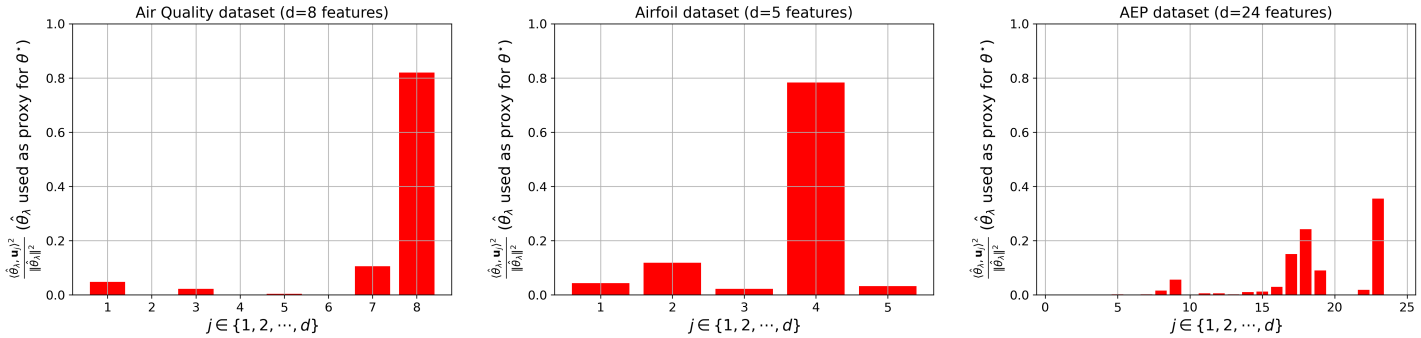
Understanding the Gains from Repeated Self-Distillation
Rebuttal material (NeurIPS 2024)



(a) Air Quality dataset ($\lambda = 0.167$) (b) Airfoil dataset ($\lambda = 0.023$) (c) AEP dataset ($\lambda = 0.237$)

Figure 1: We observe that for the AEP dataset (Table 1 of the manuscript), self-distillation provided no gain over ridge in terms of the test MSE. We provide an explanation for the flat performance. This figure is examining the alignment of $\theta^\star$ to the bases directions $\{\mathbf{u}_j\}_{j=1}^d$ for the three datasets used in the experiments (Section 5.3, Table 1). Because $\theta^\star$ is unknown for real-world tasks, we use the ridge solution $\hat{\theta}_\lambda$ (with a small λ) as a *proxy* for $\theta^\star$. The sum of all bars in a single plot is *one*, since $\{\mathbf{u}_j\}_{j=1}^d$ are unit-norm vectors that form an orthogonal basis of $\mathbb{R}^d$. We infer two things. **Firstly**, for multi-step SD to outperform ridge (as is the case for the Air Quality and Airfoil datasets), $\theta^\star$ can be well-aligned with *any* of the $\mathbf{u}_j, j \in [d]$; not necessarily $\mathbf{u}_1$. This experimentally verifies the remark from lines 203-205 in the manuscript. **Secondly**, this gives insight into why multi-step SD could not outperform ridge on the AEP task (Table 1 in the paper). Unlike the other two datasets, the $\theta^\star$ for the AEP dataset is not strongly aligned with any of the $\mathbf{u}_j, j \in [d]$. The top component in AEP only explains $\sim 35\%$ of the total $\theta^\star$ norm, whereas that number is close to $\sim 80\%$ for the Air Quality and Airfoil datasets. **Methodology of choosing** λ: We considered using the OLS solution $\hat{\theta}_{\text{OLS}} := (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}$ as the proxy for $\theta^\star$, but $(\mathbf{X}\mathbf{X}^\top)^{-1}$ was numerically unstable for these datasets, so we instead used the ridge solution $\hat{\theta}_\lambda$ with a small λ. We calculated this λ methodically for all datasets as a constant fraction of the sum of squared singular values. Explicitly, we (i) computed the SVD of the design matrix $\mathbf{X}$, and (ii) set $\lambda := C \cdot \sum_{j=1}^d s_j^2$ using the obtained singular values. The value $C := 10^{-5}$ was chosen arbitrarily (and the above trend is stable across other reasonably small values of C).

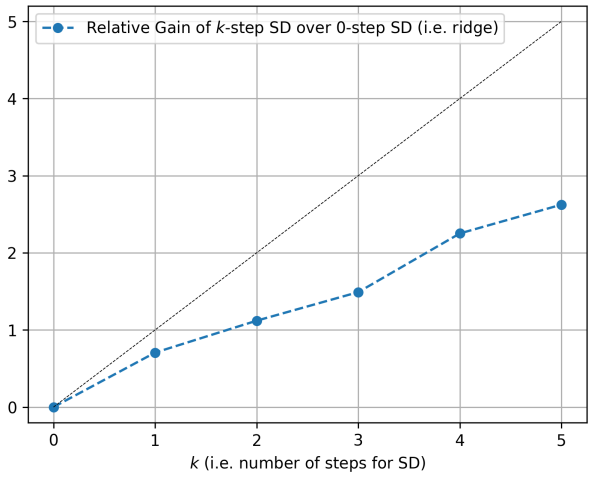


Figure 2: We empirically demonstrate that k-step SD for $k > 2$ provides a non-trivial improvement over 2-step SD. On a synthetic problem with dimension $d = 100$, $r = \text{rank}(\mathbf{X}) = 10$, noise variance $\gamma = 0.1$, and $\theta^\star = 1/\sqrt{2}(\mathbf{u}_1 + \mathbf{u}_2)$; we set the singular values of $\mathbf{X}$ with a power law from $s_1 = 1$ to $s_r = s_{10} = 0.5$ and run k-step SD for $k \in \{0, 1, 2, 3, 4, 5\}$. We plot the relative improvement of *optimal* k-step SD over *optimal* ridge (i.e. 0-step SD). Specifically, we plot the ratio $(A/B(k) - 1)$ where $A := \min_{\lambda > 0} \text{ExcessRisk}\big(\hat{\theta}(\lambda)\big)$ and $B(k) := \min_{\lambda > 0, \xi^{(k)} \in \mathbb{R}^k} \text{ExcessRisk}\big(\hat{\theta}(\lambda, \xi^{(k)})\big)$. We observe a non-trivial increase after $k = 2$, showing that multi-step SD beyond 2 steps can be valuable.