T Z = \|Z\|_{\alpha^{-1}\text{-}(q,t)},$$

where the final estimation is due to dual norm structure in Proposition 5. Since we know, $\frac{1}{1+\exp(YX^T\beta)} \leq 1$, it is easy to argue, $Var(\tilde{Z}) - Var(Z)$ is positive semidefinite, thus, we know $\|Z\|_{\alpha^{-1}\text{-}(q,t)}$ is stochastic dominated by $L_4 := \left\|\tilde{Z}\right\|_{\alpha^{-1}\text{-}(q,t)}$. Hence, we obtain $L_3 \leq L_4$. ■

### B.3. Technical Results for Optimal Regularization in GSRL for High Dimensional Linear Regression

We conclude the supplementary material by exploring the behavior of the optimal distributional uncertainty (in the sense of optimality presented in Section 3) as the dimension increases. This is an analog of the high-dimension result for SR-Lasso as Theorem 6 in Blanchet et al. (2016).

**Theorem 6 (RWP Function Asymptotic Results for High-dimension)** *Suppose that assumptions in Theorem 3 hold and select $p = 2$, $s = 1$ let us write$\sqrt{g} = \left(\sqrt{g_1}, \ldots, \sqrt{g_{\bar{d}}}\right)^T$ (so $\alpha_j = \sqrt{g_j}$) and $\tilde{g}^{-1/2} = \left(1/\sqrt{g_1}, \ldots, 1/\sqrt{g_{\bar{d}}}\right)^T$ respectively. Moreover, let us define $C(n,d)$*

$$C(n,d) = \frac{\mathbb{E}\|X\|_{\sqrt{\bar{d}}\text{-}(2,1)}}{\sqrt{n}} = \frac{\mathbb{E}\left[\max_{i=1}^{\bar{d}} \sqrt{g_i}\,\|X(G_i)\|_2\right]}{\sqrt{n}}.$$

*Assume that largest eigenvalue of $\Sigma$ is of order $o\left(nC(n,d)^2\right)$, that $\beta_*$ satisfies a weak sparsity condition, namely, $\|\beta_*\|_{\sqrt{\tilde{g}}\text{-}(2,1)} = o(1/C(n,d))$. Then,*

$$nR_n(\beta_*) \lesssim_D \frac{\|Z_n\|_{\tilde{g}^{-1/2}\text{-}(2,\infty)}}{Var(|e|)},$$

*as $n, d \to \infty$, where $Z_n := n^{-1/2}\sum_{i=1}^n e_i X_i$.*

**Proof** For linear regression model with square loss function, the RWP function is defined as in equation (10). By considering the cost function as in Theorem 1 and applying the strong duality results in the Appendix of Blanchet et al. (2016), we can write the scaled RWP function in the dual form as,

$$\begin{aligned} nR_n\left(\beta_*\right) = & \sup_{\zeta} \Big\{ -\zeta^T Z_n \\ & - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\Delta} \{ e_i \zeta^T \Delta - \left(\beta_*^T \Delta\right)\left(\zeta^T X_i\right) - \left(\sqrt{n} \left\|\Delta\right\|^2_{\tilde{g}^{-1/2}\text{-}(2,\infty)} + \left(\beta_*^T \Delta\right)\left(\zeta^T \Delta\right)\right)\}\Big\}. \end{aligned}$$

For each $i-$th inner optimization problem, we can apply Hölder inequality in Proposition 5 for the term $\left(\beta_*^T\Delta\right)\left(\zeta^T\Delta\right)$, we have an upper bound for the scaled RWP function, i.e.

$$\begin{aligned} nR_n\left(\beta_*\right) \leq & \sup_{\zeta} \Big\{ -\zeta^T Z_n \\ & - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\Delta} \{ \left(e_i \zeta - \left(\zeta^T X_i\right)\beta_*\right)^T \Delta - \sqrt{n} \left(1 - \frac{\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} \left\|\zeta\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)}}{\sqrt{n}}\right) \left\|\Delta\right\|^2_{\tilde{g}^{-1/2}\text{-}(q,t)} \}\Big\}. \end{aligned}$$

Since the coefficients for each inner optimization problem is negative and we can get an upper bound for RWP function if we do not fully optimize the inner optimization problem. For each $i$, let us take $\Delta$ to the direction satisfying the Hölder inequality in Property 5 for the term $\left(e_i\zeta - \left(\zeta^T X_i\right)\beta_*\right)^T \Delta$ and only optimize the magnitude of $\Delta$, for simplicity let us denote $\gamma = \left\|\Delta\right\|_{\tilde{g}^{-1/2}\text{-}(q,t)}$.
We have,

$$\begin{aligned} nR_n\left(\beta_*\right) \leq & \sup_{\zeta} \Big\{ -\zeta^T Z_n \\ & - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\gamma} \{ \gamma \left\|e_i \zeta - \left(\zeta^T X_i\right)\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} - \sqrt{n} \left(1 - \frac{\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} \left\|\zeta\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)}}{\sqrt{n}}\right) \gamma^2 \}\Big\}. \end{aligned}$$

For each inner optimization problem it is of quadratic form in $\gamma$, especially, when $n$ is large the coefficients for the second order term will be negative, thus, as $n \to \infty$, we can solve the inner optimization problem and obtain,

$$\begin{aligned} & nR_n\left(\beta_*\right) \\ & \leq \sup_{\zeta} \Big\{ -\zeta^T Z_n - \frac{1}{4\left(1 - \left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} \left\|\zeta\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} n^{-1/2}\right)} \frac{1}{n} \sum_{i=1}^{n} \left\|e_i\zeta - \left(\zeta^T X_i\right)\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)} \Big\} \\ & = \sup_{a \geq 0} \sup_{\zeta : \left\|\zeta\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} = 1} \Big\{ -a\zeta^T Z_n - \frac{a^2}{4\left(1 - \left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} a n^{-1/2}\right)} \frac{1}{n} \sum_{i=1}^{n} \left\|e_i\zeta - \left(\zeta^T X_i\right)\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)} \Big\}. \end{aligned}$$

The equality above is due to changing to polar coordinate for the ball under $\sqrt{\tilde{g}}\text{-}(p,s)$ norm. For the first term, $\zeta^T Z_n$, when $\left\|\zeta\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} = 1$, we can apply Hölder inequality again, i.e. $\left|\zeta^T Z_n\right| \leq \left\|Z_n\right\|_{\tilde{g}^{-1/2}\text{-}(q,t)}$. Then, only the second term in the previous display involves the direction of $\zeta$, thus we can have

$$\begin{aligned} nR_n\left(\beta_*\right) \leq & \sup_{a\geq 0}\Big\{ a\left\|Z_n\right\|_{\tilde{g}^{-1/2}\text{-}(q,t)} \\ & - \frac{a^2}{4\left(1-\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} a n^{-1/2}\right)} \inf_{\zeta:\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(p,s)}=1} \frac{1}{n}\sum_{i=1}^{n}\left\|e_i\zeta-\left(\zeta^T X_i\right)\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)}\Big\}. \end{aligned}$$

By the weak sparsity assumption, we have $\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)} n^{-1/2}\to 0$ as $n\to\infty$, the supremum over $a$ is attained at

$$a_* = \frac{2\left\|Z_n\right\|_{\tilde{g}^{-1/2}\text{-}(q,t)}}{\inf_{\zeta:\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(p,s)}=1} \frac{1}{n}\sum_{i=1}^{n}\left\|e_i\zeta-\left(\zeta^T X_i\right)\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)}} + o(1),$$

as $n\to\infty$. Therefore, we have the upper bound estimator for the scaled RWP function,

$$nR_n\left(\beta_*\right) \leq \frac{\left\|Z_n\right\|^2_{\tilde{g}^{-1/2}\text{-}(q,t)}}{\inf_{\zeta:\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(p,s)}=1} \frac{1}{n}\sum_{i=1}^{n}\left\|e_i\zeta-\left(\zeta^T X_i\right)\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)}} + o_p(1). \tag{18}$$

To get the final result, we try to find a lower bound for the infimum in the denominator. For the objective function in the denominator, since we optimize on the surface $\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(p,s)}=1$, and due to the triangle inequality analysis in Proposition 5, we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left\|e_i\zeta-\left(\zeta^T X_i\right)\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)} \geq \frac{1}{n}\sum_{i=1}^{n}\left(\left|e_i\right|\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(p,s)} - \left|\zeta^T X_i\right|\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)}\right)^2 \\ =& \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right|^2 + \left\|\beta_*\right\|^2_{\sqrt{\tilde{g}}\text{-}(p,s)}\frac{1}{n}\sum_{i=1}^{n}\left|\zeta^T X_i\right|^2 - 2\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)}\mathbb{E}\left[\left|e_i\right|\right]\frac{1}{n}\sum_{i=1}^{n}\left|\zeta^T X_i\right| - \epsilon_n\left(\zeta\right), \end{aligned}$$

where $\epsilon_n\left(\zeta\right) = 2\left\|\beta_*\right\|_{\sqrt{\tilde{g}}\text{-}(p,s)}\frac{1}{n}\sum_{i=1}^{n}\left(\left|e_i\right|-\mathbb{E}\left[\left|e_i\right|\right]\right)$. Let us denote the pseudo error to be $\tilde{e}_i = \left|e_i\right|-\mathbb{E}\left[\left|e_i\right|\right]$, which has mean zero and $Var\left[\tilde{e}_i\right]\leq Var\left[e_i\right]$. Since $e_i$ is independent of $X_i$ we have that

$$\begin{aligned} &\mathbb{E}\left[\tilde{e}_i\left|\zeta^T X_i\right|\right]=0, \\ &Var\left[\tilde{e}_i\left|\zeta^T X_i\right|\right] = Var\left[\tilde{e}_i\right]\zeta^T\Sigma\zeta \leq Var\left[e_i\right]\zeta^T\Sigma\zeta. \end{aligned}$$

By our assumptions on the eigenstructure of $\Sigma$, i.e. $\lambda_{\max}\left(\Sigma\right)=o\left(nC(n,d)^2\right)$, for the case $p=2$ and $s=1$, we have

$$\sup_{\zeta:\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(2,1)}=1}\zeta^T\Sigma\zeta \leq \sup_{\zeta:\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(2,1)}=1}\lambda_{\max}\left(\Sigma\right)\|\zeta\|_2 \leq \lambda_{\max}\left(\Sigma\right) = o\left(nC(n,d)^2\right).$$

Then, we have the variance of $\frac{1}{n}\sum_{i=1}^{n}\left|\zeta^T X_i\right|$ is of order $o\left(C(n,d)^2\right)$ uniformly on $\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(p,s)}=1$. Combining this estimate with the weak sparsity assumption that we have imposed, we have

$$\epsilon_n\left(\zeta\right)=o_p(1).$$

Since the estimate is uniform over $\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(2,1)} = 1$, we have that for $n$ sufficiently large,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \left\| e_i\zeta - \left(\zeta^T X_i\right)\beta^*\right\|_{\sqrt{\tilde{g}}\text{-}(2,1)}^2 \\
&\geq \frac{1}{n}\sum_{i=1}^{n} |e_i|^2 - \left(\mathbb{E}\left[|e_i|\right]\right)^2 + \inf_{\zeta:\|\zeta\|_{\sqrt{\tilde{g}}\text{-}(2,1)}=1} \left( \|\beta^*\|_{\sqrt{\tilde{g}}\text{-}(2,1)} \frac{1}{n}\sum_{i=1}^{n} \left|\zeta^T X_i\right| - \mathbb{E}\left[|e_i|\right]\right)^2 + o_p(1) \\
&\geq Var_n\left[|e_i|\right] + o_p(1).
\end{aligned}$$

Combining the above estimate and equation (18), when $p = q = 2$, $s = 1$ and $t = \infty$, we have that

$$nR_n\left(\beta^*\right) \leq \frac{\|Z_n\|_{\tilde{g}^{-1/2}\text{-}(2,\infty)}^2}{Var\left[|e|\right]} + o_p(1),$$

as $n \to \infty$. ■