GROKKING BEYOND THE EUCLIDEAN NORM OF MODEL PARAMETERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Grokking refers to a delayed generalization following overfitting when optimizing artificial neural networks with gradient-based methods. In this work, we demonstrate that grokking can be induced by regularization, either explicit or implicit. More precisely, we show that when there exists a model with a property P (e.g., sparse or low-rank weights) that generalizes on the problem of interest, gradient descent with a small but non-zero regularization of P (e.g., ℓ_1 or nuclear norm regularization) result in grokking. This extends previous work showing that small non-zero weight decay induces grokking. Moreover, our analysis shows that over-parameterization by adding depth makes it possible to grok or ungrok without explicitly using regularization, which is impossible in shallow cases. We further show that the ℓ_2 norm of the model parameters cannot be used as an indicator of grokking in a general setting in place of the regularized property P: the ℓ_2 norm grows in many cases where no weight decay is used, but the model generalizes anyway. We also show that grokking can be amplified through only data selection (with any other hyperparameter fixed).

1 INTRODUCTION

The optimization of machine learning models today relies entirely on gradient descent (GD). The reasons behind the ability of such a procedure to converge towards generalizing solutions are still not fully understood, particularly in over-parameterized regimes. Power et al. (2022) recently observed an even more surprising feature of this optimization procedure, *grokking*: the optimization first goes through a solution that perfectly memorizes the training data, but after a sufficiently long training time, it suddenly converges on a solution that generalizes.

Many works have shown that grokking can be observed by using a large-scale initialization and a small (but non-zero) weight decay (Liu et al., 2023a; Lyu et al., 2023). Moreover, some works have shown that the ℓ_2 norm of the weights can be used during optimization as a progression measure for generalization since it generally decreases during the transition from memorization to generalization (Liu et al., 2023a; Thilak et al., 2022; Varma et al., 2023). All these theories have left open the question of whether we always need an ℓ_2 regularization to observe generalization or whether the ℓ_2 norm of the parameter is always a good predictor of generalization in general. This paper attempts to answer these questions. We hypothesize that the dynamic of grokking goes beyond the ℓ_2 norm, that is: *If there exists a model with a property P (e.g., sparse or low-rank weights) that fits the data, then GD with a small but non-zero regularization of P (e.g., ℓ_1 or nuclear norm regularization) will also result in grokking, provided the number of training sample is large enough. Moreover, the ℓ_2 norm is no longer guaranteed to decrease with generalization when the property sought is not the ℓ_2 norm of the parameters.*

For sparsity, we first focus on a linear teacher-student setup and show that recovery of sparse vectors using gradient descent and a lasso penalty exhibits a grokking phenomenon, which is impossible using only the ℓ_2 regularization no matter the initialization scale as advocated by previous art (Lyu et al., 2023; Liu et al., 2023b). We also formally show that the generalization delay is inversely proportional to the learning rate and the ℓ_1 regularization strength and proportional to the ℓ_∞ norm of the parameters at memorization. Moreover, with a deeper over-parametrized model, there is no need to use ℓ_1, i.e., gradient descent is implicitly biased toward such a sparse solution. For the low-rank structure, we focus on matrix factorization and show that nuclear norm regularization (denoted ℓ_*)

is needed for generalization in the shallow case, and the delay between memorization and perfect recovery is inversely proportional to the strength of the ℓ_* regularization and the learning rate used, and proportional to the large singular value of the iterate at memorization. This extends previous works on matrix factorization that show that deeper linear networks can factorize low-rank matrices without explicit regularization (Arora et al., 2018; 2019). All this holds beyond shallow and/or linear networks. We show that ℓ_1 or ℓ_* can replace ℓ_2 in a more general setting and accelerate generalization, i.e., reduce grokking. We focus on a nonlinear teacher-student setup, on the algorithmic data setup (Power et al., 2022) on which grokking was first observed, with different classes of models (MLP, LSTM), and on image classification with MLP. In a setting where the ℓ_2 regularization is not used, the ℓ_2 norm of the model parameters tends to grow during training and after generalization, but optimization still produces a generalizable solution. We further observe that using ℓ_2 can worsen generalization when the property P differs from the ℓ_2 norm and is necessary for generalization.

Our contributions can be summarized as follows: (i) We show that grokking can be induced by the interplay between the sparse/low-rank structure of the solution and the ℓ_1 / ℓ_* regularization used in training, extending previous results on ℓ_2 regularization (Liu et al., 2023a; Lyu et al., 2023). (ii) For shallow linear networks, we theoretically characterize the relation between grokking time and regularization strength, showing that regularization is necessary to observe grokking on sparse or low-rank solutions. (iii) Moreover, we empirically show that in deep (non-linear) networks, the sparse/low-rank structure of the data is enough to have generalization without explicit regularization. Adding depth makes it possible to grok or ungrok simply from the implicit regularization of gradient descent. (iv) Leveraging the notion of coherence, we show that grokking can be amplified through only data selection (with any other hyperparameter fixed). (v) We show that ℓ_1 or ℓ_* can replace ℓ_2 in a more general setting and reduce grokking. Moreover, in such a scenario, and in the shallow sparse/low-rank scenario mentioned above, the ℓ_2 cannot be used as an indicator of grokking. (vi) We also show that other forms of domain-specific regularizers strongly affect the delay between memorization and generalization.

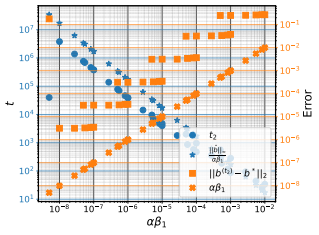


Figure 1: Generalization step t_2 and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2$ as a function of $\alpha\beta_1$. We can see that $t_2 \propto \|\hat{\mathbf{b}}\|_\infty/\alpha\beta_1$ and $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2 \propto \alpha\beta_1$, i.e. small $\alpha\beta_1$ require longer time to converge, but do so at a lower recovery error. The outlier for small $\alpha\beta_1$ is due to insufficient training (Fig. 12).

This paper is organized as follows. We study grokking on sparse recovery and low-rank matrix factorization in section 2. In section 3, we show how our result extends beyond sparse recovery and matrix factorization. We then discuss and conclude our work in section 4.

2 Grokking in Sparse Recovery and Matrix Factorization

Compressed sensing theory provides the foundation for recovering sparse signals from undersampled noisy linear measurements. Given $N \ll n$ measurements $\mathbf{y}^* = \mathcal{F}_{\mathbf{a}^*}(\mathbf{X}) + \boldsymbol{\xi}$ of a vector $\mathbf{a}^* \in \mathbb{R}^n$, where $\mathcal{F}_{\mathbf{a}}(\mathbf{X}) = \mathbf{X}\mathbf{a}$ and $\boldsymbol{\xi}$ denotes noise, we seek a reconstruction of the form $\mathbf{a} = \sum_{i=1}^n \mathbf{b}_i^* \Phi_{:,i} = \Phi\mathbf{b}$, with $\Phi \in \mathbb{R}^{n\times n}$ a dictionary and $s = \|\mathbf{b}^*\|_0 := |\{i, \mathbf{b}_i^* \neq 0\}| \ll n$. The exact recovery problem (P_0), which involves minimizing $\|\mathbf{b}\|_0$ under the constraint of the form $\|\mathcal{F}_{\Phi\mathbf{b}}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon$, is NP-hard. Therefore, we focus on the relaxed problem (P_1), minimizing $\|\mathbf{b}\|_1$ under the same constraint, commonly known as Basis Pursuit. We investigate the optimization dynamics of solving (P_1) through gradient descent by formally characterizing grokking time. More precisely, we want to minimize $f(\mathbf{b}) = \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}\|_2^2 + \beta_1\|\mathbf{b}\|_1$ using gradient descent with a learning rate α. The subgradient update rule for this problem is given by $\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha\left(G_{\beta_2}(\mathbf{b}^{(t)}) + \beta_1 h(\mathbf{b}^{(t)})\right)$ where $G_{\beta_2}(\mathbf{b}) = \nabla_{\mathbf{b}}\frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2^2 + \beta_2\mathbf{b}$ and $h(\mathbf{b}) \in \partial\|\mathbf{b}\|_1$ is any subgradient of $\|\mathbf{b}\|_1$. Intuitively, the training dynamics can be decomposed in two steps: the update $\mathbf{b}^{(t)}$ first moves near the least square solution $\hat{\mathbf{b}} := \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger \tilde{\mathbf{X}}^\top\mathbf{y}^*$ leading to memorization. Later in training, $h(\mathbf{b})$ dominates the update, leading to $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty \in \mathcal{O}(\alpha\beta_1)$ withing $\Theta(1/\alpha\beta_1)$ additional steps.

Theorem 2.1. *Assume* $\alpha < \frac{2}{\sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})+\beta_2}$ *and* $0 < \beta_1 \ll \frac{\sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})+\beta_2}{\sqrt{n}}$. *Then, there exists* $C > 0$ *and* $t_1 < \infty$ *such that* $\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_2 \leq \frac{2\alpha\beta_1 n^{1/2}}{1-\rho_2} \quad \forall t \geq t_1$ *and* $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2 \leq C\alpha\beta_1 n^{1/2} \iff t \geq t_1 + \Delta t$ *where* $\rho_2 := \sigma_{\max}\left(\mathbb{I}_n - \alpha\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\right)$ *and* $\Delta t = \Theta\left(\frac{\|\hat{\mathbf{b}}\|_\infty}{\alpha\beta_1}\right)$.

This result in valid for any ℓ_p norm ($p \in (0, \infty]$) such that $\rho_p := \left\|\mathbb{I}_n - \alpha\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\right\|_{p\to p} \in (0, 1)$, and under this condition $\|\cdot\|_2$ becomes $\|\cdot\|_p$ and $n^{1/2}$ becomes $n^{1/p}$. We also show that $f(\mathbf{b}^{(t)}) \to f(\mathbf{b}^*)$ and $\|\mathbf{b}^{(t)}\|_1 \to \|\mathbf{b}^*\|_1$ as $t \to \infty$ (Theorems C.3 and C.13). Note that when N is large enough, $\tilde{\mathbf{X}}\mathbf{b}^{(t)} = \mathbf{y}^*$ (memorization) and $\|\mathbf{b}^{(t)}\|_1 = \|\mathbf{b}^*\|_1$ are enough to conclude $\mathbf{b}^{(t)} = \mathbf{b}^*$ (generalization). In fact, after memorization, when $\|\mathbf{b}^{(t)}\|_1$ becomes too small, $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty \approx 0$ (Figure 2) since for problem of interest, the sparse solution $\mathbf{b}^*$ is the minimum ℓ_1 solution to $\|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2 \leq \epsilon$ under the sparsity constraint (section C). The smaller $\alpha\beta_1$ is, the longer it takes to recover $\mathbf{b}^*$, and the smaller is the error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty$ when $t \to \infty$ (Figures 1 and 12).

In addition to gradient descent, our results (Section C.7) extend to other iterative methods for ℓ_1 minimization, including the projected subgradient method (Section C.7) and for the proximal gradient descent method (Section C.8). Contrary to previous findings (Lyu et al., 2023; Liu et al., 2023a), we observe that in the over-parameterized regime ($N < n$), large-scale initialization and ℓ_2-regularization alone do not necessarily induce grokking (Section C.9), and instead lead to abrupt transitions in generalization error without converging to optimal solutions when sample sizes are insufficient. We term this effect "grokking without understanding", as highlighted in related work (Levi et al., 2024). Our analysis (Section C.10) demonstrates that coherence significantly impacts grokking in sparse recovery, with higher coherence delaying generalization by limiting the diversity of information captured by measurements. Furthermore, in deep linear networks (Section C.11), we find that depth $L \geq 2$ can implicitly promote sparsity and generalization, reducing the reliance on ℓ_1-regularization while mitigating generalization delays. Finally, in Section C.12, we extend these findings to realistic signals, including MNIST images, sinusoidal signals, and sparse polynomials.

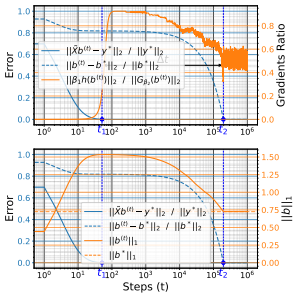


Figure 2: $G_{\beta_2}(\mathbf{b}^{(t)})$ dominates $\beta_1 h(\mathbf{b}^{(t)})$ until memorization at t_1; after which $\beta_1 h(\mathbf{b}^{(t)})$ dominates and make $\|\mathbf{b}^{(t)}\|_1$ converge to $\|\mathbf{b}^*\|_1$ at t_2, and so $\mathbf{b}^{(t_2)} = \mathbf{b}^*$.

For matrix factorization, given a low rank r matrix $\mathbf{A}^* \in \mathbb{R}^{n_1\times n_2}$, a measurement matrix $\mathbf{X} \in \mathbb{R}^{N\times n_1 n_2}$; and the measures $\mathbf{y}^* = \mathbf{X}\operatorname{vec}(\mathbf{A}^*) + \boldsymbol{\xi}$, and want to minimize $f(\mathbf{A}) = \frac{1}{2}\|\mathbf{X}\operatorname{vec}(\mathbf{A}) - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{A}\|_{\mathrm{F}} + \beta_*\|\mathbf{A}\|_*$ using gradient descent. The subgradient update rule is given by $\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \alpha\left(G_{\beta_2}(\mathbf{A}^{(t)}) + \beta_* h(\mathbf{A}^{(t)})\right)$ where $G_{\beta_2}(\mathbf{A}) = \nabla_{\mathbf{A}}\frac{1}{2}\|\mathbf{X}\operatorname{vec}\mathbf{A} - \mathbf{y}^*\|_2^2 + \beta_2\mathbf{A}$ and $h(\mathbf{A}) \in \partial\|\mathbf{A}\|_*$. Like in sparse recovery with gradient descent, the update $\mathbf{A}^{(t)}$ first moves near the least square solution $\operatorname{vec}(\hat{\mathbf{A}}) := \left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)^\dagger \mathbf{X}^\top\mathbf{y}^*$, and later in training, it converges to a solution with norm $\sigma_{\max}\left(\mathbf{A}^{(t)}\right) \in \mathcal{O}(\alpha\beta_*)$ (maximum singular value, i.e., operator norm).

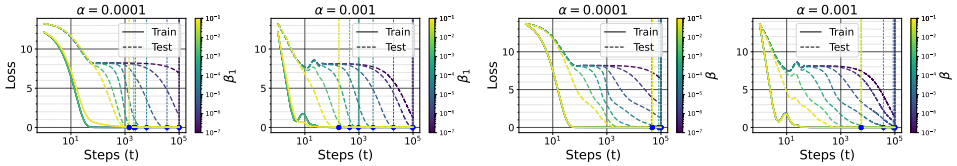
Theorem 2.2. *Assume* $\alpha < \frac{2}{\sigma_{\max}(\mathbf{X}^\top\mathbf{X})+\beta_2}$ *and* $0 < \beta_* \ll \frac{\sigma_{\max}(\mathbf{X}^\top\mathbf{X})+\beta_2}{\sqrt{\min(n_1,n_2)}}$. *For all* $p \in (0, \infty]$ *such that* $\rho_p := \left\|\mathbb{I}_n - \alpha\left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)\right\|_{p\to p} \in (0, 1)$, *there exists* $t_1 < \infty$; $\|\operatorname{vec}(\mathbf{A}^{(t)}) - \operatorname{vec}(\hat{\mathbf{A}})\|_p \leq \frac{2\alpha\beta_* n^{1/p}}{1-\rho_p} \quad \forall t \geq t_1$ *and* $\|\mathbf{A}^{(t)}\|_p \leq \alpha\beta_* n^{1/p} \iff t \geq t_2 := t_1 + \Delta t$ *with* $\Delta t = \Theta\left(\lfloor\frac{\sigma_{\max}(\hat{\mathbf{A}})}{\alpha\beta_*}\rfloor\right)$.

In particular, for $p = 2$, $\rho_2 \in (0, 1)$ since $0 < \alpha < \frac{2}{\sigma_{\max}(\mathbf{X}^\top\mathbf{X})+\beta_2}$. A choice of larger p means choosing the learning rate to have $\rho_p \in (0, \alpha_{\max})$. We also show that $f(\mathbf{A}^{(t)}) \to f(\mathbf{A}^*)$ and $\|\mathbf{A}^{(t)}\|_1 \to \|\mathbf{A}^*\|_1$ as $t \to \infty$ (Theorems D.4 and D.13). When N is large enough, $\mathbf{X}\operatorname{vec}\mathbf{A}^{(t)} = \mathbf{y}^*$ (memorization) and $\|\mathbf{A}^{(t)}\|_* = \|\mathbf{A}^*\|_*$ are enough to conclude $\mathbf{A}^{(t)} = \mathbf{A}^*$ (generalization). In fact, when $G_{\beta_2}(\mathbf{A})$ become negligeable compare to $\beta_* h(\mathbf{A})$, the singular values starts involving as $\sigma_i^{(t+1)} \approx |\sigma_i^{(t)} - \alpha|$ (Theorem D.12). This leads to a generalization through a multiscale singular

value decay phenomenon (Figure 4). The small singular value after memorization converges to $\{\sigma, 0 \leq \sigma < \alpha\beta_*\}$, followed by the next smaller one until the larger one. This process take time $\Theta\left(\lfloor \frac{\sigma_{\max}(\hat{\mathbf{A}})}{\alpha\beta_*} \rfloor\right)$. So, the smaller $\alpha\beta_*$, the longer it take to recover $\mathbf{A}^*$, and the smaller is the error $\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_\infty$ when $t \to \infty$. We also analyze the effect of coherence on grokking in matrix factorization. For matrix completion, given $\tau \in [0, 1]$, we select the first τN examples with the highest values of local coherence and select the remaining $(1 - \tau)N$ examples uniformly among the remaining. Unlike compressed sensing, where large values of τ are detrimental to generalization, here, as $\tau \to 1$, performance improves, and the number of examples required to generalize decreases exponentially, as does the time it takes the models to do so (Figures 45 and Figures 46).

3 Beyong Sparse Recovery and Low-Rank Matrix Factorization

In this section, we show that ℓ_1, ℓ_*, and domain-specific regularizers can replace ℓ_2 in a more general setting and reduce grokking. Let consider a teacher $\mathbf{y}^*(\mathbf{x}) = \mathbf{B}^* \max(\mathbf{A}^*\mathbf{x}, 0)$. We i.i.d sample N inputs output pair $\{(\mathbf{x}_i, \mathbf{y}^*(\mathbf{x}_i))\}_{i=1}^N$ and optimize the parameters $\theta = (\mathbf{A}, \mathbf{B})$ of a student $\mathbf{y}_\theta(\mathbf{x}) = \mathbf{B} \max(\mathbf{A}\mathbf{x}, 0)$ on them with the loss function $\hat{\mathcal{E}}(\theta) = \frac{1}{N}\sum_{i=1}^N \|\mathbf{y}_\theta(\mathbf{x}_i) - \mathbf{y}^*(\mathbf{x}_i)\|_2^2$ and different regularizer $\Omega_p(\theta)$ for $p \in \{1, 2, *\}$. For any $p \in \{1, 2, *\}$, the smaller is β_p and/or α, the longer is the delay between memorization and generalization (see Figures 3 for the training curve with ℓ_1, and 47, 48, 49 for more results with $\ell_{*/2}$).



(a) ℓ_1 Regularization (b) Sobolev training

Figure 3: Training and test error two layers `ReLU` teacher-student, for different values of the learning rate α and the ℓ_1 (resp. Sobolev) coefficient β_1. We can see that the smaller is α and or β_1, the longer is the delay between memorization and generalization.

Physics-Informed Neural Networks (PINNs) (Raissi et al., 2019) leverage prior knowledge from differential equations by incorporating their residuals into the loss function, ensuring that solutions remain consistent with physical laws. Sobolev training (Czarnecki et al., 2017) generalizes this idea by incorporating not only input-output pairs but also derivatives of the target function. We optimizer the student above by adding on the objective function the first order Sobolev penalty $\frac{\beta_1}{N}\sum_{i=1}^N \left\| \frac{\partial \mathbf{y}_\theta}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial \mathbf{y}^*}{\partial \mathbf{x}}(\mathbf{x}_i) \right\|_{\mathrm{F}}^2$, where the hyperparameter β_1 ensures that the model not only fits the data but also respects known smoothness constraints or differential structure. We observe that the smaller $\alpha\beta$, the longer the delay between memorization and generalization (See Figures 3 and 50).

We train a tree layers MLP and a LSMT on the addition modulo $p = 97$ problem (Power et al., 2022), and a two layers `ReLU` MLP trained on MNIST. We observe that ℓ_1 and ℓ_* have the same effect on grokking as ℓ_2, i.e., smaller regularization coefficient (and learning rate) delay generalization (more details in Sections E.3 and E.4).

4 Discussion and Conclusion

This work extends the understanding of grokking, showing that the transition from memorization to generalization can be induced not just by ℓ_2 regularization but also by sparsity or low-rank structure regularization or domain-specific regularization. These findings are particularly relevant in practice, where large-scale initialization is not always feasible, yet grokking still occurs. Our results highlight that in deep models, gradient descent implicitly drives the model towards solutions with sparse or low-rank properties, effectively mitigating overfitting (Arora et al., 2018). We also study the impact of data selection on grokking, and show that grokking can be amplified through only data selection.

REFERENCES

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization, 2018. URL https://arxiv.org/abs/1802.06509.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *CoRR*, abs/1905.13655, 2019. URL http://arxiv.org/abs/1905.13655.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *CoRR*, abs/1806.05451, 2018. URL http://arxiv.org/abs/1806.05451.

B. Barak, Benjamin L. Edelman, Surbhi Goel, S. Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2207.08799.

M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643, feb 1995. doi: 10.1088/0305-4470/28/3/018. URL https://dx.doi.org/10.1088/0305-4470/28/3/018.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 674–682, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/chenc14.html.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Sobolev training for neural networks, 2017. URL https://arxiv.org/abs/1706.04859.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, 2003. URL https://arxiv.org/abs/math/0307152.

David L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6): 797–829, 2006a. doi: https://doi.org/10.1002/cpa.20132. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20132.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ¡sup¿1¡/sup¿ minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003. doi: 10.1073/pnas.0437847100. URL https://www.pnas.org/doi/abs/10.1073/pnas.0437847100.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006b. doi: 10.1109/TIT.2006.871582.

Andreas Engel and Christian P. L. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, USA, 2001. ISBN 0521773075.

Gauthier Gidel, Francis R. Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in deep linear neural networks. *CoRR*, abs/1904.13262, 2019. URL http://arxiv.org/abs/1904.13262.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. *CoRR*, abs/1909.12051, 2019. URL http://arxiv.org/abs/1909.12051.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup*. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124010, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc61e. URL http://dx.doi.org/10.1088/1742-5468/abc61e.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation, 2020. URL https://arxiv.org/abs/1903.08560.

Noam Levi, Alon Beck, and Yohai Bar-Sinai. Grokking in linear estimators - a solvable model that groks without understanding. *International Conference on Learning Representations*, 2024.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *CoRR*, abs/2012.09839, 2020. URL https://arxiv.org/abs/2012.09839.

Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Ziming Liu, Ziqian Zhong, and Max Tegmark. Grokking as compression: A nonlinear complexity perspective. *arXiv preprint arXiv: 2310.05918*, 2023b.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, January 2021. ISSN 1095-7200. doi: 10.1137/19m1274067. URL http://dx.doi.org/10.1137/19M1274067.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. *arXiv preprint arXiv: 2311.18817*, 2023.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv: 2303.11873*, 2023.

Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale feature learning dynamics: Insights for double descent, 2021. URL https://arxiv.org/abs/2112.03215.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv: Arxiv-2201.02177*, 2022.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Holger Rauhut. *Compressive Sensing and Structured Random Matrices*, pp. 1–92. De Gruyter, Berlin, New York, 2010. ISBN 9783110226157. doi: doi:10.1515/9783110226157.1. URL https://doi.org/10.1515/9783110226157.1.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *CoRR*, abs/2005.06398, 2020. URL https://arxiv.org/abs/2005.06398.

David Saad and Sara Solla. Learning with noise and regularizers in multilayer neural networks. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52: 4225–4243, Oct 1995a. doi: 10.1103/PhysRevE.52.4225. URL https://link.aps.org/doi/10.1103/PhysRevE.52.4225.

David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995b. doi: 10.1103/PhysRevLett.74.4337. URL https://link.aps.org/doi/10.1103/PhysRevLett.74.4337.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv: Arxiv-2206.04817*, 2022.

Ryan Tibshirani. Machine learning 10-725/36-725, convex optimization: Spring 2015, lecture 8: February 9. https://www.stat.cmu.edu/~ryantibs/convexopt-S15/, 2015. Course lecture available online.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv: 2309.02390*, 2023.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery, 2019. URL https://arxiv.org/abs/1909.05122.

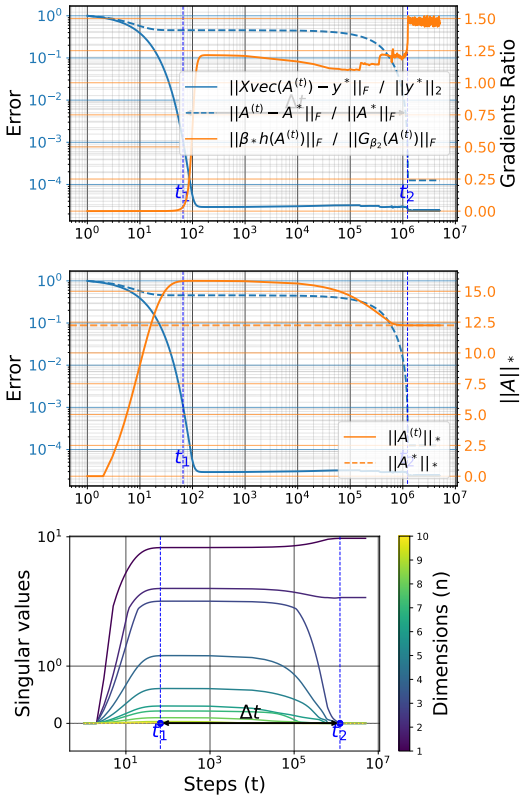


Figure 4: Relative errors, gradient ratio, the norm $\|\mathbf{A}^{(t)}\|_*$, and evolution of singular values. $G_{\beta_2}(\mathbf{A}^{(t)})$ dominates $\beta_* h(\mathbf{A}^{(t)})$ until memorization. From memorization $\beta_* h(\mathbf{A}^{(t)})$ dominates and make $\|\mathbf{A}^{(t)}\|_*$ converge to $\|\mathbf{A}^*\|_*$ at t_2, and so $\mathbf{A}^{(t_2)} = \mathbf{A}^*$. Generalization happened through a multiscale singular value decay phenomenon.

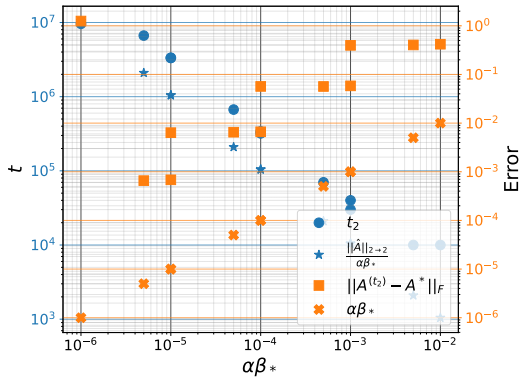


Figure 5: Generalization step t_2 (smaller t such that $\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_2/\|\mathbf{A}^*\|_{\mathrm{F}} \leq 10^{-4}$) and recovery error $\|\mathbf{A}^{(t_2)} - \mathbf{A}^*\|_2$ as a function of $\alpha\beta_*$ ($\log - \log$ plot). We can see that $t_2 \propto \|\hat{\mathbf{A}}\|_{2\to 2}/\alpha\beta_*$ and $\|\mathbf{A}^{(t_2)} - \mathbf{A}^*\|_{\mathrm{F}} \propto \alpha\beta_*$, i.e. small $\alpha\beta_*$ require longer time to converge, but do so at a lower generalization error. The outlier for very small $\alpha\beta_*$ is due to insufficient training (Figure 42).

A Related Works

Large initialization and ℓ_2 regularization Many studies in the linear teacher-student setup focus on ℓ_2 regularization, and the aim is generally to understand the classical generalization phenomenon like double descent (Hastie et al., 2020; Pezeshki et al., 2021), but not grokking. The only work on such models for grokking is Levi et al. (2024). They work on classification setting and show that the sharp increase in generalization accuracy may not imply a transition from "memorization" to "understanding" but can be an artifact of the accuracy measure. This aligns with the grokking without understanding the problem we observe in sparse recovery and low-rank matrix factorization. Our results are valid with many optimization methods for ℓ_1/ℓ_* minimization problems, such as subgradient, projected subgradient, and proximal gradient descent.

Grokking and stochasticity Our work also contradicts the hypothesis put forward when grokking was first observed, namely that grokking may be due to stochasticity or an anomaly in the optimization (Power et al., 2022; Thilak et al., 2022). Here, our algorithms are all deterministic (up to initialization).

Sparsity Barak et al. (2022) observed grokking on binary sparse parity problem, and Merrill et al. (2023) show that two subnetworks compete during training on such training, a dense (memorization) subnetwork, and a sparse (generalization) subnetwork. Since we can build a very sparse network that generalizes the sparse parity data Merrill et al. (2023), we claim that it is this sparsity that gives the models trained on this task their grokking nature.

Matrix completion To the best of our knowledge, we are the first to formally study grokking in the context of sparse recovery and low-rank matrix factorization (the shallow case). Lyu et al. (2023) show that low-rank matrix completion problems exhibit grokking with large initialization. But we prove that even on such a simple model, we do not need way decay and large initialization to observe grokking, but just $\ell_{1/*}$ regularization.

B Notations, Definitions, Preliminaries

We will optimize functions of the form $f(\theta) = \hat{\mathcal{E}}(\theta) + \beta\Omega(\theta)$, where $\hat{\mathcal{E}}$ is the square loss or cross-entropy loss function of the considered model on the training data, θ the set of model parameters, and Ω a regularizer applied to θ. It can be the standard ℓ_p norm or quasi-norm of θ, the sum of the nuclear norms of each matrix in θ (in this case, we call it ℓ_*), etc. For a vector $\mathbf{a} \in \mathbb{R}^n$, we consider the measurement operator $\mathcal{F}_{\mathbf{a}}(\mathbf{X}) = \mathbf{X}\mathbf{a} \in \mathbb{R}^N$ that take N measurement vectors $\{\mathbf{X}_i \in \mathbb{R}^n\}_{i\in[N]}$ a return the measures $\{\mathbf{X}_i^\top \mathbf{a}\}_{i\in[N]}$.

We work in $\mathbb{R}$ for compressed sensing and matrix completion, but many of our results extend easily to $\mathbb{C}$.

- We let $\mathbf{e}_k^{(n)} = [\mathbb{I}_n]_{:,k}$ be the k^{th} vector of the canonical basis of $\mathbb{R}^n$, $\mathbf{e}_{kl}^{(n)} = \delta_{kl} \forall l$. The subscript (n) will be omitted when the context will be clear
- $\odot$ is Hadamard product. For $\mathbf{Q} \in \mathbb{R}^{m\times n}$ and $\mathbf{R} \in \mathbb{R}^{m\times n}$, $(\mathbf{Q} \odot \mathbf{R})_{i,j} = \mathbf{Q}_{i,j}\mathbf{R}_{i,j}$ $(0 \leq i < m, 0 \leq j < p)$
- $\otimes$ is the Kronecker product. For $\mathbf{Q} \in \mathbb{R}^{m\times n}$ and $\mathbf{R} \in \mathbb{R}^{p\times q}$, $(\mathbf{Q}\otimes\mathbf{R})_{pr+v,qs+w} = \mathbf{Q}_{rs}\mathbf{R}_{vw}$ $(0 \leq r < m, 0 \leq v < p, 0 \leq s < n$ and $0 \leq w < q)$
- $\circ$ is the outer product, $\left(\mathbf{a}^{(1)} \circ \cdots \circ \mathbf{a}^{(n)}\right)_{i_1,\cdots,i_n} = \mathbf{a}_{i_1}^{(1)} \cdots \mathbf{a}_{i_n}^{(n)} \quad \forall (i_1, \cdots, i_n) \in [m_1] \times \cdots \times [m_n]$ for n vectors $\mathbf{a}^{(i)} \in \mathbb{R}^{m_i} \quad \forall i \in [n]$.
- $\sigma_{\max\,/\,\min}(\mathbf{A}) = \sqrt{\lambda_{\max\,/\,\min}(\mathbf{A}^\top\mathbf{A})}$ is the maximum (resp. minimum) singular value of a matrix $\mathbf{A}$, with $\lambda_{\max\,/\,\min}$ the corresponding eigenvalue
- For a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_0 = |\{i \in [n], \mathbf{x}_i \neq 0\}|$, $\|\mathbf{x}\|_p = (\sum_{i=1}^n |\mathbf{x}_i|^p)^{\frac{1}{p}}\ \forall p \in (0,\infty)$ and $\|\mathbf{x}\|_\infty = \max_{i\in[n]} |\mathbf{x}_i|$.
  We have $\frac{1}{\sqrt{n}}\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty$.
- For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the schatten p-norm of $\mathbf{A}$ is $\|\mathbf{A}\|_p = (\sum_i \sigma_i(\mathbf{A})^p)^{1/p}$. For $p = 1$, this gives the trace/nuclear norm $\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A}) = \operatorname{tr}\left(\sqrt{\mathbf{A}^\top\mathbf{A}}\right)$. The induce $p \to q$ norm of $\mathbf{A}$ is $\|\mathbf{A}\|_{p\to q} = \sup_{x\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_q}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_q$. We have $\|\mathbf{A}\|_{1\to 1} = \max_{j\in[n]} \sum_{i=1}^m |\mathbf{A}_{ij}|$ (maximum absolute column sum), $\|\mathbf{A}\|_{2\to 2} = \|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$ (operator norm, spectral norm, induced 2-norm) and $\|\mathbf{A}\|_{\infty\to\infty} = \max_{i\in[m]} \sum_{j=1}^n |\mathbf{A}_{ij}|$ (maximum absolute row sum).

$$\begin{aligned} \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{2\to 2} \leq \|\mathbf{A}\|_{1\to 1} \leq \sqrt{m}\|\mathbf{A}\|_{2\to 2} \\ \frac{1}{\sqrt{m}}\|\mathbf{A}\|_{2\to 2} \leq \|\mathbf{A}\|_{\infty\to\infty} \leq \sqrt{n}\|\mathbf{A}\|_{2\to 2} \end{aligned} \tag{1}$$

Definition B.1 (Khatri-Rao and Face-splitting products)**.** For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times n}$, the Khatri-Rao product $\mathbf{A} \star \mathbf{B} \in \mathbb{R}^{mp\times n}$ contains in each column $i \in [n]$ the matrix $\mathbf{A}_{:,i} \otimes \mathbf{B}_{:,i}$. We have the formula $\mathbf{A} \star \mathbf{B} = (\mathbf{A} \otimes 1_p) \odot (1_m \otimes \mathbf{B})$.

For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times p}$, the face-splitting product $\mathbf{A} \bullet \mathbf{B} \in \mathbb{R}^{m\times np}$ contains in each row $i \in [m]$ the matrix $\mathbf{A}_{i,:} \otimes \mathbf{B}_{i,:}$. It can be seen as the row-wise Khatri-Rao product, and we have $(\mathbf{A} \bullet \mathbf{B}) = (\mathbf{A}^\top \star \mathbf{B}^\top)^\top = \left(\mathbf{A} \otimes 1_p^\top\right) \odot (1_n^\top \otimes \mathbf{B})$.

We will generalize this operator in a higher number of vectors. If we have N vectors $A^{(k)} \in \mathbb{R}^{m\times n_k}$, then $\left(\mathbf{A}^{(1)} \bullet \mathbf{A}^{(2)} \bullet \cdots \bullet \mathbf{A}^{(N)}\right)_{i,:} = \mathbf{A}_{i,:}^{(1)} \otimes \mathbf{A}_{i,:}^{(3)} \otimes \cdots \otimes \mathbf{A}_{i,:}^{(N)} \in \mathbb{R}^{\prod_k n_k}$.

Definition B.2. A matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ can be vectorized column-wise, $\texttt{vecc}(\mathbf{M})_{in+j} = \mathbf{M}_{ij}$, or row-wise $\texttt{vecr}(\mathbf{M})_{jm+i} = \mathbf{M}_{ij}$, where $0 \leq i \leq m-1$ and $0 \leq j \leq n-1$. So $\texttt{vecc}(\mathbf{M}) = \texttt{vec}(\mathbf{M})$ and $\texttt{vecr}(\mathbf{M}) = \texttt{vec}(\mathbf{M}^\top) = \mathbb{K}_{(m,n)}\texttt{vec}(\mathbf{M})$ with $\texttt{vec}(\mathbf{M})$ the vanilla vectorization, which stack the column of $\mathbf{M}$ in a vector.

Definition B.3. A tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{m\times n\times p}$ can be vectorized column-wise, $\texttt{vecc}(\boldsymbol{\mathcal{T}})_{kmn+jm+i} = \boldsymbol{\mathcal{T}}_{ijk}$, or row-wise $\texttt{vecr}(\boldsymbol{\mathcal{T}})_{inp+jm+k} = \boldsymbol{\mathcal{T}}_{ijk}$, where $0 \leq i \leq m-1$ and $0 \leq j \leq n-1$ and $0 \leq k \leq p-1$. Note that $\boldsymbol{\mathcal{T}}$ can be vectorized in 3! ways[1].

Let $\boldsymbol{\mathcal{T}}^{(12)} = \boldsymbol{\mathcal{T}}_{(1)} \in \mathbb{R}^{m\times np}$ (mode-1 unfolding of $\boldsymbol{\mathcal{T}}$), $\boldsymbol{\mathcal{T}}^{(21)} = \boldsymbol{\mathcal{T}}_{(2)} \in \mathbb{R}^{n\times mp}$ and $\boldsymbol{\mathcal{T}}^{(32)} = \boldsymbol{\mathcal{T}}_{(3)}^\top \in \mathbb{R}^{mn\times p}$. That is

$$\boldsymbol{\mathcal{T}}^{(32)} := \begin{bmatrix} | & | & & | \\ \texttt{vecc}(\boldsymbol{\mathcal{T}}_{::1}) & \texttt{vecc}(\boldsymbol{\mathcal{T}}_{::2}) & \cdots & \texttt{vecc}(\boldsymbol{\mathcal{T}}_{::p}) \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{mn\times p}$$

[1] A tensor of order K can be vectorized in $K!$ ways.

and

$$\boldsymbol{\mathcal{T}}^{(12)} := \begin{bmatrix} -\texttt{vecr}(\boldsymbol{\mathcal{T}}_1)- \\ \cdots \\ -\texttt{vecr}(\boldsymbol{\mathcal{T}}_p)- \end{bmatrix} \in \mathbb{R}^{m \times pn}$$

We have

$$\texttt{vecc}(\boldsymbol{\mathcal{T}}) = \begin{bmatrix} | \\ \texttt{vecc}(\boldsymbol{\mathcal{T}}_{::1}) \\ | \\ \cdot \\ \cdot \\ \cdot \\ | \\ \texttt{vecc}(\boldsymbol{\mathcal{T}}_{::p}) \\ | \end{bmatrix} = \texttt{vecc}\left(\boldsymbol{\mathcal{T}}^{(32)}\right) := \boldsymbol{\mathcal{T}}^{(321)}$$

and

$$\texttt{vecr}(\boldsymbol{\mathcal{T}}) = \begin{bmatrix} | \\ \texttt{vecr}(\boldsymbol{\mathcal{T}}_1) \\ | \\ \cdot \\ \cdot \\ \cdot \\ | \\ \texttt{vecr}(\boldsymbol{\mathcal{T}}_n) \\ | \end{bmatrix} = \texttt{vecr}\left(\boldsymbol{\mathcal{T}}^{(12)}\right) := \boldsymbol{\mathcal{T}}^{(123)}$$

For $\mathbf{A} \in \mathbb{R}^{q \times p}$ and $\mathbf{B} \in \mathbb{R}^{q \times m}$, $\mathbf{A}\boldsymbol{\mathcal{T}}^{(32)} = \mathbf{A}\boldsymbol{\mathcal{T}}_{(3)} = (\boldsymbol{\mathcal{T}} \times_3 \mathbf{A})_{(3)}$ and $\mathbf{B}\boldsymbol{\mathcal{T}}^{(12)} = \mathbf{B}\boldsymbol{\mathcal{T}}_{(1)} = (\boldsymbol{\mathcal{T}} \times_1 \mathbf{B})_{(1)}$.

If we CP-decompose $\boldsymbol{\mathcal{T}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{i=1}^{R} \mathbf{A}_{:,i} \circ \mathbf{B}_{:,i} \circ \mathbf{C}_{:,i}$, with $\mathbf{A} \in \mathbb{R}^{m \times R}$, $\mathbf{B} \in \mathbb{R}^{n \times R}$ and $\mathbf{C} \in \mathbb{R}^{p \times R}$ the three mode loading matrices, then $\boldsymbol{\mathcal{T}}_{(1)} = \mathbf{A}(\mathbf{C} \star \mathbf{B})^\top$, $\boldsymbol{\mathcal{T}}_{(2)} = \mathbf{B}(\mathbf{A} \star \mathbf{C})^\top$ and $\boldsymbol{\mathcal{T}}_{(3)} = \mathbf{C}(\mathbf{B} \star \mathbf{A})^\top$.

C Sparse Recovery

C.1 Definitions and Preliminaries

Definition C.1 (Restricted Isometry Property (RIP) and Restricted Isometric Constant(RIC))**.** Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $(s, \delta_s) \in [n] \times (0, 1)$. The matrix $\mathbf{A}$ is said to satisfy the (s, δ_s)-RIP if

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \tag{2}$$

for all s-sparse vector $\mathbf{x} \in \mathbb{R}^n$ (ie $\|\mathbf{x}\|_0 \leq s$). This is equivalent to saying that for every $J \subset [n]$ with $|J| = s$

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}_{:,J}\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \tag{3}$$

for every $\mathbf{x} \in \mathbb{R}^s$; where the submatrix $\mathbf{A}_{:,J} \in \mathbb{R}^{m \times s}$ of $\mathbf{A}$ is build by selecting the columns index in J. This condition is also equivalent to the statement $\|\mathbf{A}_{:,J}^\top \mathbf{A}_{:,J} - \mathbb{I}_s\|_{2 \to 2} \leq \delta_s$, which is finally equivalent to $\texttt{Spec}\left(\mathbf{A}_{:,J}^\top \mathbf{A}_{:,J}\right) \subset [1 - \delta_s, 1 + \delta_s]$.

We say that $\mathbf{A}$ satisfies s-RIP if it satisfies (s, δ_s)-RIP with some $\delta_s \in (0, 1)$. The s-RIC of $\mathbf{A}$ is defined as the infimum $\delta_s(\mathbf{A})$ of all possible δ_s such that $\mathbf{A} \in \mathbb{R}^{m \times n}$ satisfy the (s, δ_s)-RIP.

$$\begin{aligned}
\delta_s(\mathbf{A}) &= \inf\left\{\delta_s \in (0,1) \quad | \quad (1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq s\right\} \\
&= \inf\left\{\delta_s \in (0,1) \quad | \quad (1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}_{:,J}\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^s, J \subset [n], |J| = s\right\} \\
&= \inf\left\{\delta_s \in (0,1) \quad | \quad \|\mathbf{A}_{:,J}^\top \mathbf{A}_{:,J} - \mathbb{I}_s\|_{2 \to 2} \leq \delta_s \quad \forall J \subset [n], |J| = s\right\} \\
&= \inf\left\{\delta_s \in (0,1) \quad | \quad \texttt{Spec}\left(\mathbf{A}_{:,J}^\top \mathbf{A}_{:,J}\right) \subset [1 - \delta_s, 1 + \delta_s] \quad \forall J \subset [n], |J| = s\right\}
\end{aligned}$$

So, for all $\forall J \subset [n]$ with $|J| = s$, the condition number of $\mathbf{A}_{:,J}^\top \mathbf{A}_{:,J}$ is bounds from above by $\frac{1+\delta_s(\mathbf{A})}{1-\delta_s(\mathbf{A})}$, a the one of $\mathbf{A}_{:,J}$ by $\sqrt{\frac{1+\delta_s(\mathbf{A})}{1-\delta_s(\mathbf{A})}}$.

We say that a matrix $\mathbf{A}$ satisfies the RIP if $\delta_s(\mathbf{A})$ is small for reasonably large s. All the above definitions extend to any linear map $f : \mathbb{R}^n \to \mathbb{R}^m$.

Proposition C.1. $\delta_s(\mathbf{A}) \leq \delta_{s+1}(\mathbf{A})$ for all $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $s \in [n]$.

Definition C.2 (Restricted Isometry Property)**.** Let $\mathcal{F} : \mathbb{R}^{m\times n} \to \mathbb{R}^q$ be a linear map and $(r, \delta_r) \in [n] \times (0, 1)$. f is said to satisfy (r, δ_r)-RIP if for all rank-r matrices $\mathbf{X} \in \mathbb{R}^{m\times n}$:

$$(1 - \delta_r)\|\mathbf{X}\|_{\mathrm{F}}^2 \leq \|\mathcal{F}(\mathbf{X})\|_2^2 \leq (1 + \delta_r)\|\mathbf{X}\|_{\mathrm{F}}^2 \tag{4}$$

We say that $\mathcal{F}$ satisfies r-RIP if $\mathcal{F}$ satisfies (r, δ_r)-RIP with some $\delta_r \in (0, 1)$, and the r-RIC of $\mathcal{F}$ is defined as the infimum $\delta_r(\mathcal{F})$ of all possible δ_r such that $\mathcal{F}$ satisfy the (r, δ_r)-RIP.

Definition C.3 (Coherence)**.** The coherence between two matrices $\mathbf{A} \in \mathbb{R}^{q\times m}$ and $\mathbf{B} \in \mathbb{R}^{q\times n}$ is

$$\mu(\mathbf{A}, \mathbf{B}) = \max_{i\in[m],j\in[n]} \frac{|\langle \mathbf{A}_{:,i}, \mathbf{B}_{:,j}\rangle|}{\|\mathbf{A}_{:,i}\|\|\mathbf{B}_{:,j}\|} = \max_{i\in[m],j\in[n]} \frac{|[\mathbf{A}^\top \mathbf{B}]_{i,j}|}{\|\mathbf{A}_{:,i}\|\|\mathbf{B}_{:,j}\|} \tag{5}$$

Coherence measures how similar or aligned two matrices or vectors are. Specifically, it measures how much overlap there is between the columns of $\mathbf{A}$ and $\mathbf{B}$. High coherence means they are similar or aligned, and low coherence (or incoherence) means they are very different. Incoherence is essentially the opposite of coherence. It refers to a low overlap or low similarity between the columns of $\mathbf{A}$ and $\mathbf{B}$.

The mutual coherence of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is

$$\mu(\mathbf{A}) = \max_{(i,j)\in[m]\times[n],i\neq j} \frac{|\langle \mathbf{A}_{:,i}, \mathbf{A}_{:,j}\rangle|}{\|\mathbf{A}_{:,i}\|\|\mathbf{A}_{:,j}\|} = \max_{(i,j)\in[m]\times[n],i\neq j} \frac{[\mathbf{A}^\top \mathbf{A}]_{i,j}}{\|\mathbf{A}_{:,i}\|\|\mathbf{A}_{:,j}\|} \tag{6}$$

If the coherence is small, then the columns of $\mathbf{A}$ are almost mutually orthogonal. A small coherence is desired in order to have good sparse recovery properties.

We also have the 1-coherence

$$\mu_1(\mathbf{A}, s) = \max_{i\in[n]} \max_{J\subseteq[n]\setminus i,|J|\leq s} \sum_{j\in J} \frac{|\langle \mathbf{A}_{:,i}, \mathbf{A}_{:,j}\rangle|}{\|\mathbf{A}_{:,i}\|\|\mathbf{A}_{:,j}\|} \leq s\mu(\mathbf{A})$$

Example C.1. *For the Fourier basis $\sqrt{n}\Phi_{ji} = \mathbf{e}^{-2\pi \boldsymbol{i}\frac{ji}{n}}$, we have $\mu_1(\Phi, s) = s\mu(\Phi) = s/\sqrt{n}$ (Rauhut, 2010). Each column in this basis vector corresponds to a specific frequency. For a signal $\mathbf{a}^*$, if only a few frequency components contribute significantly to $\mathbf{a}^*$, then $\mathbf{b}^* = \Phi^{-1}\mathbf{a}^*$, the Fourier transform of $\mathbf{a}^*$, will be sparse. This Φ is unitary, and its inverse is $\sqrt{n}\Phi_{ji}^{-1} = \mathbf{e}^{2\pi \boldsymbol{i}\frac{ji}{n}}$.*

Proposition C.2. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with unit norm columns, $\mu(\mathbf{A}) \geq \sqrt{\frac{n-m}{m(n-1)}}$ and $\mu_1(\mathbf{A}, s) \geq s\sqrt{\frac{n-m}{m(n-1)}}$ whenever $s \leq \sqrt{n-1}$ (Rauhut, 2010).

Proposition C.3. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with unit norm columns, $\mu(\mathbf{A}) = \delta_2(\mathbf{A})$, $\mu_1(\mathbf{A}, s) = \max_{J\in[n],|J|\leq s+1} \|\mathbf{A}_{:,J}^\top \mathbf{A}_{:,J} - \mathbb{I}\|_{1\to 1}$, and $\delta_s(\mathbf{A}) \leq \mu_1(\mathbf{A}, s-1) \leq (s-1)\mu(\mathbf{A})$ (Rauhut, 2010).

Proposition C.4 (Connexion between the coherence $\mu(\mathbf{A}, \mathbf{B})$ and $\delta_s(\mathbf{A}^\top \mathbf{B})$)**.** Let $\mathbf{A} \in \mathbb{R}^{q\times m}$, $\mathbf{B} \in \mathbb{R}^{q\times n}$ and $\mathbf{M} = \mathbf{A}^\top \mathbf{B} \in \mathbb{R}^{m\times n}$. We have

$$\max\left(\frac{1}{m\sqrt{s}}, \frac{1}{s\sqrt{m}}\right)\|\mathbf{M}_{:,J}\|_{2\to 2} \leq \mu(\mathbf{A}, \mathbf{B}) \leq \min\left(\sqrt{m}, \sqrt{n}\right)\|\mathbf{M}\|_{2\to 2} \quad \forall J \subset [n],\ |J| = s$$

and

$$\sqrt{1 - \delta_s(\mathbf{A}^\top \mathbf{B})} \leq \frac{m+s}{2}\mu(\mathbf{A}, \mathbf{B}) \tag{7}$$

Proof. For $J \subset [n]$ with $|J| = s$, we have $\mathbf{M}_{:,J} = \mathbf{A}^\top \mathbf{B}_{:,J} \in \mathbb{R}^{m\times s}$ and $\mathtt{Spec}\left(\mathbf{M}_{:,J}^\top \mathbf{M}_{:,J}\right) \subset [1 - \delta_s, 1 + \delta_s]$. This implies $\|\mathbf{M}_{:,J}\|_{2\to 2}^2 = \lambda_{\max}(\mathbf{M}_{:,J}^\top \mathbf{M}_{:,J}) \in [1 - \delta_s, 1 + \delta_s]$.

Also,

$$\|\mathbf{M}_{:,J}\|_{1\to 1} = \max_{j\in[s]} \sum_{i=1}^{m} |[\mathbf{M}_{:,J}]_{ij}| \leq m \max_{j\in[s]} \max_{i\in[m]} |[\mathbf{M}_{:,J}]_{ij}| = m\mu(\mathbf{A}, \mathbf{B}_{:,J}) \leq m\mu(\mathbf{A}, \mathbf{B})$$

$$\mu(\mathbf{A}, \mathbf{B}) = \max_{i\in[m],j\in[n]} |\mathbf{M}_{i,j}| \leq \max_{i\in[m],j\in[n]} \sum_{k=1}^{m} |\mathbf{M}_{k,j}| = \max_{j\in[n]} \sum_{k=1}^{m} |\mathbf{M}_{k,j}| = \|\mathbf{M}\|_{1\to 1}$$

and

$$\|\mathbf{M}_{:,J}\|_{\infty\to\infty} = \max_{i\in[m]} \sum_{j=1}^{s} |[\mathbf{M}_{:,J}]_{ij}| \leq s \max_{i\in[m]} \max_{j\in[s]} |[\mathbf{M}_{:,J}]_{ij}| = s\mu(\mathbf{A}, \mathbf{B}_{:,J}) \leq s\mu(\mathbf{A}, \mathbf{B})$$

$$\mu(\mathbf{A}, \mathbf{B}) = \max_{i\in[m],j\in[n]} |\mathbf{M}_{i,j}| \leq \max_{i\in[m],j\in[n]} \sum_{k=1}^{n} |\mathbf{M}_{i,k}| = \max_{i\in[m]} \sum_{k=1}^{n} |\mathbf{M}_{i,k}| = \|\mathbf{M}\|_{\infty\to\infty}$$

So

$$\max\left(\frac{\|\mathbf{M}_{:,J}\|_{1\to 1}}{m}, \frac{\|\mathbf{M}_{:,J}\|_{\infty\to\infty}}{s}\right) \leq \mu(\mathbf{A}, \mathbf{B}) \leq \min\left(\|\mathbf{M}\|_{1\to 1}, \|\mathbf{M}\|_{\infty\to\infty}\right) \tag{8}$$

For $\mathbf{C} \in \mathbb{R}^{m\times n}$,

$$\begin{aligned} \frac{1}{\sqrt{n}}\|\mathbf{C}\|_{2\to 2} \leq \|\mathbf{C}\|_{1\to 1} \leq \sqrt{m}\|\mathbf{C}\|_{2\to 2} \\ \frac{1}{\sqrt{m}}\|\mathbf{C}\|_{2\to 2} \leq \|\mathbf{C}\|_{\infty\to\infty} \leq \sqrt{n}\|\mathbf{C}\|_{2\to 2} \end{aligned} \tag{9}$$

Using 8 and 9, we obtain

$$\max\left(\frac{1}{m\sqrt{s}}, \frac{1}{s\sqrt{m}}\right) \|\mathbf{M}_{:,J}\|_{2\to 2} \leq \mu(\mathbf{A}, \mathbf{B}) \leq \min\left(\sqrt{m}, \sqrt{n}\right) \|\mathbf{M}\|_{2\to 2}$$

Combining with $\|\mathbf{M}_{:,J}\|_{2\to 2}^2 = \lambda_{\max}(\mathbf{M}_{:,J}^\top \mathbf{M}_{:,J}) \in [1-\delta_s, 1+\delta_s]$ give

$$\frac{\sqrt{\max(m,s)}}{ms}\sqrt{1-\delta_s(\mathbf{A}^\top\mathbf{B})} \leq \mu(\mathbf{A}, \mathbf{B}) \leq \sqrt{\min(m,n)}\sqrt{1+\delta_n(\mathbf{A}^\top\mathbf{B})} \tag{10}$$

Since $\|\mathbf{M}_{:,J}\|_{2\to 2} \leq \max\left(\|\mathbf{M}_{:,J}\|_{1\to 1}, \|\mathbf{M}_{:,J}\|_{\infty\to\infty}\right)$ (Rauhut, 2010), we also have

$$\sqrt{1-\delta_s(\mathbf{A}^\top\mathbf{B})} \leq \frac{m+s}{2}\mu(\mathbf{A}, \mathbf{B}) \tag{11}$$

□

C.2 The Problem

Compressed sensing theory predicts that sparse signals in high dimensions can be recovered from undersampled linear measurements. More precisely, given $N \ll n$ noisy measurements $\mathbf{y}^* = \mathcal{F}_{\mathbf{a}^*}(\mathbf{X}) + \boldsymbol{\xi} \in \mathbb{R}^N$ of a vector $\mathbf{a}^* \in \mathbb{R}^n$ (digital signal, image, etc.), we look for a reconstruction $\mathbf{a} \in \mathbb{R}^n$ that minimizes $\|\mathcal{F}_{\mathbf{a}}(\mathbf{X}) - \mathbf{y}^*\|_2$; where $\mathcal{F}_{\mathbf{a}}(\mathbf{X}) = \mathbf{X}\mathbf{a} \in \mathbb{R}^N$ is the measurement operator that take N measurement vectors $\{\mathbf{X}_i \in \mathbb{R}^n\}_{i\in[N]}$ a return the measures $\{\mathbf{X}_i^\top \mathbf{a}\}_{i\in[N]}$. Without further knowledge, this is impossible for $N < n$. This is why the sparsity of the original signal $\mathbf{a}^*$ is assumed, i.e., we can write $\mathbf{a}^* = \sum_{i=1}^n \mathbf{b}_i^* \Phi_{:,i} = \Phi\mathbf{b}^*$ with $s = \|\mathbf{b}^*\|_0 := |\{i, \mathbf{b}_i^* \neq 0\}| \ll n$, and $\Phi \in \mathbb{R}^{n\times n}$ a dictionary (see example C.1 for the Fourier transform). We assume for simplicity that Φ is an orthonormal matrix, $\Phi^\top\Phi = \mathbb{I}_n$ (Assumption C.3). In sparse coding, we aim to find $\mathbf{a} = \Phi\mathbf{b}$ under the constraint that $\|\mathbf{b}\|_0 \ll n$. This can be stated as

$$(P_0) \text{ Minimize } \|\mathbf{b}\|_0 \text{ s.t. } \|\mathcal{F}_{\Phi\mathbf{b}}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon \tag{12}$$

with ϵ an upper bound on the size of the error term $\boldsymbol{\xi} \in \mathbb{R}^N$, $\|\boldsymbol{\xi}\|_2 \leq \epsilon$. This problem is NP-hard, and the constraint $\|\mathbf{b}\|_0$ is often relaxed to an ℓ_1 regularization, and leading to the convex problem

$$(P_1) \text{ Minimize } \|\mathbf{b}\|_1 \text{ s.t. } \|\mathcal{F}_{\Phi\mathbf{b}}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon \tag{13}$$

This problem has been well studied in the signal processing literature under the name Basis Pursuit. It is well known that under certain conditions on the measurement matrix $\mathbf{X}$ (e.g., coherence with respect to Φ) and the sparsity of $\mathbf{a}^*$ in Φ, sufficiently sparse solutions of (P_1) are also solutions of (P_0) (Donoho & Elad, 2003; Candes et al., 2006). Many lower bounds on the number of measures N guaranteeing $\|\mathbf{b} - \mathbf{b}^*\|_2 \leq \epsilon$ with high probability have also been derived. Such lower bounds generally have the form $N = \Omega\left(\delta^{-\beta}\left(s\log^{\alpha}(n/s) + \log 1/\eta\right)\right)$(Rauhut, 2010), where δ capture the Restricted Isometry Property (RIP, Definition C.1) of $\tilde{\mathbf{X}} = \mathbf{X}\Phi$ and is also related to the coherence (Definition C.3) of $\mathbf{X}$ with respect to Φ (Proposition C.3), η is the percentage of error (i.e. N guaranteed a recovery with probability at least $1-\eta$), $\alpha > 0$ and $\beta > 0$ are constants. Observe that in the noiseless setting, we want $\mathbf{b}$ such that $\tilde{\mathbf{X}}\mathbf{b} = \tilde{\mathbf{X}}\mathbf{b}^*$, that is $\mathbf{b} \in \mathbf{b}^* + \text{Null}(\tilde{\mathbf{X}})$. Donoho (2006a;b) show that the nullspace $\tilde{\mathbf{X}}\mathbf{b} = 0$ has a very special structure for certain $\tilde{\mathbf{X}}$ (e.g. incoherent with any orthonormal basis): when $\mathbf{b}^*$ is sparse, the only element in the affine subspace $\mathbf{b}^* + \text{Null}(\tilde{\mathbf{X}})$ that can have a small ℓ_1 norm is $\mathbf{b}^*$ itself.

Given the measures $\mathbf{y}^* \in \mathbb{R}^N$ (possibly noisy), the measurement matrix $\mathbf{X} \in \mathbb{R}^{N\times n}$, and the sparse basis (or dictionary) $\Phi \in \mathbb{R}^{n\times n}$, we aim to solve the following problem

$$(P_0)\ \text{Minimize}\ \|\mathbf{b}\|_0\ \text{s.t.}\ \|\mathcal{F}_{\Phi\mathbf{b}}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon \tag{14}$$

and more precisely, its convex relaxation

$$(P_1)\ \text{Minimize}\ \|\mathbf{b}\|_1\ \text{s.t.}\ \|\mathcal{F}_{\Phi\mathbf{b}}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon \tag{15}$$

C.3 Assumption on the sparse basis

We will assume for simplicity that Φ is an orthonormal matrix, $\Phi^\top\Phi = \mathbb{I}_n$. It is common in sparse coding theory to consider $\Phi \in \mathbb{R}^{n\times m}$ as a dictionary with m columns referred to as atoms: and saying $\mathbf{a}^*$ is sparse means it can be written as a linear combination of a few of such atoms. But here, we assume for simplicity that we have $\mathbf{a}^* = \Phi\mathbf{b}^*$ with $\mathbf{b}^* \in \mathbb{R}^m$ and $\Phi \in \mathbb{R}^{n\times m}$ a set of $m \leq n$ linearly independent vectors (its column). Let $\Phi^\perp \in \mathbb{R}^{n\times(n-m)}$ be the orthogonal complement of Φ in $\mathbb{R}^n$, $\Psi := \begin{bmatrix}\Phi & \Phi^\perp\end{bmatrix} \in \mathbb{R}^{n\times n}$, $\tilde{\Phi} := \Psi\left(\Psi^\top\Psi\right)^{-1/2}$ the orthonormal version of Ψ, and $\tilde{\mathbf{b}}^* := \left(\Psi^\top\Psi\right)^{1/2}\begin{bmatrix}\mathbf{b}^* \\ 0\end{bmatrix}$. We have $\mathbf{a}^* = \tilde{\Phi}\tilde{\mathbf{b}}^*$, with $\|\tilde{\mathbf{b}}^*\|_0 = \|\mathbf{b}^*\|_0$ since $\Psi^\top\Psi$ is diagonal. So, assuming Φ orthonormal is without loss of generality.

C.4 The controls parameters

The incoherence between the measurement vectors (line of $\mathbf{X}$) and the sparse basis (column of Φ) is crucial for successfully recovering $\mathbf{a}^*$ (or equivalently $\mathbf{b}^*$, the sparse representation). If $\mathbf{X}$ is incoherent with Φ, each measurement captures a distinct "view" of $\mathbf{a}^*$, reducing redundancy. This diversity of information allows for the successful reconstruction of $\mathbf{b}^*$ even with fewer measurements (e.g., below the Nyquist rate for signals). Achieving low coherence (high incoherence) can be done by designing $\mathbf{X}$ to be a random matrix (e.g., Sub-Gaussian like Gaussian or Bernoulli matrices). Such random matrices are, with high probability, incoherent with any fixed orthonormal basis (Theorems C.1 and C.2).

Theorem C.1. *Le $m \leq n$ and $\Phi \in \mathbb{R}^{n\times m}$ with $\Phi^\top\Phi = \mathbb{I}_m$. For any $N \geq 1$, $\alpha > 0$ and $\beta > 1$; the matrix $\mathbf{X} \in \mathbb{R}^{N\times n}$ with $n^\alpha\mathbf{X}_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ satisfies $\mu(\mathbf{X}^\top, \Phi) \leq 2\beta\frac{\sqrt{\ln(nN)}}{n^\alpha}$ with probability at least $1 - 1/(nN)^{2\beta^2-1}$.*

Proof. Let $\sigma = n^{-\alpha}$. We have $\mathbf{X} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$, so $[\mathbf{X}\Phi]_{ij} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$ since Φ has normal columns. This implies $\mathbb{P}\left[\left|[\mathbf{X}\Phi]_{ij}\right| \geq t\right] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$, which in turn implies $\mathbb{P}\left[\max_{i,j}\left|[\mathbf{X}\Phi]_{ij}\right| \geq t\right] \leq \sum_{i,j}\mathbb{P}\left[\left|[\mathbf{X}\Phi]_{ij}\right| \geq t\right] \leq nN\exp\left(-\frac{t^2}{2\sigma^2}\right)$. Using $t = 2\beta\frac{\sqrt{\ln(nN)}}{n^\alpha}$ with $\beta > 1$, we have $t^2 = 2\left(\frac{1}{n^\alpha}\right)^2\ln\left(\frac{nN}{\eta}\right)$ with $\eta = (nN)^{1-2\beta^2}$, so $nN\exp\left(-\frac{t^2}{2\sigma^2}\right) = \eta$. □

We also have the following theorem from Rauhut (2010) about the RIP of such a matrix.

Theorem C.2. *Let $\mathbf{X} \in \mathbb{R}^{N\times n}$ be a Gaussian or Bernoulli random matrix. Let $\eta, \delta \in (0,1)$ and assume $N \geq C\delta^{-2}\left(s\ln(n/s) + \ln(1/\eta)\right)$ for a universal constant $C > 0$. Then, $\delta_s(\mathbf{X}) \leq \delta$ with probability at least $1-\eta$.*

In the rest of this section, to control the incoherence, we generate $\mathbf{X}$ for a given N by taking the first $N_1 = \min(\lfloor \tau N \rfloor, n)$ rows (with $0 \leq \tau \leq 1$, default to 0) from the first columns of Φ and the elements of the remaining $N_2 = N - N_1$ rows iid from $\mathcal{N}(0, 1/n)$ so that $\tilde{\mathbf{X}} = \mathbf{X}\Phi = \begin{bmatrix} \Phi^\top_{:,:N_1} \\ \mathbf{X}_{N_1:,:} \end{bmatrix} \Phi = \begin{bmatrix} \mathbb{I}_{N_1\times n} \\ \mathbf{X}_{N_1:,:}\Phi \end{bmatrix}$ with $\mathbf{X}_{N_1:,:} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$. The higher τ (and so N_1), the less incoherence between the measures (columns of $\mathbf{X}^\top$) and Φ. For a given s, we generate a random vector $\mathbf{b}^* \overset{iid}{\sim} \mathcal{N}(0, 1/n)$ such that $\|\mathbf{b}^*\|_0 \leq s$, and set $\mathbf{a}^* = \Phi\mathbf{b}^*$. We used $\Phi = \mathbb{I}_n$ for simplicity.

The problem (P_1) can be solved easily using convex programming library, with relative error $\|\mathbf{b} - \mathbf{b}^*\|_2 / \|\mathbf{b}^*\|_2$ of the order of 10^{-6} (Section C.5, Figures 6 and 7). As s and/or τ increases, $N_{\min}(s,\tau)$, the number of samples needs for perfect recovery increases. When $\tau \to 1$, $N_{\min}(s,\tau) \to n$ for all s.

C.5 Convex Optimization Formulations

Consider the problem of recovering $\mathbf{b}^*$ from noiseless measurements:

$$\begin{aligned} (P1\text{-noiseless}): \quad & \min_{\mathbf{b}} \|\mathbf{b}\|_1 \\ & \text{subject to} \quad \tilde{\mathbf{X}}\mathbf{b} = \mathbf{y}^*, \end{aligned} \tag{16}$$

where $\mathbf{y}^* = \tilde{\mathbf{X}}\mathbf{b}^*$. To rewrite the ℓ_1-norm objective linearly, let introduce auxiliary variables $\mathbf{t}_i$ for each component $\mathbf{b}_i$, and impose $-\mathbf{t}_i \leq \mathbf{b}_i \leq \mathbf{t}_i, \quad \mathbf{t}_i \geq 0, \quad$ for $i = 1, \ldots, n$. Then, since $\|\mathbf{b}\|_1 = \sum_{i=1}^n |\mathbf{b}_i|$, minimizing $\|\mathbf{b}\|_1$ is equivalent to minimizing $\sum_{i=1}^n \mathbf{t}_i$ subject to these constraints. The problem becomes

$$\begin{aligned} \min_{\mathbf{b},\mathbf{t}} \quad & \sum_{i=1}^n \mathbf{t}_i \\ \text{subject to} \quad & \tilde{\mathbf{X}}\mathbf{b} = \mathbf{y}^*, \\ & -\mathbf{t}_i \leq \mathbf{b}_i \leq \mathbf{t}_i, \quad i = 1, \ldots, n, \\ & \mathbf{t}_i \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{17}$$

All constraints and the objective function are linear, so this reformulation is a linear program (LP). Now assume the measurements are noisy $\mathbf{y}^* = \tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}$ and we allow for a noise tolerance $\epsilon \geq 0$. The recovery problem is

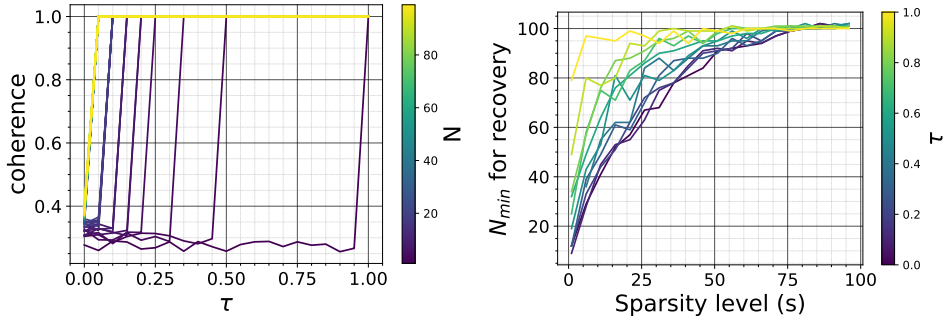
$$\begin{aligned} (P1\text{-noisy}): \quad & \min_{\mathbf{b}} \|\mathbf{b}\|_1 \\ & \text{subject to} \quad \|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2 \leq \epsilon. \end{aligned} \tag{18}$$

and by introducing the auxiliary variables, it becomes

$$\begin{aligned} \min_{\mathbf{b},\mathbf{t}} \quad & \sum_{i=1}^n \mathbf{t}_i \\ \text{subject to} \quad & \|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2 \leq \epsilon, \\ & -\mathbf{t}_i \leq \mathbf{b}_i \leq \mathbf{t}_i, \quad i = 1, \ldots, n, \\ & \mathbf{t}_i \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{19}$$

The constraints $-\mathbf{t}_i \leq \mathbf{b}_i \leq \mathbf{t}_i$ and $\mathbf{t}_i \geq 0$ are linear, while the constraint $\|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2 \leq \epsilon$ defines a second-order (quadratic) cone. Thus, the overall problem is a second-order cone program (SOCP).

We fix $n = 10^2$ and solve for different (N, s, τ) the convex problem $(P1\text{-noiseless})$ using the `cvxpy` library. As s and/or τ increases, $N_{\min}(s,\tau)$, the number of samples needs for perfect recovery increases (Figures 6 and 7). When τ converges to 1, $N_{\min}(s,\tau) \to n$ for all s. The error in those figures is the relative recovery error $\|\mathbf{b} - \mathbf{b}^*\|_2 / \|\mathbf{b}^*\|_2$. This error is usually of the order of 10^{-6}. This value gives us a basis for comparison with other methods.



(a) (b)

Figure 6: **(a)** Cohence $\mu(\mathbf{X}^\top, \Phi)$ as a function of $\tau \in (0,1)$ **(b)** Minimum number of samples for perfect recovery (relative recovery error $\leq 10^{-6}$) for $n = 10^2$ as a function of the sparsity level $s \in [n]$ and coherence parameter $\tau \in (0,1)$

```
# Input : X, Phi, y_star, n, EPSILON
import cvxpy as cp
b = cp.Variable(n)
objective = cp.Minimize(cp.norm(b, p=1))
constraints = [cp.norm(X @ (Phi @ b) - y_star, 2) <= EPSILON]
problem = cp.Problem(objective, constraints)
problem.solve()
b = b.value
```

C.6 Subgradient Descent

Let $\mathbf{y}(\mathbf{b}) = \mathcal{F}_{\mathbf{b}}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\mathbf{b}$. We have $\mathbf{y}^* = \mathcal{F}_{\mathbf{b}^*}(\tilde{\mathbf{X}}) + \boldsymbol{\xi} = \tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}$, and want to minimize $f(\mathbf{b}) = g_{\beta_2}(\mathbf{b}) + \beta_1\|\mathbf{b}\|_1$ using gradient descent, where

$$
\begin{aligned}
g_{\beta_2}(\mathbf{b}) &:= \frac{1}{2}\|\mathbf{y}(\mathbf{b}) - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}\|_2^2 \\
&= \frac{1}{2}\mathbf{b}^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^{*\top}\tilde{\mathbf{X}}\mathbf{b} + \frac{1}{2}\mathbf{y}^{*\top}\mathbf{y}^* + \frac{\beta_2}{2}\mathbf{b}^\top\mathbf{b} \\
&= \frac{1}{2}\mathbf{b}^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b} - \left(\mathbf{b}^{*\top}\tilde{\mathbf{X}}^\top + \boldsymbol{\xi}^\top\right)\tilde{\mathbf{X}}\mathbf{b} + \frac{1}{2}\left(\mathbf{b}^{*\top}\tilde{\mathbf{X}}^\top + \boldsymbol{\xi}^\top\right)\left(\tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}\right) + \frac{\beta_2}{2}\mathbf{b}^\top\mathbf{b} \\
&= \begin{cases} \frac{1}{2}\mathbf{b}^\top\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\mathbf{b} - \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b}^* + \tilde{\mathbf{X}}^\top\boldsymbol{\xi}\right)^\top\mathbf{b} + \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}\|_2^2 \\ \frac{1}{2}(\mathbf{b}-\mathbf{b}^*)^\top\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)(\mathbf{b}-\mathbf{b}^*) - \left(\tilde{\mathbf{X}}^\top\boldsymbol{\xi} - \beta_2\mathbf{b}^*\right)^\top(\mathbf{b}-\mathbf{b}^*) + \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}^*\|_2^2 \end{cases}
\end{aligned}
\tag{20}
$$

We write $F(\mathbf{b}) := G_{\beta_2}(\mathbf{b}) + \beta_1 h(\mathbf{b})$ with

$$
G_{\beta_2}(\mathbf{b}) := \nabla_{\mathbf{b}} g_{\beta_2}(\mathbf{b}) = \tilde{\mathbf{X}}^\top(\mathbf{y} - \mathbf{y}^*) + \beta_2\mathbf{b} = \begin{cases} \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\mathbf{b} - \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b}^* + \tilde{\mathbf{X}}^\top\boldsymbol{\xi}\right) \\ \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)(\mathbf{b}-\mathbf{b}^*) - \left(\tilde{\mathbf{X}}^\top\boldsymbol{\xi} - \beta_2\mathbf{b}^*\right) \end{cases}
\tag{21}
$$

and $h(\mathbf{b}) \in \partial\|\mathbf{b}\|_1$ any subgradient of $\|\mathbf{b}\|_1$, that is $h(\mathbf{b})_i = \text{sign}(\mathbf{b}_i)$ for $\mathbf{b}_i \neq 0$, and any value in $[+1, -1]$ for $\mathbf{b}_i = 0$. We used $h(\mathbf{b}) = \text{sign}(\mathbf{b})$ for simplicity and without loss of generality.

Suppose we start at some $\mathbf{b}^{(1)} := \zeta\tilde{\mathbf{b}}^{(1)}$, with $\zeta \geq 0$ the initialization scale and $\tilde{\mathbf{b}}^{(1)} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$. Using $\mathbf{F}^{(t)} := F(\mathbf{b}^{(t)})$, the subgradient update rule is

$$
\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha_t\mathbf{F}^{(t)} \quad \forall t > 1 \tag{22}
$$

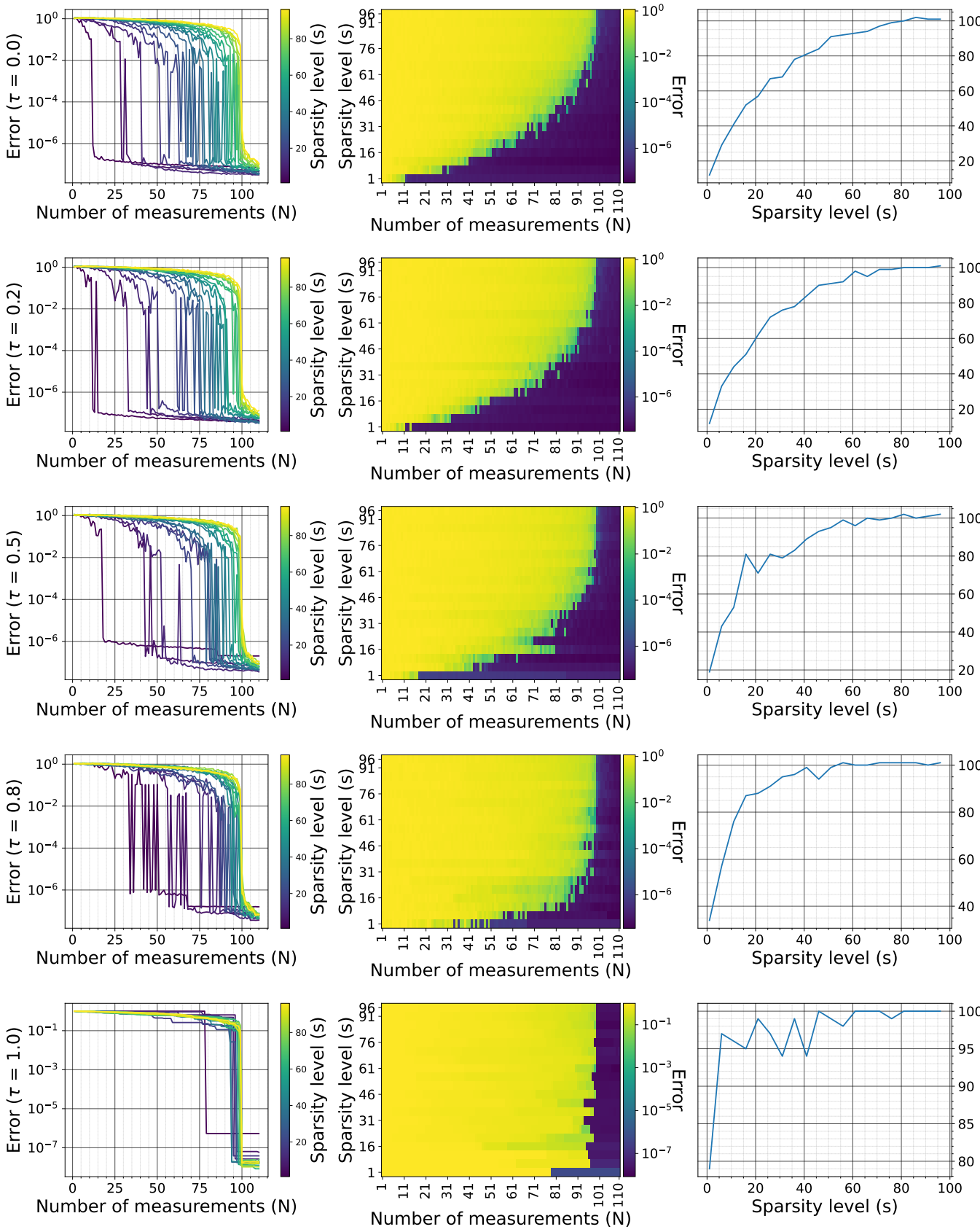


Figure 7: Relative error $\|\mathbf{b} - \mathbf{b}^*\|_2 / \|\mathbf{b}^*\|_2$ as a function of the number of measurements N, the sparsity level $s \in [n]$ and and coherence parameter $\tau \in (0, 1)$, for $n = 10^2$

with α_t the learning rate at step t. That is, using $\mathbf{h}^{(t)} = h(\mathbf{b}^{(t)})$,

$$\begin{cases} \mathbf{b}^{(t+1)} = \left[\mathbb{I}_n - \alpha_t \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right)\right] \mathbf{b}^{(t)} + \alpha_t \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi}\right) - \beta_1 \alpha_t \mathbf{h}^{(t)} \\ \mathbf{b}^{(t+1)} - \mathbf{b}^* = \left[\mathbb{I}_n - \alpha_t \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right)\right] (\mathbf{b}^{(t)} - \mathbf{b}^*) + \alpha_t \left(\tilde{\mathbf{X}}^\top \boldsymbol{\xi} - \beta_2 \mathbf{b}^*\right) - \beta_1 \alpha_t \mathbf{h}^{(t)} \end{cases} \tag{23}$$

We let $f^* = f(\mathbf{b}^*) = \beta_1 \|\mathbf{b}^*\|_1 + \frac{\beta_2}{2} \|\mathbf{b}^*\|_2^2 + \|\boldsymbol{\xi}\|_2^2$ and $f^{(t)} = f(\mathbf{b}^{(t)})$. Since the subgradient method is not a descent method, we let $\mathbf{b}_{\text{best}}^{(t)} = \arg\min_{\mathbf{b} \in \left\{\mathbf{b}^{(t')}, t' \le t\right\}} f(\mathbf{b}) = \arg\min_{\mathbf{b} \in \left\{\mathbf{b}_{\text{best}}^{(t-1)}, \mathbf{b}^{(t)}\right\}} f(\mathbf{b})$ be the best point found so far at step t, and $f_{\text{best}}^{(t)} = f(\mathbf{b}_{\text{best}}^{(t)}) = \min\left\{f_{\text{best}}^{(t-1)}, f^{(t)}\right\}$. This $\mathbf{b}_{\text{best}}^{(t)}$ can be made η-optimal for an arbitrary precision η if the step rule is chosen appropriately, as the following theorem shows.

Theorem C.3. *If $\|F(\mathbf{b})\|_2 \le L \quad \forall \mathbf{b}$ and $\|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2 \le R$, then $f_{best}^{(T)} - f^* \le \frac{R^2 + L^2 \sum_{t=1}^T \alpha_t^2}{2 \sum_{t=1}^T \alpha_t}$.*

Proof. By the definition of the subgradient $\mathbf{F}^{(T)} = F(\mathbf{b}^{(T)})$ of f at $\mathbf{b}^{(T)}$, we have $f(\mathbf{b}^{(T)}) + \left(\mathbf{b}^* - \mathbf{b}^{(T)}\right)^\top \mathbf{F}^{(T)} \le f(\mathbf{b}^*)$, i.e. $-\left(\mathbf{b}^{(T)} - \mathbf{b}^*\right)^\top \mathbf{F}^{(T)} \le -(f^{(T)} - f^*)$. So

$$\begin{aligned} 0 \le \|\mathbf{b}^{(T+1)} - \mathbf{b}^*\|_2^2 &= \|\mathbf{b}^{(T)} - \alpha_T \mathbf{F}^{(T)} - \mathbf{b}^*\|_2^2 \\ &= \|\mathbf{b}^{(T)} - \mathbf{b}^*\|_2^2 - 2\alpha_T \left(\mathbf{b}^{(T)} - \mathbf{b}^*\right)^\top \mathbf{F}^{(T)} + \alpha_T^2 \|\mathbf{F}^{(T)}\|_2^2 \\ &\le \|\mathbf{b}^{(T)} - \mathbf{b}^*\|_2^2 - 2\alpha_T \left(f^{(T)} - f^*\right) + \alpha_T^2 \|\mathbf{F}^{(T)}\|_2^2 \\ &\le \|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2^2 - 2\sum_{t=1}^T \alpha_t \left(f^{(t)} - f^*\right) + \sum_{t=1}^T \alpha_t^2 \|\mathbf{F}^{(t)}\|_2^2 \end{aligned} \tag{24}$$

This implies

$$2(f_{\text{best}}^{(T)} - f^*) \sum_{t=1}^T \alpha_t \le 2\sum_{t=1}^T \alpha_t \left(f^{(t)} - f^*\right) \le \|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2^2 + \sum_{t=1}^T \alpha_t^2 \|\mathbf{F}^{(t)}\|_2 \le R^2 + L^2 \sum_{t=1}^T \alpha_t^2 \tag{25}$$

□

The second condition of this theorem can always be satisfied by choosing an initialization appropriately. For example, if $\zeta = 0$, then we can take $R = \|\mathbf{b}^*\|_2$. The second condition will be satisfied if, for example, f satisfies the Lipschitz condition $|f(\mathbf{u}) - f(\mathbf{v})| \le L\|\mathbf{u} - \mathbf{v}\|_2$ for all $\mathbf{u}, \mathbf{v}$. But the condition is satisfied if and only if $\mathbf{b}$ (or just the $\mathbf{b}^{(t)}$) is restricted to a bounded domain since $F(\mathbf{b})$ is a linear function (up to $\gamma h(\mathbf{b})$). If $\|\mathbf{b}\|_2 \le B \quad \forall \mathbf{b}$, then $\|F(\mathbf{b})\|_2 \le \|\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\| \|\mathbf{b}\|_2 + \|\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi}\|_2 + \beta_1 \|h(\mathbf{b})\|_2 = \|\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\| B + \|\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi}\|_2 + \beta_1 \sqrt{n}$. Note that we always have $\|\mathbf{b}^{(t+1)}\|_2 \le \|\mathbb{I}_n - \alpha_t \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right)\| \|\mathbf{b}^{(t)}\|_2 + \alpha_t \|\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi}\|_2 + \beta_1 \alpha_t \|\mathbf{h}^{(t)}\|_2 \le \max_k |1 - \alpha_t \left(\sigma_k^2(\tilde{\mathbf{X}}) + \beta_2\right)| \|\mathbf{b}^{(t)}\| + \alpha_t \left(\sigma_{\max}^2(\tilde{\mathbf{X}}) \|\mathbf{b}^*\|_2 + \sigma_{\max}(\tilde{\mathbf{X}}) \|\boldsymbol{\xi}\|_2\right) + \beta_1 \alpha_t \sqrt{n}$.

That said, many step size rules lead to different accuracy.

Corollary C.1. *With a constant step size, $\alpha_t = \alpha$*

$$f_{best}^{(T)} - f^* \le \frac{R^2 + L^2 T \alpha^2}{2T\alpha} \longrightarrow_{T \to \infty} L^2 \alpha / 2 \tag{26}$$

In that case, we need a small learning rate and longer training time to achieve low errors.

With a square summable but not summable step size rule, $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$, we have

$$f_{best}^{(T)} - f^* \le \frac{R^2 + L^2 \sum_{i=1}^T \alpha_i^2}{2 \sum_{i=1}^T \alpha_i} \longrightarrow_{T \to \infty} 0 \tag{27}$$

For example, $\alpha_t = a/(b + t)$, $a > 0$ and $b \ge 0$. This method is common in practice for subgradient methods.

To explain grokking in such a setting, we will look at the landscape of the solution. Let $\tilde{\mathbf{X}} = \mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{V}^\top$ under the SVD decomposition, with $\Sigma = \mathrm{diag}(\sigma_k)_{k\in[r]}$, where $r = \mathrm{rank}(\tilde{\mathbf{X}})$ and $\sigma_{\max} = \sigma_1 \geq \cdots \sigma_k \geq \sigma_{k+1} \cdots \geq \sigma_{\min} = \sigma_r > \sigma_{r+1} = \cdots = 0$. We assume by default the SVD to be compact, i.e., $\mathbf{U} \in \mathbb{R}^{N\times r}$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$ have orthonormal columns, but we will make precision when we want it full, i.e., they also orthonormal rows, with that time $\mathbf{U} \in \mathbb{R}^{N\times N}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$. Using $\tilde{\Sigma}^{(t)} = \mathbb{I} - \alpha_t\left(\Sigma + \beta_2\mathbb{I}\right)$, the dynamics rewrites

$$\begin{cases} \mathbf{b}^{(t+1)} = \mathbf{V}\tilde{\Sigma}^{(t)}\mathbf{V}^\top\mathbf{b}^{(t)} + \alpha_t\left(\mathbf{V}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right) - \beta_1\alpha_t\mathbf{h}^{(t)} \\ \mathbf{b}^{(t+1)} - \mathbf{b}^* = \mathbf{V}\tilde{\Sigma}^{(t)}\mathbf{V}^\top(\mathbf{b}^{(t)} - \mathbf{b}^*) + \alpha_t\left(\mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi} - \beta_2\mathbf{b}^*\right) - \beta_1\alpha_t\mathbf{h}^{(t)} \end{cases}$$

We assume the step size $\alpha_t = \alpha$ satisfies $0 < \alpha < \frac{2}{\sigma_{\max}+\beta_2}$. In fact, for the dynamical system to converge, we need $\mathtt{Spec}\left[\mathbb{I}_n - \alpha_t\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\right] \subset (-1,1)$, that is $0 < \alpha_t < \frac{2}{\lambda_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})+\beta_2} = \frac{2}{\sigma^2_{\max}(\tilde{\mathbf{X}})+\beta_2} = \frac{2}{\sigma_{\max}+\beta_2}$.

For all $p > 0$, let define $\rho_p := \left\|\mathbb{I}_n - \alpha_t\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\right\|_{p\to p}$, so that $\rho_2 = \|\mathbb{I}_n - \alpha\left(\Sigma + \beta_2\mathbb{I}_n\right)\|_{2\to 2} = \max\{\max_{k\in[r]}|1-\alpha(\sigma_k+\beta_2)|, |1-\alpha\beta_2|\} \in (0,1)$.

C.6.1 MEMORIZATION

We will show that the update first moves to the least square solution of the problem, $\hat{\mathbf{b}} = \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top\mathbf{y}^* = \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\left(\Sigma\mathbf{V}^\top\mathbf{b}^* + \Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right)$; which is also the min norm solution for $N < n$ [2]. It moves exactly to $\hat{\mathbf{b}}$ (and stay there) for $\beta_1 = 0$ (Theorem C.6), and very close for β_1 small enough (Theorem C.8). If β_1 is too high, the subgradient term $h(\mathbf{b})$ dominates early, and there is no convergence, i.e., no memorization nor generalization (Theorem C.4). This $\hat{\mathbf{b}}$ can memorize (Theorem C.9) , but cannot generalize for $N < n$ (Theorem C.10).

Theorem C.4 (Oscillatory Behavior for Large β_1)**.** *Let $\mathbf{b}^{(1)} \in \mathbb{R}^n$. Consider the subgradient descent update*

$$\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha_t\left(\nabla_{\mathbf{b}}g_{\beta_2}(\mathbf{b}^{(t)}) + \beta_1 h(\mathbf{b}^{(t)})\right) \tag{28}$$

with a fixed step size $\alpha_t = \alpha > 0$, where $g_{\beta_2}(\mathbf{b}) = \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}\|_2^2$ and $h(\mathbf{b}) \in \partial\|\mathbf{b}\|_1$. If $\beta_1 > \frac{\sigma_{\max}+\beta_2}{\sqrt{n}}$ then the ℓ_1-term dominates the updates, causing the sequence $\mathbf{b}^{(t)}$ to exhibit oscillatory behavior without convergence to a minimizer of $f(\mathbf{b}) = g_{\beta_2}(\mathbf{b}) + \beta_1\|\mathbf{b}\|_1$. Consequently, neither memorization nor generalization is achieved, and both training and test errors oscillate above a suboptimal level.

Proof. We use lemma C.5 with $L = \|\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\|_{2\to 2} = \sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2$ (operator norm) be the Lipschitz constant for $G_{\beta_2}(\mathbf{b}) = \nabla_{\mathbf{b}}g_{\beta_2}(\mathbf{b}) = \tilde{\mathbf{X}}^\top(\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*) + \beta_2\mathbf{b} = \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\mathbf{b} - \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b}^* + \tilde{\mathbf{X}}^\top\boldsymbol{\xi}\right)$, since $\|G_{\beta_2}(\mathbf{u}) - G_{\beta_2}(\mathbf{v})\|_2 \leq L\|\mathbf{u}-\mathbf{v}\|_2$ for all $\mathbf{u}, \mathbf{v}$.

□

Lemma C.5. *Let $f(\mathbf{b}) = g(\mathbf{b}) + \beta_1\|\mathbf{b}\|_1$ be a convex function where g has a Lipschitz continuous gradient with Lipschitz constant $L > 0$, i.e., $\|\nabla g(\mathbf{u}) - \nabla g(\mathbf{v})\|_2 \leq L\|\mathbf{u}-\mathbf{v}\|_2$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Consider the subgradient descent update*

$$\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha\left(\nabla g(\mathbf{b}^{(t)}) + \beta_1 h(\mathbf{b}^{(t)})\right) \tag{29}$$

with a fixed step size $\alpha > 0$, where $h(\mathbf{b}^{(t)}) \in \partial\|\mathbf{b}^{(t)}\|_1$. If $\beta_1 > \frac{L}{\sqrt{n}}$ then the ℓ_1-term dominates the updates, causing the sequence $\{\mathbf{b}^{(t)}\}_{t>1}$ to exhibit oscillatory behavior without convergence

[2] Assume $\beta_2 = 0$. For $N \geq n$, the least square solution is $\hat{\mathbf{b}} = \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\right)^\dagger\tilde{\mathbf{X}}^\top\mathbf{y}^* = \mathbf{V}\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{-\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}$; and for $N < n$, the min norm solution is $\hat{\mathbf{b}} = \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^\dagger\mathbf{y}^* = \mathbf{V}\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{-\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}$

to a minimizer of f. Consequently, neither memorization nor generalization is achieved, and both training and test errors oscillate above a suboptimal level.

Proof Sketch. Since g has a Lipschitz continuous gradient with constant L, $\|\nabla g(\mathbf{b}^{(t)})\|_2 \leq L$ for all t when $\mathbf{b}^{(t)}$ is in a bounded region. Given that $\|h(\mathbf{b}^{(t)})\|_2 \approx \sqrt{n}$ at the beginning of training, if $\beta_1 > \frac{L}{\sqrt{n}}$, then

$$\beta_1 \|h(\mathbf{b}^{(t)})\|_2 \approx \beta_1 \sqrt{n} > L \geq \|\nabla g(\mathbf{b}^{(t)})\|_2 \tag{30}$$

This inequality implies that the update is dominated by the ℓ_1-term:

$$\mathbf{b}^{(t+1)} \approx \mathbf{b}^{(t)} - \alpha\beta_1 \, h(\mathbf{b}^{(t)}) \tag{31}$$

with the influence of $\nabla g(\mathbf{b}^{(t)})$ becoming negligible. Because $h(\mathbf{b}^{(t)})$ reflects the sign of $\mathbf{b}^{(t)}$, the update effectively pushes the iterates in a direction that primarily depends on sign changes rather than the curvature or detailed shape of g. This often leads to overshooting and sign flipping in each coordinate, resulting in oscillations. Consequently, the iterates do not converge to a stable minimizer of f, and the error metrics (both training and test) oscillate, remaining above some suboptimal threshold. This behavior indicates that the algorithm fails to memorize training data properly and cannot generalize well when β_1 is excessively large. □

Let us focus on reasonable values of β_1, starting with $\beta_1 = 0$.

Theorem C.6. *If $\beta_1 = 0$ and $\alpha = \alpha_t \in (0, \frac{2}{\sigma_{\max} + \beta_2})$ $\quad \forall t$, then $G_{\beta_2}(\mathbf{b}^{(t)}) \to 0$ as $t \to \infty$; where*

$$G_{\beta_2}(\mathbf{b}) = 0 \Longleftrightarrow \mathbf{b} = \hat{\mathbf{b}} + \left(\mathbb{I}_n - \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right)^\dagger \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right) \right) \mathbf{c} = \hat{\mathbf{b}} + \left(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top \right) \mathbf{c} \quad \forall \mathbf{c} \in \mathbb{R}^n \tag{32}$$

Also,

$$\|\mathbf{b}^{(t+1)} - \hat{\mathbf{b}}\|_2 \leq \rho_2^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_2 \quad \forall t \in \mathbb{N} \tag{33}$$

Proof. The solutions of $G_{\beta_2}(\mathbf{b}) = 0$ are

$$\begin{aligned}
&\begin{cases} \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right) \mathbf{b} = \tilde{\mathbf{X}}^\top \mathbf{y}^* \\ \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right) (\mathbf{b} - \mathbf{b}^*) = \left(\tilde{\mathbf{X}}^\top \boldsymbol{\xi} - \beta_2 \mathbf{b}^* \right) \end{cases} \\
&\Longleftrightarrow \begin{cases} \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right) \mathbf{b} = \tilde{\mathbf{X}}^\top \mathbf{y}^* = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi} = \mathbf{V} \Sigma \mathbf{V}^\top \mathbf{b}^* + \mathbf{V} \Sigma^{\frac{1}{2}} \mathbf{U}^\top \boldsymbol{\xi} \\ \mathbf{V} \left(\Sigma + \beta_2 \mathbb{I} \right) \mathbf{V}^\top (\mathbf{b} - \mathbf{b}^*) = \left(\mathbf{V} \Sigma^{\frac{1}{2}} \mathbf{U}^\top \boldsymbol{\xi} - \beta_2 \mathbf{b}^* \right) \end{cases} \\
&\Longleftrightarrow \begin{cases} \mathbf{b} = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right)^\dagger \tilde{\mathbf{X}}^\top \mathbf{y}^* + \left(\mathbb{I}_n - \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right)^\dagger \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right) \right) \mathbf{c} = \hat{\mathbf{b}} + \left(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top \right) \mathbf{c} \quad \forall \mathbf{c} \in \mathbb{R}^n \\ \mathbf{b} - \mathbf{b}^* = \left[\mathbf{V} \left(\Sigma + \beta_2 \mathbb{I} \right)^{-1} \Sigma \mathbf{V}^\top - \mathbb{I}_n \right] \mathbf{b}^* + \mathbf{V} \left(\Sigma + \beta_2 \mathbb{I} \right)^{-1} \Sigma^{\frac{1}{2}} \mathbf{U}^\top \boldsymbol{\xi} + \left(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top \right) \mathbf{c} \quad \forall \mathbf{c} \in \mathbb{R}^n \end{cases}
\end{aligned} \tag{34}$$

We know that

$$\begin{aligned}
\mathbf{z}^{(t+1)} &= \mathbf{A}^{(t)}\mathbf{z}^{(t)} + \mathbf{w}^{(t)} \\
&= \left(\prod_{k=t}^{t-i}\mathbf{A}^{(k)}\right)\mathbf{z}^{(t-i)} + \sum_{j=t-i}^{t-1}\left(\prod_{k=t}^{j+1}\mathbf{A}^{(k)}\right)\mathbf{w}^{(j)} + \mathbf{w}^{(t)} \quad \forall i \leq t \\
&= \begin{cases} \left(\prod_{k=t}^{1}\mathbf{A}^{(k)}\right)\mathbf{z}^{(1)} + \sum_{j=1}^{t-1}\left(\prod_{k=t}^{j+1}\mathbf{A}^{(k)}\right)\mathbf{w}^{(j)} + \mathbf{w}^{(t)} \\ \left(\prod_{k=t}^{0}\mathbf{A}^{(k)}\right)\mathbf{z}^{(0)} + \sum_{j=0}^{t-1}\left(\prod_{k=t}^{j+1}\mathbf{A}^{(k)}\right)\mathbf{w}^{(j)} + \mathbf{w}^{(t)} \end{cases} \\
&= \begin{cases} \mathbf{A}^t\mathbf{z}^{(1)} + \sum_{j=1}^{t}\mathbf{A}^{t-j}\mathbf{w}^{(j)} \\ \mathbf{A}^{t+1}\mathbf{z}^{(0)} + \sum_{j=0}^{t}\mathbf{A}^{t-j}\mathbf{w}^{(j)} \end{cases} \quad \text{if } \mathbf{A}^{(k)} = \mathbf{A} \quad \forall k \\
&= \begin{cases} \mathbf{A}^t\mathbf{z}^{(1)} + \left(\sum_{i=0}^{t-1}\mathbf{A}^i\right)\mathbf{w} \\ \mathbf{A}^{t+1}\mathbf{z}^{(0)} + \left(\sum_{i=0}^{t}\mathbf{A}^i\right)\mathbf{w} \end{cases} \quad \text{if } \mathbf{A}^{(k)} = \mathbf{A} \text{ and } \mathbf{w}^{(k)} = \mathbf{w} \quad \forall k \\
&= \begin{cases} \mathbf{A}^t\mathbf{z}^{(1)} + \left(\mathbb{I} - \mathbf{A}\right)^{\dagger}\left(\mathbb{I} - \mathbf{A}^t\right)\mathbf{w} \\ \mathbf{A}^{t+1}\mathbf{z}^{(0)} + \left(\mathbb{I} - \mathbf{A}\right)^{\dagger}\left(\mathbb{I} - \mathbf{A}^{t+1}\right)\mathbf{w} \end{cases} \quad \text{if } \mathbf{A}^{(k)} = \mathbf{A} \text{ and } \mathbf{w}^{(k)} = \mathbf{w} \quad \forall k
\end{aligned} \tag{35}$$

Let $\mathbf{A}^{(t)} = \mathbb{I}_n - \alpha_t\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right) = \mathbf{V}\tilde{\Sigma}^{(t)}\mathbf{V}^\top$ and $\mathbf{w}^{(t)} = \alpha_t\tilde{\mathbf{X}}^\top\mathbf{y}^* = \alpha_t\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b}^* + \tilde{\mathbf{X}}^\top\boldsymbol{\xi}\right) = \alpha_t\left(\mathbf{V}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right)$; so that $\mathbf{b}^{(t+1)} = \mathbf{A}^{(t)}\mathbf{b}^{(t+1)} + \mathbf{w}^{(t)}$ when $\beta_1 = 0$. For $\alpha_t = \alpha$, we let $\mathbf{A} = \mathbf{V}\tilde{\Sigma}\mathbf{V}^\top$ and $\mathbf{w} = \alpha\left(\mathbf{V}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right)$. As $t \longrightarrow \infty$, $\tilde{\Sigma}^t \longrightarrow 0$. We have

$$\begin{aligned}
\sum_{i=0}^{t-1}\mathbf{A}^i &= \mathbb{I}_n + \sum_{i=1}^{t-1}\mathbf{V}\tilde{\Sigma}^i\mathbf{V}^\top \\
&= \mathbb{I}_n - \mathbf{V}\mathbf{V}^\top + \sum_{i=0}^{t-1}\mathbf{V}\tilde{\Sigma}^i\mathbf{V}^\top \\
&= \mathbb{I}_n - \mathbf{V}\mathbf{V}^\top + \mathbf{V}\,\mathrm{diag}\left(\sum_{i=0}^{t-1}\tilde{\sigma}_k^i\right)_k\mathbf{V}^\top \\
&= \mathbb{I}_n - \mathbf{V}\mathbf{V}^\top + \mathbf{V}\,\mathrm{diag}\left(\frac{1-\tilde{\sigma}_k^t}{1-\tilde{\sigma}_k}\right)_k\mathbf{V}^\top \\
&= \mathbb{I}_n - \mathbf{V}\mathbf{V}^\top + \mathbf{V}\left(\mathbb{I} - \tilde{\Sigma}\right)^{-1}\left(\mathbb{I} - \tilde{\Sigma}^t\right)\mathbf{V}^\top \\
&\longrightarrow \mathbb{I}_n - \mathbf{V}\mathbf{V}^\top + \mathbf{V}\left(\mathbb{I} - \tilde{\Sigma}\right)^{-1}\mathbf{V}^\top = \mathbb{I}_n - \mathbf{V}\mathbf{V}^\top + \frac{1}{\alpha}\mathbf{V}\left(\Sigma + \beta_2\mathbb{I}_n\right)^{-1}\mathbf{V}^\top \text{ as } t \longrightarrow \infty
\end{aligned} \tag{36}$$

So, as $t \longrightarrow \infty$,

$$\begin{aligned}
\mathbf{b}^{(t+1)} &= \left(\sum_{i=0}^{\infty}\mathbf{A}^i\right)\mathbf{w} \\
&= \alpha\left(\mathbb{I}_r - \mathbf{V}\mathbf{V}^\top + \frac{1}{\alpha}\mathbf{V}\left(\Sigma + \beta_2\mathbb{I}_r\right)^{-1}\mathbf{V}^\top\right)\left(\mathbf{V}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right) \\
&= \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\mathbf{V}^\top\left(\mathbf{V}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right) + \left(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top\right)\mathbf{c} \text{ with } \mathbf{c} = \mathbf{w} \\
&= \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\mathbf{V}^\top\left(\mathbf{V}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{V}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right) \\
&= \hat{\mathbf{b}}
\end{aligned} \tag{37}$$

We have $\mathbf{A}\hat{\mathbf{b}} + \mathbf{c} = \hat{\mathbf{b}}$, so $\mathbf{b}^{(t+1)} - \hat{\mathbf{b}} = \mathbf{A}(\mathbf{b}^{(t)} - \hat{\mathbf{b}}) = \mathbf{A}^t(\mathbf{b}^{(1)} - \hat{\mathbf{b}})$, which implies $\|\mathbf{b}^{(t+1)} - \hat{\mathbf{b}}\|_2 \leq \|\mathbf{A}^t\|_{2\to2}\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_2$; with $\|\mathbf{A}^t\|_{2\to2} = \sigma_{\max}(\mathbf{A}^t) = \sigma_{\max}(\mathbf{A})^t = \rho_2^t$. □

We now move to a general case with $\beta_1 \geq 0$.

Lemma C.7. *For all $p > 0$ such that $\rho_p < 1$, we have*

$$\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p \leq \rho_p^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p + \alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \leq \rho_p^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p + \frac{\alpha\beta_1 n^{1/p}}{1-\rho_p} \quad \forall t \geq 1 \tag{38}$$

In particular,

$$\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_2 \leq \rho^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_2 + \alpha\beta_1 \sqrt{n} \frac{1-\rho_2^t}{1-\rho_2} \leq \rho^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_2 + \frac{\alpha\beta_1\sqrt{n}}{1-\rho_2} \quad \forall t \geq 1 \tag{39}$$

and

$$\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_\infty \leq \rho_\infty^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_\infty + \alpha\beta_1 \frac{1-\rho_\infty^t}{1-\rho_\infty} \leq \rho_\infty^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_\infty + \frac{\alpha\beta_1}{1-\rho_\infty} \quad \forall t \geq 1 \tag{40}$$

Proof. Recall

$$G_{\beta_2}(\mathbf{b}) = \tilde{\mathbf{X}}^\top(\mathbf{y} - \mathbf{y}^*) + \beta_2 \mathbf{b} = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right) \mathbf{b} - \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi}\right) \tag{41}$$

Starting from the update rule

$$\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha\Big(G_{\beta_2}(\mathbf{b}^{(t)}) + \beta_1 \mathbf{h}^{(t)}\Big) \tag{42}$$

We have

$$\mathbf{b}^{(t+1)} - \hat{\mathbf{b}} = \left(\mathbf{b}^{(t)} - \hat{\mathbf{b}}\right) - \alpha\left(G_{\beta_2}(\mathbf{b}^{(t)}) + \beta_1 \mathbf{h}^{(t)}\right) \tag{43}$$

Since $G_{\beta_2}(\hat{\mathbf{b}}) = 0$ and G_{β_2} is linear,

$$G_{\beta_2}(\mathbf{b}^{(t)}) = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right)(\mathbf{b}^{(t)} - \hat{\mathbf{b}}) \tag{44}$$

Substituting this back,

$$\begin{aligned}
\mathbf{b}^{(t+1)} - \hat{\mathbf{b}} &= \left(\mathbf{b}^{(t)} - \hat{\mathbf{b}}\right) - \alpha\left(G_{\beta_2}(\mathbf{b}^{(t)}) + \beta_1 \mathbf{h}^{(t)}\right) \\
&= \left(\mathbf{b}^{(t)} - \hat{\mathbf{b}}\right) - \alpha\left(\left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right)(\mathbf{b}^{(t)} - \hat{\mathbf{b}}) + \beta_1 \mathbf{h}^{(t)}\right) \\
&= \left[\mathbb{I}_n - \alpha\left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n\right)\right]\left(\mathbf{b}^{(t)} - \hat{\mathbf{b}}\right) - \alpha\beta_1 \mathbf{h}^{(t)}
\end{aligned} \tag{45}$$

Taking the norm; applying triangle inequality and using $\|\mathbf{h}^{(t)}\|_p \leq n^{1/p}$ give

$$\|\mathbf{b}^{(t+1)} - \hat{\mathbf{b}}\|_p \leq \rho_p \|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p + \alpha\beta_1 n^{1/p} \tag{46}$$

Repeatedly applying the recurrence,

$$\begin{aligned}
\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p &\leq \rho_p^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p + \alpha\beta_1 n^{1/p}\left(1 + \rho_p + \cdots + \rho_p^{t-1}\right) \\
&= \rho_p^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p + \alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \quad \text{for } \rho_p \neq 1 \\
&\leq \rho_p^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p + \frac{\alpha\beta_1 n^{1/p}}{1-\rho_p} \text{ for } \rho_p < 1
\end{aligned}$$

□

Theorem C.8. *Let $p > 0$ such that $\rho_p < 1$. Define*

$$t_1 := \left\lceil -\frac{\ln\left(1 + \frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p}{\alpha\beta_1 n^{1/p}}\right)}{\ln(\rho_p)} \right\rceil \tag{47}$$

Then for all $t \geq t_1$,

$$\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p \leq 2\alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \leq 2\frac{\alpha\beta_1 n^{1/p}}{1-\rho_p} \tag{48}$$

and the prediction error for $t \geq t_1$ *is bounded by*

$$\begin{aligned} \|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_p &\leq 2\alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \|\tilde{\mathbf{X}}\|_{p\to p} + \|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_p \\ &\leq 2\frac{\alpha\beta_1 n^{1/p}}{1-\rho_p} \|\tilde{\mathbf{X}}\|_{p\to p} + \|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_p \end{aligned} \tag{49}$$

Proof. The definition of t_1 ensures that for $t \geq t_1$,

$$\rho^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p \leq \alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \tag{50}$$

Thus, using lemma C.7, we have for $t \geq t_1$,

$$\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p \leq 2\alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \tag{51}$$

Using this, we derive the following

$$\begin{aligned} \|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_p &= \|\tilde{\mathbf{X}}(\mathbf{b}^{(t)} - \hat{\mathbf{b}}) + (\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*)\|_p \\ &\leq \|\tilde{\mathbf{X}}\|_{p\to p} \|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p + \|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_p \\ &\leq 2\alpha\beta_1 n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \|\tilde{\mathbf{X}}\|_{p\to p} + \|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_p \text{ for } t \geq t_1 \end{aligned} \tag{52}$$

□

Corollary C.2. *Let* $p > 0$ *such that* $\rho_p < 1$. *Define*

$$\tilde{t}_1 := \begin{cases} \left\lceil -\frac{\ln\left(\frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p}{\alpha\beta_1 n^{1/p}}\right)}{\ln(\rho_p)} \right\rceil & \text{if } \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p > \frac{\alpha\beta_1}{1-\rho_p} \quad > t_1 \\ 0 & \text{otherwise} \end{cases} \tag{53}$$

Then for all $t \geq \tilde{t}_1$,

$$\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p \leq 2\frac{\alpha\beta_1 n^{1/p}}{1-\rho_p} \tag{54}$$

and the prediction error for $t \geq \tilde{t}_1$ *is bounded by*

$$\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_p \leq \frac{2\alpha\beta_1 n^{1/p}}{1-\rho_p} \|\tilde{\mathbf{X}}\|_{p\to p} + \|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_p \tag{55}$$

Proof. The definition of $\tilde{t}_1$ ensures that for $t \geq \tilde{t}_1$,

$$\rho^t \|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_\infty \leq \frac{\alpha\beta_1 n^{1/p}}{1-\rho_p} \tag{56}$$

The rest of the proof follows from lemma C.7.

□

When the initialization $\mathbf{b}^{(1)}$ is close to $\hat{\mathbf{b}}$, it takes less time to memorize since t_1 decreases with $\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p$, as well as $\tilde{t}_1$: if $\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p \leq \frac{\alpha\beta_1 n^{1/p}}{1-\rho_p}$, $\tilde{t}_1$ is trivialy 0, otherwise it decreases with $\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p > \frac{\alpha\beta_1 n^{1/p}}{1-\rho_p}$.

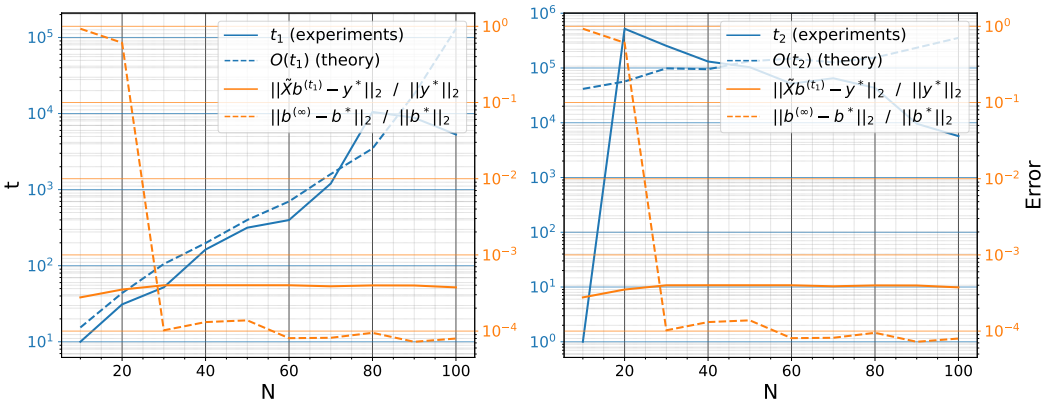


Figure 8: **(Left)** t_1 compute experimentally (when the relative training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ is reach 10^{-4}, see Figure 9) and the upper bound $-\ln\left(1 + \frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p}{\alpha\beta_1 n^{1/p}}\right) / \ln(\rho_p)$ computed in Theorem C.8, for $p = \infty$. **(Right)** Step t_2 compute experimentally (when the relative recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ reach 10^{-4} for the first time) and the upper bound $t_1 + \Delta t$. The notation $\mathbf{b}^{(\infty)}$ represent the update $\mathbf{b}^{(t)}$ at the end of training. The hyperparameters for this figure are $(n, s) = (100, 5)$, $N \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and $(\alpha, \beta_1, \beta_2) = (10^{-1}, 10^{-5}, 0)$.

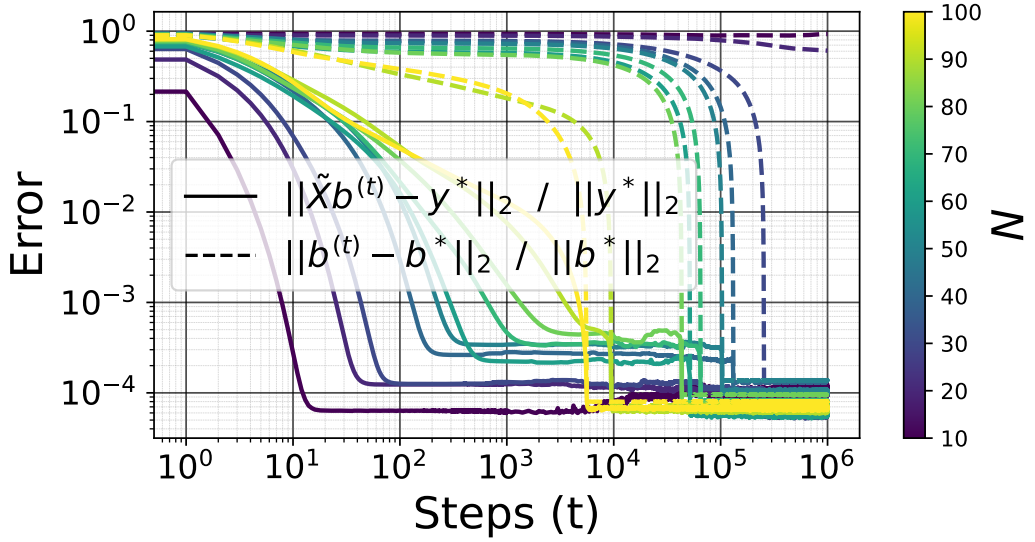


Figure 9: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the number of measurements $N \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and the subgradient descent training steps $0 \leq t \leq 2 \times 10^6$, for $(n, s) = (100, 5)$ and $(\alpha, \beta_1, \beta_2) = (10^{-1}, 10^{-5}, 0)$.

When the learning rate α alone becomes smaller, the term $\frac{\alpha\beta_1 n^{1/p}}{1-\rho_p}$ decreases, reducing the asymptotic error bound. However, a smaller α makes ρ_p closer to 1 (for example, $\rho_2 = \max\{\max_{k\in[r]}|1-\alpha(\sigma_k+\beta_2)|, |1-\alpha\beta_2|\}$), which increases t_1 and $\tilde{t}_1$. This means more iterations are needed to reach the regime where the error stabilizes near its lower bound. Another alternative for reducing the term $\frac{\alpha\beta_1 n^{1/p}}{1-\rho_p}$ and guaranteeing perfect memorization earlier is to reduce β_1. But we'll see below that this also increases the generalization delay.

Ideally, if the system, $\tilde{\mathbf{X}}\mathbf{b} = \mathbf{y}^*$ has an exact solution (and with appropriate β_2), then $\tilde{\mathbf{X}}\hat{\mathbf{b}} = \mathbf{y}^*$. In practice, due to noise in $\mathbf{y}^*$, the regularization with β_2, or model mismatch, the solution $\hat{\mathbf{b}}$ might not perfectly reproduce $\mathbf{y}^*$, resulting in a non zero residual $\|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_2$. Note that we have $\mathbf{y}^* = \tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi} = \mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{V}^\top\mathbf{b}^* + \boldsymbol{\xi}$, so

$$\begin{aligned}
\mathbf{y}(\hat{\mathbf{b}}) = \tilde{\mathbf{X}}\hat{\mathbf{b}} &= \begin{cases} \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top\mathbf{y}^* \\ \mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{V}^\top\mathbf{V}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\left(\Sigma\mathbf{V}^\top\mathbf{b}^* + \Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right) \end{cases} \\
&= \begin{cases} \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top\mathbf{y}^* \\ \mathbf{U}\Sigma^{\frac{1}{2}}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{U}\Sigma^{\frac{1}{2}}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi} \end{cases} \\
&= \begin{cases} \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top\mathbf{y}^* \\ \mathbf{U}\Sigma^{\frac{1}{2}}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{V}^\top\mathbf{b}^* + \mathbf{U}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{U}^\top\boldsymbol{\xi} \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{y}(\hat{\mathbf{b}}) - \mathbf{y}^* &= \begin{cases} \left[\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top - \mathbb{I}_N\right]\mathbf{y}^* \\ \mathbf{U}\Sigma^{\frac{1}{2}}\left[\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma - \mathbb{I}\right]\mathbf{V}^\top\mathbf{b}^* + \left[\mathbf{U}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{U}^\top - \mathbb{I}_N\right]\boldsymbol{\xi} \end{cases} \\
&= \begin{cases} \left[\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top - \mathbb{I}_N\right]\tilde{\mathbf{X}}\mathbf{b}^* + \left[\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top - \mathbb{I}_N\right]\boldsymbol{\xi} \\ \mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\left[\mathbf{U}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{U}^\top - \mathbb{I}\right]\mathbf{U}\mathbf{V}^\top\mathbf{b}^* + \left[\mathbf{U}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{U}^\top - \mathbb{I}_N\right]\boldsymbol{\xi} \end{cases}
\end{aligned}$$

Theorem C.9. *Assume* $\mathbb{E}[\boldsymbol{\xi}] = 0$ *and* $\mathrm{Cov}(\boldsymbol{\xi}) = \sigma_\xi^2\mathbb{I}_N$. *Then*

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_2^2\right] = \sum_{i=1}^{r}\left(\frac{\beta_2\sigma_i}{\sigma_i+\beta_2}\right)^2(\mathbf{V}^\top\mathbf{b}^*)_i^2 + \sum_{i=1}^{r}\left(\frac{\beta_2}{\sigma_i+\beta_2}\right)^2\sigma_\xi^2 + \sigma_\xi^2(N-r) \tag{57}$$

Proof. We have

$$\begin{aligned}
\hat{\mathbf{b}} &= \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger\tilde{\mathbf{X}}^\top\mathbf{y}^* \\
&= \mathbf{V}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\mathbf{y}^*
\end{aligned} \tag{58}$$

Next,

$$\begin{aligned}
\tilde{\mathbf{X}}\hat{\mathbf{b}} &= \mathbf{U}\Sigma^{\frac{1}{2}}\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\mathbf{y}^* \\
&= \mathbf{U}\Sigma\left(\Sigma+\beta_2\mathbb{I}\right)^{-1}\mathbf{U}^\top\mathbf{y}^* \\
&\overset{\beta_2=0}{=} \mathbf{U}\mathbf{U}^\top\mathbf{y}^*
\end{aligned} \tag{59}$$

Now consider the residual

$$\begin{aligned}
\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^* &= \mathbf{U}\Sigma\left(\Sigma+\beta_2\mathbb{I}_r\right)^{-1}\mathbf{U}^\top\mathbf{y}^* - \mathbf{U}\mathbf{U}^\top\mathbf{y}^* - (\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^* \\
&= \mathbf{U}\left[\Sigma\left(\Sigma+\beta_2\mathbb{I}_r\right)^{-1} - \mathbb{I}_r\right]\mathbf{U}^\top\mathbf{y}^* - (\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^* \\
&= -\beta_2\mathbf{U}(\Sigma+\beta_2\mathbb{I}_r)^{-1}\mathbf{U}^\top\mathbf{y}^* - (\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^* \text{ since } \Sigma(\Sigma+\beta_2\mathbb{I}_r)^{-1} - \mathbb{I}_r = -\beta_2(\Sigma+\beta_2\mathbb{I}_r)^{-1}
\end{aligned} \tag{60}$$

The first term, $\beta_2 \mathbf{U}(\Sigma + \beta_2 \mathbb{I}_r)^{-1}\mathbf{U}^\top \mathbf{y}^*$, lies in $\mathrm{Col}(\mathbf{U})$, while the second term, $(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^*$, lies in $\mathrm{Col}(\mathbf{U})^\perp$. Thus, they are orthogonal, and

$$\|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_2^2 = \|\beta_2 \mathbf{U}(\Sigma + \beta_2 \mathbb{I}_r)^{-1}\mathbf{U}^\top \mathbf{y}^*\|_2^2 + \|(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^*\|_2^2 \tag{61}$$

Let's start with the second term. Since $\mathbf{y}^* = \mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{V}^\top \mathbf{b}^* + \boldsymbol{\xi}$,

$$\begin{aligned}(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^* &= (\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{V}^\top \mathbf{b}^* + (\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\boldsymbol{\xi} \\ &= (\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\boldsymbol{\xi}\end{aligned} \tag{62}$$

So

$$\begin{aligned}\|(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^*\|_2^2 &= \|(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\boldsymbol{\xi}\|_2^2 \\ &= \boldsymbol{\xi}^\top(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\boldsymbol{\xi} \\ &= \boldsymbol{\xi}^\top(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\boldsymbol{\xi}\end{aligned} \tag{63}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\|(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{y}^*\|_2^2 &= \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{\xi}^\top(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\boldsymbol{\xi}\right] \\ &= \mathrm{tr}\left((\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)\,\mathrm{Cov}(\boldsymbol{\xi})\right) + (\mathbb{E}\boldsymbol{\xi})^\top(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top)(\mathbb{E}\boldsymbol{\xi}) \\ &= \sigma_\xi^2\,\mathrm{tr}\left(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top\right) \\ &= \sigma_\xi^2\left(N - \mathrm{tr}(\mathbf{U}\mathbf{U}^\top)\right) \\ &= \sigma_\xi^2\left(N - \mathrm{tr}(\mathbf{U}^\top\mathbf{U})\right) \\ &= \sigma_\xi^2\left(N - \mathrm{tr}(\mathbb{I}_r)\right) \\ &= \sigma_\xi^2(N - r)\end{aligned} \tag{64}$$

For the first term, we have

$$\begin{aligned}\|\beta_2 \mathbf{U}(\Sigma + \beta_2 \mathbb{I}_r)^{-1}\mathbf{U}^\top \mathbf{y}^*\|_2^2 &= \|\beta_2(\Sigma + \beta_2 \mathbb{I}_r)^{-1}\mathbf{U}^\top \mathbf{y}^*\|_2^2 \\ &= \sum_{i=1}^r \left(\frac{\beta_2}{\sigma_i + \beta_2}\right)^2 (\mathbf{U}^\top \mathbf{y}^*)_i^2 \\ &= \sum_{i=1}^r \left(\frac{\beta_2}{\sigma_i + \beta_2}\right)^2 \left(\sigma_i^{\frac{1}{2}}(\mathbf{V}^\top \mathbf{b}^*)_i + (\mathbf{U}^\top \boldsymbol{\xi})_i\right)^2 \text{ since } \mathbf{U}^\top \mathbf{y}^* = \Sigma^{\frac{1}{2}}\mathbf{V}^\top \mathbf{b}^* + \mathbf{U}^\top \boldsymbol{\xi} \\ &= \sum_{i=1}^r \left(\frac{\beta_2}{\sigma_i + \beta_2}\right)^2 \left(\sigma_i(\mathbf{V}^\top \mathbf{b}^*)_i^2 + 2\sigma_i^{\frac{1}{2}}(\mathbf{V}^\top \mathbf{b}^*)_i(\mathbf{U}^\top \boldsymbol{\xi})_i + (\mathbf{U}^\top \boldsymbol{\xi})_i^2\right)\end{aligned} \tag{65}$$

Using $\mathbb{E}_{\boldsymbol{\xi}}[(\mathbf{U}^\top \boldsymbol{\xi})_i] = 0$ and $\mathrm{Var}((\mathbf{U}^\top \boldsymbol{\xi})_i) = \sigma_\xi^2$, we get

$$\mathbb{E}_{\boldsymbol{\xi}}\|\beta_2 \mathbf{U}(\Sigma + \beta_2 \mathbb{I}_r)^{-1}\mathbf{U}^\top \mathbf{y}^*\|_2^2 = \sum_{i=1}^r \left(\frac{\beta_2}{\sigma_i + \beta_2}\right)^2 \left(\sigma_i(\mathbf{V}^\top \mathbf{b}^*)_i^2 + \sigma_\xi^2\right) \tag{66}$$

This concludes the proof. □

The expression

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_2^2\right] = \sum_{i=1}^r \left(\frac{\beta_2 \sigma_i}{\sigma_i + \beta_2}\right)^2 (\mathbf{V}^\top \mathbf{b}^*)_i^2 + \sum_{i=1}^r \left(\frac{\beta_2}{\sigma_i + \beta_2}\right)^2 \sigma_\xi^2 + \sigma_\xi^2(N - r) \tag{67}$$

offers insights into how various factors influence the prediction quality $\tilde{\mathbf{X}}\hat{\mathbf{b}}$.

Signal-to-Noise Ratio (SNR) $\|\mathbf{b}^*\|_2/\sigma_\xi$ When $\|\mathbf{b}^*\|_2$ is large compared to σ_ξ (high SNR), the signal component $(\mathbf{V}^\top\mathbf{b}^*)_i^2$ in the first sum becomes significant, and the bias introduced by regularization interacts more strongly with the true signal, so the first term largely determines the expected residual. Otherwise, the noise terms $\sum_{i=1}^r \left(\frac{\beta_2}{\sigma_i+\beta_2}\right)^2 \sigma_\xi^2 + \sigma_\xi^2(N-r)$ dominate the expected residual. In this case, noise largely drives the error, and recovering the signal becomes more challenging.

Effect of the Regularization Parameter β_2 If $\beta_2 \ll \sigma_i$ for most i, then $\frac{\beta_2}{\sigma_i+\beta_2} \approx \frac{\beta_2}{\sigma_i}$ and $\frac{\beta_2\sigma_i}{\sigma_i+\beta_2} \approx \frac{\beta_2}{\sigma_i}$. The bias and the noise contribution for dominant singular modes are both reduced, resulting in lower expected residual error. If $\beta_2 \gg \sigma_i$, then $\frac{\beta_2}{\sigma_i+\beta_2} \approx 1$ and $\frac{\beta_2\sigma_i}{\sigma_i+\beta_2} \approx \sigma_i$. Over-regularization increases bias and noise contributions, generally raising the expected residual. So β_2 controls the bias-variance tradeoff: increasing β_2 reduces variance but increases bias. The optimal β_2 minimizes the overall expected residual.

Dependence on $\tilde{\mathbf{X}}$ and its Rank. The rank r of $\tilde{\mathbf{X}}$ appears explicitly in the term $\sigma_\xi^2(N-r)$. If $\tilde{\mathbf{X}}$ is full rank (i.e., $r = N$ when $N \le n$), then the term $\sigma_\xi^2(N-r)$ vanishes, eliminating the noise component in the nullspace of $\tilde{\mathbf{X}}^\top$. For rank-deficient $\tilde{\mathbf{X}}$ ($r < N$), $\sigma_\xi^2(N-r)$ accounts for noise in directions orthogonal to the column space of $\tilde{\mathbf{X}}$. This part of the noise cannot be captured or reduced by the model, setting a lower bound on the residual error.

In practice, we run the experiment for different training data $\tilde{\mathbf{X}}$, then average the results. However, taking the expectation over the distribution of $\tilde{\mathbf{X}}$ (e.g., assuming $\tilde{\mathbf{X}}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/n)$) involves (i) Averaging over the singular values $\{\sigma_i\}$ of $\tilde{\mathbf{X}}$, which, in large dimensions, follow the Marchenko-Pastur law; (ii) Considering the distribution of singular vectors $\mathbf{U}$ and $\mathbf{V}$, which tend to be uniformly distributed over appropriate spheres. Explicit calculation of $\mathbb{E}_{\tilde{\mathbf{X}},\boldsymbol{\xi}}\left[\|\tilde{\mathbf{X}}\hat{\mathbf{b}} - \mathbf{y}^*\|_2^2\right]$ requires integrating the above expression with respect to the joint distribution of singular values and vectors, which is complex. In high-dimensional asymptotics, one typically replaces sums over singular values with integrals against the Marchenko-Pastur density and assumes uniformity in the projections $(\mathbf{V}^\top\mathbf{b}^*)_i^2$, but this does not generally yield a closed-form expression. Instead, one uses approximations or numerical simulations to understand behavior under these conditions.

So $\hat{\mathbf{b}}$ can memorize. But can it generalize? We have

$$\begin{aligned}
\hat{\mathbf{b}} - \mathbf{b}^* &= \begin{cases} \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}\right) - \mathbf{b}^* \\ \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\left(\Sigma\mathbf{V}^\top\mathbf{b}^* + \Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right) - \mathbf{b}^* \end{cases} \\
&= \begin{cases} \left[\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \mathbb{I}_n\right]\mathbf{b}^* + \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger \tilde{\mathbf{X}}^\top\boldsymbol{\xi} \\ \left[\mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\Sigma\mathbf{V}^\top - \mathbb{I}_n\right]\mathbf{b}^* + \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi} \end{cases}
\end{aligned}$$

Theorem C.10. *For $N < n$,*

$$\|\hat{\mathbf{b}} - \mathbf{b}^*\|_2^2 \ge \|(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top)\mathbf{b}^*\|_2^2 \tag{68}$$

In particular, if $\mathbf{b}^*$ has a nonzero component orthogonal to $\mathrm{Col}(\mathbf{V})$, then $\hat{\mathbf{b}}$ cannot perfectly generalize to $\mathbf{b}^*$.

Proof. Consider the regularized least-squares estimator

$$\begin{aligned}
\hat{\mathbf{b}} &= \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger \tilde{\mathbf{X}}^\top\mathbf{y}^* \\
&= \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{U}^\top\mathbf{y}^*
\end{aligned} \tag{69}$$

We have $\mathbf{V}\mathbf{V}^\top\hat{\mathbf{b}} = \hat{\mathbf{b}}$, i.e. $\hat{\mathbf{b}} \in \mathrm{Col}(\mathbf{V})$. Let decompose $\mathbf{b}^*$ into two orthogonal components:

$$\mathbf{b}^* = \mathbf{V}\mathbf{V}^\top\mathbf{b}^* + (\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top)\mathbf{b}^* = \mathbf{b}_\parallel + \mathbf{b}_\perp, \tag{70}$$

where

$$\mathbf{b}_{\|} := \mathbf{V}\mathbf{V}^\top \mathbf{b}^* \in \mathrm{Col}(\mathbf{V}), \quad \text{and} \quad \mathbf{b}_{\perp} := (\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top)\mathbf{b}^* \in \mathrm{Col}(\mathbf{V})^{\perp} \tag{71}$$

Since $\hat{\mathbf{b}} \in \mathrm{Col}(\mathbf{V})$,

$$\mathbf{V}\mathbf{V}^\top(\hat{\mathbf{b}} - \mathbf{b}_{\|}) = \hat{\mathbf{b}} - \mathbf{b}_{\|} \tag{72}$$

and $\mathbf{V}\mathbf{V}^\top \mathbf{b}_{\perp} = 0$ by orthogonality. Thus, we can express the error as

$$\begin{aligned} \hat{\mathbf{b}} - \mathbf{b}^* &= \hat{\mathbf{b}} - (\mathbf{b}_{\|} + \mathbf{b}_{\perp}) \\ &= (\hat{\mathbf{b}} - \mathbf{b}_{\|}) - \mathbf{b}_{\perp} \end{aligned} \tag{73}$$

Because $\hat{\mathbf{b}} - \mathbf{b}_{\|} \in \mathrm{Col}(\mathbf{V})$ and $\mathbf{b}_{\perp}$ lies in the orthogonal complement of $\mathrm{Col}(\mathbf{V})$, these two vectors are orthogonal. Hence,

$$\begin{aligned} \|\hat{\mathbf{b}} - \mathbf{b}^*\|_2^2 &= \|\hat{\mathbf{b}} - \mathbf{b}_{\|}\|_2^2 + \|\mathbf{b}_{\perp}\|_2^2 \\ &\geq \|\mathbf{b}_{\perp}\|_2^2 \\ &= \|(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top)\mathbf{b}^*\|_2^2. \end{aligned} \tag{74}$$

□

The theorem above shows that unless $(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top)\mathbf{b}^* = 0$, i.e., $\mathbf{b}^* \in \mathrm{Col}(\mathbf{V})$, the error $\|\hat{\mathbf{b}} - \mathbf{b}^*\|_2$ remains strictly positive. For $N < n$, $\mathbf{V}$ has rank $r < n$, so in general $\mathbf{b}^*$ will have a nonzero orthogonal component $\mathbf{b}_{\perp}$, implying that $\hat{\mathbf{b}}$ cannot fully generalize to $\mathbf{b}^*$

For $\beta_2 = 0$ (i.e., no ℓ_2 regularization), $\hat{\mathbf{b}} = \mathbf{V}\mathbf{V}^\top \mathbf{b}^* + \mathbf{V}\Sigma^{-\frac{1}{2}}\mathbf{U}^\top \boldsymbol{\xi}$. This solution memorizes the training data since $\mathbf{y}(\hat{\mathbf{b}}) - \mathbf{y}^* = \left(\mathbf{U}\mathbf{U}^\top - \mathbb{I}_N\right)\boldsymbol{\xi}$, so that $\|\mathbf{y}(\hat{\mathbf{b}}) - \mathbf{y}^*\|_2^2 = \boldsymbol{\xi}\left(\mathbb{I}_N - \mathbf{U}\mathbf{U}^\top\right)\boldsymbol{\xi} \leq \|\boldsymbol{\xi}\|_2^2 \leq \epsilon^2$. We have $\hat{\mathbf{b}} - \mathbf{b}^* = \left(\mathbf{V}\mathbf{V}^\top - \mathbb{I}_n\right)\mathbf{b}^* + \mathbf{V}\Sigma^{-\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}$, so

$$\begin{aligned} \|\hat{\mathbf{b}} - \mathbf{b}^*\|_2^2 &= \mathbf{b}^{*\top}\left(\mathbf{V}\mathbf{V}^\top - \mathbb{I}_n\right)\left(\mathbf{V}\mathbf{V}^\top - \mathbb{I}_n\right)\mathbf{b}^* + 2\mathbf{b}^{*\top}\left(\mathbf{V}\mathbf{V}^\top - \mathbb{I}_n\right)\mathbf{V}\Sigma^{-\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi} + \boldsymbol{\xi}^\top\mathbf{U}\Sigma^{-\frac{1}{2}}\mathbf{V}^\top\mathbf{V}\Sigma^{-\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi} \\ &= \mathbf{b}^{*\top}\left(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top\right)\mathbf{b}^* + \boldsymbol{\xi}^\top\mathbf{U}\Sigma^{-1}\mathbf{U}^\top\boldsymbol{\xi} \end{aligned}$$

For $N < n$, $\tilde{\mathbf{X}}$ is necessary column rank deficient, that is $\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top > 0$. In that case, $\hat{\mathbf{b}}$ can not be generalized, since $\frac{\|\hat{\mathbf{b}} - \mathbf{b}^*\|_2^2}{\|\mathbf{b}^*\|_2^2} \geq 1 + \frac{\boldsymbol{\xi}^\top\mathbf{U}\Sigma^{-1}\mathbf{U}^\top\boldsymbol{\xi}}{\|\mathbf{b}^*\|_2^2}$. For $N \geq n$, $\hat{\mathbf{b}}$ can generalize if $\tilde{\mathbf{X}}$ is full rank (e.g., if $\tau = 0$, i.e. full random Gaussian $\mathbf{X}$, then $\tilde{\mathbf{X}}$ is full rank with high probability), has small condition number $\frac{\sigma_{\max}}{\sigma_{\min}}$, and the signal to noise ratio $\|\mathbf{b}^*\|_2/\sigma_\xi$ is big enough.

C.6.2 GENERALIZATION

We now turn our attention to the generalization delay. Based on the analysis up to Theorem C.8, we now analyze the subsequent "generalization" phase, during which the iterate $\mathbf{b}^{(t)}$ transitions from memorizing the training data ($\mathbf{b}^{(t)} \approx \hat{\mathbf{b}}$) to converging toward the sparse ground truth $\mathbf{b}^*$. We focus on quantifying the additional number of iterations Δt required for this phase and bounding the generalization error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty$ as $t \to \infty$.

Lemma C.11. *Given $\alpha > 0$ and $b^{(1)} \in \mathbb{R}$, let $b^{(t+1)} = b^{(t)} - \alpha h(b^{(t)})$ for all $t \geq 1$, where $h(b) \in \partial|b|$.*

1. *A point b is stationary for this dynamical system if and only if $|b| \leq \alpha$.*
2. *We have $|b^{(t)}| \leq \alpha$ if and only if $t > \lfloor \frac{|b^{(1)}|}{\alpha} \rfloor$.*
3. *In particular, for $h(b) = \mathrm{sign}(b)$ $\forall b \in \mathbb{R}$, if $b^{(1)}/\alpha \in \mathbb{Z}$, then $b^{(t)} = 0$ for all $t > \lfloor \frac{|b^{(1)}|}{\alpha} \rfloor$.*

Proof. Let first consider the simple case $h(b) = \mathrm{sign}(b)$, so that $b^{(t+1)} = b^{(t)} - \alpha\,\mathrm{sign}(b^{(t)})$.

- If $b^{(t)} \in \{0, \alpha, -\alpha\}$, then $b^{(t+\Delta)} = 0$ for all $\Delta > 0$.
- If $b^{(t)} \in (0, \alpha)$, then $b^{(t+1)} = b^{(t)} - \alpha \in (-\alpha, 0)$, and $b^{(t+2)} = b^{(t+1)} + \alpha = b^{(t)} \in (0, \alpha)$, and so on.
- If $b^{(t)} \in (-\alpha, 0)$, then $b^{(t+1)} = b^{(t)} + \alpha \in (0, \alpha)$, and $b^{(t+2)} = b^{(t+1)} - \alpha = b^{(t)} \in (-\alpha, 0)$, and so on.
- If $b^{(t)} > \alpha$ (resp. $b^{(t)} < -\alpha$), it will be decreased (resp. increase) by α until $b^{(t)} \in (0, \alpha]$ (resp. $b^{(t)} \in [-\alpha, 0)$), and we get back to the previous cases. In that case, $|b^{(t+1)}| = |b^{(t)}| - \alpha = |b^{(1)}| - t\alpha \leq \alpha \Longrightarrow t + 1 \geq \frac{|b^{(1)}|}{\alpha}$.

Let $k = \lfloor \frac{|b^{(1)}|}{\alpha} \rfloor$. Assume $b^{(1)} \geq 0$, then $k\alpha \leq b^{(1)} < (k+1)\alpha$, so that $(k - t + 1)\alpha \leq b^{(t)} < (k - t + 2)\alpha$. Letting $k - t + 1 = 0$, we obtain $t = k + 1$ and $0 \leq b^{(t)} < \alpha$, so that $|b^{(t+\Delta)}| < \alpha$ for all $\Delta > 0$. If $b^{(1)} \leq 0$, then $-(k+1)\alpha < b^{(1)} \leq -k\alpha$, so that $(t - k - 2)\alpha < b^{(t)} \leq (t - k - 1)\alpha$. Letting $t - k - 1 = 0$, we obtain $t = k + 1$ and $-\alpha < b^{(t)} \leq 0$, so that $|b^{(t+\Delta)}| < \alpha$ for all $\Delta > 0$. This achieves the proof for $h(b) = \text{sign}(b)$.

Now consider the general dynamic $b^{(t+1)} = b^{(t)} - \alpha h(b^{(t)})$. If $b^{(1)} \neq 0$ (the case $b^{(1)} = 0$ is trivial), then the dynamic is $b^{(t+1)} = b^{(t)} - \alpha \,\text{sign}(b^{(t)})$ as long as $|b^{(t)}| \geq \alpha$, after which it will just oscillate in the ball $\{b, |b| \leq \alpha\}$ indefinitely. In fact, a fixed point b must satisfy $b = b - \alpha h(b)$; i.e. $h(b) = 0$. The only case where $0 \in \partial|b|$ is $b = 0$ or when it lies in the interval where the subgradient can be 0. However, for any b such that $|b| \leq \alpha$, it is possible to choose $h(b)$ (for instance, $h(b) = b/\alpha$) such that $b = b - \alpha h(b)$, making b a fixed point. Conversely, if $|b| > \alpha$, then $|h(b)| = 1$ and $|b - \alpha h(b)| = ||b| - \alpha| > 0$, so b is not a fixed point.

In practice, we work with the subgradient $h(b) = \text{sign}(b)$, the one provided by automatic differentiation in many optimization libraries, like Pytorch. □

Theorem C.12. *Given $\alpha > 0$ and $\mathbf{b}^{(1)} \in \mathbb{R}^n$, let $\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha h(\mathbf{b}^{(t)})$ for all $t \geq 1$, where $h(\mathbf{b}) \in \partial\|\mathbf{b}\|_1$.*

1. *A point $\mathbf{b}$ is stationary for this dynamical system if and only if $|\mathbf{b}_i| \leq \alpha \; \forall i \in [n]$. As a consequence, $\|\mathbf{b}\|_p \leq \alpha n^{1/p} \; \forall p \in [1, \infty]$.*
2. *We have $\|\mathbf{b}^{(t)}\|_\infty \leq \alpha$ if and only if $t > \lfloor \frac{\|\mathbf{b}^{(1)}\|_\infty}{\alpha} \rfloor$.*
3. *In particular, for $h(\mathbf{b}) = \text{sign}(\mathbf{b}) \; \forall \mathbf{b} \in \mathbb{R}^n$, we have $\|\mathbf{b}^{(t)}\|_0 = \left| \left\{ i \mid \mathbf{b}_i^{(1)}/\alpha \in \mathbb{Z} \right\} \right|$ for all $t > \lfloor \frac{\|\mathbf{b}^{(1)}\|_\infty}{\alpha} \rfloor$.*

Proof. By applying the Lemma C.11 coordinate wise the proof is immediat. □

Recall we have

$$\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha \left(G_{\beta_2}(\mathbf{b}^{(t)}) + \beta_1 h(\mathbf{b}^{(t)}) \right) \tag{75}$$

with

$$G_{\beta_2}(\mathbf{b}) = \tilde{\mathbf{X}}^\top(\mathbf{y} - \mathbf{y}^*) + \beta_2 \mathbf{b} = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \beta_2 \mathbb{I}_n \right) \mathbf{b} - \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{b}^* + \tilde{\mathbf{X}}^\top \boldsymbol{\xi} \right) = \beta_2 \mathbf{b}^* - \tilde{\mathbf{X}}^\top \boldsymbol{\xi} \text{ for } \mathbf{b} = \mathbf{b}^* \tag{76}$$

and $h(\mathbf{b}) \in \partial\|\mathbf{b}\|_1$. From Theorem C.8, for all $t \geq t_1 = \left\lceil -\frac{\ln\left(1 + \frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_p}{\alpha\beta_1 n^{1/p}}\right)}{\ln(\rho_p)} \right\rceil$, and for all p satisfying $\rho_p \in (0, 1)$ (e.g $p = 2$); we have $\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p \leq 2\alpha\beta_1 n^{1/p} \frac{1 - \rho_p^t}{1 - \rho_p} \leq \frac{2\alpha\beta_1 n^{1/p}}{1 - \rho_p}$, where

$\hat{\mathbf{b}} = \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)^\dagger \tilde{\mathbf{X}}^\top\mathbf{y}^*$ is the the least square solution of the problem. So

$$\begin{aligned}\|G_{\beta_2}(\mathbf{b}^{(t)})\|_p &= \|G_{\beta_2}(\mathbf{b}^{(t)}) - G_{\beta_2}(\hat{\mathbf{b}})\|_p \text{ since } G_{\beta_2}(\hat{\mathbf{b}}) = 0\\ &\leq \|\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\|_{p\to p}\|\mathbf{b}^{(t)} - \hat{\mathbf{b}}\|_p\\ &\leq 2\alpha\beta_1 n^{1/p}\|\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\|_{p\to p}\frac{1-\rho_p^t}{1-\rho_p}\\ &\leq \frac{2\alpha\beta_1 n^{1/p}}{1-\rho_p}\|\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\|_{p\to p}\end{aligned} \tag{77}$$

So, this gradient can be made much smaller than the subgradient term by choosing $\alpha\beta_1$ sufficiently small. This bound also writes

$$\begin{aligned}\|G_{\beta_2}(\mathbf{b}^{(t)})\|_2 &\leq 2\alpha\beta_1\sqrt{n}(\sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2)\frac{1-\rho_2^t}{1-\rho_2} \leq \frac{2\alpha\beta_1\sqrt{n}}{1-\rho_2}(\sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2)\\ &\leq 2\beta_1\sqrt{n}(1-\rho_2^t)\frac{\sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2}{\sigma_{\min}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2} \leq 2\beta_1\sqrt{n}\frac{\sigma_{\max} + \beta_2}{\sigma_{\min} + \beta_2} \text{ if } \tilde{\mathbf{X}} \text{ is full rank}\end{aligned} \tag{78}$$

The last line follows from the fact that if $\tilde{\mathbf{X}}$ is full rank, then $\rho_2 = 1 - \alpha(\sigma_{\min}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2)$, so that $1 - \rho_2 = \alpha(\sigma_{\min}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2)$.

Let $I := \{i \in [n] \mid b_i^* \neq 0\}$ be the support of $\mathbf{b}^*$. Since $\mathbf{b}^*$ is s-sparse, $s = |I| \ll n$. After time t_1, the contribution of the gradient G_{β_2} to the update of $\mathbf{b}_i^{(t)}$ is dominated by the ℓ_1–regularization term. Specifically, for each $i \in [n]$, the update rule approximates

$$\mathbf{b}_i^{(t+1)} \approx \mathbf{b}_i^{(t)} - \alpha\beta_1 h(\mathbf{b}_i^{(t)}) \tag{79}$$

By Theorem C.12, this lead to $\|\mathbf{b}^{(t)}\|_p \leq \alpha n^{1/p}\ \forall p \in [1,\infty]$ for (and only for) $t \geq t_2 := t_1 + \left\lfloor \frac{\|\mathbf{b}^{(1)}\|_\infty}{\alpha\beta_1} \right\rfloor$.

For $i \in I$ in particular, if $|\mathbf{b}_i^{(t_1)}| \gg |\mathbf{b}_i^*|$, then using the approximate dynamics $\mathbf{b}_i^{(t+1)} \approx \mathbf{b}_i^{(t)} - \alpha\beta_1 h(\mathbf{b}_i^{(t)} - \mathbf{b}_i^*)$, we can conclude also that $|\mathbf{b}_i^{(t)} - \mathbf{b}_i^*| \leq \alpha\beta_1$ for (and only for) $t \geq t_2$.

Note that when $\|\mathbf{b}^{(t)}\|_1$ becomes too small, $\mathbf{b}^{(t)} \approx \mathbf{b}^*$ since for problem of interest, the sparse solution $\mathbf{b}^*$ is the unique minimizer of $\|\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*\|_2$ under the sparsity constraint $s = \|\mathbf{b}^*\|_0 \ll n$ (and the RIP assumptions on $\tilde{\mathbf{X}}$). Our argument here is that the additional number of steps it takes to reach this small ℓ_1-norm solution is $\Delta t = \Theta\left(\frac{\|\hat{\mathbf{b}}\|_\infty}{\alpha\beta_1}\right)$, so that the smaller β_1 is (for α fixed), the longer it take to recover $\mathbf{b}^*$, and the smaller is the error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty$ when $t \to \infty$. If β_2 is choose such that $\|\hat{\mathbf{b}}\|_\infty \ll \alpha\beta_1$, then $\mathbf{b}^{(t)}$ will get stuck near $\hat{\mathbf{b}}$, and there will be no generalization after memorization. So a bad choice of a non-zero β_2 can be detrimental to generalization (it is better to not use β_2 on that problem unless the initialization scale is nontrivial).

By carefully choosing α and β_1, one can balance the speed of generalization (smaller Δt) with the accuracy of recovery (smaller $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty$). Appropriate step rule also guaranteed the converge of $\|\mathbf{b}^{(t)}\|_1$ to $\|\mathbf{b}^*\|_1$.

Theorem C.13. *For all $T \in \mathbb{N}^*$, we have*

$$\min_{1\leq t\leq T}\left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) \leq \frac{\|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2^2 + \left(\max_{1\leq t\leq T}\|\nabla_{\mathbf{b}} f(\mathbf{b}^{(t)})\|_2^2\right)\sum_{t=1}^T \alpha_t^2}{2\beta_1\sum_{t=1}^T \alpha_t} + \frac{\|\boldsymbol{\xi}\|_2^2 + \beta_2\|\mathbf{b}^*\|_2^2}{2\beta_1}. \tag{80}$$

Proof. We have $f(\mathbf{b}^{(t)}) = \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}^{(t)}\|_2^2 + \beta_1\|\mathbf{b}^{(t)}\|_1$ and $f(\mathbf{b}^*) = \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b}^* - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}^*\|_2^2 + \beta_1\|\mathbf{b}^*\|_1 = \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 + \frac{\beta_2}{2}\|\mathbf{b}^*\|_2^2 + \beta_1\|\mathbf{b}^*\|_1$. So for any t,

$$\begin{aligned}f(\mathbf{b}^{(t)}) - f(\mathbf{b}^*) &= \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\left(\|\mathbf{b}^{(t)}\|_2^2 - \|\mathbf{b}^*\|_2^2\right) + \beta_1\left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) - \frac{1}{2}\|\boldsymbol{\xi}\|_2^2\\ &\geq \beta_1\left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) - \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 - \frac{\beta_2}{2}\|\mathbf{b}^*\|_2^2\end{aligned} \tag{81}$$

Since

$$\frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2^2 \geq 0 \quad \text{and} \quad \frac{\beta_2}{2}\left(\|\mathbf{b}^{(t)}\|_2^2 - \|\mathbf{b}^*\|_2^2\right) \geq -\frac{\beta_2}{2}\|\mathbf{b}^*\|_2^2 \tag{82}$$

Rearranging equation equation 81 yields

$$\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1 \leq \frac{f(\mathbf{b}^{(t)}) - f(\mathbf{b}^*)}{\beta_1} + \frac{\|\boldsymbol{\xi}\|_2^2 + \beta_2\|\mathbf{b}^*\|_2^2}{2\beta_1}. \tag{83}$$

By Theorem C.3, when $\|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2 \leq R$ and $\|F(\mathbf{b}^{(t)})\|_2 \leq L \quad \forall t \leq T$,

$$\min_{1\leq t\leq T}\left(f(\mathbf{b}^{(t)}) - f(\mathbf{b}^*)\right) \leq \frac{R^2 + L^2\sum_{t=1}^T \alpha_t^2}{2\sum_{t=1}^T \alpha_t} \tag{84}$$

Substituting this into equation 83 gives

$$\min_{1\leq t\leq T}\left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) \leq \frac{R^2 + L^2\sum_{t=1}^T \alpha_t^2}{2\beta_1\sum_{t=1}^T \alpha_t} + \frac{\|\boldsymbol{\xi}\|_2^2 + \beta_2\|\mathbf{b}^*\|_2^2}{2\beta_1}. \tag{85}$$

□

So, when $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$ (e.g. $\alpha_t = a/(b+t)$, $a > 0$ and $b \geq 0$), $\|\mathbf{b}^{(t)}\|_1 \to \|\mathbf{b}^*\|_1 \to 0$ as $T \to \infty$, for $\beta_2 = 0$ in the noiseless setting.

C.6.3 Optimization landscape

We will look at the landscape of the solution. Let $I := \{i \in [n] \mid b_i^* \neq 0\}$ be the support of $\mathbf{b}^*$; $u(t) = \|\mathbf{b}_I^{(t)}\|_2$ and $v(t) = \|\mathbf{b}_{[n]\setminus I}^{(t)}\|_2$ be the norms of $\mathbf{b}^{(t)}$ restraint on its indexes in I (resp, outside I).

Figure 10 shows how $\mathbf{b}^{(t)}$ first converge to the least square solution (memorization), and from least square solution to $\mathbf{b}^*$ (N large enough) or a suboptimal solution (N too small). After memorization, when N is large enough, $v(t)$ converge to zero while $u(t)$ converge to the norm of $\mathbf{b}^*$. This is because the components of $\mathbf{b}^{(t)}$ that are not in I are shrunk at each training step until they all reach 0 (Figure 11). This convergence is impossible if $\beta_1 = 0$ (even if $\beta_2 \neq 0$).

C.6.4 Additionnal experiments

We optimize the noiseless problem ($\boldsymbol{\xi} = 0$) using the subgradient descent method with $(n, s, N, \zeta, \beta_2) = (10^2, 5, 30, 10^{-6}, 0)$ for different values of α and β_1. As expected, larger α and/or β_1 lead to fast convergence and do so at a suboptimal value of the test error (Figure 12).

We optimize the noiseless problem ($\boldsymbol{\xi} = 0$) using the subgradient descent method with $(n, \zeta, \alpha, \beta_1, \beta_2) = (10^2, 10^{-6}, 10^{-1}, 10^{-5}, 0)$, for different values of s and N. See Figures 13, 14, 15 and 16).

C.7 Projected subgradient

To ensure memorization, we can use the projected subgradient for problem (P_1) of minimizing $\|\mathbf{b}\|_1$ subject to the constraint $\mathcal{F}_{\mathbf{b}}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\mathbf{b} = \mathbf{y}^*$, where at each step the update (using now just $\beta_1 h(\mathbf{b})$ as gradient, not the whole $F(\mathbf{b})$) is projected onto the constraint set. In our case, the update write $\mathbf{b}^{(t+1)} = \Pi\left(\mathbf{b}^{(t)} - \alpha_t\beta_1 h(\mathbf{b}^{(t)})\right)$ with $\Pi(\mathbf{b}) = \mathbf{b} - \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^\dagger(\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^*) = \mathbf{P}\left(\mathbf{b} - \mathbf{b}^*\right) + \mathbf{b}^* + \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^\dagger\boldsymbol{\xi}$ the projection of $\mathbf{b}$ on the set $\{\mathbf{b}, \tilde{\mathbf{X}}\mathbf{b} = \mathbf{y}^*\}$, $\mathbf{P} = \mathbb{I}_n - \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^{-1}\tilde{\mathbf{X}}$. So $\mathbf{b}^{(t+1)} - \mathbf{b}^* = \mathbf{P}\left(\mathbf{b}^{(t)} - \mathbf{b}^*\right) - \alpha_t\beta_2\mathbf{P}h(\mathbf{b}^{(t)}) + \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^\dagger\boldsymbol{\xi}$[3]. We can also keep

[3]For a fat and full rank $\tilde{\mathbf{X}}$ ($\text{rank}(\tilde{\mathbf{X}}) = N \leq n$), if we start at $\mathbf{b}^{(1)}$ such that $\tilde{\mathbf{X}}\mathbf{b}^{(1)} = \mathbf{y}^*$, for example, the min norm solution $\mathbf{b}^{(1)} = \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^{-1}\mathbf{y}^*$, then $\mathbf{P}\left(\mathbf{b}^{(t)} - \mathbf{b}^*\right) = \mathbf{b}^{(t)} - \mathbf{b}^* - \tilde{\mathbf{X}}^\top\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)^\dagger\boldsymbol{\xi}\ \forall t \geq 1$,

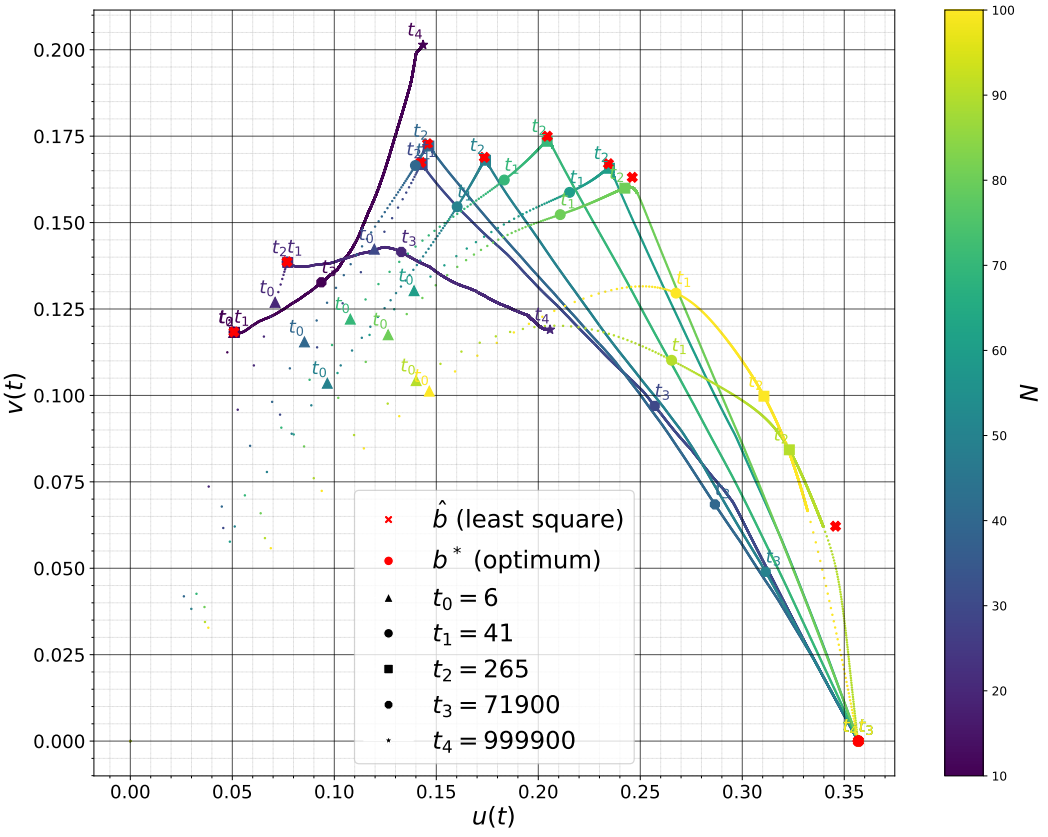


Figure 10: From initialization to least square solution (memorization), and from least square solution to $\mathbf{b}^*$ (N large enough) or a suboptimal solution (N too small). The steps t_1 and t_2 are different from those introduced above to measure memorization and generalization (respectively). They are just a means of tracing the evolution of training here. Here $N \in \{20, 30, 40, 50, 60, 70\}$, for $(n, s) = (100, 5)$ and $(\alpha, \beta_1, \beta_2) = (10^{-1}, 10^{-5}, 0)$.

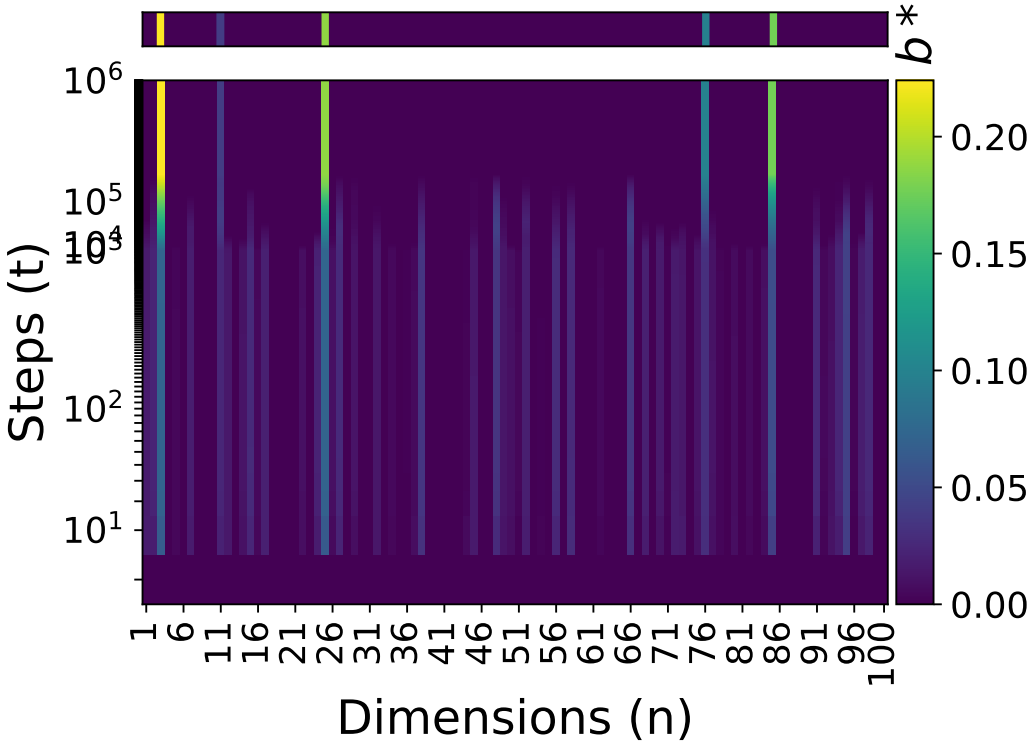


Figure 11: Convergence of $\mathbf{b}_i^{(t)}$ to $\mathbf{b}_i^*$ for each $i \in [n]$. Here $(n, s, N) = (100, 5, 30)$ and $(\alpha, \beta_1, \beta_2) = (10^{-1}, 10^{-5}, 0)$.

track of the best minimum ℓ_1 solution during training, $\mathbf{b}_{\text{best}}^{(t)} = \arg\min_{\mathbf{b} \in \left\{\mathbf{b}^{(t')}, t' \le t\right\}} \|\mathbf{b}\|_1 = \arg\min_{\mathbf{b} \in \left\{\mathbf{b}_{\text{best}}^{(t-1)}, \mathbf{b}^{(t)}\right\}} \|\mathbf{b}\|_1$. Using this, we can show that the ℓ_1 optimal gap of this method enjoys the same bound given above for the non-projected case without the requirement $\|F(\mathbf{b})\|_2 \le L \quad \forall \mathbf{b}$, with the rescale learning rate $\tilde{\alpha}_t = \beta_1 \alpha_t$; and the bound $\sqrt{n}$ on the subgradient, $\|h(\mathbf{b})\|_2^2 \le n \; \forall \mathbf{b}$. Note that we have $f^* = f(\mathbf{b}^*) = \beta_1 \|\mathbf{b}^*\|_1 + \frac{\beta_2}{2}\|\mathbf{b}^*\|_2^2 + \|\boldsymbol{\xi}\|_2^2$, and after one step of training ($t > 1$), $f^{(t)} = f(\mathbf{b}^{(t)}) = \beta_1 \|\mathbf{b}^{(t)}\|_1 + \frac{\beta_2}{2}\|\mathbf{b}^{(t)}\|_2^2$ since $\mathbf{y}(\mathbf{b}^{(t)}) = \mathbf{y}^*$.

Theorem C.14. *Let $\tilde{\alpha}_t = \beta_1 \alpha_t$. If $\|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2 \le R$, then $\|\mathbf{b}_{best}^{(t)}\|_1 - \|\mathbf{b}^*\|_1 \le \frac{R^2 + n\sum_{t=1}^T \tilde{\alpha}_t^2}{2\sum_{t=1}^T \tilde{\alpha}_t}$.*

Proof. We have

$$
\begin{aligned}
0 \le \|\mathbf{b}^{(T+1)} - \mathbf{b}^*\|_2^2 &= \|\Pi\left(\mathbf{b}^{(T)} - \alpha_T \beta_1 \cdot h(\mathbf{b}^{(T)})\right) - \mathbf{b}^*\|_2^2 \\
&\le \|\mathbf{b}^{(T)} - \mathbf{b}^* - \alpha_T \beta_1 \cdot h(\mathbf{b}^{(T)})\|_2^2 \\
&= \|\mathbf{b}^{(T)} - \mathbf{b}^*\|_2^2 - 2\alpha_T \beta_1 (\mathbf{b}^{(T)} - \mathbf{b}^*)^\top h(\mathbf{b}^{(T)}) + \beta_1^2 \alpha_T^2 \|h(\mathbf{b}^{(T)})\|_2^2 \\
&\le \|\mathbf{b}^{(T)} - \mathbf{b}^*\|_2^2 - 2\beta_1 \alpha_T \left(\|\mathbf{b}^{(T)}\|_1 - \|\mathbf{b}^*\|_1\right) + \beta_1^2 \alpha_T^2 \|h(\mathbf{b}^{(T)})\|_2^2 \text{ (by the definition of } h) \\
&\le \|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2^2 - 2\beta_1 \sum_{t=1}^T \alpha_t \left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) + \beta_1^2 \sum_{t=1}^T \alpha_t^2 \|h(\mathbf{b}^{(t)})\|_2^2
\end{aligned}
$$

and the update simplifies to $\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \alpha_t \beta_2 \mathbf{P} h(\mathbf{b}^{(t)})$. In general, even if we don't start at $\mathbf{b}^{(1)}$ satisfying $\tilde{\mathbf{X}}\mathbf{b}^{(1)} = \mathbf{y}^*$, as soon as $\tilde{\mathbf{X}}\mathbf{b}^{(t_0)} = \mathbf{y}^*$ for a certain t_1 (memorization), the next updates have the previous form. Note that $\mathbf{P}^\top = \mathbf{P}$ and $\mathbf{P}^\top \mathbf{P} = \mathbf{P}^2 = \mathbf{P}$.

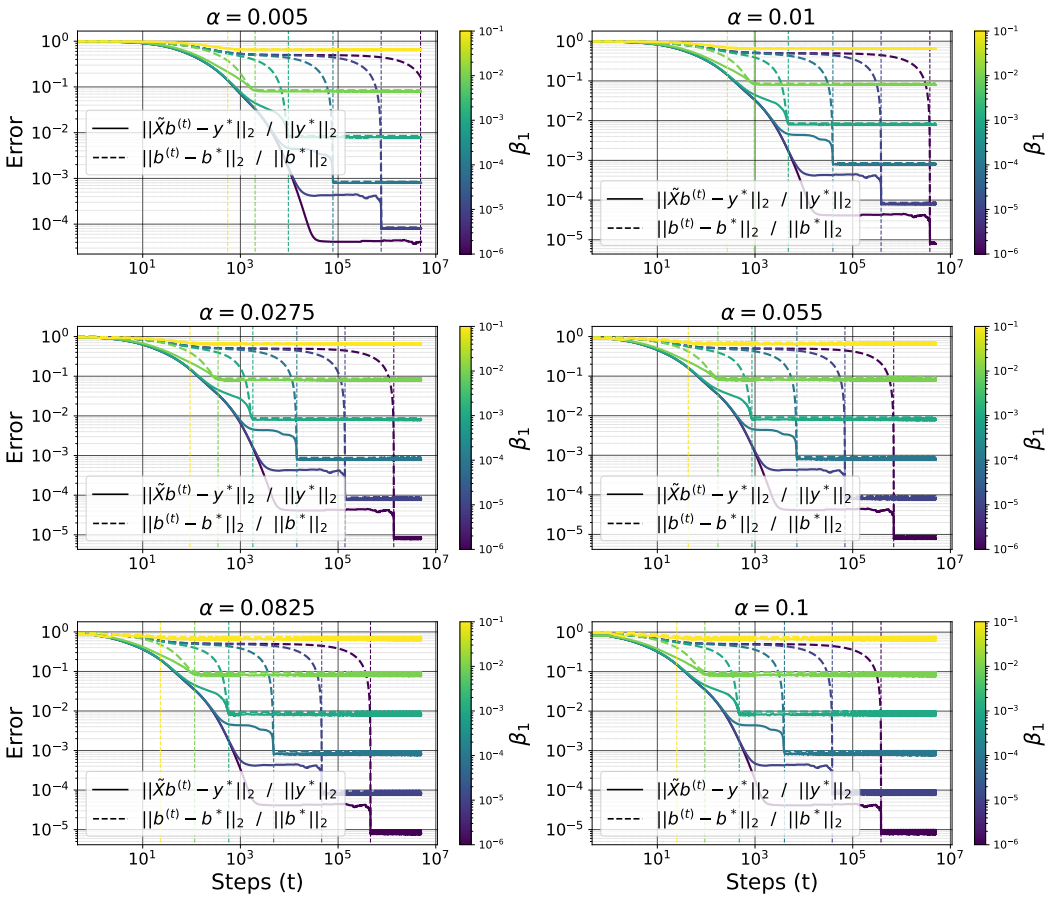


Figure 12: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the learning rate α and the ℓ_1-regularization coefficient β_1. Here $(n, s, N) = (100, 5, 30)$

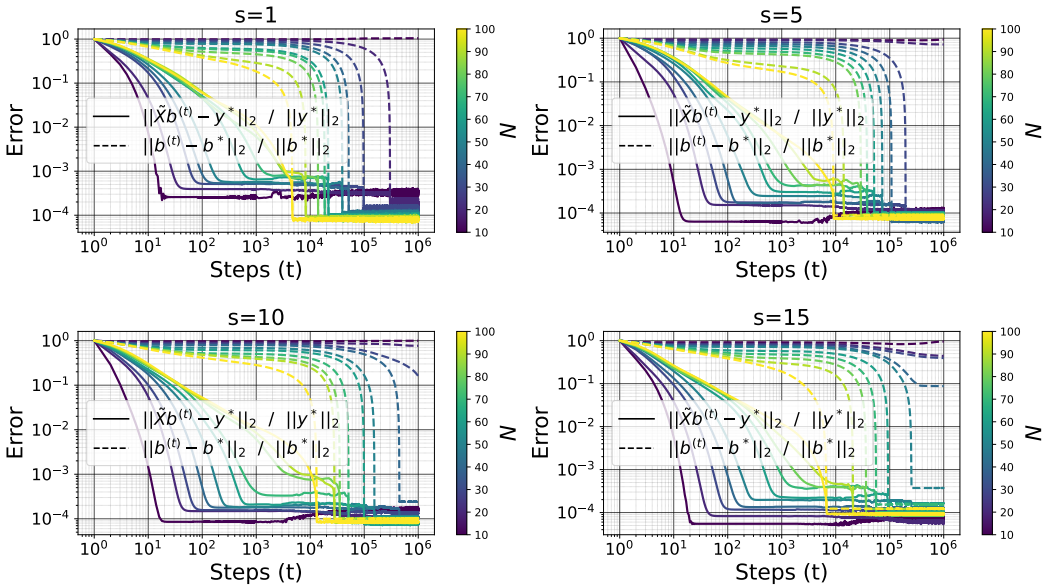


Figure 13: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the sparsity level $s \in \{1, 5, 10, 15\}$ and the measurements $N \in \{10, 20, \ldots 100\}$. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **subgradient descent**

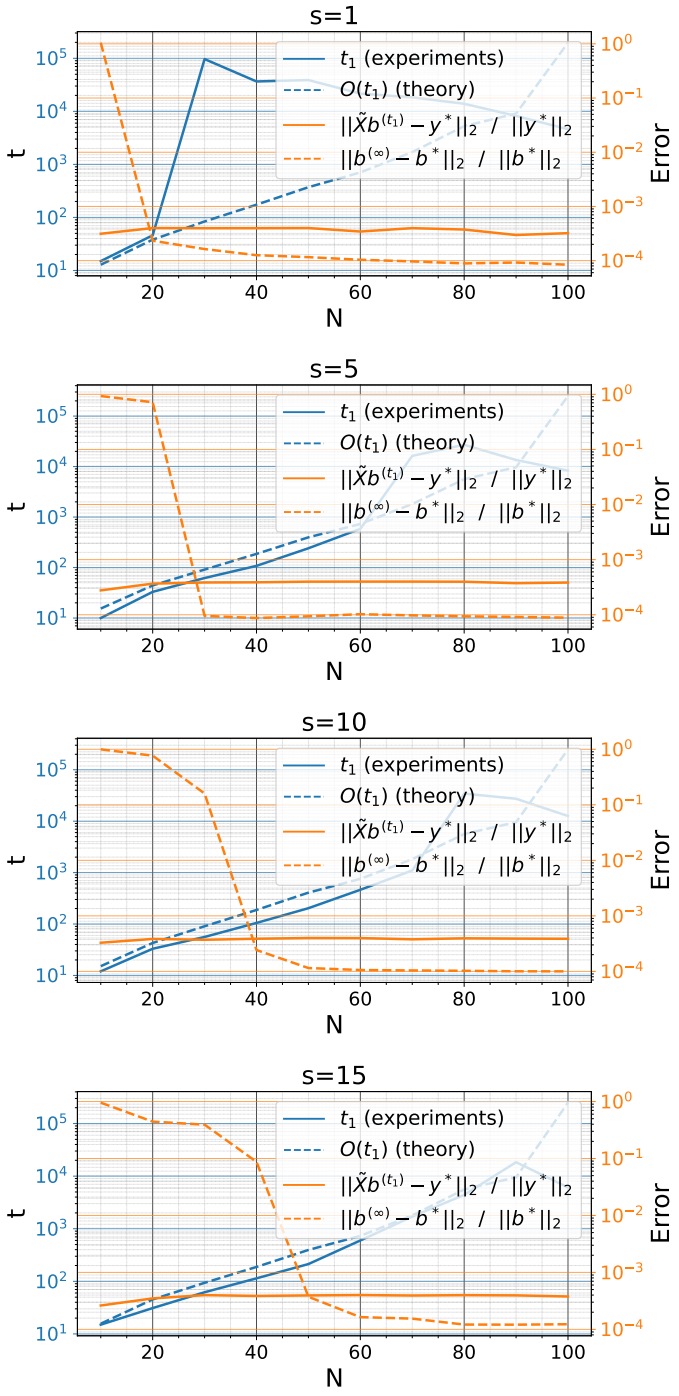


Figure 14: On the left axis, the memorization step t_1 compute experimentally (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$) and the upper bound $-\ln\left(1 + \frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_\infty}{\alpha\beta_1}\right) / \ln(\rho)$ computed in Theorem C.8. On the right axis, the error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ at step t_1 and the recovery error $\|\mathbf{b}^{(\infty)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at the end of training. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **subgradient descent**.

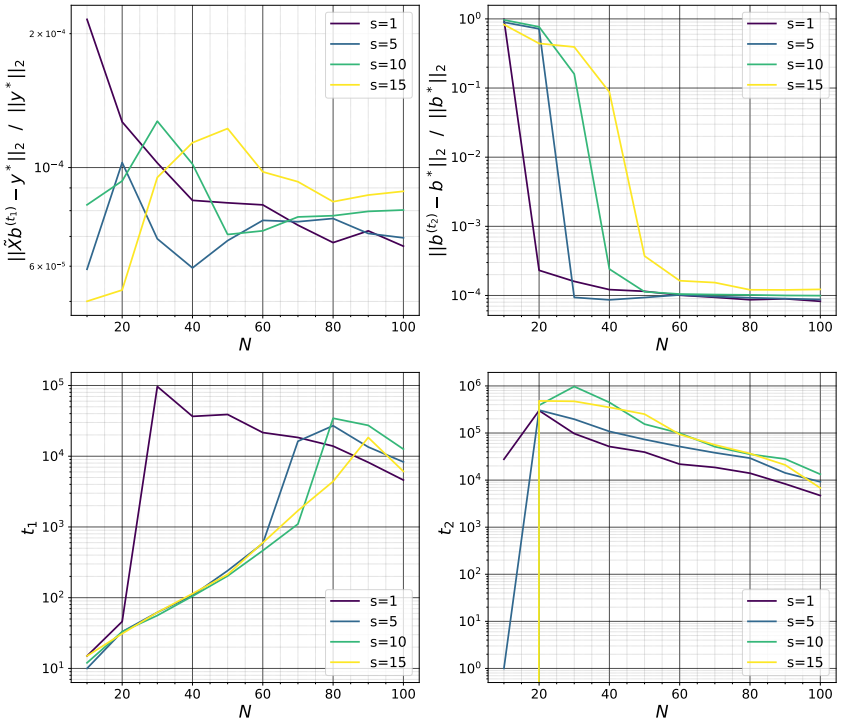


Figure 15: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ at memorization, recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at generalization, memorization step t_1 (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$), and generalization step (smaller t such that $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2 \leq 10^{-4}$ or the maximum training step). Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **subgradient descent**

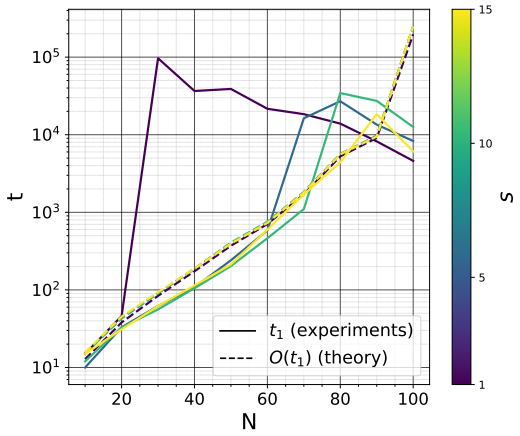


Figure 16: Memorization step t_1 compute experimentally (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$) and the upper bound computed in Theorem C.8. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **subgradient descent**.

$$\implies 2\beta_1 \left(\sum_{t=1}^{T} \alpha_t\right) \min_{t\leq T} \left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) \leq 2\beta_1 \sum_{t=1}^{T} \alpha_t \left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) \leq R^2 + \beta_1^2 \sum_{t=1}^{T} \alpha_t^2 \|h(\mathbf{b}^{(t)})\|_2^2$$

$$\iff \min_{t\leq T} \left(\|\mathbf{b}^{(t)}\|_1 - \|\mathbf{b}^*\|_1\right) \leq \frac{R^2 + \beta_1^2 \sum_{t=1}^{T} \alpha_t^2 \|h(\mathbf{b}^{(t)})\|_2^2}{2\beta_1 \sum_{t=1}^{T} \alpha_t} = \frac{R^2 + \beta_1^2 n \sum_{t=1}^{T} \alpha_t^2}{2\beta_1 \sum_{t=1}^{T} \alpha_t}$$

□

We optimize the noiseless problem ($\boldsymbol{\xi} = 0$) using the projected subgradient descent method with $(n, \zeta, \alpha, \beta_1, \beta_2) = (10^2, 10^{-6}, 10^{-1}, 10^{-5}, 0)$, for different values of s and N. We observe a grokking-like pattern similar to the subgradient case (Figures 17, 18, 19 and 20). Here, one step of training is enough to get zero training error. This further shows that generalization is driven by β_1.

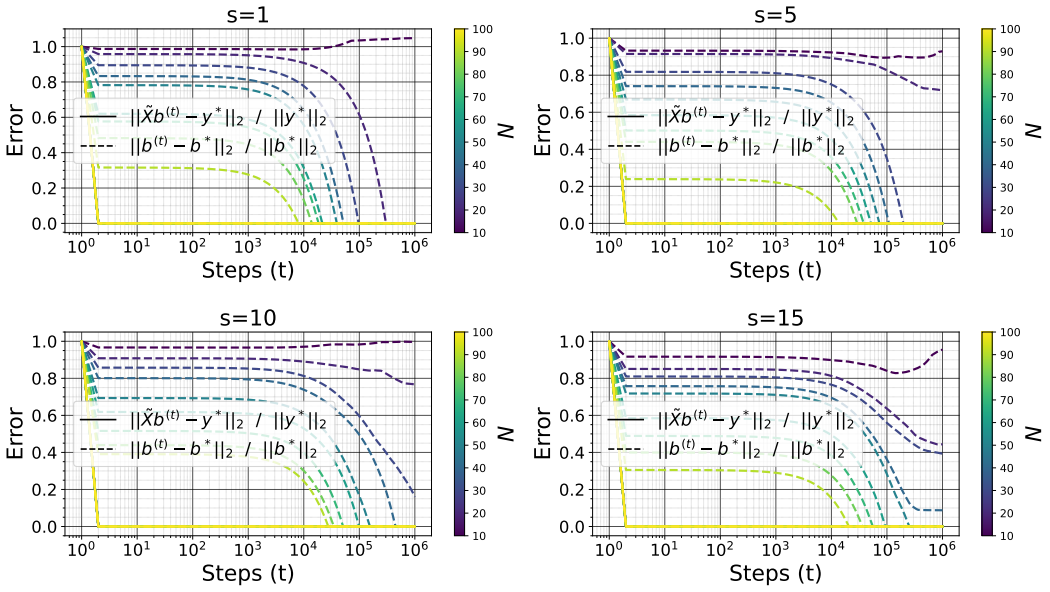


Figure 17: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the sparsity level $s \in \{1, 5, 10, 15\}$ and the measurements $N \in \{10, 20, \dots 100\}$. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **projected subgradient descent**

C.8 Proximal Gradient Descent and Iterative soft-thresholding algorithm

We have

$$\mathbf{b} - \alpha G_{\beta_2}(\mathbf{b}) = \arg\min_{\mathbf{c}} g_{\beta_2}(\mathbf{b}) + (\mathbf{c} - \mathbf{b})^\top G_{\beta_2}(\mathbf{b}) + \frac{1}{2\alpha}\|\mathbf{c} - \mathbf{b}\|_2^2$$

So

$$\begin{aligned}
\mathbf{b} - \alpha F(\mathbf{b}) &\approx \arg\min_{\mathbf{c}} g_{\beta_2}(\mathbf{b}) + (\mathbf{c} - \mathbf{b})^\top G_{\beta_2}(\mathbf{b}) + \frac{1}{2\alpha}\|\mathbf{c} - \mathbf{b}\|_2^2 + \beta_1\|\mathbf{c}\|_1 \\
&= \arg\min_{\mathbf{c}} \frac{1}{2\alpha}\left[\|\alpha G_{\beta_2}(\mathbf{b})\|_2^2 + 2\alpha(\mathbf{c} - \mathbf{b})^\top G_{\beta_2}(\mathbf{b}) + \|\mathbf{c} - \mathbf{b}\|_2^2\right] + \beta_1\|\mathbf{c}\|_1 \\
&= \arg\min_{\mathbf{c}} \frac{1}{2\alpha}\|\mathbf{c} - (\mathbf{b} - \alpha G_{\beta_2}(\mathbf{b}))\|_2^2 + \beta_1\|\mathbf{c}\|_1 \\
&= \Pi_\alpha\left(\mathbf{b} - \alpha G_{\beta_2}(\mathbf{b})\right)
\end{aligned}$$

with Π_α the proximal mapping for $\mathbf{c} \to \beta_1\|\mathbf{c}\|_1$,

$$\Pi_\alpha(\mathbf{b}) = \arg\min_{\mathbf{c}} \frac{1}{2\alpha}\|\mathbf{c} - \mathbf{b}\|_2^2 + \beta_1\|\mathbf{c}\|_1$$

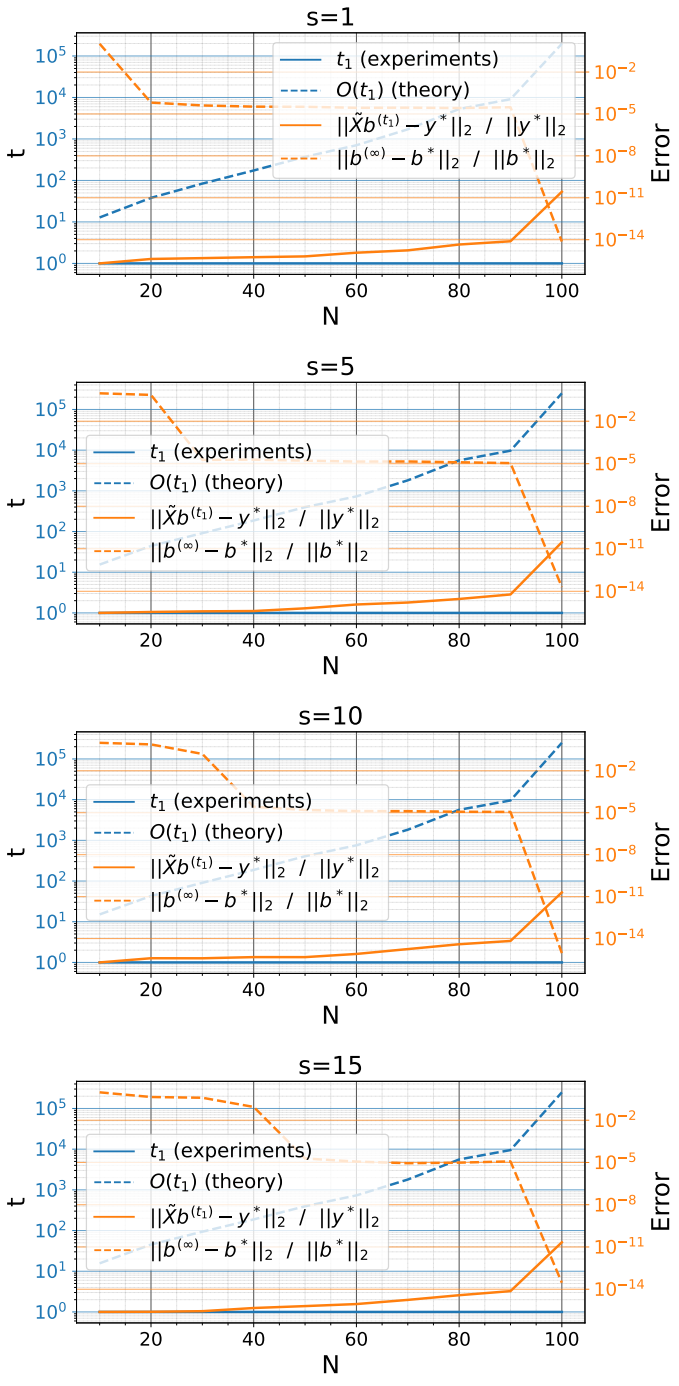


Figure 18: On the left axis, the memorization step t_1 compute experimentally (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$) and the upper bound $-\ln\left(1 + \frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_\infty}{\alpha\beta_1}\right) / \ln(\rho)$ computed in Theorem C.8. On the right axis, the error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ at step t_1 and the recovery error $\|\mathbf{b}^{(\infty)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at the end of training. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **projected subgradient descent**.

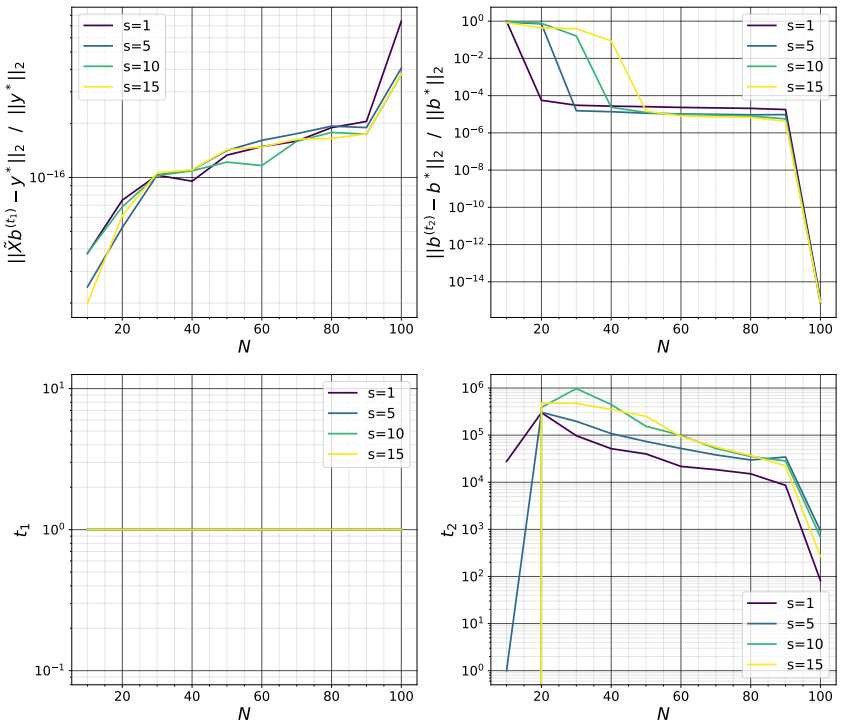


Figure 19: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ at memorization, recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at generalization, memorization step t_1 (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$), and generalization step (smaller t such that $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2 \leq 10^{-4}$ or the maximum training step). Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **projected subgradient descent**

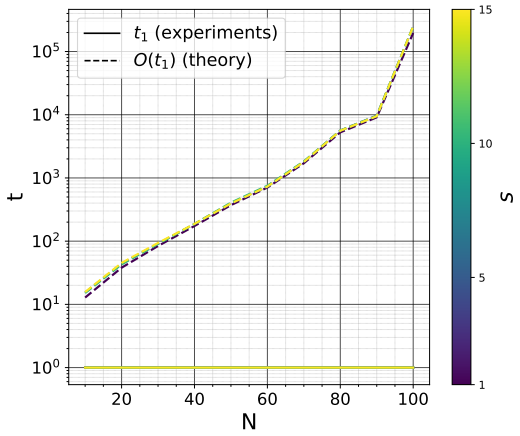


Figure 20: Memorization step t_1 compute experimentally (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$) and the upper bound computed in Theorem C.8. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **projected subgradient descent**.

Using

$$Q_\alpha(\mathbf{b}) = \frac{\mathbf{b} - \Pi_\alpha\left(\mathbf{b} - \alpha G_{\beta_2}(\mathbf{b})\right)}{\alpha}$$

The proximal update writes

$$\begin{aligned}\mathbf{b}^{(t+1)} &= \Pi_{\alpha_t}\left(\mathbf{b}^{(t)} - \alpha_t G_{\beta_2}(\mathbf{b}^{(t)})\right) \\ &= \mathbf{b}^{(t)} - \alpha_t \frac{\mathbf{b}^{(t)} - \Pi_{\alpha_t}\left(\mathbf{b}^{(t)} - \alpha_t G_{\beta_2}(\mathbf{b}^{(t)})\right)}{\alpha_t} \\ &= \mathbf{b}^{(t)} - \alpha_t Q_{\alpha_t}(\mathbf{b}^{(t)})\end{aligned}$$

This form appears similar to the standard gradient descent update but is not the most interesting in this context.

$$\begin{aligned}\Pi_\alpha(\mathbf{b}) &= \arg\min_{\mathbf{c}} \frac{1}{2\alpha}\|\mathbf{c} - \mathbf{b}\|_2^2 + \beta_1\|\mathbf{c}\|_1 \\ &= \arg\min_{\mathbf{c}} \frac{1}{2}\|\mathbf{c} - \mathbf{b}\|_2^2 + \alpha\beta_1\|\mathbf{c}\|_1 \\ &= S_{\alpha\beta_1}(\mathbf{b})\end{aligned}$$

with $S_\gamma(\mathbf{b}) = \text{sign}(\mathbf{b}) \odot \max(|\mathbf{b}| - \gamma, 0)$ the soft-thresholding operator[4],

$$S_\gamma(\mathbf{b})_i = \begin{cases} \mathbf{b}_i - \gamma & \text{if } \mathbf{b}_i > \gamma \\ 0 & \text{if } -\gamma \leq \mathbf{b}_i \leq \gamma \\ \mathbf{b}_i + \gamma & \text{if } \mathbf{b}_i < -\gamma \end{cases}$$

The final form of the update, known as the Iterative soft-thresholding algorithm (ISTA) (Daubechies et al., 2003), is then

$$\mathbf{b}^{(t+1)} = S_{\alpha_t\beta_1}\left(\mathbf{b}^{(t)} - \alpha_t G_{\beta_2}(\mathbf{b}^{(t)})\right) \quad \forall t > 1 \tag{86}$$

with

$$G_{\beta_2}(\mathbf{b}) := \nabla_{\mathbf{b}} g_{\beta_2}(\mathbf{b}) = \tilde{\mathbf{X}}^\top(\mathbf{y} - \mathbf{y}^*) + \beta_2\mathbf{b} = \begin{cases} \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)\mathbf{b} - \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b}^* + \tilde{\mathbf{X}}^\top\boldsymbol{\xi}\right) \\ \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\right)(\mathbf{b} - \mathbf{b}^*) - \left(\tilde{\mathbf{X}}^\top\boldsymbol{\xi} - \beta_2\mathbf{b}^*\right) \end{cases}$$

Theorem C.15. *Let $L = \|\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\|_{2\to 2} = \sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2$ (operator norm) be the Lipschitz constant for G_{β_2}, $|G_{\beta_2}(\mathbf{u}) - G_{\beta_2}(\mathbf{v})| \leq L\|\mathbf{u} - \mathbf{v}\|_2$ for all $\mathbf{u}$, $\mathbf{v}$. If $\|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2 \leq R$ and $\alpha_t = \alpha \leq 1/L$, then $f^{(T)} - f^* \leq \frac{R^2}{2\alpha T}$ for the ISTA.*

Proof. We applied a standard bound on proximal gradient descent (Tibshirani, 2015) for a function of the form $f = g + h : \mathbb{R}^n \to \mathbb{R}$. Such result state that the proximal gradient descent with fixed step size $\alpha_t \leq 1/L$ satisfies $f^{(T)} - f^* \leq \frac{\|\mathbf{b}^{(1)} - \mathbf{b}^*\|_2^2}{2\alpha T}$ when g is convex, differentiable, $\text{dom}(g) = \mathbb{R}^n$, ∇g is Lipschitz continuous with constant $L > 0$; and h is convex and its proximal map Π_α can be evaluated. □

We optimize the noiseless problem ($\boldsymbol{\xi} = 0$) using the soft-thresholding algorithm (ISTA) with $(n, \zeta, \alpha, \beta_1, \beta_2) = (10^2, 10^{-6}, 10^{-1}, 10^{-5}, 0)$, for different values of s and N. We observe a grokking-like pattern similar to the subgradient case (Figures 21, 22, 23 and 24).

C.9 Grokking without understanding

We start the optimization at $\mathbf{b}^{(1)} \overset{iid}{\sim} \zeta\mathcal{N}(0, 1/n)$ with $\zeta \geq 0$ the initialization scale. With a small initialization, β_1 is sufficient for generalization to happen, provided N is large enough and β_2 is not very large (if it is chosen so that $\|\hat{\mathbf{b}}\|_\infty \ll \alpha\beta_1$, it may be possible to not generalize, see section C.6.2). If the scale at initialization is large, β_2 is necessary to generalize, but is it sufficient? That is, can we generalize to the problem studied here with $\beta_1 = 0$ and $\beta_2 > 0$?

[4]On complex numbers, the soft-thresholding operator $S_\gamma(\mathbf{b}) = \text{sign}(\mathbf{b}) \odot \max(|\mathbf{b}| - \gamma, 0)$ only shrinks the magnitude and keeps the phase fixed.

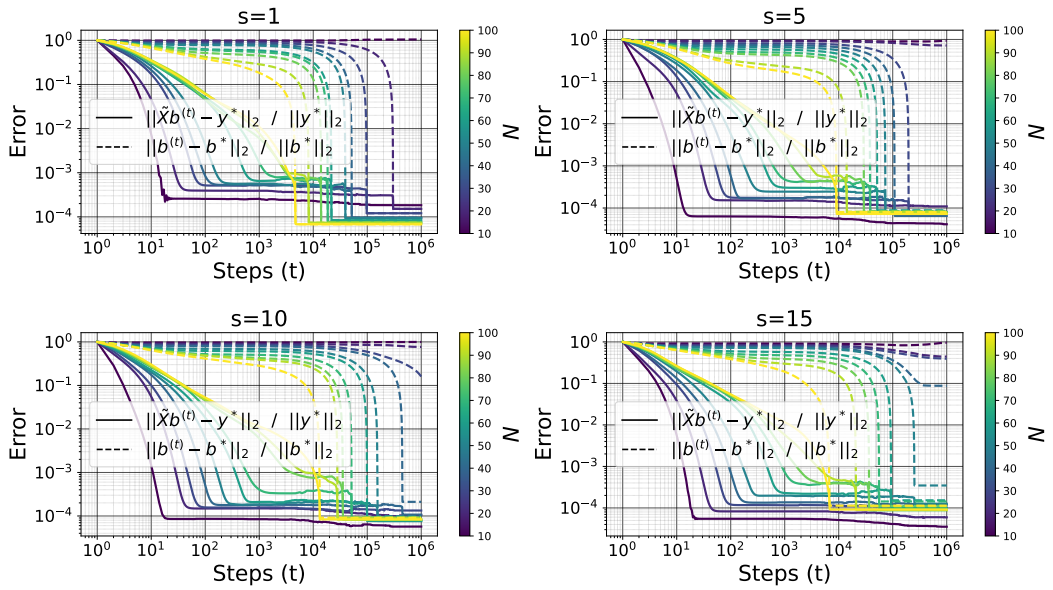


Figure 21: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the sparsity level $s \in \{1, 5, 10, 15\}$ and the measurements $N \in \{10, 20, \dots 100\}$. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **soft-thresholding algorithm (ISTA)**

As shown above, the answer to this question is no (Figures 25 and 26). But what we want to illustrate here is a phenomenon that contradicts previous art (Liu et al., 2023a; Lyu et al., 2023), namely that *in the over-parametrized regime ($N < n$ in our case), large initialization and non-zero weight decay do not always lead to grokking*. What happens is that, because of the large initialization, a more or less abrupt transition is observed in the generalization error during training, corresponding to a transition in the ℓ_2 norm of the model parameters. But this can not be called grokking because the model only converges to a sub-optimal solution. What's more, this transition appears even if the problem posed admits no solution, e.g., sparse recovery or matrix completion with a number N of examples far below the theoretical limit required for the solution to the problem posed to be the optimal solution (by any method whatsoever). This transition appears abrupt just because the training error is large at the beginning of training since the model's outputs are large. When its ℓ_2 norm becomes small, its outputs also become small, leading to a transition in error. In figure 25, without visualization of the error on a logarithmic scale, it looks like grokking has occurred, whereas this is not the case. Figure 26 futher shows the non convergence of $\mathbf{b}^{(t)}$ to $\mathbf{b}^*$: every components of $\mathbf{b}^{(t)}$ are almost 0 at the end of training.

We call this phenomenon "grokking without understanding" like Levi et al. (2024) who illustrated it in the case of linear classification. They show that the sharp increase in generalization accuracy may often not imply a transition from "memorization" to "understanding" but can be an artifact of the accuracy measure. But in our case, we are not using any significant scale at initialization (we focus on $0 \leq \zeta \leq 10^{-5}$) and are not dealing with the generalization measure problem since our test error is directly the recovery error in the function space, not the accuracy.

We hypothesize that the interplay between large initialization and small non-zero weight decay that leads to grokking as predicted (provably) by Lyu et al. (2023) does not hold in our setting because our model violates they *Assumption 3.2*. Let $y_{\mathbf{b}}(\tilde{\mathbf{x}}) = \mathbf{b}^\top \tilde{\mathbf{x}}$ denote our model.

- *Assumption 3.1* (Lyu et al., 2023): For all $\tilde{\mathbf{x}} \in \mathbb{R}^n$, the function $\mathbf{b} \to y_{\mathbf{b}}(\tilde{\mathbf{x}})$ is L-homogeneous with $L = 1$, because $y_{c\mathbf{b}}(\tilde{\mathbf{x}}) = c^L y_{\mathbf{b}}(\tilde{\mathbf{x}})$ for all $c > 0$.

- *Assumption 3.2* (Lyu et al., 2023): for $\zeta = 0$, $y_{\mathbf{b}^{(1)}}(\tilde{\mathbf{x}}) = 0$ for all $\tilde{\mathbf{x}}$ (there is generalization in this case with ℓ_1), but if $\zeta > 0$ (for instance ζ large), this is (almost surely) no longer true. So, this assumption is violated (with high probability).

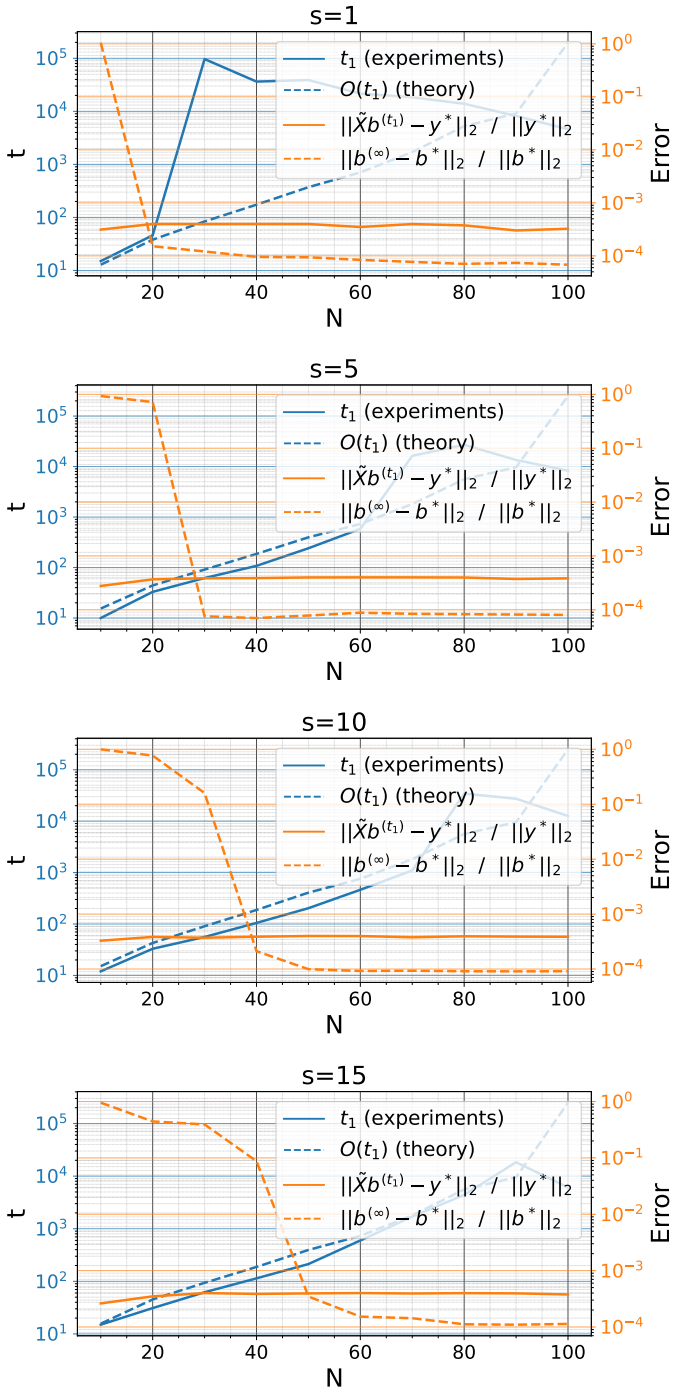


Figure 22: On the left axis, the memorization step t_1 compute experimentally (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$) and the upper bound $-\ln\left(1 + \frac{(1-\rho)\|\mathbf{b}^{(1)} - \hat{\mathbf{b}}\|_\infty}{\alpha\beta_1}\right)/\ln(\rho)$ computed in Theorem C.8. On the right axis, the error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ at step t_1 and the recovery error $\|\mathbf{b}^{(\infty)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at the end of training. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **soft-thresholding algorithm (ISTA)**.

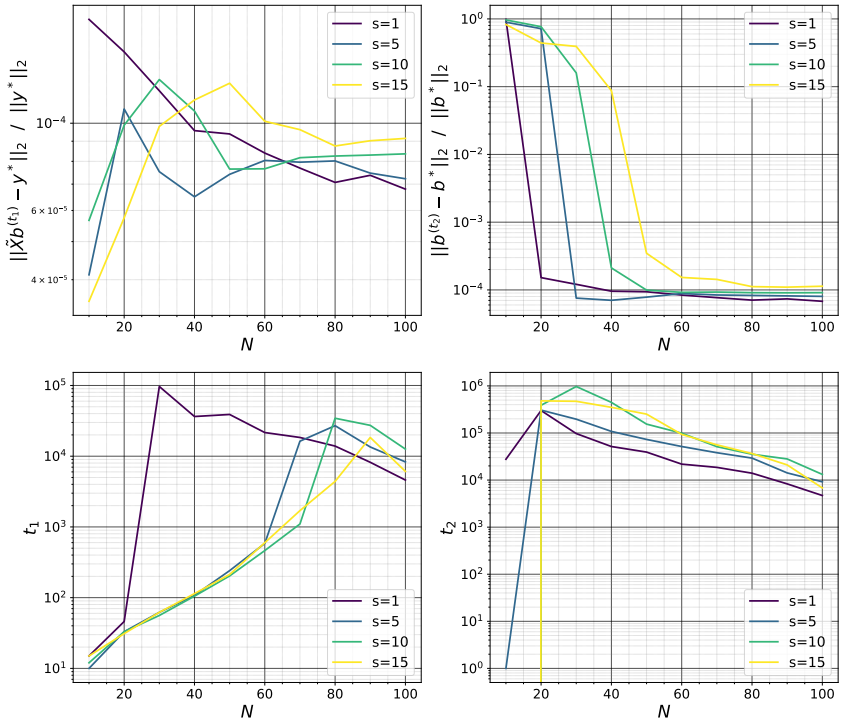


Figure 23: Training error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ at memorization, recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at generalization, memorization step t_1 (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$), and generalization step (smaller t such that $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2 \leq 10^{-4}$ or the maximum training step). Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **soft-thresholding algorithm (ISTA)**

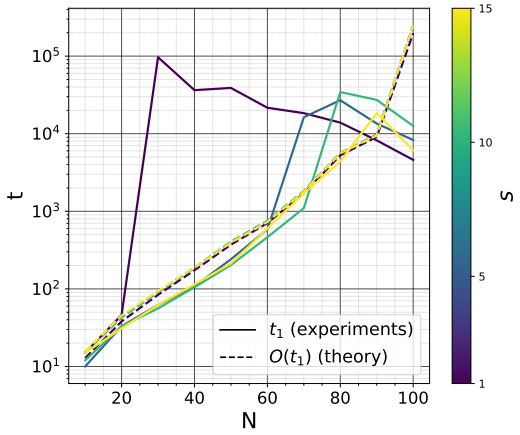


Figure 24: Memorization step t_1 compute experimentally (smaller t such that $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2 \leq 10^{-4}$) and the upper bound computed in Theorem C.8. Here $(n, \alpha, \beta_1, \beta_2) = (10^2, 10^{-1}, 10^{-5}, 0)$, with the **ISTA**.

- *Assumption 3.8* (Lyu et al., 2023): The NTK (Neural Tangent Kernel) features of training samples $\{\nabla_{\mathbf{b}} y_{\mathbf{b}}(\tilde{\mathbf{X}}_i)\}_{i\in[N]}$ are linearly independent (almost surely). In fact, $\nabla_{\mathbf{b}} y_{\mathbf{b}}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}\ \forall \tilde{\mathbf{x}}$. In the over-parametrized regime $N < n$, If $\mathbf{X} \in \mathbb{R}^{N\times n}$ has entries independent and identically distributed from a normal distribution, then the NTK features $\{\tilde{\mathbf{X}}_i\}_{i\in[N]}$ are linearly independent with high probability (because the rank of $\tilde{\mathbf{X}} = \Phi\mathbf{X}$ is N with high probability), so this assumption is verified.

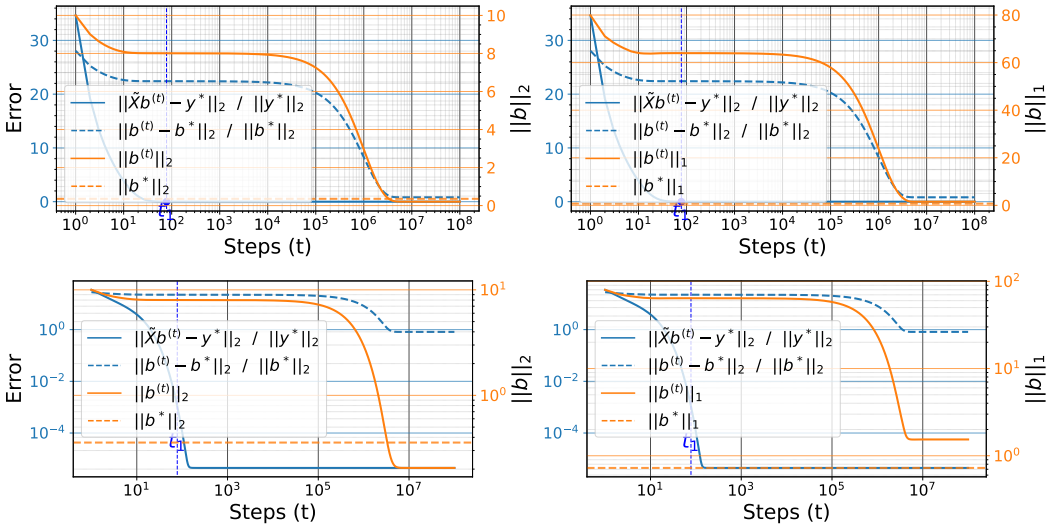


Figure 25: The figures show the relative errors and the the norm $\|\mathbf{b}^{(t)}\|_2$ (left) and $\|\mathbf{b}^{(t)}\|_1$ for $\beta_1 = 0$ and $\beta_1 \neq 0$. Here $(n, s, N) = (100, 5, 30)$ and $(\alpha, \beta_1) = (10^{-1}, 0)$; with **large initialization scale** $\zeta = 10^1$ and **small weights decay** $\beta_2 = 10^{-5}$. Without visualization of the error on a logarithmic scale (top), it looks like grokking has occurred, whereas this is not the case (bottom).

C.10 Impact of Coherence on Grokking: Amplifying Grokking through Data Selection

Above, we introduce the parameter $\tau \in [0, 1]$ that control the incoherence between the measures $\{\mathbf{X}_i\}_{i\in[N]}$ and the sparse basis (dictionary) $\{\Phi_{:,j}\}_{j\in[n]}$. $\tau = 0$ correspond to a full random gaussian $\mathbf{X}$, and correspond to the maximum incoherence, while $\tau = 1$ correspond to $\mathbf{X}_i \in \{\Phi_{:,j}\}_{j\in[n]}$ for all $i \in [N]$, and correspond minimum incoherence (coherence of 1). We also experimentally observe that when using convex programming on the problem (P_1), $N_{\min}(s, \tau)$, the number of samples needed for perfect recovery increases as s and/or τ increases. When $\tau \to 1$, $N_{\min}(s, \tau) \to n$ for all s (Section C.5).

Here, we also observe that the generalization time and the generalization delay increase with τ while the generalization error decreases with it (Figures 27 and 28 and 29). For $N < n$, when $\tau \to 1$, the generalization time $t_2 \to \infty$. This is because each measurement captures a single view (component) of $\mathbf{b}^*$, and this makes it impossible to find the optimal $\mathbf{b}^*$ by solving the equation $\mathbf{X}\Phi\mathbf{b} = \mathbf{y}^*$ (by any method whatsoever). On the other hand, as $\tau \to 0$, $\mathbf{X}$ becomes completely random, and every measurement captures a distinct "view" of $\mathbf{a}^*$, giving the best possible generalization time for the data size considered. The error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ at generalization (t_2) as a function of N and τ has the same shape as in the convex programming (Figures 6 and 7).

C.11 Deep Sparse Recovery: The Effect of Overparametrization

Let now use the parameterization $\mathbf{b} = \odot_{k=1}^{L} \mathbf{B}_k \in \mathbb{R}^n$, with $\mathbf{B} \in \mathbb{R}^{L\times n}$. This corresponds to a linear network with L layers, where each hidden layer has the parameter $\mathrm{diag}(\mathbf{B}_k) \in \mathbb{R}^{n\times n}$—with this, increasing L leads to overparameterization without altering the expressiveness of the function class $\mathbf{b} \to \mathcal{F}_{\mathbf{b}}(\mathbf{x}) = \mathbf{x}^\top\mathbf{b}$, since the model remains linear with respect to the input $\mathbf{x}$. Unlike the shallow

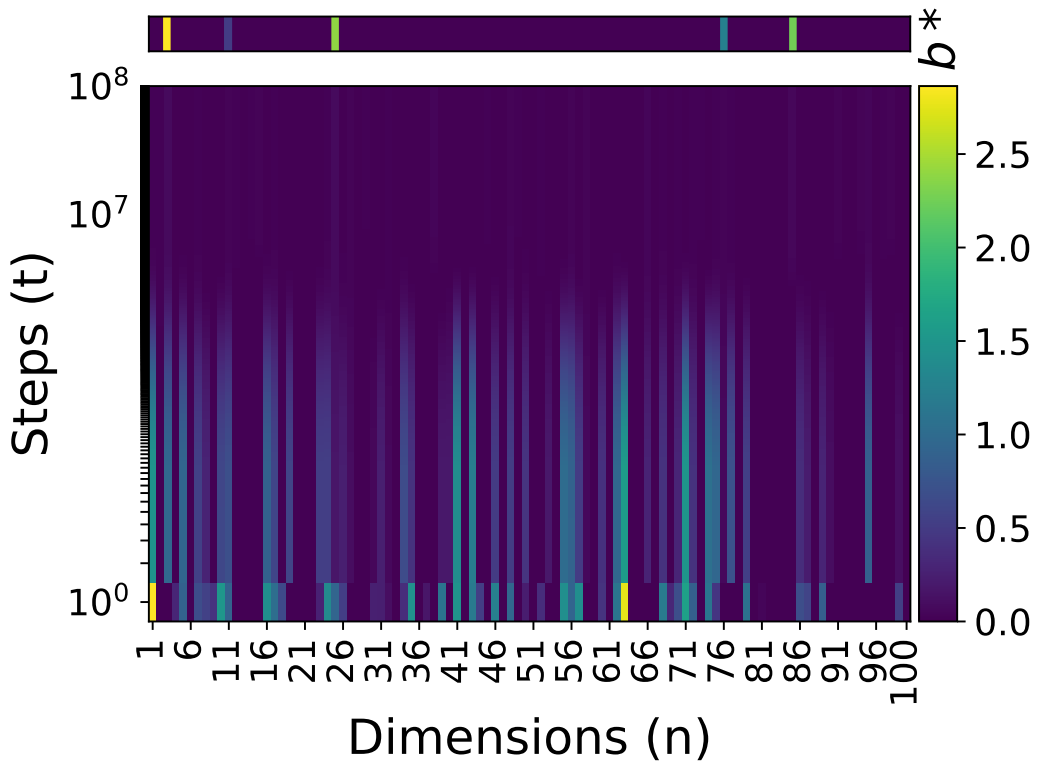


Figure 26: Non convergence of $\mathbf{b}^{(t)}$ to $\mathbf{b}^*$ for $\beta_1 = 0$ and $\beta_1 \neq 0$. Here $(n, s, N) = (100, 5, 30)$ and $(\alpha, \beta_1) = (10^{-1}, 0)$; with **large initialization scale** $\zeta = 10^1$ and **small weights decay** $\beta_2 = 10^{-5}$.

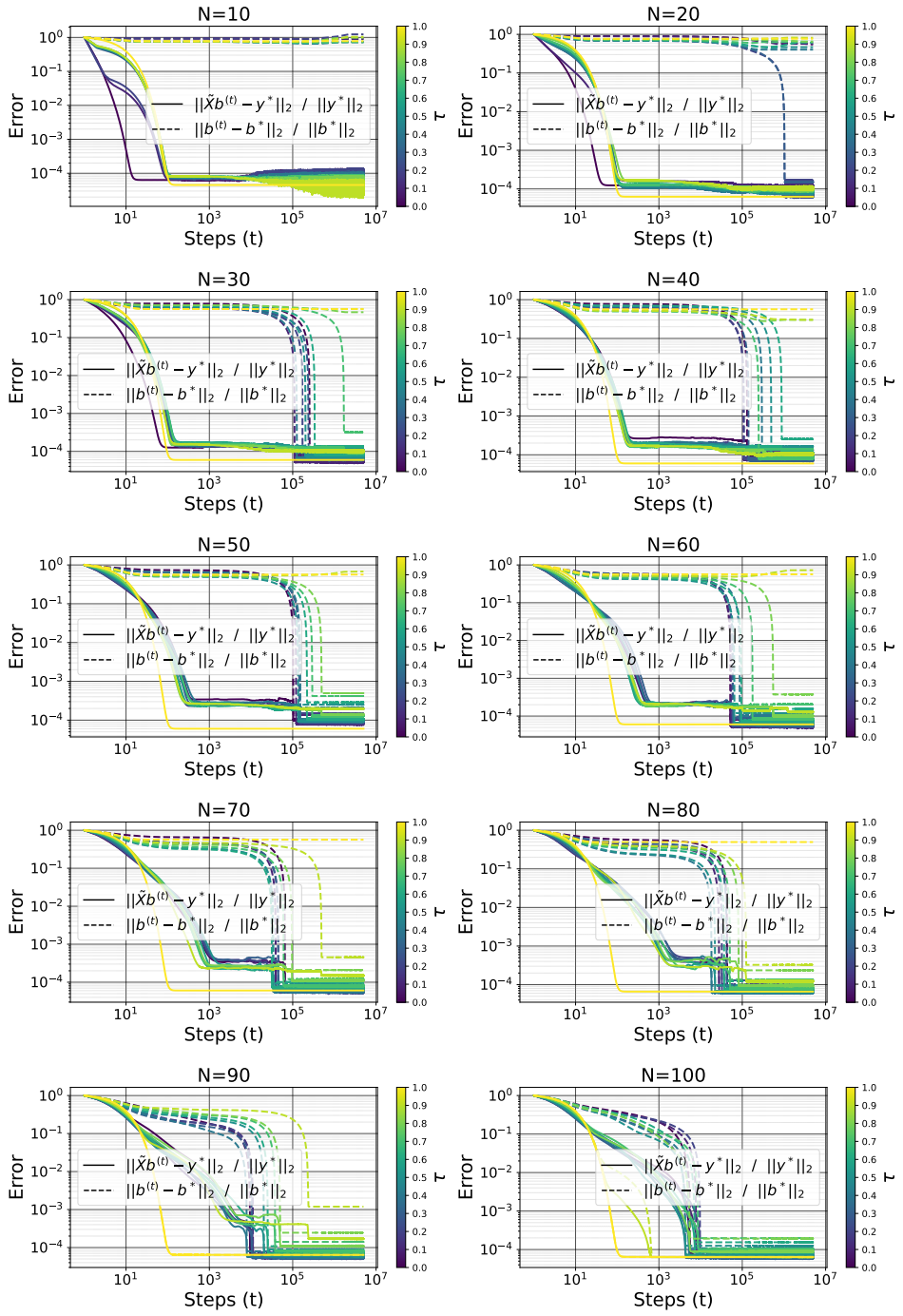


Figure 27: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the number of sample N and the coherence parameter $\tau \in [0, 1]$. Here $(n, s, \alpha, \beta_1, \beta_2, \zeta) = (10^2, 5, 10^{-1}, 10^{-5}, 0, 10^{-6})$.

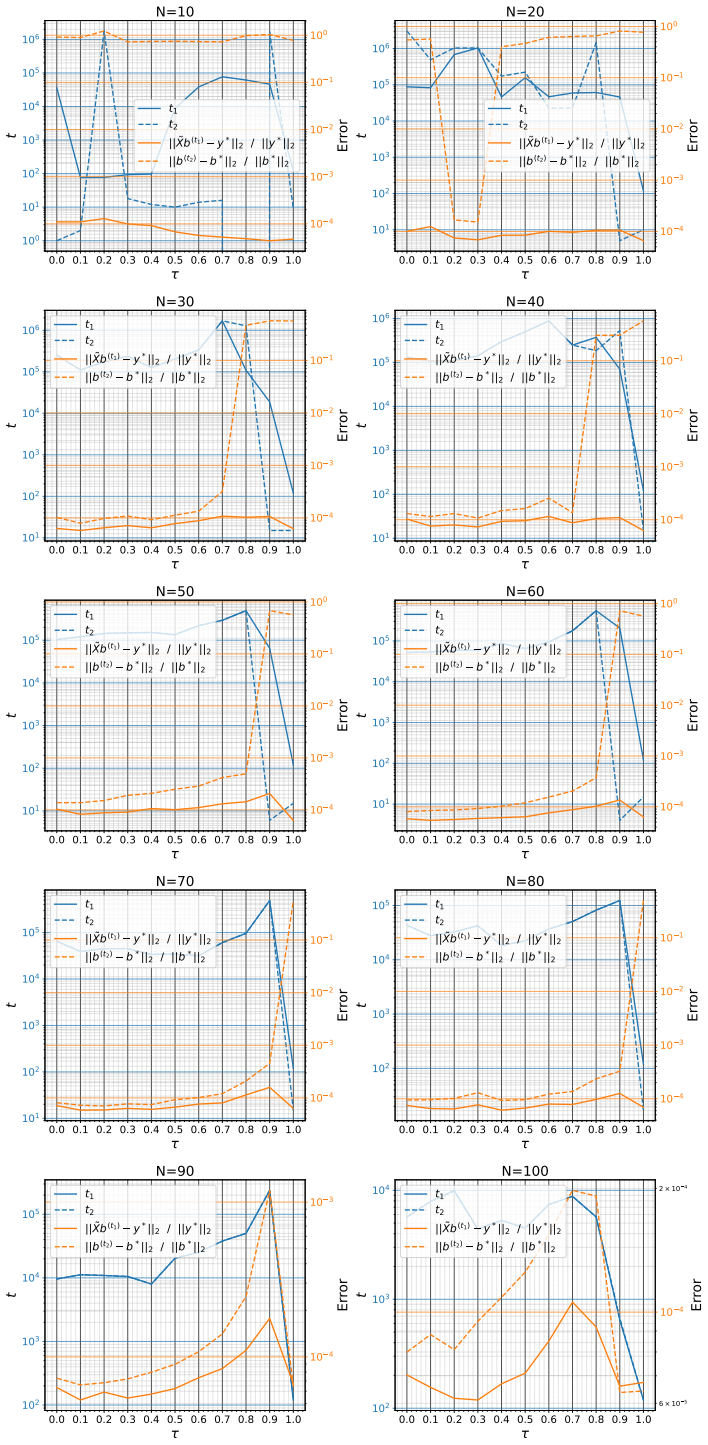


Figure 28: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the coherence parameter $\tau \in [0, 1]$. Here $(n, s, \alpha, \beta_1, \beta_2, \zeta) = (10^2, 5, 10^{-1}, 10^{-5}, 0, 10^{-6})$.

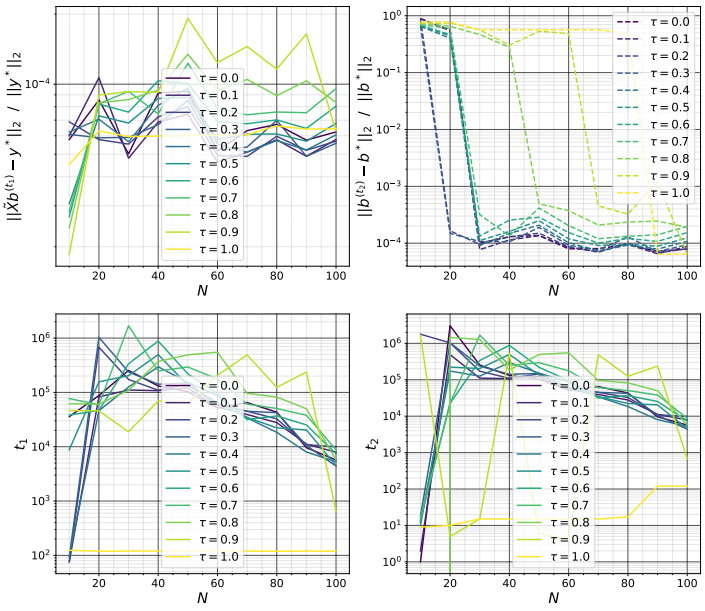


Figure 29: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the coherence parameter $\tau \in [0, 1]$. Here $(n, s, \alpha, \beta_1, \beta_2, \zeta) = (10^2, 5, 10^{-1}, 10^{-5}, 0, 10^{-6})$.

case ($L = 1$), there is no need for ℓ_1 ($\beta_1 = 0$) to generalize when $L \geq 2$ (and the initialization scale is small), as the experiments of this section suggest. With depth, the update for the whole iterate (which is now replaced by a product of matrices and a vector) is similar to the shallow case but with a preconditioner in front of the gradient. This preconditioner makes it possible to recover the sparse signal without any regularization.

We have $\mathbf{y}(\mathbf{b}) = \mathcal{F}_{\mathbf{b}}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\mathbf{b}$ and $\mathbf{y}^* = \mathcal{F}_{\mathbf{b}^*}(\tilde{\mathbf{X}}) + \boldsymbol{\xi} = \tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}$, and want to minimize $f(\mathbf{b}) = g_{\beta_2}(\mathbf{b}) + \beta_1\|\mathbf{B}\|_1$ using gradient descent, where

$$\begin{aligned} g_{\beta_2}(\mathbf{b}) &:= \frac{1}{2}\|\mathbf{y}(\mathbf{b}) - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{B}\|_{\mathrm{F}}^2 \\ &= \frac{1}{2}\mathbf{b}^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b} - \mathbf{y}^{*\top}\tilde{\mathbf{X}}\mathbf{b} + \frac{1}{2}\mathbf{y}^{*\top}\mathbf{y}^* + \frac{\beta_2}{2}\|\mathbf{B}\|_{\mathrm{F}}^2 \\ &= \frac{1}{2}\mathbf{b}^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b} - \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{b}^* + \tilde{\mathbf{X}}^\top\boldsymbol{\xi}\right)^\top\mathbf{b} + \frac{\beta_2}{2}\|\mathbf{B}\|_{\mathrm{F}}^2 + \frac{1}{2}\|\tilde{\mathbf{X}}\mathbf{b}^* + \boldsymbol{\xi}\|_2^2 \end{aligned} \tag{87}$$

Let $G(\mathbf{b}) := \frac{\partial g_{\beta_2}(\mathbf{b})}{\partial \mathbf{b}} = \tilde{\mathbf{X}}^\top(\mathbf{y}(\mathbf{b}) - \mathbf{y}^*) = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\mathbf{b} - \mathbf{b}^*) - \tilde{\mathbf{X}}^\top\boldsymbol{\xi}$. The gradient for each $\mathbf{B}_i$ is $G_{\beta_2}(\mathbf{B}_i) := \frac{\partial g_{\beta_2}(\mathbf{b})}{\partial \mathbf{B}_i} = \frac{\partial \mathbf{b}}{\partial \mathbf{B}_i}\frac{\partial g_{\beta_2}(\mathbf{b})}{\partial \mathbf{b}} + \beta_2\mathbf{B}_i = \mathrm{diag}(\prod_{k\neq i}\mathbf{B}_k)G(\mathbf{b}) + \beta_2\mathbf{B}_i$, and the update rule for each $\mathbf{B}_i$ is

$$\begin{aligned} \mathbf{B}_i^{(t+1)} &= \mathbf{B}_i^{(t)} - \alpha G_{\beta_2}(\mathbf{B}_i^{(t)}) - \alpha\beta_1 h(\mathbf{B}_i^{(t)}) \\ &= (1 - \alpha\beta_2)\mathbf{B}_i^{(t)} - \alpha\,\mathrm{diag}(\prod_{k\neq i}\mathbf{B}_k^{(t)})G(\mathbf{b}^{(t)}) - \alpha\beta_1 h(\mathbf{B}_i^{(t)}) \end{aligned} \tag{88}$$

where $h(\mathbf{B}_i) \in \partial\|\mathbf{B}_i\|_1$ any subgradient of $\|\mathbf{B}_i\|_1$, $h(\mathbf{B}_i)_k = \mathrm{sign}(\mathbf{B}_{ik})$ for $\mathbf{B}_{ik} \neq 0$, and any value in $[+1, -1]$ for $\mathbf{B}_{ik} = 0$. We start the optimization at $\mathbf{B}_i^{(1)} \overset{iid}{\sim} \zeta\mathcal{N}(0, 1/n)$ with $\zeta \geq 0$ the initialization scale.

Without ovaparametrization ($L = 1$), the gradient update for $\mathbf{b}$ writes

$$\begin{aligned} \mathbf{b}^{(t+1)} &= \mathbf{b}^{(t)} - \alpha G_{\beta_2}(\mathbf{b}^{(t)}) - \alpha\beta_1 h(\mathbf{b}^{(t)}) \\ &= (1 - \alpha\beta_2)\mathbf{b}^{(t)} - \alpha\left(G(\mathbf{b}^{(t)}) + \beta_1 h(\mathbf{b}^{(t)})\right) \end{aligned} \tag{89}$$

As we show above, for $s = \|\mathbf{b}^*\|_0 \ll n$ and $N < n$, without ℓ_1 regularization ($\beta_1 = 0$), we don't have perfect recovery. Here, the update is unconditioned and progresses uniformly in all directions. So without ℓ_1-regularization, there is no mechanism to enforce sparsity, and perfect recovery of $\mathbf{b}^*$ is impossible.

For $L = 2$, let $\mathbf{c} := \mathbf{B}_1^2 + \mathbf{B}_2^2$. If $\beta_1 = 0$, then

$$\begin{aligned}
\mathbf{b}^{(t+1)} &= \mathbf{B}_1^{(t+1)} \odot \mathbf{B}_2^{(t+1)} \\
&= \Big((1-\alpha\beta_2)\mathbf{B}_1^{(t)} - \alpha\,\mathrm{diag}(\mathbf{B}_2^{(t)})G(\mathbf{b}^{(t)})\Big) \odot \Big((1-\alpha\beta_2)\mathbf{B}_2^{(t)} - \alpha\,\mathrm{diag}(\mathbf{B}_1^{(t)})G(\mathbf{b}^{(t)})\Big) \\
&= (1-\alpha\beta_2)^2\mathbf{b}^{(t)} - \alpha(1-\alpha\beta_2)\,\mathrm{diag}(\mathbf{B}_1^{(t)} \odot \mathbf{B}_1^{(t)} + \mathbf{B}_2^{(t)} \odot \mathbf{B}_2^{(t)})G(\mathbf{b}^{(t)}) + \alpha^2\,\mathrm{diag}(\mathbf{b}^{(t)})G(\mathbf{b}^{(t)})^2 \\
&= (1-\alpha\beta_2)^2\mathbf{b}^{(t)} - \alpha(1-\alpha\beta_2)\mathbf{c}^{(t)} \odot G(\mathbf{b}^{(t)}) + \alpha^2\mathbf{b}^{(t)} \odot G(\mathbf{b}^{(t)})^2 \\
&\approx (1-2\alpha\beta_2)\mathbf{b}^{(t)} - \alpha\mathbf{c}^{(t)} \odot G(\mathbf{b}^{(t)}) \text{ for } \alpha \to 0
\end{aligned} \tag{90}$$

and

$$\begin{aligned}
\mathbf{c}^{(t+1)} &= \mathbf{B}_1^{(t+1)} \odot \mathbf{B}_1^{(t+1)} + \mathbf{B}_2^{(t+1)} \odot \mathbf{B}_2^{(t+1)} \\
&= (1-\alpha\beta_2)^2\mathbf{B}_1^{(t)} \odot \mathbf{B}_1^{(t)} - 2\alpha(1-\alpha\beta_2)\,\mathrm{diag}(\mathbf{B}_1^{(t)} \odot \mathbf{B}_2^{(t)})G(\mathbf{b}^{(t)}) + \alpha^2\,\mathrm{diag}(\mathbf{B}_2^{(t)} \odot \mathbf{B}_2^{(t)})G(\mathbf{b}^{(t)})^2 \\
&\quad + (1-\alpha\beta_2)^2\mathbf{B}_2^{(t)} \odot \mathbf{B}_2^{(t)} - 2\alpha(1-\alpha\beta_2)\,\mathrm{diag}(\mathbf{B}_2^{(t)} \odot \mathbf{B}_1^{(t)})G(\mathbf{b}^{(t)}) + \alpha^2\,\mathrm{diag}(\mathbf{B}_1^{(t)} \odot \mathbf{B}_1^{(t)})G(\mathbf{b}^{(t)})^2 \\
&= (1-\alpha\beta_2)^2\mathbf{c}^{(t)} - 4\alpha(1-\alpha\beta_2)\mathbf{b}^{(t)} \odot G(\mathbf{b}^{(t)}) + \alpha^2\mathbf{c}^{(t)} \odot G(\mathbf{b}^{(t)})^2 \\
&\approx (1-2\alpha\beta_2)\mathbf{c}^{(t)} - 4\alpha\mathbf{b}^{(t)} \odot G(\mathbf{b}^{(t)}) \text{ for } \alpha \to 0
\end{aligned} \tag{91}$$

The depth adds the preconditioning $\mathbf{P}^{(t)} = (1-\alpha\beta_2)\,\mathrm{diag}(\mathbf{c}^{(t)})$ in front of the update for $\mathbf{b}$. This preconditioning mechanism seems to implicitly favor sparsity and, thus, a perfect recovery after memorization since a sparse solution for the problem of interest is necessary $\mathbf{b}^*$ when N is large enough (with respect to $s = \|\mathbf{b}^*\|_0$ and n). In fact, when $\mathbf{c}_i^{(t)}$ goes to zero (which is the case when $\mathbf{b}_i^{(t)}$ is also small), the update becomes $\mathbf{b}_i^{(t+1)} \approx (1-2\alpha\beta_2)\mathbf{b}_i^{(t)}$, and thus push $\mathbf{b}_i^{(t+1)}$ to 0 at a geometric rate of $\mathcal{O}(1-2\alpha\beta_2)$. Otherwise, $\mathbf{c}_i^{(t)}$ (large) will amplify the gradient so that $\mathbf{c}_i^{(t)}G(\mathbf{b}^{(t)})_i$ dominates the update, which pushes $\mathbf{b}^{(t)}$ towards $\mathbf{b}^*$ (as the gradient $G(\mathbf{b}^{(t)})$ points towards a small error $\mathbf{b}^{(t)} - \mathbf{b}^*$ direction, particularly for full rank $\tilde{\mathbf{X}}$ and high signal to ratio regime).

We optimize the noiseless problem ($\boldsymbol{\xi} = 0$) using the subgradient descent method with $(n, s, \zeta, \alpha, \beta_1, \beta_2) = (10^2, 30, 10^{-2}, 10^{-1}, 10^{-5}, 0)$, for different values of N and $L \in \{1, 2, 3, 4\}$. Here, initializing $\mathbf{B}$ too close to the origin (initialization scale $\zeta \to 0$) leads $\mathbf{b}$ to not change during training. The model is able to recover the true signal $\mathbf{b}^*$, and the generalization delay becomes extremely small (compared to the shallow case with $\beta_1 \neq 0$) for $L = 2$ and disappears (ungrokking) for $L > 2$ (Figure 30). As L becomes larger, the phase transition to generalization becomes extremely abrupt. The loss decreases in a staircase fashion, with more or less long plateaus of suboptimal generalization error during training. This type of behavior is generally observed in the optimization of *Soft Committee Machines* (Biehl & Schwarze, 1995; Saad & Solla, 1995b;a; 1996; Engel & Broeck, 2001; Aubin et al., 2018; Goldt et al., 2020), which are two-layer linear or non-linear teacher-student systems, with the output layer of the student fixed to that of the teacher during training.

Also, for a fixed number N of measure, the test error decreases with L, showing that depth helps to find the signal with a smaller number of measures, albeit with a longer training time (Figures 31 and 32). So, the depth seems to have the same effect on generalization as β_1. This is in accord with the result of Arora et al. (2018) in the context of matrix factorization. They show that introducing depth effectively turns gradient descent into a shallow (single-layer) training process equipped with a built-in preconditioning mechanism. This mechanism biases updates toward directions already explored by the optimization, serving as an acceleration technique that fuses momentum with adaptive step sizes. Furthermore, they demonstrate that depth-based overparameterization can substantially speed up training, even in straightforward convex tasks like linear regression under with ℓ_p loss, $p > 2$.

Note that for $L \geq 2$, using a large scale initialization and a small but non-zero ℓ_2 regularization β_2 results in grokking (Figures 34, 35 and 33), unlike the case of $L = 1$ that gives the "grokking

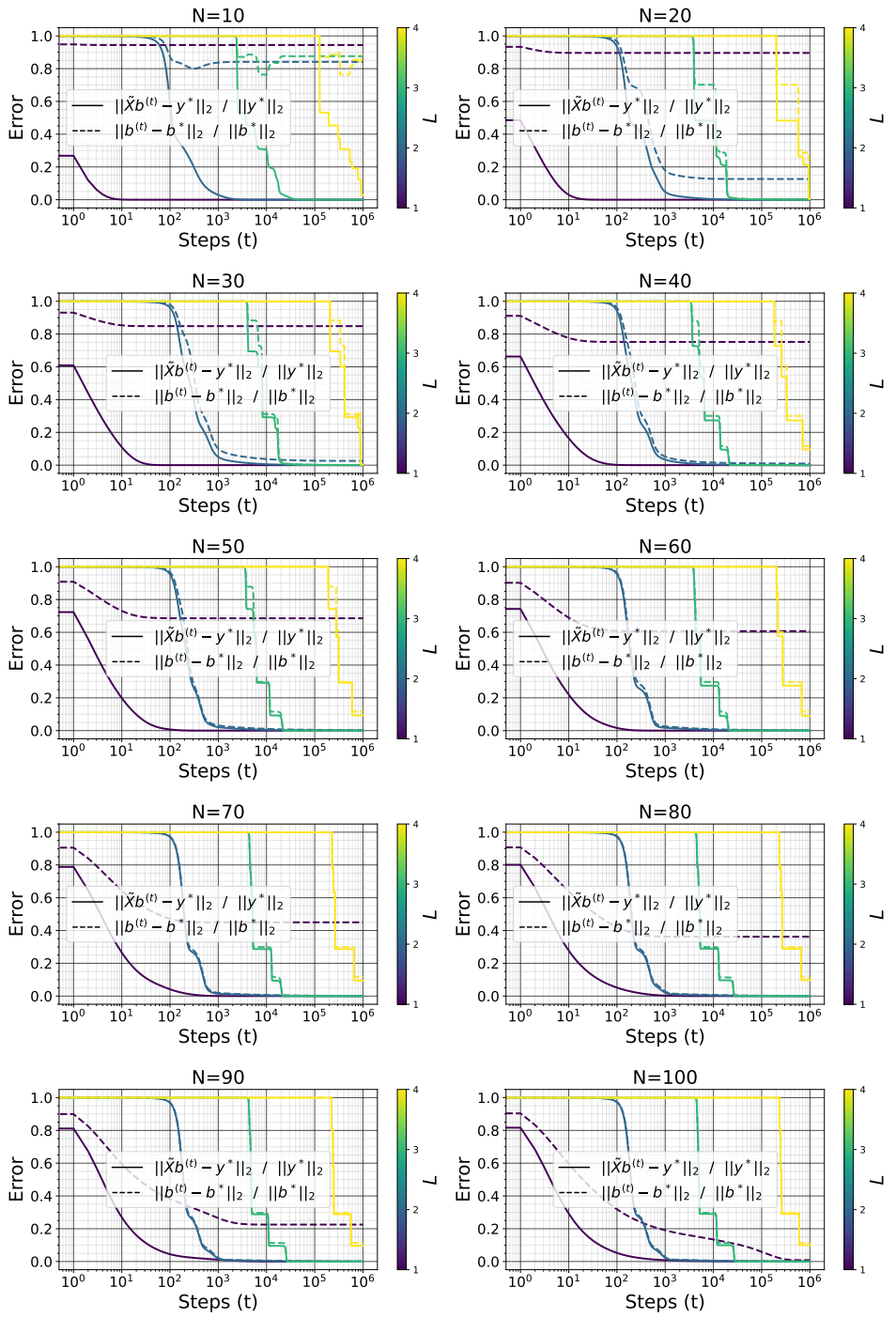


Figure 30: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the number of sample N and the depth $L \in \{1, 2, 3, 4\}$. Here $(n, s, \alpha, \beta_1, \beta_2) = (10^2, 5, 10^{-1}, 0, 0)$; with **small initialization scale** $\zeta = 10^{-6}$ for $L = 1$ and $\zeta = 10^{-2}$ for $L > 1$.

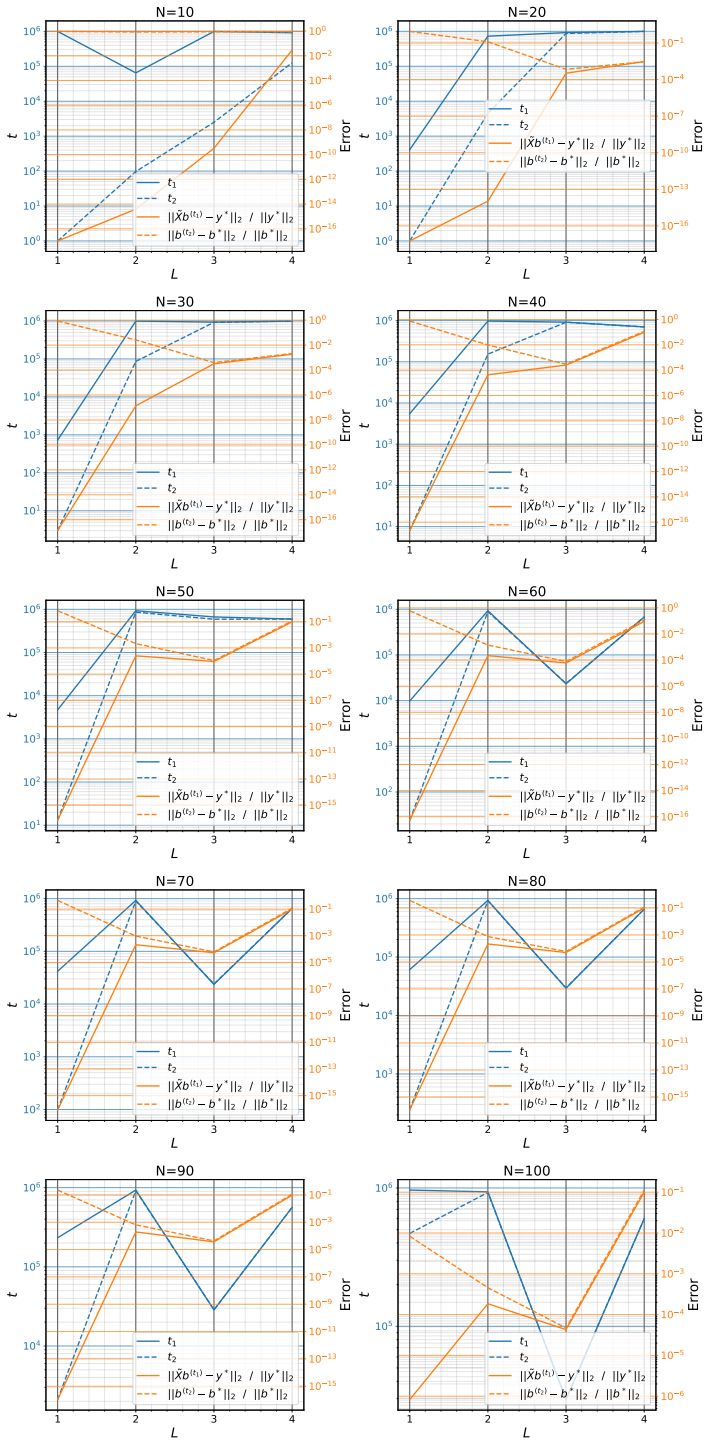


Figure 31: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the depth $L \in \{1, 2, 3, 4\}$. Here $(n, s, \alpha, \beta_1, \beta_2) = (10^2, 5, 10^{-1}, 0, 0)$; with **small initialization scale** $\zeta = 10^{-6}$ for $L = 1$ and $\zeta = 10^{-2}$ for $L > 1$.

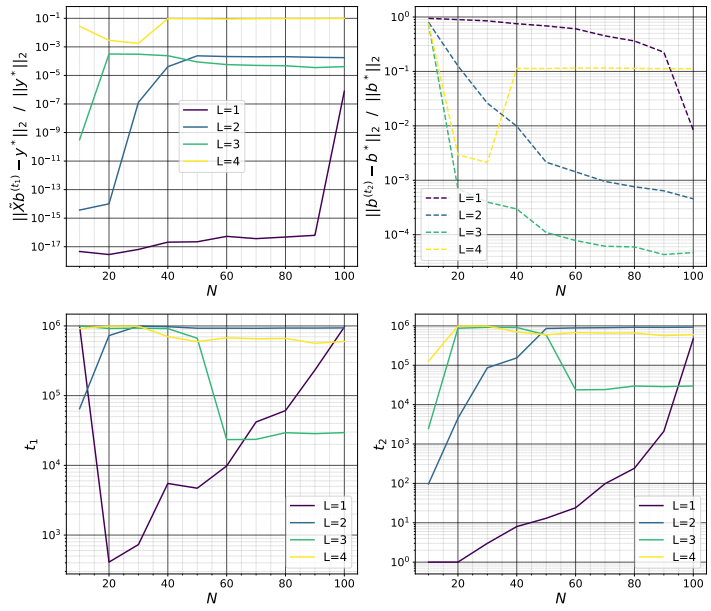


Figure 32: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the depth $L \in \{1, 2, 3, 4\}$. Here $(n, s, \alpha, \beta_1, \beta_2) = (10^2, 5, 10^{-1}, 0, 0)$; with **small initialization scale** $\zeta = 10^{-6}$ for $L = 1$ and $\zeta = 10^{-2}$ for $L > 1$. The growth (as a function of N) in the test error for $L = 4$ is simply due to the fact that we did not optimize long enough for it to decrease.

without understanding" phenomenon (Section C.9). In this regime of large intialisation and small non-zero weight decay, when L increases, the number of steps required for the model to move from memorization to generalization is reduced (grokking acceleration), and the generalization error at the end of training is considerably lower (Figure 33). Lyu et al. (2023) used a similar setup to show that an interplay between large initialization and small nonzero weights decay gives rise to grokking with the diagonal linear network $y(\mathbf{x}) = \left(\mathbf{u}^{\odot L} - \mathbf{v}^{\odot L}\right)^\top \mathbf{x}$ in the context of binary classification, but there did not study the impact of L on the generalization delay, but focus on characterizing how sharp is the transition from memorization to generalization as a function of the initialization scale and the weight decay coefficient, and how long it takes for this transition to occurs. This diagonal linear network is also often used for sparse recovery problems (Vaskevicius et al., 2019), but the focus is generally on its ability to recover the optimal solution, and not grokking.

C.12 Realistic signals

C.12.1 Recovery of an image

We consider a 8×8 digit 0 from the MNIST dataset, $n = 8^2 = 64$. The image is normalized to have values in $[0, 1]$, and the values below 0.5 are set to zero, leading to a sparsity level $s = 22$ (34.38% of n). The evaluation of the errors is shown in Figures 36, and the evolution of the reconstructed image as a function of the training steps are shown in Figure 37.

C.12.2 Recovery of a sinusoidal signal

We construct a sparse real-valued signal $\mathbf{a}^* \in \mathbb{R}^n$ from a set of sinusoidal components defined by their frequencies, amplitudes, and phases. For that, we first define the sparse frequency-domain representation $\mathbf{b}^* \in \mathbb{C}^n$ as $\mathbf{b}^*(k) = A_k e^{\mathbf{i}\varphi_k} \cdot \mathbb{1}(k \in \mathcal{F})$ where $\mathcal{F} \subset \{0, 1, \ldots, n-1\}$ is the set of selected frequency indices with $|\mathcal{F}| = s$; $A_k \in \mathbb{R}^+$ the amplitude of the sinusoid at frequency index k; $\varphi_k \in [0, 2\pi)$ the phase of the sinusoid at frequency index k; and $\mathbf{i}$ the imaginary unit ($\mathbf{i}^2 = -1$). The real-valued time-domain signal $\mathbf{a}^* \in \mathbb{R}^n$ is obtained by applying the inverse discrete Fourier

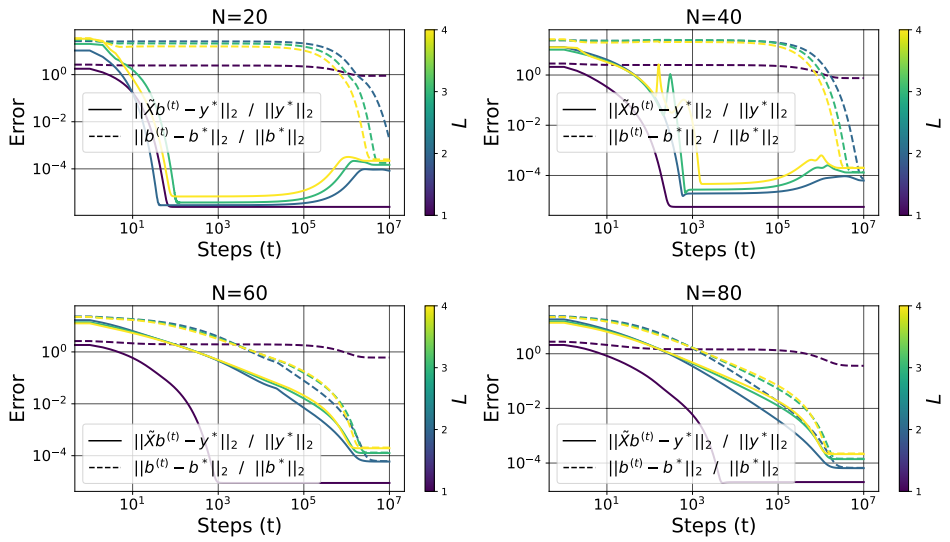


Figure 33: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ as a function of the number of sample N and the depth $L \in \{1, 2, 3, 4\}$. Here $(n, s) = (10^2, 5)$ and $(\alpha, \beta_1) = (10^{-1}, 0)$; with **large initialization scale** $\zeta = 10^0$ and **small weights decay** $\beta_2 = 10^{-5}$.

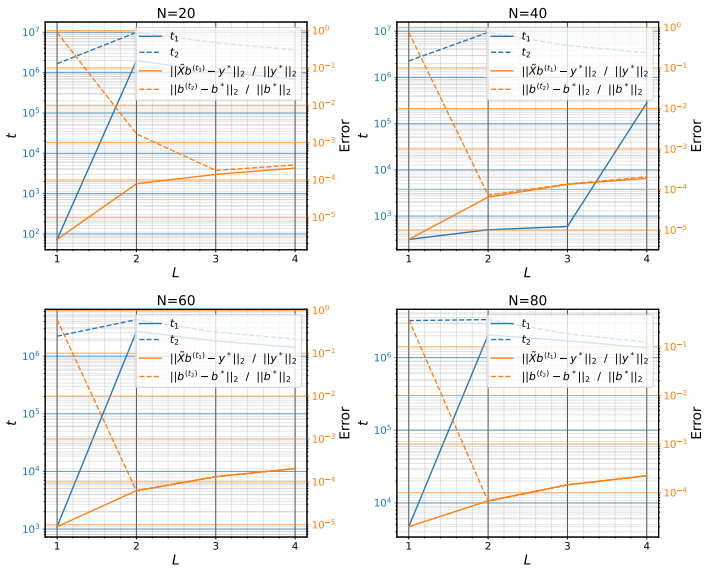


Figure 34: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the depth $L \in \{1, 2, 3, 4\}$. Here $(n, s) = (10^2, 5)$ and $(\alpha, \beta_1) = (10^{-1}, 0)$; with **large initialization scale** $\zeta = 10^0$ and **small weights decay** $\beta_2 = 10^{-5}$.

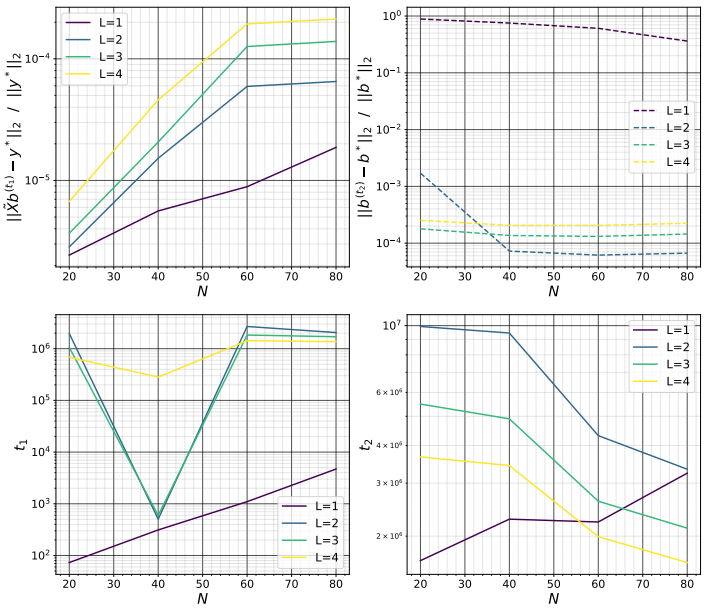


Figure 35: Training and error $\|\tilde{\mathbf{X}}\mathbf{b}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{b}^{(t_2)} - \mathbf{b}^*\|_2/\|\mathbf{b}^*\|_2$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the depth $L \in \{1, 2, 3, 4\}$. Here $(n, s) = (10^2, 5)$ and $(\alpha, \beta_1) = (10^{-1}, 0)$; with **large initialization scale** $\zeta = 10^0$ and **small weights decay** $\beta_2 = 10^{-5}$.

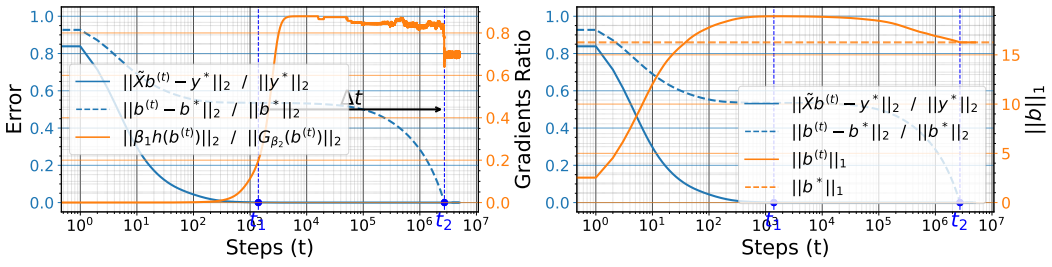


Figure 36: Reconstruction of a 8×8 digit from the MNIST dataset. The figures show the relative errors, gradient ratio, and the norm $\|\mathbf{b}^{(t)}\|_1$ (right). $G_{\beta_2}(\mathbf{b}^{(t)})$ dominates $\beta_1 h(\mathbf{b}^{(t)})$ until memorization, i.e. $\|\beta_1 h(\mathbf{b}^{(t)})\|/\|G_{\beta_2}(\mathbf{b}^{(t)})\| \ll 1$ for all $t \leq t_1$. From memorization $\beta_1 h(\mathbf{b}^{(t)})$ dominates and make $\|\mathbf{b}^{(t)}\|_1$ converge to $\|\mathbf{b}^*\|_1$ at t_2, and so $\mathbf{b}^{(t_2)} = \mathbf{b}^*$.

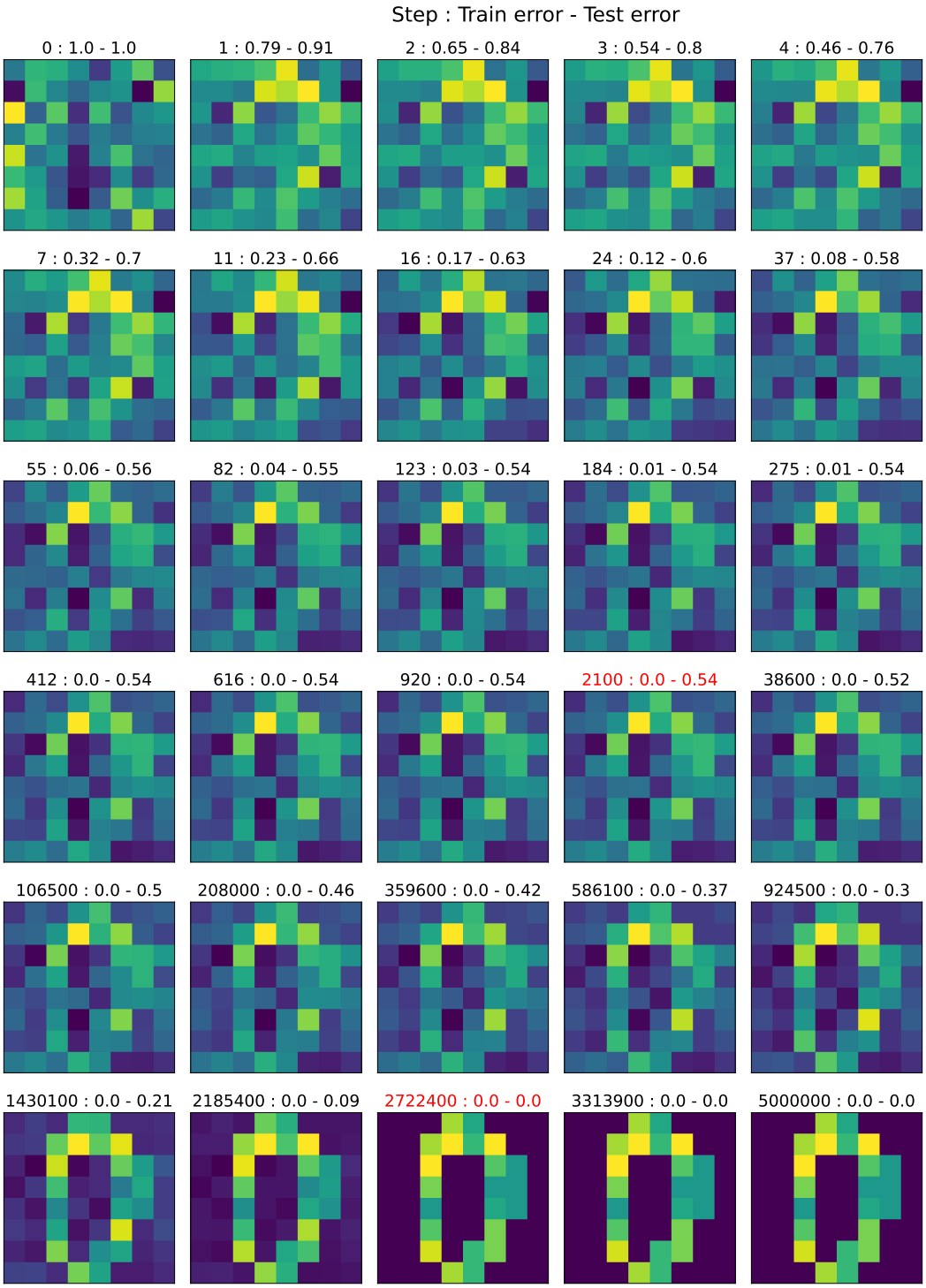


Figure 37: Reconstruction of a 8×8 digit from the MNIST dataset. The figure shows the evolution of the reconstructed image with the training step t.

transform to $\mathbf{b}^*$, scaled by a factor n to ensure consistent normalization:

$$\mathbf{a}^*(t) = n \cdot \text{Re}\left(\frac{1}{n}\sum_{k=0}^{n-1}\mathbf{b}^*[k]e^{\mathbf{i}2\pi\frac{kt}{n}}\right) = \sum_{k\in\mathcal{F}} A_k \cos\left(\frac{2\pi k}{n}t + \varphi_k\right) \quad \text{for } t = 0, \dots, n-1 \quad (92)$$

We use $(n, s) = (100, 5)$, $\mathcal{F} = \{10, 25, 40, 75, 95\}$, $A = [1.0, 0.8, 1.2, 1.5, 0.5]$ and $\varphi = [0, \pi/4, 3\pi/8, 3\pi/4, \pi]$ (Figure 38). The evaluation of the errors is shown in Figures 39, and the evolution of the reconstructed signal as a function of the training steps is shown in Figure 40.

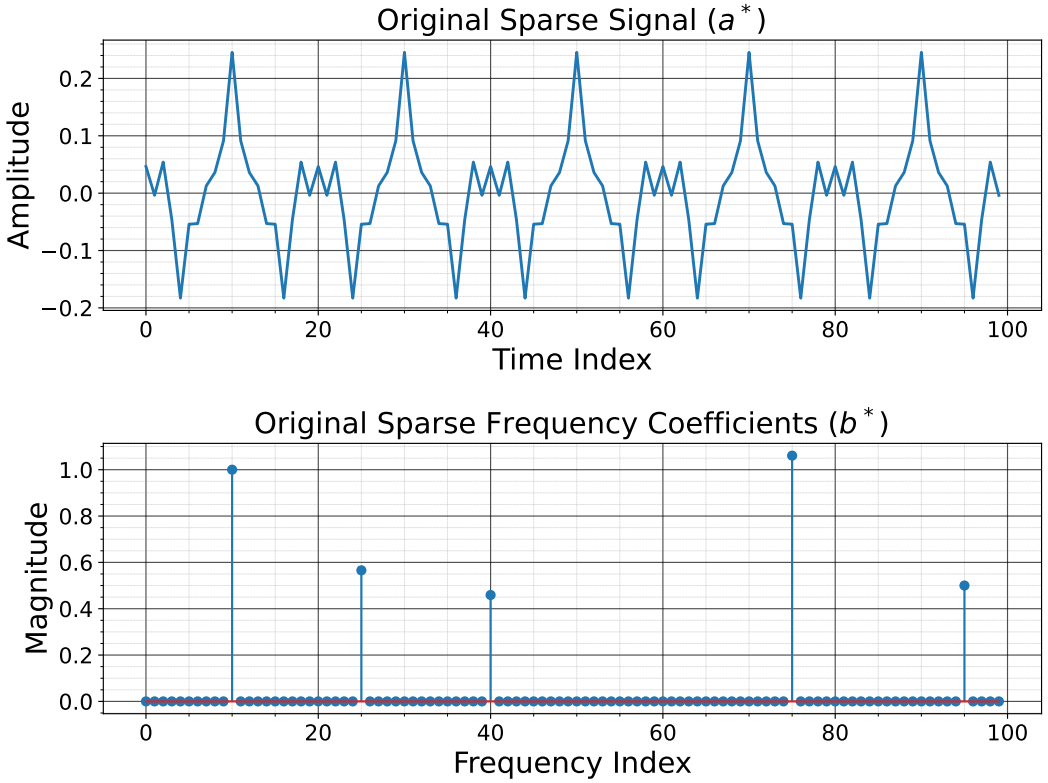


Figure 38: Reconstruction of a sinusoidal signal $\mathbf{a}^*(t) = \sum_{k\in\mathcal{F}} A_k \cos\left(\frac{2\pi k}{n}t + \varphi_k\right)$ with a sparse representation $\mathbf{b}^*(k) = A_k e^{\mathbf{i}\varphi_k} \cdot \mathbb{1}\,(k \in \mathcal{F})$, where $(n, s) = (100, 5)$, $\mathcal{F} = \{10, 25, 40, 75, 95\}$, $A = [1.0, 0.8, 1.2, 1.5, 0.5]$ and $\varphi = [0, \pi/4, 3\pi/8, 3\pi/4, \pi]$.

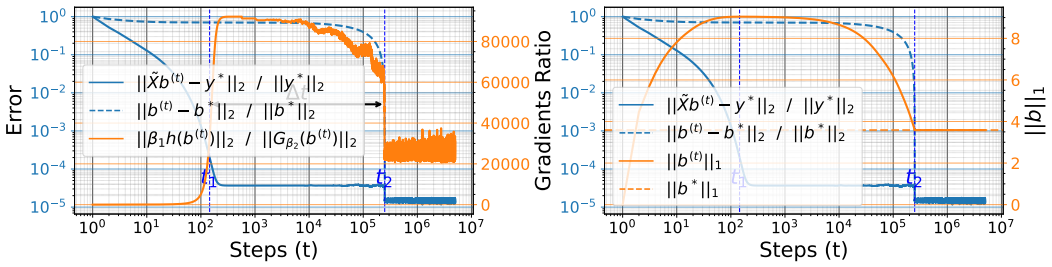


Figure 39: Reconstruction of a sinusoidal signal. The figures show the relative errors, gradient ratio, and the norm $\|\mathbf{b}^{(t)}\|_1$ (right). $G_{\beta_2}(\mathbf{b}^{(t)})$ dominates $\beta_1 h(\mathbf{b}^{(t)})$ until memorization, i.e. $\|\beta_1 h(\mathbf{b}^{(t)})\|/\|G_{\beta_2}(\mathbf{b}^{(t)})\| \ll 1$ for all $t \leq t_1$. From memorization $\beta_1 h(\mathbf{b}^{(t)})$ dominates and make $\|\mathbf{b}^{(t)}\|_1$ converge to $\|\mathbf{b}^*\|_1$ at t_2, and so $\mathbf{b}^{(t_2)} = \mathbf{b}^*$.

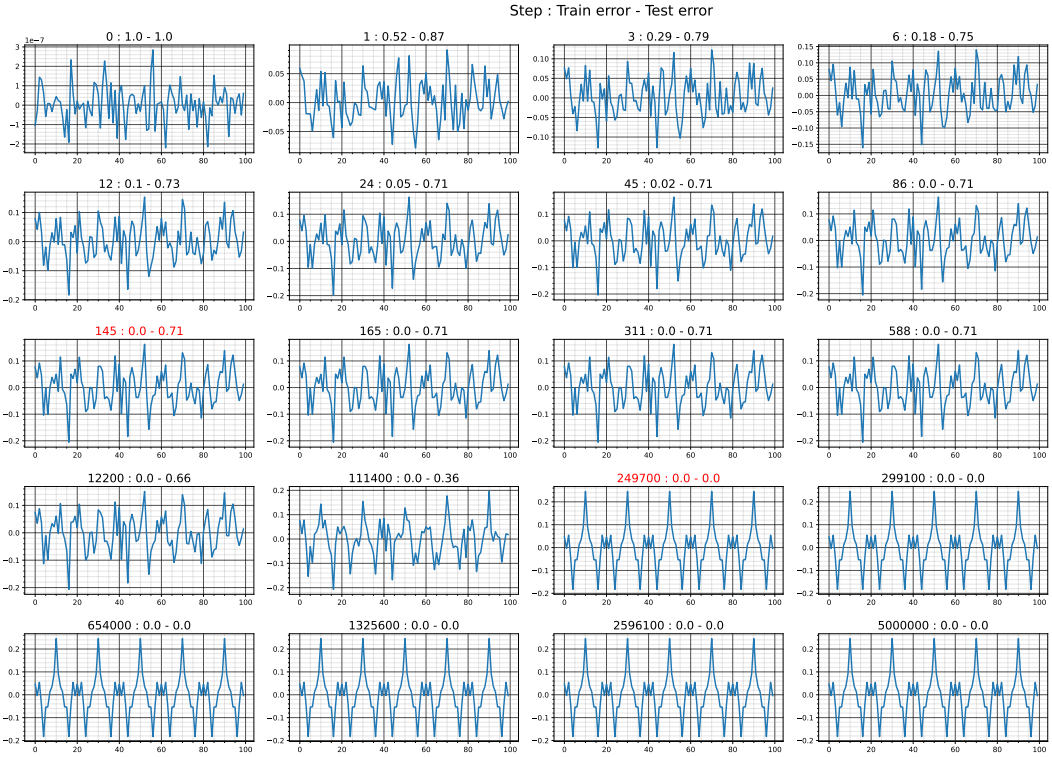


Figure 40: Reconstruction of a sinusoidal signal. The figure shows the evolution of the reconstructed image with the training step t.

C.12.3 Recovery of sparse polynomial

We consider a polynomial $p^* : \mathbb{R}^m \to \mathbb{R}$ define by $p^*(\mathbf{x}) = \mathbf{x}^\top \mathbf{M}^* \mathbf{x} + \mathbf{m}^{*\top} \mathbf{x} = (\mathbf{x} \otimes \mathbf{x})^\top \operatorname{vec} \mathbf{M}^* + \mathbf{m}^{*\top} \mathbf{x} = \mathbf{a}^{*\top} q(\mathbf{x})$ with $\mathbf{a}^* = [\operatorname{vec}(\mathbf{M}^*) \quad \mathbf{m}^*] \in \mathbb{R}^{m(m+1)}$ and $q(\mathbf{x}) = [\mathbf{x} \otimes \mathbf{x} \quad \mathbf{x}] \in \mathbb{R}^{m(m+1)}$.

To well define p^*, we make $\mathbf{M}^*$ upper triangular ($\mathbf{M}^*_{ij} = 0$ for $j < i$) so that $p^*(\mathbf{x}) = \sum_{i=1}^m \sum_{j=i}^m \mathbf{M}^*_{ij} \mathbf{x}_i \mathbf{x}_j + \sum_{i=1}^m \mathbf{m}^*_i \mathbf{x}_i$. This function has $n = \frac{(m+1)m}{2} + m = \frac{(m+3)m}{2}$ parameters, and write $p^*(\mathbf{x}) = \mathbf{a}^{*\top} q(\mathbf{x})$ with $\mathbf{a}^* = [\mathbf{M}^*_{11}, \mathbf{M}^*_{12}, \dots, \mathbf{M}^*_{1m}, \mathbf{M}^*_{22}, \dots, \mathbf{M}^*_{mm} \quad \mathbf{m}^*_1, \dots, \mathbf{m}^*_m] \in \mathbb{R}^n$ and

$$
\begin{aligned}
q(\mathbf{x}) = \big[& \mathbf{x}_1^2, \mathbf{x}_1 \mathbf{x}_2, \dots, \mathbf{x}_1 \mathbf{x}_m, \\
& \mathbf{x}_2^2, \mathbf{x}_2 \mathbf{x}_3, \dots, \mathbf{x}_2 \mathbf{x}_m, \\
& \dots, \\
& \mathbf{x}_m^2, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \big] \in \mathbb{R}^n
\end{aligned}
\tag{93}
$$

We sample $s \ll n$ of the n parameters iid from $\mathcal{N}(0, 1/n)$ and set the remaining to 0. Also, $\mathbf{x} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$.

There are two ways to have grokking on this problem :

- We can iid sample N inputs output pair $\{(\mathbf{x}_i, p^*(\mathbf{x}_i))\}_{i=1}^N$ and optimize the parameters of a student $p(\mathbf{x}) = \sum_{i=1}^n \sum_{j=i}^n \mathbf{M}_{ij} \mathbf{x}_i \mathbf{x}_j + \sum_{i=1}^n \mathbf{m}_i \mathbf{x}_i$ on them (see Section E.1 for more details).
- Or we consider that we are dealing with a compressed sensing problem, with the sparse signal $\mathbf{a}^* \in \mathbb{R}^n$ and the measurements given by $q(\mathbf{x}) \in \mathbb{R}^n$ for all $\mathbf{x} \in \mathbb{R}^m$. We optimized this version and observed grokking (Figure 41).

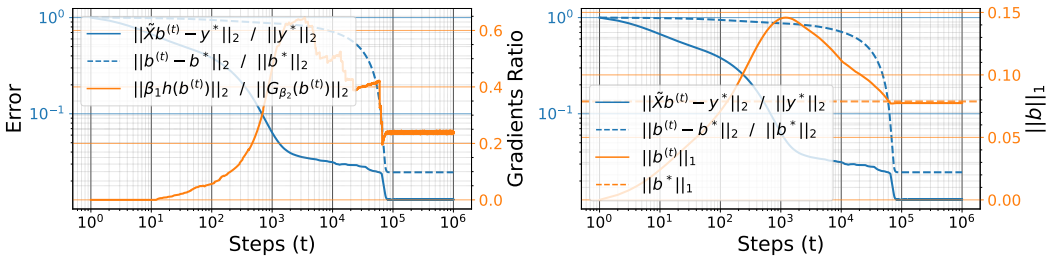


Figure 41: Reconstruction of a sparse polynomial $p^*(\mathbf{x}) = \sum_{i=1}^m \sum_{j=i}^m \mathbf{M}^*_{ij}\mathbf{x}_i\mathbf{x}_j + \sum_{i=1}^m \mathbf{m}^*_i\mathbf{x}_i$.

D Tensor Factorization

D.1 Matrix Sensing

Matrix sensing seeks to recover a low rank matrix $\mathbf{A}^* \in \mathbb{R}^{n_1\times n_2}$ from N measurement matrices $\{\mathbf{X}_i \in \mathbb{R}^{n_1\times n_2}\}_{i\in[N]}$ and measures $\mathbf{y}^* = \left(\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*)\right)_{i\in[N]}$. We have $\mathbf{y}^*_i = \mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) = \mathrm{vec}(\mathbf{X}_i)^\top \mathrm{vec}(\mathbf{A}^*) = \mathcal{F}_{\mathrm{vec}(\mathbf{A}^*)}(\mathrm{vec}(\mathbf{X}_i))$. This gives us a compressed sensing problem, with the signal vector $\mathrm{vec}(\mathbf{A}^*) \in \mathbb{R}^{n_1n_2}$ and the measurement matrix $\mathbf{X} = [\mathrm{vec}(\mathbf{X}_i)]_{i\in[N]} \in \mathbb{R}^{N\times n_1n_2}$. In fact, under full SVD $\mathbf{A}^* = \mathbf{U}^*\Sigma^*\mathbf{V}^{*\top}$, we have $\mathbf{a}^* = \mathrm{vec}(\mathbf{A}^*) = \Phi\mathbf{b}^*$; where $\mathbf{b}^* = \mathrm{vec}(\Sigma^*) \in \mathbb{R}^{n_1n_2}$, which is sparse since $\|\mathbf{b}^*\|_0 = \mathrm{rank}(\mathbf{A}^*) \le \min(n_1, n_2) \ll n_1n_2$; and $\Phi = \mathbf{V}^* \otimes \mathbf{U}^* \in \mathbb{R}^{n_1n_2\times n_1n_2}$, which has orthonormal column since $\Phi^\top\Phi = \left(\mathbf{V}^{*\top}\mathbf{V}^*\right) \otimes \left(\mathbf{U}^{*\top}\mathbf{U}^*\right) = \mathbb{I}_{n_1n_2}$. We have $\tilde{\mathbf{X}} = \mathbf{X}\Phi$.

D.2 Matrix Completion

For a matrix completion problem with matrix $\mathbf{A}^* \in \mathbb{R}^{n_1\times n_2}$, we have N measurement vectors $\left(\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}\right) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and measures $\mathbf{y}^*_i = \mathbf{X}_i^{(1)\top}\mathbf{A}^*\mathbf{X}_i^{(2)} = \left(\mathbf{X}_i^{(2)} \otimes \mathbf{X}_i^{(1)}\right)^\top \mathrm{vec}(\mathbf{A}^*) = \mathcal{F}_{\mathrm{vec}(\mathbf{A}^*)}\left(\mathbf{X}_i^{(2)} \otimes \mathbf{X}_i^{(1)}\right)$, i.e. $\mathbf{y}^* = \left(\mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)}\right)\mathrm{vec}(\mathbf{A}^*) = \mathcal{F}_{\mathrm{vec}(\mathbf{A}^*)}\left(\mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)}\right)$. This gives us a compressed sensing problem, with the signal vector $\mathrm{vec}(\mathbf{A}^*) \in \mathbb{R}^{n_1n_2}$ and the measurement matrix $\mathbf{X} = \mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)} \in \mathbb{R}^{N\times n_1n_2}$. Standard matrix completion is usually defined as recovering missing elements of a higher-order tensor from its incomplete observation. This is equivalent to requiring $\mathbf{X}_i^{(k)}$ to be selection vectors for all $k \in [2]$, i.e. $\mathbf{X}_i^{(k)}$ is the $s(i,k)^{\text{th}}$ vector of the canonical basis of $\mathbb{R}^{n_k}$ for a certain $s(i,k) \in [n_k]$. This make each $\mathbf{X}_i = \mathbf{X}_i^{(2)} \otimes \mathbf{X}_i^{(1)}$ a selection vector in $\mathbb{R}^n$, and $\mathbf{X} = \mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)}$ a selection matrix in $\mathbb{R}^{N\times n}$, so that $\mathbf{y}^*_i = \mathbf{A}^*_{s(i,1),s(i,2)} \forall i \in [N]$. So, in this formulation, each $\mathbf{X}_i^{(k)}$ is a sample from the columns of $\mathbb{I}_{n_k}$. Note that under a change of basis $\tilde{\mathbf{X}}_i^{(k)} = \mathbf{P}^{(k)}\mathbf{X}_i^{(k)}$, we have $\tilde{\mathbf{y}}^*_i = \left(\otimes_{k=1}^K \mathbf{P}^{(k)}\right)\mathbf{y}^*_i$, that is $\tilde{\mathbf{y}}^* = \mathbf{y}^*\left(\otimes_{k=1}^K \mathbf{P}^{(k)}\right)^\top$. A less standard formulation of the matrix completion task requires each $\mathbf{X}_i^{(k)}$ to be a sample from an orthonormal basis, i.e., $\mathbf{X}_i^{(k)}$ is a sample from the columns of $\mathbf{V}^{(k)} \in \mathbb{R}^{n_k\times n_k}$ with $\mathbf{V}^{(k)\top}\mathbf{V}^{(k)} = \mathbb{I}_{n_k}$. We let $\mathbf{X}_i^{(k)}$ be the $s(i,k)^{\text{th}}$ column of $\mathbf{V}^{(k)}$ for a certain $s(i,k) \in [n_k]$. Then $\mathbf{y}^*_i = \tilde{\mathbf{A}}^*_{s(i,1),\cdots,s(i,K)}$ with $\tilde{\mathbf{A}}^* = \mathbf{A}^* \times_1 \mathbf{V}^{(1)} \times_2 \mathbf{V}^{(2)}$. So, any result state of $\mathbf{A}^*$ in the standard formulation where the measurement vectors are selection vectors is valid for the tensor $\tilde{\mathbf{A}}^*$.

If we switch to a tensor $\mathbf{A}^* \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_K}$, we will have N vectors of measurements $\left(\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}, \cdots, \mathbf{X}_i^{(K)}\right) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_K}\ \forall i \in [N]$ and the measures $\mathbf{y}^*_i = \sum_{j_1,j_2,\cdots,j_K} \mathbf{A}_{j_1,\cdots,j_K}\mathbf{X}^{(1)}_{i,j_1}\mathbf{X}^{(2)}_{i,j_2}\cdots\mathbf{X}^{(K)}_{i,j_K} = \left(\mathbf{X}_i^{(K)} \otimes \mathbf{X}_i^{(K-1)} \otimes \cdots \otimes \mathbf{X}_i^{(1)}\right)^\top \mathtt{vecc}(\mathbf{A})$, i.e. $\mathbf{y}^* = \mathbf{X}\mathtt{vecc}(\mathbf{A}^*)$ with $\mathtt{vecc}(\mathbf{A}^*) = \mathbf{A}^{*(K\ldots1)} \in \mathbb{R}^n$ and $\mathbf{X} = \mathbf{X}^{(K)} \bullet \mathbf{X}^{(K-1)} \bullet \cdots \bullet \mathbf{X}^{(1)} \in \mathbb{R}^{N\times n}$; $n = \prod_{k=1}^K n_k$. Standard tensor completion is usually defined as recovering missing elements of a higher-order tensor from its incomplete observation. This is equivalent to re-

quiring $\mathbf{X}_i^{(k)}$ to be selection vectors for all $k \in [K]$, i.e. $\mathbf{X}_{i,j}^{(k)} = \delta_{j,s(i,k)} \quad \forall i,j$ for a certain $s(i,k) \in [n_k]$ ($\mathbf{X}_i^{(k)}$ is the $s(i,k)^{th}$ vector of the canonical basis of $\mathbb{R}^{n_k}$). This make each $\mathbf{X}_i = \otimes_{k=K}^{1} \mathbf{X}_i^{(k)}$ a selection vector in $\mathbb{R}^n$, and $\mathbf{X} = \bullet_{k=K}^{1} \mathbf{X}^{(k)}$ a selection matrix in $\mathbb{R}^{N \times n}$, so that $\mathbf{y}_i^* = \mathbf{A}^*_{s(i,1),\cdots,s(i,K)} \forall i \in [N]$. So, in this formulation, each $\mathbf{X}_i^{(k)}$ is a sample from the columns of $\mathbb{I}_{n_k}$. Note that under a change of basis $\tilde{\mathbf{X}}_i^{(k)} = \mathbf{P}^{(k)} \mathbf{X}_i^{(k)}$, we have $\tilde{\mathbf{y}}_i^* = \left(\otimes_{k=1}^{K} \mathbf{P}^{(k)}\right) \mathbf{y}_i^*$, that is $\tilde{\mathbf{y}}^* = \mathbf{y}^* \left(\otimes_{k=1}^{K} \mathbf{P}^{(k)}\right)^\top$. A less standard formulation of the tensor completion task requires each $\mathbf{X}_i^{(k)}$ to be a sample from an orthonormal basis $\mathbf{V}^{(k)} = \{\mathbf{v}_k^{(n_k)}\}_{k \in [n_k]}$ (i.e. ${\mathbf{v}_i^{(n_k)}}^\top \mathbf{v}_j^{(n_k)} = \delta_{ij}$). We let $\mathbf{X}_i^{(k)} = \mathbf{v}_{s(i,k)}^{(n_k)} \quad \forall i$ for a certain $s(i,k) \in [n_k]$. We can write $\mathbf{v}_\ell^{(n_k)} = \mathbf{P}^{(k)} \mathbf{e}_\ell^{(n_k)}$ with $\mathbf{P}^{(k)} \equiv \mathbf{V}^{(k)} \in \mathbb{R}^{n_k \times n_k}$ the base change matrix from the canonical basis to $\mathbf{V}^{(k)}$, which contains in each column ℓ the coordinate of $\mathbf{v}_\ell^{(n_k)}$ in $\{\mathbf{e}_k^{(n_k)}\}_{k \in [n_k]}$. So $\mathbf{X}_i^{(k)} = \mathbf{P}^{(k)} \mathbf{e}_{s(i,k)}^{(n_k)}$, and $\mathbf{y}_i^* = \left(\otimes_{k=K}^{1} \mathbf{X}_i^{(k)}\right)^\top \texttt{vecc}(\mathbf{A}^*) = \left(\otimes_{k=K}^{1} \left(\mathbf{P}^{(k)} \mathbf{e}_{s(i,k)}^{(n_k)}\right)\right)^\top \texttt{vecc}(\mathbf{A}^*) = \left(\left(\otimes_{k=K}^{1} \mathbf{P}^{(k)}\right) \left(\otimes_{k=K}^{1} \mathbf{e}_{s(i,k)}^{(n_k)}\right)\right)^\top \texttt{vecc}(\mathbf{A}^*) = \left(\otimes_{r=K}^{1} \mathbf{e}_{s(i,r)}^{(n_k)}\right)^\top \left(\otimes_{r=K}^{1} \mathbf{P}^{(k)}\right)^\top \texttt{vecc}(\mathbf{A}^*) = \left(\otimes_{k=K}^{1} \mathbf{e}_{s(i,k)}^{(n_k)}\right)^\top \texttt{vecc}\left(\tilde{\mathbf{A}}^*\right) = \tilde{\mathbf{A}^*}_{s(i,1),\cdots,s(i,K)}$ with $\tilde{\mathbf{A}}^* = \mathbf{A}^* \times_1 \mathbf{P}^{(1)} \times_2 \cdots \times_K \mathbf{P}^{(K)}$. So, any result state of $\mathbf{A}^*$ in the standard formulation where the measurement vectors are selection vectors is valid for the tensor $\tilde{\mathbf{A}}^*$.

Let us assume $K = 2$ in the following. Assume the target matrix $\mathbf{A}^*$ has rank r. Then it has $r(n_1 + n_2 - r)$ degree of freedom[5], and we need to observe at least $r(n_1 + n_2 - r)$ entries for perfect recovery. This bound can be improved by considering the structure of $\mathbf{A}^*$. Let $\mathbf{A}^* = \mathbf{U}^* \Sigma^* {\mathbf{V}^*}^\top$ be the **full** SVD of $\mathbf{A}^*$. As observed above, we are dealing with a compressed sensing problem with the signal vector $\mathbf{a}^* = \texttt{vecc}(\mathbf{A}^*) = \Phi \mathbf{b}^*$; where $\mathbf{b}^* = \texttt{vecc}(\Sigma^*) \in \mathbb{R}^{n_1 n_2}$, which is sparse since $\|\mathbf{b}^*\|_0 = r \leq \min(n_1, n_2) \ll n_1 n_2$; and $\Phi = \mathbf{V}^* \otimes \mathbf{U}^* \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$, which has orthonormal column since $\Phi^\top \Phi = \left({\mathbf{V}^*}^\top \mathbf{V}^*\right) \otimes \left({\mathbf{U}^*}^\top \mathbf{U}^*\right) = \mathbb{I}_{n_1 n_2}$. We have $\tilde{\mathbf{X}} = \mathbf{X}\Phi = \tilde{\mathbf{X}}^{(2)} \bullet \tilde{\mathbf{X}}^{(1)}$ with $\tilde{\mathbf{X}}^{(1)} = \mathbf{X}^{(1)} \mathbf{U}^*$ and $\tilde{\mathbf{X}}^{(2)} = \mathbf{X}^{(2)} \mathbf{V}^*$[6].

D.3 General framework

Given a low rank r matrix $\mathbf{A}^* \in \mathbb{R}^{n_1 \times n_2}$, a measurement matrix $\mathbf{X} \in \mathbb{R}^{N \times n_1 n_2}$; we aim to solve the following problem for $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$;

$$(P_4) \text{ Minimize } \operatorname{rank}(\mathbf{A}) \text{ subject to } \|\mathcal{F}_{\text{vec}(\mathbf{A})}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon \tag{94}$$

where $\mathbf{y}^* = \mathcal{F}_{\text{vec}(\mathbf{A}^*)}(\mathbf{X}) + \boldsymbol{\xi}$ are the measures and ϵ an upper bound on the size of the error term $\boldsymbol{\xi} \in \mathbb{R}^N$, $\|\boldsymbol{\xi}\|_2 \leq \epsilon$. As in the compressed sensing problem, this is NP-hard. The usual convex approach for matrix completion is to solve the following problem since the trace norm is a convex relaxation of the rank,

$$(P_5) \text{ Minimize } \|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A}) \text{ subject to } \|\mathcal{F}_{\text{vec}(\mathbf{A})}(\mathbf{X}) - \mathbf{y}^*\|_2 \leq \epsilon \tag{95}$$

We find the minimum nuclear norm solution since it is equivalent to minimizing the ℓ_1 norm of the corresponding sparse $\mathbf{b}$ in the sparse basis (the tensor product of the right and left singular vectors) for the solution $\mathbf{A}$ (low-rank solution). That said, many results obtained for compressed sensing can be translated to matrix completion. The main difference from standard compressed sensing is that the sparse basis is optimized jointly (and implicitly) with the signal's coordinate in that basis.

[5]The first r columns of $\mathbf{U}^*$ form an orthonormal basis for a r-dimensional subspace of $\mathbb{R}^{n_1}$ (the columns space of $\mathbf{A}^*$). Specifying this requires $r(n_1 - r)$ parameters. Similarly, the first r columns of $\mathbf{V}^*$ form an orthonormal basis for a r-dimensional subspace of $\mathbb{R}^{n_2}$ (the rows space of $\mathbf{A}^*$), and specifying this requires $r(n_2 - r)$ parameters. The r non-zero singular values are independent parameters. Thus, specifying them requires r parameters.

[6]$\tilde{\mathbf{X}} = \left(\mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)}\right)(\mathbf{V}^* \otimes \mathbf{U}^*) = \tilde{\mathbf{X}}^{(2)} \bullet \tilde{\mathbf{X}}^{(1)}$ since $\tilde{\mathbf{X}}_i = (\mathbf{V}^* \otimes \mathbf{U}^*)^\top \left(\mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)}\right)_i = \left({\mathbf{V}^*}^\top \otimes {\mathbf{U}^*}^\top\right)\left(\mathbf{X}_i^{(2)} \otimes \mathbf{X}_i^{(1)}\right) = \left({\mathbf{V}^*}^\top \mathbf{X}_i^{(2)}\right) \otimes \left({\mathbf{U}^*}^\top \mathbf{X}_i^{(1)}\right) = \left(\mathbf{V}^* \mathbf{X}^{(2)}\right)_i \otimes \left(\mathbf{U}^* \mathbf{X}^{(1)}\right)_i = \tilde{\mathbf{X}}_i^{(2)} \otimes \tilde{\mathbf{X}}_i^{(1)}$

D.4 THE CONTROL PARAMETERS

In this sub-section, we assume standard matrix completion. But the theories outlined here also apply to the general framework. The theory gives the minimal number of observations that guarantee $\mathbf{A}^*$ to be a unique solution to problem (P_5) and allow perfect recovery of $\mathbf{A}^*$ with fewer samples (Candès & Tao, 2010; Candes & Recht, 2012; Chen et al., 2014). Generally, the lower bound on N looks like $N \geq C \max(n_1, n_2)^\beta \left(r^\gamma \log^\alpha (\max(n_1, n_2)) + \log \frac{1}{\eta} \right)$ where η is the percentage of error (i.e N guaranteed perfect recovery with probability at least $1 - \eta$), $\alpha > 0$, $\beta > 0$, $\gamma > 0$ are constant, and $C > 0$ a universal constant. For example, in Candes & Recht (2012), $(\alpha, \beta, \gamma) = (1, 1.2, 1)$ for small rank $r \leq \max(n_1, n_2)^{0.2}$, and $\beta = 1.25$ for any rank. The term $\max(n_1, n_2) \log (\max(n_1, n_2))$ is due to the coupon collector effect since to recover an unknown matrix, one needs at least one observation per row and one observation per column (Candes & Recht, 2012).

Definition D.1 (Random orthogonal model (Candes & Recht, 2012))**.** For a given r, we generate two orthonormal matrices $\mathbf{U}^* \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V}^* \in \mathbb{R}^{n_2 \times r}$ with columns selected uniformly at random among all families of r orthonormal vectors; and a diagonal matrix Σ^* with only the first r diagonal element non-zero (with no assumptions about the singular values), then set $\mathbf{A}^* = \mathbf{U}^* \Sigma^* \mathbf{V}^{*\top}$.

Unless otherwise specified, we default the nonzero singular values to 1. We have the following result about the standard formulation for such matrices under the absence of noise.

Theorem D.1 (Theorem 1.1, Candes & Recht (2012))**.** *Let $\mathbf{A}^* \in \mathbb{R}^{n_1 \times n_2}$ be a matrix of rank r sampled from the random orthogonal model, and put $n = \max(n_1, n_2)$. Suppose we observe N entries of $\mathbf{A}^*$ with locations sampled uniformly at random. Then there are numerical constants C and c such that if $N \geq C n^{5/4} r \log(n)$, the minimizer to the problem (P_5) is unique and equal to $\mathbf{A}^*$ with probability at least $1 - c/n^3$; that is to say, the semidenite program (P_5) recovers all the entries of $\mathbf{A}^*$ with no error. In addition, if $r \leq n^{1/5}$, then the recovery is exact with probability at least $1 - c/n^3$ provided that $N \geq C n^{6/5} r \log(n)$.*

Assume for example $\mathbf{A}^* = \mathbf{e}_k^{(n_1)} \mathbf{e}_\ell^{(n_2)}$ for $(k, \ell) \in [n_1] \times [n_2]$. Even if this matrix ranks at 1, it has only zeros everywhere except 1 at position (i, j), so we have very little chance of reconstructing it in a high dimension by observing a portion of its inputs. The only way to guarantee observation of the input at position (i, j) is to choose measurements coherently with its singular basis $\mathbf{e}_k^{(n_2)} \otimes \mathbf{e}_\ell^{(n_1)}$. This idea is formulated more generally below.

Definition D.2. Let U be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U$ be the orthogonal projection onto U. Then, the coherence of U vis-a-vis a basis $\{\mathbf{u}_i^{(n)}\}_{i \in [n]}$ is defined by $\mu(U) = \frac{n}{r} \max_i \|\mathbf{P}_U \mathbf{u}_i^{(n)}\|^2$. We have $1 \leq \mu(U) \leq n/r$ (Candes & Recht, 2012).

For a matrix $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^\top \in \mathbb{R}^{n_1 \times n_2}$ under the **compact** SVD, the projection on the left singular value is $\mathbf{x} \to \mathbf{U}\mathbf{U}^\top \mathbf{x}$, and $\|\mathbf{U}\mathbf{U}^\top \mathbf{x}\|_2^2 = \|\mathbf{U}^\top \mathbf{x}\|_2^2$ for all $\mathbf{x}$ (similarly for the right singular value). We have the following definition of coherence, which considers each matrix entry.

Definition D.3 (Local coherence & Leverage score)**.** Let $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^\top \in \mathbb{R}^{n_1 \times n_2}$ be the **compact** SVD of a matrix $\mathbf{A}$ of rank r. The local coherences of $\mathbf{A}$ are defined by

$$\begin{aligned} \mu_i(\mathbf{A}) &= \frac{n_1}{r} \|\mathbf{U}^\top \mathbf{e}_i^{(n_1)}\|^2 = \frac{n_1}{r} \|\mathbf{U}_{i,:}\|^2 \quad \forall i \in [n_1] \\ \nu_j(\mathbf{A}) &= \frac{n_2}{r} \|\mathbf{V}^\top \mathbf{e}_j^{(n_2)}\|^2 = \frac{n_2}{r} \|\mathbf{V}_{j,:}\|^2 \quad \forall j \in [n_2] \end{aligned} \tag{96}$$

with μ_i for row i and ν_j for row j.

The quantities $\|\mathbf{U}^\top \mathbf{e}_i^{(n_1)}\|^2$ and $\|\mathbf{V}^\top \mathbf{e}_i^{(n_2)}\|^2$ are the leverage score of $\mathbf{A}$ (Chen et al., 2014), which indicate how "aligned" each row or column of the original data matrix is with the principal components (the columns of $\mathbf{U}$ or $\mathbf{V}$). For each row i, $\mu_i(A)$ measures how much this row vector projects onto the subspace spanned by the first r left singular vectors in $\mathbf{U}$. Rows with high leverage scores contribute more to the low-rank structure of $\mathbf{A}$ and are more "influential" in representing $\mathbf{A}$. Similarly, $\nu_j(A)$ measures the coherence of each column j in $\mathbf{A}$ with respect to the low-rank subspace formed by the right singular vectors in $\mathbf{V}$. High values indicate columns well-aligned with the principal directions of $\mathbf{A}$ and play a significant role in capturing its structure. Matrices with uniformly low coherence scores have rows and columns that are evenly influential. In contrast, matrices with high coherence

scores for certain rows or columns have a few specific rows or columns that dominate the low-rank structure.

In the general formulation, this definition can be extended to the set from which the measures are chosen. But in general, it leads back to the standard formulation under the change of basis.

Definition D.4 (Generalize local coherence & Leverage score)**.** We generalize the notion of coherence to any arbitrary set of vectors $\mathbf{U}^{(n_1)} = \{\mathbf{u}_i^{(n_1)}\}_{i\in[N_1]} \in \mathbb{R}^{n_1\times N_1}$ and $\mathbf{V}^{(n_2)} = \{\mathbf{v}_j^{(n_2)}\}_{j\in[N_2]} \in \mathbb{R}^{n_2\times N_2}$, and defined the generalized local coherences as

$$\begin{aligned}\mu_i(\mathbf{A}) &= \frac{n_1}{r}\|\mathbf{U}^\top \mathbf{u}_i^{(n_1)}\|^2 \quad \forall i \in [N_1]\\ \nu_j(\mathbf{A}) &= \frac{n_2}{r}\|\mathbf{V}^\top \mathbf{v}_j^{(n_2)}\|^2 \quad \forall j \in [N_2]\end{aligned} \tag{97}$$

Suppose the sets $\mathbf{U}^{(n_1)}$ and $\mathbf{V}^{(n_2)}$ are be orthonormal basis (i.e. $(N_1, N_2) = (n_1, n_2)$, ${\mathbf{u}_i^{(n_2)}}^\top \mathbf{u}_k^{(n_2)} = \delta_{ik}$ and ${\mathbf{v}_j^{(n_1)}}^\top \mathbf{v}_l^{(n_1)} = \delta_{jl}$). We can write $\mathbf{u}_i^{(n_1)} = \mathbf{P}^{(1)}\mathbf{e}_i^{(n_1)}$ and $\mathbf{v}_j^{(n_2)} = \mathbf{P}^{(2)}\mathbf{e}_j^{(n_2)}$ with $\mathbf{P}^{(k)} \in \mathbb{R}^{n_k\times n_k}$ the base change matrix from the canonical basis to $\mathbf{U}^{(n_1)}$ and $\mathbf{V}^{(n_2)}$ respectively. So

$$\begin{aligned}\mu_i(\mathbf{A}) &= \frac{n_1}{r}\|\mathbf{U}^\top\mathbf{P}^{(1)}\mathbf{e}_i^{(n_1)}\|^2 = \frac{n_1}{r}\|\tilde{\mathbf{U}}^\top\mathbf{e}_i^{(n_1)}\|^2 = \mu_i(\tilde{\mathbf{A}}) \quad \forall i \in [N_1]\\ \nu_j(\mathbf{A}) &= \frac{n_2}{r}\|\mathbf{V}^\top\mathbf{P}^{(2)}\mathbf{e}_j^{(n_2)}\|^2 = \frac{n_2}{r}\|\tilde{\mathbf{V}}^\top\mathbf{e}_j^{(n_2)}\|^2 = \nu_i(\tilde{\mathbf{A}}) \quad \forall j \in [N_2]\end{aligned} \tag{98}$$

with $\tilde{\mathbf{A}} = \mathbf{A}\times_1\mathbf{P}^{(1)}\times_2\mathbf{P}^{(2)} = \mathbf{P}^{(1)\top}\mathbf{A}\mathbf{P}^{(2)} = \mathbf{P}^{(1)\top}\mathbf{U}\Sigma\left(\mathbf{P}^{(2)\top}\mathbf{V}\right)^\top = \tilde{\mathbf{U}}\Sigma\tilde{\mathbf{V}}^\top$. That said, any result stated in the standard formulation for $\mathbf{A}$ is valid for $\tilde{\mathbf{A}}$ under the general orthonormal formulation.

Candès & Tao (2010) and Candes & Recht (2012) used mainly an upper bound μ_0 on μ_i and ν_i; $\mu_0 \geq \max\left(\max_{i\in[n_1]}\mu_i(\mathbf{A}^*), \max_{i\in[n_2]}\nu_i(\mathbf{A}^*)\right)$, and define a constant μ_1 such that the $\max_{i,j}[\mathbf{U}^*\mathbf{V}^{*\top}]_{ij} = \max_{i,j}\sum_k \mathbf{U}^*_{i,k}\mathbf{V}^*_{j,k} \leq \mu_1\sqrt{\frac{r}{n_1n_2}}$. Since $\left|\sum_k \mathbf{U}^*_{i,k}\mathbf{V}^*_{j,k}\right| \leq \sqrt{\sum_k \mathbf{U}^{*2}_{i,k}}\sqrt{\sum_k \mathbf{V}^{*2}_{j,k}} = \|\mathbf{U}^*_{i,:}\|_2\|\mathbf{V}^*_{j,:}\|_2 = \frac{r}{\sqrt{n_1n_2}}\sqrt{\mu_i(\mathbf{A}^*)\nu_j(\mathbf{A}^*)} \leq \frac{r}{\sqrt{n_1n_2}}\mu_0$ for all i, j; we can just take $\mu_1 \geq \mu_0\sqrt{r}$. From this, Candes & Recht (2012) show that if the coherence μ_0 is low, few samples are required to recover $\mathbf{A}^*$.

Theorem D.2 (Theorem 1.3, Candes & Recht (2012))**.** *Let $\mathbf{A}^* \in \mathbb{R}^{n_1\times n_2}$ be a matrix of rank r sampled from the random orthogonal model, and put $n = \max(n_1, n_2)$. Suppose we observe N entries of $\mathbf{A}^*$ with locations sampled uniformly at random. Then there are numerical constants C and c such that if $N \geq C\max\left(\mu_1^2, \mu_0^{\frac{1}{2}}\mu_1, \mu_0 n^{\frac{1}{4}}\right) nr\beta\log(n)$ for some $\beta > 2$, the minimizer to the problem (P_5) is unique and equal to $\mathbf{A}^*$ with probability at least $1 - c/n^3$. In addition, if $r \leq n^{1/5}/\mu_0$, then the recovery is exact with probability at least $1 - c/n^3$ provided that $N \geq C\mu_0 n^{6/5} r\beta\log(n)$.*

Chen et al. (2014) show that sampling the element at position (i, j) with probability $p_{ij} \in \Omega(\mu_i + \nu_j)$ allows perfect recovery of $\mathbf{A}^*$ with fewer samples, and called such sampling strategies *local coherence sampling*.

Theorem D.3 (Theorem 3.2 and Corollary 3.3, Chen et al. (2014))**.** *Let $\mathbf{A}^* \in \mathbb{R}^{n_1\times n_2}$ be a matrix of rank r with local coherence $\{\mu_i, \nu_j\}_{i\in[n_1], j\in[n_2]}$. There are universal constant $c_0, c_1, c_2 > 0$ such that if each element (i, j) is independently observed with probability $p_{ij} \geq \max\left\{\min\left\{c_0\frac{(\mu_i+\nu_j)r\log^2(n_1+n_2)}{\min(n_1,n_2)}, 1\right\}, \frac{1}{\min(n_1,n_2)^{10}}\right\}$, then $\mathbf{A}^*$ is the unique optimal solution of the nuclear minimization problem (P_5) with probability at least $1 - c_1/(n_1+n_2)^{c_2}$, for a number of sample $N \in \mathcal{O}\left(\max(n_1, n_2)r\log^2(n_1+n_2)\right)$.*

Given N and $\tau \in [0, 1]$, to control the coherence,

- For matrix factorization, we select the first $N_1 = \tau N$ examples with the highest values of $\mu_i(\mathbf{A}^*) + \nu_j(\mathbf{A}^*)$, and select the remaining $(1-\tau)N$ examples uniformly among the

remaining. The positions selected are one-hot encoded in dimensions n_1 (for row positions) and n_2 (for column positions) to have $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively.

- For matrix sensing, we generate $\mathbf{X}^{(1)}$ (resp. $\mathbf{X}^{(2)}$) by taking the first $N_1 = \min(\lfloor \tau N \rfloor, n_1)$ (resp. $N_1 = \min(\lfloor \tau N \rfloor, n_2)$) rows from the first columns of $\mathbf{U}^*$ (resp. $\mathbf{V}^*$) and the elements of the remaining $N_2 = N - N_1$ rows iid from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = 1/n_1$ (resp. $\sigma = 1/n_2$).

The higher τ (and so N_1), the less incoherence between the measures (rows of $\mathbf{X} = \mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)}$) and $\Phi = \mathbf{V}^* \otimes \mathbf{U}^*$.

D.5 Linear programming

We fix $n_1 = n_2 = 10^2$ and $\boldsymbol{\xi} = 0$ (no noise) and solve for different (N, r, τ) the convex problem (P_5) using standard linear programming (we use the `cvxpy` library). As r and/or τ increases, the number of samples needs for perfect recovery **decreases**. The relative recovery error $\|\mathbf{A} - \mathbf{A}^*\|_2 / \|\mathbf{A}^*\|_2$ obtained is usually of the order of 10^{-6} and gives us a basis for comparison with other methods. We do not include figures to save space.

D.6 Subgradient descent

We write $\mathbf{a}$ for $\mathrm{vec}(\mathbf{A})$ and $\mathbf{b}$ for $\mathrm{vec}(\Sigma)$ under full SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top \in \mathbb{R}^{n_1 \times n_2}$. The matrix is $\mathbf{A}^* = \mathbf{U}^*\Sigma^*\mathbf{V}^{*\top} \in \mathbb{R}^{n_1 \times n_2}$, the signal is $\mathbf{a}^* = \mathrm{vec}(\mathbf{A}^*)$, the sparse basis is $\Phi = \mathbf{V}^* \otimes \mathbf{U}^* \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$, the sparse coordinates are $\mathbf{b}^* = \mathrm{vec}(\Sigma^*)$. Let $\mathbf{y}(\mathbf{A}) = \mathcal{F}_{\mathbf{a}}(\mathbf{X}) = \mathbf{X}\,\mathrm{vec}(\mathbf{A})$. We have $\mathbf{y}^* = \mathcal{F}_{\mathbf{a}^*}(\mathbf{X}) + \boldsymbol{\xi} = \mathcal{F}_{\mathbf{b}^*}(\tilde{\mathbf{X}}) + \boldsymbol{\xi}$, and want to minimize $f(\mathbf{A}) = g_{\beta_2}(\mathbf{A}) + \beta_* \|\mathbf{A}\|_*$ using gradient descent, where

$$\begin{aligned} g_{\beta_2}(\mathbf{A}) &:= \frac{1}{2}\|\mathbf{y}(\mathbf{A}) - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{A}\|_{\mathrm{F}} \\ &= \frac{1}{2}\mathbf{a}^\top \mathbf{X}^\top \mathbf{X}\mathbf{a} - \mathbf{y}^{*\top}\mathbf{X}\mathbf{a} + \frac{1}{2}\mathbf{y}^{*\top}\mathbf{y}^* + \frac{\beta_2}{2}\mathbf{a}^\top\mathbf{a} \\ &= \begin{cases} \frac{1}{2}\mathbf{a}^\top \left(\mathbf{X}^\top\mathbf{X} + \beta_2 \mathbb{I}_n\right)\mathbf{a} - \left(\mathbf{X}^\top\mathbf{X}\mathbf{a}^* + \mathbf{X}^\top\boldsymbol{\xi}\right)^\top \mathbf{a} + \frac{1}{2}\|\mathbf{X}\mathbf{a}^* + \boldsymbol{\xi}\|_2^2 \\ \frac{1}{2}(\mathbf{a} - \mathbf{a}^*)^\top \left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)(\mathbf{a} - \mathbf{a}^*) - \left(\mathbf{X}^\top\boldsymbol{\xi} - \beta_2\mathbf{a}^*\right)^\top (\mathbf{a} - \mathbf{a}^*) + \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 + \frac{\beta_2}{2}\|\mathbf{a}^*\|_2^2 \end{cases} \end{aligned} \tag{99}$$

We write $F(\mathbf{A}) := G_{\beta_2}(\mathbf{A}) + \beta_* h(\mathbf{A})$ with

$$\mathrm{vec}\, G_{\beta_2}(\mathbf{A}) := \nabla_{\mathbf{a}} g_{\beta_2}(\mathbf{A}) = \mathbf{X}^\top(\mathbf{y} - \mathbf{y}^*) + \beta_2\mathbf{a} = \begin{cases} \left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)\mathbf{a} - \left(\mathbf{X}^\top\mathbf{X}\mathbf{a}^* + \mathbf{X}^\top\boldsymbol{\xi}\right) \\ \left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)(\mathbf{a} - \mathbf{a}^*) - \left(\mathbf{X}^\top\boldsymbol{\xi} - \beta_2\mathbf{a}^*\right) \end{cases} \tag{100}$$

and $h(\mathbf{A}) \in \partial\|\mathbf{A}\|_* = \left\{\mathbf{U}\mathbf{V}^\top + \mathbf{W}, \|\mathbf{W}\|_{2\to 2} \le 1, \mathbf{U}^\top\mathbf{W} = 0, \mathbf{W}\mathbf{V} = 0\right\}$ any subgradient of $\|\mathbf{A}\|_*$, with $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ under the compact SVD [7]. We use $h(\mathbf{A}) = \mathbf{U}\mathbf{V}^\top$ for simplicity and without loss of generality.

Suppose we start at some $\mathbf{A}^{(1)} := \zeta \mathbb{I}_{n_1 \times n_2}$ or $\mathbf{A}^{(1)} \overset{iid}{\sim} \zeta\mathcal{N}(0, 1/n_1 n_2)$, with $\zeta \ge 0$ the initialization scale. Using $\mathbf{F}^{(t)} := F(\mathbf{A}^{(t)})$, the subgradient update rule is

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \alpha_t \mathbf{F}^{(t)} \quad \forall t > 1 \tag{101}$$

with α_t the learning rate at step t. Using $\mathbf{a} = \mathrm{vec}\,\mathbf{A}$, we have

$$\begin{aligned} \mathbf{a}^{(t+1)} &= \mathbf{a}^{(t)} - \alpha_t\, \mathrm{vec}\, F(\mathbf{A}^{(t)}) \\ &= \mathbf{a}^{(t)} - \alpha_t \left(\mathrm{vec}\, G_{\beta_2}(\mathbf{A}) + \beta_* \,\mathrm{vec}(h(\mathbf{A}))\right) \end{aligned} \tag{102}$$

That is, using $\mathbf{h}^{(t)} = \mathrm{vec}(h(\mathbf{A}^{(t)}))$,

$$\begin{cases} \mathbf{a}^{(t+1)} = \left[\mathbb{I}_n - \alpha_t\left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)\right]\mathbf{a}^{(t)} + \alpha_t\left(\mathbf{X}^\top\mathbf{X}\mathbf{a}^* + \mathbf{X}^\top\boldsymbol{\xi}\right) - \beta_*\alpha_t\mathbf{h}^{(t)} \\ \mathbf{a}^{(t+1)} - \mathbf{a}^* = \left[\mathbb{I}_n - \alpha_t\left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)\right](\mathbf{a}^{(t)} - \mathbf{a}^*) + \alpha_t\left(\mathbf{X}^\top\boldsymbol{\xi} - \beta_2\mathbf{a}^*\right) - \beta_*\alpha_t\mathbf{h}^{(t)} \end{cases} \tag{103}$$

[7] The norm $\|\mathbf{A}\|_*$ is not differentiable everywhere because the singular values of $\mathbf{A}$ can be non-differentiable at points where they have multiplicities (e.g., when the singular values are not distinct).

We let $f^* = f(\mathbf{A}^*) = \beta_*\|\mathbf{A}^*\|_* + \frac{\beta_2}{2}\|\mathbf{a}^*\|_2^2 + \|\boldsymbol{\xi}\|_2^2$ and $f^{(t)} = f(\mathbf{A}^{(t)})$. Since the subgradient method is not a descent method, we let $\mathbf{A}_{\text{best}}^{(t)} = \arg\min_{\mathbf{A}\in\left\{\mathbf{A}^{(t')}, t'\le t\right\}} f(\mathbf{A}) = \arg\min_{\mathbf{A}\in\left\{\mathbf{A}_{\text{best}}^{(t-1)}, \mathbf{A}^{(t)}\right\}} f(\mathbf{A})$ be the best point found so far at step t, and $f_{\text{best}}^{(t)} = f(\mathbf{A}_{\text{best}}^{(t)}) = \min\left\{f_{\text{best}}^{(t-1)}, f^{(t)}\right\}$. This $\mathbf{A}_{\text{best}}^{(t)}$ can be made η-optimal for an arbitrary precision η if the step rule is chosen appropriately, as the following theorem shows.

Theorem D.4. *Suppose there exists a constant $L > 0$ such that $\|F(\mathbf{A})\|_{\text{F}} \le L$ for all $\mathbf{A}$. Let $\mathbf{A}_{\text{best}}^{(t)} = \arg\min_{1\le t'\le t} f(\mathbf{A}^{(t')})$ and $f_{\text{best}}^{(t)} = f(\mathbf{A}_{\text{best}}^{(t)})$. Then, for every $T \ge 1$, $f_{\text{best}}^{(T)} - f(\mathbf{A}^*) \le \frac{\|\mathbf{A}^{(1)}-\mathbf{A}^*\|_{\text{F}}^2 + L^2\sum_{t=1}^T \alpha_t^2}{2\sum_{t=1}^T \alpha_t}$.*

Proof. Similar to Theorem C.3 □

That said, many step size rules lead to different accuracy.

Corollary D.1. *With a constant step size, $\alpha_t = \alpha$,*

$$f_{best}^{(T)} - f^* \le \frac{\|\mathbf{A}^{(1)} - \mathbf{A}^*\|_{\text{F}}^2 + L^2T\alpha^2}{2T\alpha} \longrightarrow_{T\to\infty} L^2\alpha/2 \tag{104}$$

With a square summable but not summable step size rule, $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$, we have

$$f_{best}^{(T)} - f^* \le \frac{\|\mathbf{A}^{(1)} - \mathbf{A}^*\|_{\text{F}}^2 + L^2\sum_{i=1}^T \alpha_i^2}{2\sum_{i=1}^T \alpha_i} \longrightarrow_{T\to\infty} 0 \tag{105}$$

As in section C.6,

- We let $\mathbf{X} = \mathbf{U}\Sigma^{\frac{1}{2}}\mathbf{V}^\top$ under the compact SVD decomposition, with $\Sigma = \text{diag}(\sigma_k)_{k\in[r]}$, where $r = \text{rank}(\mathbf{X})$ and $\sigma_{\max} = \sigma_1 \ge \cdots \sigma_k \ge \sigma_{k+1} \cdots \ge \sigma_{\min} = \sigma_r > \sigma_{r+1} = \cdots = 0$
- We assume the step size $\alpha_t = \alpha$ satisfies $0 < \alpha < \frac{2}{\sigma_{\max}+\beta_2}$.
- We define $\rho_p := \left\|\mathbb{I}_n - \alpha_t\left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)\right\|_{p\to p}$ for all $p > 0$.

D.6.1 Memorization

We will show that the update first moves to the least square solution of the problem, $\hat{\mathbf{a}} = \text{vec}\,\hat{\mathbf{A}} = \left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)^\dagger\mathbf{X}^\top\mathbf{y}^* = \mathbf{V}\left(\Sigma + \beta_2\mathbb{I}\right)^{-1}\left(\Sigma\mathbf{V}^\top\mathbf{b}^* + \Sigma^{\frac{1}{2}}\mathbf{U}^\top\boldsymbol{\xi}\right)$ (Theorem TODO). If β_* is too high, the subgradient term $h(\mathbf{A})$ dominates early, and there is no convergence, i.e., no memorization nor generalization (Theorem D.5). This $\hat{\mathbf{a}}$ can memorize (Theorem TODO), but cannot generalize for $N < n$ (Theorem TODO).

Theorem D.5 (Oscillatory Behavior for Large β_*)**.** *Let $\mathbf{A}^{(1)} \in \mathbb{R}^{n_1\times n_2}$ full rank. Consider the subgradient descent update*

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \alpha_t\left(\nabla_{\mathbf{A}} g_{\beta_2}(\mathbf{A}^{(t)}) + \beta_* h(\mathbf{A}^{(t)})\right) \tag{106}$$

with a fixed step size $\alpha_t = \alpha > 0$, where $g_{\beta_2}(\mathbf{A}) = \frac{1}{2}\|\mathbf{X}\,\text{vec}\,\mathbf{A} - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\|\mathbf{A}\|_F^2$ and $h(\mathbf{A}) \in \partial\|\mathbf{A}\|_*$. If $\beta_* > \frac{\sigma_{\max}+\beta_2}{\sqrt{\min(n_1,n_2)}}$ then the ℓ_*-term dominates the updates, causing the sequence $\mathbf{b}^{(t)}$ to exhibit oscillatory behavior without convergence to a minimizer of $f(\mathbf{A}) = g_{\beta_2}(\mathbf{A}) + \beta_*\|\mathbf{A}\|_1$.

Proof. We use lemma D.6 with $L = \|\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \beta_2\mathbb{I}_n\|_{2\to 2} = \sigma_{\max}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}) + \beta_2$ (operator norm) be the Lipschitz constant for $\text{vec}\,G_{\beta_2}(\mathbf{A}) = \text{vec}\,\nabla_{\mathbf{A}} g_{\beta_*}(\mathbf{A}) = \mathbf{X}^\top(\mathbf{X}\mathbf{a} - \mathbf{y}^*) + \beta_*\mathbf{a} = \left(\mathbf{X}^\top\mathbf{X} + \beta_2\mathbb{I}_n\right)\mathbf{a} - \left(\mathbf{X}^\top\mathbf{X}\mathbf{a}^* + \mathbf{X}^\top\boldsymbol{\xi}\right)$, since $\|\,\text{vec}\,G_{\beta_2}(\mathbf{U}) - \text{vec}\,G_{\beta_2}(\mathbf{V})\|_2 \le L\|\,\text{vec}\,\mathbf{U} - \text{vec}\,\mathbf{V}\|_2$ for all $\mathbf{U}, \mathbf{V}$.

When the data-fitting gradient $\nabla_{\mathbf{A}} g_{\beta_2}(\mathbf{A}^{(t)})$ is negligible, the singular direction of $\beta_* h(\mathbf{A}^{(t)})$ (which depends on the singular vectors of $\mathbf{A}^{(t)}$) can flip across iterations in a way that prevents stable convergence (see Theorem D.12).

□

Lemma D.6. *Let $f(\mathbf{A}) = g(\mathbf{A}) + \beta_1\|\mathbf{A}\|_*$ be a convex function from $\mathbb{R}^{n_1\times n_2}$ to $\mathbb{R}$ where g has a Lipschitz continuous gradient with Lipschitz constant $L > 0$, i.e., $\|\nabla g(\mathbf{U}) - \nabla g(\mathbf{V})\|_F \le L\|\mathbf{U}-\mathbf{V}\|_F$ for all $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n_1\times n_2}$. Consider the subgradient descent update*

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \alpha\left(\nabla g(\mathbf{A}^{(t)}) + \beta_* h(\mathbf{A}^{(t)})\right) \tag{107}$$

with a fixed step size $\alpha > 0$, where $h(\mathbf{A}^{(t)}) \in \partial\|\mathbf{A}^{(t)}\|_*$. If $\beta_1 > \frac{L}{\sqrt{\min(n_1,n_2)}}$ then the ℓ_*-term dominates the updates, causing the sequence $\{\mathbf{A}^{(t)}\}_{t>1}$ to exhibit oscillatory behavior without convergence to a minimizer of f. Consequently, neither memorization nor generalization is achieved, and both training and test errors oscillate above a suboptimal level.

Proof Sketch. Since g has a Lipschitz continuous gradient with constant L, $\|\nabla g(\mathbf{A}^{(t)})\|_{\mathrm{F}} \le L$ for all t when $\mathbf{A}^{(t)}$ is within a suitable bounded region. The subgradient $h(\mathbf{A}^{(t)})$ of $\|\mathbf{A}^{(t)}\|_*$ satisfy $\|h(\mathbf{A}^{(t)})\|_* \approx \sqrt{\min(n_1,n_2)}$ at the beginning of training (full rank matrix), so $\|h(\mathbf{A}^{(t)})\|_{\mathrm{F}} \ge \|h(\mathbf{A}^{(t)})\|_*/\operatorname{rank}(h(\mathbf{A}^{(t)})) \approx \sqrt{\min(n_1,n_2)}/\min(n_1,n_2) = \sqrt{\min(n_1,n_2)}$. If $\beta_* > \frac{L}{\sqrt{\min(n_1,n_2)}}$, then

$$\beta_*\|h(\mathbf{A}^{(t)})\|_{\mathrm{F}} > \beta_*\sqrt{\min(n_1,n_2)} > L \ge \|\nabla g(\mathbf{A}^{(t)})\|_{\mathrm{F}} \tag{108}$$

This inequality implies that the update is dominated by the ℓ_*-term:

$$\mathbf{A}^{(t+1)} \approx \mathbf{A}^{(t)} - \alpha\beta_* h(\mathbf{A}^{(t)}) \tag{109}$$

with the influence of $\nabla g(\mathbf{A}^{(t)})$ becoming negligible, making the iterates swing sharply depending on the current singular-vector configuration (see Theorem D.12). This "over-regularization" effect is akin to the ℓ_1 case in vector problems, where too large causes step-to-step sign flipping. In the matrix setting, it induces rank-structure flipping or oscillations.

□

Lemma D.7. *For all $p > 0$ such that $\rho_p < 1$, we have*

$$\|\mathbf{a}^{(t)} - \hat{\mathbf{a}}\|_p \le \rho_p^t\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_p + \alpha\beta_* n^{1/p}\frac{1-\rho_p^t}{1-\rho_p} \le \rho_p^t\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_p + \frac{\alpha\beta_* n^{1/p}}{1-\rho_p} \quad \forall t \ge 1 \tag{110}$$

In particular,

$$\|\mathbf{a}^{(t)} - \hat{\mathbf{a}}\|_2 \le \rho^t\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_2 + \alpha\beta_*\sqrt{n}\frac{1-\rho_2^t}{1-\rho_2} \le \rho^t\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_2 + \frac{\alpha\beta_*\sqrt{n}}{1-\rho_2} \quad \forall t \ge 1 \tag{111}$$

and

$$\|\mathbf{a}^{(t)} - \hat{\mathbf{a}}\|_\infty \le \rho_\infty^t\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_\infty + \alpha\beta_*\frac{1-\rho_\infty^t}{1-\rho_\infty} \le \rho_\infty^t\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_\infty + \frac{\alpha\beta_*}{1-\rho_\infty} \quad \forall t \ge 1 \tag{112}$$

Proof. The proof is similar to C.7, using tha fact that $\|\operatorname{vec}(h(\mathbf{A}^{(t)})\|_p \le (n_1n_2)^{1/p} = n^{1/p}$ for all and $p > 0$ (Lemma D.11) □

Theorem D.8. *Let $p > 0$ such that $\rho_p < 1$. Define*

$$t_1 := \left\lceil -\frac{\ln\left(1 + \frac{(1-\rho)\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_p}{\alpha\beta_* n^{1/p}}\right)}{\ln(\rho_p)} \right\rceil \tag{113}$$

Then for all $t \ge t_1$,

$$\|\mathbf{a}^{(t)} - \hat{\mathbf{a}}\|_p \le 2\alpha\beta_* n^{1/p}\frac{1-\rho_p^t}{1-\rho_p} \le 2\frac{\alpha\beta_* n^{1/p}}{1-\rho_p} \tag{114}$$

and the prediction error for $t \ge t_1$ is bounded by

$$\begin{aligned}\|\tilde{\mathbf{X}}\mathbf{a}^{(t)} - \mathbf{y}^*\|_p &\le 2\alpha\beta_* n^{1/p}\frac{1-\rho_p^t}{1-\rho_p}\|\mathbf{X}\|_{p\to p} + \|\mathbf{X}\hat{\mathbf{a}} - \mathbf{y}^*\|_p \\ &\le 2\frac{\alpha\beta_* n^{1/p}}{1-\rho_p}\|\mathbf{X}\|_{p\to p} + \|\mathbf{X}\hat{\mathbf{a}} - \mathbf{y}^*\|_p\end{aligned} \tag{115}$$

Proof. The proof is similar to C.8. □

Corollary D.2. *Let $p > 0$ such that $\rho_p < 1$. Define*

$$\tilde{t}_1 := \begin{cases} \left\lceil -\frac{\ln\left(\frac{(1-\rho)\|\mathbf{a}^{(1)}-\hat{\mathbf{a}}\|_p}{\alpha\beta_* n^{1/p}}\right)}{\ln(\rho_p)} \right\rceil & \text{if } \|\mathbf{a}^{(1)}-\hat{\mathbf{a}}\|_p > \frac{\alpha\beta_*}{1-\rho_p} \quad > t_1 \\ 0 & \text{otherwise} \end{cases} \tag{116}$$

Then for all $t \geq \tilde{t}_1$,

$$\|\mathbf{a}^{(t)} - \hat{\mathbf{a}}\|_p \leq 2\frac{\alpha\beta_* n^{1/p}}{1-\rho_p} \tag{117}$$

and the prediction error for $t \geq \tilde{t}_1$ is bounded by

$$\|\tilde{\mathbf{X}}\mathbf{a}^{(t)} - \mathbf{y}^*\|_p \leq \frac{2\alpha\beta_* n^{1/p}}{1-\rho_p}\|\tilde{\mathbf{X}}\|_{p\to p} + \|\tilde{\mathbf{X}}\hat{\mathbf{a}} - \mathbf{y}^*\|_p \tag{118}$$

Proof. The proof is similar to C.2 □

Theorem D.9. *Assume $\mathbb{E}[\boldsymbol{\xi}] = 0$ and $\mathrm{Cov}(\boldsymbol{\xi}) = \sigma_\xi^2 \mathbb{I}_N$. Then*

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\mathbf{X}\hat{\mathbf{a}} - \mathbf{y}^*\|_2^2\right] = \sum_{i=1}^{r}\left(\frac{\beta_2\sigma_i}{\sigma_i+\beta_2}\right)^2 (\mathbf{V}^\top\mathbf{a}^*)_i^2 + \sum_{i=1}^{r}\left(\frac{\beta_2}{\sigma_i+\beta_2}\right)^2 \sigma_\xi^2 + \sigma_\xi^2(N-r) \tag{119}$$

Proof. The proof is similar to C.9 □

Theorem D.10. *For $N < n$,*

$$\|\hat{\mathbf{a}} - \mathbf{a}^*\|_2^2 \geq \|(\mathbb{I}_n - \mathbf{V}\mathbf{V}^\top)\mathbf{a}^*\|_2^2 \tag{120}$$

In particular, if $\mathbf{a}^*$ has a nonzero component orthogonal to $\mathrm{Col}(\mathbf{V})$, then $\hat{\mathbf{a}}$ cannot perfectly generalize to $\mathbf{a}^*$.

Proof. The proof is similar to C.10 □

Lemma D.11. *Let $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}$. We have $\|\mathrm{vec}(\mathbf{H})\|_p \leq (n_1 n_2)^{1/p}$ for all $\mathbf{H} \in \partial\|\mathbf{A}\|_*$ and $p > 0$.*

Proof. Let $\mathbf{H} \in \partial\|\mathbf{A}\|_*$. Then $\|\mathbf{H}\|_{2\to 2} \leq 1$. So by the definition of the spectral (operator) norm, we have $\|\mathbf{H}\|_{2\to 2} = \sup_{\mathbf{x}\neq 0}\frac{\|\mathbf{H}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_{\max}(\mathbf{H}) \leq 1$. Taking $\mathbf{x} = \mathbf{e}_j^{(n_2)}$, the j-th standard basis vector in $\mathbb{R}^{n_2}$, we obtain $\|\mathbf{H}_{:,j}\|_2 = \|\mathbf{H}\mathbf{e}_j^{(n_2)}\|_2 \leq 1$; which implied $\mathbf{H}_{ij} \leq \|\mathbf{H}_{:,j}\|_2 \leq 1$. So $\|\mathrm{vec}(\mathbf{H})\|_p = \left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|\mathbf{H}_{ij}|^p\right)^{1/p} \leq (n_1 n_2)^{1/p}$.

□

D.6.2 Generalization

We now turn our attention to the generalization delay. We analyse how the iterate $\mathbf{A}^{(t)}$ transitions from memorizing the training data ($\mathbf{A}^{(t)} \approx \hat{\mathbf{A}}$) to converging toward the low rank ground truth $\mathbf{A}^*$. We focus on quantifying the additional number of iterations Δt required for this phase and bounding the generalization error $\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_\infty$ as $t \to \infty$.

Theorem D.12. *Given $\alpha > 0$ and $\mathbf{A}^{(1)} = \mathbf{U}^{(1)}\Sigma^{(1)}\mathbf{V}^{(1)\top} \in \mathbb{R}^{n_1\times n_2}$ (compact SVD) with $\Sigma = \mathrm{diag}\left(\sigma_1^{(1)}, \ldots, \sigma_{r_1}^{(1)}\right)$, let*

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \alpha\mathbf{U}^{(t)}\mathbf{V}^{(t)\top} = \mathbf{U}^{(t)}\left(\Sigma^{(t)} - \alpha\mathbb{I}_{r_t}\right)\mathbf{V}^{(t)\top} \text{ for all } t \geq 1 \tag{121}$$

where $r_t = \mathrm{rank}(\mathbf{A}^{(t)})$.

1. *A point* $\mathbf{A}$ *is stationary for this dynamical system if and only if* $\|\mathbf{A}\|_{2\to 2} = \sigma_{\max}(\mathbf{A}) < \alpha$.
2. $\|\mathbf{A}^{(t)}\|_{2\to 2} < \alpha$ *if and only if* $t > \lfloor \frac{\|\mathbf{A}^{(1)}\|_{2\to 2}}{\alpha} \rfloor$.
3. *For all* $t > \lfloor \frac{\|\mathbf{A}^{(1)}\|_{2\to 2}}{\alpha} \rfloor$, $r_t = \left| \{ i \quad | \quad \sigma_i^{(1)}/\alpha \in \mathbb{Z} \} \right|$.

Proof. Equation 121 writes

$$\begin{aligned} \mathbf{A}^{(t+1)} = \mathbf{U}^{(t+1)}\Sigma^{(t+1)}\mathbf{V}^{(t+1)\top} &= \sum_{i=1}^{r_{t+1}} \sigma_i^{(t+1)} \mathbf{U}_{:,i}^{(t+1)} \mathbf{V}_{:,i}^{(t+1)\top} \\ &= \sum_{i=1}^{r_t} (\sigma_i^{(t)} - \alpha) \mathbf{U}_{:,i}^{(t)} \mathbf{V}_{:,i}^{(t)\top} \\ &= \sum_{i=1}^{r_t} |\sigma_i^{(t)} - \alpha| \cdot \operatorname{sign}(\sigma_i^{(t)} - \alpha) \mathbf{U}_{:,i}^{(t)} \mathbf{V}_{:,i}^{(t)\top} \end{aligned} \tag{122}$$

This implies

$$\sigma_i^{(t+1)} = |\sigma_i^{(t)} - \alpha| \quad \forall i \in [r_1] \tag{123}$$

So starting at $\sigma_i^{(1)}$, each σ_i decay at each step by α until $\sigma^* := \sigma_i^{(t)} \in [0, \alpha)$, and start oscillating between σ_i^* and $\alpha - \sigma_i^*$. It starts doing so when $t > t_i := \lfloor \frac{\sigma_i^{(1)}}{\alpha} \rfloor$. We take $t = \max_i t_i$. □

Like in section C.6.2, after $t_1 := \left\lceil -\frac{\ln\left(1 + \frac{(1-\rho)\|\mathbf{a}^{(1)} - \hat{\mathbf{a}}\|_p}{\alpha\beta_* n^{1/p}}\right)}{\ln(\rho_p)} \right\rceil$, $\|\mathbf{a}^{(t)} - \hat{\mathbf{a}}\|_p \le 2\alpha\beta_* n^{1/p} \frac{1-\rho_p^t}{1-\rho_p} \le 2\frac{\alpha\beta_* n^{1/p}}{1-\rho_p}$ (Theorem D.8) and

$$\begin{aligned} \|\operatorname{vec} G_{\beta_2}(\mathbf{A}^{(t)})\|_p &= \|\operatorname{vec} G_{\beta_2}(\mathbf{A}^{(t)}) - \operatorname{vec} G_{\beta_2}(\hat{\mathbf{A}})\|_p \text{ since } G_{\beta_2}(\hat{\mathbf{A}}) = 0 \\ &\le \|\mathbf{X}^\top \mathbf{X} + \beta_2 \mathbb{I}_n\|_{p\to p} \|\operatorname{vec} \mathbf{A}^{(t)} - \operatorname{vec} \hat{\mathbf{A}}\|_p \\ &\le 2\alpha\beta_* n^{1/p} \|\mathbf{X}^\top \mathbf{X} + \beta_2 \mathbb{I}_n\|_{p\to p} \frac{1-\rho_p^t}{1-\rho_p} \\ &\le \frac{2\alpha\beta_* n^{1/p}}{1-\rho_p} \|\mathbf{X}^\top \mathbf{X} + \beta_2 \mathbb{I}_n\|_{p\to p} \end{aligned} \tag{124}$$

So, this gradient can be made much smaller than the subgradient term by choosing $\alpha\beta_*$ sufficiently small. After time t_1, the contribution of the gradient G_{β_2} to the update of $\mathbf{A}^{(t)}$ is dominated by the ℓ_*–regularization term. Specifically, the update rule approximates

$$\mathbf{A}^{(t+1)} \approx \mathbf{A}^{(t)} - \alpha\beta_* \mathbf{U}^{(t)} \mathbf{V}^{(t)\top} \tag{125}$$

By theorem D.12, this converge to a solution with operator norm bound by $\alpha\beta_*$ after additional $\Delta t = \Theta\left(\lfloor \frac{\sigma_{\max}(\hat{\mathbf{A}})}{\alpha\beta_*} \rfloor\right)$ steps. Note that when $\|\mathbf{A}^{(t)}\|_*$ becomes too small, $\mathbf{A}^{(t)} \approx \mathbf{A}^*$ since for problem of interest, the minimum nuclear norm solution that fits the data is $\mathbf{A}^*$ under the low-rank constraint $r = \operatorname{rank}(\mathbf{A}) \ll \min(n_1, n_2)$ (and the coherence assumptions on $\mathbf{X}$ with respect to the eigenbasis of $\mathbf{A}^*$). The smaller $\alpha\beta_*$, the longer it take to recover $\mathbf{A}^*$, and the smaller is the error $\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_\infty$ when $t \to \infty$. Like in linear sparse recovery, if β_2 is choose such that $\sigma_{\max}(\hat{\mathbf{A}}) \ll \alpha\beta_*$, then $\mathbf{A}^{(t)}$ will get stuck near $\hat{\mathbf{A}}$, and there will be no generalization after memorization. So, a bad choice of a non-zero β_2 can be detrimental to generalization (it is better to not use β_2 on that problem unless the initialization scale is nontrivial).

Generalization appends through a multiscale singular value decay phenomenon. The small singular value after memorization converges to $\{\sigma, 0 \le \sigma < \alpha\beta_*\}$, followed by the next smaller one until the larger one. So, for $N < n_1 n_2$, if we just regularize the Frobenius norm (standard ℓ_2) without regularizing the nuclear norm (ℓ_*), we can't reach the optimal solution. On the other hand, when N is large enough, regularizing the nuclear norm is sufficient.

By carefully choosing α and β_1, one can balance the speed of generalization (smaller Δt) with the accuracy of recovery (smaller $\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_\infty$). Appropriate step rule also guaranteed the converge of $\|\mathbf{b}^{(t)}\|_1$ to $\|\mathbf{b}^*\|_1$.

Theorem D.13. *For all $T \in \mathbb{N}^*$, we have*

$$\min_{1\leq t\leq T}\left(\|\mathbf{A}^{(t)}\|_* - \|\mathbf{A}^*\|_*\right) \leq \frac{\|\mathbf{A}^{(1)} - \mathbf{A}^*\|_F^2 + \left(\max_{1\leq t\leq T}\|\nabla_{\mathbf{A}} f(\mathbf{A}^{(t)})\|_F^2\right)\sum_{t=1}^T \alpha_t^2}{2\beta_* \sum_{t=1}^T \alpha_t} + \frac{\|\boldsymbol{\xi}\|_2^2 + \beta_2\|\mathbf{A}^*\|_F^2}{2\beta_*} \tag{126}$$

Proof. The proof is similar to C.13 □

So, when $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$ (e.g. $\alpha_t = a/(b+t)$, $a > 0$ and $b \geq 0$), $\|\mathbf{A}^{(t)}\|_1 \to \|\mathbf{A}^*\|_1 \to 0$ as $T \to \infty$, for $\beta_2 = 0$ in the noiseless setting.

D.6.3 Additionnal experiments

We optimize the noiseless matrix completion problem using the subgradient descent method with $(n_1, n_2, r, N, \zeta, \beta_2) = (10, 10, 2, 70, 10^{-6}, 0)$ for different values of α and β_*. As expected, larger α and/or β_* lead to fast convergence and do so at a suboptimal value of the test error (Figure 42).

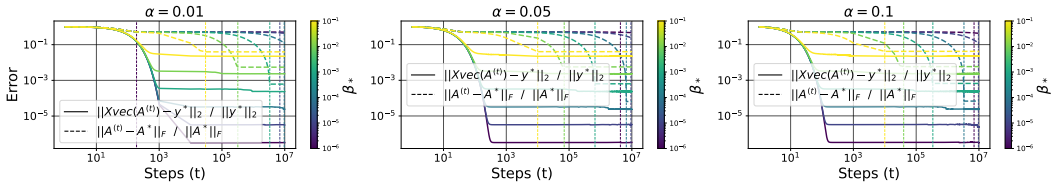


Figure 42: Training error $\|\mathbf{X}\,\mathrm{vec}\,\mathbf{A}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_\mathrm{F}/\|\mathbf{A}^*\|_\mathrm{F}$ as a function of the learning rate α and the ℓ_*-regularization coefficient β_*. Here $(n_1, n_2, r, N) = (10, 10, 2, 70)$

D.7 Projected subgradient

To ensure memorization, we can use the projected subgradient for problem (P_5) of minimizing $\|\mathbf{A}\|_*$ subject to the constraint $\mathcal{F}_{\mathrm{vec}\,\mathbf{A}}(\mathbf{X}) = \mathbf{X}\,\mathrm{vec}\,\mathbf{A} = \mathbf{y}^*$, where at each step the update (using now just $\beta_* h(\mathbf{A})$ as gradient) is projected onto the constraint set. In our case, the update write $\mathbf{A}^{(t+1)} = \Pi\left(\mathbf{A}^{(t)} - \alpha_t \beta_* h(\mathbf{A}^{(t)})\right)$ with Π the projection on the set $\{\mathbf{A}, \mathbf{X}\,\mathrm{vec}\,\mathbf{A} = \mathbf{y}^*\}$. Figure 43 shows the results for a matrix sensing problem.

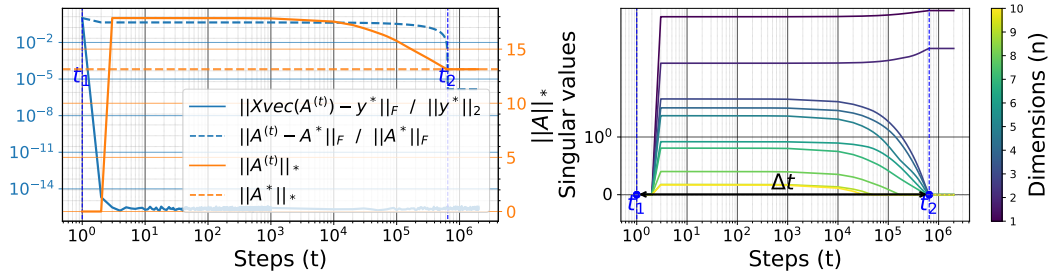


Figure 43: Relative errors, norm $\|\mathbf{A}^{(t)}\|_*$, and evolution of singular value for the **projected subgradient method**. $G_{\beta_2}(\mathbf{A}^{(t)})$ dominates $\beta_* h(\mathbf{A}^{(t)})$ until memorization. From memorization $\beta_* h(\mathbf{A}^{(t)})$ dominates and make $\|\mathbf{A}^{(t)}\|_1$ converge to $\|\mathbf{A}^*\|_1$ at t_2, and so $\mathbf{A}^{(t_2)} = \mathbf{A}^*$. Here $(n_1, n_2, r, N) = (10, 10, 2, 70)$ and $(\zeta, \alpha, \beta_*, \beta_2) = (10^{-6}, 10^{-1}, 10^{-4}, 0)$.

D.8 Proximal Gradient Descent and Iterative soft-thresholding algorithm

Similar to what we derive in section C.8, we have $\mathbf{A} - \alpha F(\mathbf{A}) = \Pi_\alpha\left(\mathbf{A} - \alpha G_{\beta_2}(\mathbf{A})\right)$ where Π_α is the proximal mapping for $\mathbf{B} \to \beta_*\|\mathbf{B}\|_*$, $\Pi_\alpha(\mathbf{A}) = \arg\min_{\mathbf{B}} \frac{1}{2\alpha}\|\mathbf{B} - \mathbf{A}\|_{\mathrm{F}}^2 + \beta_*\|\mathbf{B}\|_* = S_{\alpha\beta_*}(\mathbf{A})$ with $S_\gamma(\mathbf{A}) = \mathbf{U}\max(\Sigma - \gamma, 0)\mathbf{V}^\top$ the soft-thresholding operator for $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ under SVD, where $\max(\Sigma - \gamma, 0)_{ij} = \delta_{ij}\max(\Sigma_{ij} - \gamma, 0)$. The final form of the update is then

$$\mathbf{A}^{(t+1)} = S_{\alpha_t\beta_*}\left(\mathbf{A}^{(t)} - \alpha_t G_{\beta_2}(\mathbf{A}^{(t)})\right) \quad \forall t > 1 \tag{127}$$

Figure 44 shows the results for a matrix sensing problem.

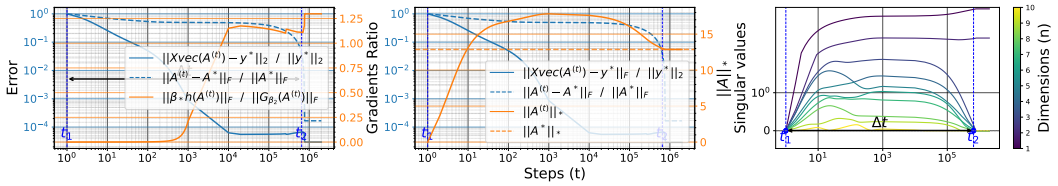


Figure 44: Gradient Ratio, relative errors, norm $\|\mathbf{A}^{(t)}\|_*$, and evolution of singular value for the **Proximal Gradient Descent**. $G_{\beta_2}(\mathbf{A}^{(t)})$ dominates $\beta_* h(\mathbf{A}^{(t)})$ until memorization. From memorization $\beta_* h(\mathbf{A}^{(t)})$ dominates and make $\|\mathbf{A}^{(t)}\|_1$ converge to $\|\mathbf{A}^*\|_1$ at t_2, and so $\mathbf{A}^{(t_2)} = \mathbf{A}^*$. Here $(n_1, n_2, r, N) = (10, 10, 2, 70)$ and $(\zeta, \alpha, \beta_*, \beta_2) = (10^{-6}, 10^{-1}, 10^{-4}, 0)$.

D.9 Grokking without understanding

Like in section C.9, there is no grokking for $N < n$ when $\beta_* \neq 0$, no matter the value of β_2 and the initialization scale $\zeta \geq 0$, $\mathbf{A}^{(1)} \overset{iid}{\sim} \zeta\mathcal{N}(0, 1/n)$. With a small initialization, β_1 is sufficient for generalization to happen, provided N is large enough and β_2 is not very large. If the scale at initialization is large, β_2 is necessary to generalize, but it is not sufficient: because of the large initialization, a transition is observed in the generalization error during training, corresponding to a transition in the ℓ_2 norm of the model parameters, but not the recovery error.

D.10 Impact of Coherence on Grokking: Amplifying Grokking through Data Selection

Above, we introduce the parameter $\tau \in [0, 1]$ that control the incoherence between the measures $\{\mathbf{X}_i\}_{i\in[N]}$ and the sparse basis (dictionary) $\{\Phi_{:,j}\}_{j\in[n]}$, with $\Phi = \mathbf{V}^* \otimes \mathbf{U}^* \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$ and $\mathbf{X} = \mathbf{X}^{(2)} \bullet \mathbf{X}^{(1)} \in \mathbb{R}^{N \times n_1 n_2}$. Unlike compressed sensing (Section C.10), where large values of τ are detrimental to generalization, here, as $\tau \to 1$, performance improves, and the number of examples required to generalize decreases exponentially, as does the time it takes the models to do so (Figures 45 and Figures 46). Note that here, for matrix completion, for a fixed τ, we select the first τN examples with the highest values of $\mu_i(\mathbf{A}^*) + \nu_j(\mathbf{A}^*)$, and select the remaining $(1 - \tau)N$ examples at random, uniformly.

D.11 Deep Matrix Factorization: The Effect of Overparametrization

Let now use the parameterization $\mathbf{A} = \prod_{k=1}^{L} \mathcal{A}_k$, with $\mathcal{A}_1 \in \mathbb{R}^{n_1 \times d}$, $\mathcal{A}_L \in \mathbb{R}^{d \times n_2}$, and $\mathcal{A}_i \in \mathbb{R}^{d \times d}$ for all $i \in (1, L)$. This corresponds to a linear network with L layers, where each hidden layer has the parameter $\mathcal{A}_k$—with this, increasing L leads to overparameterization without altering the expressiveness of the function class $\mathbf{A} \to \mathcal{F}_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}^\top \operatorname{vec}\mathbf{A}$, since the model remains linear with respect to the input $\mathbf{x}$. Like in compressed sensing, there is no need for ℓ_* ($\beta_* = 0$) to generalize when $L \geq 2$ (and the initialization scale is small), unlike the shallow case ($L = 1$). This is an observation already made and proven in previous art. (Gunasekar et al., 2017; Arora et al., 2019; Gidel et al., 2019; Gissin et al., 2019; Razin & Cohen, 2020; Li et al., 2020). Gunasekar et al. (2017); Arora et al. (2019) show increasing L implicitly bias $\mathbf{A}$ toward a low-rank solution, which oftentimes leads to more accurate recovery for sufficiently large N. In fact, with depth, the update for the

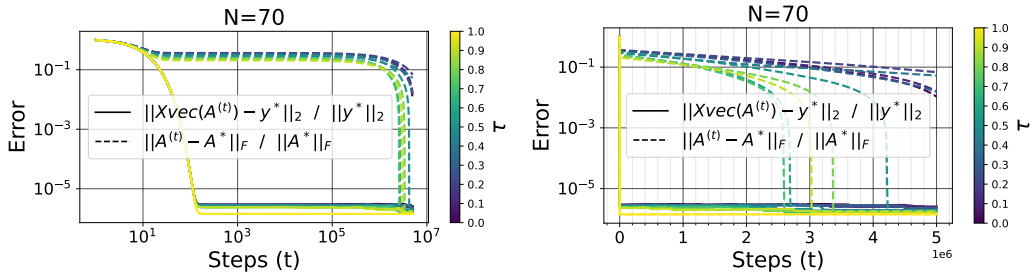


Figure 45: Training and error $\|\mathbf{X}\,\mathrm{vec}\,\mathbf{A}^{(t)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_2/\|\mathbf{A}^*\|_\mathrm{F}$ as a function of the number of sample N and the coherence parameter $\tau \in [0, 1]$. Here $(n_1, n_2, r, \alpha, \beta_1, \beta_2, \zeta) = (10, 10, 2, 10^{-1}, 10^{-5}, 0, 10^{-6})$.

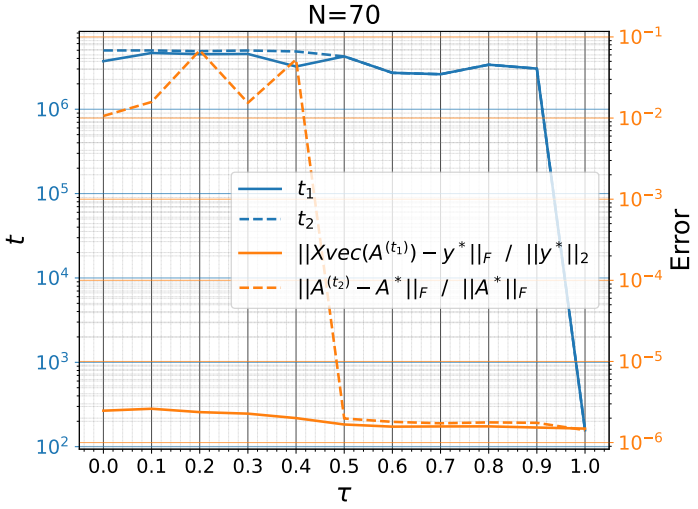


Figure 46: Training and error $\|\mathbf{X}\,\mathrm{vec}\,\mathbf{A}^{(t_1)} - \mathbf{y}^*\|_2/\|\mathbf{y}^*\|_2$ and recovery error $\|\mathbf{A}^{(t_2)} - \mathbf{A}^*\|_2/\|\mathbf{A}^*\|_\mathrm{F}$ (along with t_1 and t_2, the memorization and the generalization step) as a function of the number of sample N and the coherence parameter $\tau \in [0, 1]$. Here $(n_1, n_2, r, \alpha, \beta_1, \beta_2, \zeta) = (10, 10, 2, 10^{-1}, 10^{-5}, 0, 10^{-6})$.

whole iterate is similar to the shallow case but with a preconditioner in front of the gradient (like in section C.11). This preconditioner makes it possible to recover the low-rank matrix signal without any regularization and with fewer samples than in the shallow case (Arora et al., 2018; 2019). It is also shown specifically for this problem that initializing the model very far from the origin and using a small (but non-zero) weight decay leads to grokking (Lyu et al., 2023), i.e., the model first memorizes the observed entries, then after a long training period, converges to the sought matrices provided the number of such observe entries is large enough.

We have $\mathbf{y}(\mathbf{A}) = \mathcal{F}_{\operatorname{vec}\mathbf{A}}(\mathbf{X}) = \mathbf{X}\operatorname{vec}\mathbf{A}$ and $\mathbf{y}^* = \mathcal{F}_{\operatorname{vec}\mathbf{A}^*}(\mathbf{X}) + \boldsymbol{\xi} = \mathbf{X}\operatorname{vec}\mathbf{A}^* + \boldsymbol{\xi}$, and want to minimize $f(\mathbf{A}) = g_{\beta_2}(\mathbf{A}) + \beta_* \sum_k \|\mathcal{A}_k\|_*$ using gradient descent, where

$$g_{\beta_2}(\mathbf{A}) := \frac{1}{2}\|\mathbf{y}(\mathbf{A}) - \mathbf{y}^*\|_2^2 + \frac{\beta_2}{2}\sum_k \|\mathcal{A}_k\|_{\mathrm{F}}^2 \tag{128}$$

Let $\operatorname{vec} G(\mathbf{A}) := \frac{\partial g_{\beta_2}(\mathbf{A})}{\partial \operatorname{vec}\mathbf{A}} = \mathbf{X}^\top(\mathbf{y}(\mathbf{A}) - \mathbf{y}^*) = \mathbf{X}^\top\mathbf{X}(\operatorname{vec}\mathbf{A} - \operatorname{vec}\mathbf{A}^*) - \mathbf{X}^\top\boldsymbol{\xi}$. The gradient for each $\mathcal{A}_k$ is

$$\begin{aligned} G_{\beta_2}(\mathcal{A}_k) &:= \frac{\partial g_{\beta_2}(\mathbf{A})}{\partial \mathcal{A}_k} \\ &= \begin{cases} G(\mathbf{A})(\mathcal{A}_2\cdots\mathcal{A}_L)^\top + \beta_2\mathcal{A}_k & \text{for } k = 1 \\ (\mathcal{A}_1\cdots\mathcal{A}_{k-1})^\top G(\mathbf{A})(\mathcal{A}_{k+1}\cdots\mathcal{A}_L)^\top + \beta_2\mathcal{A}_k & \text{for } k \in (1, L) \\ (\mathcal{A}_1\cdots\mathcal{A}_{L-1})^\top G(\mathbf{A}) + \beta_2\mathcal{A}_k & \text{for } k = L \end{cases} \quad \text{(Lemma D.14)} \end{aligned} \tag{129}$$

And the update rule for each $\mathcal{A}_k$ is

$$\begin{aligned} \mathcal{A}_k^{(t+1)} &= \mathcal{A}_k^{(t)} - \alpha G_{\beta_2}(\mathcal{A}_k^{(t)}) - \alpha\beta_* h(\mathcal{A}_k^{(t)}) \\ &= (1 - \alpha\beta_2)\mathcal{A}_k^{(t)} - \alpha\left(\prod_{i<k}\mathcal{A}_i^{(t)}\right)^\top G(\mathbf{A}^{(t)})\left(\prod_{i>k}\mathcal{A}_i^{(t)}\right)^\top - \alpha\beta_* h(\mathcal{A}_k^{(t)}) \end{aligned} \tag{130}$$

where $h(\mathcal{A}_k) \in \partial\|\mathcal{A}_k\|_*$. We start the optimization at $\mathcal{A}_k^{(1)} \overset{iid}{\sim} \zeta\mathcal{N}(0, 1/n)$ with $\zeta \geq 0$ the initialization scale.

Lemma D.14. *Let $f(\mathcal{A}_1, \cdots, \mathcal{A}_L) = g(\mathbf{A}) \in \mathbb{R}$ with $\mathbf{A} = \prod_{k=1}^L \mathcal{A}_k \in \mathbb{R}^{d_0 \times d_L}$, where $\mathcal{A}_k \in \mathbb{R}^{d_{k-1}\times d_k}$ for all $k \in [L]$. We have*

$$\frac{\partial f(\mathbf{A})}{\partial \mathcal{A}_k} = \left(\prod_{i<k}\mathcal{A}_i\right)^\top \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}\left(\prod_{i>k}\mathcal{A}_i\right)^\top = \begin{cases} \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}(\mathcal{A}_2\cdots\mathcal{A}_L)^\top & \textit{for } k = 1 \\ (\mathcal{A}_1\cdots\mathcal{A}_{k-1})^\top \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}(\mathcal{A}_{k+1}\cdots\mathcal{A}_L)^\top & \textit{for } k \in (1, L) \\ (\mathcal{A}_1\cdots\mathcal{A}_{L-1})^\top \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}} & \textit{for } k = L \end{cases} \tag{131}$$

Proof. We have

$$\mathbb{R}^{d_0 d_L} \ni \operatorname{vec}\mathbf{A} = \begin{cases} \left((\mathcal{A}_2\cdots\mathcal{A}_L)^\top \otimes \mathbb{I}_{d_0}\right)\operatorname{vec}\mathcal{A}_1 & \text{for } k = 1 \\ \left((\mathcal{A}_{k+1}\cdots\mathcal{A}_L)^\top \otimes (\mathcal{A}_1\cdots\mathcal{A}_{k-1})\right)\operatorname{vec}\mathcal{A}_k & \text{for } k \in (1, L) \\ (\mathbb{I}_{d_L} \otimes (\mathcal{A}_1\cdots\mathcal{A}_{L-1}))\operatorname{vec}\mathcal{A}_L & \text{for } k = L \end{cases} \tag{132}$$

So

$$\mathbb{R}^{d_0 d_L \times d_{k-1} d_k} \ni \frac{\partial \operatorname{vec}\mathbf{A}}{\partial \operatorname{vec}\mathcal{A}_k} = \begin{cases} (\mathcal{A}_2\cdots\mathcal{A}_L)^\top \otimes \mathbb{I}_{d_0} \in \mathbb{R}^{d_L d_0 \times d_1 d_0} & \text{for } k = 1 \\ (\mathcal{A}_{k+1}\cdots\mathcal{A}_L)^\top \otimes (\mathcal{A}_1\cdots\mathcal{A}_{k-1}) \in \mathbb{R}^{d_L d_0 \times d_k d_{k-1}} & \text{for } k \in (1, L) \\ \mathbb{I}_{d_L} \otimes (\mathcal{A}_1\cdots\mathcal{A}_{L-1})\operatorname{vec}\mathcal{A}_L \in \mathbb{R}^{d_L d_0 \times d_L d_{L-1}} & \text{for } k = L \end{cases} \tag{133}$$

For $\mathbf{Q} \in \mathbb{R}^{d_0 \times d_L}$,

$$\begin{aligned} \left(\frac{\partial \operatorname{vec}\mathbf{A}}{\partial \operatorname{vec}\mathcal{A}_k}\right)^\top \operatorname{vec}\mathbf{Q} &= \begin{cases} ((\mathcal{A}_2\cdots\mathcal{A}_L) \otimes \mathbb{I}_{d_0})\operatorname{vec}\mathbf{Q} & \text{for } k = 1 \\ \left((\mathcal{A}_{k+1}\cdots\mathcal{A}_L) \otimes (\mathcal{A}_1\cdots\mathcal{A}_{k-1})^\top\right)\operatorname{vec}\mathbf{Q} & \text{for } k \in (1, L) \\ (\mathbb{I}_{d_L} \otimes (\mathcal{A}_1\cdots\mathcal{A}_{L-1}))^\top \operatorname{vec}\mathbf{Q} & \text{for } k = L \end{cases} \\ &= \begin{cases} \operatorname{vec}\left(\mathbf{Q}(\mathcal{A}_2\cdots\mathcal{A}_L)^\top\right) & \text{for } k = 1 \\ \operatorname{vec}\left((\mathcal{A}_1\cdots\mathcal{A}_{k-1})^\top\mathbf{Q}(\mathcal{A}_{k+1}\cdots\mathcal{A}_L)^\top\right) & \text{for } k \in (1, L) \\ \operatorname{vec}\left((\mathcal{A}_1\cdots\mathcal{A}_{L-1})^\top\mathbf{Q}\right) & \text{for } k = L \end{cases} \end{aligned} \tag{134}$$

So

$$\frac{\partial g(\mathbf{A})}{\partial \operatorname{vec} \mathcal{A}_k} = \left(\frac{\partial \operatorname{vec} \mathbf{A}}{\partial \operatorname{vec} \mathcal{A}_k}\right)^\top \frac{\partial g(\mathbf{A})}{\partial \operatorname{vec} \mathbf{A}} = \left(\frac{\partial \operatorname{vec} \mathbf{A}}{\partial \operatorname{vec} \mathcal{A}_k}\right)^\top \operatorname{vec} \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}$$
$$= \begin{cases} \operatorname{vec}\left(\frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}(\mathcal{A}_2 \cdots \mathcal{A}_L)^\top\right) & \text{for } k = 1 \\ \operatorname{vec}\left((\mathcal{A}_1 \cdots \mathcal{A}_{k-1})^\top \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}(\mathcal{A}_{k+1} \cdots \mathcal{A}_L)^\top\right) & \text{for } k \in (1, L) \\ \operatorname{vec}\left((\mathcal{A}_1 \cdots \mathcal{A}_{L-1})^\top \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}\right) & \text{for } k = L \end{cases} \tag{135}$$

□

E Beyong Sparse Recovery and Matrix Factorization

We will optimize functions of the form $f(\theta) = \hat{\mathcal{E}}(\theta) + \beta\Omega(\theta)$, where $\hat{\mathcal{E}}$ is the square loss or cross-entropy loss function of the considered model on the training data, θ the set of model parameters, and Ω a regularizer applied to θ. It can be the standard ℓ_p norm or quasi-norm of θ, the sum of the nuclear norms of each matrix in θ (in this case, we call it ℓ_*), etc. By normal initialization for a parameter $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, we mean $\mathbf{A}^{(0)} \overset{iid}{\sim} \mathcal{N}(0, 1/n_1)$.

E.1 Non linear Teacher-Student

We consider a teacher $\mathbf{y}^*(\mathbf{x}) = \mathbf{B}^* g(\mathbf{A}^* \mathbf{x})$ from $\mathbb{R}^d$ to $\mathbb{R}^c$ with r hidden neurons ($\mathbf{A}^* \in \mathbb{R}^{r \times d}$ and $\mathbf{B}^* \in \mathbb{R}^{c \times r}$); where $g(x) = \max(x, 0)$ and $\mathbf{x}, \mathbf{A}^*, r\mathbf{B}^* \overset{iid}{\sim} \mathcal{N}(0, 1)$. We i.i.d sample N inputs output pair $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, \mathbf{y}^*(\mathbf{x}_i))\}_{i=1}^N$ and optimize the parameters $\theta = (\mathbf{A}, \mathbf{B})$ of a student $\mathbf{y}_\theta(\mathbf{x}) = \mathbf{B} g(\mathbf{A}\mathbf{x})$ on them, starting from normal initialization, with the loss function $\hat{\mathcal{E}}(\theta) = \frac{1}{N}\sum_{i=1}^N \|\mathbf{y}_\theta(\mathbf{x}_i) - \mathbf{y}^*(\mathbf{x}_i)\|_2^2$ and different regularizer $\Omega_p(\theta)$ for $p \in \{1, 2, *\}$.

For any $p \in \{1, 2, *\}$, the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization. See Figures 47, 48 and 49 for an experiment with $(d, r, c, N) = (100, 500, 2, 10^2)$.

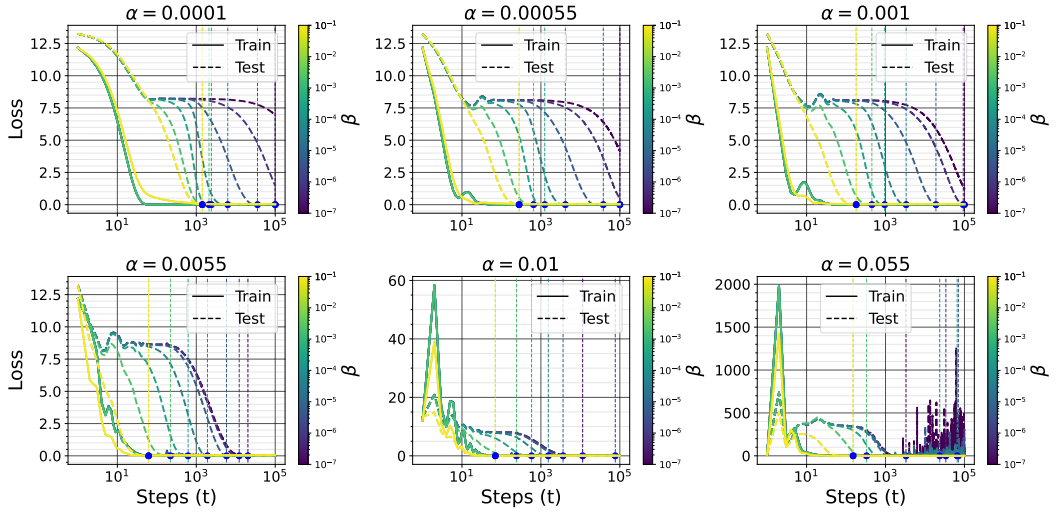


Figure 47: Training and test error two layers `ReLU` teacher-student with ℓ_1 regularization, for different values of the learning rate α and the ℓ_1 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

E.2 Domain Specific Regularization

Physics-Informed Neural Networks (PINNs) (Raissi et al., 2019) leverage prior knowledge from differential equations by incorporating their residuals into the loss function, ensuring that solutions

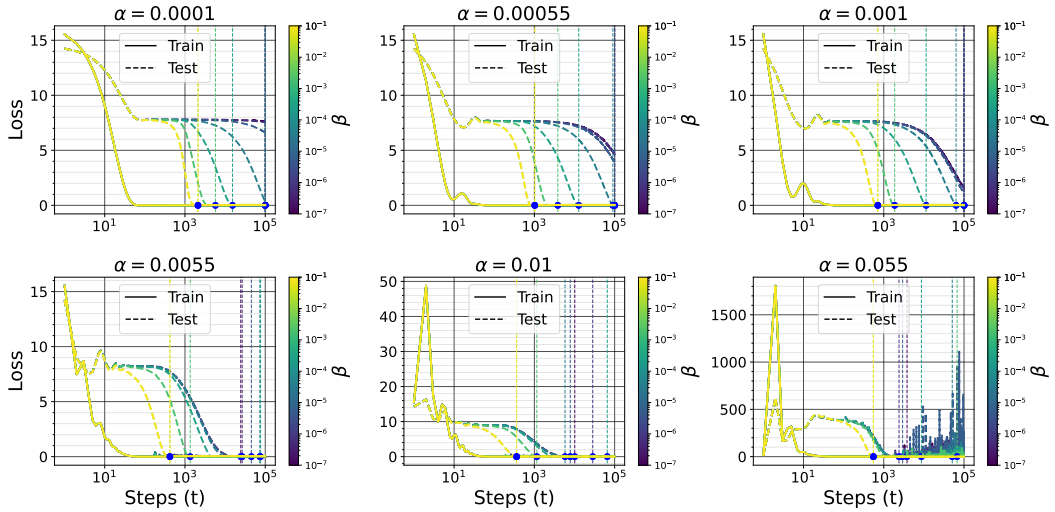


Figure 48: Training and test error two layers `ReLU` teacher-student with ℓ_2 regularization, for different values of the learning rate α and the ℓ_2 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

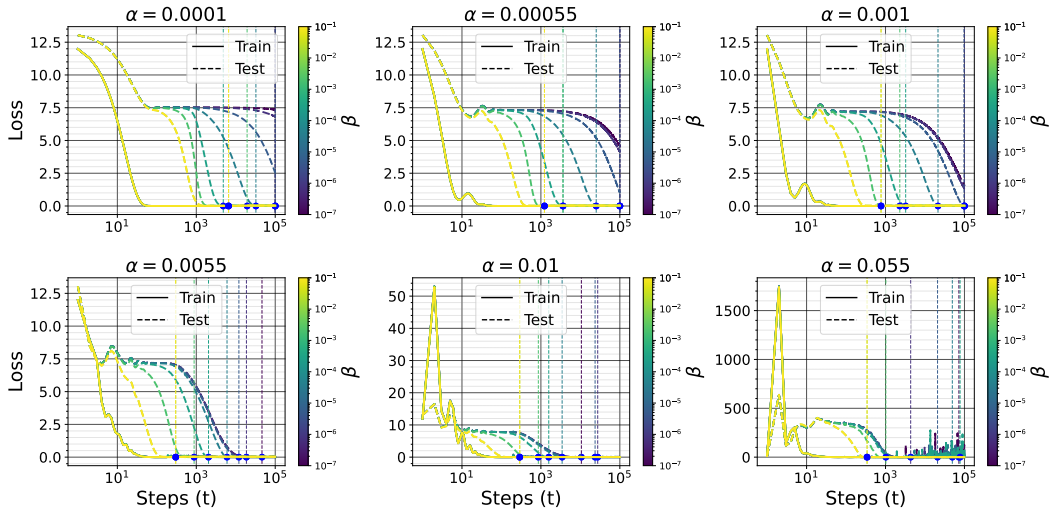


Figure 49: Training and test error two layers `ReLU` teacher-student with ℓ_* regularization, for different values of the learning rate α and the ℓ_* coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

remain consistent with physical laws. Sobolev training (Czarnecki et al., 2017) generalizes this idea by incorporating not only input-output pairs but also derivatives of the target function. More precisely, given input-output pairs $\{(\mathbf{x}_i, \mathbf{y}^*(\mathbf{x}_i)\}_{i\in[N]}$ along with known derivatives $\left\{ \frac{\partial^k \mathbf{y}^*(\mathbf{x})}{\partial \mathbf{x}^k}\Big|_{\mathbf{x}=\mathbf{x}_i} \right\}_{i\in[N]}$ for $k \in [K]$, the goal is to train a neural network $\mathbf{y}_\theta(\mathbf{x})$ that approximates both the output and its derivatives. The loss function extends the standard mean squared error (MSE) to include Sobolev penalties:

$$f(\theta) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}_\theta(\mathbf{x}_i) - \mathbf{y}^*(\mathbf{x}_i)\|^2}_{\text{data loss}} + \underbrace{\frac{\beta}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} \left\| \frac{\partial^k \mathbf{y}_\theta}{\partial \mathbf{x}^k}(\mathbf{x}_i) - \frac{\partial^k \mathbf{y}^*}{\partial \mathbf{x}^k}(\mathbf{x}_i) \right\|_{\mathrm{F}}^2}_{\text{Sobolev penalty}} \tag{136}$$

The hyperparameter β controls the contribution of the derivative alignment term. This penalty ensures that the model not only fits the data but also respects known smoothness constraints or differential structure, which is crucial in physics-based applications (Lu et al., 2021). We consider the two layers feed forward teacher $\mathbf{y}^*(\mathbf{x}) = \mathbf{B}^* g(\mathbf{A}^*\mathbf{x})$ of Section E.1, and optimize the parameters $\theta = (\mathbf{A}, \mathbf{B})$ of a student $\mathbf{y}_\theta(\mathbf{x}) = \mathbf{B} g(\mathbf{A}\mathbf{x})$ using the sobolev objectify for $K = 1$, $\frac{\partial \mathbf{y}^*(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{B}^* \operatorname{diag}\left(g'(\mathbf{A}^*\mathbf{x})\right) \mathbf{A}^*$. For any $p \in \{1, 2, *\}$, the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization. See Figure 50 for an experiment with $(d, r, c, N) = (100, 500, 2, 10^2)$.

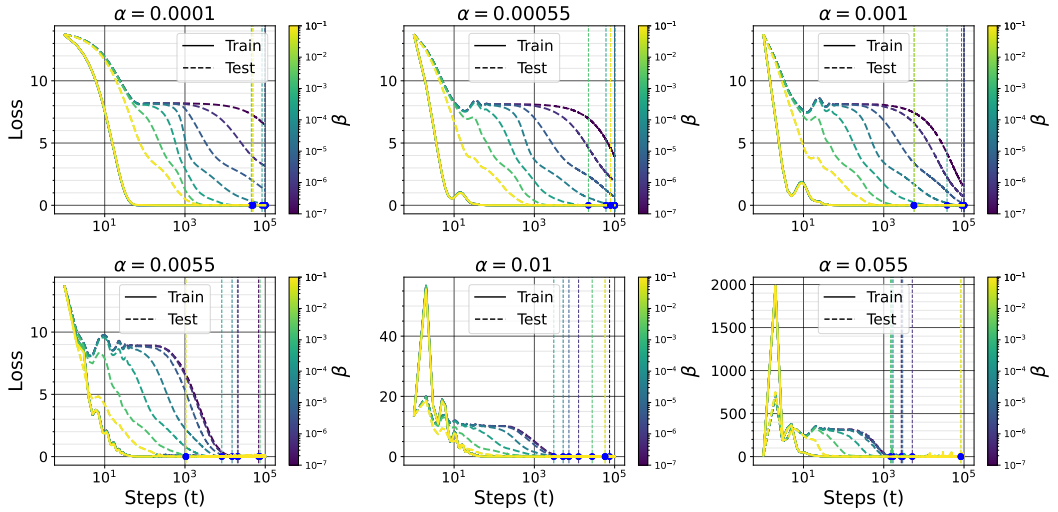


Figure 50: Training and test error two layers `ReLU` teacher-student with Sobolev training, for different values of the learning rate α and the ℓ_1 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

E.3 Algorithmic dataset

Consider a binary mathematical operator $\circ$ on $\mathcal{S} = \mathbb{Z}/p\mathbb{Z}$ for some prime integer p. We want to predict $y^*(x) = x_1 \circ x_2$ given $x = (x_1, x_2) \in \mathcal{S}^2$. The dataset $\mathcal{D} = \{(x, y^*(x)) | x \in \mathcal{S}^2\}$ is randomly partitioned into two disjoint and non-empty sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$, the training and the validation dataset respectively[8]. Let $r_{\text{train}} = |\mathcal{D}_{\text{train}}|/|\mathcal{D}|$ be the training data fraction.

For MLP, the logits for $x = (x_1, x_2)$ are given by $\mathbf{y}(x_1, x_2) = \mathbf{b}^{(2)} + \mathbf{W}^{(2)} g\left(\mathbf{b}^{(1)} + \mathbf{W}^{(1)}\left(\mathbb{E}_{\langle x_1\rangle} \circ \mathbb{E}_{\langle x_2\rangle}\right)\right)$, where $\langle x_1\rangle$ stands for the token corresponding to x_1, and $\mathbb{E}$ is the embedding matrix for all the symbols in $\mathcal{S}$, g the activation function. $\theta = \left(\mathbb{E}, \mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}\right) \in \mathbb{R}^{p\times d_1} \times \mathbb{R}^{d_2\times d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{p\times d_2} \times \mathbb{R}^p$ are the learnable

[8]It can be necessary in some contexts to consider the symmetric nature of $\circ$, so that $|\mathcal{D}| = p(p+1)/2$ if $\circ$ is symmetric (and we consider $x_1 \circ x_2$ and $x_2 \circ x_1$ as the same operation), and p^2 otherwise.

parameter, with d_1 the embedding dimension. For the LSTM, we treat a problem as a sequence classification problem, i.e., the sequence of tokens $\langle x_1 \rangle \langle \circ \rangle \langle x_2 \rangle \langle = \rangle$ is given to the model and its task is to predict $y^*(x_1, x_2)$.

We consider addition modulo $p = 97$ with $r_{\text{train}} = 40\%$. For MLP and LSTM, ℓ_1 and ℓ_* have the same effect on grokking as ℓ_2. For any $p \in \{1, 2, *\}$, the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization. See Figures 51, 52, 53, 54, 55 and 56.

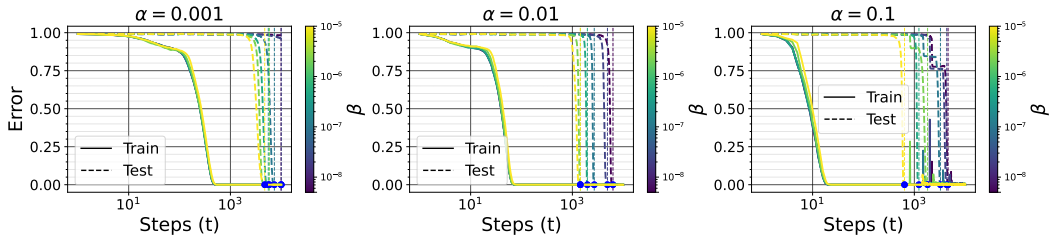


Figure 51: Training and test error ($1 - \texttt{Accuray}$) of a Multi-layer perceptron trained on the algorithmic dataset with ℓ_1 regularization for different values of the learning rate α and the ℓ_1 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

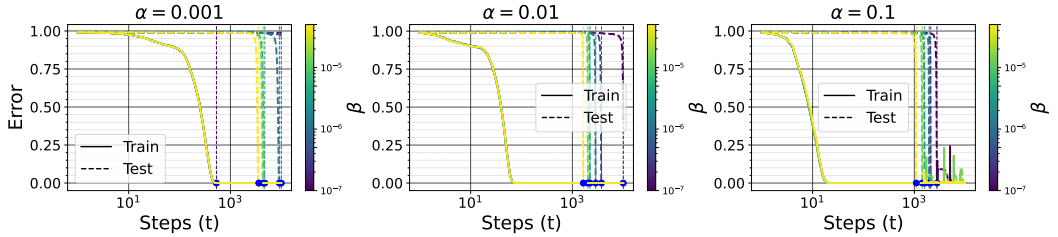


Figure 52: Training and test error ($1 - \texttt{Accuray}$) of a Multi-layer perceptron trained on the algorithmic dataset with ℓ_2 regularization for different values of the learning rate α and the ℓ_2 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

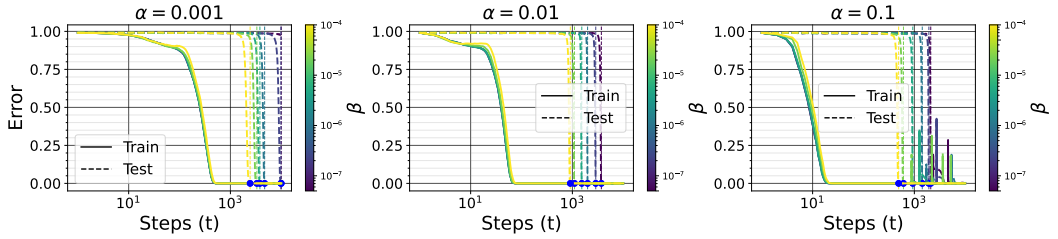


Figure 53: Training and test error ($1 - \texttt{Accuray}$) of a Multi-layer perceptron trained on the algorithmic dataset with ℓ_* regularization for different values of the learning rate α and the ℓ_* coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

E.4 Image Classification

We optimize the parameters $\theta = (\mathbf{A}, \mathbf{B})$ of a model $\mathbf{y}_\theta(\mathbf{x}) = \mathbf{B}g(\mathbf{A}\mathbf{x})$ on $N = 1000$ samples of the MNIST dataset. Figure 57 show the results for ℓ_1 : the result for ℓ_2 and ℓ_* are similar.

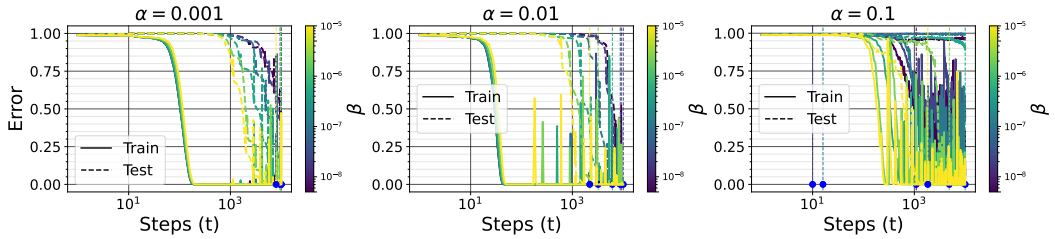


Figure 54: Training and test error ($1 - \texttt{Accuray}$) of a Long Short Term Memory trained on the algorithmic dataset with ℓ_1 regularization for different values of the learning rate α and the ℓ_1 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

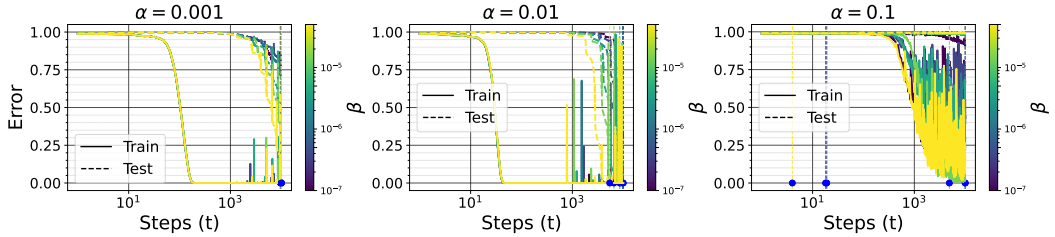


Figure 55: Training and test error ($1 - \texttt{Accuray}$) of a Long Short Term Memory trained on the algorithmic dataset with ℓ_2 regularization for different values of the learning rate α and the ℓ_2 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

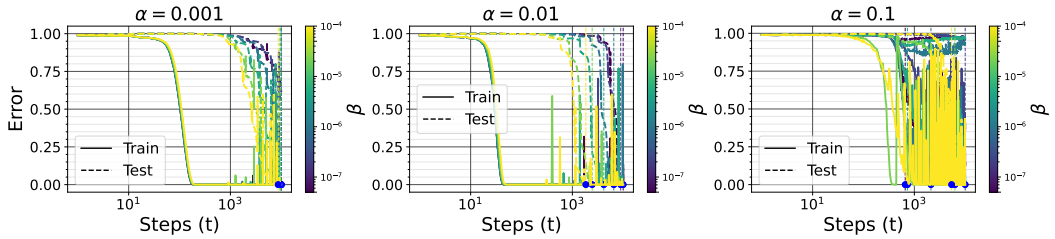


Figure 56: Training and test error ($1 - \texttt{Accuray}$) of a Long Short Term Memory trained on the algorithmic dataset with ℓ_* regularization for different values of the learning rate α and the ℓ_* coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.

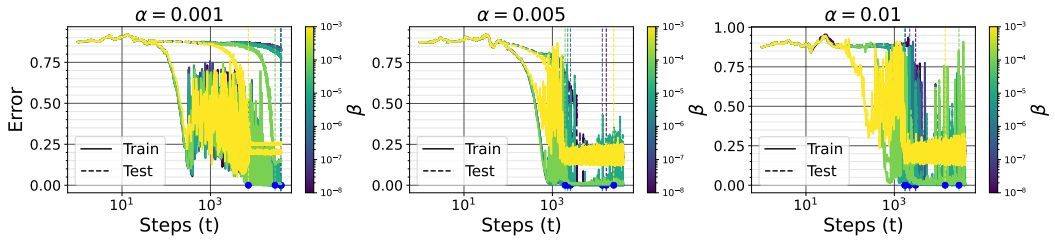


Figure 57: Training and test error ($1 - \texttt{Accuray}$) of a Multi-layer perceptron trained on MNIST with ℓ_1 regularization for different values of the learning rate α and the ℓ_1 coefficient β. We can see that the smaller is $\alpha\beta$, the longer is the delay between memorization and generalization.