A. Detailed Proofs

In this section, we state the proofs of all the theorems and claims in our manuscript.

A.1. Proof of Theorem 1

Proof. By optimality of the empirical minimizer, we have:

$$\begin{aligned}\widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) &\leq \widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}^*_n) + \lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}^*_n) + \xi \\ &\leq \widehat{\mathcal{L}}_n(\boldsymbol{\theta}^*) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \xi\end{aligned}$$

or equivalently $\widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) - \widehat{\mathcal{L}}_n(\boldsymbol{\theta}^*) \leq -\lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \xi$. By Assumption A and B, and by conveniently setting $\lambda_n = \alpha \varepsilon_{n,\delta}$ for $\alpha \geq 1$:

$$\begin{aligned}\mathcal{L}(\widehat{\boldsymbol{\theta}}_n) - \mathcal{L}(\boldsymbol{\theta}^*) &\leq \widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) - \widehat{\mathcal{L}}_n(\boldsymbol{\theta}^*) + \varepsilon_{n,\delta} c(\widehat{\boldsymbol{\theta}}_n) + \varepsilon_{n,\delta} c(\boldsymbol{\theta}^*) \\ &\leq -\lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \varepsilon_{n,\delta} c(\widehat{\boldsymbol{\theta}}_n) + \varepsilon_{n,\delta} c(\boldsymbol{\theta}^*) + \xi \\ &\leq -\lambda_n r(c(\widehat{\boldsymbol{\theta}}_n)) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \varepsilon_{n,\delta} c(\widehat{\boldsymbol{\theta}}_n) + \varepsilon_{n,\delta} c(\boldsymbol{\theta}^*) + \xi \\ &= \varepsilon_{n,\delta}(-\alpha r(c(\widehat{\boldsymbol{\theta}}_n)) + c(\widehat{\boldsymbol{\theta}}_n)) + \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi \\ &\leq \varepsilon_{n,\delta}(-r(c(\widehat{\boldsymbol{\theta}}_n)) + c(\widehat{\boldsymbol{\theta}}_n)) + \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi \\ &\leq \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi\end{aligned}$$

□

A.2. Proof of Theorem 2

Proof. By Assumption C for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$ and by Theorem 1, we have $b(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_\infty) \leq \mathcal{L}(\widehat{\boldsymbol{\theta}}_n) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi$. Since the function b is nondecreasing, its inverse function $b^\dagger$ (as defined in eq.(8)) exists and we prove our claim. □

A.3. Proof of Theorem 3

Proof. For clarity, we remove the dependence of $\widehat{\boldsymbol{\theta}}_n$ with respect to the number of samples n. That is, $\widehat{\boldsymbol{\theta}} \equiv \widehat{\boldsymbol{\theta}}_n$. By Theorem 2, we have $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq b^\dagger(\varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi) \equiv \tau$. Therefore, for all i we have $|\widehat{\theta}_i - \theta^*_i| \leq \tau$. Next, we analyze the three possible cases.

Case 1: $\theta^*_i = 0 \Rightarrow \widetilde{\theta}_i = 0$. Assume that $i \notin \mathcal{S}(\boldsymbol{\theta}^*)$. Since $\theta^*_i = 0$, we have $|\widehat{\theta}_i - \theta^*_i| = |\widehat{\theta}_i| \leq \tau$. Therefore, $\widetilde{\theta}_i = h_i(\widehat{\boldsymbol{\theta}}, \tau) = \widehat{\theta}_i 1[|\widehat{\theta}_i| > \tau] = 0$.

Case 2: $\theta^*_i > 2\tau \Rightarrow \widetilde{\theta}_i > \tau$. Assume that $i \in \mathcal{S}(\boldsymbol{\theta}^*)$ and $\theta^*_i > 2\tau$. Since $|\widehat{\theta}_i - \theta^*_i| \leq \tau$, we have $\widehat{\theta}_i \geq \theta^*_i - \tau > 2\tau - \tau = \tau$. Therefore, $\widetilde{\theta}_i = h_i(\widehat{\boldsymbol{\theta}}, \tau) = \widehat{\theta}_i 1[|\widehat{\theta}_i| > \tau] = \widehat{\theta}_i > \tau$.

Case 3: $\theta^*_i < -2\tau \Rightarrow \widetilde{\theta}_i < -\tau$. Assume that $i \in \mathcal{S}(\boldsymbol{\theta}^*)$ and $\theta^*_i < -2\tau$. Since $|\widehat{\theta}_i - \theta^*_i| \leq \tau$, we have $\widehat{\theta}_i \leq \theta^*_i + \tau < -2\tau + \tau = -\tau$. Therefore, $\widetilde{\theta}_i = h_i(\widehat{\boldsymbol{\theta}}, \tau) = \widehat{\theta}_i 1[|\widehat{\theta}_i| > \tau] = \widehat{\theta}_i < -\tau$. □

A.4. Proof of Claim i

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}) = -\langle \widehat{\mathbf{T}}_n - \mathbf{T}, \boldsymbol{\theta} \rangle$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\langle \widehat{\mathbf{T}}_n - \mathbf{T}, \boldsymbol{\theta} \rangle| \\ &\leq \|\widehat{\mathbf{T}}_n - \mathbf{T}\|_* \|\boldsymbol{\theta}\| \\ &\leq \varepsilon_{n,\delta} \|\boldsymbol{\theta}\|\end{aligned}$$

□

A.5. Proof of Claim ii

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_i t(y^{(i)})\langle \mathbf{x}^{(i)},\boldsymbol{\theta}\rangle - \frac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\langle\mathbf{x}^{(i)},\boldsymbol{\theta}\rangle)$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall\boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta})| &= |\tfrac{1}{n}\textstyle\sum_i t(y^{(i)})\langle \mathbf{x}^{(i)},\boldsymbol{\theta}\rangle - \tfrac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\langle\mathbf{x}^{(i)},\boldsymbol{\theta}\rangle| \\ &= |\langle\tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\mathbf{x}^{(i)},\boldsymbol{\theta}\rangle| \\ &\leq \|\tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\mathbf{x}^{(i)}\|_*\|\boldsymbol{\theta}\| \\ &\leq \varepsilon_{n,\delta}\|\boldsymbol{\theta}\|\end{aligned}$$

□

A.6. Proof of Claim iii

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_{ij} t(x_{ij})\theta_{ij} - \frac{1}{n}\sum_{ij}\mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)]\theta_{ij})$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall\boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta})| &= |\tfrac{1}{n}\textstyle\sum_{ij} t(x_{ij})\theta_{ij} - \tfrac{1}{n}\sum_{ij}\mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)]\theta_{ij}| \\ &= |\tfrac{1}{n}\textstyle\sum_{ij} (t(x_{ij}) - \mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)])\theta_{ij}| \\ &\leq \|\tfrac{1}{n}(t(x_{11}) - \mathbb{E}_{x\sim\mathcal{D}_{11}}[t(x)],\ldots,t(x_{n_1n_2}) - \mathbb{E}_{x\sim\mathcal{D}_{n_1n_2}}[t(x)])\|_*\|\boldsymbol{\theta}\| \\ &\leq \varepsilon_{n,\delta}\|\boldsymbol{\theta}\|\end{aligned}$$

□

A.7. Proof of Claim iv

Proof. Let K be the Lipschitz constant of f. Without loss of generality, assume that $f(0)=0$ (this can be accomplished by adding a constant factor to f). Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{ij} f(x_{ij}\theta_{ij}) - \frac{1}{n}\sum_{ij}\mathbb{E}_{x\sim\mathcal{D}_{ij}}[f(x\theta_{ij})]$. Recall that $x_{ij}\in\{-1,+1\}$. We have:

$$\begin{aligned}&(\forall\boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta})| = |\tfrac{1}{n}\textstyle\sum_{ij} (f(x_{ij}\theta_{ij}) - \mathbb{E}_{x\sim\mathcal{D}_{ij}}[f(x\theta_{ij})])| \\ &\qquad = |\tfrac{1}{n}\textstyle\sum_{ij} (1[x_{ij}=+1]f(\theta_{ij}) + 1[x_{ij}=-1]f(-\theta_{ij}) - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=+1]f(\theta_{ij}) - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=-1]f(-\theta_{ij}))| \\ &\qquad = |\tfrac{1}{n}\textstyle\sum_{ij} ((1[x_{ij}=+1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=+1])f(\theta_{ij}) + (1[x_{ij}=-1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=-1])f(-\theta_{ij}))| \\ &\qquad \leq \tfrac{1}{n}\textstyle\sum_{ij} (|1[x_{ij}=+1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=+1]|\,|f(\theta_{ij})| + |1[x_{ij}=-1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=-1]|\,|f(-\theta_{ij})|) \\ &\qquad \leq \tfrac{1}{n}\textstyle\sum_{ij} (K|\theta_{ij}| + K|\theta_{ij}|) \\ &\qquad = \tfrac{2K}{n}\|\boldsymbol{\theta}\|_1\end{aligned}$$

□

A.8. Proof of Claim v

Proof. First, we represent the function $\theta:\mathcal{X}\to\mathbb{R}$ by using the infinitely dimensional orthonormal basis. That is, $\theta(\mathbf{x}) = \sum_{j=1}^{\infty}\nu_j^{(\theta)}\psi_j(\mathbf{x}) = \langle\boldsymbol{\nu}^{(\theta)},\boldsymbol{\psi}(\mathbf{x})\rangle$, where $\boldsymbol{\nu}^{(\theta)} = (\nu_1^{(\theta)},\ldots,\nu_\infty^{(\theta)})$. In the latter, the superindex (θ) allows for associating the infinitely dimensional coefficient vector $\boldsymbol{\nu}$ with the original function θ. Then, we define the norm of the function θ with respect to the infinitely dimensional orthonormal basis. That is, $\|\theta\| = \|\boldsymbol{\nu}^{(\theta)}\|$.

Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\theta)-\mathcal{L}(\theta) = -(\frac{1}{n}\sum_i t(y^{(i)})\theta(\mathbf{x}^{(i)}) - \frac{1}{n}\sum_i\mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\theta(\mathbf{x}^{(i)}))$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall\theta)\ |\widehat{\mathcal{L}}_n(\theta)-\mathcal{L}(\theta)| &= |\tfrac{1}{n}\textstyle\sum_i t(y^{(i)})\theta(\mathbf{x}^{(i)}) - \tfrac{1}{n}\sum_i\mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\theta(\mathbf{x}^{(i)})| \\ &= |\tfrac{1}{n}\textstyle\sum_i t(y^{(i)})\langle\boldsymbol{\psi}(\mathbf{x}^{(i)}),\boldsymbol{\nu}^{(\theta)}\rangle - \tfrac{1}{n}\sum_i\mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\langle\boldsymbol{\psi}(\mathbf{x}^{(i)}),\boldsymbol{\nu}^{(\theta)}\rangle| \\ &= |\langle\tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\boldsymbol{\psi}(\mathbf{x}^{(i)}),\boldsymbol{\nu}^{(\theta)}\rangle| \\ &\leq \|\tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\boldsymbol{\psi}(\mathbf{x}^{(i)})\|_*\|\boldsymbol{\nu}^{(\theta)}\| \\ &\leq \varepsilon_{n,\delta}\|\theta\|\end{aligned}$$

□

A.9. Proof of Claim vi

Proof. Let $\| \cdot \|_*$ be the dual norm of $\| \cdot \|$. Let $\mathcal{C}^{(j,\boldsymbol{\theta})} = \{\mathbf{x} \in \mathcal{X} \mid j = \arg\min_k -\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}^{(k)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(k)})\}$. Note that $\mathcal{C}^{(1,\boldsymbol{\theta})}, \dots, \mathcal{C}^{(\infty,\boldsymbol{\theta})}$ define a partition of $\mathcal{X}$. We can rewrite the empirical loss as follows $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \sum_j \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}](-\langle \mathbf{t}(\mathbf{x}^{(i)}), \boldsymbol{\theta}^{(j)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(j)}))$. Similarly, the expected loss can be written as $\mathcal{L}(\boldsymbol{\theta}) = \sum_j \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}](-\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}^{(j)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(j)}))]$.

By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\textstyle\sum_j (\frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\langle \mathbf{t}(\mathbf{x}^{(i)}), \boldsymbol{\theta}^{(j)} \rangle - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}^{(j)} \rangle])| \\ &= |\textstyle\sum_j \langle \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\mathbf{t}(\mathbf{x})], \boldsymbol{\theta}^{(j)} \rangle| \\ &\leq \textstyle\sum_j |\langle \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\mathbf{t}(\mathbf{x})], \boldsymbol{\theta}^{(j)} \rangle| \\ &\leq \textstyle\sum_j \| \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}]\mathbf{t}(\mathbf{x})]\|_* \|\boldsymbol{\theta}^{(j)}\|\end{aligned}$$

By assumption, for all partitions $\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(\infty)}$ of $\mathcal{X}$, the dual norm fulfills $(\forall j)\ \| \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{X}^{(j)}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{X}^{(j)}]\mathbf{t}(\mathbf{x})]\|_* \leq \varepsilon_{n,\delta}$. Therefore:

$$(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| \leq \varepsilon_{n,\delta} \textstyle\sum_j \|\boldsymbol{\theta}^{(j)}\|$$

□

A.10. Proof of Claim vii

Proof. By Pinsker's inequality and Theorem 5 of (Maurer, 2004) which assumes $n \geq 8$, we have:

$$\begin{aligned}(\forall \theta)\ |\widehat{\mathcal{L}}_n(\theta) - \mathcal{L}(\theta)| &\leq \sqrt{\tfrac{1}{2}(\widehat{\mathcal{L}}_n(\theta) \log \tfrac{\widehat{\mathcal{L}}_n(\theta)}{\mathcal{L}(\theta)} + (1 - \widehat{\mathcal{L}}_n(\theta)) \log \tfrac{1 - \widehat{\mathcal{L}}_n(\theta)}{1 - \mathcal{L}(\theta)})} \\ &\leq \sqrt{\tfrac{1}{2n}(KL(\theta||\theta^{(0)}) + \log \tfrac{2\sqrt{n}}{\delta})} \\ &\leq \sqrt{\tfrac{1}{2n} \max(1, \log \tfrac{2\sqrt{n}}{\delta})} \sqrt{KL(\theta||\theta^{(0)}) + 1} \\ &\leq \sqrt{\tfrac{1}{2n} \log \tfrac{2\sqrt{n}}{\delta} (KL(\theta||\theta^{(0)}) + 1)}\end{aligned}$$

□

A.11. Proof of Claim viii

Proof. The expected loss $\mathcal{L}$ is strongly convex with parameter ν if and only if:

$$(\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{H}, \mathbf{g} \in \tfrac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_1))\ \mathcal{L}(\boldsymbol{\theta}_2) - \mathcal{L}(\boldsymbol{\theta}_1) \geq \langle \mathbf{g}, \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle + \tfrac{\nu}{2}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2$$

First, we set $\boldsymbol{\theta}_1 = \boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$. Next, note that the subdifferential vanishes at $\boldsymbol{\theta}^*$, that is $\mathbf{0} \in \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^*)$. Therefore:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) &\geq \tfrac{\nu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \\ &\geq \tfrac{\nu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty^2\end{aligned}$$

which proves our first claim. Our second claim regarding twice continuously differentiable $\mathcal{L}$ is well-known in the calculus literature. □

A.12. Proof of Claim ix

Proof. Note that by the definition of $\mathcal{M}(\boldsymbol{\theta})$, we have:

$$(\forall \boldsymbol{\theta} \in \mathcal{H}, 0 \leq \gamma \leq \mathcal{M}(\boldsymbol{\theta}))\ \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq \gamma(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 - \nu_1)$$

By the definition of ν_2, we have:

$$(\forall \boldsymbol{\theta} \in \mathcal{H})\ \nu_2 \leq \mathcal{M}(\boldsymbol{\theta})$$

By putting both statements together for $\gamma = \nu_2$, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta} \in \mathcal{H})\ \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) &\geq \nu_2(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 - \nu_1) \\ &\geq \nu_2(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty - \nu_1)\end{aligned}$$

Since $(\forall \boldsymbol{\theta} \in \mathcal{H})\ \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq 0$, we prove our claim. □

B. Discussion on Sparsistency

One way to prove sparsistency and sign consistency is to use the *primal-dual witness* method (Negahban & Wainwright, 2011; Obozinski et al., 2011; Ravikumar et al., 2008; Wainwright, 2009b; Wainwright et al., 2006). These results are specific to the given loss (linear regression (Negahban & Wainwright, 2011; Obozinski et al., 2011; Wainwright, 2009b), log-likelihood of Gaussian MRFs (Ravikumar et al., 2008), pseudo-likelihood of discrete MRFs (Wainwright et al., 2006)) as well as the specific regularizer (ℓ_1-norm (Ravikumar et al., 2008; Wainwright, 2009b; Wainwright et al., 2006), $\ell_{1,2}$-norm (Obozinski et al., 2011) and $\ell_{1,\infty}$-norm (Negahban & Wainwright, 2011)). Furthermore, due to nonuniqueness of the dual of the $\ell_{1,\infty}$-norm (Negahban & Wainwright, 2011), characterizing sign consistency by primal-dual arguments is difficult. In this paper, we prove sparsistency and sign consistency for general regularizers, besides the ℓ_1 and $\ell_{1,p}$ norms. Indeed, our results also hold for regularizers that are not norms.

Our approach is to perform thresholding of the empirical minimizer. In the context of ℓ_1-regularized linear regression, thresholding has been previously used for obtaining sparsistency and sign consistency (Meinshausen & Yu, 2009; Zhou, 2009). Note that the *primal-dual witness* method of (Negahban & Wainwright, 2011; Obozinski et al., 2011; Ravikumar et al., 2008; Wainwright, 2009b; Wainwright et al., 2006) applies only when mutual incoherence conditions hold. If such conditions are not met, sparsistency and sign consistency is not guaranteed, independently of the number of samples. In our two-step algorithm, the threshold decreases with respect to the amount of data samples. Potentially, the sparsity pattern of every true hypothesis can be recovered, even if mutual incoherence does not hold.

Seemingly contradictory results are shown in (Zhao & Yu, 2006) where mutual incoherence conditions are shown to be necessary and sufficient for ℓ_1-regularized linear regression. Note that here, we consider regularization followed by a thresholding step, which is not considered in (Zhao & Yu, 2006).

C. Dirty Multitask Prior

(Jalali et al., 2010) proposed a *dirty* multitask prior of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}^{(1)}\|_1 + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty}$ where $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)} + \boldsymbol{\theta}^{(2)}$. By the triangle inequality:

$$\begin{aligned}\|\boldsymbol{\theta}\|_{1,\infty} &= \|\boldsymbol{\theta}^{(1)} + \boldsymbol{\theta}^{(2)}\|_{1,\infty} \\ &\leq \|\boldsymbol{\theta}^{(1)}\|_{1,\infty} + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty} \\ &\leq \|\boldsymbol{\theta}^{(1)}\|_1 + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty} \\ &= \mathcal{R}(\boldsymbol{\theta})\end{aligned}$$

Thus, the *dirty* multitask prior fulfills Assumption B with $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{1,\infty}$ and $r(z) = z$.

D. Specific Dimension-Dependent Rates $\varepsilon_{n,\delta}$

D.1. Claim i for the sub-Gaussian case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall j)\ t_j(\mathbf{x})$ be sub-Gaussian with parameter σ. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t_j(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[t_j(\mathbf{x})])| > \varepsilon] \leq 2p\exp(-\frac{n\varepsilon^2}{2\sigma^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma\sqrt{2/n(\log p + \log 2/\delta)}$.

D.2. Claim i for the finite variance case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\| \cdot \|_* = \| \cdot \|_\infty$ and $\| \cdot \| = \| \cdot \|_1$. Let $(\forall j)$ $t_j(\mathbf{x})$ have variance at most σ^2. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists j) \; |\frac{1}{n} \sum_i (t_j(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[t_j(\mathbf{x})])| > \varepsilon] \leq p \, \frac{\sigma^2}{n\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma\sqrt{\frac{p}{n\delta}}$.

D.3. Claim ii for the sub-Gaussian case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\| \cdot \|_* = \| \cdot \|_\infty$ and $\| \cdot \| = \| \cdot \|_1$. Let $(\forall \mathbf{x})$ $\|\mathbf{x}\|_* \leq B$ and thus $(\forall ij)$ $|x_j^{(i)}| \leq B$. Let $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)$ be sub-Gaussian with parameter σ. Therefore $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)x_j^{(i)}$ is sub-Gaussian with parameter σB. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists j) \; |\frac{1}{n} \sum_i (t(y^{(i)}) - \mathbb{E}_{y \sim \mathcal{D}_i}[t(y)])x_j^{(i)}| > \varepsilon] \leq 2p \exp(-\frac{n\varepsilon^2}{2(\sigma B)^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{2/n(\log p + \log 2/\delta)}$.

D.4. Claim ii for the finite variance case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\| \cdot \|_* = \| \cdot \|_\infty$ and $\| \cdot \| = \| \cdot \|_1$. Let $(\forall \mathbf{x})$ $\|\mathbf{x}\|_* \leq B$ and thus $(\forall ij)$ $|x_j^{(i)}| \leq B$. Let $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)$ have variance at most σ^2. Therefore $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)x_j^{(i)}$ has variance at most $(\sigma B)^2$. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists j) \; |\frac{1}{n} \sum_i (t(y^{(i)}) - \mathbb{E}_{y \sim \mathcal{D}_i}[t(y)])x_j^{(i)}| > \varepsilon] \leq p \, \frac{(\sigma B)^2}{n\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{\frac{p}{n\delta}}$.

D.5. Claim iii for the sub-Gaussian case and ℓ_1-norm

Recall $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^{n_1 \times n_2}$ where $n = n_1 n_2$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall ij$ and $x \sim \mathcal{D}_{ij})$ $t(x)$ be sub-Gaussian with parameter σ. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists ij) \; |t(x_{ij}) - \mathbb{E}_{x \sim \mathcal{D}_{ij}}[t(x)]| > n\varepsilon] \leq 2n \exp(-\frac{(n\varepsilon)^2}{2\sigma^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \frac{\sigma}{n}\sqrt{2(\log n + \log 2/\delta)}$.

D.6. Claim iii for the finite variance case and ℓ_1-norm

Recall $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^{n_1 \times n_2}$ where $n = n_1 n_2$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall ij$ and $x \sim \mathcal{D}_{ij})$ $t(x)$ have variance at most σ^2. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists ij) \; |t(x_{ij}) - \mathbb{E}_{x \sim \mathcal{D}_{ij}}[t(x)]| > n\varepsilon] \leq n \, \frac{\sigma^2}{(n\varepsilon)^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma/\sqrt{(n\delta)}$.

D.7. Claim v for the sub-Gaussian case and ℓ_1-norm

Let $\mathbf{x} \in \mathcal{X} = \mathbb{R}^p$. Let $\| \cdot \|_* = \| \cdot \|_\infty$ and $\| \cdot \| = \| \cdot \|_1$. Let $(\forall \mathbf{x})$ $\|\boldsymbol{\psi}(\mathbf{x})\|_* \leq B$ and thus $(\forall ij)$ $|\psi_j(\mathbf{x}^{(i)})| \leq B$. Let $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)$ be sub-Gaussian with parameter σ. Therefore $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)\psi_j(\mathbf{x}^{(i)})$ is sub-Gaussian with parameter σB. The complexity of our nonparametric model grows with more samples. Let q_n be increasing with respect to the number of samples n. Assume that we have q_n orthonormal basis functions $\varphi_1, \ldots, \varphi_{q_n} : \mathbb{R} \to \mathbb{R}$. With these bases, we define $q_n p$ orthonormal basis functions of the form $\psi_j(\mathbf{x}) = \varphi_k(x_l)$ for $j = 1, \ldots, q_n p$, $k = 1, \ldots, q_n$, $l = 1, \ldots, p$. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists j) \; |\frac{1}{n} \sum_i (t(y^{(i)}) - \mathbb{E}_{y \sim \mathcal{D}_i}[t(y)])\psi_j(\mathbf{x}^{(i)})| > \varepsilon] \leq 2q_n p \exp(-\frac{n\varepsilon^2}{2(\sigma B)^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{2/n(\log p + \log q_n + \log 2/\delta)}$.

In Table 1, we set $q_n = e^{n^{2\gamma}}$ for $\gamma \in (0; 1/2)$, although other settings are possible for obtaining a decreasing rate $\varepsilon_{n,\delta}$ with respect to n.

D.8. Claim v for the finite variance case and ℓ_1-norm

Let $\mathbf{x} \in \mathcal{X} = \mathbb{R}^p$. Let $\| \cdot \|_* = \| \cdot \|_\infty$ and $\| \cdot \| = \| \cdot \|_1$. Let $(\forall \mathbf{x})$ $\|\boldsymbol{\psi}(\mathbf{x})\|_* \leq B$ and thus $(\forall ij)$ $|\psi_j(\mathbf{x}^{(i)})| \leq B$. Let $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)$ have variance at most σ^2. Therefore $(\forall i$ and $y \sim \mathcal{D}_i)$ $t(y)\psi_j(\mathbf{x}^{(i)})$ has variance at most $(\sigma B)^2$. The complexity of our nonparametric model grows with more samples. Let q_n be increasing with respect to the number of samples n. Assume that we have q_n orthonormal basis functions $\varphi_1, \ldots, \varphi_{q_n} : \mathbb{R} \to \mathbb{R}$. With these bases, we define $q_n p$ orthonormal basis functions of the form $\psi_j(\mathbf{x}) = \varphi_k(x_l)$ for $j = 1, \ldots, q_n p$, $k = 1, \ldots, q_n$, $l = 1, \ldots, p$. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists j) \; |\frac{1}{n} \sum_i (t(y^{(i)}) - \mathbb{E}_{y \sim \mathcal{D}_i}[t(y)])\psi_j(\mathbf{x}^{(i)})| > \varepsilon] \leq q_n p \, \frac{(\sigma B)^2}{n\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{\frac{q_n p}{n\delta}}$.

In Table 1, we set $q_n = n^{2\gamma}$ for $\gamma \in (0; {}^1\!/_2)$, although other settings are possible for obtaining a decreasing rate $\varepsilon_{n,\delta}$ with respect to n.

D.9. Claim vi for the sub-Gaussian case and ℓ_1-norm

In order to allow for proper estimation of the parameters $\boldsymbol{\theta}^{(j)} \in \mathbb{R}^p$ of each cluster, we assume that the hypothesis class $\mathcal{H}$ allows only for clusters containing the same number of training samples. The complexity of our nonparametric model grows with more samples. Let q_n be increasing with respect to the number of samples n. Assume that we have q_n clusters with n/q_n samples each. In order to show that for all partitions $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(\infty)}$ of $\mathcal{X}$, the dual norm fulfills $(\forall j)\ \|\frac{1}{n}\sum_i 1[\mathbf{x}^{(i)} \in \mathcal{X}^{(j)}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{X}^{(j)}]\mathbf{t}(\mathbf{x})]\|_* \leq \varepsilon_{n,\delta}$, we will show concentration for all subsets of $\{1, \ldots, n\}$ with size n/q_n. That is:

$$(\forall \mathcal{C} \subseteq \{1, \ldots, n\}, |\mathcal{C}| = n/q_n)\ \|\tfrac{1}{n}\textstyle\sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}]\mathbf{t}(\mathbf{x})]\|_* \leq \varepsilon_{n,\delta}$$

Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall j)$ $t_j(\mathbf{x})$ be sub-Gaussian with parameter σ. By the union bound, sub-Gaussianity and independence, we have:

$$\begin{aligned} \mathbb{P}[(\exists j, \mathcal{C})\ |\tfrac{1}{n/q_n}\textstyle\sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}]t_j(\mathbf{x})]| > \gamma] &\leq 2p\ \tbinom{n}{n/q_n} e^{-\frac{n/q_n \gamma^2}{2\sigma^2}} \\ &\leq 2p\ (q_n e)^{n/q_n} e^{-\frac{n/q_n \gamma^2}{2\sigma^2}} \\ &= \delta \end{aligned}$$

By solving for γ, we have $\gamma = \sigma\sqrt{2(1 + \log q_n + \frac{q_n}{n}\log p + \frac{q_n}{n}\log {}^2\!/\delta)}$. Note that $\varepsilon_{n,\delta} = \gamma/q_n$ and by setting $q_n = \sqrt{n}$ we have:

$$\begin{aligned} \varepsilon_{n,\delta} &= \sigma\sqrt{2(\tfrac{1+\log q_n}{q_n^2} + \tfrac{1}{nq_n}\log p + \tfrac{1}{nq_n}\log {}^2\!/\delta)} \\ &= \sigma\sqrt{2(\tfrac{1+\log\sqrt{n}}{n} + \tfrac{1}{n^{3/2}}\log p + \tfrac{1}{n^{3/2}}\log {}^2\!/\delta)} \end{aligned}$$

D.10. Norm inequalities to extend results to other norms

- For the k-support norm $\|\cdot\|_k^{\text{sup}}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_1 \leq \sqrt{k}\|\boldsymbol{\theta}\|_k^{\text{sup}}$.
- For the ℓ_2-norm $\|\cdot\|_2$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_2$.
- For the ℓ_∞-norm $\|\cdot\|_2$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_1 \leq p\|\boldsymbol{\theta}\|_\infty$.
- For the Frobenius norm $\|\cdot\|_{\mathfrak{F}}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})$ $\|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_{\mathfrak{F}}$.
- For the trace norm $\|\cdot\|_{\text{tr}}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})$ $\|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_{\text{tr}}$.
- For the $\ell_{1,2}$-norm $\|\cdot\|_{1,2}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})$ $\|\boldsymbol{\theta}\|_1 \leq p^{1/4}\|\boldsymbol{\theta}\|_{1,2}$.
- For the $\ell_{1,\infty}$-norm $\|\cdot\|_{1,\infty}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})$ $\|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_{1,\infty}$.
- For the $\ell_{1,2}$-norm with overlapping groups $\|\cdot\|_{1,2}^{\text{ov}}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_1 \leq \sqrt{g}\|\boldsymbol{\theta}\|_{1,2}^{\text{ov}}$ where g is the maximum group size. Additionally, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_2 \leq \|\boldsymbol{\theta}\|_{1,2}^{\text{ov}}$.
- For the $\ell_{1,\infty}$-norm with overlapping groups $\|\cdot\|_{1,\infty}^{\text{ov}}$, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_1 \leq g\|\boldsymbol{\theta}\|_{1,\infty}^{\text{ov}}$ where g is the maximum group size. Additionally, we have $(\forall\boldsymbol{\theta} \in \mathbb{R}^p)$ $\|\boldsymbol{\theta}\|_{1,2}^{\text{ov}} \leq \sqrt{g}\|\boldsymbol{\theta}\|_{1,\infty}^{\text{ov}}$.