Leveraging Low-Rank Relations Between Surrogate Tasks in Structured Prediction: Supplementary Material

The supplementary material is organized as follows:

- In Appendix A we show how the loss trick for both the vector-valued and multitask SELF estimator is derived.
- In Appendix B we carry out the theoretical analysis for trace norm estimator in the vector-valued setting.
- In Appendix C we prove the theoretical results characterizing the generalization properties of the SELF multitask estimator.
- In Appendix D we recall some results that are used in the proofs of previous sections.
- In Appendix E more details on the equivalence between Ivanov and Tikhonov regularization are provided.

A. Loss Trick(s)

In this section we discuss some aspects related to the loss trick of the SELF framework when considering different vector-valued or MTL estimators.

A.1. Loss Tricks with Matrix Factorization

In this section we provide full details of the loss trick for trace norm regularization partly discussed in Section 3. To fix the setting, recall that we are interested in studying the following surrogate problem

$$\min_{G \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} \|G\phi(x_i) - \psi(y_i)\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda \|G\|_*. \tag{19}$$

Theorem 2 (Loss Trick for Trace Norm). *Under Asm. 1, let $M, N \in \mathbb{R}^{n \times r}$ and (A_k, B_k) be the k-th iterate of gradient descent on Eq. (12) from $A_0 = \sum_{i=1}^n \phi(x_i) \otimes M^i$ and $B_0 = \sum_{i=1}^n \psi(y_i) \otimes N^i$, with M^i, N^i denoting the i-th rows of M and N respectively. Let $\hat{g}_k : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ be such that $\hat{g}_k(\cdot) = A_k B_k^* \phi(\cdot)$. Then, the structured prediction estimator $\hat{f}_k = \mathsf{d} \circ \hat{g}_k : \mathcal{X} \to \mathcal{Y}$ with decoding d in Eq. (5) is such that*

$$\hat{f}_k(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i^{\mathsf{tn}}(x)\, \ell(y, y_i)$$

for any $x \in \mathcal{X}$, with $\alpha^{\mathsf{tn}}(x) \in \mathbb{R}^n$ the output of Alg. 1 after k iterations starting from $(M_0, N_0) = (M, N)$.

Proof. We show the proof in the finite dimensional setting first and then note how it is valid in the infinite dimensional case as well. Assume $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H}_y = \mathbb{R}^T$. Let $(x_i, y_i)_{i=1}^n$ be the training set and denote by X the $\mathbb{R}^{n \times d}$ matrix containing the training inputs x_i, $i = 1, \dots, n$ and Y the $\mathbb{R}^{n \times T}$ matrix whose rows are $\psi(y_i)$, $i = 1, \dots, n$. Denote by K_x the matrix $XX^\top$ and by K_y the matrix $YY^\top$.

Using the variational form of trace norm, problem (19) can be rewritten as

$$\min_{A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{T \times r}} \frac{1}{n} \|Y - XAB^\top\|^2 + \lambda(\|A\|^2_{\mathsf{HS}} + \|B\|^2_{\mathsf{HS}}), \tag{20}$$

where $r \in \mathbb{N}$ in a further hyperparameter of the problem. In the following we will absorb the factor $1/n$ in the hyperparameter λ.

We first show that starting gradient descent algorithm with $A_0 := X^\top M_0$ for some matrix $M_0 \in \mathbb{R}^{n \times r}$ and $B_0 := Y^\top N_0$ for some matrix $N_0 \in \mathbb{R}^{n \times r}$, then at every iteration $A_k := X^\top M_k$ and $B_k := Y^\top N_k$.

Let us set

$$\mathcal{L}(A, B) := \|Y - XAB^\top\|^2 + \lambda(\|A\|^2_{\mathsf{HS}} + \|B\|^2_{\mathsf{HS}});$$

the gradients of $\mathcal{L}$ with respect to A and B are given by

1) $\nabla_A \mathcal{L}(A, B) = X^\top (XAB^\top - Y)B + \lambda A$

2) $\nabla_B \mathcal{L}(A, B) = (XAB^\top - Y)^\top XA + \lambda B.$

We show that $A_k := X^\top M_k$ and $B_k := Y^\top N_k$ by induction. Assume it is true for k and show it holds for $k+1$; denoting by ν the stepsize, we have

$$\begin{aligned} A_{k+1} &= A_k - \nu \nabla_A \mathcal{L}(A_k, B_k) = A_k - \nu(X^\top (XA_k B_k^\top - Y)B_k + \lambda A_k) \\ &= X^\top M_k - \nu(X^\top X X^\top M_k B_k^\top B_k - X^\top Y B_k) - \nu\lambda X^\top M_k \\ &= X^\top((1-\lambda\nu)M_k - \nu(K_{\mathcal{X}} M_k B_k^\top B_k - Y B_k) \\ &= X^\top \big((1-\lambda\nu)M_k - \nu(K_{\mathcal{X}} M_k N_k^\top K_{\mathcal{Y}} N_k - K_{\mathcal{Y}} N_k)\big), \end{aligned}$$

and hence $A_{k+1} = X^\top M_{k+1}$

$$M_{k+1} = (1-\lambda\nu)M_k - \nu\big(K_{\mathcal{X}} M_k N_k^\top K_{\mathcal{Y}} N_k - K_{\mathcal{Y}} N_k\big). \tag{21}$$

As for B, assume $B_k = Y^\top N_k$:

$$\begin{aligned} B_{k+1} &= B_k - \nu \nabla_B \mathcal{L}(A_k, B_k) \\ &= B_k - \nu((XA_k B_k^\top - Y)^\top XA_k + \lambda B_k) \\ &= Y^\top N_k - \nu(Y^\top N_k A_k^\top X^\top X A_k - Y^\top X A_k) - \nu\lambda Y^\top N_k \\ &= Y^\top((1-\lambda\nu)N_k - \nu(N_k(K_{\mathcal{X}} M_k)^\top K_{\mathcal{X}} M_k - K_{\mathcal{X}} M_k)) \end{aligned}$$

and hence $B_{k+1} = Y^\top N_{k+1}$ with

$$N_{k+1} = (1-\lambda\nu)N_k - \nu(N_k M_k^\top K_{\mathcal{X}}^\top K_{\mathcal{X}} M_k - K_{\mathcal{X}} M_k). \tag{22}$$

Then, denote by M and N the limits of M_k and N_k. Given a new x, the estimator is

$$\hat{g}_k(x) = x X^\top M_k N_k^\top Y.$$

Expanding the product we can rewrite

$$\hat{g}_k(x) = \sum_{i=1}^n \alpha_i^{\mathsf{tn}}(x) \psi(y_i), \qquad \alpha^{\mathsf{tn}}(x) = N_k M_k^\top X x^\top = N_k M_k^\top v_x,$$

where $v_x = Xx^\top \in \mathbb{R}^n$. Let d be the decoding map defined by

$$\mathsf{d}(h) = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle \psi(y), Vh \rangle.$$

Then

$$\hat{f}_k(x) = \mathsf{d} \circ \hat{g}_k(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^n \alpha_i^{\mathsf{tn}}(x) \langle \psi(y), V\psi(y_i) \rangle = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^n \alpha_i^{\mathsf{tn}}(x) \ell(y, y_i).$$

Note that in order to obtain the estimator $\hat{g}_k$, only the access to M_k and N_k is needed. Also, examining the updates for M_k and N_k outlined in (21) and (22) we note that the data are accessed through $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ only, which are kernels on input and output respectively. This leads to a direct extension of the argument in the infinite dimensional setting, where the RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ on input and output spaces are infinite dimensional Hilbert spaces. □

A.2. Loss Trick in the Multitask Setting

We now turn to the **multitask** case.

We recall the surrogate problem with trace norm regularization, i.e.

$$\min_{G \in \mathbb{R}^T \otimes \mathcal{H}_{\mathcal{X}}} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \|G_t \phi(x_{it}) - \psi(y_{it})\|^2 + \lambda \|G\|_*. \tag{23}$$

Proposition 6. *Let $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel with associated RKHS $\mathcal{H}_x$. Let $\hat{g} = \hat{G}\phi(\cdot)$ be the solution of problem (23), denote by $\hat{g}_t$, $t = 1, \ldots, T$ its components. Then the loss trick applies to this setting, i.e. the estimator $\hat{f} = \mathsf{d} \circ \hat{g}$ with d_T as in Eq. (16), is equivalently written as*

$$\hat{f}(x) = \operatorname*{argmin}_{c \in \mathcal{C}} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \alpha_{it}^{\mathsf{tn}}(x) \ell(c_t, y_{it}), \tag{24}$$

for some coefficients α_{it} which are derived in the proof below.

Proof. Assume $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H}_y = \mathbb{R}^T$ for the sake of clarity, so that $G\phi(x) = Gx$. For any $t = 1, \ldots, T$, let $\{(x_{it}, y_{it})\}_{i=1}^{n_t}$, be the training set for the t^{th} task.

Denote by $X \in \mathbb{R}^{n \times d}$ the matrix containing the training inputs x_{it}, and by $Y \in \mathbb{R}^{n \times T}$ the matrix whose rows are $\psi(y_{it})$; denote by X_t the $n_t \times d$ matrix containing training inputs of the t^{th} task and by Y_t the $n_t \times 1$ vector with entries $\psi(y_{it})$ $i = 1, \ldots, n_t$. We rewrite (23) using the variational form of the trace norm:

$$\min_{A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{T \times r}} \|Q \odot (Y - XAB^\top)\|^2 + \lambda(\|A\|_{\mathsf{HS}}^2 + \|B\|_{\mathsf{HS}}^2), \tag{25}$$

where $r \in \mathbb{N}$ is now a hyperparameter and Q is a mask which contains zeros in correspondence of missing data. The expression above is also equivalent to

$$\min_{A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{T \times r}} \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{n_t} \|X_t A B_t - Y_t\|^2 + \lambda(\|B_t\|_{\mathsf{HS}}^2 + \|A\|_{\mathsf{HS}}^2) \right),$$

where B_t denotes the t^{th} row of B, i.e. B_t is a $1 \times r$ vector. Thanks to this split, we can update B by updating its rows separately, via (we omit factors 2 which would come from derivatives)

$$B_{t,k+1} = B_{t,k} - \nu\big(n_t^{-1}(B_{t,k} A_k^\top X_t^\top - Y_t^\top) X_t A_k + \lambda B_{t,k}\big).$$

Initialising $B_{t,0} = Y_t^\top N_{t,0}$ for some matrix $N_{t,0} \in \mathbb{R}^{n_t \times r}$, gradient descent updates preserve the structure, and for each k, $B_{t,k} = Y_t^\top N_{t,k}$. Indeed,

$$\begin{aligned}
B_{t,k+1} &= B_{t,k} - \nu\big(n_t^{-1}(B_{t,k} A_k^\top X_t^\top - Y_t^\top) X_t A_k + \lambda B_{t,k}\big) \\
&= Y_t^\top N_{t,k} - \nu\big(n_t^{-1}(Y_t^\top N_{t,k} A_k^\top X_t^\top - Y_t^\top) X_t A_k + \lambda Y_t^\top N_{t,k}\big) \\
&= Y_t^\top \big((1 - \nu\lambda) N_{t,k} - \nu n_t^{-1}(N_{t,k} A_k^\top X_t^\top X_t A_k - X_t A_k)\big) \\
&= Y_t^\top N_{t,k+1}
\end{aligned}$$

where

$$N_{t,k+1} = (1 - \nu\lambda) N_{t,k} - \nu n_t^{-1}(N_{t,k} A_k^\top X_t^\top X_t A_k - X_t A_k).$$

Let us now focus on updates of A, and then combine the two. Set

$$\mathcal{L}(A, B) := \|Q \odot (Y - XAB^\top)\|^2 + \lambda \Bigg(\|A\|_{\mathsf{HS}}^2 + \|B\|_{\mathsf{HS}}^2 \Bigg).$$

Note that the gradient with respect to A reads as $\nabla_A \mathcal{L}(A,B) = X^\top(Q \odot (XAB^\top - Y))B + \lambda A$. Hence, initialising $A_0 = X^\top M_0$, each iterate A_k has the form $X^\top M_k$ and it is possible to perform updates on M_k only as in the proof of Thm. 2, via

$$M_{k+1} = (1 - \lambda\nu)M_k - \big((Q \odot (XX^\top M_k B_k^\top - Y)B_k\big).$$

Let us analyse the term $(Q \odot (XX^\top M_k B_k^\top - Y))B_k$: leveraging the structure of the mask,

$$(Q \odot (XX^\top M_k B_k^\top - Y))B_k = [S_1^\top, \dots, S_T^\top]^\top \tag{26}$$

where S_t is a $n_t \times r$ matrix equal to

$$S_t = X_t X^\top M_k B_{t,k}^\top B_{t,k} = X_t X^\top M_k N_{t,k}^\top Y_t Y_t^\top N_{t,k}.$$

At convergence, we will have $A = X^\top M$ and $B_t = Y_t^\top N_t$ for $t = 1, \dots, T$. Hence, the t^{th} component of the estimator is given by

$$\hat{g}_t(x) = xAB_t^\top = xX^\top M N_t^\top Y_t = \sum_{i=1}^{n_t} \alpha_{it}^{\mathsf{tn}}(x)\psi(y_{it}), \qquad \alpha_t^{\mathsf{tn}}(x) = N_t M^\top X x^\top.$$

Then, the estimator $\hat{f}_N$, with $N = (n_1, \dots, n_T)$ is given by

$$\hat{f}_N(x) = \operatorname*{argmin}_{c \in \mathcal{C}} \sum_{t=1}^{T} \langle c_t, V\hat{g}_t(x)\rangle = \operatorname*{argmin}_{c \in \mathcal{C}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} \alpha_{it}^{\mathsf{tn}}(x)\langle c_t, V\psi(y_{it})\rangle = \operatorname*{argmin}_{c \in \mathcal{C}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} \alpha_{it}^{\mathsf{tn}}(x)\ell(c_t, y_{it}),$$

and hence the loss trick holds. □

A.3. Remark on the lack of loss trick for regularizers via positive semidefinite operator

Assume $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H}_y = \mathbb{R}^T$ and let Y be the $n \times T$ matrix containing $\psi(y_i)$ in its rows. Given $A \in \mathbb{R}^{T \times T}$ symmetric positive definite, the surrogate problem with regularizer $\mathrm{tr}(GAG^\top)$ reads as

$$\frac{1}{n}\|Y - XG\|^2 + \lambda \mathrm{tr}(GAG^\top).$$

We omit the factor $1/n$ as it is does not affect what follows. The problem above has the following solution (see for instance (Alvarez et al., 2012))

$$\mathsf{vec}(G) = (I \otimes X^\top X + \lambda A \otimes I)^{-1}(I \otimes X^\top)\mathsf{vec}(Y).$$

This can be rewritten as

$$\begin{aligned}\mathsf{vec}(G) &= (A^{-1/2} \otimes I)(A^{-1} \otimes X^\top X + \lambda I)^{-1}(A^{-1/2} \otimes X^\top)\mathsf{vec}(Y)\\ &= (A^{-1} \otimes X^\top)(A^{-1} \otimes K + \lambda I)^{-1}\mathsf{vec}(Y),\end{aligned}$$

where $K = XX^\top$ is the kernel matrix. Setting $\mathsf{vec}(M(Y)) = (A^{-1} \otimes K + \lambda I)^{-1}\mathsf{vec}(Y)$,

$$\mathsf{vec}(G) = (A^{-1} \otimes X^\top)\mathsf{vec}(M(Y)) = \mathsf{vec}(X^\top M(Y)A^{-\top}) = \mathsf{vec}(X^\top M(Y)A^{-1}),$$

since A is symmetric. Then $G = X^\top M(Y)A^{-1}$. The decoding procedure yields

$$\hat{f}(x) = \mathsf{d}(\hat{g}(x)) = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle Y, V\hat{g}(x)\rangle = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle Y, VA^{-1}M(Y)^\top v_x\rangle,$$

and due to the product VA^{-1} we cannot retrieve the loss function, i.e. the loss trick.

Now, let us distinguish the following cases

1. $\mathcal{Y}$ has finite cardinality;
2. $\mathcal{Y}$ has not finite cardinality, $\mathcal{H}_y$ is infinite dimensional or ψ and V are unknown.

In the *first* case, let us set $\mathsf{N} = \{1, \dots, |\mathcal{Y}|\}$ and $\mathcal{H}_y = \mathbb{R}^{|\mathcal{Y}|}$. Let $q : \mathcal{Y} \to \mathsf{N}$ be a one-to-one function and for $y \in \mathcal{Y}$ set $Y = e_{q(y)}$ where e_i denoted the i^{th} element of the canonical basis of $\mathbb{R}^{|\mathcal{Y}|}$. Also, set $V \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ the matrix with entries $V_{ij} = \ell(q^{-1}(i), q^{-1}(j))$. Then, since A is a known matrix, ψ and V are defined as above, the estimator $\hat{f}$ can be retrieved despite the lack of loss trick.

In the *second* case, it is not clear how to manage the operation VA^{-1} since V is unknown and also, both V and A are bounded operators from an infinite dimensional space to itself. While in the standard SELF framework, the infinite dimensionality is hidden in the loss trick, and there is no need to explicitly deal with infinite dimensional objects, here it appears to be necessary due to the action of A.

B. Theoretical Analysis

Theorem 3. *Under Asm. 2, let $\mathcal{Y}$ be a compact set, let $(x_i, y_i)_{i=1}^n$ be a set of n points sampled i.i.d. and let $\hat{g}(\cdot) = \hat{G}\phi(\cdot)$ with $\hat{G}$ the solution of Eq. (13) for $\gamma = \|G_*\|_*$. Then, for any $\delta > 0$*

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \ \le \ (\mathsf{m}_y + \mathsf{M})\sqrt{\frac{4\log\frac{\mathsf{r}}{\delta}}{n}} + O(n^{-1}), \tag{14}$$

with probability at least $1-\delta$, where

$$\mathsf{M} = 2\mathsf{m}_{\mathcal{X}}\|C\|_{\mathrm{op}}^{1/2}\|G_*\|_*^2 + \mathsf{m}_{\mathcal{X}}\mathcal{R}(g_*)\|G_*\|_*, \tag{15}$$

with r a constant not depending on δ, n or G_*.

Proof. We split the error as follows:

$$\begin{aligned}\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \le & \mathcal{R}(\hat{g}) - \hat{\mathcal{R}}(\hat{g}) + \hat{\mathcal{R}}(\hat{g}) - \hat{\mathcal{R}}(g_{\gamma *}) \\ & + \hat{\mathcal{R}}(g_{\gamma *}) - \mathcal{R}(g_{\gamma *}) + \mathcal{R}(g_{\gamma *}) - \mathcal{R}(g_*).\end{aligned}$$

Now, by definition of $\hat{g}$ the term $\hat{\mathcal{R}}(\hat{g}) - \hat{\mathcal{R}}(g_{\gamma *})$ is negative. Also, denoting by $\rho_{t|\mathcal{X}}$ the marginal on $\mathcal{X}$ of the probability measure ρ_t,

$$\mathcal{R}(g_{\gamma *}) - \mathcal{R}(g_*) = \int_{\mathcal{X}} \|G_{\gamma *}\phi(x) - G_*\phi(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2 \, d\rho_{\mathcal{X}}(x) = \inf_{\{G \in \mathcal{G}_\gamma\}} \|G\phi(x) - G_*\phi(x)\|_{L^2(\rho_{\mathcal{X}})}^2 \tag{27}$$

$$\le \|(\frac{\gamma}{\|G_*\|_*})G_*\phi(x) - G_*\phi(x)\|_{L^2(\rho_{\mathcal{X}})}^2 \le \left(1 - \frac{\gamma}{\|G_*\|_*}\right)^2 \|G_*\phi\|_{L^2(\rho_{\mathcal{X}})}^2 \tag{28}$$

$$\le \left(1 - \frac{\gamma}{\|G_*\|_*}\right)^2 \mathsf{m}_{\mathcal{X}}\|G_*\|_{\mathsf{HS}}^2 \le (\|G_*\|_* - \gamma)^2\mathsf{m}_{\mathcal{X}}^2. \tag{29}$$

It remains to bound $R_1 := \mathcal{R}(\hat{g}) - \hat{\mathcal{R}}(\hat{g})$ and $R_2 := \hat{\mathcal{R}}(g_{\gamma *}) - \mathcal{R}(g_{\gamma *})$. Since

$$R_1 + R_2 \le 2 \sup_{G \in \mathcal{G}_\gamma} |\hat{\mathcal{R}}(G) - \mathcal{R}(G)|,$$

we just have to bound the term on the right hand side.

Denote

$$C = \mathbb{E}\phi(x) \otimes \phi(x), \qquad \hat{C} = \frac{1}{n}\sum_{i=1}^n \phi(x_i) \otimes \phi(x_i)$$

$$Z = \mathbb{E}\psi(y) \otimes \phi(x), \qquad \hat{Z} = \frac{1}{n}\sum_{i=1}^n \psi(y_i) \otimes \phi(x_i).$$

For any operator G in $\mathcal{H}_y \otimes \mathcal{H}_x$, we have

$$\begin{aligned}
|\hat{\mathcal{R}}(G) - \mathcal{R}(G)| &= \left| \frac{1}{n}\sum_{i=1}^n \|\psi(y_i) - G\phi(x_i)\|_{\mathcal{H}}^2 - \mathbb{E}\|\psi(y) - G\phi(x)\|_{\mathcal{H}}^2 \right| \\
&= \left| \frac{1}{n}\sum_{i=1}^n \left(\langle G^*G, \phi(x_i)\otimes\phi(x_i)\rangle_{\mathsf{HS}} - 2\,\langle G, \psi(y_i)\otimes\phi(x_i)\rangle_{\mathsf{HS}} + \|\psi(y_i)\|_{\mathcal{H}_y}^2 \right) \right. \\
&\qquad \left. - \mathbb{E}\left(\langle G^*G, \phi(x)\otimes\phi(x)\rangle_{\mathsf{HS}} - 2\,\langle G, \psi(y)\otimes\phi(x)\rangle_{\mathsf{HS}} + \|\psi(y)\|_{\mathcal{H}_y}^2 \right) \right| \\
&= \left| \left\langle G^*G, \hat{C} - C \right\rangle_{\mathsf{HS}} - 2\left\langle G, \hat{Z} - Z\right\rangle_{\mathsf{HS}} + \frac{1}{n}\sum_{i=1}^n \|\psi(y_i)\|_{\mathcal{H}_y}^2 - \mathbb{E}\|\psi(y)\|_{\mathcal{H}_y}^2 \right| \\
&\leq \|G\|_{\mathsf{HS}}^2 \|C - \hat{C}\|_{\rm op} + 2\|G\|_* \|Z - \hat{Z}\|_{\rm op} + \left| \mathbb{E}\|\psi(y)\|_{\mathcal{H}_y}^2 - \frac{1}{n}\sum_{i=1}^n \|\psi(y_i)\|_{\mathcal{H}_y}^2 \right|.
\end{aligned}$$

In the last inequality we used that $\|G^*G\|_* = \|G^*G\|_{\mathsf{HS}} = \|G\|_{\mathsf{HS}}^2$ in the first part. In the following we bound $\|C - \hat{C}\|_{\rm op}$ and $\|Z - \hat{Z}\|_{\rm op}$, in two different steps.

STEP 1 Let us start with $\|C - \hat{C}\|_{\rm op}$. We leverage the result in (Minsker, 2017) on Bernstein's inequality for self adjoint operators, which are recalled in Lemma 11 below. Let us set

$$X_i := (\phi(x_i) \otimes \phi(x_i) - C)/n$$

and note that $\mathbb{E}(X_i) = 0$. Also, resolving the square we have that

$$\mathbb{E}(X_i^2) = \frac{1}{n^2}\mathbb{E}(\langle \phi(x_i), \phi(x_i)\rangle \phi(x_i)\otimes\phi(x_i) - 2\phi(x_i)\otimes\phi(x_i)C + C^2) = \frac{1}{n^2}\mathbb{E}(\mathsf{m}_{\mathcal{X}}^2 \phi(x_i)\otimes\phi(x_i)) - C^2,$$

and hence (we assume $\mathsf{m}_{\mathcal{X}} \geq 1$)

$$\|\sum_{i=1}^n \mathbb{E}X_i^2\| \leq \frac{1}{n}(\mathsf{m}_{\mathcal{X}}^2 \|C\|_{\rm op} + \|C\|_{\rm op}^2) \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}\|C\|_{\rm op} =: \sigma^2,$$

Since $\|\phi(x_i)\| \leq \mathsf{m}_{\mathcal{X}}$ for any $i = 1, \dots, n$, we get

$$\|X_i\| \leq \frac{\mathsf{m}_{\mathcal{X}}^2 + \|C\|_{\rm op}}{n} \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n} := U.$$

Set

$$\bar{r}_1 := \frac{\mathrm{tr}\left(\sum_{i=1}^n \mathbb{E}X_i^2\right)}{\|\sum_{i=1}^n \mathbb{E}X_i^2\|_{\rm op}}.$$

Note that the quantity above is the effective rank of $\sum_{i=1}^n \mathbb{E}X_i^2$. With σ^2 and U as above, Lemma 11 yields

$$\|C - \hat{C}\|_{\rm op} \leq \frac{4}{n}\left(\frac{\mathsf{m}_{\mathcal{X}}^2}{3}\ln\left(\frac{14\bar{r}_1}{\delta}\right)\right) + \sqrt{\frac{4\mathsf{m}_{\mathcal{X}}^2\|C\|_{\rm op}}{n}\ln\left(\frac{14\bar{r}_1}{\delta}\right)}$$

with probability greater or equal to $1 - \delta$.

STEP 2 As for $\|Z - \hat{Z}\|_{\mathrm{op}}$ we proceed in a similar way: let $X_i := (\psi(y_i) \otimes \phi(x_i) - Z)/n$. Then,

$$\|X_i\| \leq \frac{\mathsf{m}_{\mathcal{Y}}\mathsf{m}_{\mathcal{X}} + \|Z\|_{\mathrm{op}}}{n} \leq \frac{2\mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y}}}{n}.$$

Also,

$$\begin{aligned} \mathbb{E}X_i^* X_i &= \frac{1}{n^2}\mathbb{E}[(\phi(x_i) \otimes \psi(y_i) - Z^*)(\psi(y_i) \otimes \phi(x_i) - Z)] & (30)\\ &= \frac{1}{n^2}\left(\mathbb{E}(\langle\psi(y_i), \psi(y_i)\rangle \phi(x_i) \otimes \phi(x_i)) - Z^* Z\right) \preceq \frac{2}{n^2}\mathbb{E}(\langle\psi(y_i), \psi(y_i)\rangle \phi(x_i) \otimes \phi(x_i)). & (31) \end{aligned}$$

Then

$$\|\sum_{i=1}^n \mathbb{E}X_i^* X_i\|_{\mathrm{op}} \leq \frac{2}{n}\|\mathbb{E}(\langle\psi(y), \psi(y)\rangle \phi(x) \otimes \phi(x))\|_{\mathrm{op}}.$$

Applying Lemma 14, we obtain

$$\|\sum_{i=1}^n \mathbb{E}X_i^* X_i\|_{\mathrm{op}} \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}(\|G_*\|_{\mathsf{HS}}^2\|C\|_{\mathrm{op}} + \mathcal{R}(g_*)).$$

Similarly,

$$\begin{aligned} \mathbb{E}X_i X_i^* &= \frac{1}{n^2}\mathbb{E}[(\psi(y_i) \otimes \phi(x_i) - Z)(\phi(x_i) \otimes \psi(y_i) - Z^*)]\\ &= \frac{1}{n^2}\left(\mathbb{E}(\langle\phi(x_i), \phi(x_i)\rangle \psi y_i) \otimes \psi(y_i)) - ZZ^*\right) \preceq \frac{2}{n^2}\mathbb{E}(\langle\phi(x_i), \phi(x_i)\rangle \psi(y_i) \otimes \psi(y_i))\\ &\preceq \frac{2}{n^2}\mathsf{m}_{\mathcal{X}}^2\mathbb{E}(\psi(y_i) \otimes \psi(y_i)) \end{aligned}$$

and

$$\|\sum_{i=1}^n \mathbb{E}X_i X_i^*\|_{\mathrm{op}} \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}\|\mathbb{E}(\psi(y) \otimes \psi(y))\|_{\mathrm{op}}.$$

Applying Lemma 13, we conclude

$$\|\sum_{i=1}^n \mathbb{E}X_i X_i^*\|_{\mathrm{op}} \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}(\|G_*\|_{\mathsf{HS}}^2\|C\|_{\mathrm{op}} + \mathcal{R}(g_*)).$$

Hence both $\|\sum_{i=1}^n \mathbb{E}X_i X_i^*\|_{\mathrm{op}}$ and $\|\sum_{i=1}^n \mathbb{E}X_i^* X_i\|_{\mathrm{op}}$ are bounded by $\frac{2\mathsf{m}_{\mathcal{X}}^2}{n}(\|G_*\|_{\mathsf{HS}}^2\|C\|_{\mathrm{op}} + \mathcal{R}(g_*))$. Moreover, let

$$\bar{r}_2 = \max\left(\frac{\mathrm{tr}(\sum_{i=1}^n \mathbb{E}X_i X_i^*)}{\|\sum_{i=1}^n \mathbb{E}X_i X_i^*\|_{\mathrm{op}}}, \frac{\mathrm{tr}(\sum_{i=1}^n \mathbb{E}X_i^* X_i)}{\|\sum_{i=1}^n \mathbb{E}X_i^* X_i\|_{\mathrm{op}}}\right),$$

which corresponds to the maximum between effective ranks of $\sum_{i=1}^n \mathbb{E}X_i X_i^*$ and $\sum_{i=1}^n \mathbb{E}X_i^* X_i$. Bernstein's inequality shown in (Minsker, 2017) (and recalled in Lemma 12) gives

$$\|Z - \hat{Z}\|_{\mathrm{op}} \leq \frac{4}{n}\left(\frac{\mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y}}}{3}\ln\left(\frac{28\bar{r}_2}{\delta}\right)\right) + \sqrt{\frac{2\mathsf{m}_{\mathcal{X}}^2(\|G_*\|_{\mathsf{HS}}^2\|C\|_{\mathrm{op}} + \mathcal{R}(g_*))}{n}\ln\left(\frac{28\bar{r}_2}{\delta}\right)}$$

with probability greater or equal to $1-\delta$. Splitting the second term we see that

$$\|Z-\hat{Z}\|_{\mathrm{op}} \leq \frac{4}{n}\Big(\frac{\mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y}}}{3}\ln\Big(\frac{28\bar{r}_2}{\delta}\Big)\Big) + \Big(\|G\|_{\mathsf{HS}}\|C\|_{\mathrm{op}}^{\frac{1}{2}}\mathsf{m}_{\mathcal{X}} + \mathsf{m}_{\mathcal{X}}\sqrt{\mathcal{R}(g_*)}\Big)\sqrt{\frac{2}{n}\ln\Big(\frac{28\bar{r}_2}{\delta}\Big)}.$$

STEP 3. Finally, by Hoeffding inequality

$$\left|\mathbb{E}\|\psi(y)\|_{\mathcal{H}_{\mathcal{Y}}}^2 - \frac{1}{n}\sum_{i=1}^{n}\|\psi(y_i)\|_{\mathcal{H}_{\mathcal{Y}}}^2\right| \leq \mathsf{m}_{\mathcal{Y}}\sqrt{\ln\Big(\frac{2}{\delta}\Big)\frac{1}{n}}$$

with probability at least $1-\delta$.

FINAL STEP. We have now all the bounds that we need. By taking $\mathsf{r} = \max(\bar{r}_1, \bar{r}_2)$ and performing an intersection bound on the three parts we conclude

$$|\hat{\mathcal{R}}(G) - \mathcal{R}(G)| \leq \gamma^2\Big(\frac{A}{n} + \frac{B}{\sqrt{n}}\Big) + \gamma\Big(\frac{A'}{n} + \frac{B'}{\sqrt{n}}\Big) + \mathsf{m}_{\mathcal{Y}}\sqrt{\ln\Big(\frac{2}{\delta}\Big)\frac{1}{n}} \tag{32}$$

with probability greater or equal than $1-3\delta$, with

$$A = 4\ln\Big(\frac{28\mathsf{r}}{\delta}\Big)\frac{\mathsf{m}_{\mathcal{X}}^2}{3}, \qquad B = (2+\sqrt{2})\mathsf{m}_{\mathcal{X}}\|C\|_{\mathrm{op}}^{\frac{1}{2}}\sqrt{\ln\Big(\frac{28\mathsf{r}}{\delta}\Big)}$$
$$A' = 4\ln\Big(\frac{28\mathsf{r}}{\delta}\Big)\frac{\mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y}}}{3}, \qquad B' = \mathsf{m}_{\mathcal{X}}\sqrt{2\mathcal{R}(g_*)}\sqrt{\ln\Big(\frac{28\mathsf{r}}{\delta}\Big)}.$$

Combining with the approximation error in Eq. (27), we obtain

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq \gamma^2\Big(\frac{A}{n} + \frac{B}{\sqrt{n}}\Big) + \gamma\Big(\frac{A'}{n} + \frac{B'}{\sqrt{n}}\Big) + \sqrt{\ln\Big(\frac{2}{\delta}\Big)\frac{\mathsf{m}_{\mathcal{Y}}^2}{n}} + (\|G_*\|_* - \gamma)^2\mathsf{m}_{\mathcal{X}}^2.$$

In principle, starting from the bound above we should optimize with respect to γ to find the optimal value, which will be between 0 and $\|G_*\|_*$. Here we consider the simpler case where $\gamma = \|G_*\|_*$. Isolating the faster terms, the bound above becomes

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq \frac{\|G_*\|_*^2}{\sqrt{n}}\Big(\mathsf{m}_{\mathcal{X}}\|C\|_{\mathrm{op}}^{\frac{1}{2}}(2+\sqrt{2}) + \|G_*\|_*\mathsf{m}_{\mathcal{X}}\sqrt{2\mathcal{R}(g_*)}\Big)\sqrt{\ln\Big(\frac{28\mathsf{r}}{\delta}\Big)} + \mathsf{m}_{\mathcal{Y}}\sqrt{\ln\Big(\frac{2}{\delta}\Big)\frac{1}{n}}$$

Rearranging we get

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq \frac{\|G_*\|_*}{\sqrt{n}}\Big[\Big((\sqrt{2}+1)\mathsf{m}_{\mathcal{X}}\|G_*\|_*\|C\|_{\mathrm{op}}^{\frac{1}{2}} + \mathsf{m}_{\mathcal{X}}\sqrt{\mathcal{R}(g_*)}\Big)\sqrt{2\ln\Big(\frac{28\mathsf{r}}{\delta}\Big)}\Big] + \mathsf{m}_{\mathcal{Y}}\sqrt{\ln\Big(\frac{2}{\delta}\Big)\frac{1}{n}} \tag{33}$$

with probability greater or equal to $1-3\delta$. Bounding $\ln\big(\frac{2}{\delta}\big)$ with $\ln\big(\frac{28\mathsf{r}}{\delta}\big)$ we get

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq (\mathsf{M} + \mathsf{m}_{\mathcal{Y}})\sqrt{\frac{\ln(\frac{\mathsf{r}}{\delta})}{n}} + O(n^{-1})$$

where $\mathsf{M} = (2+\sqrt{2})\mathsf{m}_{\mathcal{X}}\|G_*\|_*^2\|C\|_{\mathrm{op}}^{\frac{1}{2}} + \sqrt{2}\|G_*\|_*\mathsf{m}_{\mathcal{X}}\mathcal{R}(g_*)$. In the main body of the paper we bound it as

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq (\mathsf{M} + \mathsf{m}_{\mathcal{Y}})\sqrt{\frac{4\ln(\frac{\mathsf{r}}{\delta})}{n}} + O(n^{-1})$$

with $\mathsf{M} = 2\mathsf{m}_{\mathcal{X}}\|G_*\|_*^2\|C\|_{\mathrm{op}}^{\frac{1}{2}} + \|G_*\|_*\mathsf{m}_{\mathcal{X}}\mathcal{R}(g_*)$ to make it neater. □

Comparison with Hilbert-Schmidt regularization. The goal of this remark is a comparison between the constants in the bound for the trace norm estimator and in the bound we would obtain with Hilbert-Schmidt estimator.

Bound for HS-regularization. We show here the bound obtained with Hilbert-Schmidt regularization. In this case, $\mathcal{G}_\gamma := \{g(\cdot) = G\phi(\cdot) \mid \|G\|_{\mathsf{HS}} \le \gamma\}$. Note that if G is a Hilbert-Schmidt operator, then G^*G is a trace norm operator. Therefore, the term $\left\langle G^*G, \hat{C} - C \right\rangle_{\mathsf{HS}}$ can be bounded as before:

$$\|C - \hat{C}\|_{\text{op}} \le \frac{4}{n}\left(\frac{\mathsf{m}_{\mathcal{X}}^2}{3} \ln\left(\frac{14\bar{r}_1}{\delta}\right)\right) + \sqrt{\frac{4\mathsf{m}_{\mathcal{X}}^2\|C\|_{\text{op}}}{n} \ln\left(\frac{14\bar{r}_1}{\delta}\right)}$$

On the other hand, for the second term $\left\langle G, \hat{Z} - Z \right\rangle_{\mathsf{HS}}$ we have

$$\left|\left\langle G, \hat{Z} - Z \right\rangle_{\mathsf{HS}}\right| \le \|G\|_{\mathsf{HS}}\|\hat{Z} - Z\|_{\mathsf{HS}}.$$

Now, in order to bound $\|\hat{Z} - Z\|_{\mathsf{HS}}$, we note that $\|Z\|_{\mathsf{HS}}^2 \le \mathsf{m}_{\mathcal{X}}^2\mathbb{E}\|\psi(y)\|_{\mathsf{HS}}^2 = \mathsf{m}_{\mathcal{X}}^2\text{tr}(C_Y)$. Proceeding in a similar way as in Lemma 13, we obtain that $\text{tr}(C_Y) \le \mathcal{R}(g_*) + \mathsf{m}_{\mathcal{X}}^2\|G_*\|_{\mathsf{HS}}$ and hence $\|Z\|_{\mathsf{HS}}^2 \le \mathsf{m}_{\mathcal{X}}^2\mathcal{R}(g_*) + \mathsf{m}_{\mathcal{X}}^4\|G_*\|_{\mathsf{HS}}^2$. From Lemma 2 in (Smale & Zhou, 2007),

$$\|\hat{Z} - Z\|_{\mathsf{HS}} \le \sqrt{\frac{2(\mathsf{m}_{\mathcal{X}}^2\mathcal{R}(g_*) + \mathsf{m}_{\mathcal{X}}^4\|G_*\|_{\mathsf{HS}}^2)}{n}}\sqrt{\ln\left(\frac{2}{\delta}\right)} + O(n^{-1}).$$

Finally, no difference holds for the last term $\left|\mathbb{E}\|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}^2 - \frac{1}{n}\sum_{i=1}^n \|\psi(y_i)\|_{\mathcal{H}_\mathcal{Y}}^2\right|$. Hence, combining the three parts and bounding $\ln(\frac{2}{\delta})$ with $\ln(\frac{14\mathsf{r}}{\delta})$, we get

$$\mathcal{R}(\hat{g}_{\mathsf{HS}}) - \mathcal{R}(g_*) \le \frac{\|G_*\|_{\mathsf{HS}}}{\sqrt{n}}\left(\|G_*\|_{\mathsf{HS}}\|C\|_{\text{op}}^{\frac{1}{2}}2\kappa_x + \sqrt{2}\mathsf{m}_{\mathcal{X}}^2\|G_*\|_{\mathsf{HS}} + \mathsf{m}_{\mathcal{X}}\sqrt{2\mathcal{R}(g_*)} + \mathsf{m}_{\mathcal{Y}}\right)\sqrt{\ln\left(\frac{14\mathsf{r}}{\delta}\right)} + O(n^{-1}). \quad (34)$$

Note that this bound slightly refines the excess risk bounds for HS regularization provided in (Ciliberto et al., 2016).

Comparison and discussion. Let us compare the bound with HS regularization in Eq. (34) with the bound for the trace norm estimator that we derived in the proof of Thm. 3:

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \le \frac{\|G_*\|_*}{\sqrt{n}}\left(\|G_*\|_*\|C\|_{\text{op}}^{\frac{1}{2}}\mathsf{m}_{\mathcal{X}} + \mathsf{m}_{\mathcal{X}}\sqrt{2\mathcal{R}(g_*)} + \mathsf{m}_{\mathcal{Y}}\right)\sqrt{\ln\left(\frac{28\bar{r}}{\delta}\right)} + O(n^{-1}).$$

To make the comparison easier, we isolate the constants in the bounds:

$$\text{HS: } 2\mathsf{m}_{\mathcal{X}}\|G_*\|_{\mathsf{HS}}^2\|C\|_{\text{op}}^{\frac{1}{2}} + \underline{\sqrt{2}\mathsf{m}_{\mathcal{X}}\|G\|_{\mathsf{HS}}^2} + \mathsf{m}_{\mathcal{X}}\mathcal{R}(g_*)\|G_*\|_{\mathsf{HS}} + \mathsf{m}_{\mathcal{Y}} \quad \text{versus}$$
$$\text{TN: } (2 + \sqrt{2})\mathsf{m}_{\mathcal{X}}\|G_*\|_*^2\|C\|_{\text{op}}^{\frac{1}{2}} + \mathsf{m}_{\mathcal{X}}\mathcal{R}(g_*)\|G_*\|_* + \mathsf{m}_{\mathcal{Y}}.$$

We can summarize the cases as below.

- If $\|G_*\|_{\mathsf{HS}} \ll \|G_*\|_*$, then the TN bound gives no advantage over the HS one.

- Whenever $\|G_*\|_{\mathsf{HS}}$ and $\|G_*\|_*$ are of the same order, our result shows an advantage in the constant of the bound: indeed, while in trace norm case, the norm $\|G_*\|_*$ is mitigated by $\|C\|_{\text{op}}^{\frac{1}{2}}$, in the HS case is it not, because of the extra term $\|G_*\|_{\mathsf{HS}}^2\mathsf{m}_{\mathcal{X}}$. Note that $\|C\|_{\text{op}} \le \mathsf{m}_{\mathcal{X}}$ and the gap between the two can be significant: for instance, if C is the covariance operator of a uniform distribution on a d-dimensional unit sphere, $\|C\|_{\text{op}} = 1/d$ while $\mathsf{m}_{\mathcal{X}} = 1$. Hence the entity of the improvement depends on how smaller $\|C\|_{\text{op}}$ is with respect to $\text{tr}(C)$.

The point above holds true when the other quantities $(\mathsf{m}_{\mathcal{Y}}, \|G\|_* \mathcal{R}(G_*))$ in the constant do not dominate. However, this is reasonable to expect:
-$\mathcal{R}(g_*)$ is the minimum expected risk;
-$\mathsf{m}_{\mathcal{X}}$ is 1 whenever we choose a normalized kernel on the input (Gaussian);
-$\mathsf{m}_{\mathcal{Y}}$ is also typically 1: $\mathsf{m}_{\mathcal{Y}}$ is such that $\sup_{y \in \mathcal{Y}} \|k_{\mathcal{Y}}(y, \cdot)\|_{\mathcal{H}_{\mathcal{Y}}} \leq \mathsf{m}_{\mathcal{Y}}^2$ where $k_{\mathcal{Y}}$ is a reproducing kernel on the output. Whenever $\mathcal{Y}$ is finite (and hence $k_{\mathcal{Y}}(y, y') = \delta_{y==y'}$) or the loss is smooth (and hence $k_{\mathcal{Y}}$ is the Abel kernel), $\mathsf{m}_{\mathcal{Y}} = 1$.

C. Theoretical Analysis: Multitask Case

We consider the general multitask learning case which allows a different loss function for each task: the goal is to minimize the *multi-task excess risk*

$$\min_{f:\mathcal{X}\to\mathcal{C}} \mathcal{E}(f), \quad \mathcal{E}(f) = \frac{1}{T}\sum_{t=1}^{T} \int_{\mathcal{X}\times\mathbb{R}} \ell_t(f_t(x), y) d\rho_t(x, y),$$

where ρ_t is an unknown probability distribution on $\mathcal{X} \times \mathbb{R}$ that is observed via finite samples $(x_{it}, y_{it})_{i=1}^{n_t}$, for $t = 1, \dots, T$. Each ℓ_t is required to satisfy the SELF assumption in Def. 1, i.e.

$$\ell_t(y, y') = \langle \psi_t(y), V_t \psi_t(y') \rangle,$$

and for $t = 1, \dots T$ $\mathsf{m}_{\mathcal{Y},t}$ is a constant such that $\sup_{y\in\mathcal{Y}} \|\psi_t(y)\| \leq \mathsf{m}_{\mathcal{Y},t}$. In this setting the surrogate problem corresponds to

$$\min_{G:\mathcal{H}_{\mathcal{X}}\to{\mathcal{H}_{\mathcal{Y}}}^T} \mathcal{R}(G) \qquad \mathcal{R}_T(G) := \frac{1}{T}\int_{\mathcal{X}\times\mathcal{Y}} \|\psi_t(y) - G_t\phi(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2 d\rho_t(x, y),$$

and its solution is denoted with G_*. Note that each G_t is an operator in $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ and G denotes the operator from $\mathcal{H}_{\mathcal{X}}$ to ${\mathcal{H}_{\mathcal{Y}}}^T$ whose t^{th} component is G_t, $t = 1, \dots T$. Formally, $G = \sum_{t=1}^{T} G_t \otimes e_t$, with $(e_t)_{t=1}^{T}$ the canonical basis of $\mathbb{R}^T$. Since $\|G\|_{\mathsf{HS}}^2 = \sum_t \|G_t\|_{\mathsf{HS}}^2$, in case of HS regularization the surrogate problem considers each task t separately.

Here we perform regularization with trace norm of the operator G. Setting $\mathcal{G}_\gamma = \{g(\cdot) = G\phi(\cdot) \mid G : \mathcal{H}_{\mathcal{X}} \to {\mathcal{H}_{\mathcal{Y}}}^T$ is s.t. $\|G\|_* \leq \gamma\}$, we study the estimator $\hat{g}$ given by

$$\hat{g} = \operatorname*{argmin}_{g\in\mathcal{G}_\gamma} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n} \|g_t(x_{it}) - \psi_t(y_{it})\|_{\mathcal{H}_{\mathcal{Y}}}^2. \tag{35}$$

In the following we will consider $n_t = n$ for simplicity and denote $\mathcal{R}_T$ with $\mathcal{R}$, to avoid cumbersome notation. The estimator $\hat{g}$ satisfies the following excess risk bound:

Theorem 7. *For $t = 1, \dots T$, let $(x_{it}, y_{it})_{i=1}^{n}$ be an iid sample of ρ_t and $\hat{g}$ is the solution of Eq. (35) with $\gamma = \|G_*\|_*$.*

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq \frac{1}{\sqrt{nT}}\left(\|G_*\|_*^2 \|\bar{C}\|_{\mathsf{op}}^{\frac{1}{2}} \mathsf{m}_{\mathcal{X}}(2 + \sqrt{2}) + \mathsf{m}_{\mathcal{X}}\|G_*\|_* \sqrt{2\mathcal{R}(G_*)} + \bar{\mathsf{m}}_{\mathcal{Y}}\right)\sqrt{\ln\left(\frac{T\mathsf{r}}{\delta}\right)} + O((nT)^{-1}),$$

with probability greater or equal then $1 - 3\delta$, where $\bar{C}$ is the average covariance operator, $\mathcal{R}(G_*)$ the expected true risk, $\bar{\mathsf{m}}_{\mathcal{Y}} = \sqrt{\frac{1}{T}\sum_t \mathsf{m}_{\mathcal{Y},t}^2}$ and r a number independent of n, T, δ and G_*.

The section is devoted to the proof of this result, which is the formal version of theorem Thm. 5 in the main paper. We split the error as follows:

$$\begin{aligned}\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq& \mathcal{R}(\hat{g}) - \hat{\mathcal{R}}(\hat{g}) + \hat{\mathcal{R}}(\hat{g}) - \hat{\mathcal{R}}(g_{\gamma *}) \\ &+ \hat{\mathcal{R}}(g_{\gamma *}) - \mathcal{R}(g_{\gamma *}) + \mathcal{R}(g_{\gamma *}) - \mathcal{R}(g_*).\end{aligned}$$

Now, by definition of $\hat{g}$ the term $\hat{\mathcal{R}}(\hat{g}) - \hat{\mathcal{R}}(g_{\gamma *})$ is negative. Also, denoting by $\rho_{t|\mathcal{X}}$ the marginal on $\mathcal{X}$ of the probability measure ρ_t,

$$\begin{aligned}\mathcal{R}(g_{\gamma *}) - \mathcal{R}(g_*) &= \frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{X}} \|G_{t\gamma *}\phi(x) - G_{t*}\phi(x)\|_{\mathcal{H}_\mathcal{Y}}^2 \, d\rho_{t|\mathcal{X}}(x) = \inf_{\{G \in \mathcal{G}_\gamma\}} \frac{1}{T}\sum_{t=1}^{T} \|G_t\phi(x) - G_{t*}\phi(x)\|_{L^2(\rho_{t|\mathcal{X}})}^2 \\ &\leq \frac{1}{T}\sum_{t=1}^{T} \|(\frac{\gamma}{\|G_*\|_*})G_{t*}\phi(x) - G_{t*}\phi(x)\|_{L^2(\rho_{t|\mathcal{X}})}^2 \leq \Big(1 - \frac{\gamma}{\|G_*\|_*}\Big)^2 \frac{1}{T}\sum_{t=1}^{T}\|G_{t*}\phi\|_{L^2(\rho_{t|\mathcal{X}})}^2 \\ &\leq \Big(1 - \frac{\gamma}{\|G_*\|_*}\Big)^2 \frac{\mathsf{m}_\mathcal{X}^2}{T}\|G_*\|_{\mathsf{HS}}^2 \leq (\|G_*\|_* - \gamma)^2 \frac{\mathsf{m}_\mathcal{X}^2}{T}.\end{aligned}$$

It remains to bound $R_1 := \mathcal{R}(\hat{g}) - \hat{\mathcal{R}}(\hat{g})$ and $R_2 := \hat{\mathcal{R}}(g_{\gamma *}) - \mathcal{R}(g_{\gamma *})$. Since

$$R_1 + R_2 \leq 2 \sup_{G \in \mathcal{G}_\gamma} |\hat{\mathcal{R}}(G) - \mathcal{R}(G)|,$$

we just have to bound the term on the right hand side. In the following we assume $n_t = n$ for $t = 1, \dots, T$ for clarity. Also, the notation $\mathbb{E}u(x_t) \otimes v(y_t)$ is to be interpreted as $\mathbb{E}_{(x,t)\sim\rho_t} u(x) \otimes v(y)$. For $t = 1, \dots, T$, denote

$$\begin{aligned} C_t &= \mathbb{E}\phi(x_t) \otimes \phi(x_t), \qquad & \hat{C}_t &= \frac{1}{n}\sum_{i=1}^{n} \phi(x_{it}) \otimes \phi(x_{it}) \\ Z_t &= \mathbb{E}\psi(y_t) \otimes \phi(x_t), \qquad & \hat{Z}_t &= \frac{1}{n}\sum_{i=1}^{n} \psi(y_{it}) \otimes \phi(x_{it}). \end{aligned}$$

For any operator G, we have

$$\begin{aligned} |\hat{\mathcal{R}}(G) - \mathcal{R}(G)| &= \left| \frac{1}{T}\sum_{t=1}^{n}\frac{1}{n}\sum_{i=1}^{n} \|\psi(y_{it}) - G_t\phi(x_{it})\|_{\mathcal{H}_\mathcal{Y}}^2 - \mathbb{E}\|\psi(y_t) - G_t\phi(x_t)\|_{\mathcal{H}_\mathcal{Y}}^2 \right| \quad (36) \\ &= \left| \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n} \left(\langle G_t^* G_t, \phi(x_{it}) \otimes \phi(x_{it}) \rangle_{\mathsf{HS}} - 2\,\langle G_t, \psi(y_{it}) \otimes \phi(x_{it}) \rangle_{\mathsf{HS}} + \|\psi(y_{it})\|_{\mathcal{H}_\mathcal{Y}}^2 \right) \right. \quad (37) \\ &\qquad \left. - \,\mathbb{E}\left(\langle G_t^* G_t, \phi(x_t) \otimes \phi(x_t) \rangle_{\mathsf{HS}} - 2\,\langle G_t, \psi(y_t) \otimes \phi(x_t) \rangle_{\mathsf{HS}} + \|\psi(y_t)\|_{\mathcal{H}_\mathcal{Y}}^2 \right) \right| \quad (38) \\ &= \left| \frac{1}{T}\sum_{t=1}^{T} \left\langle G_t^* G_t, \hat{C}_t - C_t \right\rangle_{\mathsf{HS}} - 2\left\langle G_t, \hat{Z}_t - Z_t \right\rangle_{\mathsf{HS}} + \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n} \|\psi(y_{it})\|_{\mathcal{H}_\mathcal{Y}}^2 - \mathbb{E}\|\psi(y_t)\|_{\mathcal{H}_\mathcal{Y}}^2 \right| \quad (39) \end{aligned}$$

We analyse each term separately in the following lemmas.

Lemma 8. *The first term in Eq. (39) satisfies the following inequality:*

$$\left| \frac{1}{T}\sum_{t=1}^{T} \left\langle G_t^* G_t, \hat{C}_t - C_t \right\rangle_{\mathsf{HS}} \right| \leq \frac{1}{T}\frac{4\|G\|_{\mathsf{HS}}^2}{n}\Big(\frac{\mathsf{m}_\mathcal{X}^2}{3}\ln\Big(\frac{T\mathsf{r}_1}{\delta}\Big)\Big) + \frac{\|G\|_{\mathsf{HS}}^2}{T}\sqrt{\frac{4\mathsf{m}_\mathcal{X}^2 \max_t \|C_t\|_{\mathrm{op}}}{n}\ln\Big(\frac{T\mathsf{r}_1}{\delta}\Big)}\Big)$$

with probability $1 - \delta$, where r_1 is a constant independent of n, T, G and which is given by the problem.

Proof.

$$\frac{1}{T}\sum_{t=1}^{T} \left\langle G_t^* G_t, \hat{C}_t - C_t \right\rangle_{\mathsf{HS}} = \frac{1}{T}\mathrm{tr}(\mathbf{G}^*\mathbf{C}) \leq \frac{1}{T}\|\mathbf{G}\|_*\|\mathbf{C}_{\mathrm{op}}\|,$$

where $\mathbf{G} = \sum_{t=1}^{T}(e_t \otimes e_t) \otimes (G_t^* G_t)$ and $\mathbf{C} = \sum_{t=1}^{T}(e_t \otimes e_t) \otimes (\hat{C}_t - C_t)$. Now,

$$\|\mathbf{G}\|_* = \sum_{t=1}^{T} \|G_t^* G_t\|_* = \sum_{t=1}^{T} \|G_t G_t^*\|_{\mathsf{HS}} = \|G\|_{\mathsf{HS}}^2$$

and

$$\|\mathbf{C}\|_{\mathrm{op}} = \max_{t=1,\dots,T} \|C_t - \hat{C}_t\|_{\mathrm{op}}.$$

Using Lemma 11, we get

$$\|C_t - \hat{C}_t\|_{\mathrm{op}} \leq \frac{4}{n}\Big(\frac{\mathsf{m}_{\mathcal{X}}^2}{3} \ln\Big(\frac{14\bar{r}_t}{\delta}\Big)\Big) + \sqrt{\frac{4\mathsf{m}_{\mathcal{X}}^2 \|C_t\|_{\mathrm{op}}}{n} \ln\Big(\frac{14\bar{r}_t}{\delta}\Big)\Big)}$$

with probability greater than $1 - \delta$. Performing an intersection bound we have that for $t = 1, \dots, T$

$$\begin{aligned} \max_{t=1,\dots,T} \|C_t - \hat{C}_t\|_{\mathrm{op}} &\leq \frac{4}{n}\Big(\frac{\mathsf{m}_{\mathcal{X}}^2}{3} \ln\Big(\frac{\mathsf{r}_1}{\delta}\Big)\Big) + \sqrt{\frac{4\mathsf{m}_{\mathcal{X}}^2 \|C_t\|_{\mathrm{op}}}{n} \ln\Big(\frac{\mathsf{r}_1}{\delta}\Big)\Big)} \\ &= \frac{4}{n}\Big(\frac{\mathsf{m}_{\mathcal{X}}^2}{3} \ln\Big(\frac{\mathsf{r}_1}{\delta}\Big)\Big) + \sqrt{\frac{4\mathsf{m}_{\mathcal{X}}^2 \max_t \|C_t\|_{\mathrm{op}}}{n} \ln\Big(\frac{\mathsf{r}_1}{\delta}\Big)\Big)} \end{aligned}$$

with probability $1 - T\delta$, where $\mathsf{r}_1 = 14 \max_t \bar{r}_t$. With some abuse of notation take $\delta = \delta/T$ and we get

$$\max_{t=1,\dots,T} \|C_t - \hat{C}_t\|_{\mathrm{op}} \leq \frac{4}{n}\Big(\frac{\mathsf{m}_{\mathcal{X}}^2}{3} \ln\Big(\frac{T\mathsf{r}_1}{\delta}\Big)\Big) + \sqrt{\frac{4\mathsf{m}_{\mathcal{X}}^2 \max_t \|C_t\|_{\mathrm{op}}}{n} \ln\Big(\frac{T\mathsf{r}_1}{\delta}\Big)\Big)}$$

with probability $1 - \delta$. □

Lemma 9.

$$\frac{1}{T}\sum_{t=1}^{T}\Big\langle G_t, \hat{Z}_t - Z_t\Big\rangle \leq \frac{4\|G\|_*}{nT}\Big(\frac{\mathsf{k}_1}{3} \ln\Big(\frac{28\bar{r}}{\delta}\Big)\Big) + \frac{\mathsf{m}_{\mathcal{X}}\|G\|_*}{T}\sqrt{\frac{2T\mathcal{R}(G_*)}{n} \ln\Big(\frac{\mathsf{r}_2}{\delta}\Big)} + \frac{\|G\|_*^2}{T}\mathsf{m}_{\mathcal{X}}\sqrt{\|\sum_t C_t\|_{\mathrm{op}} \frac{2}{n} \ln\Big(\frac{\mathsf{r}_2}{\delta}\Big)}$$

with probability at least $1 - \delta$, where $\mathsf{k}_1 = \mathsf{m}_{\mathcal{X}} \max_t \mathsf{m}_{\mathcal{Y},t}$, and r_2 is independent of G_*, δ, n and T.

Proof. Let us start with the following bound

$$\frac{1}{T}\sum_{t=1}^{T}\Big\langle G_t, \hat{Z}_t - Z_t\Big\rangle = \frac{1}{T}\mathrm{tr}(G\mathbf{Z}) \leq \frac{1}{T}\|G\|_*\|\mathbf{Z}\|_{\mathrm{op}}$$

where $\mathbf{Z} = \sum_{t=1}^{T}(\hat{Z}_t - Z_t) \otimes e_t$. To bound $\|\mathbf{Z}\|_{\mathrm{op}}$ some extra work is needed. We aim to apply Lemma 12 again. Let us define

$$X_{it} = \frac{1}{n}\big(\psi(y_{it}) \otimes \phi(x_{it}) - \mathbb{E}\psi(y_t) \otimes \phi(x_t)\big) \otimes e_t,$$

so that $\sum_{i,t} X_{it} = \mathbf{Z}$. Note that

$$\|X_{it}\|_{\mathrm{op}} \leq \frac{\max_t \mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y},t} + \max_t \|Z_t\|_{\mathrm{op}}}{n} \leq \frac{2\max_t \mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y},t}}{n} \quad \text{for any } i, t.$$

In order to apply Lemma 12 we need to bound the following two quantities

$$\|\sum_{i,t}\mathbb{E}X_{it}X_{it}^*\|, \qquad \|\sum_{i,t}\mathbb{E}X_{it}^*X_{it}\|.$$

Note that

$$\mathbb{E}X_{it}X_{it}^* = \frac{1}{n^2}\big(\mathbb{E}(\psi(y_{it})^2\phi(x_{it})\otimes\phi(x_{it})) - Z_tZ_t^*\big)$$

$$\sum_{it}\mathbb{E}X_{it}X_{it}^* = \frac{1}{n}\sum_t(\frac{1}{n}\sum_{i=1}^n\mathbb{E}(\psi(y_{it})^2\phi(x_{it})\otimes\phi(x_{it})) - Z_tZ_t^*)$$

and hence

$$\|\sum_{it}\mathbb{E}X_{it}X_{it}^*\|_{\mathrm{op}} \leq \frac{2}{n}\|\sum_t\mathbb{E}(\psi(y_t)^2\phi(x_t)\otimes\phi(x_t))\|_{\mathrm{op}}. \tag{40}$$

A direct application of Lemma 16 yields

$$\|\sum_{it}\mathbb{E}X_{it}X_{it}^*\|_{\mathrm{op}} \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}(T\mathcal{R}(g_*) + \|\sum_t C_t\|_{\mathrm{op}}\|G\|_{\mathsf{HS}}^2).$$

Also,

$$X_{it}^*X_{it} = \frac{1}{n^2}(k(x_{it},x_{it})\psi(y_{it})^2 - \|\mathbb{E}(\psi(y_t)\otimes\phi(x_t)\|^2)e_t\otimes e_t$$

and

$$\begin{aligned}\sum_{it}\mathbb{E}X_{it}^*X_{it} &= \frac{1}{n}\sum_t\frac{1}{n}\sum_{i=1}^n(\mathbb{E}k(x_{it},x_{it})\psi(y_{it})^2 - \|\mathbb{E}(\psi(y_t)\otimes\phi(x_t)\|^2)e_t\otimes e_t\\ &\preceq \frac{1}{n}\sum_t(\mathsf{m}_{\mathcal{X}}^2C_{Y,t} - \|\mathbb{E}(\psi(y_t)\otimes\phi(x_t)\|^2)e_t\otimes e_t.\end{aligned}$$

Taking the operator norm, we obtain

$$\|\sum_{it}\mathbb{E}X_{it}^*X_{it}\|_{\mathrm{op}} \leq \frac{2}{n}\max_{t=1,\dots,T}(\mathsf{m}_{\mathcal{X}}^2\|C_{Y,t}\|_{\mathrm{op}}) \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}\|\sum_t C_{Y,t}\|_{\mathrm{op}} \leq \frac{2\mathsf{m}_{\mathcal{X}}^2}{n}(T\mathcal{R}(G_*) + \|G_*\|_{\mathsf{HS}}^2\|\sum_t C_t\|_{\mathrm{op}}), \tag{41}$$

where the last inequality follows by Lemma 15.

Both$\|\sum_{it}\mathbb{E}X_{it}^*X_{it}\|_{\mathrm{op}}$ and $\|\sum_{it}\mathbb{E}X_{it}X_{it}^*\|_{\mathrm{op}}$ are upper bounded by $\frac{2\mathsf{m}_{\mathcal{X}}^2}{n}(T\mathcal{R}(G_*) + \|G_*\|_{\mathsf{HS}}^2\|\sum_t C_t\|_{\mathrm{op}})$.

Then, by Lemma 12, we have

$$\|\mathbf{Z}\|_{\mathrm{op}} \leq \frac{4}{n}\Big(\frac{\max_t\mathsf{m}_{\mathcal{X}}\mathsf{m}_{y,t}}{3}\ln\Big(\frac{28\bar{r}}{\delta}\Big)\Big) + \sqrt{\frac{2\mathsf{m}_{\mathcal{X}}^2(T\mathcal{R}(G_*) + \|G_*\|_{\mathsf{HS}}^2\|\sum_t C_t\|_{\mathrm{op}})}{n}\ln\Big(\frac{28\bar{r}}{\delta}\Big)},$$

where $\bar{r}$ is the effective rank of $\mathbf{Z}$. Rearranging we get

$$\|\mathbf{Z}\|_{\mathrm{op}} \leq \frac{4}{n}\Big(\frac{\mathsf{k}_1}{3}\ln\Big(\frac{28\bar{r}}{\delta}\Big)\Big) + \mathsf{m}_{\mathcal{X}}\sqrt{\frac{2T\mathcal{R}(G_*)}{n}}\ln\Big(\frac{\mathsf{r}_2}{\delta}\Big) + \|G\|_*\mathsf{m}_{\mathcal{X}}\sqrt{\|\sum_t C_t\|_{\mathrm{op}}\frac{2}{n}\ln\Big(\frac{\mathsf{r}_2}{\delta}\Big)}$$

where $\mathsf{k}_1 = \mathsf{m}_{\mathcal{X}}\max_t\mathsf{m}_{y,t}$ and $\mathsf{r}_2 = 28\bar{r}$. □

Lemma 10. *Recall that $\|\psi(y_{it})\|^2 \leq \mathsf{m}_{y,t}^2$ for $i = 1,\dots,n$, $t = 1,\dots T$.*

$$\left|\frac{1}{T}\sum_{t=1}^T\frac{1}{n}\sum_{i=1}^n\|\psi(y_{it})\|_{\mathcal{H}_{\mathcal{Y}}}^2 - \mathbb{E}\|\psi(y_t)\|_{\mathcal{H}_{\mathcal{Y}}}^2\right| \leq \sqrt{\Big(\frac{1}{T}\sum_t\mathsf{m}_{y,t}^2\Big)\frac{1}{nT}\ln\Big(\frac{2}{\delta}\Big)}$$

with probability at least $1-\delta$.

Proof. The bound follows by a direct application of Hoeffding inequality. □

We are now ready to prove theorem Thm. 7.

Proof. Recall that

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq (\|G_*\|_* - \gamma)^2 \frac{\mathsf{m}_{\mathcal{X}}^2}{T} + 2 \sup_{G \in \mathcal{G}_\gamma} |\mathcal{R}(\hat{G}) - \mathcal{R}(G))|.$$

Recall that for any $G \in \mathcal{G}_\gamma$, $\|G\|_* \leq \gamma$. Now, combining Eq. (39) and Lemma 8, Lemma 9 and 10, we get

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq ((\|G_*\|_* - \gamma)^2 \frac{\mathsf{m}_{\mathcal{X}}^2}{T} + \gamma^2 \left(\frac{A}{n} + \frac{B}{\sqrt{n}}\right) + \gamma \left(\frac{A'}{n} + \frac{B'}{\sqrt{n}}\right) + \sqrt{\left(\frac{1}{T}\sum_t \mathsf{m}_{\mathcal{Y},t}^2\right) \frac{1}{nT} \ln\left(\frac{2}{\delta}\right)},$$

where

$$A = \frac{4}{T}\Big(\frac{\mathsf{m}_{\mathcal{X}}^2}{3} \ln\left(\frac{T\mathsf{r}_1}{\delta}\right)\Big) \qquad B = \frac{1}{T}\sqrt{4\mathsf{m}_{\mathcal{X}}^2 \max_t \|C_t\|_{\mathrm{op}} \ln\left(\frac{T\mathsf{r}_1}{\delta}\right)\Big)} + \frac{1}{T}\mathsf{m}_{\mathcal{X}} \sqrt{2\|\sum_t C_t\|_{\mathrm{op}} \ln\left(\frac{\mathsf{r}_2}{\delta}\right)}$$

$$A' = \frac{4}{T}\Big(\frac{\max_t \mathsf{m}_{\mathcal{X}}\mathsf{m}_{\mathcal{Y},t}}{3} \ln\left(\frac{\mathsf{r}_2}{\delta}\right)\Big) \qquad B' = \frac{1}{T}\sqrt{2T\mathcal{R}(G_*) \ln\left(\frac{\mathsf{r}_2}{\delta}\right)}.$$

Optimizing with respect to γ we could find the optimal parameter and compute the corresponding bound. However, in the following we choose $\gamma = \|G_*\|_*$, so that the approximation error is zero. In the following we will bound $\max_t \|C_t\|_{\mathrm{op}}$ with $\|\sum_t C_t\|_{\mathrm{op}}$ and both the logarithm terms with $\ln(\frac{T\mathsf{r}}{\delta})$ for a suitable r (e.g. $\max(\mathsf{r}_1, \mathsf{r}_2)$). Isolating the faster term we obtain

$$\begin{aligned}\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq & \|G_*\|_*^2 \mathsf{m}_{\mathcal{X}} \left[\frac{\sqrt{\|\sum_t C_t\|_{\mathrm{op}}}}{T} \frac{2+\sqrt{2}}{\sqrt{n}} + \frac{\mathsf{m}_{\mathcal{X}}\|G_*\|_*}{T}\sqrt{\frac{2T\mathcal{R}(G_*)}{n}}\right] \ln\left(\frac{T\mathsf{r}}{\delta}\right) \\ & + \sqrt{\left(\frac{1}{T}\sum_t \mathsf{m}_{\mathcal{Y},t}^2\right)\frac{1}{nT}\ln\left(\frac{2}{\delta}\right)} + O((nT)^{-1}).\end{aligned}$$

Denote by $\bar{C}$ the average of $C_1, \dots, C_T$, i.e. $\frac{1}{T}\sum_t C_t$. Then,

$$\frac{1}{T}\sqrt{\|\sum_t C_t\|_{\mathrm{op}}} = \frac{1}{\sqrt{T}}\|\bar{C}\|_{\mathrm{op}}^{\frac{1}{2}}.$$

Rearranging the terms we get the final bound

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq \frac{1}{\sqrt{nT}}\Big(\|G_*\|_*^2 \|\bar{C}\|_{\mathrm{op}}^{\frac{1}{2}} \mathsf{m}_{\mathcal{X}}(2+\sqrt{2}) + \mathsf{m}_{\mathcal{X}}\|G_*\|_*\sqrt{2\mathcal{R}(G_*)} + \bar{\mathsf{m}}_{\mathcal{Y}}\Big)\sqrt{\ln\left(\frac{T\mathsf{r}}{\delta}\right)} + O((nT)^{-1}),$$

where $\bar{\mathsf{m}}_{\mathcal{Y}} = \sqrt{\frac{1}{T}\sum_t \mathsf{m}_{\mathcal{Y},t}^2}$.

□

D. Auxiliary Lemmas

In this section we recall some auxiliary results that are used in the proofs of the work. Let us recall the definition of effective rank. Given an Hilbert space $\mathcal{H}$ let $A : \mathcal{H} \to \mathcal{H}$, be a compact operator. The effective rank of A is refined as

$$r(A) = \frac{\mathrm{tr} A}{\|A\|_{\mathrm{op}}}.$$

Lemma 11. *Let $X_1, \dots, X_n \in \mathbb{C}^{d \times d}$ a sequence of independent self adjoint random matrices such that $\mathbb{E}X_i = 0$, for $i = 1, \dots, n$ and $\sigma^2 \geq \| \sum_{i=1}^n \mathbb{E}X_i^2 \|_{\mathrm{op}}$. Assume that $\|X_i\| \leq U$ almost surely for all $1 \leq i \leq n$ and some positive $U \in \mathbb{R}$. Then, for any $t \geq \frac{1}{6}(U + \sqrt{U + 36\sigma^2})$,*

$$\mathbb{P}\Big(\| \sum_{i=1}^n X_i \|_{\mathrm{op}} > t\Big) \leq 14 r(\sum_{i=1}^n \mathbb{E}X_i^2) \exp\big(-\frac{t^2/2}{\sigma^2 + tU/3}\big), \tag{42}$$

where $r(\cdot)$ denotes the effective rank.

A similar results holds true for general matrices with no requirements on self adjointness:

Lemma 12. *Let $X_1, \dots, X_n \in \mathbb{C}^{d \times d}$ a sequence of independent random matrices such that $\mathbb{E}X_i = 0$, for $i = 1, \dots, n$ and $\sigma^2 \geq max(\| \sum_{i=1}^n \mathbb{E}X_i X_i^* \|_{\mathrm{op}}, \| \sum_{i=1}^n \mathbb{E}X_i^* X_i \|_{\mathrm{op}}$. Assume that $\|X_i\| \leq U$ almost surely for all $1 \leq i \leq n$ and some positive $U \in \mathbb{R}$. Then, for any $t \geq \frac{1}{6}(U + \sqrt{U + 36\sigma^2})$,*

$$\mathbb{P}\Big(\| \sum_{i=1}^n X_i \|_{\mathrm{op}} > t\Big) \leq 28 \tilde{d} \exp\big(-\frac{t^2/2}{\sigma^2 + tU/3}\big), \tag{43}$$

where $\tilde{d} = \max(r(\sum_{i=1}^n \mathbb{E}X_i X_i^*), r(\sum_{i=1}^n \mathbb{E}X_i^* X_i)))$ and $r(\cdot)$ denotes the effective rank.

The lemma above holds true for Hilbert Schmidt operators between separable Hilbert spaces, as shown in section 3.2 in (Minsker, 2017).

Lemma 13. *The following bound on the operator norm of the covariance operator on the output $\mathbb{E}\psi(y) \otimes \psi(y)$ holds true:*

$$\|\mathbb{E}\psi(y) \otimes \psi(y)\|_{\mathrm{op}} \leq \|G_*\|_{\mathsf{HS}}^2 \|C\|_{\mathrm{op}} + \mathcal{R}(g_*).$$

Proof. Let us start for the identity below:

$$\psi(y) \otimes \psi(y) = (\psi(y) - G_*\phi(x)) \otimes (\psi(y) - G_*\phi(x)) + G_*\phi(x) \otimes (\psi(y) - G_*\phi(x)) + \psi(y) \otimes G_*\phi(x). \tag{44}$$

Taking the expectation on the right hand side we obtain

$$\mathbb{E}((\psi(y) - G_*\phi(x)) \otimes (\psi(y) - G_*\phi(x))) + \mathbb{E}G_*\phi(x) \otimes (\psi(y) - G_*\phi(x)) + \mathbb{E}\psi(y) \otimes G_*\phi(x).$$

Note that the second term is zero, since

$$\begin{aligned} \mathbb{E}G_*\phi(x) \otimes (\psi(y) - G_*\phi(x)) &= \int_{\mathcal{X} \times \mathcal{Y}} G_*\phi(x) \otimes (\psi(y) - G_*\phi(x)) d\rho(x, y) \\ &= \int_{\mathcal{X}} G_*\phi(x) \Big(\int_{\mathcal{Y}} \psi(y) d\rho(y \mid x) - G_*\phi(x) \Big) d\rho_{\mathcal{X}} \end{aligned}$$

and $G_*\phi(x) = \int_{\mathcal{Y}} \phi(y) d\rho(y \mid x)$. As for the last term, we have

$$\mathbb{E}\psi(y) \otimes G_*\phi(x) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \psi(y) d\rho(y \mid x) \otimes G_*\phi(x) d\rho_{\mathcal{X}} = \int_{\mathcal{X}} G_*\phi(x) \otimes G_*\phi(x).$$

Taking the operator norm we get

$$\begin{aligned} \|\mathbb{E}\psi(y) \otimes \psi(y)\|_{\mathrm{op}} &\leq \|\mathbb{E}((\psi(y) - G_*\phi(x)) \otimes (\psi(y) - G_*\phi(x)))\|_{\mathrm{op}} + \|G_* C G_*^*\|_{\mathsf{HS}} \\ &\leq \mathcal{R}(g_*) + \|G_* C G_*^*\|_{\mathsf{HS}}^2 \leq \mathcal{R}(g_*) + \|G\|_{\mathsf{HS}}^2 \|C\|_{\mathrm{op}}. \end{aligned}$$

□

Lemma 14. *The following bound holds true*

$$\|\mathbb{E}(\langle\psi(y),\psi(y)\rangle\phi(x)\otimes\phi(x))\|_{\mathrm{op}} \le \mathsf{m}_{\mathcal{X}}^2(\|G_*\|_{\mathsf{HS}}^2\|C\|_{\mathrm{op}} + \mathcal{R}(g_*)).$$

Proof. Let us rewrite $\mathbb{E}(\langle\psi(y),\psi(y)\rangle\phi(x)\otimes\phi(x))$ as follows

$$\int_{\mathcal{X}\times\mathcal{Y}}\langle\psi(y),\psi(y)\rangle\phi(x)\otimes\phi(x)d\rho(x,y) = \int_{\mathcal{X}}\phi(x)\otimes\phi(x)\left(\int_{\mathcal{Y}}\langle\psi(y),\psi(y)\rangle d\rho(y\mid x)\right)d\rho_{\mathcal{X}}(x). \tag{45}$$

The inner integral corresponds to $\mathbb{E}_{y|x}\mathrm{tr}(\psi(y)\otimes\psi(y)) = \mathrm{tr}\,\mathbb{E}_{y|x}(\psi(y)\otimes\psi(y))$. Writing $\psi(y)\otimes\psi(y)$ as in Eq. (44) and integrating wrt $\rho(\cdot\mid x)$, we observe that

$$\begin{aligned}\int_{\mathcal{Y}}\psi(y)\otimes\psi(y)d\rho(y\mid x) = &\int_{\mathcal{Y}}(\psi(y)-G_*\phi(x))\otimes(\psi(y)-G_*\phi(x))d\rho(y\mid x)\\ &+G_*\phi(x)\otimes\left(\int_{\mathcal{Y}}\psi(y)d\rho(y\mid x)-G_*\phi(x)\right)\\ &+G_*\phi(x)\otimes\phi(x)G_*^*.\end{aligned}$$

Since $\int_{\mathcal{Y}}\psi(y)d\rho(y\mid x) = G_*\phi(x)$, the second term on the right hand side is zero and hence

$$\mathrm{tr}\,\mathbb{E}_{y|x}\psi(y)\otimes\psi(y) = \mathrm{tr}\,\mathbb{E}_{y|x}(\psi(y)-G_*\phi(x))\otimes(\psi(y)-G_*\phi(x)) + \mathrm{tr}\,G_*\phi(x)\otimes\phi(x)G_*^*.$$

Substituting it on the right hand side of Eq. (45) and taking the operator norm and using the triangle inequality, we obtain

$$\begin{aligned}\|\int_{\mathcal{X}}\phi(x)\otimes\phi(x)&\left(\int_{\mathcal{Y}}\|\psi(y)-G_*\phi(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2d\rho(y\mid x)\right)d\rho_{\mathcal{X}}(x)\|_{\mathrm{op}}\\ &\le\mathsf{m}_{\mathcal{X}}^2\int_{\mathcal{X}\times\mathcal{Y}}\|\psi(y)-G_*\phi(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2d\rho(y,x) = \mathsf{m}_{\mathcal{X}}^2\mathcal{R}(g_*)\end{aligned}$$

and

$$\|\int_{\mathcal{X}}\phi(x)\otimes\phi(x)\,\mathrm{tr}(G_*\phi(x)\otimes\phi(x)G_*^*)\|_{\mathrm{op}} \le \|C\|_{\mathrm{op}}\mathsf{m}_{\mathcal{X}}^2\|G_*\|_{\mathsf{HS}}^2.$$

Combining the parts together leads to the desired inequality

$$\|\mathbb{E}(\langle\psi(y),\psi(y)\rangle\phi(x)\otimes\phi(x))\|_{\mathrm{op}} \le \mathsf{m}_{\mathcal{X}}^2(\|G_*\|_{\mathsf{HS}}^2\|C\|_{\mathrm{op}} + \mathcal{R}(g_*)).$$

□

Lemma 15. *Let $C_{Y,t}$ denote the covariance on the output for the t^{th} task, that is*

$$C_{Y,t} := \mathbb{E}\psi(y_t)\otimes\psi(y_t). \tag{46}$$

Then the following inequality holds true

$$\|\sum_t C_{Y,t}\|_{\mathrm{op}} \le \|G_*\|_{\mathsf{HS}}^2\|\sum_t C_t\|_{\mathrm{op}} + T\mathcal{R}(G_*). \tag{47}$$

Proof. Let us start from the identity below:

$$\psi(y_t)\otimes\psi(y_t) = (\psi(y_t)-G_{t*}\phi(x_t))\otimes(\psi(y_t)-G_{t*}\phi(x_t)) + G_{t*}\phi(x_t)\otimes(\psi(y_t)-G_{t*}\phi(x_t)) + \psi(y_t)\otimes G_{t*}\phi(x_t).$$

Taking the expectation on the right hand side we obtain

$$\mathbb{E}((\psi(y_t)-G_{t*}\phi(x_t))\otimes(\psi(y_t)-G_{t*}\phi(x_t) + \mathbb{E}G_{t*}\phi(x_t)\otimes(\psi(y_t)-G_{t*}\phi(x_t)) + \mathbb{E}\psi(y_t)\otimes G_{t*}\phi(x_t).$$

As in Lemma 13, note that the second term is zero. As for the last term, we have

$$\begin{aligned}\mathbb{E}\psi(y_t) \otimes G_{t*}\phi(x_t) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \psi(y_t) d\rho_t(y \mid x) \otimes G_{t*}\phi(x) d\rho_{t\mathcal{X}} = \int_{\mathcal{X}} (G_{t*}\phi(x)) \otimes (G_{t*}\phi(x)) d\rho_{t,\mathcal{X}} \\ &= \int_{\mathcal{X}} (G_{t*}\phi(x) \otimes \phi(x) G_{t*}^*) \rho_{t,\mathcal{X}} = G_{t*} C_t G_{t*}^*.\end{aligned}$$

Therefore, summing on t and taking the operator norm we get

$$\begin{aligned}\|\sum_t C_{Y,t}\|_{\mathrm{op}} &\leq \|\sum_t \mathbb{E}((\psi(y_t) - G_{t*}\phi(x_t))^2\|_{\mathrm{op}} + \|G_{t*}C_tG_{t*}^*\|_{\mathsf{HS}} \\ &\leq \sum_t \mathcal{R}(G_{t*}) + \|\sum_t G_{t*}C_tG_{t*}^*\|_{\mathsf{HS}} \leq \sum_t \mathcal{R}(G_{t*}) + \|\sum_t G_{t*}\sum_s C_sG_{t*}^*\|_{\mathsf{HS}} \\ &\leq \sum_t \mathcal{R}(G_{t*}) + \sum_t \|G_{t*}\|_{\mathsf{HS}}^2 \|\sum_t C_t\|_{\mathrm{op}} \leq T\mathcal{R}(G_*) + \|G_*\|_{\mathsf{HS}}^2\|\sum_t C_t\|_{\mathrm{op}}.\end{aligned}$$

□

Lemma 16. *The following bound holds true*

$$\|\sum_t \mathbb{E}(\psi_t(y_t)^2 \phi(x_t) \otimes \phi(x_t))\|_{\mathrm{op}} \leq \mathsf{m}_{\mathcal{X}}^2(\|G_*\|_{\mathsf{HS}}^2 \|\sum_t C_t\|_{\mathrm{op}} + \mathcal{R}(g_*)).$$

Proof. It is a immediate variation of the proof of Lemma 14. □

E. Equivalence between Tikhonov and Ivanov Problems for trace norm Regularization

In this section we provide more details regarding the relation between the Tikhonov regularization problem considered in Eq. (11) and the corresponding Ivanov problem in Eq. (13). As discussed in the paper this approach guarantees that theoretical results characterizing the excess risk of the Ivanov estimator extend automatically to the Tikhonov one.

Let $(x_i, y_i)_{i=1}^n$ be a training set and consider $\Phi : \mathcal{H}_{\mathcal{X}} \to \mathbb{R}^n$ and $\Psi : \mathcal{H}_{\mathcal{Y}} \to \mathbb{R}^n$ the operators

$$\Phi = \sum_{i=1}^n e_i \otimes \phi(x_i) \qquad \text{and} \qquad \Psi = \sum_{i=1}^n e_i \otimes \psi(y_i)$$

with $e_i \in \mathbb{R}^n$ the i-th element of the canonical basis in $\mathbb{R}^n$. We can write the empirical surrogate risk in compact operatorial notation as

$$\hat{\mathcal{R}}(G) = \frac{1}{n}\sum_{i=1}^n \|G\phi(x_i) - \psi(y_i)\|_{\mathcal{H}_{\mathcal{Y}}}^2 = \frac{1}{n}\|\Phi G^* - \Psi\|_{\mathcal{H}_{\mathcal{Y}} \otimes \mathbb{R}^n}^2.$$

Proposition 17 (Representer Theorem for Trace Norm Regularization)**.** *Let $\hat{G} \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ be a minimizer of*

$$\min_{G \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} \hat{\mathcal{R}}(G) + \lambda\|G\|_*.$$

Then the range of $\hat{G}^*$ is contained in the range of Φ^*, or equivalently

$$\hat{G}(\Phi^\dagger \Phi) = \hat{G},$$

where $\Phi^\dagger$ denotes the pseudoinverse of Φ.

The proof of this result is essentially equivalent to the one in (Thm. 3 Abernethy et al., 2009). We report it here for completeness.

Proof. For any $G \in \mathcal{H}_y \otimes \mathcal{H}_x$, consider the factorization

$$G = G_0 + G_\perp \qquad \text{with} \qquad G_0 = G(\Phi^\dagger \Phi) \qquad and \qquad G_\perp = (I - (\Phi^\dagger \Phi)).$$

Note that $(\Phi^\dagger \Phi) \in \mathcal{H}_x \otimes \mathcal{H}_x$ corresponds to the orthogonal projector of $\mathcal{H}_x$ onto the range of Φ^* in $\mathcal{H}_x$ (equivalently onto the span of $(\phi(x_i))_{i=1}^n$). By construction, we have that $\Phi G^* = \Phi G_0^*$. Hence $\hat{\mathcal{R}}(G) = \hat{\mathcal{R}}(G_0)$. However, since $(\Phi^\dagger \Phi)$ is an orthogonal projector, we have that

$$\|G_0\|_* = \|G(\Phi^\dagger \Phi)\|_* \leq \|G\|_*,$$

with equality holding if and only if $G_0 = G$.

Now, if $\hat{G}$ is a minimizer of the trace norm regularized ERM we have

$$\begin{aligned} \hat{\mathcal{R}}(\hat{G}_0) + \lambda \|\hat{G}_0\|_* &\geq \hat{\mathcal{R}}(\hat{G}) + \lambda \|\hat{G}\|_* \\ &= \hat{\mathcal{R}}(\hat{G}_0) + \lambda \|\hat{G}\|_*, \end{aligned}$$

which implies $\|\hat{G}_0\|_* \geq \|\hat{G}\|_*$.

We conclude that $\hat{G} = \hat{G}_0 = \hat{G}(\Phi^\dagger \Phi)$. This corresponds to the range of G^* being contained in the range of Φ as desired. □

Proposition 18. *The empirical risk minimization for $\hat{\mathcal{R}}(G) + \lambda \|G\|_*$ admits a unique minimizer.*

Proof. According to Prop. 17, all minimizers of the trace norm regularized empirical risk minimization belong to the set

$$\mathcal{S} = \left\{ G \in \mathcal{H}_y \otimes \mathcal{H}_x \;\middle|\; G(\Phi^\dagger \Phi) = G \right\}.$$

Hence we can restrict to the optimization problem

$$\min_{G \in \mathcal{S}} \; \hat{\mathcal{R}}(G) + \lambda \|G\|_*.$$

Note that $\mathcal{S}$ is idetified by a linear relation and thus is a convex set and thus the problem above is a convex program. We now show that on $\mathcal{S}$ the ERM objective functional is actually strongly convex for the case of the least-squares loss. To see this, let us consider the Hessian of $\hat{\mathcal{R}}(\cdot)$. We have that, the gradient corresponds to

$$\nabla \hat{\mathcal{R}}(G) = \frac{2}{n} \left(G \Phi^* \Phi - \Psi^* \Phi \right),$$

and therefore the Hessian is the operator $\nabla^2 \hat{\mathcal{R}}(G) : \mathcal{H}_y \otimes \mathcal{H}_x \to \mathcal{H}_y \otimes \mathcal{H}_x$ such that

$$\nabla^2 \hat{\mathcal{R}}(G) H = \frac{2}{n} H \Phi^* \Phi,$$

for any $H \in \mathcal{H}_y \otimes \mathcal{H}_x$ (see e.g. Kollo & von Rosen, 2006). Now, we have that for any $H \in \mathcal{S}$

$$\left\langle H, \nabla^2 \hat{\mathcal{R}}(G) H \right\rangle_{\mathcal{H}_y \otimes \mathcal{H}_x} = \frac{2}{n} \left\langle H, H \Phi^* \Phi \right\rangle_{\mathcal{H}_y \otimes \mathcal{H}_x} = \frac{2}{n} \operatorname{tr}(H^* H \Phi^* \Phi).$$

Now, let $r \leq n$ be the rank of Φ and consider the singular value decomposition of $\Phi = U \Sigma V^*$, with $U \in \mathbb{R}^{n \times r}$ a matrix with orthonormal columns $V \in \mathcal{H}_x \to \mathbb{R}^r$ a linear operator such that $V^* V = I \in \mathbb{R}^{r \times r}$ and $\Sigma \in \mathbb{R}^{r \times r}$ a diagonal matrix with *all positive diagonal elements*. Then,

$$\operatorname{tr}(H^* H \Phi^* \Phi) = \operatorname{tr}(H^* H V \Sigma^2 V^*) = \operatorname{tr}(V^* H^* H V \Sigma^2) \geq \sigma_{\min}^2 \|HV\|_{\mathcal{H}_y \otimes \mathbb{R}^r}^2,$$

where σ_{min}^2 denotes the smallest singular value of Σ (equivalently, σ_{min} is the smallest singular value of Φ *greater than zero*).

Now, recall that $H \in \mathcal{S}$. Therefore

$$H = H(\Phi^\dagger \Phi) = HVV^*,$$

which implies that

$$\|H\|^2_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}} = \operatorname{tr}(H^*H) = \operatorname{tr}(VV^*H^*HVV^*) = \operatorname{tr}(V^*H^*HV(V^*V)) = \operatorname{tr}(V^*H^*HV) = \|HV\|^2_{\mathcal{H}_\mathcal{Y}\otimes\mathbb{R}^r},$$

where we have used the orthonormality $V^*V = I \in \mathbb{R}^{r\times r}$.

We conclude that

$$\left\langle H, \nabla^2\hat{\mathcal{R}}(G)H\right\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}} \geq \frac{2\sigma^2_{\mathrm{m}in}}{n}\|H\|^2_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}},$$

for any $H \in \mathcal{S}$. Note that $\sigma_{\mathrm{m}in} > 0$ is greater than zero since it is the smallest singular value of Φ greater than zero and Φ has finite rank $r \leq n$. Hence, on $\mathcal{S}$, the function $\hat{\mathcal{R}}(G)$ is strongly convex. As a consequence also the objective functional $\hat{\mathcal{R}}(G) + \lambda\|G\|_*$ is strongly convex and thus admits a unique minimizer, as desired. □

We conclude this section by reporting the result stating the equivalence between Ivanov and Tikhonov for trace norm regularization.

In the following we will denote by G_λ the minimizer of the Tikhonov regularization problem corresponding to minimizing $\hat{\mathcal{R}}(G) + \lambda\|G\|_*$ and by G^I_γ the minimizer of the Ivanov regularization problem introduced in Eq. (13), namely

$$\min_{\|G\|_*\leq\gamma} \hat{\mathcal{R}}(G).$$

We have the following.

Theorem 19. *For any $\gamma > 0$ there exists $\lambda(\gamma)$ such that $G_{\lambda(\gamma)}$ is a minimizer of Eq. (13). Moreover, for any $\lambda > 0$ there exists a $\gamma = \gamma(\lambda) > 0$ such that G_λ is a minimizer of Eq. (13).*

Proof. We first consider the case where, given a $\gamma > 0$ we want to relate a solution of the Ivanov regularization problem to that of Tikhonov regularization. We will show that there exists G^I_γ and $\lambda(\gamma)$ such that $G_\lambda(\gamma) = G^I_\gamma$. In particular we will show that such equality holds for G^I_γ the solution of minimal trace norm in the set of solutions of the Ivanov problem.

Consider again the linear subspace

$$\mathcal{S} = \left\{G \in \mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X} \;\middle|\; G = G(\Phi^\dagger\Phi)\right\}.$$

We can restrict the original Ivanov problem to

$$\min_{\substack{\|G\|_*\leq\gamma\\ G\in\mathcal{S}}} \hat{\mathcal{R}}(G).$$

Note that the above is still a convex program and attains the same minimum value of the original Ivanov problem in G^I_γ.

Moreover, we can assume $\gamma = \|G^I_\gamma\|_*$ without loss of generality. Indeed, if $\gamma > \|G^I_\gamma\|_*$ we still have that G^I_γ is a minimizer of $\hat{\mathcal{R}}(G)$ over the smaller set of operators $\|G\|_* \leq \gamma' = \|G^I_\gamma\|_*$.

Now, consider the Lagrangian associated to this constrained problem problem, namely

$$L(G,\lambda,\nu) = \hat{\mathcal{R}}(G) + \lambda(\|G\|_* - \gamma) + \nu(G - G(\Phi^\dagger\Phi)).$$

By Slater's constraint qualification (see e.g. Sec. 5 in Boyd & Vandenberghe, 2004), we have that

$$\max_{\lambda\geq0,\nu}\ \min_{G\in\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}} L(G,\lambda,\nu) = \min_{\substack{\|G\|_*\leq\gamma\\ G\in\mathcal{S}}} \hat{\mathcal{R}}(G).$$

Denote by $(\lambda(\gamma), G_{\lambda(\gamma)}, \nu_\gamma)$ the pair form which the saddle point of $L(G,\lambda,\gamma)$ is attained. Note that since γ is a constant

$$G_{\lambda(\gamma)} = \operatorname*{argmin}_{G\in\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}} \hat{\mathcal{R}}(G) + \lambda(\gamma)(\|G\|_* - \gamma) + \nu_\gamma(G - G(\Phi^\dagger\Phi)) = \operatorname*{argmin}_{G\in\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}} \hat{\mathcal{R}}(G) + \lambda(\gamma)\|G\|_*,$$

where we have made use of the representer theorem from Prop. 17, which guarantees any minimizer of $\hat{\mathcal{R}}(G) + \lambda(\gamma)\|G\|_*$ to belong to the set $\mathcal{S}$ and this satisfy $G = G(\Phi^\dagger \Phi)$. Therefore, we have

$$\hat{\mathcal{R}}(G_{\lambda(\gamma)}) + \lambda(\gamma)\|G_{\lambda,\gamma}\|_* - \lambda(\gamma)\gamma = \hat{\mathcal{R}}(G_\gamma^I),$$

recalling that $\gamma = \|G_\gamma^I\|_*$, this implies that

$$\hat{\mathcal{R}}(G_{\lambda,\gamma}) + \lambda(\gamma)\|G_{\lambda,\gamma}\|_* = \hat{\mathcal{R}}(G_\gamma^I) + \lambda(\gamma)\|G_\gamma^I\|_*.$$

Since by Prop. 18 the minimizer of $\hat{\mathcal{R}}(G) + \lambda\|G\|_*$ is unique, it follows that $G_{\lambda,\gamma} = G_\gamma^I$ as desired.

The vice-versa is straightforward: let $\lambda > 0$ and G_λ be the minimizer of the Tikhonov problem. Then, for any $G \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$

$$\hat{\mathcal{R}}(G_\lambda) + \lambda\|G_\lambda\|_* \leq \hat{\mathcal{R}}(G) + \lambda\|G\|_*.$$

If $\|G\|_* \leq \|G_\lambda\|_*$, the inequality above implies

$$\hat{\mathcal{R}}(G_\lambda) \leq \hat{\mathcal{R}}(G),$$

which implies that G_λ is a minimizer for the Ivanov problem with $\gamma(\lambda) = \|G_\lambda\|_*$, namely $G_\lambda = G_{\gamma(\lambda)}^I$ as desired. □