A Proof of Theorem 2

Proof. To bound the expectation of Ψ, we first derive an upper bound on Ψ that does not depend on $\mathbf{w}$. Let $\widehat{D}$ denote the empirical distribution related to the sample S. In the following, the expectations with respect to $\widehat{D}$ assume a fixed $\mathbf{w}$. We use directly the properties of $\mathbf{w}$ and $\mathbf{w}^\star$ as minimizers of F_S and $F^\star$. Writing $\nabla F_S(\mathbf{w}) = 0$ and $\nabla F^\star(\mathbf{w}^\star) = 0$ and taking the difference yield immediately

$$\mathbf{w}^\star - \mathbf{w} = -\frac{1}{2\lambda}\big[\mathop{\mathrm{E}}_{D}[\nabla L_z(\mathbf{w}^\star)] - \mathop{\mathrm{E}}_{\widehat{D}}[\nabla L_z(\mathbf{w})]\big] \tag{16}$$

$$= -\frac{1}{2\lambda}\big[\mathop{\mathrm{E}}_{D}[\nabla L_z(\mathbf{w}^\star)] - \mathop{\mathrm{E}}_{\widehat{D}}[\nabla L_z(\mathbf{w}^\star)] + \mathop{\mathrm{E}}_{\widehat{D}}[\nabla L_z(\mathbf{w}^\star)] - \mathop{\mathrm{E}}_{\widehat{D}}[\nabla L_z(\mathbf{w})]\big]. \tag{17}$$

Taking the inner product with $(\mathbf{w}^\star - \mathbf{w})$ and using the convexity of L_z, which implies $\big(\mathrm{E}_{\widehat{D}}[\nabla L_z(\mathbf{w}^\star) - \nabla L_z(\mathbf{w})]\big) \cdot (\mathbf{w}^\star - \mathbf{w}) \geq 0$, lead to

$$\|\mathbf{w}^\star - \mathbf{w}\|^2 \leq -\frac{1}{2\lambda}\big[\mathop{\mathrm{E}}_{D}[\nabla L_z(\mathbf{w}^\star)] - \mathop{\mathrm{E}}_{\widehat{D}}[\nabla L_z(\mathbf{w}^\star)]\big] \cdot (\mathbf{w}^\star - \mathbf{w}) \tag{18}$$

$$\leq \frac{1}{2\lambda}\big\|\mathop{\mathrm{E}}_{D}[\nabla L_z(\mathbf{w}^\star)] - \mathop{\mathrm{E}}_{\widehat{D}}[\nabla L_z(\mathbf{w}^\star)]\big\| \, \|\mathbf{w}^\star - \mathbf{w}\|. \tag{19}$$

Thus, we can write $2\lambda\|\mathbf{w}^\star - \mathbf{w}\| \leq \|\frac{1}{m}\sum_{i=1}^m \mathbf{Z}_i\|$, where $\mathbf{Z} = \nabla L_z(\mathbf{w}^\star) - \mathrm{E}[\nabla L_z(\mathbf{w}^\star)]$ and $\mathbf{Z}_i = \nabla L_{z_i}(\mathbf{w}^\star) - \mathrm{E}[\nabla L_z(\mathbf{w}^\star)]$, for all $i \in [1, m]$. Note that this upper bound does not depend on $\mathbf{w}$, which makes it easier to analyze its expectation with respect to the choice of S.

By Jensen's inequality, $2\lambda\,\mathrm{E}[\Psi] \leq \mathrm{E}\left[\|\frac{1}{m}\sum_{i=1}^m \mathbf{Z}_i\|\right] \leq \sqrt{\mathrm{E}[\|\frac{1}{m}\sum_{i=1}^m \mathbf{Z}_i\|^2]}$. Using the fact that the variables $\mathbf{Z}_i$s are i.i.d. with $\mathrm{E}[\mathbf{Z}_i] = 0$, we obtain

$$\mathrm{E}\left[\|\frac{1}{m}\sum_{i=1}^m \mathbf{Z}_i\|^2\right] = \frac{1}{m^2}\big[\sum_{i=1}^m \mathrm{E}[\|\mathbf{Z}_i\|^2] + \sum_{i \neq j} \mathrm{E}[\mathbf{Z}_i] \cdot \mathrm{E}[\mathbf{Z}_j]\big] = \frac{1}{m}\,\mathrm{E}[\|\mathbf{Z}_1\|^2] = \frac{1}{m}\,\mathrm{Var}(\mathbf{Z}_1).$$

Using the expression of $\nabla L_z(\mathbf{w}^\star)$ already derived in the proof of Theorem 1 and the elementary fact that if $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent and identically distributed, then $\mathrm{Var}(\mathbf{Z}_1) = 1/2\,\mathrm{E}[(\mathbf{Z}_1 - \mathbf{Z}_2)^2]$, this shows that $\mathrm{E}[\Psi] \leq \frac{1}{2\lambda}\sqrt{\frac{1}{2}\frac{(4R)^2}{m}} = \frac{2R}{\lambda\sqrt{2m}}$. ∎