Synthetic Differences-in-Differences with Covariates

David Hirshberg and Sylvia Klosin

October 21, 2024

DISCLAIMER: This is work in progress and content is likely to change substantially

Abstract

We propose a synthetic difference-in-difference estimator that incorporates time-varying covariates. We incorporate covariates into a high-dimensional least squares with correlated error-in-variables setting. We use results from this setting to derive conditions under which our synthetic differences-in-differences estimator is asymptotically normal with estimable variance. Monte Carlo simulations demonstrate that our estimator outperforms classic synthetic difference-in-differences in settings where covariates contain information about the outcome. We illustrate the practical performance of our estimator by studying the impact of subsidy increases on crop insurance choices within the United States Federal Crop Insurance Program (FCIP).

Keywords: panel data, synthetic differences and differences, covariates, treatment effects

1 Introduction

Applied researchers frequently use panel data to estimate treatment effects of interest. Since treatment effects are often not randomly assigned, researchers rely on the structure of panel data to create appropriate control groups for the treated units, thereby enabling valid estimation of treatment effects. Synthetic control (SC) methods, first introduced by Abadie and Gardeazabal [2003] and later expanded by Abadie et al. [2010], have become widely used among applied researchers for this purpose. Arkhangelsky et al. [2021] further extended these methods by introducing synthetic differences-in-differences (SDID). In this paper, we introduce synthetic differences-in-differences with covariates (SDIDC), which allows for the incorporation of time-varying covariates into the SDID framework.

Incorporating covariates can enhance the estimation of treatment effects, even in cases where covariates are not correlated with treatment. In our Monte Carlo simulations, which feature a data-generating process without covariate-driven treatment selection, we find that using SDIDC instead of the classic SDID reduces root mean square error (RMSE) by 50%. The improvement arises from the fact that both SDID and SDIDC estimate an unobserved latent structure from observed outcomes, Y_{it}, which are typically noisy in practice. When covariates are predictive of the outcome, controlling for them reduces noise in the outcome variable, thereby improving the precision of the estimator.

For example, in climate economics, if the outcome Y_{it} is corn yield, and weather variables, known to be exogenous to the treatment, are available and predictive of yield, controlling for these variables enhances the precision of the treatment effect estimate. Similar intuition has been demonstrated in the synthetic control context with multiple outcomes Sun et al. [2023], where controlling for multiple outcomes was shown to reduce noise, thereby leading to more accurate treatment effect estimates.

In settings where covariate selection *does* play a role, many empirical economics papers include time-varying controls in panel regressions. This practice is driven by concerns about time-varying omitted variable bias, which may be correlated with both the treatment and the outcomes. For instance, Bailey and Goodman-Bacon [2015] control for annual county-level government transfers per capita, while East et al. [2023] examine the labor market effects of police-based immigration enforcement policies, using time-varying covariates such as CZ-level annual economic conditions to account for variation in policy implementation across U.S. commuting zones (CZ) over time. In these settings it is important to allow for covariates as our method does.

From a methodological standpoint, our SDIDC approach presents an interesting challenge due to the need to simultaneously estimate the covariate coefficients (β) along with unit (ω) and time (λ) weights. Our theoretical innovation lies in combining the estimation of time and unit weights into a single minimization procedure, rather than estimating the two sets of weights separately. This joint estimation is essential because the covariate coefficient (β) is estimated concurrently with the unit and time weights.

Rather than residualizing the covariates from the outcome and then estimating the unit and time weights on the residuals, we propose a joint estimation approach. Residualizing in this way can lead to inaccurate weight estimates when covariates are highly correlated with the latent factor structure, resulting in noisy residuals. By jointly estimating β, ω, and λ, we mitigate this problem and achieve more reliable estimates.

The rest of the paper proceeds as follows. We provide our form setting in Section 2. Section 3 presents the estimator. Section 4 introduces formal results. Section 5 conducts a simulation study to illustrate the properties of our estimator. Section 6 considers an empirical example of how correcting for covariates with SDIDC impacts treatment effect estimation in an application to crop insurance.

2 Setting

In this paper we assume that the data is generated from a latent factor model, also referred to an interactive fixed-effects model (Xu [2017]). This model allows for treatment effect heterogeneity as in De Chaisemartin and d'Haultfoeuille [2020].

$$Y_{it} = \gamma_i \upsilon_t' + Z_{it}\beta + W_{it}\tau + \epsilon_{it} \tag{1}$$

For the rest of the paper we work with the matrix version of this model. To create the matrix model we stack our Y_{it} observations into an $n \times p$ matrix $\boldsymbol{Y}$ where n gives the the number of units, and p is the number of time periods. We use bold letters to denote matrices and arrays.

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{Z} \cdot \beta + \boldsymbol{W} \times \tau + \boldsymbol{\epsilon} \quad where \quad (\boldsymbol{W} \times \tau)_{it} = W_{it}\tau_{it}. \tag{2}$$

Here $\boldsymbol{L}$ is a noiseless $n \times p$ matrix that is a systemic component; we characterize it as a factor model $\boldsymbol{L} = \boldsymbol{\Gamma\Upsilon'}$. Here $\boldsymbol{\Gamma}$ is a matrix of latent unit factors, and $\boldsymbol{\Upsilon}$ is a matrix of latent time factors. The $\boldsymbol{Z}$ is our $n \times p \times k$ noiseless array of covariates, where k is the number of covariates. The β is a k-length vector that contains the parameters for the k covariates. In this paper we use the center dot notation $\cdot$ to contract the third dimension of $\boldsymbol{Z}$, and so $\boldsymbol{Z} \cdot \beta$ is a $n \times p$ matrix. The $\boldsymbol{W}$ is our $n \times p$ treatment matrix; we assume block treatment assignment, so $W_{it} = 1(\{i > N_{co}, t > T_{pre}\})$. We use subscript *co* to denote control units, *tr* for treated units, and *pre* for pre-treatment time periods, and *post* for post-treatment time periods. The last term $\boldsymbol{\epsilon}$ is the idiosyncratic component a $n \times p$ error matrix. We assume that the conditional expectation of the error matrix given $\boldsymbol{W}$, $\boldsymbol{L}$, and $\boldsymbol{Z}$ is zero. This means that treatment assignment can not depend on $\boldsymbol{\epsilon}$, but we are not assuming that $\boldsymbol{W}$ is randomized. Treatment assignment can depend on the systemic component $\boldsymbol{L}$ and the covariates $\boldsymbol{Z}$.

Our estimand of interest τ is the average treatment effect on the treated in the treated time periods. Because of the block treatment assingment, we can write τ as follows.

$$\tau = \frac{1}{N_{tr}, T_{post}} \sum_{i=N_{co+1}}^{N} \sum_{t=T_{pre+1}}^{T} \tau_{it}. \tag{3}$$

Given the block treatment structure of the data, each of our matrices can be seen as having four quadrants. We write out the quadrants of $\boldsymbol{Y}$ out explicitly below. We use the shorter notation.

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{co,pre} & \boldsymbol{Y}_{co,post} \\ \boldsymbol{Y}_{tr,pre} & \boldsymbol{Y}_{tr,post} \end{pmatrix}. \tag{4}$$

3 The Estimator

To estimate τ defined in Equation (3) we estimate unit weights $\hat{\omega}$, time weights $\hat{\lambda}$, and a coefficient vector $\hat{\beta}$. $\hat{\omega}$ is a N_{co}-length vector of unit weights; they weight pre-treatment control units to predict post-treatment outcomes absent exposure for the treated units. $\hat{\lambda}$ is a T_{pre}-length vector of time weights; they weight pre-treatment time periods for control units to predict post-treatment outcomes for control units. Our coefficient vector $\hat{\beta}$ is the regression coefficient vector for our covariates.

As in classic SDID, the estimated unit $\hat{\omega}$ and time $\hat{\lambda}$ weights are only for the control units and time periods (Arkhangelsky et al. [2021]). Treated units and treated time periods are instead averaged by being multiplied by vectors of $1/N_{tr}$ and $1/T_{post}$, to make that clear we use $\lambda_{post} \in \mathbb{R}^{T_{post}}$ and $\omega_{tr} \in \mathbb{R}^{N_{tr}}$. For any weights $\omega \in \Omega$ and $\lambda \in \Lambda$ and coefficients $\beta \in \mathbb{R}^k$ we can define an adjusted weighted double differencing estimator.

$$\hat{\tau}(\omega, \lambda, \beta) = \omega_{tr}' \boldsymbol{Y}_{tr,post} \lambda_{post} - \omega' \boldsymbol{Y}_{co,post} \lambda_{post} - \omega_{tr}' \boldsymbol{Y}_{tr,pre} \lambda + \omega' \boldsymbol{Y}_{co,pre} \lambda \tag{5}$$

Given that treated units and treatment time periods are averaged

$$\begin{pmatrix} Y_{::} & Y_{:T} \\ Y_{N:} & Y_{NT} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y}_{co,pre} & \boldsymbol{Y}_{co,post} \lambda_{post} \\ \omega_{tr}' \boldsymbol{Y}_{tr,pre} & \omega_{tr}' \boldsymbol{Y}_{tr,post} \lambda_{post} \end{pmatrix} .. \tag{6}$$

We estimate the weights for the control units and control time periods by solving the following Tikhonov-regularized least squares problem.

$$(\hat{\omega}, \hat{\lambda}, \hat{\beta}) = \underset{\Theta,\Lambda,B}{\operatorname{argmin}} \, \ell(\omega, \lambda, \beta) \tag{7}$$

$$\begin{aligned} \ell(\omega, \lambda, \beta) = {} & \|Y_{::}'\omega - Z_{:::}'\omega \cdot \beta - Y_{N:}' + Z_{N:\cdot}' \cdot \beta\|^2 + n(\eta^2 - 1)\|\Sigma_{N:}^{1/2}(\omega - \psi)\|^2 \\ & + \|Y_{::}\lambda - Z_{:::}\lambda \cdot \beta - Y_{:T} + Z_{:T\cdot} \cdot \beta\|^2 + p(\eta^2 - 1)\|\Sigma_{T:}^{1/2}(\lambda - \psi)\|^2 \end{aligned} \tag{8}$$

Where $\eta > 0$, $\Sigma_{\epsilon_i} := n^{-1}\mathbb{E}[\epsilon'\epsilon]$ is the row covariance of the errors, and $\psi := \operatorname{argmin}_{z \in \mathbb{R}^p} \mathbb{E}\|\epsilon_{N:} - \epsilon_{::}z\|^2$

We characterize our estimator by showing it converges to the deterministic model we would get by minimizing the expectation of the same loss function

Deterministic Model

$$(\tilde{\omega}, \tilde{\lambda}, \tilde{\beta}) = \underset{\Theta,\Lambda,B}{\operatorname{argmin}} \, \mathbb{E}[\ell(\omega, \lambda, \beta)]. \tag{9}$$

$$\begin{aligned} \mathbb{E}[\ell(\omega, \lambda, \beta)] = {} & \|L_{::}'\omega - Z_{:::}'\omega \cdot \beta - L_{N:}' + Z_{N:\cdot}' \cdot \beta\|^2 + n(\eta^2)\|\Sigma_{\epsilon_i}^{1/2}(\omega - \psi)\|^2 + \mathbb{E}\|\epsilon\psi - \nu\|^2 \\ & + \|L_{::}\lambda - Z_{:::}\lambda \cdot \beta - L_{:T} + Z_{:T\cdot} \cdot \beta\|^2 + p(\eta^2)\|\Sigma_{\epsilon_i}^{1/2}(\lambda - \psi)\|^2 + \mathbb{E}\|\epsilon\psi - \nu\|^2. \end{aligned} \tag{10}$$

4 Formal Results

Our main theoretical results is addapting Theorem 1 from Hirshberg [2021] for our setting. Our poof given in Appendix A.

Theorem 4.1. *Consider the setting described above with essentially gaussian and spherical noise and either the rows of $[\varepsilon, \nu]$ independent and identically distributed or the columns of $[\varepsilon, \nu]$ independent with those of ε identically distributed.*

$$\|\Sigma_{\varepsilon}^{1/2}(\hat{\theta} - \tilde{\theta})\| \leq s \quad \textit{and} \quad \|(\hat{\theta}_0 - \tilde{\theta}_0) + A(\hat{\theta} - \tilde{\theta})\| \leq \eta^{1/2} s$$

with probability $1 - c\exp\{-cu(v,s)\}$ *if* s *satisfies the fixed point condition*

$$s^2 \geq c\left[\frac{v^2\sigma^2 w^2(\Theta_s^*)}{\min(\eta_R^2, \eta_R^4)n} + \frac{(v^2\sigma^2\,\mathrm{rank}(A)/p_{\mathit{eff}})^{1/2}}{\eta_R^2 n} + \frac{\sigma\|A\hat{\theta} + \tilde{\theta}_0 - b\|w(\Theta_s^*) + v\sigma^2(n/p_{\mathit{eff}})^{1/2}w(\Theta_s^*)}{\eta_R^2 nR}\right],$$

$$\eta_R^2 = \max(0, \eta^2 - c\,\mathrm{rank}(A)/n).$$

Where $\Theta_s^* = \{\theta - \hat{\theta} : \theta \in \Theta, \|\Sigma_{\varepsilon}^{1/2}(\hat{\theta} - \tilde{\theta})\| \leq s\}$ *and* $x^{1,1/2} = x + x^{1/2}$.

Here $u(v,s) = \min\{v^2\sigma^2 w^2(\Theta_s^*)/s^2, v^2\,\mathrm{rank}(A), n\}$ *for* $v \geq 1$. *The same holds if we substitute for* $\mathrm{rank}(A)$ *a bound on approximate rank: any integer* R *for which*

$$R \geq c\min\left\{\sigma_{R+1}(A)w(\Theta_s^*)/(s + \sigma\nu p_{\mathit{eff}}^{1/2}), \sigma_{R+1}^2(A)/(v^2\sigma^2)\right\}.$$

5 Monte Carlo Evidence

5.1 Data Generating Processes

In our Monte Carlo we compare the performance of our estimator to classic SDID and TWFE under two different DGPs. The two DGPs are identical except for having different treatment assignment mechanisms. The first will have treatment uncorrelated with the covariate, and the second will have treatment related to the covariate through a logit treatment assignment function.

5.1.1 DGP 1: No selection on covariates

Our outcome is generated with additive fixed effects, interactive fixed effects, and both treatment and covariate effects.

$$Y_{it} = a_i + L_{it} + W_{it}\tau + Z_{it}\beta + \epsilon_{it}. \tag{11}$$

The individual fixed effects are drawn from a standard normal distribution.

$$a_i \sim N(0,1). \tag{12}$$

The covariate for each individual is drawn from an T length $AR(1)$ processes with an autocorrelation coefficient of .95 and standard normal errors. For our simulations $\beta = 10$.

$$Z_i \sim AR(1). \tag{13}$$

Treatment only occurs in the last time period T. The probability of treatment in the last time period is p. In our simulations for uncorrelated treatments and covariates we pick $p = .02$. For our simulation we use $\tau = 10$.

$$W_{it} = \begin{cases} 0 & \text{if } t < T \\ 0 & \text{with probability } 1 - p \text{ if } t = T \\ 1 & \text{with probability } p \text{ if } t = T. \end{cases} \tag{14}$$

For the unobserved time varying fixed effects L_{it}, we create two unobserved time varying factors $F1$ and $F2$.

$$\begin{aligned} F1_{it} &= .1 \times t + e1_{it}, \qquad & e1_{it} &\sim N(0,1) \\ F2_{it} &= .3 \times t + e2_{it}, \qquad & e2_{it} &\sim N(0,1). \end{aligned} \tag{15}$$

Treated units, that is units who in their last time period have $W_{it} = 1$, have their $L_{it} = 4F_1 + 1F_2$. Control units their $L_{it} = 4F_1 + 1F_2$ with probability 1/3 and their $L_{it} = 2F_1 + 6F_2$ with probability 2/3. This was done so that treatment was related to L_{it}, and that some control units would be better matches for the treated units than others.

$$L_{it} = \begin{cases} 4F1_{it} + 1F2_{it} & \text{for } i \text{ such that } W_{iT} = 1 \\ 4F1_{it} + 1F2_{it} & \text{with probability } 1/3 \text{ if } W_{iT} = 0 \\ 2F1_{it} + 6F2_{it} & \text{with probability } 2/3 \text{ if } W_{iT} = 0. \end{cases} \tag{16}$$

.

Our noise is also drawn from a standard normal distribution.

$$\epsilon_{it} \sim N(0,1). \tag{17}$$

5.1.2 DGP 2: Selection on covariates

This DGP is identical to the DGP above, only the treatment assignment (Equation (14)) is different. Now instead of a fixed probability of p, we have that p is a function of the covariate.

$$p = \frac{1}{1 + 3\exp\Big(11 - \frac{3}{T-1}\sum_{t=1}^{T-1} Z_{it}\Big)}. \tag{18}$$

A question. What happens if selection is based on noiseless past observations, i.e., $L_{::} + X_{::} \cdot \beta$? Seems plausible in some contexts. More generally, I'm curious about what we'd see in sim with selection trying to approximate some realistic process. Perhaps the sims we do in SDID with some auxilliary info as covariates?

$$W_{it} = \begin{cases} 0 & \text{if } t < T \\ 0 & \text{with probability } 1 - p \text{ if } t = T \\ 1 & \text{with probability } p \text{ if } t = T. \end{cases} \tag{19}$$

5.2 Results

We provide Monte Carlo results for both of the DGPs above. For each DGP we create 1000 datasets with 500 units and 40 time periods and run the following three models to calculate the treatment effect. Recall that the true treatment effect $\tau = 10$.

1. SDID with covariates
2. SDID from Arkhangelsky et al. [2021]
3. Linear two way fixed effects (TWFE)

   The TWFE model is implemented by running the following linear regression model in Equation (20).

$$Y_{it} = a_i + \gamma_t + \beta X_{it} + \tau D_{it} \tag{20}$$

Table 1: DGP 1

Method	$\hat{\tau}$	SE	Coverage	RMSE
SDIDC	9.69	3.19	0.94	10.25
SDID	10.07	4.65	0.95	21.60
TWFE	5.98	3.04	0.74	25.44

Table 2: DGP 2

Method	$\hat{\tau}$	SE	Coverage	RMSE
SDIDC	10.06	2.64	0.95	6.95
SDID	6.97	3.46	0.86	21.14
TWFE	6.00	2.71	0.68	23.36

Data generated with 500 units, 40 time periods, and true treatment effect $\tau = 10$. Tables summarize 1000 Monte Carlo simulations: $\hat{\tau}$ gives average treatment effect, SE gives the standard deviation of the estimator, Coverage gives the 95% coverage level, and RMSE is root mean square error.

The results are in line with what we would expect. Table 1 shows the results for DGP 1, when treatment is unrelated to the covariate. In this DGP, the benefit of controlling for the covariate in SDIDC comes from the fact that it removes some of the noise in the model. This leads SDIDC to have smaller standard errors (SE) and root mean square error (RMSE) than classic SDID. In Table 1 we can see that in comparison to SDID, SDIDC has 30% smaller SE and 50% smaller RMSE. Since in DGP 1 the covariate is not related to treatment assignment, SDID is not biased, since one does not need to control for the covariate for the exogenity assumption to hold. We see that the TWFE model is biased, since this model has no way of controlling for the unobserved L_{it} component that is correlated with treatment. Therefore even though TWFE is controlling for the covariate, omitting controlling for L_{it} causes a violation of the exogenity assumption.

Now in DGP 2, when the covariate is correlated with treatment, not controlling for the covariate leads to biased treatment effects in SDID. We see in Table 2 that SDID has a bias of 3 and loses proper coverage. TWFE has bias of 4 and also loses proper coverage. The only method that is able to estimate the treatmnet effect well is SDIDC, since it is able to control for both L_{it} and Z_{it}.

6 Empirical Application

We empirically study the effects of a subsidy increase in 2009 for treated crops[1] on crop insurance choice in the context of the Federal Crop Insurance Program (FCIP) following Klosin and Solomon

[1] Grainsorghum, Wheat, Soybeans, Corn, Cotton, Rice, Barley,Canola, Fluecuredtob, Pecans, Sunflowers.

[2024]. The FCIP is a government-run and financed insurance program that protects farmers against any hazard to their crops. In the FCIP, farmers can enroll their fields into separate or aggregate policies.[2] The former insures each field independently; the latter insures total yield for a given crop.

We using the universe of insurance data, at a county-crop level, to compare insurance choices and among crops treated with the policy change to crops that were not (yet) treated. Our data source is the FCIP Summary of Business [USDA, 2024]. The FCIP data record all annual crop insurance contracts. The data include contract type, acreage insured, premium paid, total potential liability, subsidy amount, insurance payout amount and loss ratios. The data are at the county x crop level, and we use data 1999 - 2014.

We run a TWFE regression to see how the policy change impacts enrollment in optional insurance. We use the number of acres insured for the county x crop in a year as our control variable.

$$\text{Optional Insurance Percent}_{c,t} = c_i + \gamma_t + \beta \text{Acres}_{c,t} + \tau \Bigg[(\text{Treated Crop}) \times (\text{Year} \geq 2009) \Bigg]_{c,t} + \epsilon_{c,t} \tag{21}$$

We then run SDIDC to see how the treatment estimate changes. The 95% confidence interval for the TWFE does not include the SDID estimates which are smaller in magnitude.

Method	$\hat{\tau}$	SE
SDIDC	0.162	0.016
TWFE	0.094	0.002

[2] In the official terminology, aggregate units are known as 'enterprise' units, and separate units are 'optional' units.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

Martha J Bailey and Andrew Goodman-Bacon. The war on poverty's experiment in public medicine: Community health centers and the mortality of older americans. *American Economic Review*, 105(3):1067–1104, 2015.

Clément De Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

Chloe N East, Annie L Hines, Philip Luck, Hani Mansour, and Andrea Velasquez. The labor market effects of immigration enforcement. *Journal of Labor Economics*, 41(4):957–996, 2023.

David A Hirshberg. Least squares with error in variables. *arXiv preprint arXiv:2104.08931*, 2021.

Sylvia Klosin and Adam Solomon. The scope of insurance: Theory and evidence from us crop insurance. *arXiv preprint arXiv:2311.16260*, 2024.

Liyang Sun, Eli Ben-Michael, and Avi Feller. Using multiple outcomes to improve the synthetic control method. *arXiv preprint arXiv:2311.16260*, 2023.

USDA. Summary of business, 2024. URL https://www.rma.usda.gov/SummaryOfBusiness.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

A Proof Theorem (4.1)

A.1 Intuition

In this paper we use a localization approuch to bound the distance between our estimated weights $(\hat{\omega}, \hat{\lambda}, \hat{\beta})$ and the oracle weights $(\tilde{\omega}, \tilde{\lambda}, \tilde{\beta})$ (Wainwright [2019], Chapter 14).

To characterize $(\hat{\omega}, \hat{\lambda}, \hat{\beta})$ we use the property that $\mathcal{L}(\delta) \leq 0$ for $\mathcal{L}$ defined in Equation (22).

$$\mathcal{L}(\delta) = \mathcal{L}(\delta_\omega, \delta_\lambda, \delta_\beta) = \ell(\tilde{\omega} + \delta_\omega, \tilde{\lambda} + \delta_\lambda, \tilde{\beta} + \delta_\beta) - \ell(\tilde{\omega}, \tilde{\lambda}, \tilde{\beta}) \tag{22}$$

We define a set $\Theta^*_{r,s}$ to be the set of all δ satisfying the following constraints.

$$\begin{aligned}
&\|\Sigma^{1/2}_{\varepsilon_{i\cdot}}\delta_\omega\| \leq s_\omega \\
&\|\Sigma^{1/2}_{\varepsilon_{\cdot j}}\delta_\lambda\| \leq s_\lambda \\
&\|L'_{::}\delta_\omega - Z'_{::\cdot}\delta_\omega \cdot \tilde{\beta} - Z'_{::\cdot}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}) \cdot \delta_\beta\| \leq r_\omega \\
&\|L_{::}\delta_\lambda - Z_{::\cdot}\delta_\lambda \cdot \tilde{\beta} - Z_{::\cdot}\delta_\lambda \cdot \delta_\beta + (Z_{:T\cdot} - Z_{::\cdot}\tilde{\lambda}) \cdot \delta_\beta\| \leq r_\lambda \\
&\|\delta_\beta\| \leq f(r,s)
\end{aligned} \tag{23}$$

characterised by two variables, $s \in \mathbb{R}$ which controls the coefficient dimension and $r \in \mathbb{R}$ which controls the prediction space.

We will show that, with high probability, $\hat{\delta}$ is in this set by (i) using an explicit characterization of $\hat{\beta}$ as a function of $\hat{\lambda}$ and $\hat{\omega}$ to establish the bound $\|\hat{\beta} - \tilde{\beta}\| \leq f(r,s)$ and (ii) showing that, for any triple $(\delta_\omega, \delta_\lambda, \delta_\beta)$ with $\|\delta_\beta\| \leq f(r,s)$, $\mathcal{L}(\delta) > 0$ unless the other constraints in (23) are satisfied, i.e., unless $\delta \in \Theta^\star_{r,s}$. The bounds we state in [some theorem] are, roughly speaking, the smallest values of r and s for which we can do this.

We will work with this lower bound on $\mathcal{L}(\delta)$, which holds for all $\delta \in \Theta^*_{r,s}$ with probability $1-p$.

$$\mathcal{L}(\delta) \geq \kappa_\omega\|A\delta_\omega\delta_\beta\|^2 + \kappa_\omega n\eta^2\|\Sigma^{1/2}\delta_\omega\|^2 + \kappa_\lambda\|A\delta_\lambda\delta_\beta\|^2 + \kappa_\lambda p\eta^2\|\Sigma^{1/2}\delta_\lambda\|^2 \tag{24a}$$

$$- 2|\max(\|A\delta_\omega\delta_\beta\|/r_\omega, 1)cKvp^{-1/2}_{\text{eff},\Sigma}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \tag{24b}$$

$$- 2|\max(\|A\delta_\lambda\delta_\beta\|/r_\lambda, 1)cKvp^{-1/2}_{\text{eff},\Sigma}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \tag{24c}$$

$$- 2|cKv\,\mathrm{RMSE}_\omega\,\mathrm{w}_\Sigma(\Theta^\star_{s_\omega})| \quad \text{for} \quad \mathrm{RMSE}_\omega := \|(L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega} \cdot \tilde{\beta} - L'_{:N} + Z'_{N:\cdot} \cdot \tilde{\beta})\| \tag{24d}$$

$$- 2|cKv\,\mathrm{RMSE}_\lambda\,\mathrm{w}_\Sigma(\Theta^\star_{s_\lambda})| \quad \text{for} \quad \mathrm{RMSE}_\lambda := \|(L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda} \cdot \tilde{\beta} - L_{:T} + Z_{:T\cdot} \cdot \tilde{\beta})\| \tag{24e}$$

$$- 2|cvK^2(n/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_s)| \tag{24f}$$

$$- 2|cvK^2(p/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_\lambda)| \tag{24g}$$

Here $p = \ldots$ and $R \in \mathbb{N}$ is the rank of some thing.

Notation

- $w(T)$ will denote the Gaussian width of $T \subset \mathbb{R}^n$
- $rad(T)$ will denote the Gaussian radius $rad(T) := \sup_{x \in T}\|x\|_2$
- The Orlicz ψ_2-norm of a random variable X denoted by $\|X\|_{\psi_2} := \inf\{t > 0, \mathbb{E}[e^{X^2/t^2}] \leq 2\}$
- $\phi = 1$
- $K = K_{row}$ for independent rows. K is a bound characterising the concentration of quantities related to ϵ
- $p^{1/2}_{eff,\Sigma}$, for independent rows is $\|\Sigma^{1/2}_{\epsilon_{i\cdot}}\tilde{\omega} - \psi\| + \|\epsilon_i\psi - \epsilon_{N:}\|_{L_2}$
- $\Theta^*_s = \{\theta - \tilde{\theta} : \theta \in \Theta, \|\Sigma^{1/2}_{\epsilon_i}(\theta - \tilde{\theta})\| \leq s\}$

- $\Theta^*_{s,r} = \{\delta_\theta := \theta - \tilde{\theta} \in \Theta^*_s : \|A\delta_\theta\| \le r\} = \{\delta_\theta := \theta - \tilde{\theta} \in \Theta^*_s : \|(L'_{::} - Z'_{::} \cdot \delta_\beta)\delta_\omega\| \le r\}$
- $\sigma_{R+1}(X)$ is the $R+1$st singular value of X. $\sigma_1(X), \sigma_2(X), \cdots$ well be the decreasing sequence of singular values of X with $\sigma_k(X) = 0$ for $k \ge rank(X)$
- $d^{1/2}_{\Sigma}(\Theta^*_s) = w_\Sigma(\Theta^*_s)/rad(\Sigma^{1/2}_{\epsilon_{i\cdot}}\Theta^*_s)$

We use

$$A\delta_\omega\delta_\beta = L'_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \tilde{\beta} - Z'_{:::}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta \tag{25}$$

$$A\delta_\lambda\delta_\beta = L_{::}\delta_\lambda - Z_{:::}\delta_\lambda \cdot \tilde{\beta} - Z_{:::}\delta_\lambda \cdot \delta_\beta + (Z'_{:T\cdot} - Z_{:::}\tilde{\lambda}) \cdot \delta_\beta \tag{26}$$

or, in stacked form,

$$\begin{aligned} &A\delta_\theta\delta_\beta = \bar{L}_{::}\delta_\theta - \bar{Z}_{:::}\delta_\theta \cdot \tilde{\beta} - \bar{Z}_{:::}\delta_\theta \cdot \delta_\beta + (\bar{Z}_{:P:} - \bar{Z}_{:::}\delta_\theta) \cdot \delta_\beta \\ &\text{for} \quad \delta_\theta = \begin{pmatrix} \delta_\omega \\ \delta_\lambda \end{pmatrix}, \quad \bar{L}_{::} = \begin{pmatrix} L'_{::} & 0 \\ 0 & L_{::} \end{pmatrix}, \quad \bar{Z}_{:::} = \begin{pmatrix} Z'_{:::} & 0 \\ 0 & Z_{:::} \end{pmatrix}, \quad \bar{Z}_{:P:} \quad = \begin{pmatrix} Z'_{N::} \\ Z_{:T:} \end{pmatrix}. \end{aligned} \tag{27}$$

A.2 Bounding δ_β in terms of r and s.

Idea: if $\omega = \tilde{\omega} + \delta_\omega, \lambda = \tilde{\lambda} + \delta_\lambda \in 27 - lastthing$, then $\hat{\beta} - \beta$ satisfies last thing.

$$\begin{aligned} \mathbb{E}(X_{uw} - X_{xz})^2 &= \mathbb{E}\{g'(A^\perp uv - A^\perp xz)\}^2 \\ &= \|A^\perp uw - A^\perp xz\|^2 \end{aligned} \tag{28}$$

What is $Auw - Axz$ for $w(u,v) = \hat{\beta}(\tilde{\omega} + u, \tilde{\lambda} + v) - \tilde{\beta}$ and $z(x,y) = \hat{\beta}(\tilde{\omega} + x, \tilde{\lambda} + y) - \tilde{\beta}$?

Let's start with this. A from (25) is related to A_ω as defined below (32) like this.

$$\begin{aligned} Auw &= L'_{::}u - A_{\tilde{\omega}+u} \cdot (\tilde{\beta} + w) - A_{\tilde{\omega}} \cdot \tilde{\beta} \\ &= [somethingwithahatmatrix] \end{aligned} \tag{29}$$

Looking at the increment $w = \{\hat{\beta}(\tilde{\omega} + u, \tilde{\lambda} + v) - \tilde{\beta}\} - z = \{\hat{\beta}(\tilde{\omega} + x, \tilde{\lambda} + y) - \tilde{\beta}\}$, here's what we see.

$$\begin{aligned} &Auw(u,v) - Axz(x,y) \\ &= A_{\tilde{\omega}+x}[A'_{\tilde{\omega}+x}A_{\tilde{\omega}+x} + A'_{\tilde{\lambda}+y}A_{\tilde{\lambda}+y}]^{-1}[b'_{\tilde{\omega}+x}A_{\tilde{\omega}+x} + b'_{\tilde{\lambda}+y}A_{\tilde{\lambda}+y}]' \\ &- [\text{same with } x, y \text{ replaced by } u, v] \\ &= H_{\tilde{\omega}+x}\{b_{\tilde{\omega}+x} + \ldots\} - H_{\tilde{\omega}+u}\{b_{\tilde{\omega}+u} + \ldots\} \quad \text{for} \quad H_{\omega,\lambda} = A_\omega[A'_\omega A_\omega + A'_\lambda A_\lambda]^{-1}A_\omega \\ &= \{H_{\tilde{\omega}+x} - H_{\tilde{\omega}+u}\}\{b_{\tilde{\omega}+x} + \ldots\} + H_{\tilde{\omega}+u}[\{b_{\tilde{\omega}+x} + \ldots\} - \{b_{\tilde{\omega}+x} + \ldots\}] \end{aligned} \tag{30}$$

But what we really want to do here is bound this.

$$R_\beta = \max_{v \in P_\beta \Theta^\star_{s,r}} \|A^\perp \Box v\| \tag{31}$$

for A acting on $\delta_\omega, \delta_\beta$ as above in (25), thought of as an operator norm on δ_ω, which we've indicated by replacing it with $\Box$. To do this, we probably want to do more or less the same argument we used below, but applied to the stochastic process $A\square\{\hat{\beta}(\omega,\lambda)-\tilde{\beta}\}$ instead of $\{\hat{\beta}(\omega,\lambda)-\tilde{\beta}\}$ itself. To bound this operator norm, we probably will want to find a sudakov-fernique-type bound on the equivalent expression

$$R_\beta = \max_{\substack{v\in P_\beta\Theta^\star_{s,r}\\ \|t\|\le 1}} A^\perp t v$$

instead of trying to bound this operator norm elementwise or whatever like we were doing below by letting q be a row of Q^{-1}.

Here $A^\perp$ is $P^\perp A$ where $P^\perp$ is a projection onto the orthogonal complement of the R-dimensional subspace where most of Ax goes—a sort of PCA thing.

Letting $\omega=\tilde{\omega}+x$, what's happening in the first term is that we're multiplying b_ω by $A_\omega(A_\omega' A_\omega + A_\lambda A_\lambda')^{-1}A_\omega'$, which blows things up even less than the usual hat matrix $A_\omega(A_\omega' A_\omega)^{-1}A_\omega'$ because the A_λ bit makes the thing we're inverting in the former bigger. This means, I think, that we should be ok: getting good approximation in the sense that $A_\omega = Z\omega - Z_N$ is small is fine because we're multiplying by as many copies of A_ω as we are A_ω^{-1}.

Let us characterize our β as a function of our weights. The idea is that we want to bound the size of the β ball. This can be be bounded by the dimention of the ball (k in our case) times the radius of the ball. What we want to do is charachtrize the radius as a function of our ω and λ.

$$\begin{aligned}\hat{\beta}(\omega,\lambda) &= \operatorname*{argmin}_\beta \|(Y_{::}'-Z_{::\cdot}'\cdot\beta)\omega-(Y_{N:}'-Z_{N:\cdot'}\cdot\beta)\|^2+\|(Y_{::}-Z_{::\cdot}\cdot\beta)\lambda-(Y_{:T}-Z_{N:\cdot}\cdot\beta)\|^2\\ &= \operatorname*{argmin}_\beta \|\underbrace{(Y_{::}'\omega-Y_{N:}')}_{b_\omega}-\underbrace{(Z_{::\cdot}'\omega-Z_{N:\cdot}')}_{A_\omega}\cdot\beta\|^2+\|\underbrace{(Y_{::}\lambda-Y_{:T})}_{b_\lambda}-\underbrace{(Z_{::\cdot}\lambda-Z_{N:\cdot})}_{A_\lambda}\cdot\beta\|^2\\ &= \operatorname*{argmin}_\beta \langle b_\omega,b_\omega\rangle+\langle b_\lambda,b_\lambda\rangle-2\langle b_\omega,A_\omega\cdot\beta\rangle-2\langle b_\lambda,A_\lambda\cdot\beta\rangle+\langle A_\omega\cdot\beta,A_\omega\cdot\beta\rangle+\langle A_\lambda\cdot\beta,A_\lambda\cdot\beta\rangle\\ &= \operatorname*{argmin}_\beta \langle b_\omega,b_\omega\rangle+\langle b_\lambda,b_\lambda\rangle-2[b_\omega' A_\omega+b_\lambda' A_\lambda]\cdot\beta+\beta'[A_\omega' A_\omega+A_\lambda' A_\lambda]\cdot\beta\end{aligned} \tag{32}$$

Here notation wise A_ω was a $p\times 1\times k$ array, but we contracted out the dimension that is just 1, and work with the resulting $p\times k$ matrix. Therefore above and below A_ω is $p\times k$ and A_λ is $n\times k$

We take the derivative wrt to β and solve for first order condition

$$\begin{aligned}-2[b_\omega' A_\omega+b_\lambda' A_\lambda]+2[A_\omega' A_\omega+A_\lambda' A_\lambda]\cdot\beta=0&\\ \beta=[A_\omega' A_\omega+A_\lambda' A_\lambda]^{-1}[b_\omega' A_\omega+b_\lambda' A_\lambda]'&\end{aligned} \tag{33}$$

A.2.1 Bound differences in β

1. define $\tilde{\beta}$ as a function of ω,λ more clearly.
2. clarify what we're bounding. Is it $\hat{\beta}(\omega,\lambda)-\tilde{\beta}(\tilde{\omega},\tilde{\lambda})$ for $(\delta_\omega,\delta_\lambda)\in P^{\omega,\lambda}\Theta^\star_{s,r}$?

So our β with noise is

$$\hat{\beta}(\omega,\lambda)=[A_\omega' A_\omega+A_\lambda' A_\lambda]^{-1}[b_\omega' A_\omega+b_\lambda' A_\lambda]' \tag{34}$$

So our β without noise at the oracle weight values is

$$\tilde{\beta}(\tilde{\omega}, \tilde{\lambda}) = [A'_{\tilde{\omega}} A_{\tilde{\omega}} + A'_{\tilde{\lambda}} A_{\tilde{\lambda}}]^{-1} [(L'_{::}\tilde{\omega} - L'_{N:})' A_{\tilde{\omega}} + (L_{::}\tilde{\lambda} - L_{:T})' A_{\lambda}]' \tag{35}$$

We bound $\|\hat{\beta}(\omega, \lambda) - \tilde{\beta}(\tilde{\omega}, \tilde{\lambda}\|_2$ here as the bound is helpful in our proofs.

$$\begin{aligned} \hat{\beta}(\omega, \lambda) - \tilde{\beta}(\tilde{\omega}, \tilde{\lambda}) &= \hat{\beta}(\omega, \lambda) \pm [A'_{\tilde{\omega}} A_{\tilde{\omega}} + A'_{\tilde{\lambda}} A_{\tilde{\lambda}}]^{-1} [(L'_{::}\omega - L'_{N:})' A_{\omega} + (L_{::}\lambda - L_{:T})' A_{\lambda}]' - \tilde{\beta}(\tilde{\omega}, \tilde{\lambda}) \\ 1) &= [A'_{\omega} A_{\omega} + A'_{\lambda} A_{\lambda}]^{-1} [(\epsilon'_{::}\omega - \epsilon'_{N:})' A_{\omega} + (\epsilon_{::}\lambda - \epsilon_{:T})' A_{\lambda}]' \\ 2) &+ \left[[A'_{\omega} A_{\omega} + A'_{\lambda} A_{\lambda}]^{-1} - [A'_{\tilde{\omega}} A_{\tilde{\omega}} + A'_{\tilde{\lambda}} A_{\tilde{\lambda}}]^{-1} \right] [(L'_{::}\tilde{\omega} - L'_{N:})' A_{\tilde{\omega}} + (L_{::}\tilde{\lambda} - L_{:T})' A_{\lambda}]' \\ 3) &+ [A'_{\tilde{\omega}} A_{\tilde{\omega}} + A'_{\tilde{\lambda}} A_{\tilde{\lambda}}]^{-1} [L'_{::}(\delta_{\omega}) A_{\omega} - (L'_{::}\tilde{\omega} - L'_{N:}) Z'_{::}\delta_{\omega} + L_{::}(\delta_{\lambda}) A_{\lambda} - (L'_{::}\tilde{\lambda} - L'_{N:}) Z_{::}\delta_{\lambda}] \end{aligned} \tag{36}$$

In particular, we'll bound the two-norm of this difference.

We call $Q(\omega, \lambda) = [A'_{\omega} A_{\omega} + A'_{\lambda} A_{\lambda}]$. And $F_1(\omega, \lambda) = [(\epsilon'_{::}\omega - \epsilon'_{N:})' A_{\omega} + (\epsilon_{::}\lambda - \epsilon_{:T})' A_{\lambda}]'$ and $f_2(\omega, \lambda) = [(L'_{::}\omega - L'_{N:})' A_{\omega} + (L_{::}\lambda - L_{:T})' A'_{\lambda}]'$. In these terms, $2 = [Q(\omega, \lambda)^{-1} - Q(\tilde{\omega}, \tilde{\lambda})^{-1}][f_2(\tilde{\omega}, \tilde{\lambda})]$.

Term 1

$$\text{Term 1} = Q(\omega, \lambda)^{-1} F_1(\omega, \lambda) \tag{37}$$

Each component of this is a subgaussian process and we want to bound the max of its two norm over $(\omega, \lambda) \in (\tilde{\omega}, \tilde{\lambda}) + P^{\omega,\lambda}\Theta^{\star}_{r,s}$. We'll do that by taking the two-norm of bounds on its components. These have the form $q(\omega, \lambda)^T F_1(\omega, \lambda)$ where $q(\omega, \lambda)$ is a row of $Q(\omega, \lambda)^{-1}$.

Via Talagrand's comparison inequality, this is bounded by a constant times the max of a corresponding gaussian process, and we can use Sudakov-Fernique to bound that max by the max of any gaussian process with increments that have larger variance. We need to find such a gaussian processes and bound its max. To do this, let's start out by calculating and bounding the increments of Term 1. We'll think of u, v and x, y being ω, λ pairs—u and x are instances of ω and v and y instances of λ.

We'll start by writing our increment and decomposing it into two parts—Like we decomposed Term 1 itself into two parts, one in which F_1 changes and another in which Q does.

$$\begin{aligned} \text{Term 1 increment} &= q(u, v) F_1(u, v) - q(x, y) F_1(x, y) \\ &= q(u, v)(F_1(u, v) - F_1(x, y)) + (q(u, v) - q(x, y)) F_1(x, y) \end{aligned} \tag{38}$$

so, because $(a + b)^2 \leq 2a^2 + 2b^2$ generally,

$$\begin{aligned} &[\text{Term 1 increment}]^2 \\ &= [\{q(u, v)(F_1(u, v) - F_1(x, y)) + (q(u, v) - q(x, y)) F_1(x, y)\}]^2 \\ &\leq 2\{q(u, \underset{1.1}{v)(F_1(u, v)} - F_1(x, y))\}^2 + 2\{(q(u, v) - q(\underset{1.2}{x, y}))F_1(x, y)\}^2. \end{aligned} \tag{39}$$

The variance of our increment is the expected value of this quantity, so we will bound the expected values of each term in our bound. We will do the calculations in subsections below. The result is the following.

$$
\begin{aligned}
\mathrm{Var}[\text{Term 1 increment}] \leq{}& \sigma(v)^2\|Z_{:::}(v-y)\cdot q\|^2 + \|\Sigma_{\varepsilon_{i.}}^{1/2}(v-y)\|^2\|(Z_{:::}v - Z_{:T:})\cdot q\|^2 \\
&+ 2\big\{(x'Z_{:::} - Z_{N::})\cdot v\big\}\big(\|x\|^2\Sigma_{\varepsilon_{i.}} + \Sigma_\nu\big)\big\{(x'Z_{:::} - Z_{N::})\cdot v\big\} \\
&+ 2\{(Z_{:::}y - Z_{:T:})\cdot v\}\Big(\{\|\Sigma_{\varepsilon_{i.}}^{1/2}(y-\psi)\|^2 + \sigma_\nu^2\}I\Big)\{(Z_{:::}y - Z_{:T:})\cdot v\}
\end{aligned}
\tag{40}
$$

Where $v = q(u,v) - q(x,y)$ and $\sigma(v)^2 = (v-\psi)'\Sigma_{\varepsilon_{i.}}(v-\psi) + \sigma_\nu^2$.

$$
\begin{aligned}
\sigma(v)^2\|Z_{:::}(v-y)\cdot q\|^2 &= \|A()(v-y)\|^2 \quad \text{for} \quad A(\cdot) = \sigma(v)Z_{:::}\cdot q \\
&\leq \underset{\cdots}{\max}\|A(\cdot)\|^2\|v-y\| = \mathbb{E}\big\{g'(v-y)\big\}^2 \quad \text{for} \quad g \sim N(0, \underset{\cdots}{\max}\|A(\cdot)\|^2 I)
\end{aligned}
\tag{41}
$$

Term 1.1 Much like we did above, we'll decompose Term 1.1 into one part in which $\omega = u\&x$ changes and another in which $\lambda = v\&y$ does.

$$
\begin{aligned}
\text{Term 1.1.} &= q(u,v)\{F_1(u,v) \pm F_1(u,y) - F_1(x,y)\} \\
&= q(u,v)\{F_1(u,v) - F_1(u,y)\} + q(u,v)\{F_1(u,y) - F_1(x,y)\}
\end{aligned}
\tag{42}
$$

and again because $(a+b)^2 \leq 2a^2 + 2b^2$,

$$
\begin{aligned}
\text{Term } 1.1^2 &\leq 2[q(u,v)\{F_1(u,v) - F_1(u,y)\}]^2 + 2[q(u,v)\{F_1(u,y) - F_1(x,y)\}]^2 \\
&= 2\{F_1(u,v) - F_1(u,y)\}^T M\{F_1(u,v) - F_1(u,y)\} \\
&+ 2\{F_1(u,y) - F_1(x,y)\}^T M\{F_1(u,y) - F_1(x,y)\} \quad \text{for} \quad M = q(u,v)q(u,v)^T.
\end{aligned}
\tag{43}
$$

We plug in the definition of F_1

$$
F_1(u,y) = (\epsilon'_{::}u - \epsilon'_{N:})(Z_{::}u - Z_{N:}) - (\epsilon_{::}y - \epsilon_{:T})(Z_{::}y - Z_{:T})
\tag{44}
$$

$$
\begin{aligned}
F_1(u,v) - F_1(u,y) &= (\epsilon'_{::}u - \epsilon'_{N:})(Z_{::}u - Z_{N:}) - (\epsilon_{::}v - \epsilon_{:T})(Z_{::}v - Z_{:T}) \\
&- [(\epsilon'_{::}u - \epsilon'_{N:})(Z_{::}u - Z_{N:}) - (\epsilon_{::}y - \epsilon_{:T})(Z_{::}y - Z_{:T})] \\
&= -(\epsilon_{::}v - \epsilon_{:T})(Z_{::}v - Z_{:T}) + (\epsilon_{::}y - \epsilon_{:T})(Z_{::}y - Z_{:T}) \\
&= -(\epsilon_{::}v - \epsilon_{:T})[(Z_{::}v - Z_{:T}) - (Z_{::}y - Z_{:T})] + [(\epsilon_{::}v - \epsilon_{:T}) - (\epsilon_{::}y - \epsilon_{:T})](Z_{::}v - Z_{:T}) \\
&= -(\epsilon_{::}v - \epsilon_{:T})[Z_{::}(v-y)] + [\epsilon_{::}(v-y)](Z_{::}v - Z_{:T})
\end{aligned}
\tag{45}
$$

Temporarily calling these two terms a and b, we plug this into 114 and apply the bound $(x+y)^T(x+y) \leq 2\|x\|^2 + 2\|y\|^2$, yielding this bound.

$$
\begin{aligned}
&\{F_1(u,v) - F_1(u,y)\}'qq'\{F_1(u,v) - F_1(u,y)\} \\
&= \{a+b\}'qq'\{a+b\} \\
&\leq 2a'qq'a + 2b'qq'b \\
&= 2(a'q)^2 + 2(b'q)^2 \\
&= \sigma(v)^2\|Z_{::}(v-y)\cdot q\|^2 + \|\Sigma_{\varepsilon_{i.}}^{1/2}(v-y)\|^2\|(Z_{::}v - Z_{:T})\cdot q\|^2
\end{aligned}
\tag{46}
$$

What remains is to calculate the expected value of these two terms. Below we will find that

$$2(a'q)^2 + 2(b'q)^2 = \sigma(v)^2 \|Z_{::}(v-y) \cdot q\|^2 + \|\Sigma_{\varepsilon_{i.}}^{1/2}(v-y)\|^2 \|(Z_{::}v - Z_{:T}) \cdot q\|^2 \tag{47}$$

The 'a' term

$$E[a'qq'a] = \mathbb{E}(q'a)^2 = \sigma^2(v)\|Z_{::}(v-y)\cdot q\|^2 \quad \text{for} \quad \sigma^2(v) = (v-\psi)'\Sigma_{\varepsilon_{i.}}(v-\psi) + \sigma_\nu^2 \tag{48}$$

using, in the last step, the characterization of $q'a$ that follows.

$$\begin{aligned}
a_i &= \{Z_{::i}(v-y)\}'\{\epsilon_{::}v - \epsilon_{:T}\} \\
&= \{Z_{::i}(v-y)\}'\Sigma^{1/2}(v)g \quad \text{where} \quad g \sim N(0,I) \quad \text{and} \\
\Sigma(v) &= \mathbb{E}\{\epsilon_{::}v - \epsilon_{:T}\}\{\epsilon_{::}v - \epsilon_{:T}\}' \\
&= \mathbb{E}[\{\epsilon_{::}(v-\psi)\} + \{\epsilon_{::}\psi - \epsilon_{:T}\}][\{\epsilon_{::}(v-\psi)\} + \{\epsilon_{::}\psi - \epsilon_{:T}\}]' \\
&= \mathbb{E}\{\epsilon_{::}(v-\psi)\}\{\epsilon_{::}(v-\psi)\}' + \mathbb{E}\{\epsilon_{::}\psi - \epsilon_{:T}\}\{\epsilon_{::}\psi - \epsilon_{:T}\}' \\
&= \mathbb{E}\epsilon_{::}(v-\psi)(v-\psi)'\epsilon_{::}' + +\mathbb{E}\{\epsilon_{::}\psi - \epsilon_{:T}\}\{\epsilon_{::}\psi - \epsilon_{:T}\}' \\
&= \sigma^2(v)I \quad \text{for} \\
\sigma^2(v) &= \mathbb{E}\epsilon_{i:}(v-\psi)(v-\psi)'\epsilon_{i:}' + \mathbb{E}\{\epsilon_{i:}\psi - \epsilon_{iT}\}\{\epsilon_{i:}\psi - \epsilon_{iT}\} \\
&= (v-\psi)'\underbrace{\mathbb{E}\epsilon_{i:}\epsilon_{i:}'}_{\Sigma_{\varepsilon_{i.}}}(v-\psi) + \sigma_\nu^2 \quad \text{transposing the scalar} \quad \epsilon_{i:}'(v-\psi).
\end{aligned} \tag{49}$$

Consequently, $q'a = \sigma(v)g'Z_{::}(v-y)\cdot q$ and

$$\begin{aligned}
\mathbb{E}(q'a)^2 = &= \sigma(v)^2\mathbb{E}\big\{g'Z_{::}(v-y)\cdot q\big\}'\big\{g'Z_{::}(v-y)\cdot q\big\} \\
&= \sigma(v)^2\{Z_{::}(v-y)\cdot q\}'\mathbb{E}gg'\{Z_{::}(v-y)\cdot q\} \\
&= \sigma(v)^2\{Z_{::}(v-y)\cdot q\}'\{Z_{::}(v-y)\cdot q\} \\
&= \sigma(v)^2\|Z_{::}(v-y)\cdot q\|^2
\end{aligned} \tag{50}$$

The 'b' term.

$$E[b'qq'b] = \mathbb{E}(q'b)^2 = \|\Sigma_{\varepsilon_{i.}}^{1/2}(v-y)\|^2\|(Z_{::}v - Z_{:T})\cdot q\|^2 \tag{51}$$

using, in the last step, the characterization of $q'b$ that follows.

$$\begin{aligned}
b_i &= [\epsilon_{::}(v-y)]'(Z_{::i}v - Z_{:Ti}) \\
&= [G\Sigma_{\varepsilon_{i.}}^{1/2}(v-y)]'(Z_{::i}v - Z_{:Ti}) \quad \text{for} \quad G_{ij} \sim N(0,1) \\
&= (v-y)'\Sigma_{\varepsilon_{i.}}^{1/2}G(Z_{::i}v - Z_{:Ti}) \quad \text{where } Z_{..i} \text{ is the part of } Z_{..} \text{ that hits } \beta_i \quad .
\end{aligned} \tag{52}$$

Consequently, $q'b = (v-y)'\Sigma_{\varepsilon_{i.}}^{1/2}G(Z_{::}v - Z_{:T})\cdot q$ and, using the identity $\mathbb{E}(x'Gy)^2 = \|x\|^2\|y\|^2$ gives us the claimed characterization.

Term 1.2

$$2\{(q(u,v) - \underset{1.2}{q(x,y)})F_1(x,y)\}^2. \tag{53}$$

$$F_1(x,y) = (x'\epsilon_{::} - \epsilon_{N:})(x'Z_{::} - Z_{N:})' - (\epsilon_{::}y - \epsilon_{:T})'(Z_{::}y - Z_{:T}) \tag{54}$$

This has the form $\mathbb{E}(v'\Sigma^{1/2}g)^2$ where $v = q(u,v) - q(x,y)$, $g \sim N(0,1)$ and $\Sigma^{1/2}$ is the covariance matrix of $F_1(x,y)$, i.e., it is $v'\Sigma v$. What is Σ? It's the expectation of an outer-product of this with itself in the 3rd (β) dimension, so to get an element we multiply a copy of this with $Z_{::} \to Z_{::i}$ and one with $Z_{::} \to Z_{::j}$.

$$\begin{aligned}\Sigma_{ij} = \mathbb{E}\big\{&(x'\epsilon_{::} - \epsilon_{N:})(x'Z_{::i} - Z_{N:i})' - (\epsilon_{::}y - \epsilon_{:T})'(Z_{::i}y - Z_{:Ti})\big\}\times \\ &\big\{(x'\epsilon_{::} - \epsilon_{N:})(x'Z_{::j} - Z_{N:j})' - (\epsilon_{::}y - \epsilon_{:T})'(Z_{::j}y - Z_{:Tj})\big\} \\ &= 1.2.1 + 2 \times 1.2.2 + 1.2.3 \quad \text{for 1.2.x as below}\end{aligned} \tag{55}$$

where

$$\begin{aligned}1.2.1 \ &= \mathbb{E}(x'Z_i - Z_{N::i})(x'\epsilon_{::} - \epsilon_{N:})'(x'\epsilon_{::} - \epsilon_{N:})(x'Z_{::j} - Z_{N:j})' \\ &= (x'Z_i - Z_{N::i})\big(\|x\|^2\Sigma_{\varepsilon_{i.}} + \Sigma_\nu\big)(x'Z_{::j} - Z_{N:j})' \\ 1.2.2 \ &= \mathbb{E}(x'Z_{::i} - Z_{N:i})(x'\epsilon_{::} - \epsilon_N)'(\epsilon_{::}y - \epsilon_T)'(Z_{::j}y - Z_{:Tj}) \\ &= (x'Z_{::i} - Z_{N:i})\big(\Sigma_{\varepsilon_{i.}}(y-\psi)x'\big)(Z_{::j}y - Z_{:Tj}) \\ 1.2.3 \ &= \mathbb{E}(Z_{::i}y - Z_{:Ti})'(\epsilon_{::}y - \epsilon_{:T})(\epsilon_{::}y - \epsilon_{:T})'(Z_{::j}y - Z_{:Tj}) \\ &= (Z_{::i}y - Z_{:Ti})'\Big(\{\|\Sigma_{\varepsilon_{i.}}^{1/2}(y-\psi)\|^2 + \sigma_\nu^2\}I\Big)(Z_{::j}y - Z_{:Tj})\end{aligned} \tag{56}$$

Therefore we have that

$$\begin{aligned}v'\Sigma v &= \big\{(x'Z_{:::} - Z_{N::})\cdot v\big\}\big(\|x\|^2\Sigma_{\varepsilon_{i.}} + \Sigma_\nu\big)\big\{(x'Z_{:::} - Z_{N::})\cdot v\big\} \\ &\quad + \big\{(x'Z_{:::} - Z_{N::})\cdot v\big\}\big(\Sigma_{\varepsilon_{i.}}(y-\psi)x'\big)\{(Z_{:::}y - Z_{:T:})\cdot v\} \\ &\quad + \{(Z_{:::}y - Z_{:T:})\cdot v\}\Big(\{\|\Sigma_{\varepsilon_{i.}}^{1/2}(y-\psi)\|^2 + \sigma_\nu^2\}I\Big)\{(Z_{:::}y - Z_{:T:})\cdot v\} \\ &\le 2\big\{(x'Z_{:::} - Z_{N::})\cdot v\big\}\big(\|x\|^2\Sigma_{\varepsilon_{i.}} + \Sigma_\nu\big)\big\{(x'Z_{:::} - Z_{N::})\cdot v\big\} \\ &\quad + 2\{(Z_{:::}y - Z_{:T:})\cdot v\}\Big(\{\|\Sigma_{\varepsilon_{i.}}^{1/2}(y-\psi)\|^2 + \sigma_\nu^2\}I\Big)\{(Z_{:::}y - Z_{:T:})\cdot v\}\end{aligned} \tag{57}$$

Steps explaining the three terms in Σ are given below.

Here each term has the form $a'Sb$ where a and b are deterministic vectors and S is the expected value of an outer product of noise vectors. We've simply substituted the appropriate matrix S here rather than showing the calculation. We'll do that now for each of our three terms.

To explain where the expressions in Equation (56) come from, we write out the steps for each of the three parts.

In 1.2.1, we have ...

$$1.2.1 \ = \mathbb{E}(x'Z_i - Z_{N::i})(x'\epsilon_{::} - \epsilon_{N:})'(x'\epsilon_{::} - \epsilon_{N:})(x'Z_{::j} - Z_{N:j})' \tag{58}$$

Focusing on the middle of this expression we have

$$\begin{aligned}\mathbb{E}(x'\epsilon_{::} - \epsilon_{N:})'(x'\epsilon_{::} - \epsilon_{N:}) &= \mathbb{E}\epsilon_{::}'xx'\epsilon_{::} + 2\mathbb{E}\epsilon_{::}'x\epsilon_{:N}' + \mathbb{E}\epsilon_{:N}\epsilon_{:N}' \\ &= \|x\|^2\Sigma_{\varepsilon_{i.}} + 0 + \Sigma_\nu.\end{aligned} \tag{59}$$

Here, in the last step, we've used the following calculation of an element of $\mathbb{E}\epsilon'_{::}xx'\epsilon_{::}$.

$$\begin{aligned}
\mathbb{E}\{\epsilon'_{::}xx'\epsilon_{::}\}_{ij} &= \mathbb{E}\epsilon'_{:i}xx'\epsilon_{:j} \\
&= \sum_{k\ell} x_k x_\ell \mathbb{E}\epsilon_{ki}\epsilon_{\ell j} \\
&= \sum_{k\ell} x_k x_\ell \quad \begin{cases} \mathbb{E}\epsilon_{ki}\epsilon_{kj} & \text{if} \quad k=\ell \\ 0 & \text{otherwise} \end{cases} \\
&= \sum_{k} x_k^2 \mathbb{E}\epsilon_{ki}\epsilon_{kj} \\
&= \|x\|^2 (\Sigma_{\varepsilon_{i\cdot}})_{ij}
\end{aligned} \tag{60}$$

Now for the steps of term 1.2.2.

$$1.2.2 \;= \mathbb{E}(x'Z_{::i} - Z_{N:i})(x'\epsilon_{::} - \epsilon_N)'(\epsilon_{::}y - \epsilon_T)'(Z_{::j}y - Z_{:Tj}) \tag{61}$$

Focusing in the middle part of this term.

$$\begin{aligned}
\mathbb{E}(x'\epsilon_{::} - \epsilon_N)'(\epsilon_{::}y - \epsilon_T)' &= \mathbb{E}(x'\epsilon_{::})'[\epsilon_{::}(y-\psi)]' + \mathbb{E}(x'\epsilon_{::})'[(\epsilon_{::}\psi - \epsilon_{:T})]' \\
&= \mathbb{E}(x'\epsilon_{::})'_i[\epsilon_{::}(y-\psi)]'_j + \mathbb{E}(\epsilon'_{::}x)_i[(\epsilon_{::}\psi - \epsilon_{:T})]'_j
\end{aligned} \tag{62}$$

Now for the steps of term 1.2.3.

$$1.2.3 \;= \mathbb{E}(Z_{::i}y - Z_{:Ti})'(\epsilon_{::}y - \epsilon_{:T})(\epsilon_{::}y - \epsilon_{:T})'(Z_{::j}y - Z_{:Tj}) \tag{63}$$

Focusing in on the middle part of this term. We'll start by adding and subtracting $\epsilon_{::}\psi$, the orthogonal projection of $\epsilon_{:T}$ onto $\epsilon_{::}$, to rewrite $\epsilon_{::}\psi - \epsilon_{:T}$ as a sum of uncorrelated terms: $\epsilon_{::}y - \epsilon_{:T} = \epsilon_{::}(y-\psi) + (\epsilon_{::}\psi - \epsilon_{:T}$ where $\mathbb{E}(\epsilon_{::}\psi - \epsilon_{:T})'\epsilon_{::} = 0$.

$$\begin{aligned}
(\epsilon_{::}y - \epsilon_{:T})(\epsilon_{::}y - \epsilon_{:T})' &= \{\epsilon_{::}(y-\psi) + (\epsilon_{::}\psi - \epsilon_{:T})\}\{\epsilon_{::}(y-\psi) + (\epsilon_{::}\psi - \epsilon_{:T})\}' \\
&= \mathbb{E}\epsilon_{::}(y-\psi)(y-\psi)'\epsilon' + \mathbb{E}(\epsilon_i\psi - \epsilon_{:T})(\epsilon_j\psi - \epsilon_T)' \\
&= I\mathbb{E}\{\epsilon_{i\cdot}(y-\psi)\}^2 + I\mathbb{E}(\epsilon_{i\cdot}\psi = \epsilon_{iT})^2 \\
&= I(y-\psi)'\Sigma_{\varepsilon_{i\cdot}}(y-\psi) + I\sigma_\nu^2 \\
&= I\{\|\Sigma_{\varepsilon_{i\cdot}}^{1/2}(y-\psi)\|^2 + \sigma_\nu^2\}
\end{aligned} \tag{64}$$

Term 2 $\quad f(X) = X^{-1}$ is locally $\|X^{-1}\|$-Lipschitz, i.e., $\|X^{-1} - Y^{-1}\| \leq \|X^{-1}\|\|X - Y\|$, so

$$\begin{aligned}
\|\text{Term 2}\| &:= \|[Q(\omega,\lambda)^{-1} - Q(\tilde{\omega},\tilde{\lambda})^{-1}] \times f_2(\tilde{\omega},\tilde{\lambda})\| \\
&\leq \|Q(\tilde{\omega},\tilde{\lambda})^{-1}\|\|Q(\omega,\lambda) - Q(\tilde{\omega},\tilde{\lambda})\| \times \|f_2(\tilde{\omega},\tilde{\lambda})\| \\
&\leq \|Q(\tilde{\omega},\tilde{\lambda})^{-1}\| \times \|f_2(\tilde{\omega},\tilde{\lambda})\| \times [\text{some function of } s, r] \quad \text{for all} \quad \delta_\omega, \delta_\lambda \in \Theta^\star_{s,r}
\end{aligned} \tag{65}$$

the some function is going to be the maximum disstance in Q .

Term 3

$$\text{Term 3 } = Q(\tilde{\omega}, \tilde{\lambda}) f_3(\omega, \lambda) \tag{66}$$

A.3 Characterizing our Radii r and s

A.4 Choosing s: Bounding $\|\Sigma^{1/2}\delta_\omega\|$ and $\|\Sigma^{1/2}\delta_\lambda\|$

Writing our version of equation 36.

$$\begin{aligned} \tilde{L}(\delta) \geq \kappa_\omega \|A\delta_\omega\delta_\beta\|^2 + \kappa_\omega n\eta^2 \|\Sigma^{1/2}\delta_\omega\|^2 - c\alpha \\ - q(\delta) cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\} \end{aligned} \tag{67}$$

$$q(\delta) = \max(\|A\delta_\omega\delta_\beta\|/r_\omega, 1) \tag{68}$$

$$\alpha = Kv\,\mathrm{RMSE}_\omega\,\mathrm{w}_\Sigma(\Theta^\star_{s_\omega}) + vK^2(n/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_s) \tag{69}$$

Now we are going to hold the s fixed

$$\begin{aligned} \tilde{L}(\delta) \geq \kappa_\omega n\eta^2 s_\omega^2 - c\alpha \\ + \kappa_\omega \|A\delta_\omega\delta_\beta\|^2 - q(\delta) cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\} \end{aligned} \tag{70}$$

What we'll do now is find a lower bound that holds for all possible values of $x = \|A\delta_\omega\delta_\beta\|$ by approximately minimizing over x. As a function of x, this second line is

$$\begin{aligned} q_1(x) &= \kappa_\omega x^2 - cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) - B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\} && \text{if } x \leq r, \\ q_2(x) &= \kappa_\omega x^2 - \frac{x}{r_\omega}\left[cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega + cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\right] && \text{if } x \geq r. \\ &= \end{aligned} \tag{71}$$

Idea. What we'll do now is choose r—which is arbitrary—to give us a relatively simple expression for this minimum. In particular, we will choose r so that the minimum is $q_1(0)$. Having done this, we can lower bound the right side of (70) by simply substituting $q_1(0)$ for its second line. This is nice because $q_1(0)$ winds up looking very simple.

Because $q_1(x)$ is increasing, its minimum on its domain is $q_1(0)$. Furthermore, so long as the leading coefficient of q_2 is positive, it has a unique global minimum, and if that minimum occurs at $x \leq r$, its minimum on its domain is $q_2(r)$. And because $q_2(r) = r^2 + q_1(0) > q_1(0)$, when the global minimum of q_2 occurs at $x \leq r$, we can lower bound the second line in (**??**) by $q_1(0)$. We choose r to make this happen: as the minimum of the polynomial $a_2x^2 - a_1x - a_0$ occurs at $x = a_1/2a_2$, it requires that

$$\frac{\left[cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega + cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\right]/r_\omega}{2\kappa_\omega} \leq r_\omega \tag{72}$$

$$\left[cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega + cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\right] \leq 2\kappa_\omega r_\omega^2 \tag{73}$$

$$-2\kappa_\omega r_\omega^2 + \left[cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega + cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\right] \leq 0 \tag{74}$$

$$\frac{cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\} + \sqrt{[cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}]^2 + 4(2\kappa_\omega)(cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda))}}{2\kappa_\omega} = r_\omega \tag{75}$$

$$\frac{cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\} + \sqrt{4(2\kappa_\omega)(cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda))}}{2\kappa_\omega} \geq r_\omega \tag{76}$$

which holds for this r This choice makes the leading coefficient of q_2 positive as required.

$$q_1(0) = -cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) - B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\} \tag{77}$$

$$\begin{aligned}\tilde{L}(\delta) \geq\ & \kappa_\omega n\eta^2 s_\omega^2 - c\alpha \\ & - cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}r_\omega - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) - B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\end{aligned} \tag{78}$$

Let $a = cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}$
and let $b^2 = cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}$
So our second line is

$$\begin{aligned}
&= -(a \times r_\omega - b^2) \\
&= -(a \times (a + \sqrt{\kappa}b)/2\kappa + b^2) \\
&= -(a^2/2\kappa + ab\sqrt{\kappa}/2\kappa + b^2) \\
&= -(a^2/2\kappa + ab/\sqrt{\kappa} + b^2) \\
&= -(a^2/2\kappa + ab\sqrt{\kappa}/2\kappa + b^2) \\
&= -(a/\sqrt{\kappa} + b)^2 \\
&\leq -[(a/\sqrt{\kappa})^2 + b^2] \\
&= -[\frac{cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}}{\sqrt{\kappa_\omega}} - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}]
\end{aligned} \tag{79}$$

Now we look at omega and lambda together to solve for s.

$$
\begin{aligned}
\tilde{L}(\delta) \geq{}& \kappa_\omega n\eta^2 s_\omega^2 + \kappa_\lambda p\eta^2 s_\lambda^2 - c\alpha_{\omega+\lambda} \\
&- \left[\frac{cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}}{\sqrt{\kappa_\omega}} - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta_{s_\omega}^\star) + B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\right] \\
&- \left[\frac{cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}}{\sqrt{\kappa_\lambda}} - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\lambda 1}\,\mathrm{w}(\Theta_{s_\lambda}^\star) + B_{\lambda 2}\sqrt{k}(s_\omega + s_\lambda)\}\right]
\end{aligned}
\tag{80}
$$

Now we put this is hashtag standard form

$\alpha'_\omega = \frac{cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}\}}{\sqrt{\kappa_\omega}} + B_{\omega 1}\,\mathrm{w}(\Theta_{s_\omega}^\star)$

$$
\begin{aligned}
\tilde{L}(\delta) \geq{}& \kappa_\omega n\eta^2 s_\omega^2 + \kappa_\lambda p\eta^2 s_\lambda^2 - c\alpha_{\omega+\lambda} - \alpha'_{\omega+\lambda} \\
&- \left[cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 2}\sqrt{k}(s_\omega + s_\lambda)\}\right] \\
&- \left[cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\lambda 2}\sqrt{k}(s_\omega + s_\lambda)\}\right] \\
&= \kappa_\omega n\eta^2 s_\omega^2 + \kappa_\lambda p\eta^2 s_\lambda^2 - a(s_\lambda + s_\omega) - \alpha - \alpha' \quad \text{for} \quad a = cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 2} + B_{\lambda 2}\}
\end{aligned}
\tag{81}
$$

A.5 Choosing r: Bounding $\|A\delta_\omega\delta_\beta\|$ and $\|A\delta_\lambda\delta_\beta\|$

Looking at (24), we have one bivariate quadradic in two variables. $\|\Sigma^{1/2}\delta_\lambda\|$ and $\|\Sigma^{1/2}\delta_\omega\|$. We want to find the minimum of the right hand side minimizing $\|\Sigma^{1/2}\delta_\lambda\|$ and $\|\Sigma^{1/2}\delta_\omega\|$, and so we will take the partial derivatives and solve for the respective r values.

We start with $\|\Sigma^{1/2}\delta_\lambda\|$.

$\frac{\partial}{\partial\|\Sigma^{1/2}\delta_\lambda\|} = 2\kappa_\omega n\eta$

The min is at zero because we dont have linear term.

So in the situation that both $\|\Sigma^{1/2}\delta_\lambda\|$ and $\|\Sigma^{1/2}\delta_\omega\|$ are zero. we have that the minimum of our problem is at

$$
\tilde{L}(\delta) \geq \kappa_\omega x^2 + \kappa_\lambda y^2 + \tag{82a}
$$

$$
- 2|\max(x/r_\omega, 1)cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta_{s_\omega}^\star) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \tag{82b}
$$

$$
- 2|\max(y/r_\lambda, 1)cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta_{s_\lambda}^\star) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \tag{82c}
$$

$$
- c_{\omega,\lambda} \tag{82d}
$$

Where

$$
\begin{aligned}
c_{\omega,\lambda} ={}& 2|cKv\,\mathrm{RMSE}_\omega\,\mathrm{w}_\Sigma(\Theta_{s_\omega}^\star)| \quad \text{for} \quad \mathrm{RMSE}_\omega := \|(L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega}\cdot\tilde{\beta} - L'_{:N} + Z'_{N:\cdot}\cdot\tilde{\beta})\| \\
&+ 2|cKv\,\mathrm{RMSE}_\lambda\,\mathrm{w}_\Sigma(\Theta_{s_\lambda}^\star)| \quad \text{for} \quad \mathrm{RMSE}_\lambda := \|(L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta})\| \\
&+ 2|cvK^2(n/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta_s^\star)| \\
&+ 2|cvK^2(p/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta_\lambda^\star)|
\end{aligned}
\tag{83}
$$

Cases

1. x is big, and we are trying to find worst case y
   (a) $x > r_\omega$ so we can solve for the r_ω as discussed, putting everything else in constant
   (b) Once we have this, we can play the advesarial case with y
      - two cases for the max term
      - what value maximizes our lower bound
2. then the other way around, but the argument is the same

Case 1 We are going to put everything that is not a function of x into the constant term. We are also in the case that x is big, so $x > r_\omega$, which takes care of the max term for ω

$$\tilde{L}(\delta) \geq \kappa_\omega x^2 + \tag{84a}$$

$$- 2|(x/r_\omega cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \tag{84b}$$

$$- c_{\omega,\lambda,y} \tag{84c}$$

We want to characterize r_ω so that the loss function is positive for all $x > r_\omega$.

$$\kappa_\omega x^2 - 2|(x/r_\omega)cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \tag{85a}$$

$$+ \text{constant}_\omega \tag{85b}$$

$$\tag{85c}$$

This will happen if r_w is larger than the larger root of this quadratic function. Triangle inequality gives the first inequality.

$$\frac{b + \sqrt{b^2 + 4ac}}{2a} \leq \frac{2(b + \sqrt{ac})}{2a} \leq r_\omega \tag{86}$$

plugging in for b, a, and c in last inequality

$$\frac{(1/r_\omega)cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\omega + B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}}{2k_\omega} + \frac{\sqrt{\kappa_\omega c_{\omega,\lambda,y}}}{2\kappa_\omega} \leq r_\omega \tag{87}$$

equivalently

$$0 \leq 2\kappa_\omega r_\omega^2 - (cKvp_{\text{eff},\Sigma}^{-1/2}\sqrt{R} + \sqrt{\kappa_\omega c_{\omega,\lambda,y}})r_\omega - cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\} \tag{88}$$

Again, this will happen if r_w is larger than the larger root of this quadratic function. Triangle inequality gives the first inequality.

$$\frac{b+\sqrt{b^2+4ac}}{2a} \leq \frac{2(b+\sqrt{ac})}{2a} \leq r_\omega \tag{89}$$

$$\frac{2(cKvp_{\text{eff},\Sigma}^{-1/2}\sqrt{R}+\sqrt{\kappa_\omega c_{\omega,\lambda,y}})+\sqrt{2\kappa_\omega \times cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega})+B_2\sqrt{k}(s_\omega+s_\lambda)\}}}{4\kappa_\omega} \leq r_\omega \tag{90}$$

ok so in order to not have to deal with c in the square root, we can also use

$$\frac{2(b^2+ac)}{4a^2} \leq r_\omega^2 \tag{91}$$

now recall that $b = (cKvp_{\text{eff},\Sigma}^{-1/2}\sqrt{R}+\sqrt{\kappa_\omega c_{\omega,\lambda,y}})$ so we can write this as $b = b_1 + b_2$. Here $b_1 = (cKvp_{\text{eff},\Sigma}^{-1/2}\sqrt{R}$ and $b_2 = \sqrt{\kappa_\omega c_{\omega,\lambda,y}})$. We have that $b \leq 2\sqrt{b_1^2+b_2^2}$. Therefore $b^2 \leq 4(b_1^2+b_2^2)$ Therefore

$$\frac{2b_1^2+2b_2^2+2ac}{4a^2} \leq r_\omega^2 \tag{92}$$

Applying this to our context

$$\frac{2(cKvp_{\text{eff},\Sigma}^{-1/2})^2R+2\kappa_\omega c_{\omega,\lambda,y}+2\kappa_\omega \times cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega})+B_2\sqrt{k}(s_\omega+s_\lambda)\}}{16\kappa_\omega} \leq r_\omega^2 \tag{93}$$

Let us write out what our constant term here is explicitly.

$$\begin{aligned} c_{\omega,\lambda,y} = {} & -\kappa_\lambda y^2 + 2|\max(y/r_\lambda,1)cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega+s_\lambda)\}| \\ & + c_{\omega,\lambda} \end{aligned} \tag{94}$$

So we need to deal with the two cases here.

1. $y > r_\lambda$
2. $y \leq r_\lambda$

Let us deal with case 1 first.

$$\begin{aligned} c_{\omega,\lambda,y} = {} & -\kappa_\lambda y^2 + 2|y/r_\lambda cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega+s_\lambda)\}| \\ & + c_{\omega,\lambda} \end{aligned} \tag{95}$$

so we want to max this, so we find the worst y for our bound. Let us take FOC wrt to y.

$$-2\kappa_\lambda y + 2|1/r_\lambda cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega+s_\lambda)\}| = 0 \tag{96}$$

therefore

$$\frac{1/r_\lambda cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}}{\kappa_\lambda} = y \tag{97}$$

Now lets think about case 2:

$$\begin{aligned} &c_{\omega,\lambda,y} = -\kappa_\lambda y^2 + 2|cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| \\ &+c_{\omega,\lambda} \end{aligned} \tag{98}$$

So, the max occurs at $y = 0$

So our solution is characterized by

So now that we have gone through these two cases, we have expressions for the ”worst case” values of y. That is they lead to the largest lower bounds on r_ω.

$$\frac{2(cKvp_{\text{eff},\Sigma}^{-1/2})^2 R + 2\kappa_\omega c_{\omega,\lambda,y} + 2\kappa_\omega \times cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}}{16\kappa_\omega} \leq r_\omega^2 \tag{99}$$

where

$$c_{\omega,\lambda,y} = +2|cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}| + c_{\omega,\lambda} \tag{100}$$

$$\begin{aligned} c_{\omega,\lambda} = &\ 2|cKv\,\mathrm{RMSE}_\omega\,\mathrm{w}_\Sigma(\Theta^\star_{s_\omega})| \quad \text{for} \quad \mathrm{RMSE}_\omega := \|(L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega}\cdot\tilde{\beta} - L'_{:N} + Z'_{N::}\cdot\tilde{\beta})\| \\ &+ 2|cKv\,\mathrm{RMSE}_\lambda\,\mathrm{w}_\Sigma(\Theta^\star_{s_\lambda})| \quad \text{for} \quad \mathrm{RMSE}_\lambda := \|(L_{::}\tilde{\lambda} - Z_{:::}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta})\| \\ &+ 2|cvK^2(n/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_s)| \\ &+ 2|cvK^2(p/p_{\text{eff},\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_\lambda)| \end{aligned} \tag{101}$$

and

$$r_\lambda > \frac{2cKvp_{\text{eff}}^{-1/2}\sqrt{R} + \sqrt{\kappa_\lambda cKvp_{\text{eff}}^{-1/2}\sqrt{R}[B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)]}}{2\kappa_\lambda} \tag{102}$$

test test

So in the EIV paper we have the ”r” condition is

$|A\delta_\theta| < \eta n^{1/2} s$

so what we just did was the ability to say

$|A\delta_\omega| < r_\omega$

and

$|A\delta_\lambda| < r_\lambda$

where r_ω and r_λ are defined by the values that satisfy 113 - 116

$$\frac{2(cKvp_{\text{eff},\Sigma}^{-1/2})^2 R + 2\kappa_\omega c_{\omega,\lambda,y} + 2\kappa_\omega \times cKvp_{\text{eff},\Sigma}^{-1/2}\{B_{\omega 1}\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}}{16\kappa_\omega} \le r_\omega^2 \tag{103}$$

$$\begin{aligned} c_{\omega,\lambda,y} = &-\kappa_\lambda\left(\frac{1/r_\lambda cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}}{\kappa_\lambda}\right)^2 \\ &+ 2|\left(\frac{1/r_\lambda cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}}{\kappa_\lambda r_\lambda}\right) cKvp_{\text{eff},\Sigma}^{-1/2}\{\sqrt{R}r_\lambda + B_{\lambda 1}\,\mathrm{w}(\Theta^\star_{s_\lambda}) \\ &+ B_2\sqrt{k}(s_\omega + s_\lambda)\}| \\ &+ c_{\omega,\lambda} \end{aligned} \tag{104}$$

$$\begin{aligned} \|X\delta_\omega\delta_\beta\|^2 &\ge \kappa\mathbb{E}\left(\|X\delta_\omega\delta_\beta\|^2\right) \\ &\ge \kappa\left(\|A\delta_\omega\delta_\beta\|^2 + n\|\Sigma^{1/2}\delta_\omega\|^2\right) \end{aligned} \tag{105}$$

A.6 Establishing the Lower Bound (24)

A.6.1 First Order Optimality Conditions

We first write the first order optimality conditions for $(\tilde{\omega}, \tilde{\lambda}, \tilde{\beta})$. They were calculated by taking the respective derivatives of Equation (10). They help us bound terms later in the proof.

$\tilde{\omega}$

$$\begin{aligned} 0 \le &\left\{L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega}\cdot\tilde{\beta} - L'_{N:} + Z'_{N:\cdot}\cdot\tilde{\beta}\right\}'\left(L'_{::} - Z'_{::\cdot}\cdot\tilde{\beta}\right)[\tilde{\omega} - \omega] \\ &+ n\eta^2(\tilde{\omega} - \psi)\Sigma_{\epsilon_i}(\tilde{\omega} - \omega) \qquad \forall\omega \in \Omega \end{aligned} \tag{106}$$

$\tilde{\lambda}$

$$\begin{aligned} 0 \le &\left\{L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta}\right\}'\left(L_{::} - Z_{::\cdot}\cdot\tilde{\beta}\right)[\tilde{\lambda} - \lambda] \\ &+ n\eta^2(\tilde{\lambda} - \psi)\Sigma_{\epsilon_i}(\tilde{\lambda} - \lambda) \qquad \forall\lambda \in \Lambda \end{aligned} \tag{107}$$

$\tilde{\beta}$

$$\begin{aligned} 0 = \mathrm{FOC}^\beta_\omega + \mathrm{FOC}^\beta_\omega \quad \text{for} \qquad &\mathrm{FOC}^\beta_\omega = \left\{L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega}\cdot\tilde{\beta} - L'_{N:} + Z'_{N:\cdot}\cdot\tilde{\beta}\right\}'\left(Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}\right) \\ \text{and} \qquad &\mathrm{FOC}^\beta_\lambda = \left\{L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta}\right\}'\left(Z_{:T\cdot} - Z_{::\cdot}\tilde{\lambda}\right) \end{aligned} \tag{108}$$

A.6.2 Break Down $\mathcal{L}(\delta_\omega, \delta_\lambda, \delta_\beta)$

We are going to do this by looking first at the first line of (8) focusing on the ω weights. In the first line of (8) the time weights λ do not show up, so we do not have to worry about them. Then we are going to look at the second line of (8) focusing on the λ weights.

We are going to shorten notation. We have $Y^0 = Y - Z \cdot \tilde{\beta}$ and $Y^\delta = Y - Z \cdot (\tilde{\beta} + \delta_\beta)$.

$$\mathcal{L}(\delta) = \mathcal{L}_\omega(\delta_\omega, \delta_\beta) + \mathcal{L}_\lambda(\delta_\lambda, \delta_\beta) \tag{109}$$

where via simple arithmetic ...

$$\begin{aligned}
\mathcal{L}_\omega(\delta_\omega, \delta_\beta) &= \|Y'^{\delta}_{::}(\tilde{\omega} + \delta_\omega) - Y'^{\delta}_{N:}\|^2 + n(\eta^2 - 1)\|\Sigma_{\epsilon_i}^{1/2}((\delta_\omega + \tilde{\omega}) - \psi_\omega)\|^2 \\
&- \left[\|Y'^{0}_{::}\tilde{\omega} - Y'^{0}_{N:}\|^2 + n(\eta^2 - 1)\|\Sigma_{\epsilon_i}^{1/2}(\tilde{\omega} - \psi)\|^2\right] \\
&= \|Y'^{\delta}_{::}(\tilde{\omega} + \delta_\omega) - Y'^{0}_{::}(\tilde{\omega}) - Y'^{\delta_\beta}_{N:} + Y'^{0}_{N:}\|^2 \\
&+ 2(Y'^{0}_{::}\tilde{\omega} - Y'^{0}_{N:})'\left[Y'^{\delta}_{::}(\tilde{\omega} + \delta_\omega) - Y'^{0}_{::}(\tilde{\omega}) - Y'^{\delta_\beta}_{N:} + Y'^{0}_{N:}\right] \\
&+ n(\eta^2 - 1)\|\Sigma_{\epsilon_i}^{1/2}\delta_\omega\|^2 + 2n(\eta^2 - 1)(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega) \\
&= \underbrace{\|Y'^{\delta}_{::}(\tilde{\omega} + \delta_\omega) - Y'^{0}_{::}(\tilde{\omega}) - Y'^{\delta_\beta}_{N:} + Y'^{0}_{N:}\|^2 - n\|\Sigma_{\epsilon_i}^{1/2}\delta_\omega\|^2}_{1} \\
&+ 2\left\{\underbrace{(Y'^{0}_{::}\tilde{\omega} - Y'^{0}_{N:})'\left[Y'^{\delta}_{::}(\tilde{\omega} + \delta_\omega) - Y'^{0}_{::}(\tilde{\omega}) - Y'^{\delta_\beta}_{N:} + Y'^{0}_{N:}\right] - (\eta^2 - 1)n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2}\right\} \\
&+ \eta^2 n\|\Sigma_{\epsilon_i}^{1/2}\delta_\omega\|^2
\end{aligned} \tag{110}$$

and arithmetic in Sec ... below yields ...

$$\mathcal{L}_\omega(\delta_\omega, \delta_\beta) = \underbrace{\|\varepsilon_{::}\delta_\omega\|^2 - n\|\Sigma_{\epsilon_i}^{1/2}\delta_\omega\|^2}_{1.1} + \underbrace{\|A\delta_\omega\delta_\beta\|^2}_{1.3} + \underbrace{2(A\delta_\omega\delta_\beta)(\varepsilon'_{::}\delta_\omega)}_{1.4} \tag{111a}$$

$$+ FOC^\omega + FOC^\beta_\omega \tag{111b}$$

$$+ \left[\underbrace{L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:.} \cdot \tilde{\beta}}_{2.1.1}\right]' \times \left\{\underbrace{Z'_{:::}\delta_\omega \cdot \delta_\beta}_{2.2.2}\right\} \tag{111c}$$

$$+ \left[\underbrace{L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:.} \cdot \tilde{\beta}}_{2.1.1}\right]' \times \left\{\underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4}\right\} \tag{111d}$$

$$+ \left[\underbrace{\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}}_{2.1.2}\right]' \times \left\{A\delta_\omega\delta_\beta\right\} \tag{111e}$$

$$+ \left[\underbrace{\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}}_{2.1.2}\right]' \times \left\{\underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4}\right\} - \underbrace{n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3} \tag{111f}$$

$$+ n\eta^2\|\Sigma_{\epsilon_i}^{1/2}\delta_\omega\|^2 \tag{111g}$$

Table:

- Term (111e)

and, analogously,

$$\mathcal{L}_\lambda(\delta_\lambda, \delta_\beta) = \dots \tag{112}$$

Where equation (110) follows from the previous equation by expanding the square[3] for the first terms on the two lines, and then again for the second terms on the two lines. Now lets look at the terms in equation (110) one by one.

1. The first term in Equation (110) is signal, and we hope it is large.
2. The second term we bound to show it is small
3. Term three is also signal and we hope is large
4. And the last term we bound to show it is small

To make notation shorter

We use

$$A\delta_\omega\delta_\beta = L'_{::}\delta_\omega - Z'_{::\cdot}\delta_\omega \cdot \tilde{\beta} - Z'_{::\cdot}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}) \cdot \delta_\beta \tag{113}$$

$$A\delta_\lambda\delta_\beta = L_{::}\delta_\lambda - Z_{::\cdot}\delta_\lambda \cdot \tilde{\beta} - Z_{::\cdot}\delta_\lambda \cdot \delta_\beta + (Z'_{:T\cdot} - Z_{::\cdot}\tilde{\lambda}) \cdot \delta_\beta \tag{114}$$

A.6.3 Term 1 in Eq (110)

This hasn't really been done. Speculatively, we've used the bound $\kappa\mathbb{E}[...] \leq [...]$.

Here's an entirely unused decomposition. Maybe relevant.

$$\begin{aligned}
\text{Term 1} &= \|Y'^{\delta}_{::}(\tilde{\omega} + \delta_\omega) - Y'^{0}_{::}(\tilde{\omega}) - Y'^{\delta_\beta}_{N:} + Y'^{0}_{N:}\|^2 - n\|\Sigma^{1/2}_{\epsilon_i}\delta_\omega\|^2 \\
&= \|Y'^{0}_{::}\delta_\omega - Z'_{::\cdot}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}) \cdot \delta_\beta\|^2 - n\|\Sigma^{1/2}_{\epsilon_i}\delta_\omega\|^2 \\
&= \|Y'_{::}\delta_\omega - Z'_{::}\delta_\omega \cdot \tilde{\beta} - Z'_{::\cdot}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}) \cdot \delta_\beta\|^2 - n\|\Sigma^{1/2}_{\epsilon_i}\delta_\omega\|^2 \\
&= \|\varepsilon_{::}\delta_\omega + A\delta_\omega\delta_\beta\|^2 - n\|\Sigma^{1/2}_{\epsilon_i}\delta_\omega\|^2 \\
&= \underbrace{\|\varepsilon_{::}\delta_\omega\|^2 - n\|\Sigma^{1/2}_{\epsilon_i}\delta_\omega\|^2}_{1.1} + \underbrace{\|A\delta_\omega\delta_\beta\|^2}_{1.3} + \underbrace{2(A\delta_\omega\delta_\beta)(\varepsilon'_{::}\delta_\omega)}_{1.4}
\end{aligned} \tag{115}$$

A.6.4 Term 2 in Eq (110)

Now we work with Term 2 in Equation (110). We are going to break Term 2 into 6 smaller terms since they are easier to bound.

[3] follows from the fact that $a^2 - b^2 = (a-b)^2 + 2b(a-b)$

$$\begin{aligned}
&\underbrace{2(Y'^0_{::}\tilde{\omega} - Y'^0_{N:})'\left[Y'^\delta_{::}(\tilde{\omega} + \delta_\omega) - Y'^0_{::}\tilde{\omega} - Y'^\delta_{N:} + Y'^0_{N:}\right] - n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2} \\
&= 2(Y'^0_{::}\tilde{\omega} - Y'^0_{N:})'\left[Y'^\delta_{::}\delta_\omega) + Y'^\delta_{::}\tilde{\omega} - Y'^0_{::}\tilde{\omega} + Y'^\delta_{N:} - Y'^0_{N:}\right] - n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega) \\
&= 2(Y'^0_{::}\tilde{\omega} - Y'^0_{N:})'\left[Y'^\delta_{::}\delta_\omega - Z'_{:::}\tilde{\omega} \cdot \delta_\beta + Z'_{N:\cdot} \cdot \delta_\beta\right] - n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega) \\
&= 2(Y'^0_{::}\tilde{\omega} - Y'^0_{N:})'\left[Y'^0_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \delta_\beta - Z'_{:::}\tilde{\omega} \cdot \delta_\beta + Z'_{N:\cdot} \cdot \delta_\beta\right] - n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega) \qquad (116) \\
&= 2(Y'^0_{::}\tilde{\omega} - Y'^0_{N:})'\left[Y'^0_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta\right] - n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega) \\
&= 2\underbrace{(Y'^0_{::}\tilde{\omega} - Y'^0_{N:})'}_{2.1}\underbrace{\left[Y'^0_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta\right]}_{2.2} - 2\underbrace{(\eta^2 - 1)n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3}
\end{aligned}$$

We are now going to break this term down into the terms that have noise and those that do not. Lets look at these terms one by one.

$$\begin{aligned}
2.1 &= \left[Y'^0_{::}\tilde{\omega} - Y'^0_{N:}\right]' \\
&= \left[L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} + \epsilon'_{::}\tilde{\omega} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta} - \epsilon'_{N:}\right]' \qquad (117) \\
&= \left[\underbrace{L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}}_{2.1.1} + \underbrace{\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}}_{2.1.2}\right]'
\end{aligned}$$

Now the last term is the only one with noise.

Now moving on to term 2.2.

$$\begin{aligned}
2.2 &= \left[Y'^0_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta\right] \\
&= \left[L'_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \tilde{\beta} + \epsilon'_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta\right]
\end{aligned} \qquad (118)$$

Now we write 2.2 into three constant parts (2.2.1) and (2.2.2) and (2.2.3) and the noise part (2.2.4).

$$\begin{aligned}
2.2 = &\underbrace{L'_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \tilde{\beta}}_{2.2.1} \\
&\underbrace{Z'_{:::}\delta_\omega \cdot \delta_\beta}_{2.2.2} \\
&\underbrace{(Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta}_{2.2.3} \\
&\underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4}
\end{aligned} \qquad (119)$$

So the reminder of Term 2 will be all the terms of 2.1 times all the terms of 2.2

$$\underbrace{(\eta^2-1)n(\omega-\psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3}=\underbrace{\eta^2 n(\omega-\psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3.1}-\underbrace{n(\omega-\psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3.2} \tag{120}$$

$$\begin{aligned}
2 = {} & (2.1.1)\times(2.2.1)-(2.3.1) && \geq 0 \text{ by FOC } \omega && (121a)\\
& +(2.1.1)\times(2.2.2) && \text{to appear in LB} && (121b)\\
& +(2.1.1)\times(2.2.3) && = FOC_\omega^\beta.\ \text{Appears in LB.} && (121c)\\
& +(2.1.1)\times(2.2.4) && \text{Mean zero noise} && (121d)\\
& +(2.1.2)\times(2.2.1+2.2.2+2.2.3) && \text{Mean zero noise} && (121e)\\
& +(2.1.1)\times(2.2.4)-(2.3.2) && \text{Centered cross noise term} && (121f)
\end{aligned}$$

A.6.5 Proofs for Term 2 Bounds

(2.1.1) × (2.2.1) **1) Terms for FOC of ω** The FOC for ω is in Equation (106). Adding terms (2.1.1) × (2.2.1) and 2.3 together we get the conditions for the FOC for ω, so we know these terms are positive, thus taking care of them.

$$\begin{aligned}
(2.1.1)\times(2.2.1) &= \Big[\underbrace{L'_{::}\tilde\omega - Z'_{:::}\tilde\omega\cdot\tilde\beta - L'_{N:} + Z'_{N:\cdot}\cdot\tilde\beta}_{2.1.1}\Big]'\times\Big\{\underbrace{L'_{::}\delta_\omega - Z'_{:::}\delta_\omega\cdot\tilde\beta}_{2.2.1}\Big\}\\
&= \Big[L'_{::}\tilde\omega - Z'_{:::}\tilde\omega\cdot\tilde\beta - L'_{N:} + Z'_{N:\cdot}\cdot\tilde\beta\Big]'\times\Big\{L'_{::} - Z'_{:::}\cdot\tilde\beta\Big\}\delta_\omega
\end{aligned} \tag{122}$$

(2.1.1) × (2.2.2) **2) No noise, and small two δ**

$$(2.1.1)\times(2.2.2) = \Big[\underbrace{L'_{::}\tilde\omega - Z'_{:::}\tilde\omega\cdot\tilde\beta - L'_{N:} + Z'_{N:\cdot}\cdot\tilde\beta}_{2.1.1}\Big]'\times\Big\{\underbrace{Z'_{:::}\delta_\omega\cdot\delta_\beta}_{2.2.2}\Big\} \tag{123}$$

This term will be small when the oracle prediction error $\|L'_{::}\tilde\omega - Z'_{:::}\tilde\omega\cdot\tilde\beta - L'_{N:} + Z'_{N:\cdot}\cdot\tilde\beta\|$ is small. It will also be small because we are multiplying it by two small δ values, δ_ω and δ_λ

(2.1.1) × (2.2.3) **2) 1/2 FOC for β** We have from the FOC of β (Equation (108)) is

$$\begin{aligned}
0 &= \{L'_{::}\tilde\omega - Z'_{:::}\tilde\omega\cdot\tilde\beta - L'_{N:} + Z'_{N:\cdot}\cdot\tilde\beta\}'\Big(Z'_{N:\cdot} - Z'_{:::}\tilde\omega\Big)\\
&+ \{L_{::}\tilde\lambda - Z_{:::}\tilde\lambda\cdot\tilde\beta - L_{:T} + Z_{:T\cdot}\cdot\tilde\beta\}'\Big(Z_{:T\cdot} - Z_{:::}\tilde\lambda\Big)
\end{aligned} \tag{124}$$

We have that (2.1.1) × (2.2.3) is the first half of the FOC for β times δ_β. Then we have the sister term for the second line. We see that is the same as the full FOC, just times δ_β

$$(2.1.1) \times (2.2.3) = \Big[\underbrace{L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:}\tilde{\omega} + Z'_{N:\cdot} \cdot \tilde{\beta}}_{2.1.1} \Big]' \times \Big\{ \underbrace{(Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}) \cdot \delta_\beta}_{2.2.3} \Big\} \tag{125}$$

$$(\text{sister term}) = \Big[\underbrace{L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda} \cdot \tilde{\beta} - L_{:T}\tilde{\lambda} + Z_{:T\cdot} \cdot \tilde{\beta}} \Big]' \times \Big\{ \underbrace{(Z_{:T\cdot} - Z_{::\cdot}\tilde{\lambda}) \cdot \delta_\beta} \Big\} \tag{126}$$

(2.1.1) × (2.2.4) **4) Mean zero noise**

$$(2.1.1) \times (2.2.4) = \Big[\underbrace{L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}}_{2.1.1} \Big]' \times \Big\{ \underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4} \Big\} \tag{127}$$

(2.1.2) × (2.2.1 + 2.2.2 + 2.2.3) **5) Mean zero noise**

$$(2.1.2) \times (2.2.1 + 2.2.2 + 2.2.3) = \Big[\underbrace{\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}}_{2.1.2} \Big]' \times \Big\{ \underbrace{L'_{::}\delta_\omega - Z'_{::\cdot}\delta_\omega \cdot \tilde{\beta} - Z'_{::\cdot}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{::\cdot}\tilde{\omega}) \cdot \delta_\beta}_{2.2.1+2.2.2+2.2.3=A\delta_\omega\delta_\beta} \Big\} \tag{128}$$

(2.1.2) × (2.2.4) **6) Cross noise term**

$$(2.1.2) \times (2.2.4) - 2.3 = \Big[\underbrace{\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}}_{2.1.2} \Big]' \times \Big\{ \underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4} \Big\} - \underbrace{n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3} \tag{129}$$

B Concentration

B.1 (2.1.1) X (2.2.4)

$$(2.1.1) \times (2.2.4) = \Big[\underbrace{L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}}_{2.1.1}\Big]' \times \Big\{\underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4}\Big\} \tag{130}$$

This term is the product of prediction error $L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}$ and the weighted error matrix and δ_ω. This will tend to be small when prediction error is small, for example when $L'_{::} - Z'_{::} \cdot \beta$ is low rank.

Let us define $z' = (L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta})'\epsilon'_{::}S^{-1/2}$. This z' is a subgaussian vector. With this notation, bounding $(2.1.1) \times (2.2.4)$ is the same as bounding $|z'S^{1/2}\delta_\omega|$ where $S^{1/2}\delta_\omega \in S^{1/2}\Theta^\star_s$. By Talagrand's majorizing measures theorem [Vershynin, 2018, Corollary 8.6.3], on an event of probability $1 - 2\exp\big(-u^2\big)$, this is bounded for all $\delta_\omega \in \Theta^\star_s$ by $c\|z\|_{\psi_2}\{\mathrm{w}(S^{1/2}\Theta^\star_s) + u\,\mathrm{rad}(S^{1/2}\Theta^\star_s)\}$.

When it has independent columns, we take $S = I$, and $\|z'|_{\psi_2} = \|(\Sigma_{\varepsilon.j}^{-1/2}\varepsilon)'\{\Sigma_{\varepsilon.j}^{1/2}(L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta})\}\|_{\psi_2} \leq cK\|\Sigma_{\varepsilon.j}^{1/2}(L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta})\|$. Thus, the following bounds hold for all $\delta_\omega \in \Theta^\star_{s_\omega}$ on an event of probability $1 - 2\exp\big(-u^2\big)$.

$$\begin{aligned}|z'S^{1/2}\delta_\omega| &\leq cK\|L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}\|(\mathrm{w}(\Sigma_{\varepsilon_i.}^{1/2}\Theta^\star_{s_\omega}) + u\,\mathrm{rad}(\Sigma_{\varepsilon_i.}^{1/2}\Theta^\star_{s_\omega})) \\ &\leq cK\|L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}\|(\mathrm{w}_\Sigma(\Theta^\star_s) + u\,\mathrm{rad}(\Sigma_{\varepsilon_i.}^{1/2}\Theta^\star_s)) \qquad \text{ind. rows}\end{aligned} \tag{131}$$

Taking $v = u\,\mathrm{rad}(\Sigma_{\varepsilon_i.}^{1/2}\Theta^\star_s)/\,\mathrm{w}_\Sigma(\Theta^\star_s)$ in the first case and $v = u\|\Sigma_{\varepsilon.j}\|^{1/2}\,\mathrm{rad}(\Theta^\star_s)/\,\mathrm{w}_\Sigma(\Theta^\star_s)$ in the second, for which we have the common notation $v = ud_\Sigma^{-1/2}(\Theta^\star_s)$, this implies that on an event of probability $1 - 2\exp\{-v^2 d_\Sigma(\Theta^\star_s)\}$,

$$|z'S^{1/2}\delta| \leq c(1+v)K\|L'_{::}\tilde{\omega} - Z'_{:::}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:\cdot} \cdot \tilde{\beta}\|\,\mathrm{w}_\Sigma(\Theta^\star_s) \quad \text{for all } \delta \in \Theta^\star_{s_\omega}. \tag{132}$$

B.2

In this section we bound two quantities appearing in [our bound],

$$\Big[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}\Big]'\Big\{A\delta_\omega\delta_\beta\Big\} \quad \text{and} \quad [\varepsilon'_{::}\delta_\omega]'\{A\delta_\omega\delta_\beta\} \quad \text{for} \quad A\delta_\omega\delta_\beta = L'_{::}\delta_\omega - Z'_{:::}\delta_\omega \cdot \tilde{\beta} - Z'_{:::}\delta_\omega \cdot \delta_\beta + (Z'_{N:\cdot} - Z'_{:::}\tilde{\omega}) \cdot \delta_\beta \tag{133}$$

The first one . Let A_Σ be $\Sigma^{1/2}A$ where Σ is the covariance matrix of $\epsilon'_{::}\tilde{\omega} - \epsilon_{N:}$, so this is just $g'A_\Sigma\delta_\omega\delta_\beta$ for an identity-covariance subgaussian vector g. This is comparable to its gaussian width (Talagrand's comparison inequality) $\mathrm{w}(\Sigma^{1/2}A\Theta^*_{s,r})$ where $A\Theta^*_{s,r} = \{A\delta_\omega\delta_\beta : (\delta_\omega, \delta_\lambda, \delta_\beta) \in \Theta^*_{s,r}\}$. To bound this quantity, we decompose $A_\Sigma\delta_\beta\delta_\omega$ into two terms: its projection on a low-dimensional subspace of $\mathbb{R}^n$ and the remainder. Letting P_R be an orthogonal projection onto arbitrary R-dimensional subspace, because $\mathrm{w}(A + B) \leq \mathrm{w}(A) + \mathrm{w}(B)$ for any sets A and B,

$$\mathrm{w}(A_\Sigma\Omega\mathcal{B}) \leq \mathrm{w}(P_R A_\Sigma\Omega\mathcal{B}) + \mathrm{w}(A^\perp\Omega\mathcal{B}) \quad \text{for} \quad A^\perp = (I - P_R)A_\Sigma.$$

$P_R A_\Sigma \Omega \mathcal{B}$ is a set of vectors of length $\leq \|\Sigma\|^{1/2} r$ in a subspace of $dim \leq R$, hence has width bound $c\|\Sigma\|^{1/2} r\sqrt{R}$. To bound the other term, we use Sudakov-Fernique, i.e.,

$$\mathbb{E}(\max X_{uv}) \leq \mathbb{E}(\max Y_{uv}) \quad \text{if} \quad \mathbb{E}(X_{uv} - X_{wz})^2 \leq \mathbb{E}(Y_{uv} - Y_{wz})^2 \quad \text{for all} \quad u, v \tag{134}$$

where $X_{uv} = g' A^\perp uv$ and $Y_{uv} = R_v g'(u - w) + R_u h'(v - z)$ for independent vectors of standard normals g, h and R_u and R_v are defined in (135) below. Here u and w are values of δ_ω and v and z are values of δ_β and we are maximizing over $(u, v) \in P_{\omega,\beta}\Theta^\star_{s,r}$, the set of pairs $(\delta_\omega, \delta_\beta)$ for which there exists some δ_λ for which the triple $(\delta_\omega, \delta_\lambda, \delta_\beta)$ is in $\Theta^\star_{s,r}$.

Let's check that the variance of the increments of Y is a bound on that of the corresponding increment of X.

$$\begin{aligned}
\mathbb{E}(X_{uv} - X_{wz})^2 &= \mathbb{E}\{g'(A^\perp uv - A^\perp wz)\}^2 \\
&= \|A^\perp uv - A^\perp wz\|^2 \\
&= \|A^\perp uv \pm A^\perp wv - A^\perp wz\|^2 \\
&\leq \|A^\perp (u - w)v\| + \|A^\perp w(v - z)\| && \text{(triangle inequality)} \\
&\leq \|A^\perp \Box v\|_{op} \|u - w\| + \|A^\perp w \Box\|_{op} \|v - z\| \\
&\leq R_\beta \|u - w\| + R_\omega \|v - z\| \quad \text{for} \\
R_\beta = \max_{v \in P_\beta \Theta^\star_{s,r}} \|A^\perp \Box v\| &\quad \text{and} \quad R_\omega = \max_{w \in P_\omega \Theta^\star_{s,r}} \|A^\perp w \Box\|.
\end{aligned} \tag{135}$$

Here $\|A^\perp \Box v\|$ and $\|A^\perp w \Box\|$ and are the operator norms of the linear maps we get by partially applying the bilinear operator $A^\perp$ with its first and second arguments respectively set to w and v. Notationally, $\Box$ stands in for the dimension that is left, indicating that partial application to v and w are along different dimensions.

The second one .

$$\delta'_\omega \varepsilon_{::}\{A\delta_\omega \delta_\beta\} \leq \sup_{u \in \Omega, (v,w) \in \Omega\mathcal{B}} u'(Avw) \leq \operatorname{rad}(\Omega)\, \mathrm{w}(A\Omega\mathcal{B}) + \mathrm{w}(\Omega)\operatorname{rad}(A\Omega\mathcal{B}) \tag{136}$$

where the second comparison is via Chevet's inequality. The first term we've just done the width calculation for and have a bound s_ω on the radius and the second is just $\mathrm{w}(\Omega_s)$ times a radius we have a bound r_ω on.

In the OG paper the idea was to control $\|A\delta\| < r_\omega$. The idea now is to see if we can do

$$\begin{bmatrix} L'_{::} - Z'_{::} \cdot \beta & 0 \\ Z'_{::} & Z_{N:} - Z'_{::} \end{bmatrix} \begin{bmatrix} \delta_\omega \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ \delta_\beta \end{bmatrix} < r_\omega \tag{137}$$

Let us call $A = \begin{bmatrix} L'_{::} - Z'_{::} \cdot \beta & 0 \\ Z'_{::} & Z_{N:} - Z'_{::} \end{bmatrix}$. We decompose A as $A_R + (A - A_R)$ where A_R is a rank-R approximation to A. Taking A_R to be the best rank-R approximation in terms of operator norm, the first R terms of the singular value decomposition $A = \sum_k \sigma_k u_k v'_k$, $\|A_R x\| \leq \|Ax\|$ for all vectors x. Thus, $A_R \Theta^\star_{s,r}$ is contained in the ball of radius r in the R-dimensional image of A_R, which has gaussian width bounded by $c\sqrt{R}r$ [e.g., Vershynin, 2018, Example 7.5.7].

Now we figure out how to bound $(A - A_R)\Theta^{\star}_{s,r}$

We rewrite the term we want to bound.

$$\begin{aligned}
&= \left[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}\right]' \times \left\{ L'_{::}\delta_\omega - Z'_{::}.\delta_\omega \cdot \tilde{\beta} - Z'_{::}\delta_\omega \cdot \delta_\beta + (Z'_{N:} - Z'_{::}\omega) \cdot \delta_\beta \right\} \\
&= \left[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}\right]' \begin{bmatrix} L'_{::} - Z'_{::} \cdot \beta & 0 \\ Z'_{::} & Z_{N:} - Z'_{::} \end{bmatrix} \begin{bmatrix} \delta_\omega \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ \delta_\beta \end{bmatrix} \\
&= \left[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}\right]' A^{\perp} uv \\
&= X_{uv}
\end{aligned} \tag{138}$$

Our term is a random process X_{uv}. We use $A^{\perp}$ to denote our square matrix of L and Z, u for the δ_ω vector, and v for the δ_β vector. This term is tricky to bound because of the quadratic term that is a multiple of both δ_ω and δ_β. To solve this we we create a new random process Y_{uv} that both 1) bounds the orginal X_{uv} process 2) is linear in δ_ω and δ_β, and we know how to bound linear terms.

We create this Y_{uv} in the following 3 steps.

1. We calculate an upper bound variance of $X_{uv} - X_{wz}$.

   $$\begin{aligned}
   Var(X_{uv} - X_{wz}) = \mathbb{E}(X_{uv} - X_{wz})^2 &= \left[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}\right]' [A^{\perp}uv - A^{\perp}wz] \\
   &= [A^{\perp}uv - A^{\perp}wz]' \Sigma_{\epsilon_i \cdot} [A^{\perp}uv - A^{\perp}wz]
   \end{aligned} \tag{139}$$

   Therefore

   $$\begin{aligned}
   \sqrt{Var(X_{uv} - X_{wz})} &= \|\Sigma^{1/2}_{\epsilon_i \cdot}[A^{\perp}uv - A^{\perp}wz]\| \\
   &= \|\Sigma^{1/2}_{\epsilon_i \cdot}[A^{\perp}uv \pm A^{\perp}wv - A^{\perp}wz]\| \\
   &\le \|\Sigma^{1/2}_{\epsilon_i \cdot}[A^{\perp}(u - w)v\| + \|\Sigma^{1/2}_{\epsilon_i \cdot}A^{\perp}w(v - z)]\| (\text{triangle inequality}) \\
   &\le \|\Sigma^{1/2}_{\epsilon_i \cdot}A^{\perp}v\| \|(u - w)\| + \|\Sigma^{1/2}_{\epsilon_i \cdot}A^{\perp}w\| \|(v - z)\| (\text{Cauchy–Schwarz})
   \end{aligned} \tag{140}$$

   Let us first focus on the first term of the last line.

   $$\|\Sigma^{1/2}_{\epsilon_i \cdot}A^{\perp}v\| = \tag{141}$$

2. We create a Gaussian process Y_{uv} so that $Y_{uv} - Y_{wz}$ a) is linear in δ_ω and δ_β b) has variance at least as large of the variance of $X_{uv} - X_{wz}$.

   $$Y_{uv} := \langle g, u \rangle \,\mathrm{rad}() + \langle h, v \rangle \,\mathrm{rad}() \tag{142}$$

   where

   $$g \sim N(0, I) \text{ and } h \sim N(0, I) \tag{143}$$

   This formula for Y_{uv} comes from Theorem 8.7.1 (Sub-gaussian Chevets inequality) of Vershynin [2018].

3. The Sudakov-Fernique inequality tells that

   So we have that Y_{uv} bounds our original process.

B.3 Term 6

(2.1.2) × (2.2.4) **6) Cross noise term**

$$(2.1.2) \times (2.2.4) - 2.3 = \Big[\underbrace{\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}}_{2.1.2}\Big]' \times \Big\{\underbrace{\epsilon'_{::}\delta_\omega}_{2.2.4}\Big\} - \underbrace{n(\omega - \psi)'(\Sigma_{\epsilon_i}\delta_\omega)}_{2.3} \tag{144}$$

Term (2.1.2)×(2.2.4) is not mean zero noise because it is not mean zero. Thankfully we can center it by subtracting term 2.3, which is its mean. The mean of (2.1.2)×(2.2.4) is $\mathbb{E}[(\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:})' \times \{\epsilon'_{::}\delta_\omega\}] = (\omega - \psi)' n(\Sigma_{\epsilon_i}\delta_\omega)$ in order to create a mean zero noise term. We see that this term is not a function of any of the β or Z terms, so it's bounds follow directly from Hirshberg [2021]. We have that with probability $1 - 6\exp[-c\min\{vd_\Sigma(\Theta^*_s), n\}]$,

$$\sup_{\delta\in\Theta^*_s} \|[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}]' \times \{\epsilon'_{::}\delta_\omega\} - \mathbb{E}[\epsilon'_{::}\tilde{\omega} - \epsilon'_{N:}]' \times \{\epsilon'_{::}\delta_\omega\}]\| \le cvK^2(n/p_{eff,\Sigma})^{1/2} w_\Sigma(\Theta^*_s) \tag{145}$$

C Bringing Bounds together

Our proof relies on a few high-probability bounds on the terms in $\tilde{\mathcal{L}}(\delta)$. Choose $s_\omega, s_\lambda, r_\omega, r_\lambda \in \mathbb{R}$ and $R \in \mathbb{N}$

Notation:

and define $\Theta^* := \Theta - \tilde{\omega} - \tilde{\lambda} - \tilde{\beta}$

and $\Theta^\star_{s_\omega} := \{\delta \in \Theta^\star : \|\Sigma^{1/2}_{\varepsilon.j}\delta_\omega\| \le s_\omega\}$ and $\Theta^\star_{s_\lambda} := \{\delta \in \Theta^\star : \|\Sigma^{1/2}_{\varepsilon_i.}\delta_\lambda\| \le s_\lambda\}$.

and $\Theta^\star_s := \{\delta \in \Theta^\star : \|\Sigma^{1/2}_{\varepsilon_i.}\delta_\omega\| \le s_\omega \text{ and } \|\Sigma^{1/2}_{\varepsilon_i.}\delta_\lambda\| \le s_\lambda\}$.

We use

$$A\delta_\omega\delta_\beta = L'_{::}\delta_\omega - Z'_{:::.}\delta_\omega \cdot \tilde{\beta} - Z'_{:::.}\delta_\omega \cdot \delta_\beta + (Z'_{N:.} - Z'_{:::.}\tilde{\omega}) \cdot \delta_\beta \tag{146}$$

$$A\delta_\lambda\delta_\beta = L_{::}\delta_\lambda - Z_{:::.}\delta_\lambda \cdot \tilde{\beta} - Z_{:::.}\delta_\lambda \cdot \delta_\beta + (Z'_{:T.} - Z_{:::.}\tilde{\lambda}) \cdot \delta_\beta \tag{147}$$

terms for the ω parts: On an event of probability $1 - c\exp\big[-c\min\{v^2\phi^{-4}\,\mathrm{w}^2_\Sigma(\Theta^\star)/s^2,\ v^2R,\ n\}\big]$, all $\delta_\omega \in \Theta^\star_{s_\omega}$ satisfy

$$\|X\delta_\omega\delta_\beta\|^2 \ge \kappa\|A\delta_\omega\delta_\beta\|^2 + \kappa n\|\Sigma^{1/2}_{\varepsilon_i.}\delta_\omega\|^2. \tag{148}$$

Bounds on terms

$$|(\varepsilon'_{::}\tilde{\omega} - \epsilon'_{N:})' A\delta_\omega\delta_\beta| \le \max(\|A\delta\|/r, 1) cKvp^{-1/2}_{eff,\Sigma}\{\sqrt{R}r + B_1\,\mathrm{w}(\Theta^\star_{s_\omega}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}. \tag{149}$$

$$|(L'_{::}\tilde{\omega} - Z'_{:::.}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:.} \cdot \tilde{\beta})'\varepsilon\delta_\omega| \le cKv\|(L'_{::}\tilde{\omega} - Z'_{:::.}\tilde{\omega} \cdot \tilde{\beta} - L'_{N:} + Z'_{N:.} \cdot \tilde{\beta})\|\,\mathrm{w}_\Sigma(\Theta^\star_{s_\omega}) \tag{150}$$

$$|[(\varepsilon'_{::}\tilde{\omega} - \epsilon'_{N:})'\varepsilon'_{::} - \mathbb{E}(\varepsilon'_{::}\tilde{\omega} - \epsilon'_{N:})'\varepsilon'_{::}]\delta_\omega| \le cvK^2(n/p_{eff,\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_s). \tag{151}$$

$$\begin{aligned}\text{constant bit} &= \left[L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega}\cdot\tilde{\beta} - L'_{N:} + Z'_{N:\cdot}\cdot\tilde{\beta}\right]' \times \left\{Z'_{::\cdot}\delta_\omega \cdot \delta_\beta\right\} \\ &\leq \|L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega}\cdot\tilde{\beta} - L'_{N:} + Z'_{N:\cdot}\cdot\tilde{\beta}\|\|Z'_{::}s_\omega(s_\omega + s_\lambda)\|\end{aligned} \tag{152}$$

terms for the λ On an event of probability $1 - c\exp\bigl[-c\min\{v^2\phi^{-4}\,\mathrm{w}^2_\Sigma(\Theta^\star)/s^2,\ v^2R,\ n\}\bigr]$, all $\delta_\lambda \in \Theta^\star_{s_\lambda}$ satisfy

$$\|X\delta_\lambda\delta_\beta\|^2 \geq \kappa\|A\delta_\lambda\delta_\beta\|^2 + \kappa n\|\Sigma^{1/2}_{\varepsilon_{i\cdot}}\delta_\lambda\|^2. \tag{153}$$

$$|(\varepsilon_{::}\tilde{\lambda} - \epsilon_{:T})'A\delta_\lambda\delta_\beta| \leq \max(\|A\delta\|/r, 1)cKvp^{-1/2}_{eff,\Sigma}\{\sqrt{R}r + \sigma_{R+1} + B_1\,\mathrm{w}(\Theta^\star_{s_\lambda}) + B_2\sqrt{k}(s_\omega + s_\lambda)\}. \tag{154}$$

$$|(L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z'_{:T\cdot}\cdot\tilde{\beta})'\varepsilon\delta_\lambda| \leq cKv\|(L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta})\|\,\mathrm{w}_\Sigma(\Theta^\star_{s_\lambda}) \tag{155}$$

$$|[(\varepsilon_{::}\tilde{\lambda} - \epsilon_{:T})'\varepsilon_{::} - \mathbb{E}(\varepsilon_{::}\tilde{\lambda} - \epsilon_{:T})'\varepsilon_{::}]\delta_\lambda| \leq cvK^2(n/p_{eff,\Sigma})^{1/2}\,\mathrm{w}_\Sigma(\Theta^\star_s). \tag{156}$$

$$\begin{aligned}\text{constant bit} &= \left[L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta}\right]' \times \left\{Z_{::\cdot}\delta_\lambda \cdot \delta_\beta\right\} \\ &\leq \|L_{::}\tilde{\lambda} - Z'_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L'_{:T} + Z'_{:T\cdot}\cdot\tilde{\beta}\|\|Z_{::}s_\lambda(s_\omega + s_\lambda)\|\end{aligned} \tag{157}$$

D D.2

In equation 30 we had defined $\tilde{L}$ and then we spent appendix A bounding the terms. Therefore we can lower bound $\tilde{L}$ with the terms on the right hand side of 79.

$$\begin{aligned}\tilde{L}(\delta) \geq{}& \kappa_\omega\|A\delta_\omega\delta_\beta\|^2 + \kappa_\omega n\|\Sigma^{1/2}\delta_\omega\|^2 + \kappa_\lambda\|A\delta_\lambda\delta_\beta\|^2 + \kappa_\lambda n\|\Sigma^{1/2}\delta_\lambda\|^2 && \text{(158a)} \\ &- 2|(\varepsilon'_{::}\tilde{\omega} - \epsilon'_{N:})'A\delta_\omega\delta_\beta| && \text{(158b)} \\ &- 2|(\varepsilon_{::}\tilde{\lambda} - \epsilon_{:T})A\delta_\lambda\delta_\beta| && \text{(158c)} \\ &- 2|(L'_{::}\tilde{\omega} - Z'_{::\cdot}\tilde{\omega}\cdot\tilde{\beta} - L'_{N:} + Z'_{N:\cdot}\cdot\tilde{\beta})'\varepsilon\delta_\omega| && \text{(158d)} \\ &- 2|(L_{::}\tilde{\lambda} - Z_{::\cdot}\tilde{\lambda}\cdot\tilde{\beta} - L_{:T} + Z_{:T\cdot}\cdot\tilde{\beta})\varepsilon\delta_\lambda| && \text{(158e)} \\ &- 2|[(\varepsilon'_{::}\tilde{\omega} - \epsilon'_{N:})'\varepsilon'_{::} - \mathbb{E}(\varepsilon'_{::}\tilde{\omega} - \epsilon'_{N:})'\varepsilon'_{::}]\delta_\omega| && \text{(158f)} \\ &- 2|[(\varepsilon_{::}\tilde{\lambda} - \epsilon_{:T})\varepsilon_{::} - \mathbb{E}(\varepsilon_{::}\tilde{\lambda} - \epsilon'_{:T})\varepsilon_{::}]\delta_\lambda| && \text{(158g)}\end{aligned}$$