Is Supervised Learning Really That Different From Unsupervised?

Oskar Allerbo
KTH Royal Institute of Technology

Thomas B. Schön
Uppsala University

Abstract

We demonstrate how supervised learning can be decomposed into a two-stage procedure, where (1) all model parameters are selected in an unsupervised manner, and (2) the outputs $\boldsymbol{y}$ are added to the model, *without changing the parameter values*. This is achieved by a new model selection criterion that—in contrast to cross-validation—can be used also without access to $\boldsymbol{y}$. For linear ridge regression, we bound the asymptotic out-of-sample risk of our method in terms of the optimal asymptotic risk. We also demonstrate that versions of linear and kernel ridge regression, smoothing splines, k-nearest neighbors, random forests, and neural networks, trained without access to $\boldsymbol{y}$, perform similarly to their standard $\boldsymbol{y}$-based counterparts. Hence, our results suggest that the difference between supervised and unsupervised learning is less fundamental than it may appear.

1 INTRODUCTION

In contrast to supervised learning, unsupervised learning is performed without access to the outputs $\boldsymbol{y}$, i.e. the labels (for classification) or the response (for regression). The success of unsupervised learning raises the question: Are the outputs really needed when training supervised models? Or is it possible to successfully train a supervised model without $\boldsymbol{y}$, not including the labels/response until after all model parameters are selected and kept fixed? If this is indeed the case, it suggests that—under the hood—supervised learning is closely related to unsupervised learning.

In Figure 1, we give an example where supervised classification can be replaced by a two-stage procedure, based on unsupervised learning. First, the model parameters are selected in an unsupervised manner, splitting the input data into four classes, with (yet) unknown labels. Then, the training labels are revealed, allowing the model to assign labels to the four classes, but without altering the borders between classes.

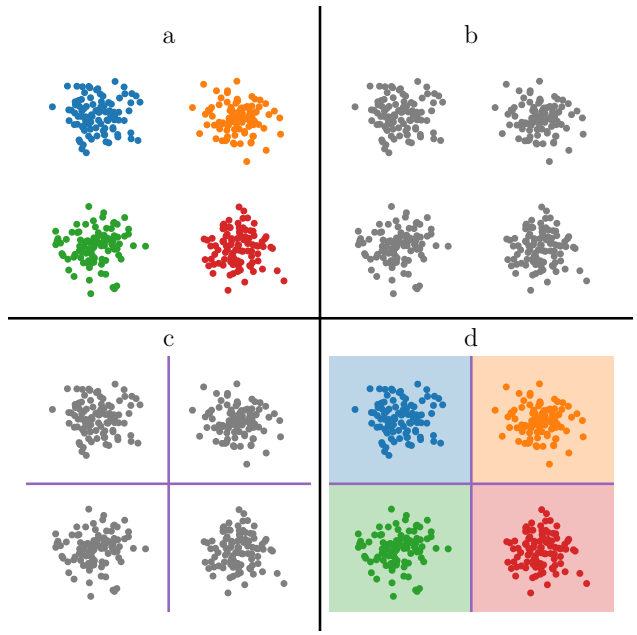

Figure 1: Illustrating training without labels: After the label information (a) has been removed (b), an unsupervised algorithm is used to split the input space into four classes, separated by the purple cross, without using the label information (c). After fixing the class borders, the labels are revealed, enabling the model to label each class (d).

More generally, the key to training without $\boldsymbol{y}$ is to express the model as a smoother, i.e. on the form $\hat{f}^* = \boldsymbol{s}^{*\top}\boldsymbol{y}$, where $\hat{f}^* = \hat{f}(\boldsymbol{x}^*)$ denotes the prediction for covariate $\boldsymbol{x}^*$, and $\boldsymbol{s}^*$ denotes a smoother vector. This implies that each prediction is a weighted sum of the n observations in $\boldsymbol{y}$, where the weights are given by the smoother vector. If $\boldsymbol{s}^*$ only depends on the covariates and not on $\boldsymbol{y}$, the model is linear in $\boldsymbol{y}$ (but *not*

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

necessarily in $\boldsymbol{x}$). Such a model is referred to as a linear smoother. If $\boldsymbol{s}^*$ additionally depends on $\boldsymbol{y}$, it is called a nonlinear smoother. For many models, including linear and kernel ridge regression, $\boldsymbol{s}^*$ follows immediately from the closed-form expression of the model, while for others, it is more complicated, but still possible. For instance, Jeffares et al. (2024) recently showed how neural networks can be expressed as nonlinear smoothers based on the neural tangent kernel (NTK) framework. Since the smoother contains all the model parameters, whenever $\boldsymbol{s}^*$ is selected independently of $\boldsymbol{y}$, training the model is $\boldsymbol{y}$-free.

The possibility of training without $\boldsymbol{y}$ has an interesting implication: It suggests that training a supervised model amounts to finding a good representation of $\boldsymbol{x}^*$, and that this can be done without the information in $\boldsymbol{y}$. To obtain the predictions, $\boldsymbol{y}$ must be included in the model; however, it is sufficient to do this *after* the representation of $\boldsymbol{x}^*$ is constructed. It can even be hypothesized that the optimal model can always be written as a linear smoother, meaning that the optimal predictions are weighted sums of $\boldsymbol{y}$. The weights depend on the training covariates $\boldsymbol{X}$ and $\boldsymbol{x}^*$. Hence, the weights depend on how the out-of-sample data relate to the training data, but not on $\boldsymbol{y}$.

In this work, we introduce a model-independent way to train supervised models, including neural networks, as linear smoothers. We also demonstrate that the obtained models compare very well to the corresponding models trained with the information in $\boldsymbol{y}$ in terms of performance. We thus illustrate both the close relation between supervised and unsupervised learning and the plausibility of the hypothesis above.

Our **main contribution** is to **demonstrate** that **standard supervised models** for classification and regression can be **successfully trained in an unsupervised manner**. We do this by

- introducing a *model selection criterion* that is based on the out-of-sample model predictions, and thus works for any classification/regression model. We demonstrate that this criterion, in contrast to cross-validation, also *works without access to* $\boldsymbol{y}$.
- bounding the asymptotic risk of our method in terms of the optimal asymptotic risk for linear ridge regression.

The aim of this paper is to demonstrate the feasibility of training supervised models without access to $\boldsymbol{y}$, thereby highlighting the parallels between supervised and unsupervised learning. It is not intended to suggest that our method outperforms existing $\boldsymbol{y}$-based approaches, nor that models should be trained without $\boldsymbol{y}$ in practical applications.

All proofs are deferred to Appendix F and the code is available at `https://github.com/allerbo/training_without_y`.

2 NOTATION

We denote scalars with lowercase non-bold letters, vectors with lowercase bold letters, and matrices with uppercase bold letters. The in-sample (training) data are denoted as $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$. The out-of-sample quantities are denoted with a star: $\boldsymbol{x}^* \in \mathbb{R}^d$ denotes any out-of-sample covariate, while $\boldsymbol{X}_v^* \in \mathbb{R}^{n_v^*\times d}$ denotes out-of-sample validation covariates. Predictions are denoted as $\hat{\boldsymbol{f}} = \hat{\boldsymbol{f}}(\boldsymbol{X}) \in \mathbb{R}^n$ (in-sample), $\hat{f}^* = \hat{f}^*(\boldsymbol{x}^*) \in \mathbb{R}$ (out-of-sample), and $\hat{\boldsymbol{f}}_v^* = \hat{\boldsymbol{f}}_v^*(\boldsymbol{X}_v^*) \in \mathbb{R}^{n_v^*}$ (validation), while the true function values are denoted as $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{X}) \in \mathbb{R}^n$ (in-sample) and $f^* = f^*(\boldsymbol{x}^*) \in \mathbb{R}$ (out-of-sample). $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\boldsymbol{X}) \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\varphi}^* = \boldsymbol{\varphi}^*(\boldsymbol{x}^*) \in \mathbb{R}^p$ denote p-dimensional feature expansions of the in- and out-of-sample data respectively, while $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{X}) \in \mathbb{R}^{n\times n}$, $\boldsymbol{s}^* = \boldsymbol{s}^*(\boldsymbol{x}^*, \boldsymbol{X}) \in \mathbb{R}^n$, and $\boldsymbol{S}_v^* = \boldsymbol{S}_v^*(\boldsymbol{X}_v^*, \boldsymbol{X}) \in \mathbb{R}^{n_v^*\times n}$ denote the in-sample, out-of-sample and validation smoother matrices (or vectors). $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix.

3 RELATED WORK

Unsupervised Learning: Unsupervised learning comprises a broad class of methods that share the common goal of learning the structure of unlabeled data. Classical examples of unsupervised learning include *clustering*, such as different flavors of k-means (MacQueen, 1967; Lloyd, 1982), spectral (Ng et al., 2001; von Luxburg, 2007), density-based (Ester et al., 1996; Bhattacharjee and Mitra, 2021), and hierarchical (Murtagh and Contreras, 2012; Campello et al., 2013) clustering; *dimensionality reduction*, including different flavors of principal component analysis (Pearson, 1901; Schölkopf et al., 1997; Zou et al., 2006), autoencoders (Kramer, 1991; Ng et al., 2011; Kingma and Welling, 2013; Allerbo and Jörnsten, 2021) and graph-based methods (Maaten and Hinton, 2008; McInnes et al., 2018); and different approaches to (non-negative) *matrix factorization* (Lee and Seung, 2000; Mnih and Salakhutdinov, 2007; Koren et al., 2009). Another example of unsupervised learning is (deep) *generative models*, including variational autoencoders (Kingma and Welling, 2013), generative adversarial networks (Goodfellow et al., 2014), normalizing flows (Rezende and Mohamed, 2015), and diffusion models (Sohl-Dickstein et al., 2015), where the goal is to learn the distribution of the data, to enable sampling of realistic, synthetic data.

Selfsupervised Learning: In self-supervised learning, rather than using externally provided training labels, the supervisory signal is constructed from the covariate data. This is often accomplished by augmenting or transforming the data to create new samples that are related by construction. Contrastive learning (see e.g. Chen et al. (2020), He et al. (2020), and Wang and Isola (2020)) enforces data generated from the same instance to have similar latent representations, while instance discrimination (see e.g. Wu et al. (2018), and Ibrahim et al. (2024)) assigns pseudo-labels to the augmented data, where the data generated from the same original instance share labels. Using self-supervised methods for identifying the data-generating process has been studied by, e.g., Zimmermann et al. (2021), Hyvärinen et al. (2019), and Reizinger et al. (2024), where the latter two use non-linear independent component analysis (see e.g. Hyvärinen et al. (2023) for an overview). Our work differs from self-supervised learning in that we do not use any supervisory signal during training (not even one based on augmented data). Furthermore, we do not make any assumptions about a generating process that we try to identify.

Model Selection without Validation Data: Hyperparameter selection is usually based on some sort of cross-validation, i.e. by training the model on parts of the data and validating it on other parts. There are, however, many methods for selecting hyperparameters without evaluating the performance on validation data, but, in contrast to our work, they still use the outputs $\boldsymbol{y}$. The most classic methods are probably Mallows's C_p (Mallows, 1964), the Akaike information criterion (AIC) (Akaike, 1973), and the Bayesian information criterion (BIC) (Schwarz, 1978). They all estimate the in-sample prediction error, but since then, many alternatives have been proposed. These include approaches based on the gradient of the loss function (Mahsereci et al., 2017); estimated marginal likelihood (Duvenaud et al., 2016; Lyle et al., 2020); stability of predictions (Lange et al., 2002; Yuan et al., 2025); the neural tangent kernel (Jacot et al., 2018; Xu et al., 2021; Chen et al., 2021; Zhu et al., 2022); unlabeled test data (Garg et al., 2021; Deng and Zheng, 2021; Peng et al., 2024); training speed (Ru et al., 2021; Zhang et al., 2024); topological properties of the model (Corneanu et al., 2020; Ballester et al., 2024); and desired properties of the inferred function (Allerbo and Jörnsten, 2022).

Training with Random Labels: One way to make a model independent of $\boldsymbol{y}$ is to replace $\boldsymbol{y}$ with a random vector. This was investigated by Zhang et al. (2016), who showed that overparameterized neural networks can achieve zero training error when trained with random labels. However, in contrast to our work, in terms of generalization, the resulting models were at the level of random guessing. Since then, training on partly, or completely, random labels has received a lot of attention, where the main focus has been on partly random labels, see e.g. Zhang et al. (2016), Jiang et al. (2018), Zhang and Sabuncu (2018), Han et al. (2019), Song et al. (2022), Wei et al. (2024), Nguyen et al. (2024), and Wang et al. (2024). Training with completely random labels has successfully been used for neural architecture search (Zhang et al., 2021); estimating model complexity (Becker and Risse, 2024); partitioning models across GPUs during training (Karadag and Topaloglu, 2025); and model pretraining (Pondenkandath et al., 2018; Maennel et al., 2020; Behrens, 2025).

Smoother Models: Linear smoothers have a long history and have been studied by e.g. Watson (1964), Reinsch (1967), Rosenblatt (1971), Silverman (1985), and Buja et al. (1989). Due to their more complicated form, nonlinear smoothers are harder to analyze theoretically but have been studied by e.g. Mallows (1980). More recently, Jeffares et al. (2024) showed that deep neural networks trained with the squared loss and infinitesimal learning rate can be formulated as smoothers, where the smoother matrix is a function of the neural tangent kernel, NTK. In some settings, the NTK is constant during training (Jacot et al., 2018; Lee et al., 2019), and thus independent of $\boldsymbol{y}$, in which case the neural network is a linear smoother. However, in practice, the NTK changes during training (Liu et al., 2020; Fort et al., 2020), making it dependent on $\boldsymbol{y}$, and the neural network thus becomes a nonlinear smoother. This stands in contrast to our work, where the neural network smoother matrix is always independent of $\boldsymbol{y}$. Due to the high computational cost of calculating the NTK, computationally efficient methods have been proposed by Novak et al. (2022), Wei et al. (2022), Wang et al. (2022), and Mohamadi et al. (2023), where the last three rely on approximations of the kernel.

In- and Out-of-Sample Performance: A classical measure of model capacity is the effective number of parameters, also known as the effective degrees of freedom, (Efron, 1986, 2004). This is computed using only in-sample data, which limits its usefulness in some situations, see e.g. Janson et al. (2015). Generalizations of the effective number of parameters that also include out-of-sample data have been proposed by Rosset and Tibshirani (2020), Luan et al. (2021), Curth et al. (2023), and Patil et al. (2024). A related topic that has gained a lot of attention is when in- and out-of-sample data follow different distributions, see e.g. Hendrycks et al. (2021), Liu et al. (2021), Harun et al. (2024), Yang et al. (2024), and Cui et al. (2025).

4 MODEL SELECTION WITHOUT USING THE OUTPUTS

In this section, we first discuss why in-sample model complexity is not necessarily a reliable basis for model selection. We then propose a model selection criterion based on the variance of the out-of-sample predictions, which we make independent of $\boldsymbol{y}$.

4.1 Limitations of Model Selection Based on In-Sample Model Complexity

Let us consider model selection by pre-defining the desired model complexity in terms of the effective number of parameters. For linear smoothers, where $\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{y}$, the effective number of parameters is obtained as the trace of $\boldsymbol{S}$. The model complexity thus depends on $\boldsymbol{X}$ and the model parameters, but not on $\boldsymbol{y}$ and $\boldsymbol{x}^*$. Although often working relatively well, this method has some obvious limitations: First, it is not clear *which complexity to choose* to obtain good generalization. Second, when there is more than one model parameter, there may exist several different models with the *same complexity*, but with *different performance*. This is illustrated in Figure 2, where we have used kernel ridge regression with a Gaussian kernel to fit the data, which means that there are two parameters: regularization and kernel bandwidth. All three functions have the same in-sample complexity, and they all perfectly interpolate the training data. However, while the green function provides reasonable out-of-sample predictions, the red and yellow functions greatly over- or underfit between observations. Consequently, in-sample model complexity, even in a two-parameter model, is not sufficient to predict out-of-sample performance.

4.2 Model Selection Based on Out-of-Sample Predictions: Controlling the Variance

For the green function in Figure 2, which generalizes well compared to the red and yellow ones, the out-of-sample predictions roughly follow the same distribution as the observations, i.e. $\hat{f}^* \overset{d}{\approx} y$ (where $\overset{d}{\approx}$ denotes approximately equal in distribution). This also makes sense: for a model to generalize well, the behavior of the predictions (including the out-of-sample predictions) must not deviate too much from that of the data.

Given out-of-sample validation covariates $\boldsymbol{X}_v^*$, for which no $\boldsymbol{y}$-data exists, we propose model selection by matching the empirical distributions of $\boldsymbol{y}$ and $\hat{\boldsymbol{f}}_v^*$ through matching of moments. For centered data, $\frac{1}{n}\sum_{i=1}^{n} y_i = 0$ and $\frac{1}{n_v^*}\sum_{i=1}^{n_v^*} \hat{f}_{v,i}^* \approx 0$, which means that the first sample moments already approximately match. We additionally require the second sample moments,

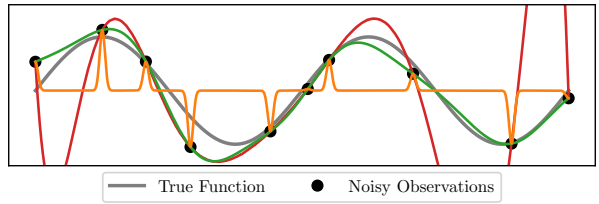

Figure 2: Illustrating the limitations of in-sample model complexity. All three function estimates (green, orange, and red) have the same in-sample model complexity, in terms of the effective number of parameters, but behave very differently between the observations. Details are given in Appendix A.

and hence the sample variances, to match, disregarding higher-order moments for simplicity. For a smoother, model selection thus amounts to minimizing

$$\begin{aligned} &\left| \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n_v^*}\sum_{i=1}^{n_v^*} {\hat{f}_{v,i}^*}^2 \right| \\ &= \left| \boldsymbol{y}^\top \left(\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n_v^*}\boldsymbol{S}_v^{*\top}\boldsymbol{S}_v^* \right) \boldsymbol{y} \right|, \end{aligned} \tag{1}$$

a method we refer to as matching of sample variances, MSV. So far, MSV still depends on $\boldsymbol{y}$; we address this in Section 4.3.

To evaluate MSV, we need to calculate the out-of-sample validation smoother matrix $\boldsymbol{S}_v^*$, for which we need $\boldsymbol{X}_v^*$. In principle, we would like to use all possible future $\boldsymbol{x}$-values, which is, of course, unrealistic, and we need to settle for a sub-sample. One way to create such a sample would be to split $\boldsymbol{X}$ into two parts and use one for training and the other for validation. However, this comes with at least two limitations: First, observations in $\boldsymbol{y}$ corresponding to the validation covariates would then not be considered at all during training, which would be wasteful. Second, we would like the model to generalize well to all possible future values $\boldsymbol{x}$, also those not previously seen. Both of these issues can be solved by sampling $\boldsymbol{X}_v^*$ from a generative model, trained on $\boldsymbol{X}$, something that allows us to create as large an $\boldsymbol{X}_v^*$ as we wish. Throughout this paper, we generate $\boldsymbol{X}_v^*$ by sampling from a multivariate normal distribution, parameterized by the sample mean and covariance of $\boldsymbol{X}$. As we will see, even this extremely simple generative model works well in practice. An alternative to sample $\boldsymbol{X}_v^*$ is to replace the second sample moment, $\frac{1}{n_v^*}\boldsymbol{S}_v^{*\top}\boldsymbol{S}_v^*$, in Equation 1 with the true second moment, $\mathbb{E}_{\boldsymbol{x}^*}[\boldsymbol{s}^*\boldsymbol{s}^{*\top}]$. Whenever $\mathbb{E}_{\boldsymbol{x}^*}[\boldsymbol{s}^*\boldsymbol{s}^{*\top}]$ can be obtained in closed form, which is not always the case, sampling of $\boldsymbol{X}_v^*$ is no longer needed.

4.3 Making Model Selection $\boldsymbol{y}$-free

As can be seen from Equation 1, model selection through MSV amounts to minimizing a closed-form expression of the form $|\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}|$, where $\boldsymbol{A}$ is a symmetric matrix that does not depend on $\boldsymbol{y}$ unless $\boldsymbol{S}_v^*$ does. We can make this expression independent of $\boldsymbol{y}$ by simply replacing $\boldsymbol{y}$ with a random vector, $\boldsymbol{y}_R$. It can be shown (see Appendix F) that when the elements in $\boldsymbol{y}_R$ are uncorrelated, with zero mean and variance σ_y^2, then

$$\left|\mathbb{E}_y\left(\boldsymbol{y}_R^\top \boldsymbol{A}\boldsymbol{y}_R\right)\right| = |\mathrm{Tr}\left(\boldsymbol{A}\right)| \cdot \sigma_y^2 =: \|\boldsymbol{A}\|_T \cdot \sigma_y^2. \quad (2)$$

We can thus obtain $\boldsymbol{y}$-free model selection, by minimizing

$$\left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n_v^*}\boldsymbol{S}_v^{*\top}\boldsymbol{S}_v^*\right\|_T = \left|1 - \frac{1}{n_v^*}\mathrm{Tr}\left(\boldsymbol{S}_v^{*\top}\boldsymbol{S}_v^*\right)\right|,$$

or $\left|1 - \mathrm{Tr}\left(\mathbb{E}(\boldsymbol{s}^*\boldsymbol{s}^{*\top})\right)\right| = \left|1 - \mathbb{E}(\|\boldsymbol{s}^*\|_2^2)\right|$ if the expectation is available, something we refer to as MSV-Tr.[1] An alternative to Equation 2 for making MSV independent of $\boldsymbol{y}$ is to note that

$$\begin{aligned}\left|\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}\right| &= \left\|\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}\right\|_2 \leq \|\boldsymbol{A}\|_2 \cdot \|\boldsymbol{y}\|_2^2 \\ &\leq \|\boldsymbol{A}\|_F \cdot \|\boldsymbol{y}\|_2^2 \leq \|\boldsymbol{A}\|_* \cdot \|\boldsymbol{y}\|_2^2,\end{aligned} \quad (3)$$

where $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_*$ denote the spectral, Frobenius, and nuclear matrix norms, respectively. The expressions in Equations 2 and 3 are very similar, but differ in one important aspect: In contrast to the three proper norms, unless $\boldsymbol{A}$ is positive semi-definite (PSD), $\|\boldsymbol{A}\|_T = |\mathrm{Tr}(\boldsymbol{A})|$ is a seminorm, which means that it may be 0 for matrices other than the zero matrix. When $\boldsymbol{A}$ is PSD, its eigenvalues and singular values coincide, and $\|\boldsymbol{A}\|_T = \|\boldsymbol{A}\|_*$.

5 WHEN DOES $\boldsymbol{y}$-FREE MODEL SELECTION WORK?

In this section, we analyse the MSV-Tr model selection criterion in terms of the bias-variance decomposition of the conditional out-of-sample prediction risk. We also compare the asymptotic risk, as $n, d \to \infty$, for linear ridge regression with regularization selected by MSV-Tr to the theoretically optimal risk. We finally show that, in contrast to MSV, $\boldsymbol{y}$-free generalized cross-validation cannot be used for model selection.

5.1 Analysis of $\boldsymbol{y}$-free MSV-Tr

Assuming that the data is generated as $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, where the elements in $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ are i.i.d. with zero mean and variance σ_ε^2, the conditional out-of-sample risk is then given by

$$R_{\boldsymbol{X}} := \mathbb{E}\left((\hat{f}^* - f^*)^2 | \boldsymbol{X}\right) = B_{\boldsymbol{X}} + V_{\boldsymbol{X}},$$

with bias and variance components $B_{\boldsymbol{X}}$ and $V_{\boldsymbol{X}}$.

According to Proposition 1, MSV-Tr selects the model so that the variance of the inferred out-of-sample functions equals the variance of the measurement noise, which is very reasonable. Since we must not use $\boldsymbol{y}$ when selecting the model parameters, the bias cannot be estimated, and thus the risk must be assessed only through its variance component. Since, in general, both a too small variance (which generally leads to a too high bias) and a too large variance lead to a high risk, using the same variance as the noise is often a good compromise. One may argue that $\boldsymbol{I}_n/n$ in Equation 1 should be multiplied by some $a > 0$ to obtain $|a \cdot \sigma_\varepsilon^2 - V_{\boldsymbol{X}}|$ in Equation 4. The model could then be made more robust (or flexible) by tuning a. This would, however, require estimating the signal-to-noise ratio (SNR), for which we need access to $\boldsymbol{y}$.

Proposition 1. *For a linear smoother $\hat{f}^* = \boldsymbol{s}^*(\hat{\boldsymbol{\theta}})^\top \boldsymbol{y}$, with parameters $\hat{\boldsymbol{\theta}}$,*

$$\begin{aligned}\hat{\boldsymbol{\theta}}_T &:= \arg\min_{\hat{\boldsymbol{\theta}}} \left|1 - \mathbb{E}(\|\boldsymbol{s}^*(\hat{\boldsymbol{\theta}})\|_2^2)\right| \\ &= \arg\min_{\hat{\boldsymbol{\theta}}} |\sigma_\varepsilon^2 - V_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}})|.\end{aligned} \quad (4)$$

Even if Proposition 1 relates the selected model to the observation noise, an explicit value for $\hat{\boldsymbol{\theta}}_T$ is generally inaccessible. However, for linear ridge regression in the asymptotic setting, we can obtain a closed-form expression for the regularization λ, as is done in Theorem 1. We also show that the out-of-sample risk when selecting λ based on MSV-Tr is bounded in terms of the theoretically optimal (inaccessible) risk. To distinguish the asymptotic quantities from the non-asymptotic, we denote them with a superscript bar.

Theorem 1. *For linear ridge regression, where $\boldsymbol{s}^*(\lambda) := \boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top$, if $n, d \to \infty$, such that $d/n \to \gamma \in (0, \infty)$, and the distribution of $\boldsymbol{x}$ has zero mean, unit variance and a finite moment of order $4 + \eta$ for some $\eta > 0$, then*

$$\begin{aligned}\overline{\lambda_T} &:= \arg\min_{\lambda \geq 0} \left|\sigma_\varepsilon^2 - \overline{V_{\boldsymbol{X}}}(\lambda)\right| \\ &= \begin{cases} 3\sqrt{\frac{\gamma}{2}} - \gamma - 1, & \gamma \in \left(\frac{1}{2}, 2\right) \\ 0, & \gamma \in \left(0, \frac{1}{2}\right] \cup [2, \infty), \end{cases}\end{aligned}$$

where $\overline{V_{\boldsymbol{X}}}$ denotes the variance component of the asymptotic conditional out-of-sample risk.

Additionally, for SNR $\in [1, 80]$ *and* $\gamma \in (0, \infty)$,

$$\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}) < 2.45, \quad (5)$$

[1] MSV-Tr is closely related to the generalized effective number of parameters introduced by Curth et al. (2023): $\left|1 - \frac{1}{n_v^*}\mathrm{Tr}\left(\boldsymbol{S}_v^{*\top}\boldsymbol{S}_v^*\right)\right| = \left|1 - n \cdot p_{\hat{\boldsymbol{S}}}^0\right|$, where $p_{\hat{\boldsymbol{S}}}^0$ denotes their generalization of the effective number of parameters.

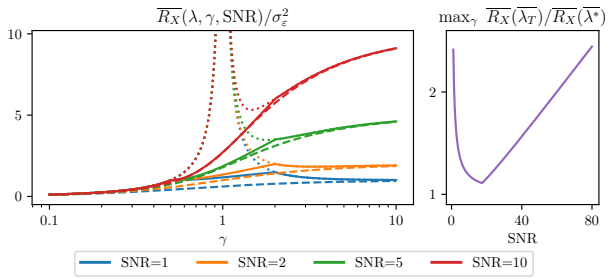

Figure 3: Left: Asymptotic risks, $\overline{R_{\boldsymbol{X}}}$, for different values of $\gamma = \lim_{n,d\to\infty} \frac{d}{n}$ and the signal-to-noise ratio, SNR. For $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})$ we use solid, for $\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*})$ dashed, and for $\overline{R_{\boldsymbol{X}}}(0)$ dotted lines. For $\gamma \notin \left(\frac{1}{2}, 2\right)$, $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})$ coincides with $\overline{R_{\boldsymbol{X}}}(0)$, but for $\gamma \in \left(\frac{1}{2}, 2\right)$, where $\overline{R_{\boldsymbol{X}}}(0)$ tends to deviate substantially form $\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*})$ (and even diverge for $\gamma = 1$), $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})$ stays much closer to $\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*})$. Right: $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*})$ for SNR in $[1, 80]$. In this interval, the quotient is always less than 2.45 (and, for some values of SNR, close to 1), which is expected from Theorem 1.

where $\mathrm{SNR} := \|\boldsymbol{\beta}^*\|_2^2/\sigma_\varepsilon^2$ *(for* $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}^*$*) denotes the signal-to-noise ratio and* $\overline{R_{\boldsymbol{X}}}(\lambda)$ *denotes the asymptotic risk, which is minimized for* $\overline{\lambda^*} = \gamma/\mathrm{SNR}$.

Remark 1. *The assumption of zero mean and unit variance is quite weak, since this can always be obtained by properly rescaling and rotating the data.*

In the left panel of Figure 3, we plot $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})$ for different values of γ and SNR together with $\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*})$ and $\overline{R_{\boldsymbol{X}}}(\lambda = 0)$. $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T})$ follows $\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*})$ quite closely, avoiding the divergence at $\gamma = 1$ that occurs for $\overline{R_{\boldsymbol{X}}}(0)$. In the right panel, we plot the maximum ratio of the MSV-Tr to the optimal asymptotic risk as a function of the SNR. For SNR $\in [1, 80]$, the quotient is less than 2.45, which is in accordance with Equation 5. It is, however, easy to verify that the quotient diverges both for SNR$\to 0$ and SNR$\to \infty$, which suggests that MSV-Tr does not work well for extreme SNRs, unless modified by introducing the a discussed above.

5.2 What About $\boldsymbol{y}$-free Cross Validation?

Cross-validation has no closed-form in general, and can thus not be made independent of $\boldsymbol{y}$ in the same way as MSV. However, leave-one-out cross-validation, LOOCV (Allen, 1974), and its close relative generalized cross-validation, GCV (Golub et al., 1979), have closed-form solutions, where the latter amounts to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}_i}{1 - \mathrm{Tr}(\boldsymbol{S})/n}\right)^2 = n \cdot \frac{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{S})^\top(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{y}}{\left(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S})\right)^2},$$

which can be made independent of $\boldsymbol{y}$ in the same ways as MSV.[2]

However, in Theorem 2, we show that, in contrast to MSV, $\boldsymbol{y}$-free, norm-based GCV cannot be used for selecting the regularization strength λ for ridge regression *regardless* of the feature expansion. The theorem thus includes, e.g., linear and kernel ridge regression.

Theorem 2.
Let $\boldsymbol{S}_\lambda := \boldsymbol{\Phi}\boldsymbol{\Phi}_\lambda^+$ *and* $\boldsymbol{s}_\lambda^* := \boldsymbol{\varphi}^{*\top}\boldsymbol{\Phi}_\lambda^+$*, where*

$$\boldsymbol{\Phi}_\lambda^+ := (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \lambda\boldsymbol{I}_n)^{-1},$$

and let $\|\cdot\|$ *denote the trace seminorm, the nuclear norm, or the Frobenius norm. Then*

(a) for GCV, $\arg\min_{\lambda\geq 0} \frac{\|(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\|}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_\lambda))^2} = \infty$.

(b) for in-sample MSV,
$\arg\min_{\lambda\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda^\top\boldsymbol{S}_\lambda\right\| =$
$\arg\min_{\lambda\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda\right\| = 0$.

(c) for out-of-sample MSV, if $\mathbb{E}(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*\top}) = \boldsymbol{I}_p$*, then*
$\arg\min_{\lambda\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\| < \infty$.
If, in addition, $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$*, then*
$\arg\min_{\lambda\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\| > 0$.

Remark 2. *Thanks to its close connection to ridge regression, it is not difficult to extend Theorem 2 to gradient flow, i.e., gradient descent with infinitesimal learning rate, as is done in Appendix E. Through this extension, we also include, e.g., neural networks via the NTK framework.*

Remark 3. *For the spectral norm, the analog of Theorem 2 becomes slightly more complicated, but the conclusions are the same. See Appendix E for details.*

For the two in-sample methods, GCV and in-sample MSV, $\boldsymbol{y}$-free model selection always selects either infinite (for GCV) or zero (for in-sample MSV) regularization, regardless of the data and the feature expansion, and thus cannot be used for parameter selection. (Note that a training time of 0 corresponds to an infinitely regularized model.) For out-of-sample MSV, things become more interesting. Assuming that the features are isotropic, which, for linear features, can be obtained by properly rescaling and rotating the data, we see that the optimal regularization is always less than infinity, and larger than zero if $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$, which is consistent with Proposition 1.

[2] Since LOOCV has a more complicated form than GCV, it is less amenable to theoretical analysis. However, given the similarities between the methods, it should behave very similarly to GCV. This is also our experience from empirical experiments.

6 $\boldsymbol{y}$-FREE TRAINING OF NEURAL NETWORKS, NEAREST NEIGHBORS, AND RANDOM FORESTS

To apply our developments to neural networks, k-nearest neighbors (kNN), and random forests (RF), these models must first be formulated as linear smoothers. In this section, we show how to do that.

6.1 Neural Networks

Jeffares et al. (2024) expressed neural networks as non-linear smoothers, based on the NTK framework. Our formulation extends their work with two important distinctions: First, in addition to the squared loss, our formulation allows for using the cross-entropy loss, which we use for classification. Second, to obtain a linear smoother, we train the neural network on *non-informative labels/response*, not adding the true $\boldsymbol{y}$ until after all network parameters have been fixed. The details are given in Appendix B.

Calculating the smoother matrix requires calculating the NTK in every iteration when training the network, which can be expensive both in terms of computation and storage. The purpose of formulating a neural network as a linear smoother is, however, not to obtain more efficient training, but to show that it is possible to successfully train the network without using the information in $\boldsymbol{y}$. Another limitation of the smoother formulation is that, for neural networks, the smoother matrix cannot be obtained in closed form, but it is instead computed iteratively during training. Thus, the $\boldsymbol{x}^*$-data of all desired future predictions need to be included during training.

6.2 k-Nearest Neighbors and Random Forests

In the smoother formulations of k-nearest neighbors and decision trees, $\boldsymbol{s}^*(\boldsymbol{x}^*, \boldsymbol{x_i})$ is non-zero only if $\boldsymbol{x_i}$ belongs to the neighborhood of $\boldsymbol{x}^*$ (for k-nearest neighbors), or is in the same leaf as $\boldsymbol{x}^*$ (for decision trees). For random forests, the total smoother is an aggregation of the smoothers of the individual trees. The details are given in Appendix C.

7 EXPERIMENTS

In this section, we investigate $\boldsymbol{y}$-free training for a smoothing spline, linear ridge regression (LRR), kernel ridge regression (KRR), neural network regression (NNR), neural network classification (NNC), k-nearest neighbor regression (kNNR), k-nearest neighbor classification (kNNC), random forest regression (RFR), and random forest classification (RFC) on one synthetic and six real-world data sets, which are presented in Table 1.

Table 1: Real data sets, with number of features, d.

Data Set	d
MNIST (LeCun et al., 1998)	784
CIFAR-10 (Krizhevsky et al., 2009)	3072
Energy consumption in steel production (Sathishkumar et al., 2021)	6
Run time of CPUs (Revow, 1996)	21
Critical temperature of superconductors (Hamidieh, 2018)	81
Power consumption of Tetouan City (Salam and El Hibaoui, 2018)	7

The exact experimental setups are given in Appendix A; here follows a summary: For KRR, we used the Gaussian kernel. For NNR and NNC, we used a one-hidden-layer feed-forward neural network, trained with gradient descent with momentum and the squared (for NNR) or cross-entropy (for NNC) loss.

For the real data sets, we randomly sampled 500 training and 100 test observations. For MSV, we sampled 500 validation observations (only covariates, $\boldsymbol{X}_v^*$) from a multivariate normal distribution parameterized by the sample mean and covariance of the training data. This was repeated 10 times for different subsets of the data.

To select regularization/stopping time/number of neighbors, we compare 10-fold $\boldsymbol{y}$-based cross-validation to Frobenius norm-based MSV. For NNR and NNC on the real data, $\boldsymbol{y}$-based training was instead obtained by using $\boldsymbol{y}$ during training and setting aside 20% of the training data for validating when to stop training (henceforth referred to as the standard method). To train the neural networks and random forests without $\boldsymbol{y}$, we used a random $\boldsymbol{y}_R \in \mathbb{R}^n$, where, for NNR and RFR, $\boldsymbol{y}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and, for NNC and RFC, $(\boldsymbol{y}_R)_i \sim \mathrm{Cat}(c)$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution and $\mathrm{Cat}(c)$ the categorical distribution for c categories of equal probability. For the neural networks, we monitored the Frobenius norm-based MSV and GCV to decide when to stop training. For the random forests, we used the default parameters of the scikit-learn implementation.

The results are displayed in Figure 4 and Table 2. In the table, for each method, we present the median (Q2) and first (Q1) and third (Q3) quartiles, over the 10 realizations, of the accuracy (for classification) or R^2 (for regression) [3] on the test data. In all cases, $\boldsymbol{y}$-free MSV

[3] $R^2 := 1 - \|\boldsymbol{y} - \hat{\boldsymbol{f}}\|_2^2 / \|\boldsymbol{y} - \bar{y}\|_2^2 \leq 1$, where $\bar{y}$ denotes

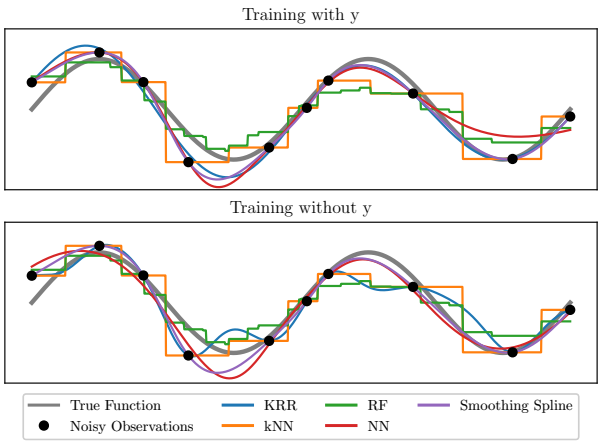

Figure 4: The results of training different models with and without $\boldsymbol{y}$ on synthetic data. The models trained without $\boldsymbol{y}$ perform similarly to those trained with $\boldsymbol{y}$. In the bottom panel, all five models are expressed in the form $\hat{f}^* = \boldsymbol{s}^{*\top}\boldsymbol{y}$, where $\boldsymbol{s}^*$ is constructed independently of $\boldsymbol{y}$.

performs much better than random guessing, most of the time at the level of the $\boldsymbol{y}$-based methods, and sometimes even slightly better. $\boldsymbol{y}$-free GCV performs at the level of random guessing, which corresponds to an accuracy of 0.1 for classification, since both MNIST and CIFAR-10 have ten different classes, and to an R^2 of 0 for regression. This is expected from Theorem 2, according to which, $\boldsymbol{y}$-free GCV-based training uses infinite regularization (for neural networks in the form of zero training time).[4,5]

The larger gap between $\boldsymbol{y}$-based and $\boldsymbol{y}$-free training of random forests compared to the other models can be attributed to the importance of $\boldsymbol{y}$ when selecting the splits in the decision trees.[6] The limited performance of NNC on MNIST and CIFAR-10, compared to state-of-the-art, can be attributed to the fact that we only use 500 training observations, due to the high computational cost of calculating the smoother matrix for neural networks.

the mean of $\boldsymbol{y}$, measures the proportion of the variance in $\boldsymbol{y}$ that is explained by the model, and is a normalized version of the mean squared error. $R^2 = 1$ corresponds to a perfect fit, while a model that predicts $\hat{f}_i = \bar{y}$ for all i results in $R^2 = 0$; thus a negative value of R^2 is possible, but it corresponds to a model that performs worse than always predicting the mean of the observed data.

[4]Technically, Theorem 2 does not apply to kNN. However, k is a sort of regularization, since a larger k results in a more stable model. For kNN, $\boldsymbol{y}$-free GCV selects $k = n$, which is analog to $\lambda = \infty$.

[5]$\boldsymbol{y}$-free LOOCV performs identically to $\boldsymbol{y}$-free GCV on all methods and data, and is thus not reported separately.

[6]We also investigated using a vector of only zeros, rather than $\boldsymbol{y}_R$, for $\boldsymbol{y}$-free training. For neural networks, the results were virtually identical, but for RF the results of using $\mathbf{0}$ were at the level of random guessing, which is not surprising: If all y values are identical, the decision tree consists of a single leaf containing all training data, and the regression model simply predicts the mean, while the classification model always predicts the same class.

8 CONCLUSIONS

We demonstrated how standard supervised models, including neural networks, can be successfully trained in an unsupervised manner. We did this by formulating each model as a smoother, i.e. on the form $\hat{f}^* = \boldsymbol{s}^{*\top}\boldsymbol{y}$, and constructing the smoother, $\boldsymbol{s}^*$, without using the information in $\boldsymbol{y}$. To construct the smoother independently of $\boldsymbol{y}$, we introduced a model selection criterion, MSV, that is based on the out-of-sample predictions of the model, but independent of $\boldsymbol{y}$. For iteratively trained models, including neural networks, we replaced the true outputs, $\boldsymbol{y}$, with either a random vector or the zero vector during training.

We showed how MSV can be expressed in terms of the variance component of the out-of-sample risk, and, for linear ridge regression with $n, d \to \infty$, bounded the asymptotic risk of MSV in terms of the optimal asymptotic risk.

Using both real and synthetic data, we showed how five different linear and nonlinear models for classification and regression, trained without $\boldsymbol{y}$, perform on par with the standard, $\boldsymbol{y}$-based versions, and much better than random guessing. This suggests that, under the hood, supervised and unsupervised learning are closely related, and also supports the hypothesis that, given $\boldsymbol{X}$, $\boldsymbol{x}^*$, and $\boldsymbol{y}$, the optimal model is a weighted sum of $\boldsymbol{y}$, with weights that depend on $\boldsymbol{X}$ and $\boldsymbol{x}^*$ only.

We do not suggest that $\boldsymbol{y}$ should be excluded from the training process in practice, partly because the smoother formulation scales poorly with the data. This holds especially true for neural networks, where the smoother matrix depends on the neural tangent kernel, which needs to be repeatedly recalculated during training.

While our model selection criterion performs well in practice, we have no evidence that it is optimal, leaving room for future improvements in $\boldsymbol{y}$-free training.

Acknowledgements

This research was supported by the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by Knut and Alice Wallenberg Foundation, and by the *Kjell och Märta Beijer Foundation.*

Table 2: Median and first and third quartile, over the 10 splits, of the accuracy (for classification) or R^2 (for regression) on the test data. The models trained with and without $\boldsymbol{y}$ perform similarly. $\boldsymbol{y}$-free GCV performs at the level of random guessing.

Data	Model	Method	Accuracy/R^2 Q2 (Q1, Q3)
MNIST	NNC	Standard	0.82 (0.78, 0.85)
		MSV	0.75 (0.74, 0.79)
		GCV	0.11 (0.10, 0.12)
	kNNC	Standard	0.78 (0.76, 0.79)
		MSV	0.76 (0.73, 0.79)
		GCV	0.09 (0.08, 0.13)
	RFC	Standard	0.83 (0.79, 0.85)
		Random $\boldsymbol{y}$	0.29 (0.25, 0.34)
CIFAR-10	NNC	Standard	0.32 (0.30, 0.34)
		MSV	0.33 (0.29, 0.38)
		GCV	0.09 (0.07, 0.11)
	kNNC	Standard	0.23 (0.19, 0.25)
		MSV	0.25 (0.18, 0.26)
		GCV	0.12 (0.10, 0.14)
	RFC	Standard	0.33 (0.31, 0.37)
		Random $\boldsymbol{y}$	0.21 (0.16, 0.25)
Steel Production	LRR	CV	0.98 (0.98, 0.99)
		MSV	0.99 (0.98, 0.99)
		GCV	0.00 (-0.02, 0.00)
	KRR	CV	1.00 (1.00, 1.00)
		MSV	0.95 (0.92, 0.96)
		GCV	0.00 (-0.02, 0.00)
	NNR	Standard	0.95 (0.95, 0.96)
		MSV	0.98 (0.97, 0.98)
		GCV	0.00 (-0.02, 0.00)
	kNNR	CV	0.98 (0.98, 0.99)
		MSV	0.98 (0.98, 0.99)
		GCV	0.98 (0.98, 0.99)
		GCV	0.00 (-0.03, 0.00)
	RFR	Standard	0.99 (0.99, 0.99)
		Random $\boldsymbol{y}$	0.82 (0.79, 0.85)
CPU Run Time	LRR	CV	0.68 (0.63, 0.72)
		MSV	0.71 (0.67, 0.74)
		GCV	0.00 (0.00, 0.00)
	KRR	CV	0.91 (0.89, 0.93)
		MSV	0.65 (0.51, 0.76)
		GCV	0.00 (0.00, 0.00)

Data	Model	Method	Accuracy/R^2 Q2 (Q1, Q3)
CPU Run Time	NNR	Standard	0.69 (0.63, 0.72)
		MSV	0.71 (0.64, 0.80)
		GCV	0.00 (0.00, 0.00)
	kNNR	CV	0.85 (0.65, 0.93)
		MSV	0.83 (0.68, 0.92)
		GCV	0.00 (0.00, 0.00)
	RFR	Standard	0.96 (0.95, 0.98)
		Random $\boldsymbol{y}$	0.34 (0.25, 0.42)
Superconductors	LRR	CV	0.71 (0.65, 0.73)
		MSV	0.71 (0.65, 0.74)
		GCV	0.01 (0.01, 0.01)
	KRR	CV	0.76 (0.71, 0.81)
		MSV	0.73 (0.70, 0.76)
		GCV	0.00 (0.00, 0.00)
	NNR	Standard	0.67 (0.65, 0.70)
		MSV	0.72 (0.68, 0.73)
		GCV	0.00 (0.00, 0.00)
	kNNR	CV	0.73 (0.71, 0.74)
		MSV	0.74 (0.72, 0.76)
		GCV	0.00 (0.00, 0.00)
	RFR	Standard	0.80 (0.77, 0.83)
		Random $\boldsymbol{y}$	0.51 (0.47, 0.53)
Power Consumption	LRR	CV	0.61 (0.57, 0.68)
		MSV	0.61 (0.57, 0.68)
		GCV	0.00 (0.00, 0.00)
	KRR	CV	0.79 (0.77, 0.81)
		MSV	0.75 (0.74, 0.79)
		GCV	0.00 (0.00, 0.00)
	NNR	Standard	0.69 (0.67, 0.73)
		MSV	0.68 (0.59, 0.71)
		GCV	0.00 (0.00, 0.00)
	kNNR	CV	0.73 (0.69, 0.76)
		MSV	0.74 (0.71, 0.76)
		GCV	0.00 (-0.01, 0.00)
	RFR	Standard	0.77 (0.76, 0.80)
		Random $\boldsymbol{y}$	0.52 (0.49, 0.54)

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *technometrics*, 16(1):125–127.

Allerbo, O. and Jörnsten, R. (2021). Non-linear, sparse dimensionality reduction via path lasso penalized autoencoders. *Journal of Machine Learning Research*, 22(283):1–28.

Allerbo, O. and Jörnsten, R. (2022). Bandwidth selection for gaussian kernel ridge regression via jacobian control. *arXiv preprint arXiv:2205.11956*.

Ballester, R., Clemente, X. A., Casacuberta, C., Madadi, M., Corneanu, C. A., and Escalera, S. (2024). Predicting the generalization gap in neural networks using topological data analysis. *Neurocomputing*, 596:127787.

Becker, M. and Risse, B. (2024). Learned random label predictions as a neural network complexity metric. *arXiv preprint arXiv:2411.19640*.

Behrens, F. (2025). Knowledge distillation for random data: Soft labels and similarity scores may contain memorized information. In *New Frontiers in Associative Memories*.

Bhattacharjee, P. and Mitra, P. (2021). A survey of density based clustering algorithms. *Frontiers of Computer Science*, 15:1–27.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Chen, W., Gong, X., and Wang, Z. (2021). Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. *arXiv preprint arXiv:2102.11535*.

Corneanu, C. A., Escalera, S., and Martinez, A. M. (2020). Computing the testing error without a testing set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2685.

Cui, T., Tang, C., Zhou, D., Li, Y., Gong, X., Ouyang, W., Su, M., and Zhang, S. (2025). Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. *Nature Communications*, 16(1):1891.

Curth, A., Jeffares, A., and van der Schaar, M. (2023). A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, 36:55932–55962.

Deng, W. and Zheng, L. (2021). Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15069–15078.

Duvenaud, D., Maclaurin, D., and Adams, R. (2016). Early stopping as nonparametric variational inference. In *Artificial intelligence and statistics*, pages 1070–1077. PMLR.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, 81(394):461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Fort, S., Dziugaite, G. K., Paul, M., Kharaghani, S., Roy, D. M., and Ganguli, S. (2020). Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861.

Garg, S., Balakrishnan, S., Kolter, Z., and Lipton, Z. (2021). Ratt: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning*, pages 3598–3609. PMLR.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354.

Han, J., Luo, P., and Wang, X. (2019). Deep self-learning from noisy labels. In *Proceedings of the*

IEEE/CVF international conference on computer vision, pages 5138–5147.

Harun, M. Y., Lee, K., Gallardo, G., Krishnan, G., and Kanan, C. (2024). What variables affect out-of-distribution generalization in pretrained models? *Advances in Neural Information Processing Systems*, 37:56479–56525.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349.

Hyvärinen, A., Khemakhem, I., and Morioka, H. (2023). Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10).

Hyvärinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR.

Ibrahim, M., Klindt, D., and Balestriero, R. (2024). Occam's razor for self supervised learning: What is sufficient to learn good representations? *arXiv preprint arXiv:2406.10743*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Janson, L., Fithian, W., and Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485.

Jeffares, A., Curth, A., and van der Schaar, M. (2024). Deep learning through a telescoping lens: A simple model provides empirical insights on grokking, gradient boosting & beyond. *Advances in Neural Information Processing Systems*, 37:123498–123533.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR.

Karadag, C. V. and Topaloglu, N. (2025). Partitioned neural network training via synthetic intermediate labels. *Multimedia Tools and Applications*, pages 1–20.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Lange, T., Braun, M., Roth, V., and Buhmann, J. (2002). Stability-based model selection. *Advances in neural information processing systems*, 15.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32.

Liu, C., Zhu, L., and Belkin, M. (2020). On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33:15954–15964.

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021). Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137.

Luan, B., Lee, Y., and Zhu, Y. (2021). Predictive model degrees of freedom in linear regression. *arXiv preprint arXiv:2106.15682*.

Lyle, C., Schut, L., Ru, R., Gal, Y., and van der Wilk, M. (2020). A bayesian perspective on training speed and model selection. *Advances in neural information processing systems*, 33:10396–10408.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298. University of California press.

Maennel, H., Alabdulmohsin, I. M., Tolstikhin, I. O., Baldock, R., Bousquet, O., Gelly, S., and Keysers, D. (2020). What do neural networks learn when trained with random labels? *Advances in Neural Information Processing Systems*, 33:19693–19704.

Mahsereci, M., Balles, L., Lassner, C., and Hennig, P. (2017). Early stopping without a validation set. *arXiv preprint arXiv:1703.09580*.

Mallows, C. (1964). Choosing variables in a linear regression: A graphical aid. In *Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS, 1964*.

Mallows, C. L. (1980). Some theory of nonlinear smoothers. *The Annals of statistics*, pages 695–715.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mnih, A. and Salakhutdinov, R. R. (2007). Probabilistic matrix factorization. *Advances in neural information processing systems*, 20.

Mohamadi, M. A., Bae, W., and Sutherland, D. J. (2023). A fast, well-founded approximation to the empirical neural tangent kernel. In *International Conference on Machine Learning*, pages 25061–25081. PMLR.

Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

Ng, A. et al. (2011). Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19.

Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Nguyen, T., Ibrahim, S., and Fu, X. (2024). Noisy label learning with instance-dependent outliers: Identifiability via crowd wisdom. *Advances in Neural Information Processing Systems*, 37:97261–97298.

Novak, R., Sohl-Dickstein, J., and Schoenholz, S. S. (2022). Fast finite width neural tangent kernel. In *International Conference on Machine Learning*, pages 17018–17044. PMLR.

Patil, P., Du, J.-H., and Tibshirani, R. J. (2024). Revisiting optimism and model complexity in the wake of overparameterized machine learning. *arXiv preprint arXiv:2410.01259*.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Peng, R., Zou, H., Wang, H., Zeng, Y., Huang, Z., and Zhao, J. (2024). Energy-based automated model evaluation. *arXiv preprint arXiv:2401.12689*.

Pondenkandath, V., Alberti, M., Puran, S., Ingold, R., and Liwicki, M. (2018). Leveraging random label memorization for unsupervised pre-training. *arXiv preprint arXiv:1811.01640*.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183.

Reizinger, P., Bizeul, A., Juhos, A., Vogt, J. E., Balestriero, R., Brendel, W., and Klindt, D. (2024). Cross-entropy is all you need to invert the data generating process. *arXiv preprint arXiv:2410.21869*.

Revow, M. (1996). comp-activ dataset. `http://www.cs.toronto.edu/~delve/data/comp-activ/desc.html`.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Rosenblatt, M. (1971). Curve estimates. *The Annals of Mathematical Statistics*, 42(6):1815–1842.

Rosset, S. and Tibshirani, R. J. (2020). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*.

Ru, R., Lyle, C., Schut, L., Fil, M., van der Wilk, M., and Gal, Y. (2021). Speedy performance estimation for neural architecture search. *Advances in Neural Information Processing Systems*, 34:4079–4092.

Salam, A. and El Hibaoui, A. (2018). Comparison of machine learning algorithms for the power consumption prediction:-case study of Tetouan City-. In *2018 6th international renewable and sustainable energy conference (IRSEC)*, pages 1–5. IEEE.

Sathishkumar, V. E., Shin, C., and Cho, Y. (2021). Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city. *Building Research & Information*, 49(1):127–143.

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17:395–416.

Wang, H., Huang, W., Wu, Z., Tong, H., Margenot, A. J., and He, J. (2022). Deep active learning by leveraging training dynamics. *Advances in Neural Information Processing Systems*, 35:25171–25184.

Wang, J., Xia, X., Lan, L., Wu, X., Yu, J., Yang, W., Han, B., and Liu, T. (2024). Tackling noisy labels with network parameter additive decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wei, A., Hu, W., and Steinhardt, J. (2022). More than a toy: Random matrix models predict how real-world neural representations generalize. In *International conference on machine learning*, pages 23549–23588. PMLR.

Wei, T., Li, H.-T., Li, C., Shi, J.-X., Li, Y.-F., and Zhang, M.-L. (2024). Vision-language models are strong noisy label detectors. *Advances in Neural Information Processing Systems*, 37:58154–58173.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Xu, J., Zhao, L., Lin, J., Gao, R., Sun, X., and Yang, H. (2021). KNAS: green neural architecture search. In *International Conference on Machine Learning*, pages 11613–11625. PMLR.

Yang, J., Zhou, K., Li, Y., and Liu, Z. (2024). Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662.

Yuan, S., Feng, L., and Liu, T. (2025). Early stopping against label noise without validation data. *arXiv preprint arXiv:2502.07551*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, X., Hou, P., Zhang, X., and Sun, J. (2021). Neural architecture search with random labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10907–10916.

Zhang, Y., Hsieh, J., Li, X., Chang, M.-C., Lee, C.-C., and Fan, K.-C. (2024). MOTE-NAS: Multi-objective training-based estimate for efficient neural architecture search. *Advances in Neural Information Processing Systems*, 37:100845–100869.

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhu, Z., Liu, F., Chrysos, G., and Cevher, V. (2022). Generalization properties of NAS under activation and skip connection search. *Advances in Neural Information Processing Systems*, 35:23551–23565.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International conference on machine learning*, pages 12979–12990. PMLR.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

In Appendices A, B, and D, we use as superscript * to denote the concatenation of in-sample (training) and out-of-sample quantities, i.e., $\hat{\boldsymbol{f}}^\star := \begin{bmatrix} \hat{\boldsymbol{f}} \\ \hat{\boldsymbol{f}}^* \end{bmatrix}$, $\boldsymbol{X}^\star := \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}^* \end{bmatrix}$, and $\boldsymbol{S}^\star := \begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{S}^* \end{bmatrix}$.

A EXPERIMENTAL DETAILS

A.1 Details for Figures 2 and 4

The true function is given by $y = \sin(2\pi x)$. We sampled $n = 10$ training observations according to $x_i \sim \mathcal{U}(-1, 1)$, $y_i \sim \mathcal{N}(\sin(2\pi x_i), 0.3^2)$, where $\mathcal{U}(a, b)$ denotes the uniform distribution on $[a, b]$, and $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. In addition, we used $n^* = 1000$ uniformly spaced out-of-sample data points on the interval [-1,1].

For kernel ridge regression, we used the Gaussian kernel, $k(\boldsymbol{x}, \boldsymbol{x'}, \sigma) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x'}\|_2^2}{2\sigma^2}\right)$. The predictions are given in closed form by $\hat{\boldsymbol{f}}^\star = \boldsymbol{K}^\star \left(\boldsymbol{K} + \lambda \boldsymbol{I}_n\right)^{-1} \boldsymbol{y}$, where $\boldsymbol{K}^\star = [\boldsymbol{K}^\top, \boldsymbol{K}^{*\top}]^\top \in \mathbb{R}^{(n+n^*)\times n}$, and $\boldsymbol{K}^\star_{ij} = k(\boldsymbol{x}^\star_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{j}}, \sigma)$.

For Figure 2 we used $\lambda = 0$ and $\sigma \in \{1.3, 0.16, 0.01\}$.

For the smoothing spline, we used B-splines of degree 3 to create the basis matrix $\boldsymbol{B}^\star = [\boldsymbol{B}^\top, \boldsymbol{B}^{*\top}]^\top \in \mathbb{R}^{(n+n^*)\times(n+4)}$, and the penalty matrix $\boldsymbol{\Omega} \in \mathbb{R}^{(n+4)\times(n+4)}$, see e.g. Chapter 5 in Hastie et al. (2009) for details. The predictions are given in closed form by $\hat{\boldsymbol{f}}^\star = \boldsymbol{B}^\star \left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{y}$.

For the neural network, we used one hidden layer with 20 nodes and the tanh activation function, trained with gradient descent with momentum, with learning rate $\eta = 0.01$, momentum $\gamma = 0.95$, and the squared loss.

We used 100 logarithmically spaced candidate values between 10^{-4} and 1 for λ and σ respectively, and $k \in [1, 10]$ for k-nearest neighbors. For $\boldsymbol{y}$-based training, we selected λ, σ, and k via leave-one-out cross-validation, and the stopping epoch via a 80/20% split of the data into training/validation sets. For $\boldsymbol{y}$-free training, we used Frobenius norm-based MSV and trained the neural network and the random forest on $\boldsymbol{y}_R \in \mathbb{R}^n$, where $\boldsymbol{y}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.

A.2 Details for the Real Data in Section 7

For KRR we used the Gaussian kernel, $k(\boldsymbol{x}, \boldsymbol{x'}, \sigma) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x'}\|_2^2}{2\sigma^2}\right)$. For NNR and NNC, we used one hidden layer with 200 nodes and the tanh activation function, trained with gradient descent with momentum, with learning rate $\eta = 10^{-4}$, and momentum $\gamma = 0.7$. For NNR, we used the squared loss, and for NNC, the cross-entropy loss. For each data set, the covariates, $\boldsymbol{X}^\star$, were standardized to zero mean and unit variance. For regression, the response, $\boldsymbol{y}$, was standardized to zero mean, and for classification, the labels were one-hot encoded. We randomly sampled 500 training and 100 test observations. For MSV, we sampled 500 validation observations (only covariates, $\boldsymbol{X}^*_v$) from a multivariate normal distribution parameterized by the sample mean and covariance of the training data. This was repeated 10 times for different subsets of the data.

For model selection, we compare generalized (for the synthetic data) or 10-fold (for the real data) $\boldsymbol{y}$-based cross-validation to Frobenius norm-based MSV. For NNR and NNC on the real data, $\boldsymbol{y}$-based training was instead obtained by using $\boldsymbol{y}$ during training and setting aside 20% of the training data for validating when to stop training (henceforth referred to as the standard method). To train the neural networks and random forests without $\boldsymbol{y}$, we used a random $\boldsymbol{y}_R \in \mathbb{R}^n$, where, for NNR and RFR, $\boldsymbol{y}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and, for NNC and RFC, $(\boldsymbol{y}_R)_i \sim \mathrm{Cat}(k)$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution and $\mathrm{Cat}(k)$ the categorical distribution for k categories of equal probability. For the neural networks, we monitored the Frobenius norm-based MSV and GCV to decide when to stop training. For the random forests, we used the default parameters of the scikit-learn implementation.

For LRR, KRR, and kNN, we compared three different model selection methods: 10-fold, $\boldsymbol{y}$-based cross-validation, and $\boldsymbol{y}$-free, Frobenius norm-based MSV and GCV. For LRR and KRR, we evaluated 200 logarithmically spaced candidate values between 10^{-4} and 20, plus 10^6, for λ and σ, respectively. For kNN, we investigated k between 2 and 30, plus $k = n$ (for 10-fold cross-validation the maximum k was $9/10 \cdot n$). For $\boldsymbol{y}$-based training of the neural network models, we used $\boldsymbol{y}$ and set aside 20% of the training data for validating when to stop training (henceforth referred to as the standard method). To train the neural networks and random forests without $\boldsymbol{y}$, we used a random $\boldsymbol{y}_R \in \mathbb{R}^n$, where, for NNR and RFR, $\boldsymbol{y}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and, for NNC and RFC, $(\boldsymbol{y}_R)_i \sim \mathrm{Cat}(c)$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution and $\mathrm{Cat}(c)$ the categorical distribution for c categories of equal

probability. For NNC, $\boldsymbol{y}_R$, was transformed into a $n \times (c-1)$ one-hot matrix (see Appendix D.2 for details). For the neural networks, we monitored the Frobenius norm-based MSV and GCV to decide when to stop training. For the random forests, we used the default parameters of Python's scikit-learn implementation.

Each experiment took between 0.1 seconds for LRR and 45 minutes for NNC on the CIFAR-10 data to run on an Intel i9-13980HX processor with access to 32 GB RAM. The total run time, for all experiments, was approximately 10 hours.

The MNIST data are available at `http://yann.lecun.com/exdb/mnist/` under the CC-BY-SA 3.0 license.
The CIFAR-10 data are available at `https://www.cs.toronto.edu/~kriz/cifar.html` under the MIT license.
The Steel Production data are available at
`https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption`
under the CC-BY 4.0 license. We used the features `Lagging_Current_Reactive.Power_kVarh`, `Leading_Current_Reactive_Power_kVarh`, `CO2(tCO2)`, `Lagging_Current_Power_Factor`, `Leading_Current_Power_Factor`, and `NSM` to predict `Usage_kWh`.
The CPU Run Time data are available at
`http://www.cs.toronto.edu/~delve/data/comp-activ/desc.html`. We excluded the features `usr`, `sys`, `wio`, and `idle`.
The Superconductor data are available at
`https://archive.ics.uci.edu/dataset/464/superconductivty+data`
under the CC-BY 4.0 license.
The Power Consumption data are available at
`https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city`
under the CC-BY 4.0 license. We excluded the `DateTime` feature, using the remaining features to predict `Zone 3 Power Consumption`.

B $\boldsymbol{y}$-FREE TRAINING OF NEURAL NETWORKS IN MORE DETAIL

In Theorem 3, we show how we can construct a neural network smoother matrix iteratively during training. We thus obtain a model in the form $\boldsymbol{S}^\star(\hat{\boldsymbol{\theta}}) \cdot \boldsymbol{y}$, where $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ denotes the model parameters, that approximates $\hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}})$ to arbitrary precision, given a small enough learning rate.

Since $\hat{\boldsymbol{\theta}}$ depends on $\boldsymbol{y}$, so does $\boldsymbol{S}^\star$. To make $\boldsymbol{S}^\star$ independent of $\boldsymbol{y}$, we can simply replace $\boldsymbol{y}$ with a random, or zero, label/response vector during training, i.e. instead of $\hat{\boldsymbol{f}}^\star = \boldsymbol{S}^\star(\hat{\boldsymbol{\theta}}(\boldsymbol{y})) \cdot \boldsymbol{y}$, we use $\hat{\boldsymbol{f}}^\star = \boldsymbol{S}^\star(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_R)) \cdot \boldsymbol{y}$, where $\boldsymbol{y}_R$ is a vector sampled at random, or $\hat{\boldsymbol{f}}^\star = \boldsymbol{S}^\star(\hat{\boldsymbol{\theta}}(\boldsymbol{0})) \cdot \boldsymbol{y}$. To decide when to stop training, i.e. which $\boldsymbol{S}^\star(\hat{\boldsymbol{\theta}}_k)$ to use, we can monitor the $\boldsymbol{y}$-free version of MSV during training.

Theorem 3.
Let the neural network be trained with gradient descent with momentum, with learning rate $\eta > 0$, and momentum $\gamma \in [0, 1)$, with either the squared or the cross-entropy loss.
Let $\boldsymbol{S}_k^\star$ be updated according to

$$\begin{aligned}
&\boldsymbol{S}_{-1}^\star = \boldsymbol{S}_0^\star = \boldsymbol{0}, \\
&\boldsymbol{S}_{k+1}^\star = \boldsymbol{S}_k^\star + \gamma \cdot \left(\boldsymbol{S}_k^\star - \boldsymbol{S}_{k-1}^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_{k+1}^\star \cdot \left(\boldsymbol{I}_n - \boldsymbol{S}_k\right), \\
&k = 0, 1, \ldots
\end{aligned}$$

where $\tilde{\boldsymbol{K}}_k^\star$ denotes a generalized time-dependent NTK (see Remark 5).
Then, if all derivatives are bounded, there exists a constant, $C < \infty$, such that

$$\left\| \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_k) - \left(\boldsymbol{S}_k^\star \left(\boldsymbol{y} - \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}_0) \right) + \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_0) \right) \right\|_\infty \leq \eta \cdot C. \tag{6}$$

Remark 4. *If $\hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_0) = \boldsymbol{0}$, Equation 6 simplifies to $\left\| \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{S}_k^\star \boldsymbol{y} \right\|_\infty \leq \eta \cdot C$.*

Remark 5. *For the square loss, $\tilde{\boldsymbol{K}}_k^\star = \boldsymbol{K}_k^{\star\mathrm{NTK}} = \left(\partial_{\hat{\boldsymbol{\theta}}_k} \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_k)\right) \left(\partial_{\hat{\boldsymbol{\theta}}_k} \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}_k)\right)^\top$ is the standard time-dependent NTK. For the cross-entropy loss, $\tilde{\boldsymbol{K}}_k^\star = \boldsymbol{K}_k^{\star\mathrm{NTK}} \cdot \tilde{\boldsymbol{F}}_k$, where $\tilde{\boldsymbol{F}}_k$ is a square, block-diagonal matrix that depends on $\hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}_k)$. The details are given in Appendix D.*

Remark 6. *Since* $\tilde{\boldsymbol{K}}_k^\star = \tilde{\boldsymbol{K}}^\star(\hat{\boldsymbol{\theta}}_k)$, $\boldsymbol{S}_k^\star = \boldsymbol{S}^\star\left(\{\tilde{\boldsymbol{K}}_l^\star\}_{l=0}^k\right) = \boldsymbol{S}^\star\left(\{\hat{\boldsymbol{\theta}}_l\}_{l=0}^k\right)$, *i.e.* $\boldsymbol{S}_k^\star$ *depends on all historical values of* $\hat{\boldsymbol{\theta}}$ *during training.*

Remark 7. *When* $\tilde{\boldsymbol{K}}^\star$ *is constant during training, the smoother formulation is exact, i.e.* $C = 0$ *in Equation 6. This is the case, for instance, when solving linear or kernel regression with gradient descent.*

C $\boldsymbol{y}$-FREE TRAINING OF k-NEAREST NEIGHBORS AND RANDOM FORESTS

We formulate k-nearest neighbors (kNN) and random forests (RF) as linear smoothers for regression, in which case the predicted value is the average of the training data. For classification, the average is simply replaced by the mode (i.e. the most common value).

For k-nearest neighbors, each prediction is defined as the average of the k closest training points:

$$\hat{f}_{NN}^*(\boldsymbol{x}^*) := \frac{1}{k} \cdot \sum_{i:\boldsymbol{x_i} \in N_k(\boldsymbol{x}^*)} y_i,$$

where $N_k(\boldsymbol{x_i})$ is the neighborhood of $\boldsymbol{x}^*$, consisting of the k closest training points $\boldsymbol{x_i}$.

Defining

$$(\boldsymbol{s}_{NN}^*(\boldsymbol{x}^*, \boldsymbol{X}))_i = \boldsymbol{s}_{NN}^*(\boldsymbol{x}^*, \boldsymbol{x_i}) := \begin{cases} \frac{1}{k} \text{ if } \boldsymbol{x_i} \in N_k(\boldsymbol{x}^*), \\ 0 \text{ if } \boldsymbol{x_i} \notin N_k(\boldsymbol{x}^*), \end{cases}$$

we obtain

$$\hat{f}_{NN}^*(\boldsymbol{x}^*) = \frac{1}{k} \cdot \sum_{i:\boldsymbol{x_i} \in N_k(\boldsymbol{x}^*)} y_i = \boldsymbol{s}_{NN}^{*\top} \boldsymbol{y}.$$

For decision trees, the input space is split into M non-overlapping regions, where each prediction is defined as the average of the training points that belong to the same region as the point of interest:

$$\hat{f}_T^*(\boldsymbol{x}^*) := \frac{1}{|R(\boldsymbol{x}^*)|} \cdot \sum_{i:\boldsymbol{x_i} \in R(\boldsymbol{x}^*)} y_i,$$

where $R(\boldsymbol{x}^*)$ is the region that $\boldsymbol{x}^*$ belongs to, and $|R(\boldsymbol{x}^*)|$ is the number of training points $\boldsymbol{x_i}$ in $R(\boldsymbol{x}^*)$. Defining

$$(\boldsymbol{s}_T^*(\boldsymbol{x}^*, \boldsymbol{X}))_i = \boldsymbol{s}_T^*(\boldsymbol{x}^*, \boldsymbol{x_i}) := \begin{cases} \frac{1}{|R(\boldsymbol{x}^*)|} \text{ if } \boldsymbol{x_i} \in R(\boldsymbol{x}^*), \\ 0 \text{ if } \boldsymbol{x_i} \notin R(\boldsymbol{x}^*), \end{cases}$$

we obtain

$$\hat{f}_T^*(\boldsymbol{x}^*) = \frac{1}{|R(\boldsymbol{x}^*)|} \cdot \sum_{i:\boldsymbol{x_i} \in R(\boldsymbol{x}^*)} y_i = \boldsymbol{s}_T^{*\top} \boldsymbol{y}.$$

The random forest prediction is the average of the predictions of n_T trees, trained on different bootstrap samples of the training data:

$$\hat{f}_{RF}^*(\boldsymbol{x}^*) := \frac{1}{n_T} \cdot \sum_{t=1}^{n_T} \left(\frac{1}{|R_t(\boldsymbol{x}^*)|} \cdot \sum_{i:\boldsymbol{x_i} \in R_t(\boldsymbol{x}^*)} y_i \right) = \underbrace{\frac{1}{n_T} \cdot \sum_{t=1}^{n_T} \boldsymbol{s}_{T,t}^{*\top}}_{=:\boldsymbol{s}_{RF}^{*\top}} \boldsymbol{y} = \boldsymbol{s}_{RF}^{*\top} \boldsymbol{y}.$$

D NEURAL NETWORKS AS SMOOTHERS IN EVEN MORE DETAIL

In this section, we describe in more detail how neural networks can be expressed as smoothers. First, we briefly review the neural tangent framework, which the smoother formulation builds upon. We then generalize the NTK, and thus the smoother, framework to hold also for the cross-entropy loss.

D.1 Review of the Neural Tangent Kernel

For univariate outputs, updating the parameters of a neural network (or actually any iteratively trained regression model) with loss function $L(\hat{\boldsymbol{f}}, \boldsymbol{y})$, trained with gradient descent, amounts to

$$\hat{\boldsymbol{\theta}}(t+\Delta t) = \hat{\boldsymbol{\theta}}(t) - \Delta t \cdot \frac{\partial L(\hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t)), \boldsymbol{y})}{\partial \hat{\boldsymbol{\theta}}(t)} = \hat{\boldsymbol{\theta}}(t) - \Delta t \cdot \left(\frac{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))}{\partial \hat{\boldsymbol{\theta}}(t)}\right)^{\top} \cdot \frac{\partial L(\hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t)), \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))},$$

where we have used the chain rule, and where Δt denotes the learning rate. Rearranging and letting $\Delta t \to 0$, we obtain the gradient flow update for $\hat{\boldsymbol{\theta}}$,

$$\frac{\partial \hat{\boldsymbol{\theta}}(t)}{\partial t} = -\left(\frac{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))}{\partial \hat{\boldsymbol{\theta}}(t)}\right)^{\top} \cdot \frac{\partial L(\hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t)), \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))},$$

and with another application of the chain rule, the gradient flow update for $\hat{\boldsymbol{f}}^{\star}$ is obtained as

$$\begin{aligned}\frac{\partial \hat{\boldsymbol{f}}^{\star}(\hat{\boldsymbol{\theta}}(t))}{\partial t} &= \frac{\partial \hat{\boldsymbol{f}}^{\star}(\hat{\boldsymbol{\theta}}(t))}{\partial \hat{\boldsymbol{\theta}}(t)} \cdot \frac{\partial \hat{\boldsymbol{\theta}}(t)}{\partial t} = -\frac{\partial \hat{\boldsymbol{f}}^{\star}(\hat{\boldsymbol{\theta}}(t))}{\partial \hat{\boldsymbol{\theta}}(t)} \cdot \left(\frac{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))}{\partial \hat{\boldsymbol{\theta}}(t)}\right)^{\top} \cdot \frac{\partial L(\hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t)), \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))} \\ &=: -\boldsymbol{K}^{\star\mathrm{NTK}}(\hat{\boldsymbol{\theta}}(t)) \cdot \frac{\partial L(\hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t)), \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}(t))},\end{aligned} \tag{7}$$

where

$$\boldsymbol{K}^{\star\mathrm{NTK}}(t) := \frac{\partial \hat{\boldsymbol{f}}^{\star}(t)}{\partial \hat{\boldsymbol{\theta}}} \cdot \left(\frac{\partial \hat{\boldsymbol{f}}(t)}{\partial \hat{\boldsymbol{\theta}}}\right)^{\top}.$$

For the squared loss, $L(\hat{\boldsymbol{f}}(t), \boldsymbol{y}) = \frac{1}{2}\left\|\boldsymbol{y} - \hat{\boldsymbol{f}}(t)\right\|_2^2$, $\frac{\partial L(\hat{\boldsymbol{f}}(t), \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}} = \hat{\boldsymbol{f}}(t) - \boldsymbol{y}$, and

$$\frac{\partial \hat{\boldsymbol{f}}(t)}{\partial t} = \boldsymbol{K}^{\star\mathrm{NTK}}(t) \cdot \left(\boldsymbol{y} - \hat{\boldsymbol{f}}(t)\right). \tag{8}$$

D.1.1 Multivariate Outputs

For multivariate outputs, $\boldsymbol{y}$ and $\hat{\boldsymbol{f}}^{\star}$ are no longer vectors, but matrices, i.e. $\boldsymbol{y} \in \mathbb{R}^{n}$, $\hat{\boldsymbol{f}}^{\star} \in \mathbb{R}^{n+n^*}$, generalize to $\boldsymbol{Y} \in \mathbb{R}^{n \times d_y}$, $\hat{\boldsymbol{F}}^{\star} \in \mathbb{R}^{(n+n^*) \times d_y}$, and Equation 7 generalizes to

$$\begin{aligned}\left(\frac{\partial \hat{F}^{\star}(t)}{\partial t}\right)^{i_N i_y} &= -\left(\frac{\partial \hat{F}^{\star}(t)}{\partial \hat{\theta}}\right)_{i_p}^{i_N i_y} \cdot \left(\frac{\partial \hat{F}(t)}{\partial \hat{\theta}}\right)_{j_n j_y}^{i_p} \cdot \left(\frac{\partial L(t)}{\partial \hat{F}}\right)^{j_n j_y} \\ &=: -K^{\star\mathrm{NTK}}(t)_{j_n j_y}^{i_N i_y} \cdot \left(\frac{\partial L(t)}{\partial \hat{F}}\right)^{j_n j_y},\end{aligned} \tag{9}$$

where $\left(\frac{\partial \hat{F}^{\star}(t)}{\partial t}\right)^{i_N i_y} \in \mathbb{R}^{(n+n^*) \times d_y}$, $\left(\frac{\partial \hat{F}^{\star}(t)}{\partial \hat{\theta}}\right)_{i_p}^{i_N i_y} \in \mathbb{R}^{(n+n^*) \times d_y \times p}$, $\left(\frac{\partial \hat{F}(t)}{\partial \hat{\theta}}\right)_{i_p}^{i_n i_y} \in \mathbb{R}^{n \times d_y \times p}$, $\left(\frac{\partial L(t)}{\partial \hat{F}}\right)^{i_n i_y} \in \mathbb{R}^{n \times d_y}$, and $K^{\star\mathrm{NTK}}(t)_{j_n j_y}^{i_N i_y} \in \mathbb{R}^{(n+n^*) \times d_y \times n \times d_y}$, and upper and lower indices are summed over according to Einstein's notation. However, by vectorizing $\boldsymbol{Y}$ and $\hat{\boldsymbol{F}}^{\star}$ into vectors of length $n \cdot d_y$ and $(n+n^*) \cdot d_y$, i.e. $\boldsymbol{Y} = [\boldsymbol{y_1},\ \boldsymbol{y_2}, \dots \boldsymbol{y_n}]^{\top} \in \mathbb{R}^{n \times d_y} \mapsto \boldsymbol{y} = \left[\boldsymbol{y_1}^{\top},\ \boldsymbol{y_2}^{\top}, \dots \boldsymbol{y_n}^{\top}\right]^{\top} \in \mathbb{R}^{n d_y}$, and analogously for $\hat{\boldsymbol{F}}^{\star}$ and $\hat{\boldsymbol{f}}^{\star}$, Equation 9 can be written on the same form as Equation 7, but for $\boldsymbol{K}^{\star\mathrm{NTK}} \in \mathbb{R}^{(n+n^*) d_y \times n d_y}$, rather than $\boldsymbol{K}^{\star\mathrm{NTK}} \in \mathbb{R}^{(n+n^*) \times n}$. To avoid a four-dimensional $\boldsymbol{K}^{\star\mathrm{NTK}}$, and thus a four-dimensional $\boldsymbol{S}^{\star}$, we use this formulation whenever dealing with multi-dimensional outputs.

D.2 Classification with the Cross-Entropy Loss

For classification with more than two classes, we use compact one-hot encoding, where observation i is represented as a vector $\boldsymbol{y_i} \in \mathbb{R}^{c-1}$, where c is the number of classes. Note that it is enough with $c-1$ elements to represent c classes: the first $c-1$ classes are encoded in the standard one-hot way (with one element in $\boldsymbol{y_i}$ being 1 and the remaining 0) while the last class is encoded as $\boldsymbol{y_i} = \mathbf{0}$, i.e., it is none of the first $c-1$ classes. For $c = 2$, this becomes the standard encoding for binary labels, with $y_i \in \{0, 1\}$.

For the corresponding predictions, $\hat{\boldsymbol{f}}_{\boldsymbol{i}}^{\star} \in \mathbb{R}^{c-1}$, we have $(\hat{\boldsymbol{f}}_{\boldsymbol{i}}^{\star})_j =: \hat{f}_{ij}^{\star} \in [0, 1]$, where the probability of the last class is obtained as $\hat{f}_{ic}^{\star} = 1 - \sum_{j=1}^{c-1} \hat{f}_{ij}^{\star}$. In this formulation, the cross-entropy loss is given by

$$
\begin{aligned}
L(\hat{\boldsymbol{f}}, \boldsymbol{y}) &= \sum_{i=1}^{n} \tilde{L}(\hat{\boldsymbol{f}}_{\boldsymbol{i}}, \boldsymbol{y_i}) := \\
&- \sum_{i=1}^{n} \left(\sum_{j=1}^{c-1} y_{ij} \cdot \log(\hat{f}_{ij}) + \left(1 - \sum_{j=1}^{c-1} y_{ij}\right) \cdot \log\left(1 - \sum_{j=1}^{c-1} \hat{f}_{ij}\right) \right),
\end{aligned}
\tag{10}
$$

where $y_{ij} := (\boldsymbol{y_i})_j$ and

$$
\tilde{L}(\hat{\boldsymbol{f}}_{\boldsymbol{i}}, \boldsymbol{y_i}) = - \left(\sum_{j=1}^{c-1} y_{ij} \cdot \log(\hat{f}_{ij}) + \left(1 - \sum_{j=1}^{c-1} y_{ij}\right) \cdot \log\left(1 - \sum_{j=1}^{c-1} \hat{f}_{ij}\right) \right).
$$

Remark 8. *With $\hat{f}_{ij} = \left(e^{g(\boldsymbol{x_{ij}})}\right) / \left(1 + \sum_{k \neq j} e^{g(\boldsymbol{x_{ik}})}\right)$, for some linear or nonlinear function $g(\cdot)$, for $c = 2$, this formulation becomes exactly logistic regression.*

In Proposition 2, we show that there are matrices $\tilde{\boldsymbol{F}}_{\boldsymbol{i}}$ such that $\frac{\partial \tilde{L}(\hat{\boldsymbol{f}}_{\boldsymbol{i}}, \boldsymbol{y_i})}{\partial \hat{\boldsymbol{f}}_{\boldsymbol{i}}}$ can be written on the form $\tilde{\boldsymbol{F}}_{\boldsymbol{i}} \cdot (\hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y_i})$.

Proposition 2.
With $\tilde{L}(\hat{\boldsymbol{f}}_{\boldsymbol{i}}, \boldsymbol{y_i})$ defined according to Equation 10,

$$
\frac{\partial \tilde{L}(\hat{\boldsymbol{f}}_{\boldsymbol{i}}, \boldsymbol{y_i})}{\partial \hat{\boldsymbol{f}}_{\boldsymbol{i}}} = \left((\hat{\boldsymbol{F}}_{\boldsymbol{i}})^{-1} + \frac{1}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1}\mathbf{1}^{\top} \right) \cdot \left(\hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y_i} \right) =: \hat{\boldsymbol{F}}_{\boldsymbol{i}} \cdot (\boldsymbol{y_i} - \hat{\boldsymbol{f}}_{\boldsymbol{i}}),
$$

where $(\hat{\boldsymbol{F}}_{\boldsymbol{i}})^{-1} \in \mathbb{R}^{(c-1)\times(c-1)}$ denotes the diagonal matrix with $\left((\hat{\boldsymbol{F}}_{\boldsymbol{i}})^{-1}\right)_{jj} = \frac{1}{\hat{f}_{ij}}$, and $\mathbf{1} \in \mathbb{R}^{c-1}$ denotes a vector of only ones.

Using the vectorized forms, $\boldsymbol{y} := \mathrm{vec}(\boldsymbol{Y}) \in \mathbb{R}^{n \cdot (c-1)}$, where $\boldsymbol{Y} \in \mathbb{R}^{n \times (c-1)}$, and equivalently for $\hat{\boldsymbol{f}}$, as described in Section D.1.1, we thus obtain

$$
\frac{\partial L(\hat{\boldsymbol{f}}, \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}} = \tilde{\boldsymbol{F}} \cdot (\boldsymbol{y} - \hat{\boldsymbol{f}}),
\tag{11}
$$

where $\tilde{\boldsymbol{F}} = \mathrm{diag}(\tilde{\boldsymbol{F}}_{\boldsymbol{1}}, \tilde{\boldsymbol{F}}_{\boldsymbol{2}}, \ldots \tilde{\boldsymbol{F}}_{\boldsymbol{n}})$, $\mathbb{R}^{n(c-1)\times n(c-1)}$ is a block diagonal matrix, with the $\tilde{\boldsymbol{F}}_{\boldsymbol{i}}$s along the diagonal. Thus, by replacing $\boldsymbol{K}^{\star \mathrm{NTK}}(t)$ with $\boldsymbol{K}^{\star \mathrm{NTK}}(t) \cdot \tilde{\boldsymbol{F}}(t)$ in Equation 8, the smoother framework trivially extends to the cross-entropy loss.

E GENERALIZATIONS OF THEOREM 2

In this section, we generalize Theorem 2 to include gradient flow and the spectral norm.

E.1 Theorem 2 for Gradient Flow

Gradient flow, i.e., gradient descent with an infinitesimal learning rate, has the closed-form solution

$$
\hat{\boldsymbol{\beta}}_{\mathrm{GF}} = (\boldsymbol{I}_d - \exp(-t\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}))(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{y},
$$

where t denotes the training time and exp the matrix exponential. Replacing the matrix exponential with its first-order Taylor expansion, i.e.,

$$\exp(-t\boldsymbol{A}) = \exp(t\boldsymbol{A})^{-1} \approx (\boldsymbol{I} + t\boldsymbol{A})^{-1},$$

we obtain

$$\hat{\boldsymbol{\beta}}_{\mathrm{GF}} \approx (\boldsymbol{I}_d - (\boldsymbol{I}_p + t\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1})(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y},$$

which, for $t = 1/\lambda$, coincides with the ridge regression solution,

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\boldsymbol{I}_d)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} = (\boldsymbol{I}_p - (\boldsymbol{I}_p + 1/\lambda\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1})(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}.$$

The equivalence of Theorem 2 for gradient flow is given by Theorem 4.

Theorem 4.
Let $\boldsymbol{S}_\lambda := \boldsymbol{\Phi}\boldsymbol{\Phi}_t^+$ and $\boldsymbol{s}_\lambda^* := \boldsymbol{\varphi}^{*\top}\boldsymbol{\Phi}_t^+$, where

$$\boldsymbol{\Phi}_t^+ := (\boldsymbol{I}_p - \exp(-t\boldsymbol{\Phi}^\top\boldsymbol{\Phi}))(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}(\boldsymbol{I}_n - \exp(-t\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)),$$

and let $\|\cdot\|$ denote the trace seminorm, the nuclear norm, or the Frobenius norm. Then

(a) for GCV, $\arg\min_{t\geq 0} \frac{\|(\boldsymbol{I}_n - \boldsymbol{S}_t)^\top(\boldsymbol{I}_n - \boldsymbol{S}_t)\|}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_t))^2} = 0.$

(b) for in-sample MSV,
$\arg\min_{t\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\| =$
$\arg\min_{t\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_t\right\| = \infty.$

(c) for out-of-sample MSV, if $\mathbb{E}(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*\top}) = \boldsymbol{I}_p$, then
$\arg\min_{t\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{s}_t^*\boldsymbol{s}_t^{*\top})\right\| > 0.$
If, in addition, $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$, then
$\arg\min_{t\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{s}_t^*\boldsymbol{s}_t^{*\top})\right\| < \infty.$

E.2 Theorems 2 and 4 for the Spectral Norm

Theorem 5 is the equivalent of Theorems 2 and 4 for the spectral norm. The results are the same as for the other three (semi)norms, with two exceptions: For in-sample MSV, and $n > p$, any value of λ or t is optimal, which is equally useless as always selecting $\lambda = 0$ or $t = \infty$. For out-of-sample MSV, a very reasonable additional assumption is needed when $n > p$.

Theorem 5.
Let

$$\boldsymbol{\Phi}_\lambda^+ := (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \lambda\boldsymbol{I}_n)^{-1}$$
$$\boldsymbol{\Phi}_t^+ := (\boldsymbol{I}_p - \exp(-t\boldsymbol{\Phi}^\top\boldsymbol{\Phi}))(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}(\boldsymbol{I}_n - \exp(-t\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)),$$

so that the in- and out-of-sample smoothers are $\boldsymbol{S}_\lambda := \boldsymbol{\Phi}\boldsymbol{\Phi}_\lambda^+$ and $\boldsymbol{s}_\lambda^* := \boldsymbol{\varphi}^{*\top}\boldsymbol{\Phi}_\lambda^+$ for ridge regression in feature space, and $\boldsymbol{S}_t := \boldsymbol{\Phi}\boldsymbol{\Phi}_t^+$ and $\boldsymbol{s}_t^* := \boldsymbol{\varphi}^{*\top}\boldsymbol{\Phi}_t^+$ for gradient flow in feature space. Then

(a) For GCV, $\arg\min_{\lambda\geq 0} \frac{\|(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\|_2}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_\lambda))^2} = \infty$ *and* $\arg\min_{t\geq 0} \frac{\|(\boldsymbol{I}_n - \boldsymbol{S}_t)^\top(\boldsymbol{I}_n - \boldsymbol{S}_t)\|_2}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_t))^2} = 0.$

(b) For in-sample MSV,
if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is non-singular, then
$\arg\min_{\lambda\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda^\top\boldsymbol{S}_\lambda\right\|_2 = \arg\min_{\lambda\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda\right\|_2 = 0$ *and*
$\arg\min_{t\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\|_2 = \arg\min_{t\geq 0} \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_t\right\|_2 = \infty.$
if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is singular, then $\left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda^\top\boldsymbol{S}_\lambda\right\|_2 = \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda\right\|_2 = \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\|_2 = \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_t\right\|_2 = 1$ *for all λ and t.*

(c) *For out-of-sample MSV, if* $\mathbb{E}(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*\top}) = \boldsymbol{I}_p$,
if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ *is non-singular, then* $\arg\min_{\lambda\geq 0}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 < \infty$ *and* $\arg\min_{t\geq 0}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_t^*\boldsymbol{s}_t^{*\top})\right\|_2 > 0$.
if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ *is singular, then* $\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 = \frac{1}{n}$ *for* $\lambda \geq \max_{i=1,\dots n}\left(\sqrt{\frac{n}{2}}s_i - s_i^2\right)$ *and* $\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_t^*\boldsymbol{s}_t^{*\top})\right\|_2 = \frac{1}{n}$ *for* $t \leq \min_{i=1,\dots n}\left(\log\left(\frac{\sqrt{n}}{\sqrt{n}-\sqrt{2}s_i}\right)\cdot\frac{1}{s_i^2}\right)$, *where* $\{s_i\}_{i=1}^n$ *denote the singular values of* $\boldsymbol{\Phi}$.
If, in addition, $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$, *and* $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ *is non-singular, or* $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ *is singular, and at least one of its singular values is in* $(0, n/2)$, *then* $\arg\min_{\lambda\geq 0}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 > 0$ *and* $\arg\min_{t\geq 0}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_t^*\boldsymbol{s}_t^{*\top})\right\|_2 < \infty$.

F PROOFS

Proof of Equation 2.

$$\left|\mathbb{E}_y\left(\boldsymbol{y}_R^\top\boldsymbol{A}\boldsymbol{y}_R\right)\right| \overset{(a)}{=} \left|\mathrm{Tr}\left(\mathbb{E}_y\left(\boldsymbol{y}_R^\top\boldsymbol{A}\boldsymbol{y}_R\right)\right)\right| \overset{(b)}{=} \left|\mathrm{Tr}\left(\mathbb{E}_y\left(\boldsymbol{y}_R\boldsymbol{y}_R^\top\right)\cdot\boldsymbol{A}\right)\right| \overset{(c)}{=} \sigma_y^2\cdot\left|\mathrm{Tr}\left(\boldsymbol{A}\right)\right|,$$

where we have used *(a)* that the trace of a scalar is the scalar itself, *(b)* the cyclic property of the trace, and *(c)* that $\mathbb{E}(\boldsymbol{y}_R\boldsymbol{y}_R^\top) = \sigma_y^2\boldsymbol{I}_n$. □

Proof of Proposition 1.
For a linear smoother, $\hat{f}^* = \boldsymbol{s}^{*\top}\boldsymbol{y}$, where $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, standard bias-variance calculations yield

$$\begin{aligned}
V_{\boldsymbol{X}} &= \mathbb{E}_{x^*}\left(\mathbb{E}_\varepsilon((\hat{f}^* - \mathbb{E}_\varepsilon(\hat{f}^*))^2) = \mathbb{E}_\varepsilon(\hat{f}^{*2}) - \mathbb{E}_\varepsilon(\hat{f}^*)^2\right)\\
&= \mathbb{E}_{x^*}\left(\mathbb{E}_\varepsilon\left(\boldsymbol{s}^{*\top}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{s}^*\right) - \mathbb{E}_\varepsilon\left(\boldsymbol{s}^{*\top}\boldsymbol{y}\right)\right)^2\right) = \mathbb{E}_{x^*}\left(\mathbb{E}_\varepsilon\left(\boldsymbol{s}^{*\top}(\boldsymbol{f}+\boldsymbol{\varepsilon})(\boldsymbol{f}+\boldsymbol{\varepsilon})^\top\boldsymbol{s}^*\right) - \mathbb{E}_\varepsilon\left(\boldsymbol{s}^{*\top}(\boldsymbol{f}+\boldsymbol{\varepsilon})\right)^2\right)\\
&= \mathbb{E}_{x^*}\left(\boldsymbol{s}^{*\top}\boldsymbol{f}\boldsymbol{f}^\top\boldsymbol{s}^* + \boldsymbol{s}^{*\top}\boldsymbol{f}\mathbb{E}_\varepsilon(\boldsymbol{\varepsilon}^\top)\boldsymbol{s}^* + \boldsymbol{s}^{*\top}\mathbb{E}_\varepsilon(\boldsymbol{\varepsilon})\boldsymbol{f}^\top\boldsymbol{s}^* + \boldsymbol{s}^{*\top}\mathbb{E}_\varepsilon(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\boldsymbol{s}^* - \left(\boldsymbol{s}^{*\top}\boldsymbol{f} + \boldsymbol{s}^{*\top}\mathbb{E}_\varepsilon(\boldsymbol{\varepsilon})\right)^2\right)\\
&= \mathbb{E}_{x^*}\left(\boldsymbol{s}^{*\top}\mathbb{E}_\varepsilon(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\boldsymbol{s}^*\right) = \mathbb{E}_{x^*}\left(\boldsymbol{s}^{*\top}\boldsymbol{s}^*\cdot\sigma_\varepsilon^2\right) = \mathbb{E}_{x^*}\left(\|\boldsymbol{s}^*\|_2^2\right)\cdot\sigma_\varepsilon^2.
\end{aligned}$$

Thus,

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\hat{\boldsymbol{\theta}}}\left|1 - \mathbb{E}(\|\boldsymbol{s}^*(\hat{\boldsymbol{\theta}})\|_2^2)\right| = \arg\min_{\hat{\boldsymbol{\theta}}}\left|1 - V_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}})/\sigma_\varepsilon^2\right| = \arg\min_{\hat{\boldsymbol{\theta}}}\left|\sigma_\varepsilon^2 - V_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}})\right|.$$

□

Proof of Theorem 1.
To alleviate notation, we denote the SNR as s during this proof. According to Corollary 5 by Hastie et al. (2022), almost surely

$$\begin{aligned}
V_{\boldsymbol{X}}(\lambda,\gamma) &\to \sigma_\varepsilon^2\gamma\left(m_F(-\lambda,\gamma) - \lambda m_F'(-\lambda,\gamma)\right) =: \overline{V_{\boldsymbol{X}}}(\lambda)\\
R_{\boldsymbol{X}}(\lambda,\gamma) &\to \sigma_\varepsilon^2\gamma\left(m_F(-\lambda,\gamma) - \lambda(1 - s\cdot\lambda/\gamma)m_F'(-\lambda,\gamma)\right) =: \overline{R_{\boldsymbol{X}}}(\lambda),
\end{aligned} \tag{12}$$

where $m_F(z,\gamma)$ and $m_F'(z,\gamma) := \partial_z m(z,\gamma)$ are the Stieltjes transform of the Marchenko-Pastur law of $\hat{\boldsymbol{\Sigma}}$, and its derivative (with respect to z). For isotropic features, when $\boldsymbol{\Sigma} = \boldsymbol{I}$, we obtain the closed forms

$$\begin{aligned}
m_F(z,\gamma) &= \frac{1-\gamma-z-\sqrt{(1-\gamma-z)^2-4\gamma z}}{2\gamma z}\\
m_F'(z,\gamma) &= \frac{\frac{z(1+\gamma-z)}{\sqrt{(1-\gamma-z)^2-4\gamma z}} - 1 + \gamma + \sqrt{(1-\gamma-z)^2-4\gamma z}}{2\gamma z^2}.
\end{aligned}$$

Plugging this into Equation 12, we obtain

$$\overline{V_{\boldsymbol{X}}}(\lambda,\gamma) = \sigma_\varepsilon^2\gamma\left(m_F(-\lambda,\gamma) - \lambda m_F'(-\lambda,\gamma)\right) = \frac{\sigma_\varepsilon^2}{2}\left(\frac{1+\gamma+\lambda}{\sqrt{(1-\gamma+\lambda)^2+4\gamma\lambda}} - 1\right).$$

Solving for λ, we obtain

$$\frac{\sigma_\varepsilon^2}{2}\left(\frac{1+\gamma+\lambda}{\sqrt{(1-\gamma+\lambda)^2+4\gamma\lambda}}-1\right)=\sigma_\varepsilon^2 \iff \lambda = 3\sqrt{\frac{\gamma}{2}}-\gamma-1$$

$$\text{where } \lambda \geq 0 \text{ for } \frac{1}{2}\leq\gamma\leq 2 \implies \overline{\lambda_T}=\begin{cases}3\sqrt{\frac{\gamma}{2}}-\gamma-1, & \gamma\in\left(\frac{1}{2},2\right)\\ 0, & \gamma\in\left(0,\frac{1}{2}\right]\cup[2,\infty)\end{cases}$$

and

$$\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)=\begin{cases}\sigma_\varepsilon^2\cdot\left(1+s\frac{(\sqrt{2\gamma}-1)^2}{\gamma}\right), & \gamma\in\left(\frac{1}{2},2\right)\\ \overline{R_{\boldsymbol{X}}}(0,\gamma), & \gamma\in\left(0,\frac{1}{2}\right]\cup[2,\infty)\,.\end{cases}$$

According to Lemma 1, the maximum of $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*},\gamma)$ occurs for some $\gamma\in\left[\frac{1}{2},2\right]$. We make a grid in $\gamma\in\left[\frac{1}{2},2\right]$ and $s\in[1,80]$ with $\Delta\gamma=\Delta s=10^{-5}$, and numerically calculate the maximum of $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*},\gamma)$ on this grid. The maximum value is less than 2.445.

We then use the bounds on $\left|\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}(s),\gamma)}{\partial\gamma}\right|$ and $\left|\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}(s),\gamma)}{\partial s}\right|$ from Lemma 2 to bound the maximum deviation of the function value between the grid points. The deviation is less than

$$\frac{1}{2}\cdot\sqrt{\left(\max_{\gamma\in\left[\frac{1}{2},2\right]}\left|\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}(s),\gamma)}{\partial\gamma}\right|\right)^2\cdot(\Delta\gamma)^2+\left(\max_{s\in[1,80]}\left|\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}(s),\gamma)}{\partial s}\right|\right)^2\cdot(\Delta s)^2}$$
$$\leq\frac{10^{-5}}{2}\cdot\sqrt{752.2^2+10.5^2}\leq 0.003762.$$

Thus, the maximum is less than $2.445+0.003762\leq 2.449$.

□

Lemma 1.

$$\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*},\gamma)}{\partial\gamma}\begin{cases}>0 \textit{ for } \gamma\in\left(0,\frac{1}{2}\right]\\ <0 \textit{ for } \gamma\in[2,\infty)\,.\end{cases}$$

Lemma 2. *For $\gamma\in\left[\frac{1}{2},2\right]$, and $s\in[1,80]$,*

$$\left|\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}(s),\gamma)}{\partial\gamma}\right|\leq 752.2$$

and

$$\left|\frac{\partial\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}(s),\gamma)}{\partial s}\right|\leq 10.5.$$

Proof of Lemma 1.

According to Hastie et al. (2022),

$$\overline{R_{\boldsymbol{X}}}(0,\gamma)=\begin{cases}\sigma_\varepsilon^2\cdot\frac{\gamma}{1-\gamma}, & \gamma\leq 1\\ \sigma_\varepsilon^2\cdot\left(s\left(1-\frac{1}{\gamma}\right)+\frac{1}{\gamma-1}\right), & \gamma>1\end{cases}$$
$$\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*},\gamma)=\frac{\sigma_\varepsilon^2}{2}\left(s-s/\gamma-1+\sqrt{4s+(1-s+s/\gamma)^2}\right),$$

and straightforward calculations yield

$$\frac{\partial\left(\overline{R_{\boldsymbol{X}}}(0,\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*},\gamma)\right)}{\partial\gamma}=\begin{cases}s\cdot\frac{1+s\cdot(\frac{1}{\gamma}-1)-\sqrt{4s+\left(1+s\cdot(\frac{1}{\gamma}-1)\right)^2}}{2\gamma^2\sqrt{4s+\left(1+s\cdot(\frac{1}{\gamma}-1)\right)^2}}+\frac{1}{(\gamma-1)^2}, & \text{for } \gamma<1\\ s\cdot\frac{1-s\cdot(1-\frac{1}{\gamma})+\sqrt{4s+\left(1-s\cdot(1-\frac{1}{\gamma})\right)^2}}{2\gamma^2\sqrt{4s+\left(1-s\cdot(1-\frac{1}{\gamma})\right)^2}}-\frac{1}{(1-\gamma)^2}, & \text{for } \gamma>1.\end{cases}$$

We first show that, for $\gamma < 0$,

$$s \cdot \frac{1 + s \cdot (\frac{1}{\gamma} - 1) - \sqrt{4s + \left(1 + s \cdot (\frac{1}{\gamma} - 1)\right)^2}}{2\gamma^2 \sqrt{4s + \left(1 + s \cdot (\frac{1}{\gamma} - 1)\right)^2}} + \frac{1}{(\gamma - 1)^2} > 0,$$

using the intermediate variables $a := \frac{1}{\gamma} - 1 > 0$, $b := 1 + sa > 0$ and $c := \sqrt{4s + b^2} > 0$:

$$\begin{aligned}
&s \cdot \frac{1 + s \cdot (\frac{1}{\gamma} - 1) - \sqrt{4s + \left(1 + s \cdot (\frac{1}{\gamma} - 1)\right)^2}}{2\gamma^2 \sqrt{4s + \left(1 + s \cdot (\frac{1}{\gamma} - 1)\right)^2}} + \frac{1}{\gamma^2 a^2} \\
&= s \cdot \frac{b - c}{2\gamma^2 c} + \frac{1}{\gamma^2 a^2} = s \cdot \frac{b - c}{2\gamma^2 c} \cdot \frac{b + c}{b + c} + \frac{1}{\gamma^2 a^2} = s \cdot \frac{b^2 - c^2}{2\gamma^2 c(b + c)} + \frac{1}{\gamma^2 a^2} \\
&= \frac{-2s^2}{\gamma^2 c(b + c)} + \frac{1}{\gamma^2 a^2}.
\end{aligned}$$

Since $c \geq b > 0$, and $b > sa$, we have $c(b + c) \geq b(b + b) = 2b^2 > 2s^2 a^2$, and

$$\frac{1}{\gamma^2 a^2} - \frac{2s^2}{\gamma^2 c(b + c)} > \frac{1}{\gamma^2 a^2} - \frac{2s^2}{\gamma^2 \cdot 2s^2 a^2} = \frac{1}{\gamma^2 a^2} - \frac{1}{\gamma^2 a^2} = 0.$$

We then show that, for $\gamma > 1$,

$$s \cdot \frac{1 - s \cdot (1 - \frac{1}{\gamma}) + \sqrt{4s + \left(1 - s \cdot (1 - \frac{1}{\gamma})\right)^2}}{2\gamma^2 \sqrt{4s + \left(1 - s \cdot (1 - \frac{1}{\gamma})\right)^2}} - \frac{1}{(\gamma - 1)^2} < 0.$$

Using the intermediate variables $a := 1 - \frac{1}{\gamma} \in (0, 1)$, $b := 1 - s \cdot a < 1$ and $c := \sqrt{4s + b^2} > 0$, we obtain

$$\begin{aligned}
&s \cdot \frac{1 - s \cdot (1 - \frac{1}{\gamma}) + \sqrt{4s + \left(1 - s \cdot (1 - \frac{1}{\gamma})\right)^2}}{2\gamma^2 \sqrt{4s + \left(1 - s \cdot (1 - \frac{1}{\gamma})\right)^2}} - \frac{1}{(\gamma - 1)^2} \quad = \frac{s(b + c)}{2\gamma^2 c} - \frac{1}{\gamma^2 a^2} \\
&= \frac{1}{2\gamma^2 c} \left(sb + c \left(s - \frac{2}{a^2} \right) \right).
\end{aligned}$$

We thus need to show that

$$\underbrace{\left(\frac{2}{a^2} - s \right)}_{=:d} c > sb.$$

We split the analysis into the following three cases: $s \in \left(0, \frac{1}{a}\right]$, $s \in \left(\frac{1}{a}, \frac{2}{a^2}\right]$, and $s \in \left(\frac{2}{a^2}, \infty\right)$.

For $s \in \left(0, \frac{1}{a}\right]$, we have $b \geq 0$, $d > 0$, and $c > 0$, and, using $c^2 = 4s + b^2$, we obtain

$$dc > sb \iff d^2 c^2 > s^2 b^2 \iff d^2(4s + b^2) > s^2 b^2 \iff b^2(d^2 - s^2) + 4sd^2 > 0.$$

Using the conjugate rule and the definition of d, we further obtain

$$b^2(d^2 - s^2) + 4sd^2 = b^2(d + s)(d - s) + 4sd^2 = b^2 \cdot \frac{2}{a^2} \cdot \left(\frac{2}{a^2} - 2s \right) + 4sd^2.$$

Since $4sd^2 > 0$, we just need to show that $\left(\frac{2}{a^2} - 2s\right) \geq 0$. Since $s \leq \frac{1}{a}$, and $a \in (0, 1)$,

$$\left(\frac{2}{a^2} - 2s \right) = 2 \cdot \left(\frac{1}{a^2} - s \right) \geq 2 \cdot \left(\frac{1}{a^2} - \frac{1}{a} \right) = 2 \cdot \frac{1 - a}{a^2} > 0.$$

For $s \in \left(\frac{1}{a}, \frac{2}{a^2}\right]$, we have $b < 0$, $d \geq 0$, and $c > 0$. Thus, trivially $cd > sb$.

For $s \in \left(\frac{2}{a^2}, \infty\right)$, we have $b < 0$, $d < 0$, and $c > 0$, and

$$dc > sb \iff d^2c^2 < s^2b^2 \iff s^2b^2 - d^2c^2 > 0.$$

Using the definitions of b, c, and d, we obtain

$$\begin{aligned} f(s) := s^2b^2 - d^2c^2 &= s^2(1-sa)^2 - \left(\frac{2}{a^2} - s\right)^2 \left(4s + (1-sa)^2\right) \\ &= \frac{4}{a^4}\left((3-2a)a^2s^2 + (a^2+2a-4)s - 1\right). \end{aligned}$$

Since $f(x)$ has a positive second derivative, if $f(2/a^2) > 0$ and $f'(2/a^2) > 0$, then $f(s) > 0$ for $s \geq 2/a^2$. Straightforward calculations yield

$$\begin{aligned} f(2/a^2) &= \frac{4}{a^6}(a-2)^2 > 0, \\ f'(2/a^2) &= \frac{4}{a^4}\left(2(3-2a)a^2s + a^2 + 2a - 4\right), \ f'(2/a^2) = \frac{4}{a^4}\left(a^2 - 6a + 8\right) > \frac{4}{a^4}\left(a^2 - 6 + 8\right) > 0. \end{aligned}$$

□

Proof of Lemma 2.
Let

$$\begin{aligned} f_1(\gamma) &:= 1 + s \cdot \frac{(\sqrt{2\gamma} - 1)^2}{\gamma} \\ f_2(\gamma) &:= \frac{1}{2} \cdot \left(s - s/\gamma - 1 + \sqrt{4s + (1 - s + s/\gamma)^2}\right), \end{aligned}$$

so that $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T}, \gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*}, \gamma) = f_1(\gamma)/f_2(\gamma)$.

According to the quotient rule and the triangle inequality,

$$\begin{aligned} \left|\left(\frac{f_1(\gamma)}{f_2(\gamma)}\right)'\right| &= \left|\frac{f_1'(\gamma) \cdot f_2(\gamma) - f_1(\gamma) \cdot f_2'(\gamma)}{f_2^2(\gamma)}\right| \leq \left|\frac{f_1'(\gamma) \cdot f_2(\gamma)}{f_2^2(\gamma)}\right| + \left|\frac{f_1(\gamma) \cdot f_2'(\gamma)}{f_2^2(\gamma)}\right| \\ &= |f_1'(\gamma)| \cdot \frac{1}{f_2(\gamma)} + \frac{|f_2'(\gamma)|}{f_2(\gamma)} \cdot \frac{f_1(\gamma)}{f_2(\gamma)}, \end{aligned}$$

where we have used that $f_1, f_2 \geq 0$. According to straightforward calculations,

$$\begin{aligned} f_1'(\gamma) &= s \cdot \frac{\sqrt{2\gamma} - 1}{\gamma^2} \\ f_2'(\gamma) &= \frac{s}{2\gamma^2} \cdot \left(1 - \frac{1 - s + s/\gamma}{\sqrt{4s + (1 - s + s/\gamma)^2}}\right) = \frac{s}{\gamma^2\sqrt{4s + (1 - s + s/\gamma)^2}} \cdot f_2(\gamma). \end{aligned}$$

Let us first bound $|f_1'(\gamma)|$. For $\gamma \in \left[\frac{1}{2}, 2\right]$, $|f_1'(\gamma)| = f_1'(\gamma)$ and

$$0 = f_1''(\gamma) = s \cdot \frac{4 - 3\sqrt{2\gamma}}{2\gamma^3} \iff \gamma \in \left\{\frac{8}{9}, \infty\right\}$$

Since $f_1'(8/9) = 27/64 = 0.421875 > f_1'(2) = 1/4 > f_1'(1/2) = 0$, we have $|f_1'(\gamma)| = f_1'(\gamma) \leq 27/64$.

Let us now bound $|f_2'(\gamma)|/f_2(\gamma) = s/\left(\gamma^2\sqrt{4s + (1 - s + s/\gamma)^2}\right)$.

$$\gamma^2 \geq 1/4 \text{ and } \sqrt{4s + (1 - s + s/\gamma)^2} \geq \sqrt{4s} = 2\sqrt{s}.$$

Thus

$$\frac{|f_2'(\gamma)|}{f_2(\gamma)} = \frac{s}{\left(\gamma^2\sqrt{4s+(1-s+s/\gamma)^2}\right)} \leq \frac{s}{1/4\cdot 2\sqrt{s}} = 2\sqrt{s}.$$

Thus, so far we have

$$\left|\left(\frac{f_1(\gamma)}{f_2(\gamma)}\right)'\right| \leq \frac{\frac{27}{64}+2\sqrt{s}\cdot f_1(\gamma)}{f_2(\gamma)}.$$

Since $f_1'(\gamma) > 0$ for $s > 0$ and $\gamma > 1/2$, $f_1(\gamma)$ obtains its maximum for $\gamma = 2$, i.e., $f_1(\gamma) \leq 1+s/2$. Since $f_2'(\gamma) > 0$ for $s > 0$ and $\gamma > 0$, $f_2(\gamma)$ obtains its minimum for $\gamma = 1/2$, i.e.,

$$\begin{aligned} f_2(\gamma) &\geq \frac{\sqrt{s^2+6s+1}-s-1}{2} = \frac{\left(\sqrt{s^2+6s+1}-s-1\right)\cdot\left(\sqrt{s^2+6s+1}+s+1\right)}{2\cdot\left(\sqrt{s^2+6s+1}+s+1\right)} = \frac{s^2+6s+1-(s+1)^2}{2\cdot\left(\sqrt{s^2+6s+1}+s+1\right)} \\ &\geq \frac{s^2+6s+1-(s+1)^2}{2\cdot\left(\sqrt{s^2+6s+9}+s+1\right)} = \frac{4s}{2\cdot\left(\sqrt{(s+3)^2}+s+1\right)} = \frac{s}{s+2}. \end{aligned}$$

Putting it together, we obtain

$$\left|\left(\frac{f_1(\gamma)}{f_2(\gamma)}\right)'\right| \leq \left(\frac{27}{64}+2\sqrt{s}\cdot(1+s/2)\right)\cdot\frac{s+2}{s} = \left(\frac{27}{64}+\sqrt{s}\cdot(s+2)\right)\cdot\frac{s+2}{s}.$$

Since $\left(\frac{27}{64}+\sqrt{s}\cdot(s+2)\right)\cdot\frac{s+2}{s} > 0$ for $s > 0.8757$, it obtains it maximum at $s = 80$, where it is 752.198.

Now, we instead let

$$\begin{aligned} f_1(s) &:= 1 + s\cdot\frac{(\sqrt{2\gamma}-1)^2}{\gamma} \\ f_2(s) &:= \frac{1}{2}\cdot\left(s - s/\gamma - 1 + \sqrt{4s+(1-s+s/\gamma)^2}\right), \end{aligned}$$

so that $\overline{R_{\boldsymbol{X}}}(\overline{\lambda_T},\gamma)/\overline{R_{\boldsymbol{X}}}(\overline{\lambda^*},\gamma) = f_1(s)/f_2(s)$.

According to the quotient rule and the triangle inequality,

$$\begin{aligned} \left|\left(\frac{f_1(s)}{f_2(s)}\right)'\right| &= \left|\frac{f_1'(s)\cdot f_2(s) - f_1(s)\cdot f_2'(s)}{f_2^2(s)}\right| \leq \left|\frac{f_1'(s)\cdot f_2(s)}{f_2^2(s)}\right| + \left|\frac{f_1(s)\cdot f_2'(s)}{f_2^2(s)}\right| \\ &= |f_1'(s)|\cdot\frac{1}{f_2(s)} + |f_2'(s)|\cdot\frac{f_1(s)}{f_2^2(s)}, \end{aligned}$$

where we have again used that $f_1, f_2 \geq 0$. According to straightforward calculations,

$$\begin{aligned} f_1'(s) &= \frac{(\sqrt{2\gamma}-1)^2}{\gamma} \\ f_2'(s) &= \frac{1-(1/\gamma-1)\cdot f_2(s)}{\sqrt{4s+(1-s+s/s)^2}}. \end{aligned}$$

Let us first bound $|f_1'(s)|$. For $\gamma \in \left[\frac{1}{2}, 2\right]$, $|f_1'(s)| = f_1'(s) \geq 0$, so the maximum is obtained for $\gamma = 2$. Thus

$$|f_1'(s)| \leq \frac{(\sqrt{2\cdot 2}-1)^2}{2} = \frac{1}{2}.$$

Let us now bound $|f_2'(\gamma)| = (1-(1/\gamma-1)\cdot f_2(s))/\sqrt{4s+(1-s+s/\gamma)^2}$. For $\gamma \in \left[\frac{1}{2}, 2\right]$,

$$1/\gamma - 1 \in \left[-\frac{1}{2}, 1\right] \text{ and } \sqrt{4s+(1-s+s/\gamma)^2} \geq \sqrt{4s} = 2\sqrt{s}.$$

Thus

$$|f_2'(s)| = \left| \frac{1 - (1/\gamma - 1) \cdot f_2(s)}{\sqrt{4s + (1 - s + s/s)^2}} \right| \leq \frac{1 + f_2(s)}{2\sqrt{s}}.$$

From above, we know that $f_1(s) \leq 1 + s/2$, and $f_2(s) \geq s/(s+2)$.

Putting it together, we obtain

$$\begin{aligned} \left| \left(\frac{f_1(s)}{f_2(s)} \right)' \right| &\leq |f_1'(s)| \cdot \frac{1}{f_2(s)} + |f_2'(s)| \cdot \frac{f_1(s)}{f_2^2(s)} \leq \frac{1}{2} \cdot \frac{1}{f_2(s)} + \frac{1 + f_2(s)}{2\sqrt{s}} \cdot \frac{1 + s/2}{f_2^2(s)} \\ &= \frac{1}{2} \cdot \frac{s+2}{s} + \frac{1 + \frac{s+2}{s}}{4\sqrt{s}} \cdot (2+s) \cdot \frac{(s+2)^2}{s^2} = \frac{s+2}{2s} \cdot \left(1 + \frac{(s+1)(s+2)}{s\sqrt{s}} \right), \end{aligned}$$

which, for $s \in [1, 80]$, obtains it maximum for $s = 1$, where its value is 10.5.

□

Proof of Theorems 2, 4, and 5.
Let $\{s_i^2\}_{i=1}^n$ denote the singular values of $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$, where, if $n > p$, $s_i = 0$ for $i > p$, and where $\{s_i\}_{i=1}^{\min(n,p)}$ are the singular values of $\boldsymbol{\Phi}$.

We first note that when a matrix $\boldsymbol{A}$ is positive semi-definite, its eigenvalues and singular values coincide, and thus $\|\boldsymbol{A}\|_T = |\mathrm{Tr}(\boldsymbol{A})| = \|\boldsymbol{A}\|_*$. This is the case for parts (a) and (b).

We will first prove all ridge regression expressions, i.e. expressions depending on λ. The proofs for gradient flow, which depend on t, are very similar, and done afterwards.

Ridge Regression (λ)
Part (a)
For part *(a)*, let $a_i := 1 - \frac{s_i^2}{s_i^2 + \lambda} \geq 0$. Then, the singular values of $(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top (\boldsymbol{I}_n - \boldsymbol{S}_\lambda)$ and $\boldsymbol{I}_n - \boldsymbol{S}_\lambda$ are $\left\{ \left(1 - \frac{s_i^2}{s_i^2+\lambda} \right)^2 \right\}_{i=1}^n = \{a_i^2\}_{i=1}^n$, and $\left\{ 1 - \frac{s_i^2}{s_i^2+\lambda} \right\}_{i=1}^n = \{a_i\}_{i=1}^n$ respectively. Since the $a_i \geq 0$ for all i, $\boldsymbol{I}_n - \boldsymbol{S}_\lambda$ is positive semi-definite, which means that its singular and eigenvalues coincide and thus that its trace equals the sum of its singular values. Consequently

$$\frac{\left\| (\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top (\boldsymbol{I}_n - \boldsymbol{S}_\lambda) \right\|}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_\lambda))^2} = \begin{cases} \frac{\sum_{i=1}^n a_i^2}{\left(\sum_{i=1}^n a_i\right)^2} & \text{for the nuclear (and trace semi) norm.} \\ \sqrt{\frac{\sum_{i=1}^n a_i^4}{\left(\sum_{i=1}^n a_i\right)^4}} & \text{for the Frobenius norm.} \\ \frac{\max_{i=1,\dots n} a_i^2}{\left(\sum_{i=1}^n a_i\right)^2} & \text{for the spectral norm.} \end{cases} \tag{13}$$

For the nuclear and Frobenius norms, we can use that for any vector, $\boldsymbol{v} \in \mathbb{R}^n$, for $d > 1$,

$$\|\boldsymbol{v}\|_d = \left(\sum_{i=1}^n |v_i|^d \right)^{1/d} \geq n^{1-1/d} \cdot \|\boldsymbol{v}\|_1 = n^{1-1/d} \cdot \sum_{i=1}^n |v_i|,$$

with equality iff $v_i = 1$ for all i. Thus the two corresponding expressions in Equation 13 are minimized when $a_i = 1$ for all i, i.e. for $\lambda = \infty$, in which case their values are

$$\lim_{\lambda \to \infty} \left(\frac{\left\| (\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top (\boldsymbol{I}_n - \boldsymbol{S}_\lambda) \right\|_*}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_\lambda))^2} \right) = \frac{1}{n}$$

$$\lim_{\lambda \to \infty} \left(\frac{\left\| (\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top (\boldsymbol{I}_n - \boldsymbol{S}_\lambda) \right\|_F}{(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_\lambda))^2} \right) = \frac{1}{n\sqrt{n}}.$$

For the spectral norm, let $a_m := \max_{i=1,\dots n} a_i$. Then the corresponding expression in Equation 13 becomes $a_m^2 / \left(\sum_{i=1}^n a_i\right)^2 = 1/\left(\sum_{i=1}^n \frac{a_i}{a_m}\right)^2$. Here, $\frac{a_i}{a_m} \leq 1$ is maximized when $a_i = a_m$ for all i, i.e. when $\lambda = \infty$ and $a_i = 1$ for all i. In this case,

$$\lim_{\lambda\to\infty}\left(\frac{\left\|(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^\top(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\right\|_2}{\left(\mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\right)^2}\right) = \frac{1}{n^2}.$$

Part (b)

For part *(b)*,

$$\begin{aligned}
&n \cdot \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda\right\| \\
=&n \cdot \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Phi}^\top\right\| \\
=&\begin{cases}
\sum_{i=1}^n \left|1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right| & \text{for the nuclear (and trace semi) norm.} \\
\sum_{i=1}^n \left(1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right)^2 & \text{for the squared Frobenius norm.} \\
\max_{i=1,\dots n}\left|1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right| & \text{for the spectral norm.}
\end{cases}
\end{aligned}$$

If $s_i > 0$, $\frac{s_i^4}{(s_i^2+\lambda)^2} \in [0,1]$ for $\lambda \geq 0$, and equals 1 only for $\lambda = 0$. Hence, $\left|1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right|$ (and thus $\left(1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right)^2$) is minimized for $\lambda = 0$, in which case it is 0. If $s_i = 0$, $\left|1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right| = 1$ regardless of λ. For the nuclear and Frobenius norms, since each term of the sum obtains its minimum for $\lambda = 0$, so does the sum. For the spectral norm, the minimum is obtained either for $\lambda = 0$, if $\arg\max_{s_i}\left|1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right| > 0$, or for any value of λ if $\arg\max_{s_i}\left|1 - \frac{s_i^4}{(s_i^2+\lambda)^2}\right| = 0$.

Analogously,

$$\begin{aligned}
n \cdot \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{S}_\lambda\right\| &= n \cdot \left\|\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Phi}^\top\right\| \\
&= \begin{cases}
\sum_{i=1}^n \left|1 - \frac{s_i^2}{s_i^2+\lambda}\right| & \text{for the nuclear (and trace semi) norm.} \\
\sum_{i=1}^n \left(1 - \frac{s_i^2}{s_i^2+\lambda}\right)^2 & \text{for the squared Frobenius norm.} \\
\max_{i=1,\dots n}\left|1 - \frac{s_i^2}{s_i^2+\lambda}\right| & \text{for the spectral norm.}
\end{cases}
\end{aligned}$$

where, for $\lambda \geq 0$, $\frac{s_i^2}{s_i^2+\lambda} \in [0,1]$ equals 1 only for $\lambda = 0$, if $s_i > 0$, and $\frac{s_i^2}{s_i^2+\lambda} = 0$ regardless of λ if $s_i = 0$. Again, for the nuclear and Frobenius norms, each term of the sums obtains its minimum for $\lambda = 0$, and so do the sums. For the spectral norm, the minimum is obtained either for $\lambda = 0$, if $\arg\max_{s_i}\left|1 - \frac{s_i^2}{s_i^2+\lambda}\right| > 0$, or for any value of λ if $\arg\max_{s_i}\left|1 - \frac{s_i^2}{s_i^2+\lambda}\right| = 0$.

Part (c)

For part *(c)*, we want to show that the derivative of $\left\|\frac{1}{n}\boldsymbol{I} - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|$ with respect to λ is strictly negative for $\lambda = 0$, strictly positive for $M_1 < \lambda < M_2$ and non-negative for $\lambda \geq M_2$, in which case the minimum of the norm is obtained for $\lambda \in (0, M_1)$.

Since

$$\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top}) = \frac{1}{n}\boldsymbol{I}_n - (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{\Phi}\underbrace{\mathbb{E}(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*\top})}_{=\boldsymbol{I}_p}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \lambda\boldsymbol{I}_n)^{-1},$$

$$
\begin{aligned}
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_T &= \left|\sum_{i=1}^n \left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right)\right| \\
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_* &= \sum_{i=1}^n \left|\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right| \\
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_F &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right)^2 \\
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 &= \max_{i=1,\dots n} \left|\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right| = \left|\frac{1}{n} - \frac{s_m^2}{(s_m^2+\lambda)^2}\right|,
\end{aligned}
\tag{14}
$$

where $s_m := \arg\max_{s_i} \left|\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right|$.

The needed derivatives with respect to λ are

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) &= \frac{2s_i^2}{(s_i^2+\lambda)^3} \\
\frac{\partial}{\partial\lambda}\left|\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right| &= \frac{2s_i^2}{(s_i^2+\lambda)^3}\cdot \operatorname{sign}\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) \\
\frac{\partial}{\partial\lambda}\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right)^2 &= \frac{4s_i^2}{(s_i^2+\lambda)^3}\cdot\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right),
\end{aligned}
$$

and hence

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_T &= \sum_{i=1}^n \frac{2s_i^2}{(s_i^2+\lambda)^3}\cdot \operatorname{sign}\left(\sum_{i=1}^n\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right)\right) \\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_* &= \sum_{i=1}^n \frac{2s_i^2}{(s_i^2+\lambda)^3}\cdot \operatorname{sign}\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) \\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_F &= \sum_{i=1}^n \frac{4s_i^2}{(s_i^2+\lambda)^3}\cdot\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) \\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 &= \frac{2s_m^2}{(s_m^2+\lambda)^3}\cdot \operatorname{sign}\left(\frac{1}{n} - \frac{s_m^2}{(s_m^2+\lambda)^2}\right).
\end{aligned}
\tag{15}
$$

We first consider large λ's. For the trace seminorm and the nuclear and Frobenius norm, we note that $\lambda > \sqrt{n}s_i - s_i^2 \implies \left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) > 0 \implies \sum_{i=1}^n \left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) > 0$, and hence, if $\max_{i=1,\dots n}\left(\sqrt{n}s_i - s_i^2\right) =: M_1 < \lambda < \infty$, each term in the three sums in Equation 15 sum is strictly positive for $s_i > 0$, and zero for $s_i = 0$. Consequently, for $M_1 < \lambda < \infty$,

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_T &> 0 \\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_* &> 0 \\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_F &> 0.
\end{aligned}
$$

For the spectral norm, to obtain $\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 > 0$ when $\sqrt{n}s_m - s_m^2 =: M_m < \lambda < \infty$, we also need $s_m > 0$. If $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is non-singular, all its singular values are non-zero, and this is trivially fulfilled. However, if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is singular, at least one s_i is zero. We call this singular value s_0 (i.e. $s_0 = 0$) and note that $\left|\frac{1}{n} - \frac{s_0^2}{(s_0^2+\lambda)^2}\right| = \frac{1}{n}$ regardless of λ. Furthermore, if $\lambda \geq \sqrt{\frac{n}{2}}s_i - s_i^2$ for all s_i, then $\left|\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right| \leq \frac{1}{n}$. This means that for for $\lambda \geq \max_{i=1,\dots n}\sqrt{\frac{n}{2}}s_i - s_i^2$, $s_m = 0$ and $\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2 = \frac{1}{n}$.

We now consider $\lambda = 0$. For the trace seminorm and for the nuclear and Frobenius norms, if $s_i = 0$, the corresponding terms in the sums in Equation 15 are zero, and thus do not contribute when calculating the derivative. For terms where $s_i > 0$, $\lim_{\lambda\to 0}\left(\frac{1}{n} - \frac{s_i^2}{(s_i^2+\lambda)^2}\right) = \left(\frac{1}{n} - \frac{1}{s_i^2}\right)$. Since $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$ by assumption, $s_i^2 \in (0, n)$ for all i, and we can write $s_i^2 = n - a_i$ for $0 < a_i < n$. Hence,

$$\left(\frac{1}{n} - \frac{1}{s_i^2}\right) = \left(\frac{1}{n} - \frac{1}{n - a_i}\right) = \left(-\frac{a_i}{n(n - a_i)}\right) < 0.$$

Thus, since all terms in the three sums in Equation 15 are either 0, if $s_i = 0$, or negative, if $s_i > 0$, the sums, and hence the derivatives, are negative: For $\lambda = 0$,

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_T &< 0\\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_* &< 0\\
\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_F &< 0.
\end{aligned}$$

For the spectral norm, we again need $s_m > 0$. If $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is non-singular, this is trivial, while if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is singular, there exists, by assumption, at least one $s_i^2 \in (0, n/2)$, for which $\frac{1}{s_i^2} - \frac{1}{n} > \frac{2}{n} - \frac{1}{n} = \frac{1}{n}$. Thus, since the maximum in Equation 14 is obtained for an i such that $s_i > 0$, we have $s_m > 0$, and

$$\frac{\partial}{\partial\lambda}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\lambda^*\boldsymbol{s}_\lambda^{*\top})\right\|_2\Bigg|_{\lambda=0} = \frac{2}{s_m^4}\cdot\operatorname{sign}\left(\frac{1}{n} - \frac{1}{s_m^2}\right).$$

Again, since $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$, $s_m^2 \in (0, n)$ and we can write $s_m^2 = n - a_m$ for $0 < a_m < n$, and

$$\left(\frac{1}{n} - \frac{1}{s_m^2}\right) = \left(\frac{1}{n} - \frac{1}{n - a_m}\right) = \left(-\frac{a_m}{n(n - a_m)}\right) < 0.$$

Gradient Flow (t)

Part (a)

For part *(a)*, let $a_i := e^{-ts_i^2} \geq 0$. Then, the singular values of $(\boldsymbol{I}_n - \boldsymbol{S}_t)^\top(\boldsymbol{I}_n - \boldsymbol{S}_t)$ and $\boldsymbol{I}_n - \boldsymbol{S}_t$ are $\left\{e^{-ts_i^4}\right\}_{i=1}^n = \{a_i^2\}_{i=1}^n$, and $\left\{e^{-ts_i^2}\right\}_{i=1}^n = \{a_i\}_{i=1}^n$ respectively. Since the $a_i \geq 0$ for all i, $\boldsymbol{I}_n - \boldsymbol{S}_t$ is positive semi-definite, which means that its singular and eigenvalues coincide and thus that its trace equals the sum of its singular values. Consequently

$$\frac{\left\|(\boldsymbol{I}_n - \boldsymbol{S}_t)^\top(\boldsymbol{I}_n - \boldsymbol{S}_t)\right\|}{\left(\operatorname{Tr}(\boldsymbol{I}_n - \boldsymbol{S}_t)\right)^2} = \begin{cases} \frac{\sum_{i=1}^n a_i^2}{\left(\sum_{i=1}^n a_i\right)^2} & \text{for the nuclear (and trace semi) norm.}\\ \sqrt{\frac{\sum_{i=1}^n a_i^4}{\left(\sum_{i=1}^n a_i\right)^4}} & \text{for the Frobenius norm.}\\ \frac{\max_{i=1,\dots n} a_i^2}{\left(\sum_{i=1}^n a_i\right)^2} & \text{for the spectral norm.}\end{cases} \tag{16}$$

For the nuclear and Frobenius norms, we can use that for any vector, $\boldsymbol{v} \in \mathbb{R}^n$, for $d > 1$,

$$\|\boldsymbol{v}\|_d = \left(\sum_{i=1}^n |v_i|^d\right)^{1/d} \geq n^{1-1/d}\cdot\|\boldsymbol{v}\|_1 = n^{1-1/d}\cdot\sum_{i=1}^n |v_i|,$$

with equality iff $v_i = 1$ for all i. Thus the two corresponding expressions in Equation 16 are minimized when $a_i = 1$ for all i, i.e. for $t = 0$, in which case their values are

$$\lim_{t\to 0}\left(\frac{\left\|(\boldsymbol{I}_n-\boldsymbol{S}_t)^\top(\boldsymbol{I}_n-\boldsymbol{S}_t)\right\|_*}{(\mathrm{Tr}(\boldsymbol{I}_n-\boldsymbol{S}_t))^2}\right)=\frac{1}{n}$$

$$\lim_{t\to 0}\left(\frac{\left\|(\boldsymbol{I}_n-\boldsymbol{S}_t)^\top(\boldsymbol{I}_n-\boldsymbol{S}_t)\right\|_F}{(\mathrm{Tr}(\boldsymbol{I}_n-\boldsymbol{S}_t))^2}\right)=\frac{1}{n\sqrt{n}}.$$

For the spectral norm, let $a_m := \max_{i=1,\dots n} a_i$. Then the corresponding expression in Equation 16 becomes $a_m^2/\left(\sum_{i=1}^n a_i\right)^2 = 1/\left(\sum_{i=1}^n \frac{a_i}{a_m}\right)^2$. Here, $\frac{a_i}{a_m}\le 1$ is maximized when $a_i = a_m$ for all i, i.e. when $t=0$ and $a_i=1$ for all i. In this case,

$$\lim_{t\to 0}\left(\frac{\left\|(\boldsymbol{I}_n-\boldsymbol{S}_t)^\top(\boldsymbol{I}_n-\boldsymbol{S}_t)\right\|_2}{(\mathrm{Tr}(\boldsymbol{I}_n-\boldsymbol{S}_t))^2}\right)=\frac{1}{n^2}.$$

Part (b)

For part *(b)*,

$$n\cdot\left\|\frac{1}{n}\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\| = n\cdot\left\|\frac{1}{n}\boldsymbol{I}_n-\frac{1}{n}\left(\boldsymbol{I}_n-\exp(-t\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)\right)^2\right\|$$

$$=\begin{cases}\sum_{i=1}^n\left|1-\left(1-e^{-ts_i^2}\right)^2\right| & \text{for the nuclear (and trace semi) norm.}\\ \sum_{i=1}^n\left(1-\left(1-e^{-ts_i^2}\right)^2\right)^2 & \text{for the squared Frobenius norm.}\\ \max_{i=1,\dots n}\left|1-\left(1-e^{-ts_i^2}\right)^2\right| & \text{for the spectral norm.}\end{cases}$$

If $s_i>0$, $e^{-ts_i^2}\in[0,1]$ for $t\ge 0$, and equals 0 only for $t=\infty$. Hence, $\left|1-\left(1-e^{-ts_i^2}\right)^2\right|$ (and thus $\left(1-\left(1-e^{-ts_i^2}\right)^2\right)^2$) is minimized for $t=\infty$, in which case it is 0. If $s_i=0$, $\left|1-\left(1-e^{-ts_i^2}\right)^2\right|=1$ regardless of t. For the nuclear and Frobenius norms, since each term of the sum obtains its minimum for $t=\infty$, so does the sum. For the spectral norm, the minimum is obtained either for $t=\infty$, if $\arg\max_{s_i}\left|1-\left(1-e^{-ts_i^2}\right)^2\right|>0$, or for any value of t if $\arg\max_{s_i}\left|1-\left(1-e^{-ts_i^2}\right)^2\right|=0$.

Analogously,

$$n\cdot\left\|\frac{1}{n}\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{S}_t\right\| = n\cdot\left\|\frac{1}{n}\boldsymbol{I}_n-\frac{1}{n}\left(\boldsymbol{I}_n-\exp(-t\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)\right)\right\|$$

$$=\begin{cases}\sum_{i=1}^n\left|e^{-ts_i^2}\right| & \text{for the nuclear (and trace semi) norm.}\\ \sum_{i=1}^n\left(e^{-ts_i^2}\right)^2 & \text{for the squared Frobenius norm.}\\ \max_{i=1,\dots n}\left|e^{-ts_i^2}\right| & \text{for the spectral norm.}\end{cases}$$

where, for $t\ge 0$, $e^{-ts_i^2}\in[0,1]$ equals 0 only for $t=\infty$, if $s_i>0$, and $e^{-ts_i^2}=1$ regardless of t if $s_i=0$. Again, for the nuclear and Frobenius norms, each term of the sums obtains its minimum for $t=\infty$, and so do the sums. For the spectral norm, the minimum is obtained either for $t=\infty$, if $\arg\max_{s_i} e^{-ts_i}>0$, or for any value of t if $\arg\max_{s_i} e^{-ts_i}=0$.

Part (c)

For part *(c)*, we define $\tau = 1/t$, and want to show that the derivative of $\left\|\frac{1}{n}\boldsymbol{I}-\mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|$ with respect to τ is strictly negative for $\tau=0$, strictly positive for $M_1<\tau<M_2$ and non-negative for $\tau\ge M_2$, in which case the minimum of the norm is obtained for $\tau\in(0,M_1)$, i.e. for $t>1/M_1$.

Since

$$\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top}) = \frac{1}{n}\boldsymbol{I}_n - \left(\boldsymbol{I}_n - \exp(-\boldsymbol{\Phi}\boldsymbol{\Phi}^\top/\tau)\right)(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}\underbrace{\mathbb{E}(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*\top})}_{=\boldsymbol{I}_p}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}(\boldsymbol{I}_n - \exp(-\boldsymbol{\Phi}\boldsymbol{\Phi}^\top/\tau))$$

$$= \frac{1}{n}\boldsymbol{I}_n - \left(\boldsymbol{I}_n - \exp(-\boldsymbol{\Phi}\boldsymbol{\Phi}^\top/\tau)\right)^2(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1},$$

$$\begin{aligned}
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_T &= \left|\sum_{i=1}^n\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right)\right| \\
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_* &= \sum_{i=1}^n\left|\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right| \\
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_F &= \sum_{i=1}^n\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right)^2 \\
\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_2 &= \max_{i=1,\dots n}\left|\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right| = \left|\frac{1}{n} - \left(\frac{1-e^{-s_m^2/\tau}}{s_i}\right)^2\right|,
\end{aligned} \tag{17}$$

where $s_m := \arg\max_{s_i}\left|\frac{1}{n} - \left(\frac{1-e^{-ts_i^2}}{s_i}\right)^2\right|$.

The needed derivatives with respect to τ are

$$\begin{aligned}
\frac{\partial}{\partial\tau}\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) &= \frac{2}{\tau^2}\cdot\left(1-e^{-s_i^2/\tau}\right)e^{-s_i^2/\tau} \\
\frac{\partial}{\partial\tau}\left|\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right| &= \frac{2}{\tau^2}\cdot\left(1-e^{-s_i^2/\tau}\right)e^{-s_i^2/\tau}\cdot\operatorname{sign}\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) \\
\frac{\partial}{\partial\tau}\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right)^2 &= \frac{4}{\tau^2}\cdot\left(1-e^{-s_i^2/\tau}\right)e^{-s_i^2/\tau}\cdot\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right)
\end{aligned}$$

and hence

$$\begin{aligned}
\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_T &= \sum_{i=1}^n\frac{2}{\tau^2}\cdot\left(1-e^{-s_i^2/\tau}\right)e^{-s_i^2/\tau}\cdot\operatorname{sign}\left(\sum_{i=1}^n\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right)\right) \\
\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_* &= \sum_{i=1}^n\frac{2}{\tau^2}\cdot\left(1-e^{-s_i^2/\tau}\right)e^{-s_i^2/\tau}\cdot\operatorname{sign}\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) \\
\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_F &= \sum_{i=1}^n\frac{4}{\tau^2}\cdot\left(1-e^{-s_i^2/\tau}\right)e^{-s_i^2/\tau}\cdot\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) \\
\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_2 &= \frac{2}{\tau^2}\cdot\left(1-e^{-s_m^2/\tau}\right)e^{-s_m^2/\tau}\cdot\operatorname{sign}\left(\frac{1}{n} - \left(\frac{1-e^{-s_m^2/\tau}}{s_i}\right)^2\right).
\end{aligned} \tag{18}$$

We first consider large τ's. For the trace seminorm and the nuclear and Frobenius norm, we note that $\tau > \frac{s_i^2}{\log\left(\sqrt{n}/(\sqrt{n}-s_i)\right)} \implies \left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) > 0 \implies \sum_{i=1}^n\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) > 0$, and hence, if

$\max_{i=1,\dots n}\left(\frac{s_i^2}{\log\left(\sqrt{n}/(\sqrt{n}-s_i)\right)}\right) =: M_1 < \tau < \infty$, each term in the three sums in Equation 18 sum is strictly positive for $s_i > 0$, and zero for $s_i = 0$. Consequently, for $M_1 < \tau < \infty$,

$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_T > 0$$
$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_* > 0$$
$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_F > 0.$$

For the spectral norm, to obtain $\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_2 > 0$ when $s_m^2/\log\left(\sqrt{n}/(\sqrt{n}-s_i)\right) =: M_m < \tau < \infty$, we also need $s_m > 0$. If $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is non-singular, all its singular values are non-zero, and this is trivially fulfilled. However, if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is singular, at least one s_i is zero. We call this singular value s_0 (i.e. $s_0 = 0$) and note (by Taylor expanding $e^{-s_0^2/\tau}$) that $\left|\frac{1}{n} - \left(\frac{1-e^{-s_0^2/\tau}}{s_0}\right)^2\right| = \frac{1}{n}$ regardless of τ. Furthermore, if $\tau \geq s_i^2/\log\left(\sqrt{n}/(\sqrt{n}-s_i\sqrt{2})\right)$ for all s_i, then $\left|\frac{1}{n} - \left(1-e^{-s_i^2/\tau}\right)^2\right| \leq \frac{1}{n}$. This means that for for $\tau \geq \max_{i=1,\dots n} s_i^2/\log\left(\sqrt{n}/(\sqrt{n}-s_i\sqrt{2})\right)$, $s_m = 0$ and $\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_2 = \frac{1}{n}$.

We now consider $\tau = 0$. For the trace seminorm and for the nuclear and Frobenius norms, if $s_i = 0$, the corresponding terms in the sums in Equation 18 are zero, and thus do not contribute when calculating the derivative. For terms where $s_i > 0$, $\lim_{\tau\to 0}\left(\frac{1}{n} - \left(\frac{1-e^{-s_i^2/\tau}}{s_i}\right)^2\right) = \left(\frac{1}{n} - \frac{1}{s_i^2}\right)$. Since $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$ by assumption, $s_i^2 \in (0,n)$ for all i, and we can write $s_i^2 = n - a_i$ for $0 < a_i < n$. Hence,

$$\left(\frac{1}{n} - \frac{1}{s_i^2}\right) = \left(\frac{1}{n} - \frac{1}{n-a_i}\right) = \left(-\frac{a_i}{n(n-a_i)}\right) < 0.$$

Thus, since all terms in the three sums in Equation 18 are either 0, if $s_i = 0$, or negative, if $s_i > 0$, the sums, and hence the derivatives, are negative: For $\tau = 0$,

$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_T < 0$$
$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_* < 0$$
$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_F < 0.$$

For the spectral norm, we again need $s_m > 0$. If $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is non-singular, this is trivial, while if $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is singular, there exists, by assumption, at least one $s_i^2 \in (0, n/2)$, for which $\frac{1}{s_i^2} - \frac{1}{n} > \frac{2}{n} - \frac{1}{n} = \frac{1}{n}$. Thus, since the maximum in Equation 17 is obtained for an i such that $s_i > 0$, we have $s_m > 0$, and

$$\frac{\partial}{\partial\tau}\left\|\frac{1}{n}\boldsymbol{I}_n - \mathbb{E}(\boldsymbol{s}_\tau^*\boldsymbol{s}_\tau^{*\top})\right\|_2\Bigg|_{\tau=0} = \frac{2}{s_m^4}\cdot\operatorname{sign}\left(\frac{1}{n} - \frac{1}{s_m^2}\right).$$

Again, since $\|\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\|_2 < n$, $s_m^2 \in (0,n)$ and we can write $s_m^2 = n - a_m$ for $0 < a_m < n$, and

$$\left(\frac{1}{n} - \frac{1}{s_m^2}\right) = \left(\frac{1}{n} - \frac{1}{n-a_m}\right) = \left(-\frac{a_m}{n(n-a_m)}\right) < 0.$$

□

Proof of Theorem 3.
For gradient descent with momentum, $\hat{\boldsymbol{\theta}}$ is updated according to

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}_{-1} = \hat{\boldsymbol{\theta}}_0, \\
&\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \gamma \cdot \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right) - \eta \cdot \frac{\partial L(\hat{\boldsymbol{f}}_k, \boldsymbol{y})}{\partial \boldsymbol{\theta}_k},\ k = 0, 1, \ldots,
\end{aligned}$$

where, according to the chain rule,

$$\frac{\partial L(\hat{\boldsymbol{f}}_k, \boldsymbol{y})}{\partial \boldsymbol{\theta}_k} = \left(\frac{\partial \hat{\boldsymbol{f}}_k}{\partial \boldsymbol{\theta}_k}\right)^\top \frac{\partial L(\hat{\boldsymbol{f}}_k, \boldsymbol{y})}{\partial \hat{\boldsymbol{f}}_k} = \left(\frac{\partial \hat{\boldsymbol{f}}_k}{\partial \boldsymbol{\theta}_k}\right)^\top \boldsymbol{I}_{\tilde{\boldsymbol{F}}_k} \cdot (\hat{\boldsymbol{f}}_k - \boldsymbol{y}),$$

where $\boldsymbol{I}_{\tilde{\boldsymbol{F}}_k} = \boldsymbol{I}_n$ for the squared loss and $\boldsymbol{I}_{\tilde{\boldsymbol{F}}_k} = \tilde{\boldsymbol{F}}_k$ for the cross-entropy error (see Equation 11). That is, $\hat{\boldsymbol{\theta}}$ is updated according to

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}_{-1} = \hat{\boldsymbol{\theta}}_0, \\
&\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \gamma \cdot \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right) + \eta \left(\frac{\partial \hat{\boldsymbol{f}}_k}{\partial \boldsymbol{\theta}_k}\right)^\top \boldsymbol{I}_{\tilde{\boldsymbol{F}}_k} \cdot (\boldsymbol{y} - \hat{\boldsymbol{f}}_k),\ k = 0, 1, \ldots.
\end{aligned}$$

The proof consists of two steps: First, we define the function $\hat{\boldsymbol{f}}_k'^\star$ according to

$$\begin{aligned}
&\hat{\boldsymbol{f}}_{-1}'^\star = \hat{\boldsymbol{f}}_0'^\star = \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_0) \\
&\hat{\boldsymbol{f}}_{k+1}'^\star = \hat{\boldsymbol{f}}_k'^\star + \gamma \cdot \left(\hat{\boldsymbol{f}}_k'^\star - \hat{\boldsymbol{f}}_{k-1}'^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_{k+1}^\star \cdot \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k'\right),\ k = 0, 1, \ldots
\end{aligned}$$

and show that

$$\begin{aligned}
&\boldsymbol{S}_{-1}^\star = \boldsymbol{S}_0^\star = \boldsymbol{0} \\
&\boldsymbol{S}_{k+1}^\star = \boldsymbol{S}_k^\star + \gamma \cdot \left(\boldsymbol{S}_k^\star - \boldsymbol{S}_{k-1}^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_{k+1}^\star \cdot \left(\boldsymbol{I}_n - \boldsymbol{S}_k\right),\ k = 0, 1, \ldots
\end{aligned}$$

implies that

$$\hat{\boldsymbol{f}}_k'^\star = \boldsymbol{S}_k^\star \left(\boldsymbol{y} - \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}_0)\right) + \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_0). \tag{19}$$

Then, we show that there exists a $C < \infty$, such that

$$\left\| \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_k) - \hat{\boldsymbol{f}}_k'^\star \right\|_\infty \leq \eta \cdot C. \tag{20}$$

To alleviate notation, we will write $\hat{\boldsymbol{f}}_k^\star := \hat{\boldsymbol{f}}^\star(\hat{\boldsymbol{\theta}}_k)$.

We first show Equation 19 by induction:

For $k = 0$,

$$\begin{aligned}
&\boldsymbol{S}_0^\star = \boldsymbol{0} \\
&\hat{\boldsymbol{f}}_0'^\star = \hat{\boldsymbol{f}}_0 = \boldsymbol{0} = \boldsymbol{0} \cdot (\boldsymbol{y} - \boldsymbol{0}) + \boldsymbol{0} = \boldsymbol{S}_0^\star \cdot \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_0\right) + \hat{\boldsymbol{f}}_0^\star.
\end{aligned}$$

For $k = 1$,

$$\begin{aligned}
\boldsymbol{S}_1^\star &= \boldsymbol{S}_0^\star + \gamma \cdot \left(\boldsymbol{S}_0^\star - \boldsymbol{S}_{-1}^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_1^\star \left(\boldsymbol{I}_n - \boldsymbol{S}_0\right) = \boldsymbol{0} + \gamma \cdot (\boldsymbol{0} - \boldsymbol{0}) + \eta \cdot \tilde{\boldsymbol{K}}_1^\star \left(\boldsymbol{I}_n - \boldsymbol{0}\right) = \eta \cdot \tilde{\boldsymbol{K}}_1^\star \\
\hat{\boldsymbol{f}}_1'^\star &= \hat{\boldsymbol{f}}_0'^\star + \gamma \cdot \left(\hat{\boldsymbol{f}}_0'^\star - \hat{\boldsymbol{f}}_{-1}'^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_1^\star \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_0'\right) = \hat{\boldsymbol{f}}_0^\star + \gamma \cdot \left(\hat{\boldsymbol{f}}_0^\star - \hat{\boldsymbol{f}}_0^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_1^\star \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_0\right) \\
&= \boldsymbol{S}_1^\star \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_0\right) + \hat{\boldsymbol{f}}_0^\star.
\end{aligned}$$

For $k + 1 \geq 2$,

$$
\begin{aligned}
\hat{\boldsymbol{f}}'^{\star}_{k-1} =& \boldsymbol{S}^{\star}_{k-1}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\hat{\boldsymbol{f}}^{\star}_0\\
\hat{\boldsymbol{f}}'^{\star}_{k} =& \boldsymbol{S}^{\star}_{k}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\hat{\boldsymbol{f}}^{\star}_0\\
\boldsymbol{S}^{\star}_{k+1} =& \boldsymbol{S}^{\star}_{k}+\gamma\cdot\left(\boldsymbol{S}^{\star}_{k}-\boldsymbol{S}^{\star}_{k-1}\right)+\eta\cdot\tilde{\boldsymbol{K}}^{\star}_{k+1}\cdot\left(\boldsymbol{I}_n-\boldsymbol{S}_k\right)\\
\hat{\boldsymbol{f}}'^{\star}_{k+1} =& \hat{\boldsymbol{f}}'^{\star}_{k}+\gamma\cdot\left(\hat{\boldsymbol{f}}'^{\star}_{k}-\hat{\boldsymbol{f}}'^{\star}_{k-1}\right)+\eta\cdot\tilde{\boldsymbol{K}}^{\star}_{k+1}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}'_k\right)\\
=& \boldsymbol{S}^{\star}_{k}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\hat{\boldsymbol{f}}^{\star}_0+\gamma\cdot\left(\boldsymbol{S}^{\star}_{k}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\hat{\boldsymbol{f}}^{\star}_0-\boldsymbol{S}^{\star}_{k-1}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)-\hat{\boldsymbol{f}}^{\star}_0\right)\\
&+\eta\cdot\tilde{\boldsymbol{K}}^{\star}_{k+1}\left(\boldsymbol{y}-\boldsymbol{S}^{\star}_{k}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)-\hat{\boldsymbol{f}}^{\star}_0\right)\\
=& \left(\boldsymbol{S}^{\star}_{k}+\gamma\cdot\left(\boldsymbol{S}^{\star}_{k}-\boldsymbol{S}^{\star}_{k-1}\right)\right)\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\eta\cdot\tilde{\boldsymbol{K}}^{\star}_{k+1}\left(\boldsymbol{I}_n-\boldsymbol{S}_k\right)\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\hat{\boldsymbol{f}}^{\star}_0\\
=& \boldsymbol{S}^{\star}_{k+1}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}_0\right)+\hat{\boldsymbol{f}}^{\star}_0.
\end{aligned}
$$

The proof of Equation 20 is to a large extent based on the mean value inequality,

$$
\|\boldsymbol{f}(x+\Delta x)-\boldsymbol{f}(x)\|_2 \le \max_x\left\|\frac{\partial\boldsymbol{f}(x)}{\partial x}\right\|_2\cdot|\Delta x|, \tag{21}
$$

and the second-order Taylor expansion,

$$
f(\boldsymbol{x}+\boldsymbol{\Delta x})=f(\boldsymbol{x})+\boldsymbol{\Delta x}^{\top}\left(\frac{\partial f(\boldsymbol{x})}{\partial\boldsymbol{x}}\right)+r,\ |r|\le\frac{1}{2}\|\boldsymbol{\Delta x}\|_2^2\cdot\max_{\boldsymbol{x}}\left\|\frac{\partial^2 f(\boldsymbol{x})}{\partial\boldsymbol{x}^2}\right\|_2. \tag{22}
$$

We first show that

$$
\begin{aligned}
&\hat{\boldsymbol{f}}^{\star}_{k+1}=\hat{\boldsymbol{f}}^{\star}_{k}+\gamma\cdot\left(\hat{\boldsymbol{f}}^{\star}_{k}-\hat{\boldsymbol{f}}^{\star}_{k-1}\right)+\eta\cdot\tilde{\boldsymbol{K}}^{\star}_{k+1}\cdot\left(\boldsymbol{y}-\hat{\boldsymbol{f}}^{\star}_{k}\right)+\eta^2\cdot\boldsymbol{r}\\
&\text{where } \|\boldsymbol{r}\|_\infty\le M_1:=\frac{1}{2}M_{f_{\theta^2}}\left(\gamma M_{\theta_t}^2+\left(\gamma M_{\theta_t}+M_yM_{f_\theta}\right)^2\right),
\end{aligned} \tag{23}
$$

for

$$
\begin{aligned}
&M_{\theta_t}:=\max_t\left\|\frac{\partial\hat{\boldsymbol{\theta}}(t)}{\partial t}\right\|_2,\ M_{f_\theta}:=\max_t\left\|\frac{\partial\hat{\boldsymbol{f}}^{\star}(t)}{\partial\hat{\boldsymbol{\theta}}}\right\|_2,\ M_{f_{\theta^2}}:=\max_t\left\|\frac{\partial^2\hat{f}^{\star}(t)}{\partial\hat{\boldsymbol{\theta}}^2}\right\|_2,\\
&M_y:=\max_t\left\|\boldsymbol{y}-\hat{\boldsymbol{f}}^{\star}_k\right\|_2.
\end{aligned}
$$

To keep track of η, we switch between reminder terms of the form $\tilde{r}$, which include η, and r, which do not, i.e. $\tilde{r}=\eta r$.

By applying Equation 21 to $\hat{\boldsymbol{\theta}}(t)$, where $\hat{\boldsymbol{\theta}}_k=\hat{\boldsymbol{\theta}}(k\eta)$, we obtain

$$
\left\|\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right\|_2=\left\|\hat{\boldsymbol{\theta}}\left(k\eta\right)-\hat{\boldsymbol{\theta}}\left((k-1)\eta\right)\right\|_2\le\max_t\left\|\frac{\partial\hat{\boldsymbol{\theta}}(t)}{\partial t}\right\|_2\cdot|k\eta-(k-1)\eta|=M_{\theta_t}\cdot\eta
$$

Applying Equation 22 element-wise to $\hat{\boldsymbol{f}}^{\star}_k$, in combination with Equation 21, we obtain

$$
\begin{aligned}
\hat{f}^{\star}_k-\hat{f}^{\star}_{k-1}&=\hat{f}^{\star}\left(\hat{\boldsymbol{\theta}}_k\right)-\hat{f}^{\star}\left(\hat{\boldsymbol{\theta}}_{k-1}\right)=\hat{f}^{\star}\left(\hat{\boldsymbol{\theta}}_k\right)-\hat{f}^{\star}\left(\hat{\boldsymbol{\theta}}_k-\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right)\right)\\
&=\hat{f}^{\star}\left(\hat{\boldsymbol{\theta}}_k\right)-\left(\hat{f}^{\star}\left(\hat{\boldsymbol{\theta}}_k\right)-\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right)^{\top}\left(\frac{\partial\hat{f}^{\star}_k}{\partial\hat{\boldsymbol{\theta}}}\right)+\tilde{r}_1\right)\\
&=\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right)^{\top}\left(\frac{\partial\hat{f}^{\star}_k}{\partial\hat{\boldsymbol{\theta}}_k}\right)-\tilde{r}_1\iff\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right)^{\top}\left(\frac{\partial\hat{f}^{\star}_k}{\partial\hat{\boldsymbol{\theta}}_k}\right)=\hat{f}^{\star}_k-\hat{f}^{\star}_{k-1}+\tilde{r}_1,\\
|\tilde{r}_1|&\le\frac{1}{2}\left\|\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right\|_2^2\cdot\max_t\left\|\frac{\partial^2\hat{f}^{\star}(t)}{\partial\hat{\boldsymbol{\theta}}_k^2}\right\|_2\le\frac{1}{2}\eta^2M_{\theta_t}^2M_{f_{\theta^2}}\\
\iff&\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\theta}}_{k-1}\right)^{\top}\left(\frac{\partial\hat{f}^{\star}_k}{\partial\hat{\boldsymbol{\theta}}_k}\right)=\hat{f}^{\star}_k-\hat{f}^{\star}_{k-1}+\eta^2\cdot r_1,\ |r_1|\le\frac{1}{2}M_{\theta_t}^2M_{f_{\theta^2}}.
\end{aligned}
$$

With $\tilde{\boldsymbol{k}}_k^\star := \boldsymbol{I}_{\tilde{\boldsymbol{F}}_k}^\top \cdot \frac{\partial \hat{\boldsymbol{f}}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k} \cdot \frac{\partial \hat{f}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}$, we obtain

$$
\begin{aligned}
\hat{f}_{k+1}^\star =& \hat{f}^\star\left(\hat{\boldsymbol{\theta}}_{k+1}\right) = \hat{f}^\star\left(\hat{\boldsymbol{\theta}}_k + \gamma \cdot \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right) + \eta \cdot \left(\frac{\partial \hat{\boldsymbol{f}}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right)^\top \boldsymbol{I}_{\tilde{\boldsymbol{F}}_k}\left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right)\right) \\
=& \hat{f}^\star\left(\hat{\boldsymbol{\theta}}_k\right) + \gamma \cdot \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right)^\top \left(\frac{\partial \hat{f}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right) + \left(\eta \cdot \left(\frac{\partial \hat{\boldsymbol{f}}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right)^\top \boldsymbol{I}_{\tilde{\boldsymbol{F}}_k}\left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right)\right)^\top \left(\frac{\partial \hat{f}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right) + \tilde{r}_2 \\
=& \hat{f}_k^\star + \gamma \cdot \left(\hat{f}_k^\star - \hat{f}_{k-1}^\star\right) + \gamma \cdot \eta^2 \cdot r_1 + \eta \cdot \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right)^\top \tilde{\boldsymbol{k}}_k^\star + \tilde{r}_2 \\
=& \hat{f}_k^\star + \gamma \cdot \left(\hat{f}_k^\star - \hat{f}_{k-1}^\star\right) + \eta \cdot \tilde{\boldsymbol{k}}_k^{\star\top}\left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right) + \underbrace{\tilde{r}}_{=:\gamma\cdot\eta^2\cdot r_1 + \tilde{r}_2}, \\
|\tilde{r}_2| \leq& \frac{1}{2} \cdot \left\| \gamma \cdot \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right) + \eta \cdot \left(\frac{\partial \hat{\boldsymbol{f}}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right)^\top \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right) \right\|^2 \cdot M_{f_{\theta^2}} \\
\leq& \frac{1}{2} \cdot \left(\gamma^2 \cdot \left\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right\|^2 + 2\gamma \cdot \left\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right\| \cdot \eta \left\|\frac{\partial \hat{\boldsymbol{f}}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right\| \cdot \left\|\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right\| \right. \\
& \left. + \eta^2 \left\|\frac{\partial \hat{\boldsymbol{f}}_k^\star}{\partial \hat{\boldsymbol{\theta}}_k}\right\|^2 \cdot \left\|\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right\|^2 \right) \cdot M_{f_{\theta^2}} \\
\leq& \frac{1}{2}\left(\gamma^2\eta^2 M_{\theta_t}^2 + 2\gamma\eta^2 M_{\theta_t} M_{f_\theta} M_y + \eta^2 M_{f_\theta}^2 M_y^2\right) \cdot M_{f_{\theta^2}} \\
=& \frac{1}{2}\eta^2 M_{f_{\theta^2}}\left(\gamma^2 M_{\theta_t}^2 + 2\gamma M_{\theta_t} M_{f_\theta} M_y + M_{f_\theta} M_y^2\right) = \frac{1}{2}\eta^2 M_{f_{\theta^2}}\left(\gamma M_{\theta_t} + M_y M_{f_\theta}\right)^2 \\
|\tilde{r}| \leq& \eta^2 \cdot \gamma \cdot |r_1| + |\tilde{r}_2| \leq \eta^2 \cdot \gamma \cdot \frac{1}{2} M_{\theta_t}^2 M_{f_{\theta^2}} + \frac{1}{2}\eta^2 M_{f_{\theta^2}}\left(\gamma M_{\theta_t} + M_y M_{f_\theta}\right)^2 \\
=& \frac{\eta^2}{2} \cdot M_{f_{\theta^2}}\left(\gamma M_{\theta_t}^2 + \left(\gamma M_{\theta_t} + M_y M_{f_\theta}\right)^2\right) \\
\iff& \hat{f}_{k+1}^\star = \hat{f}_k^\star + \gamma \cdot (\hat{f}_k^\star - \hat{f}_{k-1}^\star) + \eta \cdot \tilde{\boldsymbol{k}}_k^{\star\top}\left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right) + \eta^2 \cdot r, \\
& |r| \leq \frac{\eta^2}{2} \cdot M_{f_{\theta^2}}\left(\gamma M_{\theta_t}^2 + \left(\gamma M_{\theta_t} + M_y M_{f_\theta}\right)^2\right) \\
\implies& \hat{\boldsymbol{f}}_{k+1}^\star = \hat{\boldsymbol{f}}_k^\star + \gamma \cdot \left(\hat{\boldsymbol{f}}_k^\star - \hat{\boldsymbol{f}}_{k-1}^\star\right) + \eta \cdot \tilde{\boldsymbol{K}}_k^\star\left(\boldsymbol{y} - \hat{\boldsymbol{f}}_k\right) + \eta^2 \cdot \boldsymbol{r} \\
& \|\boldsymbol{r}\|_\infty \leq \frac{\eta^2}{2} \cdot M_{f_{\theta^2}}\left(\gamma M_{\theta_t}^2 + \left(\gamma M_{\theta_t} + M_y M_{f_\theta}\right)^2\right) =: \eta^2 \cdot M_1,
\end{aligned}
$$

which proves Equation 23.

To prove Equation 20, we first apply Equation 22 element-wise to $\hat{\boldsymbol{f}}_k^\star$, to obtain

$$
\begin{aligned}
& \left(\hat{f}_{k+1}^\star - \hat{f}_k^\star\right) - \left(\hat{f}_k^\star - \hat{f}_{k-1}^\star\right) = \left(\hat{f}^\star\left((k+1)\eta\right) - \hat{f}^\star\left(k\eta\right)\right) - \left(\hat{f}^\star\left(k\eta\right) - \hat{f}^\star\left((k-1)\eta\right)\right) \\
& = \left(\hat{f}^\star\left(k\eta\right) + \eta \cdot \partial_t \hat{f}^\star\left(k\eta\right) + \tilde{r}_{f_1} - \hat{f}^\star\left(k\eta\right)\right) - \left(\hat{f}^\star\left(k\eta\right) - \hat{f}^\star\left(k\eta\right) + \eta \cdot \partial_t \hat{f}^\star\left(k\eta\right) - \tilde{r}_{f_2}\right) \\
& = \tilde{r}_{f_1} + \tilde{r}_{f_2} =: \tilde{r}_f,\ |\tilde{r}_f| \leq 2 \cdot \frac{1}{2}\eta^2 \cdot \max_t \left|\frac{\partial^2 \hat{f}^\star(t)}{\partial t^2}\right| \\
\iff& \hat{f}_k^\star - \hat{f}_{k-1}^\star = \hat{f}_{k+1}^\star - \hat{f}_k^\star + \eta^2 \cdot r_f,\ |r_f| \leq M_{f_t^2} \\
\implies& \hat{\boldsymbol{f}}_k^\star - \hat{\boldsymbol{f}}_{k-1}^\star = \hat{\boldsymbol{f}}_{k+1}^\star - \hat{\boldsymbol{f}}_k^\star + \eta^2 \cdot \boldsymbol{r_f},\ \|\boldsymbol{r_f}\|_\infty \leq M_{f_t^2},
\end{aligned}
$$

where

$$
M_{f_t^2} := \max\left(\max_t \left|\frac{\partial^2 \hat{f}^\star(t)}{\partial t^2}\right|,\ \max_t \left|\frac{\partial^2 \hat{f}'^\star(t)}{\partial t^2}\right|\right).
$$

Thus,

$$
\begin{aligned}
\hat{\boldsymbol{f}}^{\star}_{k+1} - \hat{\boldsymbol{f}}'^{\star}_{k+1} =& \hat{\boldsymbol{f}}^{\star}_{k} + \gamma \cdot (\hat{\boldsymbol{f}}^{\star}_{k} - \hat{\boldsymbol{f}}^{\star}_{k-1}) + \eta \cdot \tilde{\boldsymbol{K}}^{\star}_{k} \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_{k} \right) + \eta^2 \cdot \boldsymbol{r} \\
& - \hat{\boldsymbol{f}}'^{\star}_{k} - \gamma \cdot (\hat{\boldsymbol{f}}'^{\star}_{k} - \hat{\boldsymbol{f}}'^{\star}_{k-1}) - \eta \cdot \tilde{\boldsymbol{K}}^{\star}_{k} \left(\boldsymbol{y} - \hat{\boldsymbol{f}}'_{k} \right) \\
=& \hat{\boldsymbol{f}}^{\star}_{k} + \gamma \cdot (\hat{\boldsymbol{f}}^{\star}_{k+1} - \hat{\boldsymbol{f}}^{\star}_{k}) + \eta \cdot \tilde{\boldsymbol{K}}^{\star}_{k} \left(\boldsymbol{y} - \hat{\boldsymbol{f}}_{k} \right) \\
& - \hat{\boldsymbol{f}}'^{\star}_{k} - \gamma \cdot (\hat{\boldsymbol{f}}'^{\star}_{k+1} - \hat{\boldsymbol{f}}'^{\star}_{k}) - \eta \cdot \tilde{\boldsymbol{K}}^{\star}_{k} \left(\boldsymbol{y} - \hat{\boldsymbol{f}}'_{k} \right) + \eta^2 \cdot \underbrace{\left(\boldsymbol{r_f} + \boldsymbol{r} - \boldsymbol{r_{f'}} \right)}_{=: \boldsymbol{r_3}} \\
=& \hat{\boldsymbol{f}}^{\star}_{k} - \hat{\boldsymbol{f}}'^{\star}_{k} + \gamma \cdot \left(\left(\hat{\boldsymbol{f}}^{\star}_{k+1} - \hat{\boldsymbol{f}}'^{\star}_{k+1} \right) - \left(\hat{\boldsymbol{f}}^{\star}_{k} - \hat{\boldsymbol{f}}'^{\star}_{k} \right) \right) - \eta \cdot \tilde{\boldsymbol{K}}^{\star}_{k} \left(\hat{\boldsymbol{f}}^{\star}_{k} - \hat{\boldsymbol{f}}'^{\star}_{k} \right) + \eta^2 \cdot \boldsymbol{r_3} \\
\|\boldsymbol{r_3}\|_{\infty} \leq& \eta^2 \cdot M_1 + \eta^2 \cdot 2 M_{f_t^2} =: \eta^2 \cdot \tilde{C}.
\end{aligned}
\tag{24}
$$

For $\hat{\boldsymbol{e}}_k := \hat{\boldsymbol{f}}^{\star}_{k} - \hat{\boldsymbol{f}}'^{\star}_{k}$, and $\hat{e}_k := \|\hat{\boldsymbol{e}}_k\|_{\infty}$, we obtain, from Equation 24,

$$
\begin{aligned}
\hat{\boldsymbol{e}}_{k+1} &= \hat{\boldsymbol{e}}_k + \gamma \cdot (\hat{\boldsymbol{e}}_{k+1} - \hat{\boldsymbol{e}}_k) - \eta \cdot \tilde{\boldsymbol{K}}^{\star}_{k} \cdot \hat{\boldsymbol{e}}_k + \eta^2 \cdot \boldsymbol{r_3} \\
(1-\gamma)\hat{e}_{k+1} &\leq (1-\gamma)\hat{e}_k + \eta \cdot M^2_{f_\theta} \cdot \hat{e}_k + \eta^2 \tilde{C} \\
\hat{e}_{k+1} &\leq \left(1 + \frac{\eta}{1-\gamma} \cdot M^2_{f_\theta} \right) \hat{e}_k + \frac{\eta^2}{1-\gamma} \tilde{C}.
\end{aligned}
$$

Since, by definition, $\hat{e}_0 = \left\| \hat{\boldsymbol{f}}^{\star}_{0} - \hat{\boldsymbol{f}}'^{\star}_{0} \right\|_{\infty} = \left\| \hat{\boldsymbol{f}}^{\star}_{0} - \hat{\boldsymbol{f}}^{\star}_{0} \right\|_{\infty} = 0$, recursively applying this inequality above renders the geometric series

$$
\begin{aligned}
\hat{e}_k &\leq \sum_{l=0}^{k} \left(1 + \frac{\eta}{1-\gamma} \cdot M^2_{f_\theta} \right)^l \cdot \frac{\eta^2}{1-\gamma} \tilde{C} = \frac{\left(1 + \frac{\eta}{1-\gamma} \cdot M^2_{f_\theta} \right)^k - 1}{1 + \frac{\eta}{1-\gamma} \cdot M^2_{f_\theta} - 1} \cdot \frac{\eta^2}{1-\gamma} \tilde{C} \\
&= \frac{\left(1 + \frac{\eta}{1-\gamma} \cdot M^2_{f_\theta} \right)^k - 1}{M^2_{f_\theta}} \cdot \eta \tilde{C} \leq \frac{e^{\frac{1}{1-\gamma} \cdot M^2_{f_\theta} \cdot \eta k}}{M^2_{f_\theta}} \cdot \eta \tilde{C} =: \eta \cdot C,
\end{aligned}
$$

for

$$
C := \frac{e^{\frac{1}{1-\gamma} \cdot M^2_{f_\theta} \cdot \eta k}}{M^2_{f_\theta}} \cdot \left(M_1 + 2 M_{f_t^2} \right),
$$

where in the last inequality we have used that for $x > -1$, $(1+x)^k \leq e^{xk}$.

□

Proof of Proposition 2.
Let $\oslash$ denote element-wise division. Then

$$
\begin{aligned}
\frac{\partial \tilde{L}(\hat{\boldsymbol{f}}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{i}})}{\partial \hat{f}_{ij}} &= -\frac{\partial}{\partial \hat{f}_{ij}} \left(\sum_{j=1}^{c-1} y_{ij} \cdot \log(\hat{f}_{ij}) + \left(1 - \sum_{j=1}^{c-1} y_{ij}\right) \cdot \log\left(1 - \sum_{j=1}^{c-1} \hat{f}_{ij}\right) \right) \\
&= -\left(\frac{y_{ij}}{\hat{f}_{ij}} - \frac{1 - \sum_{j=1}^{c-1} y_{ij}}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \right) \implies \frac{\partial L(\hat{\boldsymbol{f}}_{\boldsymbol{i}})}{\partial \hat{\boldsymbol{f}}_{\boldsymbol{i}}} = \frac{1 - \sum_{j=1}^{c-1} y_{ij}}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1} - \boldsymbol{y}_{\boldsymbol{i}} \oslash \hat{\boldsymbol{f}}_{\boldsymbol{i}} \\
&= \frac{1 - \sum_{j=1}^{c-1} y_{ij} - \left(1 - \sum_{j=1}^{c-1} \hat{f}_{ij}\right)}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1} + \frac{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1} - \boldsymbol{y}_{\boldsymbol{i}} \oslash \hat{\boldsymbol{f}}_{\boldsymbol{i}} \\
&= \frac{\sum_{j=1}^{c-1} \hat{f}_{ij} - \sum_{j=1}^{c-1} y_{ij}}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1} + \mathbf{1} - \boldsymbol{y}_{\boldsymbol{i}} \oslash \hat{\boldsymbol{f}}_{\boldsymbol{i}} \\
&= \frac{\mathbf{1}^\top (\hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1} + \hat{\boldsymbol{f}}_{\boldsymbol{i}} \oslash \hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}} \oslash \hat{\boldsymbol{f}}_{\boldsymbol{i}} \\
&= \frac{1}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1}\mathbf{1}^\top (\hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}}) + (\hat{\boldsymbol{F}}_{\boldsymbol{i}})^{-1} (\hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}}) \\
&= \left((\hat{\boldsymbol{F}}_{\boldsymbol{i}})^{-1} + \frac{1}{1 - \sum_{j=1}^{c-1} \hat{f}_{ij}} \cdot \mathbf{1}\mathbf{1}^\top \right) \left(\hat{\boldsymbol{f}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}} \right).
\end{aligned}
$$

$\square$