E-Companion

EC.1. Motivating Example

We recall the one-dimensional example provided in Figure 1. The data-collection distribution $P_{Y|X}$ is given by

$$\mathbb{P}_P[X=1]=3/4, \quad \mathbb{P}_P[X=0]=1/4,$$

and

$$\mathbb{P}_P[Y \mid X=x]=\begin{cases} y_x & \text{w.p. } 2/3 \\ -y_x & \text{w.p. } 1/3 \end{cases},$$

where $y_0=4, y_1=10$.

First, we observe that the fully-saturated model $h:\{0,1\}\to\mathbb{R}$ the minimizes the mean-squared error solves

$$\inf_{h(x)\in\mathbb{R}} \frac{2}{3}\cdot(y_x-h(x))^2+\frac{1}{3}\cdot(-y_x-h(x))^2.$$

We note that this model h will solve the first-order condition

$$2(y_x-h(x))+(-y_x-h(x))=0,$$

which yields $h_{\text{naive}}(x)=\frac{y_x}{3}$.

Second, we can compute the robust fully-saturated model that minimizes the worst-case mean-squared error under the conditional biased sampling model (4) by solving a linear program. Let $S_\Gamma(P,Q_X)$ be as defined in the first paragraph of Section 2. We solve

$$\inf_{h:\mathcal{X}\to\mathbb{R}} \sup_{Q:\mathcal{S}_\Gamma(P,Q_X)} \mathbb{E}_Q\left[(Y-h(X))^2\right]. \tag{EC.1}$$

For a fixed $h(x)$, we can solve the inner maximization using a linear program

$$\begin{aligned} \text{maximize } & q^+\cdot(y_x-h(x))^2+q^-\cdot(-y_x-h(x))^2 \\ \text{subject to } & \Gamma^{-1}\cdot\frac{2}{3}\le q^+\le \min(1,\Gamma\cdot\frac{2}{3}) \\ & \Gamma^{-1}\cdot\frac{1}{3}\le q^-\le \min(1,\Gamma\cdot\frac{1}{3}) \\ & q^++q^-=1. \end{aligned}$$

Solving this program for a discrete grid of $h(x)\in[-y_x,y_x]$ and selecting $h(x)$ the yields the minimum objective value yields the solution for each $x\in\{0,1\}$.

To obtain a closed form expression, we can re-write the linear program above as follows

$$\begin{aligned} \text{maximize } & q^+\cdot(y_x-h(x))^2+(1-q^+)\cdot(-y_x-h(x))^2 \\ \text{subject to } & \max(\Gamma^{-1}\cdot\frac{2}{3},1-\min(1,\Gamma\cdot\frac{1}{3}))\le q^+\le \min(1,\Gamma\cdot\frac{2}{3},1-\Gamma^{-1}\cdot\frac{1}{3}). \end{aligned}$$

We note that the objective is linear in q^+. In particular, the objective is given by $-4 \cdot h(x) \cdot y_x \cdot q^+ + (y_x + h(x))^2$. Since the objective is linear and the constraint is an interval, then q^+ is an endpoint of the interval. If $y_x \cdot h(x) > 0$, then the optimal q^+ is given by the left endpoint $\max(\Gamma^{-1} \cdot \frac{2}{3}, 1 - \min(1, \Gamma \cdot \frac{1}{3}))$. If $y_x \cdot h(x) < 0$, then the optimal q^+ is given by the right endpoint $\min(1, \Gamma \cdot \frac{2}{3}, 1 - \Gamma^{-1} \cdot \frac{1}{3})$. If $y_x \cdot h(x) = 0$, then any feasible q^+ is optimal.

We can use these results to see that if $1 \le \Gamma \le \frac{3}{2}$, then $h_{\Gamma,\text{cond}}(x) = (h_{\text{naive}}(x) - \frac{2}{3} \cdot (\Gamma - 1) \cdot y_x)_+$ and if $\Gamma > \frac{3}{2}$, then $h_{\Gamma,\text{cond}}(x) = 0$.

Third, we can compute the robust fully-saturated model that minimizes the worst-case mean-squared error under the unconditional biased sampling model Definition 2. Let $\tilde{S}_\Gamma(P)$ be the set of distributions that can generate P under Definition 2. We solve

$$\inf_{h:\mathcal{X}\to\mathbb{R}} \sup_{Q:\tilde{S}_\Gamma(P)} \mathbb{E}_Q\left[(Y - h(X))^2\right]. \tag{EC.2}$$

First, we show that we can restrict the decision rules h that we optimize over to the following set

$$\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R} \mid h(1) \in [-y_1, y_1],\, h(0) \in [-y_0, y_0]\}.$$

Consider $h' \notin \mathcal{H}$ and let $\tilde{Q}$ be any probability distribution over (X, Y) with mass on four points $(0, y_0), (0, -y_0), (1, y_1), (1, -y_1)$. Let $\bar{h}$ be defined as follows: $\bar{h}(1)$ is the projection of $h'(1)$ on $[-y_1, y_1]$ and $\bar{h}(0)$ be the projection of $h'(0)$ on $[-y_0, y_0]$. Then, it is straightforward to see that $\mathbb{E}_{\tilde{Q}}\left[(Y - \bar{h}(X))^2\right] \le \mathbb{E}_Q\left[(Y - h'(X))^2\right]$.

We can now consider the following optimization problem

$$\inf_{h\in\mathcal{H}} \sup_{Q:\tilde{S}_\Gamma(P)} \mathbb{E}_Q\left[(Y - h(X))^2\right]. \tag{EC.3}$$

We consider solving the inner maximization problem for a fixed choice of $h \in \mathcal{H}$. We must solve

$$\begin{aligned}
\text{maximize } & q_1^+ \cdot (y_1 - h(1))^2 + q_1^- (-y_1 - h(1))^2 + q_0^+ \cdot (y_0 - h(0))^2 + q_0^- \cdot (-y_0 - h(0))^2 \\
\text{subject to } & \sum_{i\in\{0,1\}} q_i^+ + q_i^- = 1 \\
& \frac{\Gamma^{-1}}{2} \le q_1^+ \le \min(\frac{\Gamma}{2}, 1) \\
& \frac{\Gamma^{-1}}{4} \le q_1^- \le \min(\frac{\Gamma}{4}, 1) \\
& \frac{\Gamma^{-1}}{6} \le q_0^+ \le \min(\frac{\Gamma}{6}, 1) \\
& \frac{\Gamma^{-1}}{12} \le q_0^- \le \min(\frac{\Gamma}{12}, 1).
\end{aligned}$$

This program can be written as a standard LP with three free variables.

$$\begin{aligned}
\text{maximize } & q_1^- \cdot [(y_1 + h(1))^2 - (y_1 - h(1))^2] \\
& + q_0^+ \cdot [(y_0 - h(0))^2 - (y_1 - h(1))^2] \\
& + q_0^- \cdot [(y_0 + h(0))^2 - (y_1 - h(1))^2] \\
\text{subject to } & 1 - \min(\frac{\Gamma}{2}, 1) \le q_1^- + q_0^+ + q_0^- \le 1 - \frac{\Gamma^{-1}}{2}, \\
& \frac{\Gamma^{-1}}{4} \le q_1^- \le \min(\frac{\Gamma}{4}, 1), \\
& \frac{\Gamma^{-1}}{6} \le q_0^+ \le \min(\frac{\Gamma}{6}, 1), \\
& \frac{\Gamma^{-1}}{12} \le q_0^- \le \min(\frac{\Gamma}{12}, 1).
\end{aligned} \tag{EC.4}$$

Note that the lower bound of the first constraint does not bind because $1 - \min(\frac{\Gamma}{2}, 1) \le \frac{\Gamma^{-1}}{2}$ for all $\Gamma \ge 1$, so we can ignore it.

We assume for the sake of contradiction that $h_{\Gamma,\text{uncond}}$ lies in the set $\mathcal{H}_1^- = \{h \in \mathcal{H} \mid h(1) < 0\}$. Note that if this holds, the coefficient of q_1^- must be negative. Thus, the worst-case distribution must require that q_1^- takes on its lowest possible value $\frac{\Gamma^{-1}}{4}$. In this parameter regime, we note that

$$\frac{2\Gamma^{-1}}{3} \le q_1^- + q_0^+ + q_0^- \le \min(\frac{\Gamma^{-1}}{4} + \min(\frac{\Gamma}{6}, 1) + \min(\frac{\Gamma}{12}, 1), 1 - \frac{\Gamma^{-1}}{2}). \tag{EC.5}$$

As a result, we must have that when $1 \le \Gamma < 3$,

$$q_1^+ \ge 1 - \frac{\Gamma^{-1} + \Gamma}{4}, \tag{EC.6}$$

and when $\Gamma \ge 3$,

$$q_1^+ \ge \frac{\Gamma^{-1}}{2}. \tag{EC.7}$$

We note that a policy h that minimizes the risk under any distribution that satisfies (EC.6) when $1 \le \Gamma < 3$ and (EC.7) when $\Gamma > 3$ and has $q_1^- = \frac{\Gamma^{-1}}{4}$ must have $h(1) \ge 0$. This is a contradiction, so $h_{\Gamma,\text{uncond}} \notin \mathcal{H}_1^-$.

As a result, $h_{\Gamma,\text{uncond}} \in \mathcal{H} \setminus \mathcal{H}_1^-$. Note that any $h \in \mathcal{H} \setminus \mathcal{H}_1^-$ has the properties that $h(0) \in [-4, 4]$ and $h(1) \in [0, 10]$. We characterize the solution to (EC.4) when $h = h_{\Gamma,\text{uncond}}$. First, since $h_{\Gamma,\text{uncond}}(1) \ge 0$, the solution must have the property that

$$\frac{q_1^+}{q_1^-} \ge 1.$$

As a result, we can add this as a constraint to the optimization problem (EC.4) as follows:

$$\begin{aligned}
\text{maximize } & q_1^- \cdot [(y_1 + h(1))^2 - (y_1 - h(1))^2] \\
& + q_0^+ \cdot [(y_0 - h(0))^2 - (y_1 - h(1))^2] \\
& + q_0^- \cdot [(y_0 + h(0))^2 - (y_1 - h(1))^2] \\
\text{subject to } & q_0^+ + q_0^- \le \min(1 - \frac{\Gamma^{-1}}{2} - q_1^-, 1 - 2q_1^-), \\
& \frac{\Gamma^{-1}}{4} \le q_1^- \le \min(\frac{\Gamma}{4}, 1), \\
& \frac{\Gamma^{-1}}{6} \le q_0^+ \le \min(\frac{\Gamma}{6}, 1), \\
& \frac{\Gamma^{-1}}{12} \le q_0^- \le \min(\frac{\Gamma}{12}, 1).
\end{aligned} \tag{EC.8}$$

To solve this optimization problem, we observe that the coefficient of q_1^- in the objective is guaranteed to be strictly greater than the coefficients of q_0^+, q_0^- when $h \in \mathcal{H} \setminus \mathcal{H}_1^-$. To see this, observe that $y_1 + h_{\Gamma,\text{uncond}}(1) \in [10, 20]$ and $y_0 + h_{\Gamma,\text{uncond}}(0), y_0 - h_{\Gamma,\text{uncond}}(0) \in [0, 8]$.

As a result, the optimal solution to (EC.8) when $h \in \mathcal{H} \setminus \mathcal{H}_1^-$ will allocate the maximum possible mass to q_1^- subject to feasibility. The maximum possible mass is allocated to q_1^- when the minimum possible mass is allocated to q_0^+, q_0^-. Allocating the minimum possible mass to q_0^+, q_0^- is feasible

$$(q_0^+, q_0^-) = (\frac{\Gamma^{-1}}{6}, \frac{\Gamma^{-1}}{12}).$$

The remaining constraints are that

$$\begin{aligned}
& \frac{\Gamma^{-1}}{4} \le 1 - \frac{\Gamma^{-1}}{2} - q_1^-, \\
& \frac{\Gamma^{-1}}{4} \le 1 - 2q_1^-, \\
& \frac{\Gamma^{-1}}{4} \le q_1^- \le \min(\frac{\Gamma}{4}, 1).
\end{aligned}$$

We note that these constraints can be simplified to yield

$$q_1^- \le \begin{cases} \frac{\Gamma}{4} & 1 \le \Gamma < 1 + \frac{\sqrt{2}}{2} \\ \frac{1}{2} - \frac{\Gamma^{-1}}{8} & \Gamma \ge 1 + \frac{\sqrt{2}}{2} \end{cases}.$$

Thus, the optimal solution to (EC.8) when $h \in \mathcal{H} \setminus \mathcal{H}_1^-$ will set q_1^- to its upper bound above. Since this solution allocates the minimum possible mass to q_0^+, q_0^-, it is optimal.

Thus, the worst-case distribution are as follows. For $1 \le \Gamma < 1 + \frac{\sqrt{2}}{2}$,

$$dQ(x, y) = \begin{cases} (1, y_1) & \text{w.p. } 1 - \frac{\Gamma + \Gamma^{-1}}{4} \\ (1, -y_1) & \text{w.p. } \frac{\Gamma}{4} \\ (0, y_0) & \text{w.p. } \frac{\Gamma^{-1}}{6} \\ (0, -y_0) & \text{w.p. } \frac{\Gamma^{-1}}{12} \end{cases}.$$

For $\Gamma \geq 1 + \frac{\sqrt{2}}{2}$,

$$dQ(x,y) = \begin{cases} (1, y_1) & \text{w.p. } \frac{1}{2} - \frac{\Gamma^{-1}}{8} \\ (1, -y_1) & \text{w.p. } \frac{1}{2} - \frac{\Gamma^{-1}}{8} \\ (0, y_0) & \text{w.p. } \frac{\Gamma^{-1}}{6} \\ (0, -y_0) & \text{w.p. } \frac{\Gamma^{-1}}{12} \end{cases}.$$

Thus,

$$h_{\Gamma,\text{uncond}}(x) = \begin{cases} \frac{y_0}{3} & x = 0 \\ ((1 - \frac{2\Gamma^2}{4\Gamma - 1}) \cdot y_1)_+ & x = 1 \end{cases}$$

Equivalently,

$$h_{\Gamma,\text{uncond}}(x) = \begin{cases} h_{\text{naive}}(x) & x = 0 \\ (h_{\text{naive}}(x) - \frac{60\Gamma^2 - 20(4\Gamma - 1)}{3 \cdot (4\Gamma - 1)})_+ & x = 1 \end{cases}.$$

EC.2. Simulation Examples

We start by presenting results on two stylized one-dimensional examples to further elucidate the difference of RU Regression and Unconditional RU Regression. For simplicity, these examples do not involve covariate shift. Our first setting uses a data-generating process P

$$X_i \sim \text{Uniform}[0,6], \quad U_i \sim \text{Bernoulli}(p), \quad Y_i \mid X_i, U_i \sim N(\sin(X_i) + 5 \cdot U_i,\, 1). \tag{EC.9}$$

where $p = 0.2$ and U_i is unobservable. Our second setting uses

$$\begin{aligned} &X_i \sim \text{Uniform}[0,10], \quad U_i \sim \text{Bernoulli}(p), \quad Z_i \sim \text{Bernoulli}(0.5), \\ &Y_i \mid X_i, U_i, Z_i \sim \begin{cases} N(\sqrt{X_i} + U_i \cdot (3\sqrt{X_i} + 1),\, 1) & X_i \leq 6 \\ N(10 \cdot (2Z_i - 1),\, 1) & X_i > 6 \end{cases}, \end{aligned} \tag{EC.10}$$

where again $p = 0.2$ and U_i, Z_i are unobservables. In both cases, we seek to learn predictive rules with sampling-bias-robust guarantees under mean-squared error.

In both settings, data is split into two "bands" governed by an unobserved parameter U_i. In the observed data, in both cases, data from the lower band ($U_i = 0$) has been oversampled relative to the upper band, and basic empirical-risk minimization will thus focus on accurate prediction for this lower bound. In contrast, one might qualitatively expect sampling-bias-robust methods to seek predictions that make a more even compromise between accuracy in the upper and lower bands, just in case data from the upper band proves to be more prevalent in the target population.

Results for the first setting are shown in Figure EC.1. As expected, both conditional and unconditional RU regression shift the prediction towards the (less sampled) upper band—and do so more aggressively as we increase Γ. The right panel of Figure EC.1 shows the learned quantile function

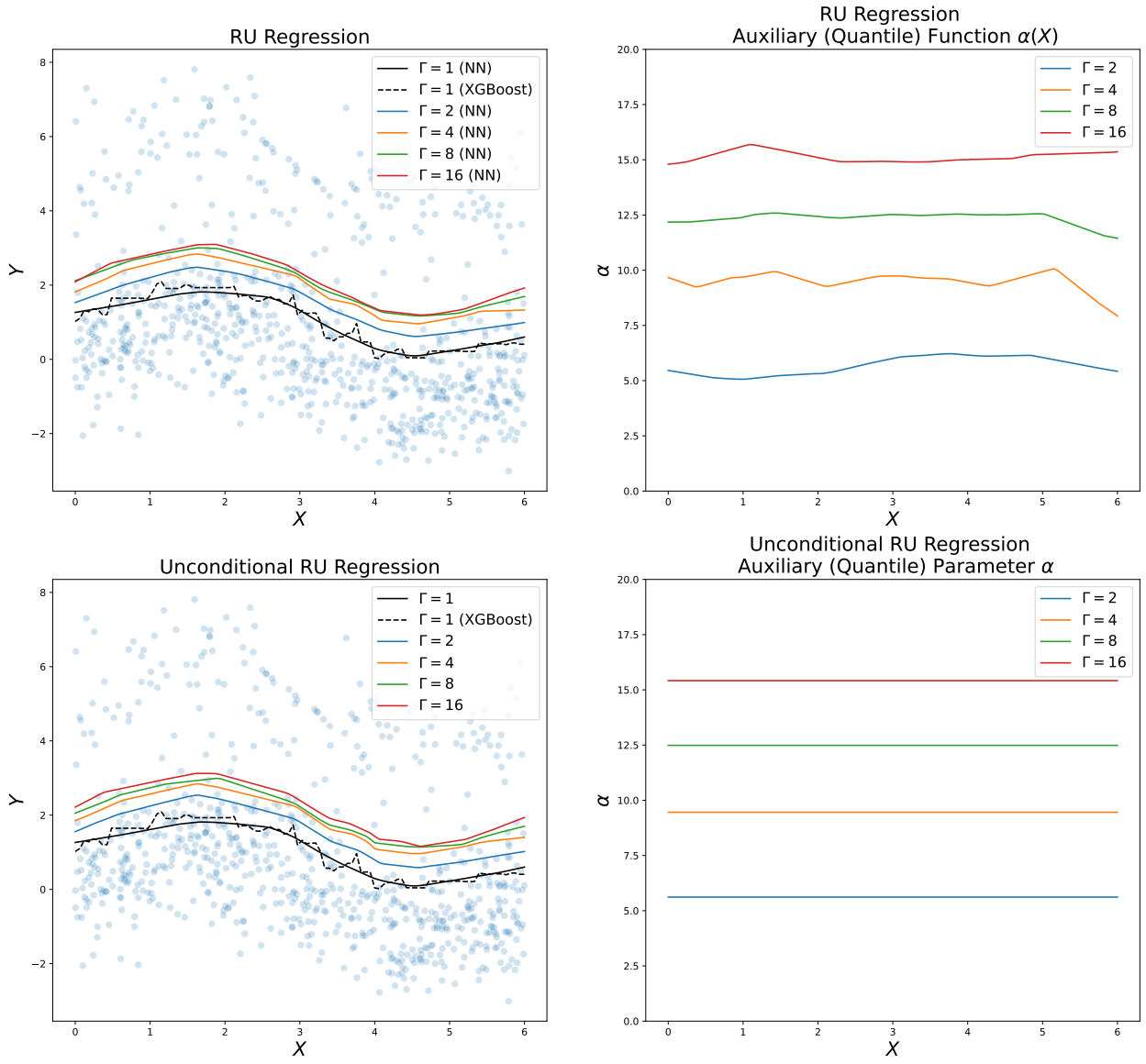


Figure EC.1 **We plot regression models that are learned via RU Regression and Unconditional RU Regression and the training distribution. We also visualized the learned auxiliary parameter α. Compared to the previous example, the distribution over conditional losses does not change too much with x.**

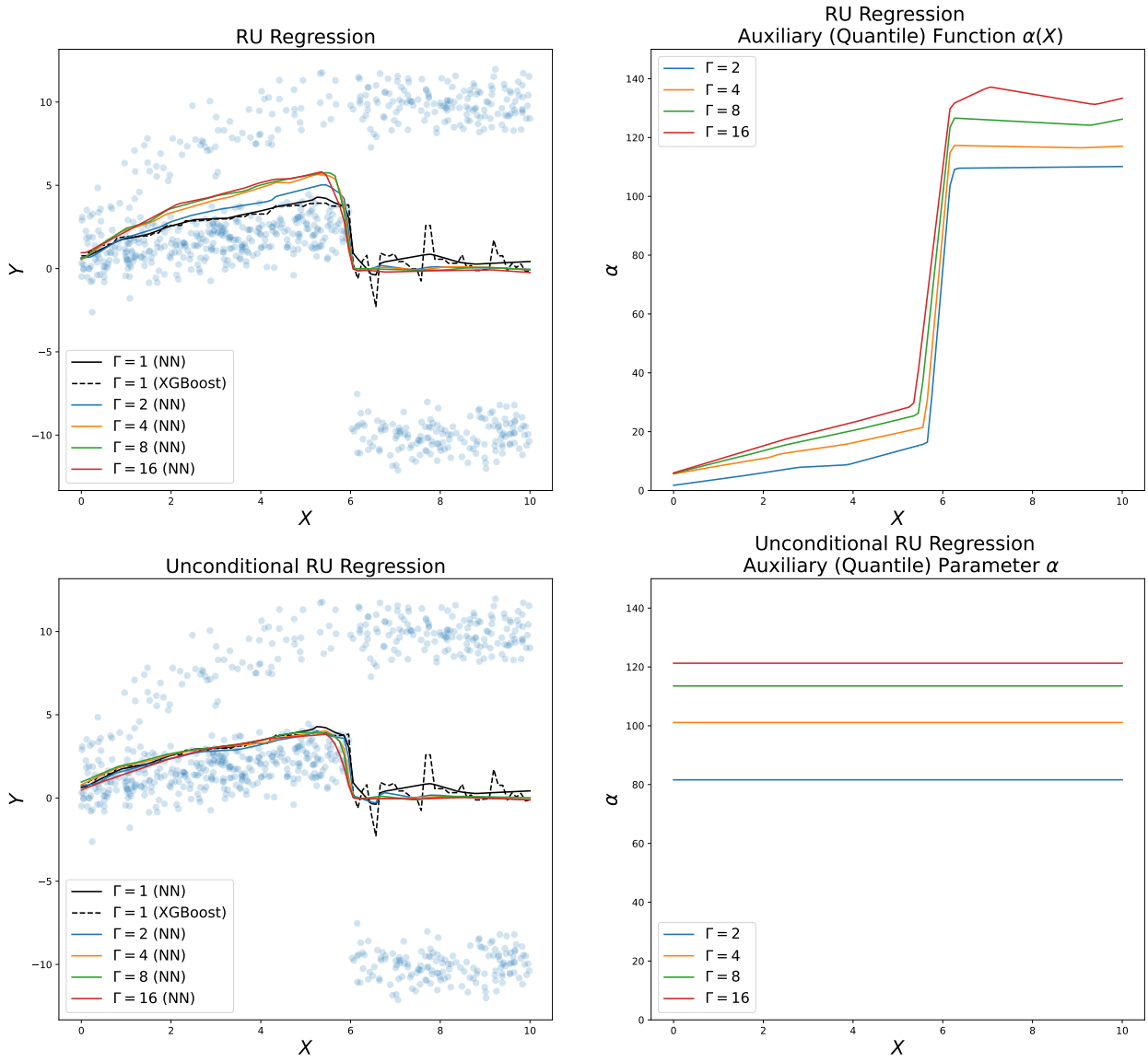


Figure EC.2 **We plot regression models that are learned via RU Regression and Unconditional RU Regression and the training distribution. We also visualized the learned auxiliary parameter α. We note that the loss distribution differs significantly on from the region $X_i < 6$ to $X_i > 6$.**

for the conditional loss, $\alpha(x)$. We see that the learned quantile function for conditional RU regression is essentially constant in x, which helps understand why its behavior here closely matched that of unconditional RU regression (which pre-specifies use of a constant α).

Results for the second setting are shown in Figure EC.2. To understand the results, it is important to recognize how the data-generating distribution changes when $X_i < 6$ versus $X_i > 6$. In the former "easy" region, the outcomes Y_i can be clustered into two bands depending on U_i as described above; and outcomes in both bands can be predicted in terms of X_i. On the other hand, in the latter "difficult" region, the outcomes Y_i are either approximately 10 or -10 with equal probability, and X_i is not predictive at all—thus any predictive rule will have large mean-squared error in this region.

Given these preliminiaries, we see that conditional RU regression again behaves qualitatively like in the first setting: In the easy regime, where there are two predictable bands but one is oversampled, conditional RU regression moves predictions towards the less sampled band. In the difficult regime, where there is nothing to predict, conditional RU regression leaves predictions unmoved relative to empirical risk minimization. Conversely, in this setting, unconditional RU regression essentially doesn't change the predictions made by empirical risk minimization. While this may appear paradoxical, what's going on here is that unconditional RU regression is most concerned about distributional shifts in X-space that change the relative sampling frequency of the easy and difficult regions, and it's less focused on shifts in the distribution of the unobserved U_i. But, because our neural network can flexibly adjust for X_i, re-weighting the sample between the easy and difficult regions (which can be identified in terms of X_i) doesn't change the nature of the optimal prediction rule at all. It thus appears that, in this example, using unconditional RU regression makes us overall more pessimistic about how much loss we might face—but doesn't give us useful guidance in how to change predictions.

EC.3. Additional Details on Numerical Experiments

EC.3.1. Online Health Survey Experiment

EC.3.1.1. Datasets The Household Pulse Survey (HPS) (Bureau 2021) is online health survey deployed by the Census Bureau to collect core demographic household characteristics, as well as topics such as mental health and food sufficiency. We treat HPS 2021 as our biased training population because it is known to have a 2-10% response rate and may not be as representative as traditional offline surveys.

The Behavioral Risk Factor Surveillance System (BRFSS) (CDC 2021) is a telephone survey administered by the CDC to collect state data about US residents regarding their health-related risk behaviors, chronic health conditions, and use of preventative services. We treat BRFSS 2021 as our target population.

We also use data from the American Community Survey (Bureau 2006) to obtain state-level characteristics.

EC.3.1.2. Features and Dataset Preprocessing

Features. We use 44 features which are a combination of individual-level attributes and state-level characteristics. The individual-level characteristics that we use are age group, gender, education level, income level, race/ethnicity, marital status, any insurance, insurance from their employer, Medicare insurance, Medicaid insurance, other insurance, household size, and state. The state-level characteristics that we have is the proportion of the state that attened a 4 year college, proportion with health insurance, average household size, educated at the high school level or less, proportion that only speak English speak, proportion of females never married, fertility rate, proportion on food stamps, proportion with a graduate degree, proportion of households with a computer, proportion of households with internet, proportion of males never married, mean income, median house value, median income, median rent, proportion in poverty, proportion with private health insurance, proportion attended some college, unemployment rate, proportion US born, proportion veterans, Republican percentage, and total population.

Relationship Between Mental Health Outcomes. The BRFSS and HPS collect different mental health indicators. The BRFSS measures 30-day prevalence of clinically significant anxiety and depression while the HPS administers the scale PHQ-4. The Depression and Anxiety module of the 2018 BRFSS administered the PHQ-4 survey to BRFSS survey respondents from Guam, Oregon, Tennessee, and Ohio, in addition to measuring their 30-day prevalence of clinically significant anxiety and depression. Based on Kessler et al. (2022b), who learn a mapping between dichotomized versions of these two mental health outcomes, we use this data to learn a mapping between the two mental health outcomes.

Since we expect the PHQ-4 score to be monotonically increasing in 30-day prevalence of anxiety-depression, we fit an isotonic regression model to the 2018 BRFSS data. We plot the learned mapping in Figure EC.3.

Covariate Weights. We estimate the weights using $\tilde{X}_i$, a subset of the covariates corresponding to age group, gender, education category, income level, and race/ethnicity. We create a dataset $\{(\tilde{X}_i, Z_i)\}$ taking covariates from the HPS train and validation set and assigning them the label $Z_i = 0$ and taking covariates from the BRFSS dataset and assigning them the label $Z_i = 1$. Following Sugiyama et al. (2008), Menon and Ong (2016), we train a class-balanced probabilistic classifier $f : \tilde{\mathcal{X}} \to [0, 1]$ and compute the density ratio

$$r(x_s) = \frac{f(x_s)}{1 - f(x_s)}.$$

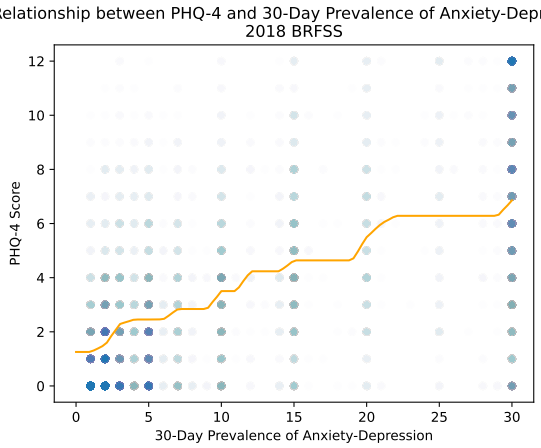


Figure EC.3 We use the BRFSS 2018 data to learn a mapping from 30-day prevalence of anxiety-depression to PHQ-4 score.

When the full covariates X_i are used, this approach yields the true density ratio

$$r(x) = \frac{dQ_X(x)}{dP_X(x)} = \frac{\mathbb{P}\left[X = x \mid Z = 1\right]}{\mathbb{P}\left[X = x \mid Z = 0\right]} = \frac{f(x)}{1 - f(x)}.$$

In our experiments, we set f is a decision tree classifier.

EC.3.1.3. Training Details

Models. In the first set of experiments (h is a linear model), we represent h as a one-layer neural network. For RU Regression and Weighted RU Regression, α is represented by a neural network with 1 hidden layer with 64 units and ReLU activation.

In the second set of experiments (h is a neural network), we represent h as a neural network with 2 hidden layers and 64 units per hidden layer and ReLU activation. For RU Regression and Weighted RU Regression, α is a neural network with 1 hidden layer with 64 units and ReLU activation.

We also compare to an XGBoost baseline (h is an ensemble of decision trees). With the exception of the number of boosting rounds and the learning rate, we use the default hyperparameter settings of the Python XGBoost package. We set the number of boosting rounds to be at most 5000 and learning rate to be 0.001. The default parameters settings are as follows: the step size shrinkage is 0.3, minimum loss reduction required to make a further partition on a leaf node of the tree is 0, the maximum depth of the trees is 6, the minimum sum of instance weight (hessian) needed in a child is 1, the maximum delta step we allow each leaf output is unconstrained, the weight of ℓ_2 regularization is set to 1, and ℓ_1 regularization is not used.

Dataset Splits. The HPS dataset is split into train, validation, and test sets consisting of 403636, 269091, and 448486 samples, respectively. The BRFSS dataset consists of 423807 samples, only used at test-time.

Training Procedure . The neural network models are trained for a maximum of 20 epochs with batch size equal to 20000 and we use the Adam optimizer with learning rate 1e-2. Each epoch we check the loss obtained on the validation set and select the model that minimizes the loss on the validation set. The XGBoost model is trained for a maximum of 5000 boosting rounds. Each round we check the loss obtained on the validation set and stop training the model after the validation loss has not improved in 20 boosting rounds.

EC.3.2. One-Dimensional Toy Example

Models. For RU Regression, we jointly train two neural networks to learn the regression function h and the quantile function α, respectively. The neural network for h has 2 hidden layers and 64 units per hidden layer and ReLU activation. The neural network for α has 1 hidden layer and 64 units per hidden layer and ReLU activation.

For Unconditional RU regression, we jointly train a neural network to learn the regression function h and a one-dimensional parameter α. The neural network for h has 2 hidden layers and 64 units per hidden layer and ReLU activation.

We also compare to an XGBoost baseline (h is an ensemble of decision trees). With the exception of the number of boosting rounds, the learning rate, and the maximum depth, we use the default hyperparameter settings of the Python XGBoost package. We set the maximum depth to be 3, the number of boosting rounds to be at most 5000, and learning rate to be 0.001. The default parameters settings are as follows: the step size shrinkage is 0.3, minimum loss reduction required to make a further partition on a leaf node of the tree is 0, the minimum sum of instance weight (hessian) needed in a child is 1, the maximum delta step we allow each leaf output is unconstrained, the weight of ℓ_2 regularization is set to 1, and ℓ_1 regularization is not used.

Dataset Splits. For all methods, the train and validation sets consists of 10000 and 2000 samples, respectively. The train and validation sets are generated via the data models specified in Section EC.2.

Training Procedure. The models are trained for a maximum of 400 epochs with batch size equal to 2500 and we use the Adam optimizer with learning rate 1e-2. Each epoch we check the loss obtained on the validation set and select the model that minimizes the loss on the validation set. The XGBoost model is trained for a maximum of 5000 boosting rounds. Each round we check the loss obtained on the validation set and stop training the model after the validation loss has not improved in 20 boosting rounds.

EC.3.3. MIMIC-III Experiment

EC.3.3.1. Dataset Medical Information Mart for Intensive Care III (MIMIC-III) is a freely accessible medical database of critically ill patients admitted to the intensive care unit (ICU)

at Beth Israel Deaconess Medical Center (BIDMC) from 2001 to 2012 (Johnson et al. 2016b, Goldberger et al. 2000). During that time, BIDMC switched clinical information systems from Carevue (2001-2008) to Metavision (2008-2012). To ensure data consistency, only data archived via the Metavision system was used in the dataset.

EC.3.3.2. Feature Selection and Data Preprocessing We use the patient features and imputed values as in Harutyunyan et al. (2019). A total of 17 variables were extracted from the chartevents table to include in the dataset - capillary refill rate, blood pressure (systolic, diastolic, and mean), fraction of inspired oxygen, Glasgow Coma Score (eye opening response, motor response, verbal response, and total score), serum glucose, heart rate, respiratory rate, oxygen saturation, respiratory rate, temperature, weight, arterial pH. We also include additional demographic features including age, gender, and ethnicity. For each unique ICU stay, values were extracted for the first 24 hours upon admission to the ICU and averaged. Normal values were imputed for missing variables as shown in Table EC.1.

Variable	MIMIC-III item ids from chartevents table	Imputed value
Capillary refill rate	(223951, 224308)	0
Diastolic blood pressure	(220051, 227242, 224643, 220180, 225310)	59.0
Systolic blood pressure	(220050, 224167, 227243, 220179, 225309)	118.0
Mean blood pressure	(220052, 220181, 225312)	77.0
Fraction inspired oxygen	(223835)	0.21
GCS eye opening	(220739)	4
GCS motor response	(223901)	6
GCS verbal response	(223900)	5
GCS total	(220739 + 223901 + 223900)	15
Glucose	(228388, 225664, 220621, 226537)	128.0
Heart Rate	(220045)	86
Height	(226707, 226730)	170.0
Oxygen saturation	(220227, 220277, 228232)	98.0
Respiratory rate	(220210, 224688, 224689, 224690)	19
Temperature	(223761, 223762)	97.88
Weight	(224639, 226512, 226531)	178.6
pH	(223830)	7.4

Table EC.1 Variables included in dataset

Following the cohort selection procedure in Wang et al. (2020), we further restrict to patients with covariates within physiologically valid range of measurements and length-of-stay less than or equal to 10 days.

EC.3.3.3. Training Details Models. For RU Regression, we jointly train two neural networks to learn the regression function h and the quantile function α, respectively. The neural network for h has 2 hidden layers and 64 units per hidden layer and ReLU activation. The neural network for α has 1 hidden layer and 64 units per hidden layer and ReLU activation.

For Unconditional RU regression, we jointly train a neural network to learn the regression function h and a one-dimensional parameter α. The neural network for h has 2 hidden layers and 64 units per hidden layer and ReLU activation.

We also compare to an XGBoost baseline (h is an ensemble of decision trees). With the exception of the number of boosting rounds and the learning rate, we use the default hyperparameter settings of the Python XGBoost package. We set the number of boosting rounds to be at most 5000 and learning rate to be 0.001. The default parameters settings are as follows: the step size shrinkage is 0.3, minimum loss reduction required to make a further partition on a leaf node of the tree is 0, the maximum depth of the trees is 6, the minimum sum of instance weight (hessian) needed in a child is 1, the maximum delta step we allow each leaf output is unconstrained, the weight of ℓ_2 regularization is set to 1, and ℓ_1 regularization is not used.

Dataset Splits. The MIMIC-III dataset is split into train, validation, and test sets consisting of 7045, 4697, and 7829 samples, respectively. First, a kernel density estimator of dQ_Y is fit using pooled data from the original MIMIC-III train and validation sets. This kernel density estimator is used to compute the weight function π_e that generates the biased training populations. Recall that training sets for the biased training population P_e are generated by sampling with replacement from the original MIMIC-III training set using weight function π_e. The size of the generated training sets are 7045, the same size as the original training set.

Training Procedure. The models are trained for a maximum of 20 epochs with batch size equal to 1761 and we use the Adam optimizer with learning rate 1e-2. Each epoch we check the loss obtained on the validation set and select the model that minimizes the loss on the validation set. The XGBoost model is trained for a maximum of 5000 boosting rounds. Each round we check the loss obtained on the validation set and stop training the model after the validation loss has not improved in 20 boosting rounds.

EC.3.4. Selection of Sensitivity Parameter

In our experiments, we did not focus on selection of Γ, and instead, we treated Γ as a sensitivity parameter and evaluated the procedure for a range of Γ. However, in practice, we may need to commit to some choice of sensitivity parameter without access to the target conditional distribution. We propose the following heuristic procedure for selecting the sensitivity parameter when the target covariate distribution Q_X is available. This procedure is analogous to Appendix F of Chernozhukov et al. (2022).

1. For each covariate $j = 1, 2, \dots d$, we can treat X_j as the outcome and compute the maximal and minimal values of $r_j(x_j \mid x_{-j}) = \frac{dQ_{X_j|X_{-j}=x_{-j}}(x_j)}{dP_{X_j|X_{-j}=x_{-j}}(x_j)}$ for various choices of x_{-j}. In practice, we may instead want to consider x_{-j} in our training data, and the 1% and 99% quantiles of

the probabilities as the upper and lower bounds. We can take these values to compute Γ_j, the appropriate sensitivity parameter for this "outcome."

2. We can form an estimate of the sensitivity parameter Γ using either the maximal or certain quantile of $\Gamma_1, \ldots, \Gamma_d$.

To estimate the conditional density ratio r_j, one approach is to estimate the two conditional density functions $dQ_{X_j|X_{-j}}, dP_{X_j|X_{-j}}$ and then compute the density ratio.

Alternatively, we can estimate the conditional density ratio r_j by estimating the "joint" density ratio $r(x_{-j}, x_j) = \frac{dQ_{X_{-j},X_j}(x_{-j},x_j)}{dP_{X_{-j},X_j}(x_{-j},x_j)}$ and the "covariate" density ratio $r_{-j}(x_{-j}) = \frac{dQ_{X_{-j}}(x_{-j})}{dP_{X_{-j}}(x_{-j})}$ by the balanced-classifier approach described in Section EC.3.1.2. We can combine these two density ratios as follows to obtain the desired conditional density ratio

$$r_j(x_j \mid x_{-j}) = \frac{r(x_{-j}, x_j)}{r_{-j}(x_{-j})}.$$

EC.4. Preliminaries

EC.4.1. Notation

EC.4.1.1. Loss function-related notation We introduce notation that is used in the proofs and technical lemmas.

$$L^{\Gamma}_{\mathrm{RU},1}(z,y) := \Gamma^{-1} L(z,y) \tag{EC.11}$$

$$L^{\Gamma}_{\mathrm{RU},2}(a) := (1-\Gamma^{-1})a \tag{EC.12}$$

$$L^{\Gamma}_{\mathrm{RU},3}(z,y,a) := (\Gamma - \Gamma^{-1}) \cdot (L(z,y)-a)_+. \tag{EC.13}$$

When we consider loss functions L that satisfy Assumption 2, we define

$$\ell_1(y) := \begin{cases} \ell(y) & y > 0 \\ 0 & y \le 0 \end{cases}, \quad \ell_2(y) := \begin{cases} 0 & y > 0 \\ \ell(y) & y \le 0 \end{cases}, \tag{EC.14}$$

$$T_{1,x}(c) := \mathbb{E}_{P_{Y|X=x}}\left[\ell(Y-c) \mid X = x\right], \tag{EC.15}$$

$$T_{3,x}(c,d) := \begin{cases} \mathbb{E}_{P_{Y|X=x}}\left[(\ell(Y-c)-d)\mathbb{I}(\ell(Y-c)>d) \mid X=x\right] & d > 0 \\ \mathbb{E}_{P_{Y|X=x}}\left[\ell(Y-c)-d \mid X=x\right] & d \le 0 \end{cases}. \tag{EC.16}$$

EC.4.1.2. Absolute Constants We define the following absolute constants that will be used in the proofs. Let

$$M^+_{\Gamma} = \sup_{h\in\mathcal{H}, x\in\mathcal{X}} q^L_{\eta(\Gamma)}(x; h(x)) + 1, \tag{EC.17}$$

where $\mathcal{H} = \{h \in L^2(P_X, \mathcal{X}) \mid ||h||_\infty \le 2B\}$. Note that M^+_Γ depends on B, Γ, L, and the conditional data distribution $P_{Y|X}$. Further, define

$$C_{M^+_\Gamma} := \sup_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(M^+_\Gamma))|, \tag{EC.18}$$

$$C_{M^-,\delta} := \inf_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(M^- - \delta))| \tag{EC.19}$$

$$C_{M^-} := \inf_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(M^-))|. \tag{EC.20}$$

where M^+_Γ is defined in (EC.17) and M^- is defined in Theorem 2.

EC.4.1.3. Useful Sets Define

$$R_{f,c} := \{x \in \mathcal{X} \mid f(x) < c\} \tag{EC.21}$$

$$S_{f,c} := \{x \in \mathcal{X} \mid f(x) > c\}. \tag{EC.22}$$

Let $\Theta = \mathcal{H} \times \mathcal{A}$ where

$$\mathcal{H} = \{h \in L^2(P_X, \mathcal{X}) \mid ||h||_\infty \le 2B\}, \quad \mathcal{A} = \{\alpha \in L^2(P_X, \mathcal{X}) \mid 0 \le \alpha \le M_\Gamma^+\},$$

and M_Γ^+ is defined in (EC.17).

EC.4.2. Technical Lemmas

Our main results rely on the following technical lemmas.

LEMMA 4. *Under Assumption 2, $T_{1,x}(c)$ is twice-differentiable in c and*

$$\mathbb{E}_P\left[L^\Gamma_{RU,1}(h(X), Y)\right] = \Gamma^{-1}\mathbb{E}_{P_X}\left[T_{1,X}(h(X))\right].$$

Proof in Appendix EC.6.1.

LEMMA 5. *Under Assumption 2, 3, $T_{3,x}(c,d)$ is differentiable in c, d. In particular,*

$$T^d_{3,x}(c,d) = \begin{cases} -\Pr(\ell(Y-c) > d \mid X = x) & d > 0 \\ -1 & d \le 0 \end{cases}.$$

Equivalently,

$$T^d_{3,x}(c,d) = \begin{cases} -1 + P_{Y|X=x}(c + \ell_1^{-1}(d)) - P_{Y|X=x}(c + \ell_2^{-1}(d)) & d > 0 \\ -1 & d \le 0 \end{cases}.$$

In addition, $T_{3,x}(c,d)$ is twice-differentiable in c, d when $d > 0$. The second derivatives are

$$\begin{aligned} T^{cc}_{3,x}(c,d) &= \sum_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(c + \ell_i^{-1}(d)) + \mathbb{E}_{P_{Y|X}}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c) > d)\right], \\ T^{dd}_{3,x}(c,d) &= \sum_{i\in\{1,2\}} \frac{p_{Y|X=x}(c + \ell_i^{-1}(d))}{|\ell'(\ell_i^{-1}(d))|}, \\ T^{cd}_{3,x}(c,d) &= p_{Y|X=x}(c + \ell_1^{-1}(d)) - p_{Y|X=x}(c + \ell_2^{-1}(d)), \end{aligned}$$

where ℓ_1^{-1} is the inverse of $\ell(z)$ when $z > 0$ and ℓ_2^{-1} is the inverse of $\ell(z)$ when $z < 0$.

Also,

$$\mathbb{E}_P\left[L^\Gamma_{RU,3}(h(X), \alpha(X), Y)\right] = (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[T_{3,X}(h(X), \alpha(X))\right].$$

Proof in Appendix EC.6.2.

LEMMA 6. *Under Assumptions 1, 2, 3, $\mathbb{E}_P\left[L^\Gamma_{RU}(h(X), \alpha(X), Y)\right]$ is strictly convex in (h, α) on Θ. Proof in Appendix EC.6.3.*

LEMMA 7. *Under Assumption 2, 3, there are symmetric matrices $A_x(c,d), B_x(c,d)$ such that*

$$A_x(c,d) \preceq \nabla^2 T_{3,x}(c,d) \preceq B_x(c,d)$$

when $d>0$. The entries of $A_x(c,d)$ are given by

$$\begin{aligned}
A_{x,11}(c,d) &= \sum_{i\in\{1,2\}} |\ell_i'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(c+\ell_i^{-1}(d)) + C\cdot \Pr(\ell(Y-c)>d \mid X=x)\\
A_{x,22}(c,d) &= \sum_{i\in\{1,2,\}} \frac{p_{Y|X=x}(c+\ell_i^{-1}(d))}{|\ell_i'(\ell_i^{-1}(d))|},\\
A_{x,12}(c,d) &= p_{Y|X=x}(c+\ell_1^{-1}(d)) - p_{Y|X=x}(c+\ell_2^{-1}(d)).
\end{aligned}$$

The entries of $B_x(c,d)$ are given by

$$\begin{aligned}
B_{x,11}(c,d) &= \sum_{i\in\{1,2\}} |\ell_i'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(c+\ell_i^{-1}(d)) + \mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-c) \mid X=x\right],\\
B_{x,22}(c,d) &= \sum_{i\in\{1,2\}} \frac{p_{Y|X=x}(c+\ell_i^{-1}(d))}{|\ell_i'(\ell_i^{-1}(d))|},\\
B_{x,12}(c,d) &= p_{Y|X=x}(c+\ell_1^{-1}(d)) - p_{Y|X=x}(c+\ell_2^{-1}(d)).
\end{aligned}$$

Proof in Appendix EC.6.4.

LEMMA 8. *Under Assumption 1, 2, 3, $\mathbb{E}_P[L_{RU}^{\Gamma}(h(X),\alpha(X),Y)]$ is Gâteaux differentiable in (h,α) on $L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})$ and twice-Gâteaux differentiable in (h,α) on $\mathcal{C}$, where*

$$\mathcal{C} = \{(h,\alpha)\in\Theta \mid \alpha(x)>0 \quad \forall x\in\mathcal{X}\}.$$

Proof in Appendix EC.6.5.

LEMMA 9. *Under Assumptions 1, 2, 4, 5, 7, for any $h\in\Lambda_c^p(\mathcal{X})$, there exists $\bar{L}(X,Y)$ such that*

$$|L(h(x),y) - L(h_\Gamma^*(x),y))| \le \bar{L}(x,y)\cdot|h(x)-h_\Gamma^*(x)|,$$

where $\sup_{x\in\mathcal{X}} \mathbb{E}_{P_{Y|X}}\left[\bar{L}(x,Y)^2 \mid X=x\right] \le M<\infty$. Proof in Appendix EC.6.6.

LEMMA 10. *Let A be a 2×2 symmetric matrix with $\operatorname{tr}(A)>0$ and $\det(A)\ge 0$. Then*

$$\lambda_{\min}(A) \ge \frac{\det A}{\operatorname{tr} A}, \quad \lambda_{\max}(A) \le \operatorname{tr} A.$$

Proof in Appendix EC.6.7.

EC.5. Proof of Main Results

EC.5.1. Proof of Lemma 1

First, suppose that Q generates P via conditional Γ-biased sampling. We show that (10) holds and that $\sup_{x\in\mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < C$ for some $C < \infty$. By definition, we have that

$$\begin{aligned}\frac{dQ_{Y|X=x}(y)}{dP_{Y|X=x}(y)} &= \frac{d\tilde{Q}_{Y|X}(y)}{d\tilde{Q}_{Y|X,S=1}(y)} \\ &= \frac{d\tilde{Q}_{X,Y}(x,y)}{d\tilde{Q}_X(x)} \cdot \frac{d\tilde{Q}_{X|S=1}(x)}{d\tilde{Q}_{X,Y|S=1}(x,y)} \\ &= \frac{d\tilde{Q}_{X,Y}(x,y)}{d\tilde{Q}_{X,Y|S=1}(x,y)} \cdot \frac{d\tilde{Q}_{X|S=1}(x)}{d\tilde{Q}_X(x)}.\end{aligned}$$

Multiplying numerator and denominator by $\mathbb{P}_{\tilde{Q}}[S=1]$ yields

$$\begin{aligned}\frac{dQ_{Y|X=x}(y)}{dP_{Y|X=x}(y)} &= \Big(\frac{d\tilde{Q}_{X|S=1}(x)}{d\tilde{Q}_X(x)} \cdot \mathbb{P}_{\tilde{Q}}[S=1]\Big) \cdot \Big(\frac{d\tilde{Q}_{X,Y}(x,y)}{d\tilde{Q}_{X,Y|S=1}(x,y)} \cdot \frac{1}{\mathbb{P}_{\tilde{Q}}[S=1]}\Big) && \text{(EC.23)} \\ &= \frac{\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x]}{\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x, Y=y]} && \text{(EC.24)} \\ &\in [\Gamma^{-1}, \Gamma]. && \text{(EC.25)}\end{aligned}$$

(EC.24) follows from Bayes' Rule. (EC.25) follows from (4). So, (10) holds. We also show that the covariate density ratio between P and Q is bounded. By Bayes' rules

$$\mathbb{P}_{\tilde{Q}}[S=1,\, X=x] = \mathbb{P}_{\tilde{Q}}[S=1]\, dP_X(x) = dQ_X(x)\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x],$$

and so

$$\frac{dP_X(x)}{dQ_X(x)} = \frac{\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x]}{\mathbb{P}_{\tilde{Q}}[S=1]} \le \frac{1}{\mathbb{P}_{\tilde{Q}}[S=1]}$$

is in fact uniformly bounded above for all $x \in \mathcal{X}$.

Second, we show the converse. Let Q be a distribution over (X,Y) that satisfies (10). We define $\tilde{Q}$ to be a distribution over (X,Y,S), where $X \in \mathcal{X}, Y \in \mathcal{Y}, S \in \{0,1\}$. We set $\tilde{Q}_{X,Y} = Q$ and define that

$$\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x, Y=y] = \frac{1}{N} \cdot \frac{dP(x,y)}{dQ(x,y)}, \tag{EC.26}$$

where $N \le \frac{1}{C\Gamma}$. Note that $\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x, Y=y] \in [0,1]$ because (10) holds and $\sup_{x\in\mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < C$.

To show that converse holds, we must verify (4) holds for $\tilde{Q}$ and that $\tilde{Q}_{X,Y|S=1} = P$. First, we verify (4). We compute $\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x]$.

$$\mathbb{P}_{\tilde{Q}}[S=1 \mid X=x] = \mathbb{E}_{\tilde{Q}}\left[\frac{1}{N} \cdot \frac{dP(X,Y)}{dQ(X,Y)} \mid X=x\right] = \frac{1}{N} \cdot \frac{dP_X(x)}{dQ_X(x)}. \tag{EC.27}$$

Note that

$$\mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x, Y=y\right] = \frac{1}{N} \cdot \frac{dP(x,y)}{dQ(x,y)} = \mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x\right] \cdot \frac{dP_{Y|X=x}(y)}{dQ_{Y|X=x}(y)}.$$

From (10), we have that

$$\frac{dP_{Y|X=x}(y)}{dQ_{Y|X=x}(y)} \in [\Gamma^{-1}, \Gamma].$$

So, we have that (4) holds for $\tilde{Q}$.

Now, we can verify that $\tilde{Q}_{X,Y|S=1} = P$. We aim to verify that

$$d\tilde{Q}_{X,Y|S=1}(x,y) = dP(x,y). \tag{EC.28}$$

We have that

$$\begin{aligned} d\tilde{Q}_{X,Y,S=1}(x,y) &= \mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x, Y=y\right] \cdot \mathbb{P}_{\tilde{Q}}\left[X=x, Y=y\right] \\ &= \frac{1}{N} \cdot \frac{dP_{X,Y}(x,y)}{dQ_{X,Y}(x,y)} \cdot dQ_{X,Y}(x,y) \\ &= \frac{1}{N} \cdot dP(x,y). \end{aligned}$$

In addition, from (EC.27), we have that

$$\mathbb{P}_{\tilde{Q}}\left[S=1\right] = \mathbb{E}_{\tilde{Q}_X}\left[\mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x\right]\right] = \mathbb{E}_{\tilde{Q}_X}\left[\frac{1}{N} \cdot \frac{dP_X(X)}{dQ_X(X)}\right] = \frac{1}{N}.$$

Thus, we have that

$$d\tilde{Q}_{X,Y|S=1}(x,y) = \frac{d\tilde{Q}_{X,Y,S=1}(x,y)}{\mathbb{P}_{\tilde{Q}}\left[S=1\right]} = dP_{X,Y}(x,y).$$

Therefore, we have that Q can generate P under conditional Γ-biased sampling.

EC.5.2. Proof of Lemma 2

Let $h^* \in L^2(Q_X, \mathcal{X})$ be the solution to (5). Let the function $\tilde{h}$ almost surely be minimizer of (16) at every x. Since $\tilde{h}$ almost surely solves (16) for every $x \in \text{supp}(Q_X)$,

$$\sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(\tilde{h}(X), Y) \mid X=x\right] \leq \sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(h^*(X), Y) \mid X=x\right].$$

Given any marginal distribution Q_X, we can marginalize over X to see that

$$\mathbb{E}_{Q_X}\left[\sup_{Q_{Y|X}:Q \in S_\Gamma(P)} \mathbb{E}_{Q_{Y|X}}\left[L(\tilde{h}(X), Y) \mid X\right]\right] \leq \mathbb{E}_{Q_X}\left[\sup_{Q_{Y|X}:Q \in S_\Gamma(P)} \mathbb{E}_{Q_{Y|X}}\left[L(h^*(X), Y) \mid X=x\right]\right].$$

Based on our definition of $S_\Gamma(P, Q_X)$, we note that for any $h \in L^2(Q_X, \mathcal{X})$

$$\sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h(X), Y)\right] = \mathbb{E}_{Q_X}\left[\sup_{Q_{Y|X}:Q \in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(h(X), Y) \mid X\right]\right].$$

Thus, we have that

$$\sup_{Q\in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(\tilde{h}(X),Y)\right] \le \sup_{Q\in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h^*(X),Y)\right].$$

Finally, by definition of h^* we must also have that

$$\sup_{Q\in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h^*(X),Y)\right] \le \sup_{Q\in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(\tilde{h}(X),Y)\right].$$

These last two inequalities yield the desired equivalence.

EC.5.3. Proof of Theorem 2

First, we establish the existence and uniqueness of the minimizer of the population RU risk. After that, we give a characterization of the optimal α^*_Γ. Finally, we show that there are upper and lower bounds on the value of $\alpha^*_\Gamma(x)$ for all x.

EC.5.3.1. Existence and Uniqueness of The Minimizer We use the following lemma, whose proof is provided at the end of this section.

LEMMA 11. *Suppose Assumption 1 and 2 hold. For any $(h,\alpha)\notin\Theta$, there exists $(\bar{h},\bar{\alpha})\in\Theta$ such that $\mathbb{E}_P\left[L^\Gamma_{RU}(\bar{h}(X),\bar{\alpha}(X),Y)\right] < \mathbb{E}_P\left[L^\Gamma_{RU}(h(X),\alpha(X),Y)\right]$ Proof in Appendix EC.5.3.4.*

Further, Lemma 6 implies that at most one minimizer exists over Θ. Thus, it remains to prove the existence of a minimizer over Θ.

To show this result, we use the following lemma.

LEMMA 12 **(Theorem 2.6.1, Balakrishnan (2012))**. *A continuous convex functional defined on a Hilbert space achieves its minimum on every convex closed bounded set.*

Now we prove that there exists a minimizer of the population RU risk over $\mathcal{H}\times\mathcal{A}$. Clearly, the population RU risk is continuous. We have the RU loss is convex from the first part of Theorem 1, so the population RU risk is also convex in (h,α). In addition, $\Theta\subset L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})$, which is a Hilbert space. In addition, since L^∞ balls are closed in $L^2(P_X,\mathcal{X})$, and Θ consists of a product of L^∞ balls (one of which is not centered at 0), so Θ is closed in $L^2(P_X,\mathcal{X})$. Also, Θ is convex and bounded. Thus Lemma 12 holds, so $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ must achieve a minimum on Θ.

EC.5.3.2. Characterization of The Minimizer Let $L(h,\alpha)=\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ as the population RU risk. Since $L(h,\alpha)$ is Gâteaux differentiable (Lemma 8) and has a unique minimizer at $(h^*_\Gamma,\alpha^*_\Gamma)$ (first part of Theorem 2), the Gâteaux derivative in the direction ϕ is equal to 0 for all $\phi\in L^2(P_X,\mathcal{X})$, i.e.

$$L'_\alpha(h^*_\Gamma,\alpha^*_\Gamma;\phi)=0,\quad \forall\phi\in L^2(P_X,\mathcal{X}).$$

Recall that from Lemma 8, we have that

$$L'_\alpha(h,\alpha;\phi) = (1-\Gamma^{-1})\mathbb{E}_{P_X}\left[\phi(X)\right] + (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[T^d_{3,X}(h^*_\Gamma(X),\alpha^*_\Gamma(X))\phi(X)\right].$$

So, at $(h^*_\Gamma,\alpha^*_\Gamma)$, we have that

$$\mathbb{E}_{P_X}\left[\phi(X)\cdot\left(\frac{1-\Gamma^{-1}}{\Gamma-\Gamma^{-1}} + T^d_{3,X}(h^*_\Gamma(X),\alpha^*_\Gamma(X))\right)\right] = 0, \quad \forall \phi \in L^2(P_X,\mathcal{X}).$$

We note that by Lemma 5,

$$T^d_{3,x}(h(x),\alpha(x)) = -1 + F_{x;h(x)}(\alpha(x)),$$

where $F_{x;h(x)}$ is the distribution over $L(h(x),Y)$ where Y is distributed according to $P_{Y|X=x}$. So, we have that

$$\mathbb{E}_{P_X}\left[\phi(X)\cdot(-\eta(\Gamma) + F_{X,h^*_\Gamma(X)}(\alpha^*_\Gamma(X))\right] = 0, \quad \forall \phi \in L^2(P_X,\mathcal{X}).$$

So, $-\eta(\Gamma) + F_{x,h^*_\Gamma(x)}(\alpha^*_\Gamma(x))$ must be equal to 0 almost everywhere for the above equation to hold for all ϕ. Therefore, we conclude that

$$\alpha^*_\Gamma(x) = F^{-1}_{x;h^*_\Gamma(x)}(\eta(\Gamma)) = q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x)).$$

EC.5.3.3. Bounds on The Minimizer Now, with this definition of α^*_Γ, we can show that there exists $M^- > 0$ such that $\alpha^*_\Gamma(x) > M^-$ for all $x \in \mathcal{X}$. Recall that Lemma 11 gives us that any minimizer of the population risk must lie in the bounded function class Θ So, we must have that

$$\alpha^*_\Gamma(x) = q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x)) \geq \inf_{x\in\mathcal{X},h\in\mathcal{H}} q^L_{\eta(\Gamma)}(x;h(x)).$$

So, it suffices to bound the term on the right. For convenience, we define

$$m(x;h(x)) := q^L_{\frac{1}{2}}(x;h(x)).$$

We note that $\eta(\Gamma) > \frac{1}{2}$. So,

$$q^L_{\eta(\Gamma)}(x;h(x)) \geq m(x;h(x)).$$

We have that for any $x \in \mathcal{X}, h \in \mathcal{H}$,

$$\Pr(L(h(X),Y) \leq m(X;h(X)) \mid X = x) = \frac{1}{2}.$$

Recall that under Assumption 2, $L(h(x),y) = \ell(y-h(x))$. We can apply Assumption 2 to see that

$$\Pr\left(Y \in \left[h(x) + \ell_2^{-1}(m(x;h(x))), h(x) + \ell_1^{-1}(m(x;h(x)))\right] \mid X = x\right) = \frac{1}{2}.$$

Recall under Assumption 3, the density of $p_{Y|X=x}$ is uniformly upper bounded by $P_{\max}$. As a result, we have that

$$P_{\max} \cdot \left\{ \ell_1^{-1}(m(x;h(x))) - \ell_2^{-1}(m(x;h(x))) \right\} \geq \frac{1}{2}.$$

Rearranging, we have that

$$\ell_1^{-1}(m(x;h(x))) - \ell_2^{-1}(m(x;h(x))) \geq \frac{1}{2P_{\max}}.$$

So,

$$\max\{\ell_1^{-1}(m(x;h(x))), -\ell_2^{-1}(m(x;h(x)))\} \geq \frac{1}{4P_{\max}}.$$

Applying ℓ to both sides, we conclude that

$$m(x;h(x)) \geq \ell\left(\frac{1}{4P_{\max}}\right).$$

Since $\frac{1}{4P_{\max}} > 0$, we have that $m(x;h(x))$ is lower bounded by a positive constant for any choice of $h \in \mathcal{H}, x \in \mathcal{X}$ and $\Gamma > 1$. Thus, we have that

$$\alpha^*_\Gamma(x) = q^L_{\eta(\Gamma)}(x;h(x)) \geq \inf_{x\in\mathcal{X}, h\in\mathcal{H}} q^L_{\frac{1}{2}}(x;h(x)) \geq \ell\left(\frac{1}{4P_{\max}}\right).$$

So, let $M^- = \ell(\frac{1}{4P_{\max}})/2$. Then $\alpha^*(x) > M^-$ for all $x \in \mathcal{X}$.

In addition, it is straightforward to see that

$$\begin{aligned}
\alpha^*(x) &= q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x)) \\
&\leq \sup_{h:|h|_\infty \leq 2B} q^L_{\eta(\Gamma)}(x;h(x)) \\
&< \sup_{h:|h|_\infty \leq 2B} q^L_{\eta(\Gamma)}(x;h(x)) + 1 \\
&= M^+_\Gamma,
\end{aligned}$$

where M^+_Γ is defined in (EC.17).

EC.5.3.4. Proof of Lemma 11 We note that $M^+_\Gamma < \infty$ because $\mathcal{X}$ is compact and $\mathcal{H}$ is bounded. Suppose for the sake of contradiction (h,α) is a minimizer of the population RU risk and $(h,\alpha) \notin \Theta$. There are three cases

1. $(h,\alpha) \in \mathcal{H}^c \times \mathcal{A}$,
2. $(h,\alpha) \in \mathcal{H} \times \mathcal{A}^c$,
3. $(h,\alpha) \in \mathcal{H}^c \times \mathcal{A}^c$.

First, we focus on the case where $(h,\alpha) \in \mathcal{H}^c \times \mathcal{A}$. We consider $\bar{h}$,

$$\bar{h}(x) = \begin{cases} h(x) & h(x) \in [-2B, 2B] \\ 2B & h(x) > 2B \\ -2B & h(x) < -2B \end{cases}.$$

We note that $(\bar{h}, \alpha) \in \Theta$. We define $R_{h,-2B}$ and $S_{h,2B}$ following (EC.21) and (EC.22).

$$\begin{aligned}
&\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right] - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y)\right] \\
&= \mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y)\mathbb{I}(R_{h,-2B})\right] \\
&\quad + \mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y)\mathbb{I}(S_{h,2B})\right]
\end{aligned}$$

because they only differ on $R_{h,-2B}$ and $S_{h,2B}$. Analyzing the second term on the right side above, we see that

$$\begin{aligned}
&\mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y))\cdot\mathbb{I}(S_{h,2B})\right] \\
&= \mathbb{E}_{P_X}\left[\Big(\Gamma^{-1}T_{1,X}(h(X),\alpha(X))) + (\Gamma-\Gamma^{-1})\cdot T_{3,X}(h(X),\alpha(X))\Big)\mathbb{I}(S_{h,2B})\right],
\end{aligned}$$

where $T_{1,X}, T_{3,X}$ are defined in (EC.15) and (EC.16), respectively. For $x \in S_{h,2B}$,

$$\begin{aligned}
&\Gamma^{-1}T_{1,x}(h(x),\alpha(x)) + (\Gamma-\Gamma^{-1})\cdot T_{3,x}(h(x),\alpha(x)) - \Gamma^{-1}T_{1,x}(\bar{h}(x),\alpha(x)) - (\Gamma-\Gamma^{-1})\cdot T_{3,x}(\bar{h}(x),\alpha(x)) \\
&= (h(x)-\bar{h}(x))\cdot\Big(\Gamma^{-1}T^{c}_{1,x}(\tilde{h}(x),\alpha(x)) + (\Gamma-\Gamma^{-1})\cdot T^{c}_{3,X}(\tilde{h}(x),\alpha(x))\Big) \quad \tilde{h}(x)\in[\bar{h}(x),h(x)] && \text{(EC.29a)} \\
&= (h(x)-\bar{h}(x))\cdot\mathbb{E}_{P_{Y|X=x}}\left[\Gamma^{-1}\cdot(-\ell'(Y-\tilde{h}(x))) + (\Gamma-\Gamma^{-1})\cdot(-\ell'(Y-\tilde{h}(x)))\cdot\mathbb{I}(\ell(Y-\tilde{h}(x))>\alpha(x))\right] && \text{(EC.29b)} \\
&\geq (h(x)-\bar{h}(x))\cdot\mathbb{E}_{P_{Y|X=x}}\left[\Gamma^{-1}\cdot(-\ell'(Y-\tilde{h}(x)))\right] && \text{(EC.29c)} \\
&> 0. && \text{(EC.29d)}
\end{aligned}$$

(EC.29a) follows from the Mean Value Theorem, the differentiability of $T_{1,x}$ (Lemma 4), and the differentiability of $T_{3,x}$ (Lemma 5). (EC.29b) follows from Lemma 4 and Lemma 5. The inequality in (EC.29c) comes from the observation that for $x \in S_{h,2B}$, we have that $Y - \tilde{h}(x) \leq -B$ because $Y \in [-B, B]$ and $\tilde{h}(x) \in [2B, h(x)]$. So, $-\ell'(Y-\tilde{h}(x)) > 0$. Meanwhile, $h(x) - \bar{h}(x) > 0$. So, the product of $-\ell'(Y-\tilde{h}(x))\cdot(h(x)-\bar{h}(x)) > 0$. Since $\Pr(\ell(Y-\tilde{h}(x)) > \alpha(x)|X=x) \geq 0$, (EC.29c) holds. For the same reason, (EC.29d) holds as well. Thus, if $S_{h,2B}$ has positive measure, then

$$\mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y))\mathbb{I}(X\in S_{h,2B})\right] > 0.$$

An analogous argument can be used to show that for $R_{h,-2B}$ with positive measure,

$$\mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y))\mathbb{I}(X\in R_{h,-2B})\right] > 0.$$

Thus, as long as $R_{h,-2B}\cup S_{h,2B}$ has positive measure, which must be the case under our assumption that the minimizer $(h,\alpha)\in\mathcal{H}^c\times\mathcal{A}$, then there is $(\bar{h},\alpha)\in\Theta$ that achieves lower population RU risk. This is a contradiction, so the minimizer cannot be in $\mathcal{H}^c\times\mathcal{A}$.

Now, we consider the next case that the minimizer $(h,\alpha)\in\mathcal{H}\times\mathcal{A}^c$. Consider $\bar{\alpha}\in\mathcal{A}$,

$$\bar{\alpha}(x)=\begin{cases}0 & \alpha(x)<0\\ \alpha(x) & 0\le\alpha(x)\le M_\Gamma^+\\ M_\Gamma^+ & \alpha(x)>M_\Gamma^+\end{cases}$$

Note that $(h,\bar{\alpha})\in\Theta$. We define $R_{\alpha,0}$ and S_{α,M_Γ^+} according to (EC.21) and (EC.22), respectively. We have that

$$\begin{aligned}&\mathbb{E}_P\left[L_{\mathrm{RU}}^\Gamma(h(X),\alpha(X),Y)\right]-\mathbb{E}_P\left[L_{\mathrm{RU}}^\Gamma(h(X),\bar{\alpha}(X),Y)\right]\\&=\mathbb{E}_P\left[(L_{\mathrm{RU}}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU}}^\Gamma(h(X),\bar{\alpha}(X),Y))\mathbb{I}(R_{\alpha,0})\right]\\&\quad+\mathbb{E}_P\left[(L_{\mathrm{RU}}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU}}^\Gamma(h(X),\bar{\alpha}(X),Y))\mathbb{I}(S_{\alpha,M_\Gamma^+})\right].\end{aligned}$$

because they only differ on $R_{\alpha,0}$ and S_{α,M_Γ^+}. We find that

$$\begin{aligned}&\mathbb{E}_P\left[(L_{\mathrm{RU}}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU}}^\Gamma(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(R_{\alpha,0})\right]\\&=(1-\Gamma^{-1})\mathbb{E}_P\left[(\alpha(X)-\bar{\alpha}(X))\mathbb{I}(R_{\alpha,0})\right]+(\Gamma-\Gamma^{-1})\mathbb{E}_P\left[(L(h(X),Y)-\alpha(X))_+\cdot\mathbb{I}(R_{\alpha,0})\right]\\&\quad-(\Gamma-\Gamma^{-1})\mathbb{E}_P\left[(L(h(X),Y)-\bar{\alpha}(X))_+\cdot\mathbb{I}(R_{\alpha,0})\right]\\&=(1-\Gamma^{-1})\mathbb{E}_X\left[\alpha(X)\mathbb{I}(R_{\alpha,0})\right]+(\Gamma-\Gamma^{-1})\mathbb{E}_P\left[(L(h(X),Y)-\alpha(X))\mathbb{I}(R_{\alpha,0})\right]\\&\quad-(\Gamma-\Gamma^{-1})\mathbb{E}_P\left[L(h(X),Y)\mathbb{I}(R_{\alpha,0})\right]\\&=(1-\Gamma)\mathbb{E}_P\left[\alpha(X)\cdot\mathbb{I}(R_{\alpha,0})\right].\end{aligned}$$

If $R_{\alpha,0}$ has positive measure, then

$$\mathbb{E}_P\left[(L_{\mathrm{RU}}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU}}^\Gamma(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(R_{\alpha,0})\right]>0$$

because on $R_{\alpha,0}$, we have that $\alpha(X)<0$ and also $(1-\Gamma)<0$. In addition,

$$\begin{aligned}&\mathbb{E}_P\left[(L_{\mathrm{RU}}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU}}^\Gamma(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(S_{\alpha,M_\Gamma^+})\right]\\&=\mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[L_{\mathrm{RU},2}^\Gamma(\alpha(X))-L_{\mathrm{RU},2}^\Gamma(\bar{\alpha}(X))+L_{\mathrm{RU},3}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU},3}^\Gamma(h(X),\bar{\alpha}(X),Y)\mid X\right]\mathbb{I}(S_{\alpha,M_\Gamma^+})\right].\end{aligned}$$

For $x\in S_{\alpha,M_\Gamma^+}$, we compute

$$\begin{aligned}&\mathbb{E}_{P_{Y|X}}\left[L_{\mathrm{RU},2}^\Gamma(\alpha(X))-L_{\mathrm{RU},2}^\Gamma(\bar{\alpha}(X))+L_{\mathrm{RU},3}^\Gamma(h(X),\alpha(X),Y)-L_{\mathrm{RU},3}^\Gamma(h(X),\bar{\alpha}(X),Y)\mid X=x\right]&&\text{(EC.30a)}\\&=\mathbb{E}_{P_{Y|X=x}}\left[(1-\Gamma^{-1})(\alpha(X)-\bar{\alpha}(X))\mid X=x\right]&&\text{(EC.30b)}\\&\quad+\mathbb{E}_{P_{Y|X=x}}\left[(\Gamma-\Gamma^{-1})\Big(T_{3,X}(h(X),\alpha(X))-T_{3,X}(h(X),\bar{\alpha}(X))\Big)\mid X=x\right]&&\text{(EC.30c)}\\&=(1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x))+(\Gamma-\Gamma^{-1})\Big(T_{3,x}(h(x),\alpha(x))-T_{3,x}(h(x),\bar{\alpha}(x))\Big)&&\text{(EC.30d)}\end{aligned}$$

$$= (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\cdot(\alpha(x)-\bar{\alpha}(x))\cdot T^d_{3,x}(h(x),\tilde{\alpha}(x)) \quad \tilde{\alpha}(x)\in[\bar{\alpha}(x),\alpha(x)] \tag{EC.30e}$$

$$= (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\cdot(\alpha(x)-\bar{\alpha}(x))\cdot(-1+F_{x;h(x)}(\tilde{\alpha}(x))) \tag{EC.30f}$$

$$> (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\cdot(\alpha(x)-\bar{\alpha}(x))\cdot(-1+\eta(\Gamma)) \tag{EC.30g}$$

$$= 0. \tag{EC.30h}$$

In the above derivation, we have that (EC.30d) follows from Lemma 5 and Assumption 2. Next, we apply the Mean Value Theorem to $T_{3,x}(c,d)$ to arrive at (EC.30e). After that, we use the definition of $T^d_{3,x}(c,d)$ for $d>0$ from Lemma 5, where $\tilde{\alpha}(x)>0$. Finally, we recall that $F_{x;h(x)}$ is the distribution over $L(h(x),Y)=\ell(Y-h(x))$ when Y is distributed according to $P_{Y|X=x}$. We can show (EC.30g) as follows. Since $\tilde{\alpha}(x)\in[\bar{\alpha}(x),\alpha(x)]$ and $x\in S_{\alpha,M^+_\Gamma}$, we have that

$$F_{x;h(x)}(\tilde{\alpha}(x)) \geq F_{x;h(x)}(\bar{\alpha}(x)) = F_{x;h(x)}(M^+_\Gamma),$$

and we have that

$$q^L_{\eta(\Gamma)}(x;h(x)) = F^{-1}_{x;h(x)}(\eta(\Gamma)) < M^+_\Gamma$$

by the definition of M^+_Γ (EC.17). So, we see that $\eta(\Gamma)<F_{x;h(x)}(M^+_\Gamma)$. In addition, we note that $\alpha(x)-\bar{\alpha}(x)>0$ for $x\in S_{\alpha,M^+_\Gamma}$ and $\Gamma-\Gamma^{-1}>0$. We conclude that if S_{α,M^+_Γ} has positive measure, then

$$\mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(S_{\alpha,M^+_\Gamma})\right] > 0.$$

Thus, as long as $R_{\alpha,0}\cup S_{\alpha,M^+_\Gamma}$ has positive measure, which must be the case because we assumed that $(h,\alpha)\in\mathcal{H}\times\mathcal{A}^c$, there is $(h,\bar{\alpha})\in\Theta$ that achieves lower population RU risk than the minimizer (h,α). This is a contradiction, so any minimizer cannot be in $\mathcal{H}\times\mathcal{A}^c$.

Combining the two previous arguments, we can show that any minimizer also cannot be in $\mathcal{H}^c\times\mathcal{A}^c$. Thus, any minimizer of the population RU risk must lie in Θ.

EC.5.4. Proof of Theorem 3

We define some new notation. Recall that under Assumption 2, we can rewrite $L(z,y)=\ell(y-z)$. Let ℓ_1^{-1} be the inverse of $\ell(z)$ where $z>0$. Let ℓ_2^{-1} be the inverse of $\ell(z)$ where $z\leq 0$. Further, let $0<\epsilon<\frac{1-\eta(\Gamma)}{2P_{\max}}$.

Proof of Strong Convexity. We demonstrate that the population RU risk is strongly convex with constant $\kappa_1(\epsilon)$ in an $||\cdot||_\infty$ ball about minimizer with radius $\delta(\epsilon)$. We show that we can pick the radius $\delta(\epsilon)$ so that for $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$, we have that $0<M^- -\delta(\epsilon)<\alpha<M^+_\Gamma$, where M^+_Γ is as defined in (EC.17).

First, we find a radius $\delta_1(\epsilon)$ where the lower bound will be satisfied. Define

$$g_i(x;h,\alpha) = h(x) + \ell_i^{-1}(\alpha(x)). \tag{EC.31}$$

By continuity, for every $\epsilon > 0$, there exists $0 < \delta_1(\epsilon) < M^-$ such that if $(h,\alpha) \in \mathcal{C}_{\delta_1(\epsilon)}$, then

$$\sup_{x\in\mathcal{X}, i\in\{1,2\}} |g_i(x;h,\alpha) - g_i(x;h^*_\Gamma,\alpha^*_\Gamma)| < \epsilon.$$

Note that for $(h,\alpha) \in C_{\delta_1(\epsilon)}$, we have that $\alpha > 0$ because $||\alpha - \alpha^*_\Gamma||_\infty \le \delta_1(\epsilon)$, $\alpha^*_\Gamma(x) > M^-$ for all $x \in \mathcal{X}$, and finally, $\delta_1(\epsilon) < M^-$.

Second, we find a radius δ_2 where the upper bound will be satisfied. By Theorem 2 and the definition of M^+_Γ in (EC.17),

$$\alpha^*_\Gamma(x) = q^L_{\eta(\Gamma)}(x; h^*_\Gamma(x)) \le \sup_{x\in\mathcal{X}, h:||h||_\infty \le 2B} q^L_{\eta(\Gamma)}(x; h^*_\Gamma(x))$$
$$M^+_\Gamma = \sup_{x\in\mathcal{X}, h:||h||_\infty \le 2B} q^L_{\eta(\Gamma)}(x; h^*_\Gamma(x)) + 1.$$

Let $\delta_2 = \frac{1}{2}$. We note that for $(h,\alpha) < C_{\delta_2}$, we have that $\alpha(x) < \alpha^*_\Gamma(x) + \frac{1}{2} < M^+_\Gamma$ for all $x \in \mathcal{X}$.

Thus, if we pick the radius of the $||\cdot||_\infty$ to be $\delta(\epsilon) := \min(\delta_1(\epsilon), \delta_2)$, then both of these properties will be satisfied. As a result, $\delta(\epsilon)$ is constant that depends on $M^+_\Gamma, M^-, \epsilon$.

Let $L(h,\alpha), L_1(h,\alpha), L_3(h,\alpha)$ be shorthand for the population RU risk, the first term of the population RU risk, and the third term of the population RU risk, respectively.

$$L(h,\alpha) = \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right],$$
$$L_1(h,\alpha) = \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right],$$
$$L_3(h,\alpha) = \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right].$$

We compute the second Gâteaux derivative of the population RU risk.

$$\langle L''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \tag{EC.32a}$$
$$= \langle L_1''(h,\alpha;\psi,\phi) + L_3''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \tag{EC.32b}$$
$$\ge \langle L_3''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \tag{EC.32c}$$
$$= (\Gamma - \Gamma^{-1})\cdot \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix} \nabla^2 T_{3,X}(h(X),\alpha(X)) \begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] \tag{EC.32d}$$
$$\ge (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix} A_X(h(X),\alpha(X)) \begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] \tag{EC.32e}$$

(EC.32b) follows from Lemma 8. (EC.32c) holds because $\langle L_1''(h,\alpha;(\psi,\phi),(\psi,\phi))\rangle \ge 0$ because $L_1(h,\alpha)$ is strongly convex in h (see Section EC.6.3.2 for a proof) and does not depend on α. Next, (EC.32d) follows from Lemma 5. Finally, $A_x(c,d)$ is the lower bound on the Hessisan matrix of

$T_{3,x}(c,d)$ defined in Lemma 7. To develop a lower bound for $\langle L''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle$, we aim to apply Lemma 10 to $A_x(h(x),\alpha(x))$.

Before we verify the conditions and apply Lemma 10, we bound quantities that appear in the trace and determinant of $A_x(h(x),\alpha(x))$ when $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$. These quantities include $a_{i,x}$ for $i\in\{1,2\}, \sum_{i\in\{1,2\}} f_{i,x}$, and $1-F_{x;h(x)}(\alpha(x))$, where

$$a_{i,x} := |\ell_i'(\ell_i^{-1}(\alpha(x)))| \quad i=1,2, \tag{EC.33a}$$

$$f_{i,x} := p_{Y|X=x}(h(x)+\ell_i^{-1}(\alpha(x))) \quad i=1,2. \tag{EC.33b}$$

First, we focus on $a_{i,x}$. We note that for $(h,\alpha)\in C_{\delta(\epsilon)}$, we have $\alpha < M_\Gamma^+$. Since $|\ell_i'(\ell_i^{-1}(y))|$ is strictly increasing in y and on $C_{\delta(\epsilon)}$, $\alpha(x)\geq M^- - \delta(\epsilon)$ for all $x\in\mathcal{X}$. So, we have

$$0 < C_{M^-,\delta(\epsilon)} \leq a_{i,x} \leq C_{M_\Gamma^+} < \infty \quad i=1,2, x\in\mathcal{X},$$

where $C_{M^-,\delta}$ is defined in (EC.19) and $C_{M_\Gamma^+}$ is defined in (EC.18). We note that $C_{M^-,\delta(\epsilon)}$ is a constant that depends on M^- and $\delta(\epsilon)$ and $C_{M_\Gamma^+}$ is a constant that depends on M_Γ^+.

Second, we aim to show that $\sum_{i=\{1,2\}} f_{i,x}$ is similarly upper and lower bounded. The upper bound is straightforward from Assumption 3. To obtain the lower bound, we first analyze $\sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(\alpha_\Gamma^*(x))$, which can be written as $\sum_{i\in\{1,2\}} p_{Y|X=x}(g_i(x;h_\Gamma^*,\alpha_\Gamma^*))$ using the definition of g in (EC.31).

For some quantile $c_{i,x}$, we can write that

$$\ell_i^{-1}(q^L_{\eta(\Gamma)}(x;h_\Gamma^*(x))) = q^Y_{c_{i,x}}(x) - h_\Gamma^*(x),$$

where the first term corresponds to the $c_{i,x}$-th quantile of Y, where Y is distributed following $P_{Y|X=x}$. We realize that

$$\begin{aligned}
\sum_{i\in\{1,2\}} p_{Y|X=x}(g_i(x;h_\Gamma^*,\alpha_\Gamma^*)) &= \sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(\alpha_\Gamma^*(x)) \\
&= \sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(q^L_{\eta(\Gamma)}(x;h_\Gamma^*(x)))) \\
&= \sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+q^Y_{c_{i,x}}(x)-h_\Gamma^*(x)) \\
&= \sum_{i\in\{1,2\}} p_{Y|X=x}(q^Y_{c_{i,x}}(x)).
\end{aligned}$$

Furthermore, we realize that either $c_{1,x}$ or $c_{2,x}$ lies in $[1-\frac{\eta(\Gamma)}{2}, 1+\frac{\eta(\Gamma)}{2}]$. Because $q^L_{\eta(\Gamma)}(x;h_\Gamma^*(x))$ corresponds to the $\eta(\Gamma)$-th quantile of the conditional losses, we must have that

$$c_{1,x} - c_{2,x} = \eta(\Gamma). \tag{EC.34}$$

In addition, $c_{1,x} \leq 1$, so $c_{2,x} \leq 1-\eta(\Gamma)$. So, $c_{2,x} \in [0, 1-\eta(\Gamma)]$. Suppose that $c_{2,x} \in [1-\frac{\eta(\Gamma)}{2}, 1-\eta(\Gamma)]$, then clearly the desired claim holds. If $c_{2,x} \notin [1-\frac{\eta(\Gamma)}{2}, 1-\eta(\Gamma)]$, this means that $c_{2,x} \in [0, 1-\frac{\eta(\Gamma)}{2})$. So, we must have that $c_{1,x} \in [\eta(\Gamma), 1+\frac{\eta(\Gamma)}{2})$. Thus, we have that at least one of $c_{1,x}, c_{2,x}$ lies in the interval $[1-\frac{\eta(\Gamma)}{2}, 1+\frac{\eta(\Gamma)}{2}]$.

Now, we have that

$$\sum_{i\in\{1,2\}} f_{i,x} = \sum_{i\in\{1,2\}} p_{Y|X=x}(g_i(x;h,\alpha)),$$

and $\delta(\epsilon)$ was chosen so that for $(h,\alpha) \in \mathcal{C}_{\delta(\epsilon)}$

$$\sup_{x\in\mathcal{X}, i\in\{1,2\}} |g_i(x;h,\alpha) - g_i(x;h^*_\Gamma,\alpha^*_\Gamma)| = \sup_{x\in\mathcal{X}, i\in\{1,2\}} |g_i(x;h,\alpha) - q^Y_{c_{i,x}}(x)| < \epsilon.$$

Thus, we realize that for $(h,\alpha) \in \mathcal{C}_{\delta(\epsilon)}$,

$$g_i(x;h,\alpha) = q^Y_{c_{i,x}}(x) + b_i(x), \quad b_i(x) \in (-\epsilon,\epsilon), i \in \{1,2\}, x \in \mathcal{X}. \tag{EC.35}$$

So, for $(h,\alpha) \in \mathcal{C}_{\delta(\epsilon)}$, we realize that a lower bound on $\sum_{i\in\{1,2\}} f_{i,x} = \sum_{i\in\{1,2\}} p_{Y|X}(g_i(x;h,\alpha))$ is given by

$$P_{\min,\Gamma,\epsilon} := \inf_{c\in[1-\frac{\eta(\Gamma)}{2}, 1+\frac{\eta(\Gamma)}{2}], b\in[-\epsilon,\epsilon], x\in\mathcal{X}} p_{Y|X=x}(q^Y_c(x)+b). \tag{EC.36}$$

Thus, we have that

$$0 < P_{\min,\Gamma,\epsilon} \leq \sum_{i\in\{1,2\}} f_{i,x} \leq 2P_{\max} < \infty \quad i = 1,2, x \in \mathcal{X},$$

and clearly each $f_{i,x}$ must be nonnegative.

Third, we aim to show that $1 - F_{x;h(x)}(\alpha(x))$ is similarly upper and lower bounded on $\mathcal{C}_{\delta(\epsilon)}$. Clearly, an upper bound on this quantity is 1. To compute the lower bound, we see that for $(h,\alpha) \in \mathcal{C}_{\delta(\epsilon)}$,

$$\begin{aligned}
1 - F_{x;h(x)}(\alpha(x)) &= 1 - P_{Y|X=x}(g_1(x;h,\alpha)) + P_{Y|X=x}(g_2(x;h,\alpha)) \\
&= 1 - P_{Y|X=x}(q^Y_{c_{1,x}}(x) + b_1(x)) + P_{Y|X=x}(q^Y_{c_{2,x}}(x) + b_2(x)) \quad b_1(x), b_2(x) \in (-\epsilon,\epsilon) \\
&\geq 1 - c_{1,x} - P_{\max}\cdot\epsilon + c_{2,x} - P_{\max}\cdot\epsilon \\
&= 1 - \eta(\Gamma) - 2P_{\max}\epsilon \\
&> 0.
\end{aligned}$$

The first line follows from the definition of F and g_i from (EC.31). In the second line, we apply (EC.35). In the third line, we note that the c.d.f. of $P_{Y|X=x}$ at $q^Y_{c_{i,x}}(x) + b_i(x)$ can be closely approximated by the value of the c.d.f. at $q^Y_{c_{i,x}}(x)$. Next, we apply (EC.34). The last line follows because $\epsilon < \frac{1-\eta(\Gamma)}{2P_{\max}}$. Thus, we have that

$$1 - F_{x;h(x)}(\alpha(x)) \geq 1 - \eta(\Gamma) - 2P_{\max}\epsilon > 0. \tag{EC.37}$$

Now, we finally verify the conditions of Lemma 10. We note that $A_x(h(x),\alpha(x))$ is a symmetric matrix by definition. We realize that $\operatorname{tr} A_x(h(x),\alpha(x)) \geq 0$ because

$$\begin{aligned}
\operatorname{tr} A_x(h(x),\alpha(x)) &= A_{x,11}(h(x),\alpha(x)) + A_{x,22}(h(x),\alpha(x)) && \text{(EC.38)}\\
&= \sum_{i\in\{1,2\}} a_{i,x}\cdot f_{i,x} + \sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} + C(1-F_{x;h(x)}(\alpha(x))) && \text{(EC.39)}\\
&\geq C(1-F_{x;h(x)}(\alpha(x))) && \text{(EC.40)}\\
&> 0. && \text{(EC.41)}
\end{aligned}$$

(EC.40) follows from the observation that $f_{i,x}, a_{i,x} \geq 0$. (EC.41) follows from (EC.37). In addition, we see that $\det A_x(h(x),\alpha(x)) \geq 0$ because

$$\begin{aligned}
&\det A_x(h(x),\alpha(x)) && \text{(EC.42a)}\\
&= A_{x,11}(h(x),\alpha(x))\cdot A_{x,22}(h(x),\alpha(x)) - (A_{x,12}(h(x),\alpha(x)))^2 && \text{(EC.42b)}\\
&= \Big(\sum_{i\in\{1,2\}} a_{i,x}\cdot f_{i,x} + C(1-F_{x;h(x)}(\alpha(x)))\Big)\cdot\Big(\sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}}\Big) - (f_{1,x}-f_{2,x})^2 && \text{(EC.42c)}\\
&= \Big(\frac{a_{1,x}}{a_{2,x}} + \frac{a_{2,x}}{a_{1,x}} + 2\Big)\cdot f_{1,x}\cdot f_{2,x} + C(1-F_{x;h(x)}(\alpha(x)))\cdot\Big(\sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}}\Big) && \text{(EC.42d)}\\
&\geq C\cdot\Big(\sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}}\Big)\cdot(1-F_{x;h(x)}(\alpha(x))) && \text{(EC.42e)}\\
&\geq C\cdot\frac{1}{C_{M_\Gamma^+}}\cdot\Big(\sum_{i\in\{1,2\}} f_{i,x}\Big)\cdot(1-\eta(\Gamma)-2P_{\max}\epsilon) && \text{(EC.42f)}\\
&\geq C\cdot\frac{1}{C_{M_\Gamma^+}}\cdot P_{\min,\Gamma,\epsilon}\cdot(1-\eta(\Gamma)-2P_{\max}\epsilon) && \text{(EC.42g)}\\
&> 0. && \text{(EC.42h)}
\end{aligned}$$

Thus, we can apply Lemma 10 to $A_x(h(x),\alpha(x))$ to see that

$$\lambda_{\min}(A_x(h(x),\alpha(x)) \geq \frac{\det A_x(h(x),\alpha(x))}{\operatorname{tr} A_x(h(x),\alpha(x))}.$$

We can combine the lower bound on $\det A$ from (EC.42g) with the following upper bound on $\operatorname{tr} A$ to find a lower bound on $\lambda_{\min}(A_x(h(x),\alpha(x))$ that does not depend on the choice of $x\in\mathcal{X}$ and $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$.

$$\begin{aligned}
\operatorname{tr} A_x(h(x),\alpha(x)) &= \sum_{i\in\{1,2\}} a_{i,x}\cdot f_{i,x} + \sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} + C(1-F_{x;h(x)}(\alpha(x))) && \text{(EC.43a)}\\
&\leq 2P_{\max}(C_{M_\Gamma^+} + \frac{1}{C_{M^-,\delta(\epsilon)}}) + C && \text{(EC.43b)}\\
&= \frac{2P_{\max}(C_{M_\Gamma^+}\cdot C_{M^-,\delta(\epsilon)}+1) + C\cdot C_{M^-,\delta(\epsilon)}}{C_{M^-,\delta(\epsilon)}} && \text{(EC.43c)}
\end{aligned}$$

Therefore, applying(EC.43c) and (EC.42g), we find that

$$\begin{aligned}
&\lambda_{\min}(A_x(h(x),\alpha(x))\\
&\geq \frac{\det A_x(h(x),\alpha(x))}{\operatorname{tr} A_x(h(x),\alpha(x))}\\
&\geq \left(C\cdot(1-\eta(\Gamma)-2P_{\max}\epsilon)\cdot\frac{1}{C_{M_\Gamma^+}}\cdot P_{\min,\Gamma,\epsilon}\right)\cdot\left(\frac{C_{M^-,\delta(\epsilon)}}{2P_{\max}(C_{M_\Gamma^+}\cdot C_{M^-,\delta(\epsilon)}+1)+C\cdot C_{M^-,\delta(\epsilon)}}\right)\\
&\geq \frac{C\cdot(1-\eta(\Gamma)-2P_{\max}\cdot\epsilon)\cdot P_{\min,\Gamma,\epsilon}}{2P_{\max}\cdot(C_{M_\Gamma^+}\cdot C_{M^-,\delta(\epsilon)}+1)+C\cdot C_{M^-,\delta(\epsilon)}}\cdot\frac{C_{M^-,\delta(\epsilon)}}{C_{M_\Gamma^+}}.
\end{aligned}$$

We define $\kappa_1(\epsilon)$ as follows

$$\kappa_1(\epsilon):=(\Gamma-\Gamma^{-1})\cdot\frac{C\cdot(1-\eta(\Gamma)-2P_{\max}\cdot\epsilon)\cdot P_{\min,\Gamma,\epsilon}}{2P_{\max}\cdot(C_{M_\Gamma^+}\cdot C_{M^-,\delta(\epsilon)}+1)+C\cdot C_{M^-}}\cdot\frac{C_{M^-,\delta(\epsilon)}}{C_{M^+}}.$$

We realize that for all $x\in\mathcal{X},(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$, we have that

$$(\Gamma-\Gamma^{-1})\cdot A_x(h(x),\alpha(x))\succeq\kappa_1(\epsilon)\cdot I_2.$$

Revisiting (EC.32e), we have that

$$\begin{aligned}
\langle L''(h,\alpha;(\psi,\phi)),(\psi,\phi)\rangle &\geq (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X)&\phi(X)\end{bmatrix}A_X(h(X),\alpha(X))\begin{bmatrix}\psi(X)\\\phi(X)\end{bmatrix}\right] &\text{(EC.44)}\\
&\geq \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X)&\phi(X)\end{bmatrix}\kappa_1(\epsilon)I_2\begin{bmatrix}\psi(X)\\\phi(X)\end{bmatrix}\right] &\text{(EC.45)}\\
&= \kappa_1(\epsilon)\mathbb{E}_{P_X}\left[\psi(X)^2+\phi(X)^2\right] &\text{(EC.46)}\\
&= \kappa_1(\epsilon)||(\psi,\phi)||^2. &\text{(EC.47)}
\end{aligned}$$

Thus, $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ is $\kappa_1(\epsilon)$ strongly convex in (h,α) on $\mathcal{C}_{\delta(\epsilon)}$.

We investigate the behavior of $\kappa_1(\epsilon)$ as $\epsilon\to 0$. We can show that as $\epsilon\to 0$, $\kappa_1(\epsilon)\to\kappa$, where

$$\kappa_1:=(1-\Gamma^{-1})\cdot\frac{C\cdot P_{\min,\Gamma}}{2P_{\max}\cdot(C_{M_\Gamma^+}\cdot C_{M^-}+1)+C\cdot C_{M^-}}\cdot\frac{C_{M^-}}{C_{M_\Gamma^+}}, \tag{EC.48}$$

where C_{M^-} is defined in (EC.20).

To see this, note that if $\epsilon\to 0$, then $\delta(\epsilon)\to 0$. So, we have that as $\epsilon\to 0$, $C_{M^-,\delta(\epsilon)}\to C_{M^-}$, $P_{\min,\Gamma,\epsilon}\to P_{\min,\Gamma}$, and $\epsilon\cdot P_{\max}\to 0$. So, we have that

$$\begin{aligned}
\lim_{\epsilon\to 0}\kappa_1(\epsilon) &= \lim_{\epsilon\to 0}(\Gamma-\Gamma^{-1})\frac{C\cdot(1-\eta(\Gamma)-2P_{\max}\cdot\epsilon)\cdot P_{\min,\Gamma,\epsilon}}{2P_{\max}\cdot(C_{M_\Gamma^+}\cdot C_{M^-,\delta(\epsilon)}+1)+C\cdot C_{M^-,\delta(\epsilon)}}\cdot\frac{C_{M^-,\delta(\epsilon)}}{C_{M_\Gamma^+}}\\
&= (\Gamma-\Gamma^{-1})\cdot\frac{C\cdot(1-\eta(\Gamma))\cdot P_{\min,\Gamma}}{2P_{\max}\cdot(C_{M_\Gamma^+}\cdot C_{M^-}+1)+C\cdot C_{M^-}}\cdot\frac{C_{M^-}}{C_{M_\Gamma^+}}\\
&= (1-\Gamma^{-1})\cdot\frac{C\cdot P_{\min,\Gamma}}{2P_{\max}\cdot(C_{M_\Gamma^+}\cdot C_{M^-}+1)+C\cdot C_{M^-}}\cdot\frac{C_{M^-}}{C_{M_\Gamma^+}}\\
&= \kappa_1.
\end{aligned}$$

Thus, as $\epsilon \to 0$, then $\kappa_1(\epsilon) \to \kappa_1$, where κ_1 is defined in (EC.48). Finally, we note that M_Γ^+ is a constant that depends on B, Γ (as defined in (EC.17)) and loss function L, and M^- is a constant that depends on $P_{\max}$ and loss function L, C_{M^-} depends on M^-, and $C_{M_\Gamma^+}$ depends on M_Γ^+, so κ_1 is a constant that depends on $B, C, P_{\max}, P_{\min,\Gamma}, \Gamma$, and the loss function L.

Proof of Smoothness. To show that the population RU risk is κ_2-smooth on $\mathcal{C}_{\delta(\epsilon)}$, we show that

$$\langle L''_{h\alpha}(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \le \kappa_2 ||(\psi,\phi)||^2_{L^2(P_X,\mathcal{X})}.$$

We have that

$$\begin{aligned}
&\langle L''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \\
&= \langle L_1''(h,\alpha;\psi,\phi) + L_3''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \\
&\le \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix}\cdot\left(\Gamma^{-1}\nabla^2 T_{1,X}(h(X),\alpha(X)) + (\Gamma-\Gamma^{-1})\nabla^2 T_{3,X}(h(X),\alpha(X))\right)\begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] \\
&\le \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix}\cdot\left(\Gamma^{-1}\nabla^2 T_{1,X}(h(X),\alpha(X)) + (\Gamma-\Gamma^{-1})\nabla^2 B_X(h(X),\alpha(X))\right)\begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right].
\end{aligned}$$

The second line follows from Lemma 8. The matrix $B_x(h(x),\alpha(x))$ is as defined in Lemma 7. It suffices to show that there is $\kappa_2(\epsilon)$ such that

$$\Gamma^{-1}\nabla^2 T_{1,x}(h(x),\alpha(x)) + (\Gamma-\Gamma^{-1})B_x(h(x),\alpha(x)) \preceq \kappa_2(\epsilon) I_2 \quad \forall x \in \mathcal{X}.$$

Applying Lemma 4 and Assumption 4, we have that

$$\nabla^2 T_{1,x}(h(x),\alpha(x)) = \begin{bmatrix}\mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-h(x))\right] & 0\\ 0 & 0\end{bmatrix} \preceq D I_2. \tag{EC.49}$$

From the proof of strong convexity, for $0 < \epsilon < \frac{1-\eta(\Gamma)}{2P_{\max}}$, there exists $0 < \delta(\epsilon) < M^-$ so that for $(h,\alpha) \in C_{\delta(\epsilon)}$, $(\Gamma-\Gamma^{-1})\nabla^2 T_{3,x}(h(x),\alpha(x))$ is positive definite. So, on this set $\mathcal{C}_{\delta(\epsilon)}$, $B_x(h(x),\alpha(x))$ is also certainly positive definite. So, $B_x(h(x),\alpha(x))$ satisfies the conditions of Lemma 10, so we can conclude that $\lambda_{\max}(B_x(h(x),\alpha(x))) \le \operatorname{tr} B_x(h(x),\alpha(x))$. We can write that

$$\operatorname{tr} B_x(h(x),\alpha(x)) = \sum_{i\in\{1,2\}} a_{i,x}\cdot f_{i,x} + \sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} + \mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-h(x))\right],$$

where $a_{i,x}$ and $f_{i,x}$ are defined in (EC.33a) and (EC.33b), respectively.

We note that for $(h,\alpha) \in \mathcal{C}_{\delta(\epsilon)}$, we have that $0 < M^- - \delta(\epsilon) < \alpha(x) < M_\Gamma^+$ for all $x \in \mathcal{X}$. We can use this property to upper bound $\operatorname{tr} B_x(h(x),\alpha(x))$. Recall $C_{M^-,\delta}$ from (EC.19) and $C_{M_\Gamma^+}$ from (EC.18).

$$\begin{aligned}\operatorname{tr} B_x(h(x),\alpha(x)) &= \sum_{i\in\{1,2\}} a_{i,x}\cdot f_{i,x} + \sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} + \mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-h(x))\right] \\ &\le 2P_{\max}(C_{M_\Gamma^+} + \frac{1}{C_{M^-,\delta(\epsilon)}}) + D.\end{aligned}$$

We arrive at the second inequality by recalling the definition of $P_{\max}$ from Assumption 3, $C_{M_\Gamma^+}$ from (EC.18), $C_{M^-,\delta}$ from (EC.19), and D from Assumption 4. So, we have that

$$B_x(h(x),\alpha(x)) \preceq \Big(2P_{\max}(C_{M_\Gamma^+} + \frac{1}{C_{M^-}}) + D\Big) I_2. \tag{EC.50}$$

Let

$$\kappa_2(\epsilon) := (\Gamma - \Gamma^{-1})\cdot\Big(2P_{\max}\Big(C_{M_\Gamma^+} + \frac{1}{C_{M_-,\delta(\epsilon)}}\Big)\Big) + \Gamma\cdot D. \tag{EC.51}$$

Combining the constants from (EC.49) and (EC.50), we have that for $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$

$$\Gamma^{-1}\nabla^2 T_{1,x}(h(x),\alpha(x)) + (\Gamma-\Gamma^{-1})B_x(h(x),\alpha(x)) \preceq \kappa_2(\epsilon) I_2 \quad \forall x\in\mathcal{X}.$$

Thus, we conclude that $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ is $\kappa_2(\epsilon)$-smooth in (h,α) on $\mathcal{C}_{\delta(\epsilon)}$. Let

$$\kappa_2 := (\Gamma - \Gamma^{-1})\cdot\Big(2P_{\max}\Big(C_{M_\Gamma^+} + \frac{1}{C_{M^-}}\Big)\Big) + \Gamma\cdot D. \tag{EC.52}$$

As $\epsilon\to 0$, $\delta(\epsilon)\to 0$. So, $C_{M_-,\delta(\epsilon)}\to C_{M^-}$. This implies that $\kappa_2(\epsilon)\to\kappa_2$ as the radius of the $||\cdot||_\infty$-ball shrinks. Finally, we note that M_Γ^+ is a constant that depends on B and Γ (as defined in (EC.17)), M^- is a constant that depends on $P_{\max}$, and C_{M^-} depends on M^-, and $C_{M_\Gamma^+}$ depends on M_Γ^+, so κ_2 is a constant that depends on $B, D, P_{\max}, \Gamma$ and the loss function L.

EC.5.5. Proof of Lemma 3

We demonstrate the smoothness of α^*_Γ through an application of Theorem 5, an implicit function theorem that we derive for Hölder-smooth functions. First, we specify a function v that yields an implicit representation of the optimal α^*_Γ. Second, we verify that v satisfies the conditions of Theorem 5, which requires (1) defining a neighborhood where the derivative of v with respect to one of its arguments is a bounded away from zero and (2) guaranteeing that v lies in a p-Hölder ball with radius b. Finally, we apply Theorem 5 to guarantee that α^*_Γ has the desired smoothness.

EC.5.5.1. Specification of Implicit Function By Theorem 2, the population RU risk has a unique minimizer $(h^*_\Gamma,\alpha^*_\Gamma)$ over $L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})$. By Theorem 2, we have that $\alpha^*_\Gamma = q^L_{\eta(\Gamma)}(x;h(x))$ and $\alpha^*_\Gamma(x)\in[M^-,M^+_\Gamma]$.

Since $\alpha^*_\Gamma = q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x))$, we have that

$$\mathbb{P}_{Y|X=x}\left[L(h^*_\Gamma(x),Y)\le\alpha^*_\Gamma(x)\right] = \eta(\Gamma) \quad \forall x\in\mathcal{X}.$$

Under Assumption 2, this condition is equivalent to

$$\mathbb{P}_{Y|X=x}\left[h_\Gamma^*(x)+\ell_2^{-1}(\alpha_\Gamma^*(x)) \le Y \le h_\Gamma^*(x)+\ell_1^{-1}(\alpha_\Gamma^*(x))\right] = \eta(\Gamma) \quad \forall x \in \mathcal{X}$$

where ℓ_1, ℓ_2 are defined as in (EC.14).

In other words, let $v : \mathcal{X} \times \mathcal{T}$, where $\tilde{\mathcal{T}} = [M^-, M_\Gamma^+]$. We define

$$v(x,t) := P_{Y|X=x}(h_\Gamma^*(x)+\ell_1^{-1}(t)) - P_{Y|X=x}(h_\Gamma^*(x)+\ell_2^{-1}(t)) - \eta(\Gamma).$$

Note that α_Γ^* is the unique solution to $v(x,t(x))=0$ by Theorem 2 and $\alpha_\Gamma^*(x)$ takes value in the interior of $\mathcal{X}\times\tilde{\mathcal{T}}$. Thus, v yields an implicit representation of α_Γ^*.

EC.5.5.2. Derivative of Implicit Function Bounded Away From Zero We aim to define a neighborhood $\mathcal{X}\times\mathcal{T} \subset \mathcal{X}\times\tilde{\mathcal{T}}$ where $\frac{\partial}{\partial t}v$ is lower bounded away from zero, i.e. there exists constants K_1, K_2 such that $0 < K_1 \le \frac{\partial}{\partial t}v(x,t) \le K_2 < \infty$.

We have that v is differentiable and

$$\begin{aligned}\frac{\partial}{\partial t}v(x,t) &= p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(t))\cdot\frac{d}{dt}\ell_i^{-1}(t) - p_{Y|X=x}(h_\Gamma^*(x)+\ell_2^{-1}(t))\cdot\frac{d}{dt}\ell_2^{-1}(t)\\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(t))\cdot\frac{1}{|\ell'(\ell_i^{-1}(t))|},\end{aligned}$$

where the second line follow from the differentiability of ℓ, which gives that ℓ_i^{-1} is differentiable with $\frac{d}{dt}\ell_i^{-1}(t) = \frac{1}{\ell'(\ell_i^{-1}(t))}$. Since ℓ is strictly increasing on $[M^-, M_\Gamma^+]$ and ℓ is strictly decreasing on $[-M_\Gamma^+, -M^-]$, $\frac{d}{dt}\ell_i^{-1}(t)$ is bounded away from zero on $\tilde{\mathcal{T}}$.

First, it is straightforward to see that $\frac{\partial}{\partial t}v(x,t)$ is bounded above by a constant K_2 that depends on $P_{\max}, M^-, M_\Gamma^+$ for all $x\in\mathcal{X}, t\in\tilde{\mathcal{T}}$.

Second, we aim to show that there exists a neighborhood containing $(x, \alpha_\Gamma^*(x))$ where $\frac{\partial}{\partial t}v(x,t)$ is lower bounded away from zero.

As an intermediate step, we first show that $\frac{\partial}{\partial t}v(x,t)$ is lower bounded away from zero at $(x,\alpha_\Gamma^*(x))$ for all $x\in\mathcal{X}$. It suffices to show that $f(x,\alpha_\Gamma^*(x))$, where $f(x,t) = \sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(t))$ is lower bounded away from zero by $P_{\min,\Gamma}$ for all $x\in\mathcal{X}$. This result holds by an argument that is also used in the proof of Theorem 3. For some quantile $c_{i,x}$, we can write that

$$\ell_i^{-1}(q^L_{\eta(\Gamma)}(x; h_\Gamma^*(x))) = q^Y_{c_{i,x}}(x) - h_\Gamma^*(x),$$

where the first term corresponds to the $c_{i,x}$-th conditional quantile of Y. Then, we have that

$$\begin{aligned}&f(x,\alpha^*(x))\\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(h_\Gamma^*(x)+\ell_i^{-1}(\alpha_\Gamma^*(x))\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in\{1,2\}} p_{Y|X=x}(h^*_\Gamma(x) + \ell_i^{-1}(q^L_{\eta(\Gamma)}(x; h^*_\Gamma(x))))\\
&= \sum_{i\in\{1,2\}} p_{Y|X=x}(q^Y_{c_{i,x}}(x))\\
&\geq P_{\min,\Gamma}.
\end{aligned}
$$

The last inequality follows because that either $c_{1,x}$ or $c_{2,x}$ must lie in $[1-\frac{\eta(\Gamma)}{2}, 1+\frac{\eta(\Gamma)}{2}]$ (full argument provided in the proof of Theorem 3).

Furthermore, we show that $f(x,t) = \sum_{i\in\{1,2\}} p_{Y|X=x}(h^*_\Gamma(x) + \ell_i^{-1}(t)) \in \Lambda^p(\mathcal{X}\times\tilde{\mathcal{T}})$. Under Assumption 5, we have that $P_{Y|X} \in \Lambda^{p+1}(\mathcal{X}\times\mathcal{Y})$, so $p_{Y|X} \in \Lambda^p(\mathcal{X}\times\mathcal{Y})$. In addition, since ℓ_1^{-1} is strictly increasing and ℓ_2^{-1} is strictly decreasing, we can apply Assumption 5 and Corollary 3, we have that $\ell_i^{-1} \in \Lambda^p_{C_{M^-,M^+_\Gamma,p}}(\mathcal{T})$ for $i=1,2$ for a constant $C_{M^-,M^+,p}$ that depends on M^-, M^+_Γ, p. Since $p_{Y|X}, h^*_\Gamma, \ell_i^{-1}$ are all p-Hölder, f is also p-Hölder by Lemma 18. Since f is p-Hölder and $\alpha^*_\Gamma(x)$ takes value in the interior of $\tilde{\mathcal{T}}$ for all $x\in\mathcal{X}$, there must exist a compact neighborhood $\mathcal{X}\times\mathcal{U}\subset\mathcal{X}\times\tilde{\mathcal{T}}$ that contains $\{(x,\alpha^*_\Gamma(x))\}_{x\in\mathcal{X}}$ where f is lower bounded away from zero by a constant that depends on $P_{\min,\Gamma}, M^-$, and M^+_Γ. This implies that there exists a neighborhood $\mathcal{X}\times\mathcal{T}$ containing $\{(x,\alpha^*_\Gamma(x))\}_{x\in\mathcal{X}}$ where $\frac{\partial}{\partial t}v(x,t)$ is lower bounded away from zero by a constant K_1 that depends on $P_{\min,\Gamma}, M^-$, and M^+_Γ.

Therefore, there exists a neighborhood $\mathcal{X}\times\mathcal{T}\subset\mathcal{X}\times\tilde{\mathcal{T}}$ containing $(x,\alpha^*_\Gamma(x))$ where $\frac{\partial}{\partial t}v(x,t)$ is bounded above and below by positive constants K_1, K_2.

EC.5.5.3. Bound on the Norm of Implicit Function We show that $v\in\Lambda^p_b(\mathcal{X}\times\mathcal{T})$ for some constant b that depends on $M^-, M^+_\Gamma, P_{\min,\Gamma}, p$.

To bound the norm of v, we express v as a composition of functions:

$$
v(x,t) = \sum_{i\in\{1,2\}} g(x, y_i(x,t)) - \eta(\Gamma),
$$

where

$$
y_i(x,t) := h^*_\Gamma(x) + \ell_i^{-1}(t) \quad i=1,2
$$

and

$$
g(x,y) = P_{Y|X=x}(y).
$$

Case 1: $p\leq 1$. If $p\leq 1$, then

$$
\begin{aligned}
||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} &\leq 2(||g||_{\Lambda^p(\mathcal{X}\times\mathcal{Y})} + K_2\cdot||y_i||_{\Lambda^p(\mathcal{X}\times\mathcal{T})})\\
&\leq 2(c + K_2||h^*_\Gamma(x) + \ell_i^{-1}(t)||_{\Lambda^p(\mathcal{X}\times\mathcal{T})})\\
&\leq 2c\cdot(||h^*_\Gamma||_{\Lambda^p(\mathcal{X})} + ||\ell_i^{-1}||_{\Lambda^p(\mathcal{T})}))\\
&\leq 2c\cdot(c + K_2\cdot(c + C_{M^+,M^-_\Gamma,p})).
\end{aligned}
$$

The first line follows from Lemma 18. The second line follows from Assumption 5. The third line follows from the definition of $y(x,t)$. The fourth line follows from the assumption that $h_\Gamma^* \in \Lambda_c^p(\mathcal{X})$ and the result that $\ell_i^{-1} \in \Lambda^p_{C_{M^+,M^-,p}}(\mathcal{T})$.

Case 2: $p > 1$. Similarly, if $p > 1$, then we have

$$\begin{aligned}
||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} &\leq 2||g||_{\Lambda^p(\mathcal{X}\times\mathcal{Y})} \cdot ||y||^p_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \\
&\leq 2c \cdot ||h_\Gamma^*(x) + \ell_i^{-1}(t)||^p_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \\
&\leq 2c \cdot (||h_\Gamma^*||_{\Lambda^p(\mathcal{X})} + ||\ell_i^{-1}||_{\Lambda^p(\mathcal{T})})^p \\
&\leq 2c \cdot (c + C_{M^-,M_\Gamma^+,p})^p.
\end{aligned}$$

The first line follows from Lemma 18. The second line follows from Assumption 5. The third line follows from the definition of $y(x,t)$. The fourth line follows from the assumption that $h_\Gamma^* \in \Lambda_c^p(\mathcal{X})$ and the result that $\ell_i^{-1} \in \Lambda^p_{C_{M^+,M^-,p}}(\mathcal{T})$. Thus, for any p, there exists a constant b that depends on p, c, M^-, M^+, K_2 such that $v \in \Lambda_b^p(\mathcal{X}\times\mathcal{T})$.

EC.5.5.4. Application of Implicit Function Theorem Finally, we can apply Theorem 5. In the case where $p \leq 1$, the we have already demonstrated that the necessary conditions are satisfied, so we can conclude that $\alpha_\Gamma^* \in \Lambda_{c'}^p(\mathcal{X})$ for c' that depends on p, b, K_1, K_2. Since b depends on c, p, M^-, M_Γ^+, we have that c' depends on $p, c, M^-, M_\Gamma^+, P_{\min,\Gamma}, P_{\max}$ and loss function L overall.

In the case where $p > 1$, we additionally must check that $|\frac{\partial}{\partial x_j} v(x,t)| < L$ for some $L > 0$. Since we have already shown that $||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \leq b$ for $p > 1$, we certainly have that this condition holds for $L = b$. Thus, we can apply Theorem 5 and observe that $\alpha_\Gamma^* \in \Lambda_{c'}^p(\mathcal{X})$ for a constant c' that depends on $p, c, d, M^-, M_\Gamma^+, P_{\min,\Gamma}, P_{\max}$ and loss function L overall.

EC.5.6. Details of Sieve Estimators and Proof of Theorem 4

EC.5.6.1. Truncated Sieve Spaces We denote the empirical risk minimizer $\hat{\theta}_n := (\hat{h}_n, \hat{\alpha}_n)$. We have that the minimizer of the population RU Risk $\theta^* := (h_\Gamma^*, \alpha_\Gamma^*)$ is unique (Theorem 2) and lies in a Hölder ball (Lemma 3). Overloading notation, we define the radius of this ball to be a constant c. In particular, by Theorem 2, θ^* lies in a *truncated* Hölder ball (EC.70) given by $\Theta := \Lambda_c^p(\mathcal{X}, -2B, 2B) \times \Lambda_c^p(\mathcal{X}, 0, M_\Gamma^+)$. In addition, the population RU risk is strictly convex over this set by Lemma 6.

Leveraging the observation that minimizer lies in a truncated Hölder space, we consider polynomial $\mathrm{Pol}(J_n)$ or univariate spline sieves $\mathrm{Spl}(r, J_n)$ and following Jin et al. (2022), we apply *truncation* so that the output range of functions in the sieve space is restricted to a bounded range $[a, b]$. We denote such sieves as $\mathrm{Pol}(J_n, a, b)$ and $\mathrm{Spl}(r, J_n, a, b)$, respectively.

A *truncated polynomial sieve* for estimating $(\hat{h}_n, \hat{\alpha}_n)$ is denoted $\Theta_{J_n} = \mathcal{H}_{J_n} \times \mathcal{A}_{J_n}$, where

$$\mathcal{H}_{J_n} = \{x \mapsto \min(\max(\prod_{k=1}^{d} f_k(x_k), -2B), 2B) : f_k \in \text{Pol}(J_n), k = 1, \dots d\},$$

and

$$\mathcal{A}_{J_n} = \{x \mapsto \min(\max(\prod_{k=1}^{d} f_k(x_k), 0), M_\Gamma^+) : f_k \in \text{Pol}(J_n)\}.$$

A *truncated univariate spline sieve* for estimating for estimating $(\hat{h}_n, \hat{\alpha}_n)$ is $\Theta_{J_n} = \mathcal{H}_{J_n} \times \mathcal{A}_{J_n}$, where

$$\mathcal{H}_{J_n} = \{x \mapsto \min(\max(\prod_{k=1}^{d} f_k(x_k), -2B), 2B) : f_k \in \text{Spl}(r, J_n), k = 1, \dots d\},$$

and

$$\mathcal{A}_{J_n} = \{x \mapsto \min(\max(\prod_{k=1}^{d} f_k(x_k), 0), M_\Gamma^+) : f_k \in \text{Spl}(r, J_n)\}.$$

We consider a sequence of such sieves $\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_J \subseteq \cdots \subseteq \Theta$.

EC.5.6.2. Consistency of Sieve Estimators We can show the consistency result via an application of the following theorem, given in Remark 3.3 of Chen (2007). For completeness, we state the full theorem below:

PROPOSITION 1. *Let Z_i be distributed i.i.d. following a distribution P. Let $\theta^* \in \Theta$ be the population risk minimizer*

$$\theta^* = \underset{\theta \in \Theta}{\text{argmin}}\, \mathbb{E}_P\left[l(\theta, Z_i)\right].$$

Let $\hat{\theta}_n$ be the empirical risk minimizer given by

$$\frac{1}{n}\sum_{i=1}^{n} l(\hat{\theta}_n, Z_i) \leq \inf_{\theta \in \Theta_{J_n}} \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) + o_P(1).$$

Assume that

1. $\mathbb{E}_P\left[l(\theta^*, Z_i)\right] < \infty$ *and if* $\mathbb{E}_P\left[l(\theta^*, Z_i)\right] = -\infty$, *then* $\mathbb{E}_P\left[l(\theta, Z_i)\right] > -\infty$ *for all* $\theta \in \Theta_J \setminus \{\theta^*\}$ *for* $J \geq 1$.
2. *There are a nonincreasing positive function δ and a positive function g such that for all $\epsilon > 0$ and for all $J \geq 1$,*

$$\inf_{\theta \in \Theta_J : ||\theta - \theta^*||_{L^2(P_X, \mathcal{X})} > \epsilon} \mathbb{E}_P\left[l(\theta, Z_i)\right] - \mathbb{E}_P\left[l(\theta^*, Z_i)\right] \geq \delta(J) \cdot g(\epsilon) > 0$$

   and $\liminf_J \delta(J) > 0$.
3. *Sieves spaces satisfy*

$$\Theta_J \subseteq \Theta_{J+1} \subseteq \Theta$$

   for all J, and there exists a sequence $\pi_J(\theta^*)$ such that $||\theta^* - \pi_J(\theta^*)||_{L^2(P_X, \mathcal{X})} \to 0$ as $J \to 0$ where $\pi_J : \Theta \to \Theta_J$ is the projection of $\theta \in \Theta$ onto the J-th sieve space Θ_J. Sieves spaces are compact under $||\cdot||_{L^2(P_X, \mathcal{X})}$.

4. $\mathbb{E}_P[l(\theta, Z_i)]$ *is continuous at* θ^*.
5. $\mathbb{E}_P\left[\sup_{\theta\in\Theta_J}|l(\theta, Z_i)|\right]$ *is bounded.*
6. *There are a finite* $s>0$ *and a random variable* $U(Z_i)$ *with* $\mathbb{E}_P[U(Z_i)]<\infty$ *such that*

$$\sup_{\theta,\theta'\in\Theta_J:||\theta-\theta'||_{L^2(P_X,\mathcal{X})}<\delta} |l(\theta, Z_i) - l(\theta, Z_i)| < \delta^s U(Z_i).$$

7. $\log N(\delta^{1/s}, \Theta_{J_n}, ||\cdot||_{L^2(P_X,\mathcal{X})}) = o(n)$ *for all* $\delta>0$.

Then $||\hat{\theta}_n - \theta^*||_{L^2(P_X,\mathcal{X})} = o_p(1)$.

It remains to verify that the conditions of this theorem hold in our setting. First, we note that our observed data (X_i, Y_i) is i.i.d. We take the loss function l to be the RU Loss. We verify the conditions of the theorem.

Verifying Condition 1. Under our assumptions, the minimizer of the population RU Risk $\theta^* := (h^*_\Gamma, \alpha^*_\Gamma)$ is unique (Theorem 2) and lies in a Hölder ball (Lemma 3). So, the first condition holds.

Verifying Condition 2. It holds because the minimizer of the population RU risk is strictly convex on the sieve spaces. To see this, we recall that by Lemma 6, the population RU risk is strictly convex on Θ and the sieve spaces its subsets. Thus, the minimizer of the population RU risk on this set must be well-separated.

Verifying Condition 3. The sieve spaces by definition satisfy $\Theta_J \subseteq \Theta_{J+1} \subseteq \Theta$ and are bounded polynomial or univariate splines sieves, so they are compact for all $J \leq 1$ in $L^2(P_X, \mathcal{X})$. We also have that the projection property holds for our truncated sieve spaces, as well. Let $\tilde{\pi}_J : \Theta \to \tilde{\Theta}_J$ be the projection of a parameter $\theta \in \Theta$ to $\tilde{\Theta}_J$, which is the sieve space without truncation. By Section 5.3.1 of Timan (1963) we have that

$$||\tilde{\pi}_{J_n}(\theta^*) - \theta^*||_\infty = O(J_n^{-p}).$$

The truncation is a contraction map to the true minimizer, so

$$\begin{aligned}
||\pi_{J_n}(\theta^*) - \theta^*||_{L^2(P_X,\mathcal{X})} &\leq ||\pi_{J_n}(\theta^*) - \theta^*||_\infty \\
&\leq ||\tilde{\pi}_{J_n}(\theta^*) - \theta^*||_\infty \\
&= O(J_n^{-p}) \\
&= O\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big) \to 0
\end{aligned}$$

which implies the third property.

Verifying Condition 4. It holds because population RU risk is continuous.

Verifying Condition 5. It holds because the sieve spaces Θ_J are bounded and $\mathcal{Y}$ is bounded and the absolute value of L_{RU} is continuous.

Verifying Condition 6. We consider $\theta \in \mathcal{B}_\delta$, where

$$\mathcal{B}_\delta = \{\theta \in \Theta_n \mid ||\theta - \theta^*||^2_{L^2(P_X,\mathcal{X})} \leq \delta\}.$$

We apply Lemma 9.

$$|L^\Gamma_{\text{RU}}(\theta(x), y) - L^\Gamma_{\text{RU}}(\theta^*(x), y)| \lesssim |\bar{L}(x,y) \cdot (h(x) - h^*(x))| + |\alpha(x) - \alpha^*(x)| \tag{EC.53}$$

$$\lesssim |\bar{L}(x,y)| \cdot ||\theta - \theta^*||_\infty \tag{EC.54}$$

$$\lesssim |\bar{L}(x,y)| \cdot ||\theta - \theta^*||^{\frac{2p}{2p+d}}_{L^2(\lambda)} \tag{EC.55}$$

$$\lesssim |\bar{L}(x,y)| \cdot ||\theta - \theta^*||^{\frac{2p}{2p+d}}_{L^2(P_X,\mathcal{X})}. \tag{EC.56}$$

Above, since $(h^*_\Gamma, \alpha^*_\Gamma)$ lie in Hölder balls with radius c, we can apply Lemma 20 to see that for $\theta \in \Theta$, $||\theta||_\infty \lesssim ||\theta||^{\frac{2p}{2p+d}}_{L^2(\lambda)}$, where λ is the Lebesgue measure. This gives (EC.55). Under Assumption 6, $||\theta - \theta'||_{L^2(P_X,\mathcal{X})} \asymp ||\theta - \theta'||_{L^2(\lambda)}$, which gives (EC.56). Thus, we find that this condition holds with $s = \frac{2p}{2p+d}$ and $U(X_i, Y_i) = |\bar{L}(X_i, Y_i)|$.

Verifying Condition 7. Recall that $\tilde{\Theta}_{J_n}$ is the standard sieve space without truncation. We note that the covering number of the truncated sieve space Θ_{J_n} is upper bounded by the covering number of $\tilde{\Theta}_{J_n}$, so we have that

$$\log N(w^{1+\frac{d}{2p}}, \Theta_n, ||\cdot||_{L^2(P_X,\mathcal{X})}) \leq \log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_{J_n}, ||\cdot||_{L^2(P_X,\mathcal{X})}).$$

In addition, for the finite-dimensional linear sieves, such as non-truncated polynomials $\text{Pol}(J_n)$ and univariate splines $\text{Spl}(r, J_n)$, we have that $\dim(\tilde{\Theta}_{J_n}) = O(J_n^d)$

$$\log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_{J_n}, ||\cdot||_{L^2(P_X,\mathcal{X})}) \lesssim \dim(\tilde{\Theta}_{J_n}) \log(\frac{1}{w}) \asymp J_n^d \log(\frac{1}{w}).$$

from Van de Geer and van de Geer (2000).

EC.5.6.3. Proof of Theorem 4 The main goal of this proof is to show that Theorem 3.2 of Chen (2007) applies to our setting. We write the full theorem statement below.

PROPOSITION 2. *Let Z_i be distributed i.i.d. following a distribution P. Let $\theta^* \in \Theta$ be the population risk minimizer*

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_P\left[l(\theta, Z_i)\right].$$

Let $\hat{\theta}_n$ be the empirical risk minimizer given by

$$\frac{1}{n}\sum_{i=1}^{n} l(\hat{\theta}_n, Z_i) \le \inf_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) + O_P(\epsilon_n^2).$$

Let $\mathcal{F}_n = \{l(\theta, Z_i) - l(\theta^*, Z_i) : ||\theta - \theta^*|| \le \delta, \theta \in \Theta_n\}$. For some constant $b > 0$, let

$$\delta_n = \inf\left\{\delta \in (0,1) : \frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta} \sqrt{H_{[]}(w^{1+\frac{d}{2p}}, \mathcal{F}_n, ||\cdot||)}dw \le 1\right\},$$

where $H_{[]}(w, \mathcal{F}_n, ||\cdot||_r)$ is the $L^r(P)$ metric entropy with bracketing of the class $\mathcal{F}_n$.

Assume that the following conditions hold.

1. $||\hat{\theta}_n - \theta^*|| = o_P(1)$.
2. *In a neighborhood of θ^*,* $\mathbb{E}\left[l(\theta, Z_i) - l(\theta^*, Z_i)\right] \asymp ||\theta - \theta^*||^2$.
3. *There is $C_1 > 0$ s.t. for all small $\epsilon > 0$*

$$\sup_{\theta\in\Theta_n : ||\theta-\theta^*||\le\epsilon} \mathrm{Var}\left[l(\theta, Z_i) - l(\theta^*, Z_i)\right] \le C_1\epsilon^2.$$

4. *For any $\delta > 0$, there exists a constant $s \in (0,2)$ such that*

$$\sup_{\theta\in\Theta_n : ||\theta-\theta^*||\le\delta} |l(\theta, Z_i) - l(\theta^*, Z_i)| \le \delta^s U(Z_i)$$

with $\mathbb{E}\left[U(Z_i)^\gamma\right] \le C_2$ for some $\gamma \ge 2$.

Then $||\hat{\theta}_n - \theta^*|| = O_P(\epsilon_n)$, where

$$\epsilon_n = \max\left\{\delta_n, \inf_{\theta\in\Theta_n} ||\theta^* - \theta||\right\}.$$

It remains to verify that the conditions of Proposition 2 hold in our setting. For the metric, we will use $||\cdot||_{L^2(P_X,\mathcal{X})}$. Note that our observed data (X_i, Y_i) is i.i.d.. Since any function $\theta \in \Theta$ only depends on X, $||\cdot||_{L^2(P_X,\mathcal{X})} = ||\cdot||_{L^2(P,\mathcal{X}\times\mathcal{Y})}$.

Verifying Condition 1. This is proved in Section EC.5.6.2.

Verifying Condition 2. We note that by Theorem 3, the population RU risk is strongly convex

and smooth in a $||\cdot||_\infty$-ball about the minimizer θ^*. We note that all θ in this $||\cdot||_\infty$-ball about θ^* also must lie in a $||\cdot||_{L^2(P_X,\mathcal{X})}$-ball about θ^*. So, in a $L^2(P_X,\mathcal{X})$-neighborhood of θ^*, we have that

$$\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X),Y)\right]-\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(\theta^*(X),Y)\right]\asymp||\theta-\theta^*||^2_{L^2(P_X,\mathcal{X})}.$$

Verifying Condition 3. First, we show the following three intermediate results.

$$\mathbb{E}_P\left[(L(h(X),Y)-L(h^*_\Gamma(X),Y))^2\right]\lesssim||h-h^*_\Gamma||^2_{L^2(P_X,\mathcal{X})}. \tag{EC.57}$$

$$\mathbb{E}_P\left[(\alpha(X)-\alpha^*_\Gamma(X))^2\right]\asymp||\alpha-\alpha^*_\Gamma||^2_{L^2(P_X,\mathcal{X})}. \tag{EC.58}$$

$$\mathbb{E}_P\left[((L(h(X),Y)-\alpha(X))_+-(L(h^*_\Gamma(X),Y)-\alpha^*_\Gamma(X))_+)^2\right]\lesssim||\theta-\theta^*||^2_{L^2(P_X,\mathcal{X})}. \tag{EC.59}$$

Eq. (EC.57) can be shown by apply Lemma 9.

$$\begin{aligned}
\mathbb{E}_P\left[(L(h(X),Y)-L(h^*_\Gamma(X),Y))^2\right]&=\mathbb{E}_P\left[\bar{L}(X,Y)^2\cdot(h(X)-h^*_\Gamma(X))^2\right]\\
&=\mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[\bar{L}(X,Y)^2\cdot(h(X)-h^*_\Gamma(X))^2\mid X=x\right]\right]\\
&\le\sup_{x\in\mathcal{X}}\mathbb{E}_{P_{Y|X}}\left[\bar{L}(x,Y)^2\mid X=x\right]\cdot||h-h^*_\Gamma||^2_{L^2(P_X,\mathcal{X})}\\
&\asymp||h-h^*_\Gamma||^2_{L^2(P_X,\mathcal{X})}.
\end{aligned}$$

Eq. (EC.58) is true by definition. So, we proceed to show (EC.59). We use (EC.57), (EC.58).

$$\begin{aligned}
&\mathbb{E}_P\left[((L(h(X),Y)-\alpha(X))_+-(L(h^*_\Gamma(X),Y)-\alpha^*_\Gamma(X))_+)^2\right]\\
&\le\mathbb{E}_P\left[((L(h(X),Y)-\alpha(X))-(L(h^*_\Gamma(X),Y)-\alpha^*_\Gamma(X)))^2\right]\\
&=\mathbb{E}_P\left[((L(h(X),Y)-L(h^*_\Gamma(X),Y))-(\alpha(X)-\alpha^*_\Gamma(X)))^2\right]\\
&\le2\mathbb{E}_P\left[(L(h(X),Y)-L(h^*_\Gamma(X),Y))^2\right]+2\mathbb{E}_P\left[(\alpha(X)-\alpha^*_\Gamma(X))^2\right]\\
&\lesssim||h-h^*_\Gamma||^2_{L^2(P_X,\mathcal{X})}+||\alpha-\alpha^*_\Gamma||^2_{L^2(P_X,\mathcal{X})}\\
&=||\theta-\theta^*||^2_{L^2(P_X,\mathcal{X})}.
\end{aligned}$$

Now, we consider $\theta\in\mathcal{B}_\epsilon$ where

$$\mathcal{B}_\epsilon=\{\theta\in\Theta_n\mid||\theta-\theta^*||_{L^2(P_X,\mathcal{X})}\le\epsilon\}.$$

We aim to show that $\mathrm{Var}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X),Y)-L^\Gamma_{\mathrm{RU}}(\theta^*(X),Y)\right]\lesssim\epsilon^2$ when $||\theta-\theta^*||_{L^2(P_X,\mathcal{X})}\le\epsilon$.

$$\begin{aligned}
&\mathrm{Var}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X),Y)-L^\Gamma_{\mathrm{RU}}(\theta^*(X),Y)\right]\\
&\le\mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(\theta(X),Y)-L^\Gamma_{\mathrm{RU}}(\theta^*(X),Y))^2\right]\\
&\le3\mathbb{E}_P\left[(L(h(X),Y)-L(h^*_\Gamma(X),Y))^2\right]+3\mathbb{E}_P\left[(\alpha(X)-\alpha^*_\Gamma(X))^2\right]\\
&\quad+3\mathbb{E}_P\left[((L(h(X),Y)-\alpha(X))_+-(L(h^*_\Gamma(X),Y)-\alpha^*_\Gamma(X))_+)^2\right]\\
&\lesssim\|h-h^*_\Gamma\|^2_{L^2(P_X,\mathcal{X})}+\|\alpha-\alpha^*_\Gamma\|^2_{L^2(P_X,\mathcal{X})}+\|\theta-\theta^*\|^2_{L^2(P_X,\mathcal{X})}\\
&\lesssim\|\theta-\theta^*\|^2_{L^2(P_X,\mathcal{X})}.
\end{aligned}$$

The second line comes from the Cauchy-Schwarz inequality and the second last line comes from (EC.57), (EC.58), and (EC.59). This prove the third condition.

Verifying Condition 4. This is the same as Condition 6 in Proposition 1.

Deriving the rate for $||\hat{\theta}_n - \theta^*||_{L^2(P_X,\mathcal{X})}$. By Proposition 2, we have that

$$||\hat{\theta}_n - \theta^*||_{L^2(P_X,\mathcal{X})} = O_P\left(\max\left\{\delta_n, \inf_{\theta\in\Theta_n} ||\theta - \theta^*||_{L^2(P_X,\mathcal{X})}\right\}\right).$$

We will show

$$\delta_n \preceq \sqrt{\frac{J_n^d \log n}{n}}, \quad \inf_{\theta\in\Theta_n} ||\theta^* - \theta||_{L^2(P_X,\mathcal{X})} \preceq J_n^{-p}$$

By setting $J_n = (\frac{n}{\log n})^{\frac{1}{2p+d}}$. Thus, we have that $||\hat{\theta} - \theta^*||_{L^2(P_X,\mathcal{X})} = O_P\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big)$.

Bounding δ_n. Let $\mathcal{F}_n = \{L^{\Gamma}_{\mathrm{RU}}(\theta(X_i), Y_i) - L^{\Gamma}_{\mathrm{RU}}(\theta^*(X_i), Y_i) : ||\theta - \theta^*||_{L^2(P_X,\mathcal{X})} \leq \delta, \theta \in \Theta_n\}$. Let $H_{[]}(w, \mathcal{F}_n, ||\cdot||_{L^2(P_X,\mathcal{X})})$ be the $L^2(P_X,\mathcal{X})$-metric entropy with bracketing of the class $\mathcal{F}_n$. Since in our setting, we satisfy the fourth condition of Theorem 3.2 from Chen (2007) with $s = \frac{2p}{2p+d}$,

$$H_{[]}(w, \mathcal{F}_n, ||\cdot||_2) \leq \log N(w^{1+\frac{d}{2p}}, \Theta_n, ||\cdot||_{L^2(P_X,\mathcal{X})}).$$

Recall that $\tilde{\Theta}_n$ is the sieve space *without truncation*. We note that the covering number of Θ_n is upper bounded by the covering number of $\tilde{\Theta}_n$, so we have that

$$H_{[]}(w, \mathcal{F}_n, ||\cdot||_2) \leq \log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_n, ||\cdot||_{L^2(P_X,\mathcal{X})}).$$

For the finite-dimensional linear sieves, such as univariate splines and polynomials, we have that

$$\log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_n, ||\cdot||_{L^2(P_X,\mathcal{X})}) \lesssim \dim(\tilde{\Theta}_n) \log(\frac{1}{w})$$

from Van de Geer and van de Geer (2000). Then, we have that

$$\frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta} \sqrt{\log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_n, ||\cdot||_{L^2(P_X,\mathcal{X})})}dw \lesssim \frac{1}{\delta}\sqrt{\frac{\dim(\tilde{\Theta}_n)}{n}\log\frac{1}{\delta}}.$$

We realize that

$$\delta_n \preceq \sqrt{\frac{\dim(\tilde{\Theta}_n)\log n}{n}}.$$

We note that $\tilde{\Theta}_n = \tilde{\mathcal{H}}_n \times \tilde{\mathcal{A}}_n$. We have that $\dim(\tilde{\Theta}_n) = 2J_n^d = O(J_n^d)$. Plugging this in, we have that

$$\delta_n \preceq \sqrt{\frac{J_n^d \log n}{n}}.$$

Bounding the approximation error $\inf_{\theta\in\Theta_n}||\theta^*-\theta||_{L^2(P_X,\mathcal{X})}$. Since the truncation of the sieve space is a contraction map to the true minimizer, we have that

$$\inf_{\theta\in\Theta_n}||\theta^*-\theta||_{L^2(P_X,\mathcal{X})}\leq\inf_{\theta\in\tilde{\Theta}_n}||\theta^*-\theta||_\infty\preceq J_n^{-p},$$

where the last inequality follows from Section 5.3.1 of Timan (1963).

Deriving the rate for $||\hat{\theta}_n-\theta^*||_{L^2(Q_X,\mathcal{X})}$**.** We can show that for $Q_X\ll P_X$, if $\sup_{x\in\mathcal{X}}\frac{dQ_X(x)}{dP_X(x)}<C$ for some $C<\infty$, then the same rate of convergence holds. We have that

$$\begin{aligned}
||\hat{\theta}_n-\theta^*||_{L^2(Q_X,\mathcal{X})}&=\left(\mathbb{E}_{Q_X}\left[(\hat{\theta}_n(x)-\theta^*(x))^2\right]\right)^{\frac{1}{2}}\\
&=\left(\mathbb{E}_{P_X}\left[(\hat{\theta}_n(x)-\theta^*(x))^2\cdot\frac{dP_X(x)}{dQ_X(x)}\right]\right)^{\frac{1}{2}}\\
&\leq\left(\mathbb{E}_{P_X}\left[(\hat{\theta}_n(x)-\theta^*(x))^2\right]\right)^{\frac{1}{2}}\cdot\left(\sup_{x\in\mathcal{X}}\left|\frac{dP_X(x)}{dQ_X(x)}\right|\right)^{\frac{1}{2}}\\
&=||\hat{\theta}_n-\theta^*||_{L^2(P_X,\mathcal{X})}\cdot\sqrt{C}\\
&=O_P\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big).
\end{aligned}$$

EC.5.7. Proof of Corollary 2

We find that

$$\begin{aligned}
\sup_{Q\in\mathcal{S}_\Gamma(P,Q_X)}\mathbb{E}_Q\left[L(h(X),Y)\right]&=\mathbb{E}_{Q_X}\left[\inf_{\alpha(x)\in\mathbb{R}}\mathbb{E}_{P_{Y|X}}\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(x),Y)\mid X=x\right]\right]\\
&=\inf_{\alpha\in L^2(P_X,\mathcal{X})}\mathbb{E}_{Q_X}\left[\mathbb{E}_{P_{Y|X}}\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\mid X=x\right]\right]\\
&=\inf_{\alpha\in L^2(P_X,\mathcal{X})}\mathbb{E}_P\left[r(X)\cdot L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right].
\end{aligned}$$

The first line follows from the proof of Theorem 1. The second line follows from the fact that a function $\alpha\in L^2(P_X,\mathcal{X})$ minimizes $\mathbb{E}_{Q_X}\left[\mathbb{E}_{P_{Y|X}}\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\mid X\right]\right]$ if and only if it minimizes $\mathbb{E}_{P_{Y|X}}\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(x),Y)\mid X=x\right]$ for every x, which can be seen through an identical argument as the proof of Lemma 2. The last line follows from the definition of r.

EC.6. Proof of Technical Lemmas

EC.6.1. Proof of Lemma 4

We have that $T^c_{1,x}(c)=-\mathbb{E}_{P_{Y|X=x}}\left[\ell'(Y-c)\right]$. In addition, $T^{cc}_{1,x}(c)=\mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-c)\right]$. So, $T_{1,x}$ is twice differentiable in c. In addition, we realize that

$$\begin{aligned}
\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right]&=\mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X=x}}\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right]\right]\\
&=\Gamma^{-1}\mathbb{E}_{P_X}\left[T_{X,1}(h(X))\right].
\end{aligned}$$

EC.6.2. Proof of Lemma 5

EC.6.2.1. Main Proof First, we compute the first derivatives of $T_{3,x}(c,d)$. Second, we compute the second derivatives of $T_{3,x}(c,d)$ when $d>0$. Finally, we show that $T_{3,x}$ can be used to express $\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]$.

Computing derivatives for the $d\leq 0$ case is straightforward. When $d\leq 0$, we have that

$$T^c_{3,x}(c,d) = -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mid X=x\right]$$
$$T^d_{3,x}(c,d) = -1.$$

Now, we consider the $d>0$ case. To compute the derivatives of

$$T_{3,x}(c,d) = \mathbb{E}_{P_{Y|X}}\left[(\ell(Y-c)-d)\mathbb{I}(\ell(Y-c)>d)\mid X=x\right],$$

we first identify when the condition $\ell(Y-c)>d$ is satisfied. The strong convexity of ℓ given by Assumption 2 implies that ℓ is strictly increasing on $y>0$ and ℓ is strictly decreasing on $y<0$. We define ℓ_1^{-1} to be the inverse of $\ell(y)$ on $y>0$. We define ℓ_2^{-1} to be the inverse of $\ell(y)$ on $y<0$. By the Inverse Function Theorem, we have that

$$(\ell_i^{-1})'(z) = \frac{1}{\ell'(\ell_i^{-1}(z))} \quad i=1,2. \tag{EC.60}$$

We note that $\ell_1^{-1}(z)>0$, and $\ell(y)$ strictly increasing on $y>0$, so $\ell'(\ell_1^{-1}(z))>0$. By (EC.60), we have that $(\ell_1^{-1})'(z)>0$. This means that ℓ_1^{-1} is strictly increasing on its domain. By an analogous argument, we have that $(\ell_2^{-1})'(z)<0$ and ℓ_2^{-1} is strictly decreasing on its domain.

Based on the results above, we realize that for $d>0$,

$$\{y\in\mathbb{R}\mid \ell(y-c)>d\} = \{y\in\mathbb{R}\mid y-c>\ell_1^{-1}(d)\}\cup\{y\in\mathbb{R}\mid y-c<\ell_2^{-1}(d)\}.$$

Thus, we can rewrite $T_{3,x}(c,d)$ for $d>0$ as follows

$$T_{3,x}(c,d) = \mathbb{E}_{Y|X=x}\left[(\ell(Y-c)-d)\mathbb{I}(Y-c>\ell_1^{-1}(d))\right] + \mathbb{E}_{Y|X=x}\left[(\ell(Y-c)-d)\mathbb{I}(Y-c<\ell_2^{-1}(d))\right].$$

Now, we can compute the derivatives of $T_{3,x}(c,d)$ on $d>0$ using the lemma below. The proof of this lemma can be found at the end of this subsection.

Lemma 13. *Let $Z\sim F$, where F has continuous density. Then,*

$$\frac{d}{ds}\mathbb{E}_P\left[(Z-s)\mathbb{I}(Z-s\geq 0)\right] = -\mathbb{P}_P\left[Z\geq s\right].$$

$$\begin{aligned}T^d_{3,x}(c,d) &= \mathbb{E}_{P_{Y|X}}\left[-1\cdot\mathbb{I}(\ell(Y-c)>d)\mid X=x\right]\\ &= -\Pr(Y>c+\ell_1^{-1}(d)\mid X=x)-\Pr(Y<c+\ell_2^{-1}(d)\mid X=x)\\ &= -1+P_{Y|X=x}(c+\ell_1^{-1}(d))-P_{Y|X=x}(c+\ell_2^{-1}(d)).\end{aligned}$$

Another way to express $T^d_{3,x} = -\Pr(\ell(Y-c)>d|X=x)$.

Since $P_{Y|X}$ is continuous,

$$\lim_{d\to 0^+} T^d_{3,x}(c,d) = -1+P_{Y|X=x}(-c)-P_{Y|X=x}(-c) = -1 = \lim_{d\to 0^-} T^d_{3,x}(c,d),$$

so $T_{3,x}(c,d)$ is differentiable at $d=0$. Also,

$$\begin{aligned}T^c_{3,x}(c,d) &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(\ell(Y-c)>d)\mid X=x\right]\\ &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\cdot\mathbb{I}(Y-c>\ell_1^{-1}(d))\right]-\mathbb{E}_{Y|X=x}\left[\ell'(Y-c)\cdot\mathbb{I}(Y-c<\ell_2^{-1}(d))\mid X=x\right]\end{aligned}$$

We realize that

$$\lim_{d\to 0^+} T^c_{3,x}(c,d) = -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mid X=x\right] = \lim_{d\to 0^-} T^c_{3,x}(c,d),$$

so $T_{3,x}(c,d)$ is differentiable with respect to c.

Second, we compute the second derivatives of $T_{3,x}(c,d)$ when $d>0$. It is straightforward to see that

$$T^{dc}_{3,x}(c,d) = p_{Y|X=x}(c+\ell_1^{-1}(d))-p_{Y|X=x}(c+\ell_2^{-1}(d)).$$

In addition, we have that

$$\begin{aligned}T^{dd}_{3,x}(c,d) &= p_{Y|X=x}(c+\ell_1^{-1}(d))\cdot\frac{1}{\ell'(\ell_1^{-1}(d))}-p_{Y|X=x}(c+\ell_2^{-1}(d))\cdot\frac{1}{\ell'(\ell_2^{-1}(d))}\\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(c+\ell_i^{-1}(d))\cdot\frac{1}{|\ell'(\ell_i^{-1}(d))|}.\end{aligned}$$

The second line follows because $\ell'(\ell_2^{-1}(y))<0$. Finally, we compute $T^{cc}_{3,x}(c,d)$. First, we recall $T^c_{3,x}(c,d)$ from Lemma 5 and simplify it as follows.

$$\begin{aligned}T^c_{3,x}(c,d) &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(\ell(Y-c)>d)\mid X=x\right]\\ &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(Y>\ell_1^{-1}(d)+c)\mid X=x\right]-\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(Y<\ell_2^{-1}(d)+c)\mid X=x\right]\\ &= -\int_{\ell_1^{-1}(d)+c}^{\infty}\ell'(y-c)p_{Y|X=x}(y)dy-\int_{-\infty}^{\ell_2^{-1}(d)+c}\ell'(y-c)p_{Y|X=x}(y)dy\\ &= -\int_{\ell_1^{-1}(d)}^{\infty}\ell'(y)p_{Y|X=x}(y+c)dy-\int_{-\infty}^{\ell_2^{-1}(d)}\ell'(y)p_{Y|X=x}(y+c)dy.\end{aligned}$$

Now, we compute $T^{cc}_{3,x}(c,d)$ by differentiating with respect to c and applying integration by parts.

$$\begin{aligned}
T^{cc}_{3,x}(c,d) &= -\int_{\ell_1^{-1}(d)}^{\infty} \ell'(y) p'_{Y|X=x}(y+c)dy - \int_{-\infty}^{\ell_2^{-1}(d)} \ell'(y) p'_{Y|X=x}(y+c)dy \\
&= -\Big(\ell'(y)p_{Y|X=x}(y+c)\Big|_{\ell_1^{-1}(d)}^{\infty} - \int_{\ell_1^{-1}(d)}^{\infty} p_{Y|X=x}(y+c)\ell''(y)dy\Big) \\
&\quad - \Big(\ell'(y)p_{Y|X=x}(y+c)\Big|_{-\infty}^{\ell_2^{-1}(d)} - \int_{-\infty}^{\ell_2^{-1}(d)} p_{Y|X=x}(y+c)\ell''(y)dy\Big) \\
&= \ell'(\ell_1^{-1}(d))p_{Y|X=x}(\ell_1^{-1}(d)+c) + \int_{\ell_1^{-1}(d)}^{\infty} p_{Y|X=x}(y+c)\ell''(y)dy \\
&\quad - \ell'(\ell_2^{-1}(d))p_{Y|X=x}(\ell_2^{-1}(d)+c) + \int_{-\infty}^{\ell_2^{-1}(d)} p_{Y|X=x}(y+c)\ell''(y)dy \\
&= \ell'(\ell_1^{-1}(d))p_{Y|X=x}(\ell_1^{-1}(d)+c) + \int_{c+\ell_1^{-1}(d)}^{\infty} p_{Y|X=x}(y)\ell''(y-c)dy \\
&\quad - \ell'(\ell_2^{-1}(d))p_{Y|X=x}(\ell_2^{-1}(d)+c) + \int_{-\infty}^{c+\ell_2^{-1}(d)} p_{Y|X=x}(y)\ell''(y-c)dy \\
&= \sum_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(\ell_i^{-1}(d)+c) + \mathbb{E}_{P_{Y|X}}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c)>d) \mid X=x\right].
\end{aligned}$$

Thus, when $d>0$, $T_{3,x}(c,d)$ is twice differentiable in (c,d).

Lastly, we find that

$$\begin{aligned}
&\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right] \\
&= (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_P\left[(\ell(Y-h(X))-\alpha(X))_+\right] \\
&= (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[\ell(Y-h(X))-\alpha(X))_+ \mid X\right]\right] \\
&= (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[\begin{cases}\mathbb{E}_{P_{Y|X}}\left[(\ell(Y-h(X))-\alpha(X))\mathbb{I}(\ell(Y-h(X))>\alpha(X))\right] & \alpha(X)>0\\ \mathbb{E}_{P_{Y|X}}\left[(\ell(Y-h(X))-\alpha(X))\right] & \alpha(X)\le 0\end{cases}\right] \\
&= (\Gamma-\Gamma^{-1})\mathbb{E}_{P_X}\left[T_{3,X}(h(X),\alpha(X))\right].
\end{aligned}$$

EC.6.2.2. Proof of Lemma 13 The proof follows from the continuity of f and the Fundamental Theorem of Calculus.

$$\begin{aligned}
\frac{d}{ds}\mathbb{E}_F\left[(Z-s)\mathbb{I}(Z\ge s)\right] &= \frac{d}{ds}\int_s^{\infty}(z-s)f(z)dz \\
&= \frac{d}{ds}\int_s^{\infty} zf(z)dz - \frac{d}{ds}(s\int_s^{\infty} f(z)dz) \\
&= -sf(s) - \frac{d}{ds}(s\cdot\mathbb{P}_F\left[Z\ge s\right]) \\
&= -sf(s) - \mathbb{P}_F\left[Z\ge s\right] + sf(s) \\
&= -\mathbb{P}_F\left[Z\ge s\right].
\end{aligned}$$

EC.6.3. Proof of Lemma 6

We start with two lemmas with proofs presented at the end of this subsection.

LEMMA 14. *Let $H(h,\alpha)=G(h)+F(h,\alpha)$, where G is strongly convex and Gâteaux differentiable in h and F is jointly convex in (h,α), strictly convex in α, and Gâteaux differentiable in (h,α). Then H is strictly convex in (h,α).*

LEMMA 15. *Under Assumption 2, $\mathbb{E}_P\left[L^{\Gamma}_{RU,1}(h(X),Y)\right]$ is strongly convex in h and strong convexity is defined using the $L^2(P_X,\mathcal{X})$ norm.*

Let

$$\begin{aligned} F(h,\alpha) &= \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right] \\ G(h) &= \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},1}(h(X),Y)\right] \\ H(h,\alpha) &= \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},1}(h(X),Y)\right]+\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]. \end{aligned}$$

Note that

$$\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]=H(h,\alpha)+\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},2}(\alpha(X))\right]. \tag{EC.61}$$

Since the population RU risk is the sum of H and a function that is convex in (h,α), then it suffices to show that H is strictly convex. It remains to show that the conditions of Lemma 14 hold so that we can conclude that H, as defined above, is strictly convex in (h,α). Strong convexity of G follows from Lemma 15.

Verifying Gâteaux differentiability, strong convexity of G, and convexity of F. First, we note that F,G are both Gâteaux differentiable by Lemma 8.

By the first part of Theorem 1, F is jointly convex in (h,α).

Verifying the strict convexity of F in α on $\mathcal{A}$. Let $L_3(h,\alpha)=\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]$. By Lemma 8, we have that $\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]$ is Gâteaux differentiable in α with

$$L'_{3,\alpha}(h,\alpha;\phi)=(\Gamma-\Gamma^{-1})\cdot\mathbb{E}_X\left[T^d_{3,X}(h(X),\alpha(X))\cdot\phi(X)\right].$$

We need to show that for $\alpha_1,\alpha_2\in\mathcal{A}$ that differ on a set of positive measure, we have that

$$\mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X),\alpha_1(X))-T^d_{3,X}(h(X),\alpha_2(X))\cdot(\alpha_1(X)-\alpha_2(X))\right]>0.$$

From Lemma 5, we have that

$$T^d_{3,x}(h(x),\alpha(x))=\begin{cases} t^d_{3,x}(h(x),\alpha(x)) & \alpha(x)>0 \\ -1 & \alpha(x)\le 0 \end{cases},$$

where

$$t^d_{3,x}(h(x),\alpha(x)) = -1 + P_{Y|X=x}(h(x)+\ell_1^{-1}(\alpha(x))) - P_{Y|X=x}(h(x)+\ell_2^{-1}(\alpha(x))).$$

By the definition of ℓ_1^{-1}, ℓ_2^{-1} from Lemma 5, we have that

$$\ell_1^{-1}(\alpha(x)) > \ell_2^{-1}(\alpha(x)).$$

Under Assumption 3, we have that $P_{Y|X=x}$ is strictly increasing, so

$$P_{Y|X=x}(h(x)+\ell_1^{-1}(\alpha(x))) - P_{Y|X=x}(h(x)+\ell_2^{-1}(\alpha(x))) > 0,$$

which implies that

$$t^d_{3,x}(h(x),\alpha(x)) > -1. \tag{EC.62}$$

Under Assumption 2, ℓ_1^{-1} is strictly increasing and ℓ_2^{-1} is strictly decreasing. We realize that if $\alpha_1(x) > \alpha_2(x)$, then

$$\begin{aligned} P_{Y|X=x}(h(x)+\ell_1^{-1}(\alpha_1(x))) &> P_{Y|X=x}(h(x)+\ell_1^{-1}(\alpha_2(x))) \\ P_{Y|X=x}(h(x)+\ell_2^{-1}(\alpha_1(x))) &< P_{Y|X=x}(h(x)+\ell_2^{-1}(\alpha_2(x))), \end{aligned}$$

so

$$t^d_{3,x}(h(x),\alpha_1(x)) > t^d_{3,x}(h(x),\alpha_2(x)). \tag{EC.63}$$

Let $D = \{x \in \mathcal{X} \mid \alpha_1(x) \neq \alpha_2(x)\}$. Now, we compute

$$\mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X),\alpha_1(X)) - T^d_{3,X}(h(X),\alpha_2(X)) \cdot (\alpha_1(X)-\alpha_2(X))\right] \tag{EC.64a}$$

$$= \mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X),\alpha_1(X)) - T^d_{3,X}(h(X),\alpha_2(X)) \cdot (\alpha_1(X)-\alpha_2(X))\mathbb{I}(D)\right] \tag{EC.64b}$$

$$= \mathbb{E}_{P_X}\left[((t^d_{3,X}(h(X),\alpha_1(X)) - t^d_{3,X}(h(X),\alpha_2(X))) \cdot (\alpha_1(X)-\alpha_2(X)) \cdot \mathbb{I}(S_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D)\right] \tag{EC.64c}$$

$$+ \mathbb{E}_{P_X}\left[(t^d_{3,X}(h(X),\alpha_1(X)) + 1)(\alpha_1(X)-\alpha_2(X)) \cdot \mathbb{I}(S_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D)\right] \tag{EC.64d}$$

$$+ \mathbb{E}_{P_X}\left[(-1 - t^d_{3,X}(h(X),\alpha_2(X)))(\alpha_1(X)-\alpha_2(X)) \cdot \mathbb{I}(S^c_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D)\right]. \tag{EC.64e}$$

The first line holds because $(T^d_{3,x}(h(x),\alpha_1(x)) - T^d_{3,x}(h(x),\alpha_2(x)) \cdot (\alpha_1(x)-\alpha_2(x)) = 0$ on D^c. The decomposition into (EC.64c), (EC.64d), (EC.64e) holds because $T^d_{3,x}(h(x),\alpha_1(x)) - T^d_{3,x}(h(x),\alpha_2(x) = 0$ when $\alpha_1(x) \leq 0$ and $\alpha_2(x) \leq 0$.

Since we have that $\alpha_1, \alpha_2 \in \mathcal{A}$ and D has positive measure, we can show that $P(S^c_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D) < P(D)$. We consider two cases 1) $S_{\alpha_1,0} \cap D$ has positive measure and 2) $S_{\alpha_1,0} \cap D = \emptyset$. Suppose $S_{\alpha_1,0} \cap D$ has positive measure, then clearly

$$P(S^c_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D) \leq P(S^c_{\alpha_1} \cap D) < P(D).$$

If $S_{\alpha_1,0} \cap D$ empty, this means that $\alpha_1(x) \leq 0$ for all $x \in D$. At the same time, we have that for all $\alpha \in \mathcal{A}$, $\alpha(x) \geq 0$ for every $x \in \mathcal{X}$. So, we must have that $\alpha_1 = 0$ on D. We must have that $\alpha_2(x) > 0$ on D, because α_1, α_2 must differ on D and $\alpha_2(x) \geq 0$ for all $x \in \mathcal{X}$. So, this means that $S_{\alpha_2,0} \cap D$ has positive measure, so

$$P(S^c_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D) \leq P(S^c_{\alpha_2,0} \cap D) < P(D).$$

So, at least at least one of the sets $S_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D$, $S_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D$, $S^c_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D$ has positive measure.

Suppose $S_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D$ has positive measure. WLOG, if $\alpha_1(x) > \alpha_2(x)$, then $T^d_{3,x}(h(x), \alpha_1(x) - T^d_{3,x}(h(x), \alpha_2(x)) > 0$. In addition, if $\alpha_1(x) < \alpha_2(x)$, then $T^d_{3,x}(h(x), \alpha_1(x) - T^d_{3,x}(h(x), \alpha_2(x)) < 0$. Then (EC.64c) must be positive. We can use a similar argument to verify that (EC.64d) will be positive if $S_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D$ has positive measure and (EC.64e) will be positive if $S^c_{\alpha_1,0} \cap S_{\alpha_2,0}$ has positive measure. Thus, we conclude that

$$\mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X), \alpha_1(X)) - T^d_{3,X}(h(X), \alpha_2(X)) \cdot (\alpha_1(X) - \alpha_2(X))\right] > 0$$

so $\alpha \mapsto \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X), \alpha(X), Y)\right]$ is strictly convex on $\mathcal{A}$.

EC.6.3.1. Proof of Lemma 14 Since H is Gâteaux differentiable with derivative equal to $H'_{h,\alpha}$, we aim to show that, for any $h, \tilde{h}, \alpha, \tilde{\alpha}$,

$$\langle H'_{h,\alpha}(h, \alpha) - H'_{\tilde{h},\tilde{\alpha}}(0,0), (h - \tilde{h}, \alpha - \tilde{\alpha})\rangle > 0$$

to establish strict convexity. Without loss of generality, we assume that $(\tilde{h}, \tilde{\alpha}) = (0,0)$. Define the Gâteaux derivative of F and G in (h, α) to be $F'_{h,\alpha}, G'_{h,\alpha}$, respectively. Let the Gâteaux derivative of G with respect to h be G'_h. We have that

$$\begin{aligned}\langle H'_{h,\alpha}(h, \alpha) - H'_{h,\alpha}(0,0), (h, \alpha)\rangle &= \langle F'_{h,\alpha}(h, \alpha) + G'_{h,\alpha}(h, \alpha) - F'_{h,\alpha}(0,0) - G'_{h,\alpha}(0,0), (h, \alpha)\rangle \\ &= \langle F'_{h,\alpha}(h, \alpha) - F'_{h,\alpha}(0,0), (h, \alpha)\rangle + \langle G'_h(h, \alpha) - G'_h(0,0), h\rangle.\end{aligned}$$

Note that G does not depend on α, so the Gâteaux derivative is $G'_\alpha(h, \alpha) = G'_\alpha(0,0) = 0$. Since F is jointly convex in (h, α) and G is strongly convex in h and does not depend on α, both terms above are nonnegative.

If $h \neq 0$, then we have that

$$\begin{aligned}\langle H'_{h,\alpha}(h, \alpha) - H'_{h,\alpha}(0,0), (h, \alpha)\rangle &\geq \langle G'_h(h, \alpha) - G'_h(0,0), h) \\ &\geq \mu_1 ||h||^2 > 0,\end{aligned}$$

where the last line follows follows from G's strong convexity in h. If $h=0$ and $\alpha \neq 0$, then we have that

$$\begin{aligned}\langle H'_{h,\alpha}(h,\alpha) - H'_{h,\alpha}(0,0), (h,\alpha)\rangle &= \langle H'_{h,\alpha}(0,\alpha) - H'_{h,\alpha}(0,0), (0,\alpha)\rangle \\ &= \langle F'_{h,\alpha}(0,\alpha) - F'_{h,\alpha}(0,0), (0,\alpha)\rangle \\ &> 0,\end{aligned}$$

where the last inequality follows due to the strict convexity of F in α. Thus, H is strictly convex in (h,α).

EC.6.3.2. Proof of Lemma 15 Strong convexity follows from

$$\begin{aligned}L''_{1,h}(h,\alpha;\psi,\psi) &= \Gamma^{-1}\mathbb{E}_{P_X}\left[T^{cc}_{1,X}(h(X))\cdot(\psi(X))^2\right]. \\ &\geq \Gamma^{-1}\cdot C_{L,l}||\psi||^2_{L^2(\mathcal{X},P_X)}\end{aligned}$$

for $\psi \in L^2(P_X,\mathcal{X})$. The last line follows from Assumption 2, where we assume that ℓ is strongly convex.

EC.6.4. Proof of Lemma 7

Now, define a symmetric 2×2 matrix $A_x(c,d)$ where

$$\begin{aligned}A_{x,11}(c,d) &= T^{cc}_{3,x}(c,d) - \mathbb{E}_{Y|X=x}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c)>d)\right] + C_{L,l}\cdot\Pr(\ell(Y-c)>d\,|\,X=x) \\ A_{x,22}(c,d) &= T^{dd}_{3,x}(c,d) \\ A_{x,12}(c,d) &= T^{dc}_{3,x}(c,d),\end{aligned}$$

where F is the distribution over $\ell(Y-c)$ where Y follows $P_{Y|X=x}$. Under Assumption 2, we have that ℓ is $C_{\ell,l}$-strongly convex, so

$$\mathbb{E}_{Y|X=x}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c)>d)\right] - C_{L,l}\cdot\Pr(\ell(Y-c)>d|X=x)\geq 0.$$

Thus, we have that

$$\begin{aligned}\nabla^2 T_{3,x}(c,d) - A_x(c,d) &= \begin{bmatrix}\mathbb{E}_{Y|X=x}\left[(\ell''(Y-c)-C_{L,l})\mathbb{I}(\ell(Y-c)>d)\right] & 0\\ 0 & 0\end{bmatrix} \\ &= \begin{bmatrix}\mathbb{E}_{Y|X=x}\left[(\ell''(Y-c)\mathbb{I}(\ell(Y-c)>d)\right] - C_{L,l}\Pr(\ell(Y-c)>d\,|\,X=x) & 0\\ 0 & 0\end{bmatrix} \\ &\succeq 0.\end{aligned}$$

So,

$$\nabla^2 T_{3,x}(c,d) \succeq A_x(c,d). \tag{EC.65}$$

We can also define a symmetric 2×2 matrix $B_x(c,d)$ where

$$\begin{aligned}
B_{x,11}(c,d) &= T^{cc}_{3,x}(c,d) + \mathbb{E}_{Y|X=x}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c) \le d)\right] \\
B_{x,22}(c,d) &= T^{dd}_{3,x}(c,d) \\
B_{x,12}(c,d) &= T^{dc}_{3,x}(c,d).
\end{aligned}$$

Under Assumption 2, we have that ℓ is strongly convex, so

$$\mathbb{E}_{Y|X=x}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c) \le d)\right] \ge C_{\ell,l} \cdot \Pr(\ell(Y-c) \le d | X = x) \ge 0.$$

Thus, we have that

$$B_x(c,d) - \nabla^2 T_{3,x}(c,d) = \begin{bmatrix} \mathbb{E}_{Y|X=x}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c) \le d)\right] & 0 \\ 0 & 0 \end{bmatrix} \succeq 0.$$

So,

$$\nabla^2 T_{3,x}(c,d) \preceq B_x(c,d). \tag{EC.66}$$

Combining (EC.65) and (EC.66) yields the desired result.

EC.6.5. Proof of Lemma 8

Let $L(h,\alpha) = \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$. First, we verify Gâteaux differentiability with respect to α. We show that the directional derivative of $L(h,\alpha)$ with respect to α in the direction ϕ exists for all $\phi \in L^2(P_X,\mathcal{X})$. We note that the directional derivative with respect to α in the direction ϕ is given by

$$L'_{\alpha}(h,\alpha;\phi) = \lim_{\theta \to 0^+} \frac{L(h,\alpha+\theta\phi) - L(h,\alpha)}{\theta}.$$

We simplify the numerator as follows

$$\begin{aligned}
&L(h,\alpha+\theta\phi) - L(h,\alpha) \\
&= \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},2}((\alpha+\theta\phi)(X))\right] + \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),(\alpha+\theta\phi)(X),Y)\right] \\
&\quad - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},2}(\alpha(X))\right] - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right] \\
&= \theta(1-\Gamma^{-1}) \cdot \mathbb{E}_{P_X}\left[\phi(X)\right] + (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_{P_X}\left[T_{3,X}(h(X),(\alpha+\theta\phi)(X)) - T_{3,X}(h(X),\alpha(X))\right].
\end{aligned}$$

The first line follows because only the second and third term of the RU loss depend on α. The second line follows by Lemma 5. We analyze the second term on the right side of the above equation. We note that by Lemma 5, the map $T_{3,x}(c,d)$ is differentiable with respect to c,d. So,

$$\lim_{\theta \to 0^+} \frac{T_{3,x}(h(x),\alpha(x)+\theta\phi(x)) - T_{3,x}(h(x),\alpha(x))}{\theta} = T^{d}_{3,x}(h(x),\alpha(x))\phi(x).$$

Therefore, we have that

$$\begin{aligned}
L'_\alpha(h,\alpha;\phi) &= \lim_{\theta\to 0^+}(1-\Gamma^{-1})\cdot\frac{\theta\mathbb{E}_{P_X}\left[\phi(X)\right]}{\theta} \\
&\quad+\lim_{\theta\to 0^+}(\Gamma-\Gamma^{-1})\cdot\frac{\mathbb{E}_{P_X}\left[T_{3,X}(h(X),\alpha(x)+\theta\phi(X))-T_{3,X}(h(X),\alpha(X))\right]}{\theta} \\
&= (1-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[\phi(X)\right]+(\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[T^d_{3,X}(h(X),\alpha(X))\phi(X)\right] \\
&= \mathbb{E}_{P_X}\left[((1-\Gamma^{-1})+(\Gamma-\Gamma^{-1})\cdot T^d_{3,X}(h(X),\alpha(X))\cdot\phi(X)\right].
\end{aligned}$$

Since the directional derivative of $L(h,\alpha)$ with respect to α and in the direction ϕ exists for all $\phi\in\mathcal{A}$, then $L(h,\alpha)$ is Gâteaux differentiable in α.

We use a similar technique to verify Gâteaux differentiability with respect to h. We show that the directional derivative of $L(h,\alpha)$ with respect to h in the direction ψ exists for $\psi\in\mathcal{H}$. We recall that the directional derivative of $L(h,\alpha)$ with respect to h in the direction ψ is given by

$$L'_h(h,\alpha;\psi)=\lim_{\theta\to 0^+}\frac{L(h+\theta\psi,\alpha)-L(h,\alpha)}{\theta}. \tag{EC.67}$$

We simplify the directional derivative in (EC.67) as follows.

$$\begin{aligned}
L'_h(h,\alpha;\psi) &= \lim_{\theta\to 0^+}\frac{L(h+\theta\psi,\alpha)-L(h,\alpha)}{\theta} \\
&= \frac{\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}((h+\theta\psi)(X),Y)-L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right]}{\theta} \\
&\quad+\lim_{\theta\to 0^+}\frac{\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}((h+\theta\psi)(X),\alpha(X),Y)-L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]}{\theta} \\
&= \lim_{\theta\to 0^+}\Gamma^{-1}\cdot\frac{\mathbb{E}_{P_X}\left[T_{1,X}(h(X)+\theta\psi(X))-T_{1,X}(h(X))\right]}{\theta} \\
&\quad+\lim_{\theta\to 0^+}(\Gamma-\Gamma^{-1})\cdot\frac{\mathbb{E}_{P_X}\left[T_{3,X}(h(X)+\theta\psi(X),\alpha(X))-T_{3,X}(h(X),\alpha(X))\right]}{\theta} \\
&= \mathbb{E}_{P_X}\left[(\Gamma^{-1}\cdot T^c_{1,X}(h(X))+(\Gamma-\Gamma^{-1})\cdot T^c_{3,X}(h(X),\alpha(X))\cdot\psi(X)\right].
\end{aligned}$$

The first line follows because only the first and third terms of the RU loss depend on h. The second line follows because of Lemma 4 and Lemma 5. The third line follows from the differentiability of $T_{1,x},T_{3,x}$, which is given by Lemmas 4 and 5. Since the directional derivative of $L(h,\alpha)$ with respect to h and in the direction ψ exists for all $\psi\in L^2(P_X,\mathcal{X})$, and the directional derivative can be expressed as a continuous linear function (given the inner product on $L^2(P_X,\mathcal{X})$), then $L(h,\alpha)$ is Gâteaux differentiable in h.

We can compute second derivatives of $L(h,\alpha)$ on $\mathcal{C}$ by applying Lemma 4 and Lemma 5. Note that $T_{3,x}$ is twice-differentiable when $d>0$. For $(h,\alpha)\in\mathcal{C}$, we have that $\alpha(x)\geq 0$. We note that the restriction of $\mathcal{C}$ to the coordinate that corresponds to h is $L^2(P_X,\mathcal{X})$. Let $\mathcal{A}'$ be the resitrction of

$\mathcal{C}$ to the coordinate that corresponds to α. In the following result, we consider $\psi_1, \psi_2 \in L^2(P_X, \mathcal{X})$ and $\phi_1, \phi_2 \in \mathcal{A}'$. We find that

$$\begin{aligned}
L''_{hh}(h,\alpha;\psi_1,\psi_2) &= L''_{1,hh}(h,\alpha;\psi_1,\psi_2) + L''_{3,hh}(h,\alpha;\psi_1,\psi_2) \\
&= \Gamma^{-1}\mathbb{E}_{P_X}\left[T^{cc}_{1,X}(h(X))\psi_1(X)\psi_2(X)\right] + (\Gamma-\Gamma^{-1})\mathbb{E}_{P_X}\left[T^{cc}_{3,X}(h(X),\alpha(X))\psi_1(X)\psi_2(X)\right]. \\
L''_{h\alpha}(h,\alpha;\psi_1,\phi_1) &= L''_{3,h\alpha}(h,\alpha;\psi_1,\phi_1) \\
&= (\Gamma-\Gamma^{-1})\mathbb{E}_{P_X}\left[T^{cd}_{3,X}(h(X),\alpha(X))\psi_1(X)\phi_1(X)\right]. \\
L''_{\alpha\alpha}(h,\alpha;\phi_1,\phi_2) &= L''_{3,\alpha\alpha}(h,\alpha;\phi_1,\phi_2) \\
&= (\Gamma-\Gamma^{-1})\mathbb{E}_{P_X}\left[T^{dd}_{3,X}(h(X),\alpha(X))\phi_1(X)\phi_2(X)\right].
\end{aligned}$$

EC.6.6. Proof of Lemma 9

First, by the Mean Value Theorem, we have that for any $z \in \mathbb{R}$,

$$|\ell'(z)| = |\ell'(z) - \ell'(0)| \le |\ell''(\tilde{z})| \cdot |z|,$$

where $\tilde{z}$ is between z and 0. By Assumption 4, $|\ell''(\tilde{z})| \le C_{L,u}$, so

$$|\ell'(z)| \le C_{L,u} \cdot |z|. \tag{EC.68}$$

Again, by the Mean Value Theorem, we have that for any $h \in \Lambda^p_c(\mathcal{X})$ and $x \in \mathcal{X}$,

$$\begin{aligned}
|L(h(x),y) - L(h^*(x),y)| &= |\ell(y-h(x)) - \ell(y-h^*(x))| \\
&= |\ell'(y - (\lambda(x)\cdot h(x) + (1-\lambda(x))\cdot h^*(x)))| \cdot |h(x) - h^*(x)| \quad \lambda(x) \in [0,1].
\end{aligned}$$

We can define $\bar{L}(x,y) = |\ell'(y - (\lambda(x)\cdot h(x) + (1-\lambda(x))\cdot h^*(x)))|$. Now, we aim to verify that there exists some $0 < M < \infty$ such that

$$\sup_{x\in\mathcal{X}} \mathbb{E}_{P_{Y|X}}\left[\bar{L}(X,Y)^2 \mid X = x\right] < M.$$

We apply (EC.68).

$$\begin{aligned}
\mathbb{E}_{P_{Y|X}}\left[\bar{L}(x,Y)^2 \mid X = x\right] &= \mathbb{E}_{P_{Y|X}}\left[(\ell'(Y - (\lambda(x)\cdot h(x) + (1-\lambda(x)\cdot h^*(x))\cdot h^*(x))))^2 \mid X = x\right] \\
&= \mathbb{E}_{P_{Y|X}}\left[(\ell'(Y - (\lambda(x)\cdot h(x) + (1-\lambda(x))\cdot h^*(x)))\cdot h^*(x))^2 \mid X = x\right] \\
&= \mathbb{E}_{P_{Y|X}}\left[C^2_{L,u}\cdot((Y - (\lambda(x)\cdot h(x) + (1-\lambda(x))\cdot h^*(x)))\cdot h^*(x))^2 \mid X = x\right] \\
&\lesssim \mathbb{E}_{P_{Y|X}}\left[Y^2 \mid X = x\right] + h(x)^2 + h^*(x)^2 \\
&\lesssim \sup_{x\in\mathcal{X}} \mathbb{E}_{P_{Y|X}}\left[Y^2 \mid X = x\right] + c^2 \\
&< \infty.
\end{aligned}$$

The last two lines follow from Assumption 5 and 7. Assumption 5 gives that $h, h^* \in \Lambda^p_c(x)$, so $|h(x)| \le c$ and $|h^*(x)| \le c$. Assumption 7 gives that $\sup_{x\in\mathcal{X}} \mathbb{E}_{P_{Y|X}}[Y^2 \mid X = x]$ is finite. Thus, $\sup_{x\in\mathcal{X}} \mathbb{E}_{P_{Y|X}}\left[\bar{L}(X,Y)^2 \mid X = x\right] < \infty$.

EC.6.7. Proof of Lemma 10

For any 2×2 matrix A,

$$\lambda_{\max}(A)+\lambda_{\min}(A)=\operatorname{tr} A,\quad \lambda_{\max}(A)\cdot\lambda_{\min}(A)=\det A.$$

Since $\operatorname{tr} A\geq 0$ and $\det A\geq 0$, it must be that $\lambda_{\max}(A)\geq\lambda_{\min}(A)\geq 0$. Thus,

$$\lambda_{\max}(A)\leq \operatorname{tr} A,$$

and

$$\lambda_{\min}(A)=\frac{\det A}{\lambda_{\max}(A)}\geq\frac{\det A}{\operatorname{tr} A}.$$

EC.7. Results on Hölder Spaces

Recall the definition of p-Hölder smooth functions (Definition 3).

Overloading notation, we also say that g is uniformly Hölder-continuous with $\gamma\in(0,1]$ if the quantity

$$||g||_{\Lambda^\gamma(\mathcal{X})}=\sup_{\substack{x,x'\in\mathcal{X},\\ x\neq x'}}\frac{|g(x)-g(y)|}{|x-y|^\gamma}$$

is finite.

An important property of the Hölder norm is that

$$||h||_{\Lambda^p(\mathcal{X})}\leq\sum_{|\beta|_1\leq m-1}||D^\beta h||_{\Lambda^{p-1}(\mathcal{X})}. \tag{EC.69}$$

We can also define a truncated Hölder space:

$$\Lambda^p_c(\mathcal{X},a,b):=\{x\mapsto\min(\max(f(x),a,b)),f\in\Lambda^p_c(\mathcal{X})\}. \tag{EC.70}$$

We state a few key results on Hölder spaces and provide proofs of these results.

Lemma 16. *If $f,g\in\Lambda^p(\mathcal{X})$, then $||fg||_{\Lambda^p(\mathcal{X})}\leq C_p\cdot||f||_{\Lambda^p(\mathcal{X})}\cdot||g||_{\Lambda^p(\mathcal{X})}$, where C_p is a constant that depends only on p. In particular, $fg\in\Lambda^p(\mathcal{X})$.*

Lemma 17. *Define the following notation: For a vector $z\in\mathbb{R}^s$ and $k\in\mathbb{Z}^s_+$, let $[z]^k=\prod_{i=1}^s z_i^{k_i}$. Let $v(x,t)\in\Lambda^p(\mathcal{X}\times\mathcal{T})$ for some compact sets $\mathcal{X}\subset\mathbb{R}^d$ and $\mathcal{T}\subset\mathbb{R}$. Let $t:\mathcal{X}\to\mathbb{R}$ be a function in $\Lambda^p(\mathcal{X})$. Denote $u(x):=v(x,t(x))$ and $v^{(\beta,r)}(x):=\frac{\partial^{|\beta|_1+r}}{\partial x_1^{\beta_1}\ldots\partial x_d^{\beta_d}\partial t^r}$. Let $g(x)=(g_1(x),g_2(x),\ldots g_d(x),t(x))$, where $g_i(x)=x_i$. Let $q=\prod_{i=1}^d(\beta_i+1)-1$. Let $\beta\in\mathbb{Z}^d_+$. Let $\ell_1,\ell_2,\ldots\ell_q$ be a complete listing of all $\ell\in\mathbb{Z}^{d+1}_+$ such that $\ell\leq\beta$ and $|\ell|_1>0$.*

$$D^\beta u(x)=\sum_{\substack{\beta'\in\mathbb{Z}^d_+,r\in\mathbb{Z}_+:\\ 1\leq|\beta'|+r\leq|\beta|_1}}\sum_{P(\beta,\beta',r)}C(\beta,\beta',r,k_1,k_2,\ldots k_q)\cdot v^{(\beta',r)}(x_0)\prod_{j=1}^q[D^{\ell_j}g(x_0)]^{k_j}, \tag{EC.71}$$

where

$$P(\beta,\beta',r)=\left\{(k_1,k_2,\ldots k_q)\middle|k_i\in\mathbb{Z}^{d+1}_+,\sum_{i=1}^q k_i=(\beta',r),\sum_{i=1}^q|k_i|_1\cdot\ell_i=\beta\right\}.$$

LEMMA 18. *Let $v(x,t) \in \Lambda^p(\mathcal{X} \times \mathcal{T})$ for some compact sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{T} \subset \mathbb{R}$. Let $t : \mathcal{X} \to \mathbb{R}$ be a function in $\Lambda^p(\mathcal{X})$. Denote $u(x) := v(x, t(x))$. Then $u \in \Lambda^{\min\{p,p^2\}}(\mathcal{X})$. If furthermore $\frac{\partial}{\partial t} v(x,t)$ exists and satisfies $|\frac{\partial}{\partial t} v(x,t)| < L$, then $u \in \Lambda^p(\mathcal{X})$ and*

$$||u||_{\Lambda^p(\mathcal{X})} \leq (L \cdot ||t||_{\Lambda^p(\mathcal{X})} + ||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})}).$$

In addition, if $p > 1$, then

$$||u||_{\Lambda^p(\mathcal{X})} \leq C_p \cdot ||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \cdot ||t||^m_{\Lambda^p(\mathcal{X})}.$$

LEMMA 19. *If $t \in \Lambda^p(\mathcal{X})$ and*

$$0 < K_1 \leq t(x) \leq K_2 < \infty \quad \forall x \in \mathcal{X}$$

for some constants K_1, K_2, then $1/t \in \Lambda^p(\mathcal{X})$. In addition, if $p \leq 1$, there exists a constant C_{K_1} such that $||\frac{1}{t}||_{\Lambda^p(\mathcal{X})} \leq C_{K_1} ||t||_{\Lambda^p(\mathcal{X})}$, and if $p > 1$, there exists a constant C_{K_1,K_2} such that $||\frac{1}{t}||_{\Lambda^p(\mathcal{X})} \leq C_{K_1,K_2} \cdot ||t||^m_{\Lambda^p(\mathcal{X})}$.

THEOREM 5. *Let $v(x,t) \in \Lambda^p(\mathcal{X} \times \mathcal{T})$ for some compact sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{T} \subset \mathbb{R}$ and suppose v is differentiable with respect to its arguments. Assume that*

$$0 < K_1 \leq \frac{\partial}{\partial t} v(x,t) \leq K_2 < \infty \quad \forall x \in \mathcal{X}, t \in \mathcal{T}$$

for constants K_1, K_2. Furthermore, let $t(x)$ be the unique solution of $v(x,t) = 0$. If $p \leq 1$, then $t \in \Lambda^p(\mathcal{X})$ and

$$||t||_{\Lambda^p(\mathcal{X})} \leq C_{K_1,K_2,p} \cdot ||v||_{\Lambda^p(\mathcal{X})}.$$

If $p > 1$ and we additionally have

$$\Big|\Big|\frac{\partial}{\partial x_j} v(x,t)\Big|\Big|_\infty \leq L \quad \forall j \in [d]$$

for a constant $L > 0$, then $t \in \Lambda^p(\mathcal{X})$ and

$$||t||_{\Lambda^p(\mathcal{X})} \leq C_{K_1,K_2,L,p,d} \cdot ||v||^{2m!}_{\Lambda^p(\mathcal{X}\times\mathcal{T})}. \tag{EC.72}$$

COROLLARY 3. *Let $\mathcal{T} \subset \mathbb{R}$ be a compact set. If $f \in \Lambda^p_c(\mathcal{T})$ and f is differentiable on $\mathcal{T}$, is strictly increasing, and*

$$0 < K_1 \leq f'(t) \leq K_2 < \infty \quad \forall t \in \mathcal{T},$$

then $f^{-1} \in \Lambda^p_{c'}(\mathcal{X})$, where $\mathcal{X}$ is the image of f on $\mathcal{T}$ and c' is a constant that depends on p, d, K_1, K_2, and $||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})}$ for $v(x,t) := f(t) - x$.

LEMMA 20. *For $\theta \in \Lambda^p_c(\mathcal{X})$, we have that $||\theta||_\infty \leq C \cdot (2c)^{1-\frac{2p}{2p+d}} \cdot ||\theta||^{\frac{2p}{2p+d}}_{L^2(\mathcal{X})}$, where constant C that depends on p, d but not θ or c.*

EC.7.1. Proof of Lemma 16

We note that

$$||fg||_{\Lambda^p(\mathcal{X})} = \sum_{|\beta|_1 \le m} ||D^\beta fg||_\infty + \sum_{|\beta|_1 = m} ||D^\beta fg||_{\Lambda^\gamma(\mathcal{X})}.$$

By the product rule, we have that

$$D^\beta fg = \sum_{\lambda \le \beta} C(\beta, \lambda) \cdot \partial^{\beta-\lambda} f \cdot \partial^\lambda g$$

for absolute constants $C(\beta, \lambda) \ge 0$.

We note that

$$||\partial^{\beta-\lambda} f \cdot \partial^\lambda g||_\infty \le ||\partial^{\beta-\lambda} f||_\infty \cdot ||\partial^\lambda g||_\infty.$$

In addition, we have that

$$\begin{aligned}
||\partial^{\beta-\lambda} f \cdot \partial^\lambda g||_{\Lambda^\gamma(\mathcal{X})} &= \sup_{x,x' \in \mathcal{X}, x \neq x'} \frac{|\partial^{\beta-\lambda} f(x) \cdot \partial^\lambda g(x) - \partial^{\beta-\lambda} f(x') \cdot \partial^\lambda g(x')|}{|x - x'|^\gamma} \\
&= \sup_{x,x' \in \mathcal{X}, x \neq x'} \frac{|\partial^{\beta-\lambda} f(x) \cdot (\partial^\lambda g(x) - \partial^\lambda g(x')) + (\partial^{\beta-\lambda} f(x) - \partial^{\beta-\lambda} f(x')) \cdot \partial^\lambda g(x')|}{|x - x'|^\gamma} \\
&\le ||\partial^{\beta-\lambda} f||_\infty \cdot ||\partial^\lambda g||_{\Lambda^\gamma(\mathcal{X})} + ||\partial^\lambda g||_\infty \cdot ||\partial^{\beta-\lambda} f||_{\Lambda^\gamma(\mathcal{X})}.
\end{aligned}$$

So, we have that

$$\begin{aligned}
||fg||_{\Lambda^p(\mathcal{X})} &= \sum_{|\beta|_1 \le m} ||D^\beta fg||_\infty + \sum_{|\beta|_1 = m} ||D^\beta fg||_{\Lambda^\gamma(\mathcal{X})} \\
&= \sum_{|\beta|_1 \le m} ||\sum_{\lambda \le \beta} C(\beta, \lambda) \cdot \partial^{\beta-\lambda} f \cdot \partial^\lambda g||_\infty + \sum_{|\beta|_1 = m} ||\sum_{\lambda \le \beta} C(\beta, \lambda) \cdot \partial^{\beta-\lambda} f \cdot \partial^\lambda g||_{\Lambda^\gamma(\mathcal{X})} \\
&\le \sum_{|\beta|_1 \le m} \sum_{\lambda \le \beta} C(\beta, \lambda) ||\partial^{\beta-\lambda} f||_\infty \cdot ||\partial^\lambda g||_\infty \\
&\quad + \sum_{|\beta|_1 = m} \sum_{\lambda \le \beta} C(\beta, \lambda) \cdot (||\partial^{\beta-\lambda} f||_\infty \cdot ||\partial^\lambda g||_{\Lambda^\gamma(\mathcal{X})} + ||\partial^\lambda g||_\infty \cdot ||\partial^{\beta-\lambda} f||_{\Lambda^\gamma(\mathcal{X})}) \\
&\le C_m \cdot \Big(\sum_{|\beta|_1 \le m} \sum_{\lambda \le \beta} ||\partial^{\beta-\lambda} f||_\infty \cdot ||\partial^\lambda g||_\infty \\
&\quad + \sum_{|\beta|_1 = m} \sum_{\lambda \le \beta} ||\partial^{\beta-\lambda} f||_\infty \cdot ||\partial^\lambda g||_{\Lambda^\gamma(\mathcal{X})} + ||\partial^\lambda g||_\infty \cdot ||\partial^{\beta-\lambda} f||_{\Lambda^\gamma(\mathcal{X})} \Big) \\
&\le C_m \cdot \Big(\sum_{|\beta|_1 \le m} \sum_{|\lambda|_1 \le m} ||\partial^\beta f||_\infty \cdot ||\partial^\lambda g||_\infty + \sum_{|\beta|_1 \le m} \sum_{|\lambda|_1 \le m} ||\partial^\beta f||_\infty \cdot ||\partial^\lambda g||_{\Lambda^\gamma(\mathcal{X})} + ||\partial^\lambda g||_\infty \cdot ||\partial^\beta f||_{\Lambda^\gamma(\mathcal{X})} \Big) \\
&\le C'_m \cdot ||f||_{\Lambda^p(\mathcal{X})} \cdot ||g||_{\Lambda^p(\mathcal{X})}.
\end{aligned}$$

EC.7.2. Proof of Lemma 17

We use the Faa di Bruno formula for the chain rule.

EC.7.2.1. Multivariate Faa di Bruno formula We use Theorem 2.1 from (Savits 2006). Define the following notation $k! = \prod_{i=1}^{s} k_i!$. Let $\beta \in \mathbb{Z}_+^d$. Let $h : \mathbb{R}^d \to \mathbb{R}$. Let $x_0 \in \mathbb{R}^d$. Assume that $g = (g_1, g_2, \dots g_m)$, where $g_i : \mathbb{R}^d \to \mathbb{R}$ and $g_i \in C^m(x_0)$ and $f \in C^m(y_0)$, where $y_0 = (g_1(x_0), \dots g_m(x_0))$. Define that for $\ell \in \mathbb{Z}_+^d$, we define $D^\ell g := (D^\ell g_1, D^\ell g_2, \dots, D^\ell g_m)$. Let $h(x_1, x_2, \dots x_d) = f(g_1(x_0), \dots, g_m(x_0))$. Let $q = \prod_{i=1}^{d}(\beta_i + 1) - 1$, and let $\ell_1, \ell_2, \dots \ell_q$ be a complete listing of all $\ell \in \mathbb{Z}_+^d$ such that $\ell \le \beta$ with $|\ell|_1 > 0$.

Under these conditions $D^\beta h(x_0)$ exists and is given by

$$D^\beta h(x_0) = \sum_{\lambda \in \mathbb{Z}_m^+ : 1 \le |\lambda|_1 \le |\beta|_1} f^\lambda(y_0) \sum_{P(\beta,\lambda)} \beta! \cdot \prod_{j=1}^{q} \frac{[D^{\ell_j} g(x_0)]^{k_j}}{(k_j!)[\ell_j!]^{k_j}},$$

where

$$P(\beta, \lambda) = \Big\{ (k_1, k_2, \dots k_q) \Big| k_i \in \mathbb{Z}_+^m,\ |k_i|_1 \ge 0,\ \sum_{i=1}^{q} k_i = \lambda,\ \sum_{i=1}^{q} |k_i|_1 \ell_i = \beta \Big\}.$$

In other words,

$$D^\beta h(x_0) = \sum_{\lambda : 1 \le |\lambda|_1 \le |\beta|_1} \sum_{P(\beta,\lambda)} C(\beta, \lambda, k_1, k_2, \dots k_q) \cdot f^{(\lambda)}(y_0) \prod_{j=1}^{q} [D^{\ell_j} g(x_0)]^{k_j}$$

for some universal constants $C(\beta, \lambda, k_1, k_2, \dots k_q)$.

Let $x_0 \in \mathcal{X}$. Let $y_0 = t(x_0)$. Let $u(x) := v(x, t(x))$. Let $v^{(\beta,r)}(x_0) = \frac{\partial^{|\beta|_1 + r}}{\partial x_1^{\beta_1} \dots \partial x_d^{\beta_d} \partial t^r} v(x, y_0)$ for any β, r with $|\beta|_1 + r \le m$. Let $g(x) := (g_1(x), g_2(x), \dots g_d(x), t(x))$, where $g_i(x) = x_i$. Define $\{\ell_i\}_{i=1}^q$ as the complete listing of vectors $\ell \in \mathbb{Z}_+^d$ with $\ell \le \beta$ where $[\ell] > 0$.

We can express λ from multivariate Faa di Bruno formula as $\lambda := (\beta', r)$, where $\beta' \in \mathbb{Z}_d^+$ and $r \in \mathbb{Z}_+$. Applying multivariate Faa di Bruno gives that

$$D^\beta u(x_0) = \sum_{\substack{(\beta', r) \in \mathbb{Z}_+^{d+1}: \\ 1 \le |\beta'| + r \le |\beta|_1}} \sum_{P(\beta, \beta', r)} C(\beta, \beta', r, k_1, k_2, \dots k_q) \cdot v^{(\beta', r)}(x_0) \prod_{j=1}^{q} [D^{\ell_j} g(x_0)]^{k_j},$$

where

$$P(\beta, \beta', r) = \left\{ (k_1, k_2, \dots k_q) \Big| k_i \in \mathbb{Z}_+^{d+1},\ |k_i|_1 \ge 0,\ \sum_{i=1}^{q} k_i = (\beta', r),\ \sum_{i=1}^{q} |k_i|_1 \cdot \ell_i = \beta \right\}.$$

EC.7.3. Proof of Lemma 18

When $p \le 1$, we show that $u \in \Lambda^{p^2}(\mathcal{X})$. Let $x, x' \in \mathcal{X}$.

$$\begin{aligned}
|u(x') - u(x)| &= |v(x', t(x')) - v(x, t(x))| \\
&\le ||v||_{\Lambda^p(\mathcal{X} \times \mathcal{T})} \cdot |(x, t(x)) - (x', t(x'))|_2^p \\
&\le C||v||_{\Lambda^p(\mathcal{X} \times \mathcal{T})} \cdot (|x - x'|_2^p + |t(x) - t(x')|_2^p) \\
&\le C||v||_{\Lambda^p(\mathcal{X} \times \mathcal{T})} \cdot (|x - x'|_2^p + ||t||_{\Lambda^p(\mathcal{X})}^p \cdot |x - x'|_2^{p^2})
\end{aligned}$$

$$
\begin{aligned}
&\leq C||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \cdot |x-x'|_2^{p-p^2} \cdot |x-x'|_2^{p^2} + C||v||_{\Lambda^p(\mathcal{X})} \cdot ||t||^p_{\Lambda^p(\mathcal{X})} |x-x'|_2^{p^2} \\
&\leq C||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \cdot \Big(\sup_{x,x'\in\mathcal{X}} |x-x'|_2^{p-p^2} \Big) \cdot |x-x'|_2^{p^2} + C||v||_{\Lambda^p(\mathcal{X})} \cdot ||t||^p_{\Lambda^p(\mathcal{X})} |x-x'|_2^{p^2} \\
&\leq K|x-x'|_2^{p^2}.
\end{aligned}
$$

The first line holds because $v \in \Lambda^p(\mathcal{X}\times\mathcal{T})$. The second line follows because $||z_1+z_2||^p \leq 2^p(||z_1||^p + ||z_2||^p)$ for any $p \geq 0$. The third line holds because $t \in \Lambda^p(\mathcal{X})$. Finally, we use the compactness of $\mathcal{X}$ to demonstrate that there exists some constant $K \in \mathbb{R}_+$ such that $|u(x')-u(x)| \leq K||x-x'||_2^{p^2}$. Thus, $u \in \Lambda^{p^2}(\mathcal{X})$.

With no additional assumptions, we can conclude that $u \in \Lambda^{\min\{p,p^2\}}(\mathcal{X})$.

If we additionally have that $p \leq 1$ and $|\frac{\partial}{\partial t} v(x,t)| \leq L$, we have that

$$
\begin{aligned}
|u(x')-u(x)| &= |v(x',t(x'))-v(x,t(x))| \\
&= |v(x',t(x'))-v(x',t(x))+v(x',t(x))-v(x,t(x))| \\
&\leq |v(x',t(x'))-v(x',t(x))| + |v(x',t(x))-v(x,t(x))| \\
&\leq L\cdot|t(x')-t(x)| + ||v||_{\Lambda^p(\mathcal{X})}|x'-x|_2^\gamma \\
&= L\cdot||t||_{\Lambda^p(\mathcal{X})}\cdot|x'-x|_2^\gamma + ||v||_{\Lambda^p(\mathcal{X})}|x'-x|_2^\gamma \\
&= (L\cdot||t||_{\Lambda^p(\mathcal{X})} + ||v||_{\Lambda^p(\mathcal{X})})\cdot|x-x'|_2^\gamma.
\end{aligned}
$$

Thus, $u \in \Lambda^p(\mathcal{X})$ and $||u||_{\Lambda^p(\mathcal{X})} \leq ||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} + L\cdot||t||_{\Lambda^p(\mathcal{X})}$.

When $p > 1$, we show that $u \in \Lambda^p(\mathcal{X})$. Clearly, $u \in C^m(\mathcal{X})$ because $v \in C^m(\mathcal{X}\times\mathcal{T})$ and $t \in C^m(\mathcal{X})$. By Lemma 16, it remains to show that for any $\beta \in \mathbb{Z}_+^d$ where $|\beta|_1 = m$, $D^\beta u$ is a sum of product of functions that lie in $\Lambda^\gamma(\mathcal{X})$, so this implies that $D^\beta u \in \Lambda^\gamma(\mathcal{X})$. From Lemma 17, $D^\beta u(x)$ is the sum of products of $v^{(\beta',r)}$, $D^{\ell_j}t(x)$, and $D^{\ell_j}g_i(x)$, where these terms are defined in Lemma 17. The values of (β',r) involved in the sum satisfy $|\beta'|+r \leq |\beta|_1 = m$, so we have that $v^{(\beta',r)} \in \Lambda^{p-|\beta'|-r} \subset \Lambda^\gamma(\mathcal{X})$. We note that by definition $\ell_j \leq \beta$, so $[\ell_j] \leq m$. Since $t(x) \in \Lambda^p(\mathcal{X})$, we have that $D^{\ell_j}t \in \Lambda^{p-[\ell_j]}(\mathcal{X}) \subset \Lambda^\gamma(\mathcal{X})$. In addition $g_i(x) \in \Lambda^q(\mathcal{X})$ for any $q < \infty$ for all $i = 1,2,\dots d$, so we also have that $D^{\ell_j}g_i(x) \in \Lambda^\gamma(\mathcal{X})$. Thus, we have that $D^\beta u \in \Lambda^\gamma(\mathcal{X})$, so $u \in \Lambda^p(\mathcal{X})$.

We use the property (EC.69); we bound on $||u||_{\Lambda^p(\mathcal{X})}$ by obtaining a bound on $||D^\beta u||_{\Lambda^{p-1}(\mathcal{X})}$, where $\beta \in \mathbb{Z}_+^d$. We recall from Lemma 17 that

$$
D^\beta u = \sum_{\substack{\beta'\in\mathbb{Z}_+^d, r\in\mathbb{Z}_+: \\ 1\leq|\beta'|+r\leq|\beta|_1}} \sum_{P(\beta,\beta',r)} C(\beta,\beta',r,k_1,k_2,\dots k_q)\cdot v^{(\beta',r)}(x_0)\prod_{j=1}^q [D^{\ell_j}g(x_0)]^{k_j},
$$

where $g(x) = (x,t(x))$, $\ell_j \in \mathbb{Z}_+^d$, $\ell_j \leq \beta$, and $\sum_{j=1}^q [k_j]\cdot\ell_j = \beta$.

As an intermediate step to bounding the Hölder norm of $||D^\beta u||_{\Lambda^{p-1}(\mathcal{X})}$, we bound $||D^{\ell_j}g_i||_{\Lambda^p(\mathcal{X})}$ and $||\prod_{j=1}^q [D^{\ell_j}g]^{k_j}||_{\Lambda^{p-1}(\mathcal{X})}$. for some choice of β', r and $k_1,k_2,\dots k_q \in P(\beta,\beta',r)$. Since we have

that $g_i(x) = x_i$, we have that $||D^{\ell_j} g_i||_{\Lambda^{p-1}(\mathcal{X})}$ is a constant that depends only on $||x_i||_\infty$. So, we can say that there exists $C > 0$ such that $||D^{\ell_j} g_i||_{\Lambda^{p-1}(\mathcal{X})} \le C$. We have that

$$\begin{aligned}
\Big|\Big|\prod_{j=1}^{q}[D^{\ell_j} g]^{k_j}\Big|\Big|_{\Lambda^{p-1}(\mathcal{X})} &= \Big|\Big|\prod_{j=1}^{q}\Big((D^{\ell_j} t)^{k_{j,d+1}} \cdot \prod_{i=1}^{d}(D^{\ell_j} g_i)^{k_{j,i}}\Big)\Big|\Big|_{\Lambda^{p-1}(\mathcal{X})} \\
&\le C_{p-1} \cdot \prod_{j=1}^{q} ||D^{\ell_j} t||^{k_{j,d+1}}_{\Lambda^{p-1}(\mathcal{X})} \cdot \prod_{i=1}^{d} ||(D^{\ell_j} g_i)||^{k_{j,i}}_{\Lambda^{p-1}(\mathcal{X})} \\
&\le C'_{p-1} \cdot \prod_{j=1}^{q} ||D^{\ell_j} t||^{k_{j,d+1}}_{\Lambda^{p-1}(\mathcal{X})} \\
&\le C'_{p-1} \cdot ||D^{\ell_j} t||^{m}_{\Lambda^{p-1}(\mathcal{X})}.
\end{aligned}$$

The second line follows from Lemma 16. The third line follows from the observation that $D^{\ell_j} g_i = 1$ if $\ell_j = e_i$, where e_i is the i-th unit vector in $\mathbb{Z}^d_+$, otherwise $D^{\ell_j} g_i = 0$, so $||D^{\ell_j} g_i||_{\Lambda^{p-1}(\mathcal{X})} \le 1$, so we can view $\prod_{i=1}^{d} ||(D^{\ell_j} g_i)||^{k_{j,i}}_{\Lambda^{p-1}(\mathcal{X})}$ as a constant with no dependence on d.

We bound the Hölder norm of the RHS.

$$\begin{aligned}
||D^\beta u||_{\Lambda^{p-1}(\mathcal{X})} &\le \Big|\Big| \sum_{\substack{\beta' \in \mathbb{Z}^d_+, r \in \mathbb{Z}_+: \\ 1 \le |\beta'| + r \le |\beta|_1}} \sum_{P(\beta,\beta',r)} C(\beta,\beta',r,k_1,k_2,\dots k_q) \cdot v^{(\beta',r)} \prod_{j=1}^{q} [D^{\ell_j} g]^{k_j} \Big|\Big|_{\Lambda^{p-1}(\mathcal{X})} \\
&\le \sum_{\substack{\beta' \in \mathbb{Z}^d_+, r \in \mathbb{Z}_+: \\ 1 \le |\beta'| + r \le |\beta|_1}} \sum_{P(\beta,\beta',r)} C(\beta,\beta',r,k_1,k_2,\dots k_q) \cdot ||v^{(\beta',r)} \prod_{j=1}^{q} [D^{\ell_j} g]^{k_j}||_{\Lambda^{p-1}(\mathcal{X})} \\
&\le \sum_{\substack{\beta' \in \mathbb{Z}^d_+, r \in \mathbb{Z}_+: \\ 1 \le |\beta'| + r \le |\beta|_1}} \sum_{P(\beta,\beta',r)} C(\beta,\beta',r,k_1,k_2,\dots k_q) \cdot ||v^{(\beta',r)} \prod_{j=1}^{q} [D^{\ell_j} g]^{k_j}||_{\Lambda^{p-1}(\mathcal{X})} \\
&= \sum_{\substack{\beta' \in \mathbb{Z}^d_+, r \in \mathbb{Z}_+: \\ 1 \le |\beta'| + r \le |\beta|_1}} \sum_{P(\beta,\beta',r)} C(\beta,\beta',r,k_1,k_2,\dots k_q) \cdot C_{p-1} ||v^{(\beta',r)}||_{\Lambda^{p-1}(\mathcal{X})} \cdot ||\prod_{j=1}^{q} [D^{\ell_j} g]^{k_j}||_{\Lambda^{p-1}(\mathcal{X})} \\
&\le C_\beta \cdot ||v^{(\beta',r)}||_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})} \cdot C'_{p-1} \cdot ||D^{\ell_j} t||^m_{\Lambda^{p-1}(\mathcal{X})} \\
&= C'_{\beta,p-1} \cdot ||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \cdot ||t||^p_{\Lambda^m(\mathcal{X})}.
\end{aligned}$$

Thus, there exists $C_p > 0$ such that

$$||u||_{\Lambda^p(\mathcal{X})} \le C_p \cdot ||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \cdot ||t||^m_{\Lambda^p(\mathcal{X})}.$$

EC.7.4. Proof of Lemma 19

We consider $p \le 1$, so $m = 0$ and $\gamma \in (0, 1]$. For any $x, x' \in \mathcal{X}$, we have that

$$\Big|\frac{1}{t(x)} - \frac{1}{t(x')}\Big| = \Big|\frac{t(x') - t(x)}{t(x)t(x')}\Big|$$

$$= \frac{|t(x') - t(x)|}{|t(x)| \cdot |t(x')|}$$
$$\leq \frac{1}{(K_1)^2} \cdot ||t||_{\Lambda^p(\mathcal{X})} \cdot |x' - x|_2^p.$$

Thus, we also have that $1/t \in \Lambda^p(\mathcal{X})$. In addition, there exists a constant C_{K_1} such that $||\frac{1}{t}||_{\Lambda^p(\mathcal{X})} \leq C_{K_1} \cdot ||t||_{\Lambda^p(\mathcal{X})}$.

Now we consider the case where $p > 1$. Let $v(x,t) = 1/t$, where $K = ||t(x)||_\infty$ which is finite because $\mathcal{X}$ is compact and t is continuous and define $\mathcal{T} = [K_1, K_2]$. Then we have that $v(x,t) \in \Lambda^p(\mathcal{X} \times \mathcal{T})$. Thus, we can consider the function $u(x) := v(x, t(x)) = \frac{1}{t(x)}$ and apply Lemma 18 to see that $u = 1/t \in \Lambda^p(\mathcal{X})$ and $||u||_{\Lambda^p(\mathcal{X})} \leq C_p \cdot ||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} \cdot ||t||^m_{\Lambda^p(\mathcal{X})}$. Since the function v only depends on its argument t, we realize that $||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})} = ||v||_{\Lambda^p(\mathcal{T})}$ and consequently, there exists $C_{K_1,K_2} > 0$ such that $||v||_{\Lambda^p(\mathcal{T})} < C_{K_1,K_2}$. Thus, we have that for $p > 1$, $||\frac{1}{t}||_{\Lambda^p(\mathcal{X})} \leq C_p \cdot C_{K_1,K_2} \cdot ||t||^m_{\Lambda^p(\mathcal{X})}$.

EC.7.5. Proof of Theorem 5

We prove the result by induction. We start with the case that $p \leq 1$, so $m = 0$ and $\gamma \in (0, 1]$. Let $x, x' \in \mathcal{X}$. By the Mean Value Theorem and the fact that v is strictly increasing in t, we have that

$$|v(x, t(x)) - v(x, t(x'))| \geq K_1 \cdot |t(x') - t(x)|.$$

Also, since $v \in \Lambda^p(\mathcal{X} \times \mathcal{T})$

$$|v(x', t(x')) - v(x, t(x'))| \leq ||v||_{\Lambda^\gamma(\mathcal{X})} \cdot |x - x'|_2^\gamma.$$

Combining these results, we have that

$$c \cdot |t(x') - t(x)| \leq |v(x, t(x)) - v(x, t(x'))|.$$

We can use the fact that $v(x, t(x)) = v(x', t(x')) = 0$ to rewrite the RHS of the above inequality.

$$\begin{aligned} |v(x, t(x)) - v(x, t(x'))| &= |v(x', t(x')) - v(x, t(x'))| \\ &= |v(x', t(x')) - v(x, t(x'))| \\ &\leq ||v(\cdot, t(x'))||_{\Lambda^\gamma(\mathcal{X})} \cdot |x - x'|_2^\gamma. \end{aligned}$$

Thus,

$$|t(x') - t(x)| \leq \frac{||v(\cdot, t(x'))||_{\Lambda^\gamma(\mathcal{X})}}{K_1} \cdot |x - x'|_2^\gamma,$$

so $t \in \Lambda^\gamma(\mathcal{X})$ and there is a constant $C_{K_1,K_2,\gamma}$ such that

$$||t||_{\Lambda^\gamma(\mathcal{X})} \leq C_{K_1,K_2,\gamma} \cdot ||v||_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})}. \tag{EC.73}$$

Next we show that when $m=1$, $D^\beta t$ exists for any β with $|\beta|_1=1$ and $j=1,2,\ldots,d$,

$$\frac{\partial}{\partial x_j}t(x) = -\frac{\frac{\partial}{\partial x_j}v(x,t)\big|_{t=t(x)}}{\frac{\partial}{\partial t}v(x,t)\big|_{t=t(x)}}. \tag{EC.74}$$

We note that the previous case proves that t is Lipschitz because $t\in\Lambda^\gamma(\mathcal{X})$ for $\gamma=1$. Again, let $x,x'\in\mathcal{X}$. We have that $|t(x')-t(x)|=O(|x'-x|_2)$. We note that

$$v(x',t(x'))-v(x',t(x))+v(x',t(x))-v(x,t(x))=0.$$

Since $m=1$, we can apply the Mean Value Theorem twice to the above equation to see that

$$\frac{\partial}{\partial t}v(x',t)\big|_{t=t(x)}\cdot(t(x')-t(x))+\sum_{j=1}^d\frac{\partial}{\partial x_j}v(x,t)\big|_{t=t(x)}\cdot(x_j-x_j')=o(|x-x'|_2). \tag{EC.75}$$

Since $v\in\Lambda^{1+\gamma}(\mathcal{X}\times\mathcal{T})$, then $\frac{\partial}{\partial t}v(x,t)\in\Lambda^\gamma(\mathcal{X}\times\mathcal{T})$, so

$$\frac{\partial}{\partial t}v(x',t)\big|_{t=t(x)}=\frac{\partial}{\partial t}v(x,t)\big|_{t=t(x)}+O(|x-x'|_2^\gamma).$$

Plugging this into (EC.75), we have that

$$\Big(\frac{\partial}{\partial t}v(x,t)\big|_{t=t(x)}+O(|x-x'|_2^\gamma)\Big)\cdot(t(x')-t(x))+\sum_{j=1}^d\frac{\partial}{\partial x_j}v(x,t)\big|_{t=t(x)}\cdot(x_j-x_j')=o(|x-x'|_2). \tag{EC.76}$$

Since $t(x)-t(x')=O(|x-x'|_2)$, (EC.76) simplifies to

$$\frac{\partial}{\partial t}v(x,t)\big|_{t=t(x)}\cdot(t(x')-t(x))+\sum_{j=1}^d\frac{\partial}{\partial x_j}v(x,t)\big|_{t=t(x)}\cdot(x_j-x_j')=o(|x-x'|_2).$$

The above equation implies (EC.74). We this form (EC.74) and the property (EC.69) to obtain a bound on $||t||_{\Lambda^{1+\gamma}(\mathcal{X})}$ by obtaining an upper bound on $||\frac{\partial}{\partial x_j}t||_{\Lambda^\gamma(\mathcal{X})}$.

As a preliminary, we upper bound

$$\left|\left|\frac{\frac{\partial}{\partial x_j}v(x,t)}{\frac{\partial}{\partial t}v(x,t)}\right|\right|_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})}.$$

Denote $f(x,t)=\frac{\partial}{\partial x_j}v(x,t)$ and $g(x,t)=\frac{\partial}{\partial t}v(x,t)$. Recall that $f,g\in\Lambda^\gamma(\mathcal{X}\times\mathcal{T})$, $|f|\le L$ and $g(x,t)\in[K_1,K_2]$. By (EC.69), we must have that $||f||_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})},||g||_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})}\le||v||_{\Lambda^{1+\gamma}(\mathcal{X}\times\mathcal{T})}$.

$$\begin{aligned}
\Big|\frac{f(x,t)}{g(x,t)}-\frac{f(x',t')}{g(x',t')}\Big| &= \Big|\frac{f(x,t)}{g(x,t)}-\frac{f(x',t')}{g(x,t)}+\frac{f(x',t')}{g(x,t)}-\frac{f(x',t')}{g(x',t')}\Big| \\
&\le \frac{1}{g(x,t)}\cdot|f(x,t)-f(x',t')|+|f(x',t')|\cdot\frac{|g(x,t)-g(x',t')|}{|g(x,t)|\cdot|g(x',t')|} \\
&\le \frac{1}{K_1}\cdot|f(x,t)-f(x',t')|+L\cdot\frac{|g(x,t)-g(x',t')|}{(K_1)^2} \\
&= C_{K_1,K_2,L}\cdot(||f||_{\Lambda^\gamma(\mathcal{X})}+||g||_{\Lambda^\gamma(\mathcal{X})})\cdot|(x,t)-(x',t')|_2^\gamma \\
&= C_{K_1,K_2,L}\cdot||v||_{\Lambda^{1+\gamma}(\mathcal{X})}\cdot|(x,t)-(x',t')|_2^\gamma
\end{aligned}$$

This implies that

$$\Big|\Big|\frac{f}{g}\Big|\Big|_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})} = \Big|\Big|\frac{\frac{\partial}{\partial x_j}v(x,t)}{\frac{\partial}{\partial t}v(x,t)}\Big|\Big|_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})} \leq C_{K_1,K_2,L}||v||_{\Lambda^{1+\gamma}(\mathcal{X})}. \tag{EC.77}$$

We can use this result to show that $||\frac{\partial}{\partial x_j}t||_{\Lambda^\gamma(\mathcal{X})} \leq ||v||_{\Lambda^{1+\gamma}(\mathcal{X})}$.

We have that

$$\begin{aligned}
\Big|\Big|\frac{\partial}{\partial x_j}t\Big|\Big|_{\Lambda^\gamma(\mathcal{X})} &= \left|\left|\frac{\frac{\partial}{\partial x_j}v(x,t)}{\frac{\partial}{\partial t}v(x,t)}\Big|_{t=t(x)}\right|\right|_{\Lambda^\gamma(\mathcal{X})} \\
&\leq \left|\left|\frac{\frac{\partial}{\partial x_j}v(x,t)}{\frac{\partial}{\partial t}v(x,t)}\right|\right|_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})} + K_2\cdot||t||_{\Lambda^\gamma(\mathcal{X})} \\
&\leq C_{K_1,K_2,L}||v||_{\Lambda^{1+\gamma}(\mathcal{X}\times\mathcal{T})} + K_2||t||_{\Lambda^\gamma(\mathcal{X})} \\
&\leq C_{K_1,K_2,L}||v||_{\Lambda^{1+\gamma}(\mathcal{X}\times\mathcal{T})} + C_{K_1,K_2,p-1}\cdot||v||_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})} \\
&= C'_{K_1,K_2,L,p}||v||_{\Lambda^{1+\gamma}(\mathcal{X}\times\mathcal{T})}.
\end{aligned}$$

The first line follows from (EC.74). The second line is due to the norm of the composition of Hölder functions in Lemma 18. The third line is due to (EC.77). The fourth line is due to (EC.73). The fifth line holds because $||v||_{\Lambda^\gamma(\mathcal{X}\times\mathcal{T})} \leq ||v||_{\Lambda^{1+\gamma}(\mathcal{X}\times\mathcal{T})}$. Thus, there exists $C_{K_1,K_2,L,p,d} > 0$ such that

$$||t||_{\Lambda^{1+\gamma}(\mathcal{X})} \leq C_{K_1,K_2,L,p,=d}\cdot||v||_{\Lambda^{1+\gamma}(\mathcal{X})}.$$

We are left to prove the result for $p > 1$. We prove a stronger claim that

$$||t||_{\Lambda^p(\mathcal{X})} \leq ||v||^{k_p}_{\Lambda^p(\mathcal{X})}, \tag{EC.78}$$

where $k_p = 1$ for any $p < 2$ and $k_p = m + (m-1)\cdot k_{p-1}$. We note that showing that this claim holds yields our desired result in (EC.72) because $k_p \geq 1$, so

$$k_p = m(k_{p-1}+1) - 1 \leq m(k_{p-1}+1) \leq m!(k_{p-m}+1) \leq 2m!.$$

We prove by induction that under the conditions of Theorem 5, (EC.78) holds. Suppose that this result holds for $p-1$. Since $v(x,t) \in \Lambda^p(\mathcal{X}\times\mathcal{T}) \subset \Lambda^{p-1}(\mathcal{X}\times\mathcal{T})$, the inductive hypothesis implies that $t \in \Lambda^{p-1}(\mathcal{X})$ and

$$||t||_{\Lambda^{p-1}(\mathcal{X})} \leq C_{K_1,K_2,L,p-1,d}\cdot||v||^{k_{p-1}}_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}. \tag{EC.79}$$

We aim to show that $t \in \Lambda^p(\mathcal{X})$ and the desired bound on the norm (EC.72) holds. Note that

$$||t||_{\Lambda^p(\mathcal{X})} \leq \sum_{j=1}^{d}\Big|\Big|\frac{\partial}{\partial x_j}t(x)\Big|\Big|_{\Lambda^{p-1}(\mathcal{X})}$$

by (EC.69). So, we bound the norm by bounding $||\frac{\partial}{\partial x_j}t(x)||_{\Lambda^{p-1}(\mathcal{X})}$ and we recall that $\frac{\partial}{\partial x_j}t(x)$ is given by (EC.74). As a preliminary, we note that since $v \in \Lambda^p(\mathcal{X}\times\mathcal{T})$, we have that $\frac{\partial}{\partial x_j}v(x,t), \frac{\partial}{\partial t}v(x,t) \in \Lambda^{p-1}(\mathcal{X}\times\mathcal{T})$. We analyze

$$\begin{aligned}
\Big|\Big|\frac{\partial}{\partial x_j}v(x,t)\cdot\frac{1}{\frac{\partial}{\partial t}v(x,t)}\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})} &\le C_{p-1}\Big|\Big|\frac{\partial}{\partial x_j}v(x,t)\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\cdot\Big|\Big|\frac{1}{\frac{\partial}{\partial t}v(x,t)}\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\\
&\le C_{p-1}\Big|\Big|\frac{\partial}{\partial x_j}v(x,t)\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\cdot C_{K_1}\cdot\Big|\Big|\frac{\partial}{\partial t}v(x,t)\Big|\Big|^{m-1}_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\\
&\le C_{K_1,p-1}\cdot||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})}\cdot||v||^{m-1}_{\Lambda^p(\mathcal{X}\times\mathcal{T})}\\
&\le C_{K_1,p-1}||v||^m_{\Lambda^p(\mathcal{X}\times\mathcal{T})}.
\end{aligned}$$

The first line follows from the product rule for Hölder functions in Lemma 16. The second line follows from the rule for reciprocals of Hölder functions in Lemma 19. The third line holds by definition of the Hölder norm. Thus, we have that

$$\Big|\Big|\frac{\partial}{\partial x_j}v(x,t)\cdot\frac{1}{\frac{\partial}{\partial t}v(x,t)}\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})} \le C_{K_1,p-1}||v||^m_{\Lambda^p(\mathcal{X}\times\mathcal{T})}. \tag{EC.80}$$

Now, we can obtain a bound on $||\frac{\partial}{\partial x_j}t||_{\Lambda^{p-1}(\mathcal{X})}$.

$$\begin{aligned}
\Big|\Big|\frac{\partial}{\partial x_j}t\Big|\Big|_{\Lambda^{p-1}(\mathcal{X})} &= \Big|\Big|\frac{\partial}{\partial x_j}v(x,t)\cdot\frac{1}{\frac{\partial}{\partial t}v(x,t)}\Big|_{t=t(x)}\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\\
&\le C_{p-1}\cdot\Big|\Big|\frac{\partial}{\partial x_j}v(x,t)\cdot\frac{1}{\frac{\partial}{\partial t}v(x,t)}\Big|\Big|_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\cdot||t||^{m-1}_{\Lambda^{p-1}(\mathcal{X})}\\
&\le C'_{p-1}||v||^m_{\Lambda^p(\mathcal{X}\times\mathcal{T})}\cdot||t||^{m-1}_{\Lambda^{p-1}(\mathcal{X})}\\
&\le C''_{p-1}||v||^m_{\Lambda^p(\mathcal{X}\times\mathcal{T})}\cdot\left(C_{K_1,K_2,L,p-1,d}\cdot||v||^{k_{p-1}}_{\Lambda^{p-1}}\right)^{m-1}\\
&\le C''_{K_1,K_2,L,p-1,d}\cdot||v||^{m+k_{p-1}\cdot(m-1)}_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}\\
&= C''_{K_1,K_2,L,p-1,d}\cdot||v||^{k_p}_{\Lambda^{p-1}(\mathcal{X}\times\mathcal{T})}.
\end{aligned}$$

The first line follows by definition. The second line follows from the composition of Hölder functions in Lemma 18. The third line is due to (EC.80). The fourth line is due to the inductive hypothesis (EC.79). Thus, we show by induction that (EC.78) holds and as a consequence, (EC.72) holds and $t \in \Lambda^p(\mathcal{X})$.

EC.7.6. Proof of Corollary 3

Note that $\mathcal{X}$ is compact because f is continuous and $\mathcal{T}$ is compact. Since f is strictly increasing, it is invertible, so $f(f^{-1}(x)) - x = 0$. Let $v : \mathcal{X}\times\mathcal{T} \to \mathbb{R}$, where $v(x,t) := f(t) - x$ Clearly, $v(x,t) \in \Lambda^p(\mathcal{X}\times\mathcal{T})$. In addition, $\frac{\partial}{\partial t}v(x,t) = f'(t) > 0$. Since $\mathcal{T}$ is compact, we have that $\inf_{t\in\mathcal{T}} f'(t) > 0$. In addition, v has bounded derivative with respect to x. We note that $||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})}$ can be upper bounded by a constant that depends on c. By Theorem 5, $f^{-1} \in \Lambda^p(\mathcal{X})$. If $p \le 1$, then $||f^{-1}||_{\Lambda^p} \le C_{K_1,K_2,p}\cdot||v||_{\Lambda^p(\mathcal{X}\times\mathcal{T})}$. If $p > 1$, then a bound on $||f^{-1}||_{\Lambda^p(\mathcal{X})}$ is given by a constant that depends on c, K_1, K_2, p, d by applying (EC.72).

EC.7.7. Proof of Lemma 20

This result holds through an application of Theorem 1 of Brezis and Mironescu (2019). For completeness, we state the full theorem below.

Let $W^{r,q}$ denote the fractional-order Sobolev space (Brezis and Mironescu 2018) where $0<r<1$ and $1\le q\le\infty$. Let $\mathcal{X}$ be a compact set $\mathbb{R}^d$. Let $s_1,s_2,r,p_1,p_2,q,\theta,d$ satisfy the following conditions

$$\begin{aligned}&0\le s_1\le s_2,\quad r\ge 0,\quad 1\le p_1,p_2,q\le\infty,\quad (s_1,p_1)\ne(s_2,p_2),\quad \theta\in(0,1),\\ &r<s:=\theta s_1+(1-\theta)s_2,\\ &\frac{1}{q}=\Big(\frac{\theta}{p_1}+\frac{1-\theta}{p_2}\Big)-\frac{s-r}{d}.\end{aligned}$$

Then the Gagliardo-Nirenberg-Sobolev (GNS) inequality holds (with a few exceptions):

$$||f||_{W^{r,q}(\mathcal{X})}\le C\cdot||f||^{\theta}_{W^{s_1,p_1}(\mathcal{X})}\cdot||f||^{1-\theta}_{W^{s_2,p_2}(\mathcal{X})}\quad\forall f\in W^{s_1,p_1}(\mathcal{X})\cap W^{s_2,p_2}(\mathcal{X}),$$

where C is a constant that depends on s_1,s_2,r,p_1,p_2,q,d but not θ. The exceptions include

1. $d=1$, s_2 is an integer ≥ 1, $1<p_1<\infty$, $p_2=1$, $s_1=s_2-1+\frac{1}{p_1}$, $s_2+\frac{\theta}{p_1}-1<r<s_2+\frac{\theta}{p_1}-\theta$, and $r\ge s_1$.
2. $d\ge 1$, $p_1=\infty$, $1<p_2<\infty$, $q=\infty$, $s_1=r\ge 0$ is an integer, $s_2=r+\frac{d}{p_2}$ (for every $\theta\in(0,1)$).

First, we note that we can express the spaces $L^\infty(\mathcal{X}), L^2(\mathcal{X})$, and $\Lambda^p(\mathcal{X})$ as (fractional order) Sobolev spaces (Brezis and Mironescu 2018). In particular, $L^q(\mathcal{X})=W^{0,q}(\mathcal{X})$ and $\Lambda^p(\mathcal{X})=W^{p,\infty}(\mathcal{X})$. The definition of the norm on general $W^{r,q}(\mathcal{X})$ spaces can be found in Brezis and Mironescu (2018), though we will not use it.

We apply Theorem 1 of Brezis and Mironescu (2019) with the following choice of constants:

$$r=s_1=0,\quad q=p_2=\infty,\quad p_1=2,\quad s_2=p,\quad \theta=\frac{2p}{2p+d},\quad s=\frac{pd}{2p+d}.$$

It is straightforward to verify that these constants satisfy the conditions of the theorem (and do not fall under the exceptions): $s_1,s_2,r\in\mathbb{R}_+$, $1\le p_1,p_2,q\le\infty$, $\theta\in(0,1)$, $r<s$,

$$\theta s_1+(1-\theta)s_2=(1-\theta)p=\Big(1-\frac{2p}{2p+d}\Big)\cdot p=\frac{pd}{2p+d}=s,$$

and

$$\Big(\frac{\theta}{p_1}+\frac{1-\theta}{p_2}\Big)-\frac{s-r}{d}=\frac{p}{2p+d}-\frac{1}{d}\cdot\frac{pd}{2p+d}=0=\frac{1}{q}.$$

Thus, we have that for $f\in W^{0,2}(\mathcal{X})\cap W^{p,\infty}(\mathcal{X})$

$$||f||_{W^{0,\infty}(\mathcal{X})}\le C\cdot||f||^{\frac{2p}{2p+d}}_{W^{0,2}(\mathcal{X})}\cdot||f||^{1-\frac{2p}{2p+d}}_{W^{p,\infty}},$$

where C is a constant that does not depend on f. The above inequality is equivalent to

$$||f||_{L^\infty(\mathcal{X})}\le C\cdot||f||^{\frac{2p}{2p+d}}_{L^2(\mathcal{X})}\cdot||f||^{1-\frac{2p}{2p+d}}_{\Lambda^p(\mathcal{X})}.$$

Since $f \in \Lambda_c^p(\mathcal{X})$, we have that $||f||_{\Lambda^p(\mathcal{X})} \leq 2c$. Thus, we have that the above inequality is equivalent to

$$||f||_{L^\infty(\mathcal{X})} \leq C \cdot ||f||_{L^2(\mathcal{X})}^{\frac{2p}{2p+d}} \cdot (2c)^{1-\frac{2p}{2p+d}},$$

which yields the desired result.