

A Proofs

The combination of the following 2 Lemmas is a generalization of the geometric inequality proved by Liang et al. [LRS15]. In many respects the scheme of the proof is similar.

Lemma A.1. (*Geometric inequality for the exact $Star_d$ estimator in the second step*)

Let $\hat{g}_1 \dots \hat{g}_d$ be Δ_1 -empirical risk minimizers from the first step of the $Star_d$ procedure, \tilde{f} be the exact minimizer from the second step of the $Star_d$ procedure. Then, for $c_{A.1} = \frac{1}{18}$ the following inequality holds:

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\tilde{f} - Y)^2 \geq c_{A.1} \widehat{\mathbb{E}}(\tilde{f} - h)^2 - 2\Delta_1. \quad (9)$$

Proof. For any function f, g we denote the empirical ℓ_2 distance to be $\|f\|_n := \left[\widehat{\mathbb{E}} f^2 \right]^{\frac{1}{2}}$, empirical product to be $\langle f, g \rangle_n := \widehat{\mathbb{E}} [fg]$ and the square of the empirical distance between \mathcal{F} and Y as r_1 . By definition of $Star_d$ estimator for some $\lambda \in [0; 1]$ we have:

$$\tilde{f} = (1 - \lambda)\hat{g} + \lambda f,$$

where \hat{g} lies in a convex hull of Δ_1 -empirical risk minimizers $\{\hat{g}_i\}_{i=1}^d$. Denote the balls centered at Y to be $\mathcal{B}_1 := \mathcal{B}(Y, \sqrt{r_1})$, $\mathcal{B}'_1 := \mathcal{B}(Y, \|\hat{g} - Y\|_n)$ and $\mathcal{B}_2 := \mathcal{B}(Y, \|f - Y\|_n)$. The corresponding spheres will be called $\mathcal{S}_1, \mathcal{S}'_1, \mathcal{S}_2$. We have $\mathcal{B}_2 \subseteq \mathcal{B}_1$ and $\mathcal{B}_2 \subseteq \mathcal{B}'_1$. Denote by \mathcal{C} the conic hull of \mathcal{B}_2 with origin \hat{g} and define the spherical cap outside the cone \mathcal{C} to be $\mathcal{S} = \mathcal{S}'_1 \setminus \mathcal{C}$.

First, $\tilde{f} \in \mathcal{B}_2$ and it is a contact point of \mathcal{C} and \mathcal{S}_2 . Indeed, \tilde{f} is necessarily on a line segment between \hat{g} and a point outside \mathcal{B}_1 that does not pass through the interior of \mathcal{B}_2 by optimality of \tilde{f} . Let K be the set of all contact points of \mathcal{C} and \mathcal{S}_2 – potential locations of \tilde{f} .

Second, for any $h \in \mathcal{F}$, we have $\|h - Y\|_n \geq \sqrt{r_1}$ i.e. any $h \in \mathcal{F}$ is not in the interior of \mathcal{B}_1 . Furthermore, let \mathcal{C}' be bounded subset cone \mathcal{C} cut at K . Thus $h \in (\text{int}\mathcal{C})^c \cap (\mathcal{B}_1)^c$ or $h \in \mathcal{T}$, where $\mathcal{T} := (\text{int}\mathcal{C}') \cap (\mathcal{B}_1)^c$.

For any $h \in \mathcal{F}$ consider the two dimensional plane \mathcal{L} that passes through three points \hat{g}, Y, h , depicted in Figure 2. Observe that the left-hand side of the desired inequality (9) is constant as \tilde{f} ranges over K . The maximization of $\|h - f'\|_n^2$ over $f' \in K$ is achieved by $f' \in K \cap \mathcal{L}$. Hence, to prove the desired inequality, we can restrict our attention to the plane \mathcal{L} and f' . Let h_\perp be the projection of h onto the shell $L \cap \mathcal{S}'_1$. By the geometry of the cone and triangle inequality we have:

$$\|f' - \hat{g}\|_n \geq \frac{1}{2} \|\hat{g} - h_\perp\|_n \geq \frac{1}{2} (\|f' - h_\perp\|_n - \|f' - \hat{g}\|_n),$$

and, hence, $\|f' - \hat{g}\|_n \geq \|f' - h_\perp\|_n/3$. By the Pythagorean theorem,

$$\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2 = \|\hat{g} - Y\|_n^2 - \|f' - Y\|_n^2 = \|f' - \hat{g}\|_n^2 \geq \frac{1}{9} \|f' - h_\perp\|_n^2.$$

We can now extend this claim to h . Indeed, due to the geometry of the projection $h \rightarrow h_\perp$ and the fact that $h \in (\text{int}\mathcal{C})^c \cap (\text{int}\mathcal{B}_1)^c$ or $h \in \mathcal{T}$ there are 2 possibilities:

a) $h \in (\mathcal{B}'_1)^c$. Then $\langle h_\perp - Y, h_\perp - h \rangle_n \leq 0$;

b) $h \in \mathcal{B}'_1$. Then, since $h \in (\mathcal{B}_1)^c$, we have

$$\langle h_\perp - Y, h_\perp - h \rangle_n \leq (\|h - Y\| + \|h - h_\perp\|) \|h - h_\perp\| \leq \|h_\perp - Y\|_n^2 - \|h - Y\|_n^2 \leq \Delta_1.$$

In both cases, the following inequality is true

$$\begin{aligned} \|h - Y\|_n^2 - \|f' - Y\|_n^2 &= \|h_\perp - h\|_n^2 - 2\langle h_\perp - Y, h_\perp - h \rangle_n + (\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2) \\ &\geq \|h_\perp - h\|_n^2 - 2\Delta_1 + \frac{1}{9} \|f' - Y\|_n^2 \geq \frac{1}{18} \|f' - h\|_n^2 - 2\Delta_1. \end{aligned}$$

407

□

Combining everything together, we get the required result for the constant $c_{A.2} = \frac{c_{A.1}}{2} = \frac{1}{36}$:

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 \geq c_{A.2} \cdot \widehat{\mathbb{E}}(\widehat{f} - h)^2 - 2(1 + c_{A.2})[\Delta_1 + \Delta_2].$$

420

□

For convenience, we introduce a Δ -excess risk

$$\mathcal{E}_\Delta(\widehat{g}) := \mathbb{E}(\widehat{g} - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2 - 2(1 + c_{A.2})[\Delta_1 + \Delta_2],$$

421 then the following 2 statements are direct consequences of the corresponding statements from the
 422 article [LRS15]. The only difference is that in our case the geometric inequality has terms on the
 423 right side with minimization errors Δ_1, Δ_2 . Also our definition of the set \mathcal{H} is different, but all that
 424 was needed from it was the property that \widehat{f} lies in $\mathcal{H} + f^*$. For brevity, we will not repeat the proofs,
 425 but only indicate the numbers of the corresponding results in the titles of the assertions. We will also
 426 proceed for statements whose proofs we slightly modify or use without changes.

Corollary A.3 (Corollary 3). *Conditioned on the data $\{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$, we have deterministic upper bound for the Star_d estimator:*

$$\mathcal{E}_\Delta(\widehat{f}) \leq (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - \widehat{f})] + \mathbb{E}(f^* - \widehat{f})^2 - (1 + c_{A.2}) \cdot \widehat{\mathbb{E}}(f^* - \widehat{f})^2.$$

Theorem A.4 (Theorem 4). *The following expectation bound on excess loss of the Star_d estimator holds:*

$$\mathbb{E} \mathcal{E}_\Delta(\widehat{f}) \leq (2F' + F(2 + c_{A.2})/2) \cdot \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i h(\mathbf{X}_i) - c_{A.4} h(\mathbf{X}_i)^2 \right\},$$

427 where $\sigma_1, \dots, \sigma_n$ are independent Rademacher random variables, $c_{A.4} = \min \left\{ \frac{c_{A.2}}{4F'}, \frac{c_{A.2}}{4F'(2+c_{A.2})} \right\}$,
 428 $F = \sup_{f \in \mathcal{F}} |f|_\infty$ and $F' = \sup_{\mathcal{F}} |Y - f|_\infty$ almost surely.

Theorem A.5 (Theorem 7). *Assume the lower isometry bound in Definition 3.2 holds with $\eta_{lib} = c_{A.2}/4$ and some $\delta_{lib} < 1$ and \mathcal{H} is the set defined in 7. Let $\xi_i = Y_i - f^*(\mathbf{X}_i)$. Define*

$$A := \sup_{h \in \mathcal{H}} \frac{\mathbb{E} h^4}{(\mathbb{E} h^2)^2} \text{ and } B := \sup_{\mathbf{X}, Y} \mathbb{E} \xi^4.$$

Then there exist two absolute constants $c'_{A.5}, \tilde{c}_{A.5} > 0$ (which only depend on $c_{A.2}$), such that

$$\mathbb{P} \left(\mathcal{E}_\Delta(\widehat{f}) > 4u \right) \leq 4\delta_{lib} + 4\mathbb{P} \left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \xi_i h(\mathbf{X}_i) - \tilde{c}_{A.5} h(\mathbf{X}_i)^2 > u \right)$$

for any

$$u > \frac{32\sqrt{AB}}{c'_{A.5}} \frac{1}{n}$$

429 as long as $n > \frac{16(1-c'_{A.5})^2 A}{c_{A.5}^2} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.2}/4)$.

430

Lemma A.6 (Lemma 15). *The offset Rademacher complexity for \mathcal{H} is bounded as:*

$$\mathbb{E}_\sigma \sup_{\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i \xi_i h(\mathbf{X}_i) - C h(\mathbf{X}_i)^2 \right\} \leq K(C)\varepsilon + M(C) \cdot \frac{\log \mathcal{N}_2(\mathcal{H}, \varepsilon)}{n}$$

and with probability at least $1 - \delta$

$$\sup_{\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i \xi_i h(\mathbf{X}_i) - C h(\mathbf{X}_i)^2 \right\} \leq K(C)\varepsilon + M(C) \cdot \frac{\log \mathcal{N}_2(\mathcal{H}, \varepsilon) + \log \frac{1}{\delta}}{n},$$

431 where

$$K(C) := 2 \left(\sqrt{\sum_{i=1}^n \xi_i^2 / n} + C \right), \quad M(C) := \sup_{h \in \mathcal{H} \setminus \{0\}} 4 \frac{\sum_{i=1}^n h(\mathbf{X}_i)^2 \xi_i^2}{C \sum_{i=1}^n h(\mathbf{X}_i)^2}. \quad (10)$$

432 *Proof.* Let $N_2(\mathcal{H}, \varepsilon)$ be the ε -net of the \mathcal{H} of size at most $\mathcal{N}_2(\mathcal{H}, \varepsilon)$ and $v[h]$ be the closest point from
 433 this net for function $h \in \mathcal{H}$, i.e. $\|h - v[h]\|_2 \leq \varepsilon$. By using the inequality $v[h]_i^2 \leq 2(h_i^2 + (v[h]_i -$
 434 $h_i)^2)$, we can get next upper bound:

$$\begin{aligned} & \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i \xi_i h(\mathbf{X}_i) - Ch(\mathbf{X}_i)^2 \right\} \\ & \leq \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i \xi_i (h(\mathbf{X}_i) - v[h](\mathbf{X}_i)) + C(v[h]^2(\mathbf{X}_i)/2 - h^2(\mathbf{X}_i)) \right\} \\ & \quad + \frac{1}{n} \sup_{v \in N_2(\mathcal{H}, \varepsilon)} \left\{ \sum_{i=1}^n 2\sigma_i \xi_i v(\mathbf{X}_i) - \frac{C}{2} v(\mathbf{X}_i)^2 \right\} \\ & \leq 2\varepsilon \left(\sqrt{\sum_{i=1}^n \xi_i^2 / n} + C \right) + \frac{1}{n} \sup_{v \in N_2(\mathcal{H}, \varepsilon)} \left\{ \sum_{i=1}^n 2\sigma_i \xi_i v(\mathbf{X}_i) - \frac{C}{2} v(\mathbf{X}_i)^2 \right\}. \end{aligned}$$

435 The right summarand is supremum over set of cardinality not more than $\mathcal{N}_2(\mathcal{H}, \varepsilon)$. By using Lemma
 436 A.11, we acquire the expected estimates. \square

437 We have now obtained, using the offset Rademacher complexity technique, the upper bound on excess
 438 risk in terms of the coverage size of the set \mathcal{H} . To get the desired result, we need to obtain an upper
 439 bound on the size of the cover \mathcal{H} in terms of the size of the cover \mathcal{F} .

Lemma A.7. *For any scale $\varepsilon > 0$, the covering number of $\mathcal{F} \subseteq V(L+1) \cdot \mathcal{B}_2$ (where \mathcal{B}_2 is a sphere of radius one in space with norm $\|\cdot\|_n$) and that of \mathcal{H} are bounded in the sense:*

$$\log \mathcal{N}_2(\mathcal{F}, \varepsilon) \leq \log \mathcal{N}_2(\mathcal{H}, \varepsilon) \leq (d+2) \left[\log \mathcal{N}_2 \left(\mathcal{F}, \frac{\varepsilon}{3(d+1)} \right) + \log \frac{6(d+1)V(L+1)}{\varepsilon} \right].$$

Proof. If we define as $N(\mathcal{F}, \varepsilon)$ the ε -net cardinality no more than $\mathcal{N}(\mathcal{F}, \varepsilon)$, then the following is true: $N(\mathcal{F}_1, \varepsilon_1) + N(\mathcal{F}_2, \varepsilon_2)$ is $(\varepsilon_1 + \varepsilon_2)$ -net for $\mathcal{F}_1 + \mathcal{F}_2$. Hence, $\mathcal{N}(\mathcal{F}_1 + \mathcal{F}_2, \varepsilon_1 + \varepsilon_2) \leq \mathcal{N}(\mathcal{F}_1, \varepsilon_1) \cdot \mathcal{N}(\mathcal{F}_2, \varepsilon_2)$. With this we can obtain the following upper bound

$$\mathcal{N}_2(\mathcal{H}, \varepsilon) \leq \mathcal{N}_2(\mathcal{F} + \text{Hull}_d, \varepsilon) \leq \mathcal{N}_2 \left(\mathcal{F}, \frac{\varepsilon}{3} \right) \cdot \mathcal{N}_2 \left(\text{Hull}_d, \frac{2\varepsilon}{3} \right).$$

But since Hull_d is the sum of $d+1$ functions from \mathcal{F} with coefficients in $[-1; 1]$, by the inequality (3), we can cover this with a net of size no more than

$$\left[\mathcal{N}_2 \left(\mathcal{F}, \frac{\varepsilon}{3(d+1)} \right) \cdot \frac{6(d+1)V(L+1)}{\varepsilon} \right]^{d+1}.$$

440 \square

441 Note that to obtain the required orders, we only need coverage with $\varepsilon = 1/n$.

Corollary A.8. *Let \mathcal{H} defined in 7 for $\mathcal{F} = \mathcal{F}(L, \mathbf{p}, s)$, then for V defined in 5 holds*

$$\log \mathcal{N}_2 \left(\mathcal{H}, \frac{1}{n} \right) \leq c_{A.8} d s \log(V L n d),$$

442 where $c_{A.8}$ is a some independent constant.

443 *Proof.* By lemma A.7 and inequality 4, we have

$$\begin{aligned} \log \mathcal{N}_2(\mathcal{H}, 1/n) & \leq (d+2) \left[\log \mathcal{N}_2 \left(\mathcal{F}(L, \mathbf{p}, s), \frac{1}{3n(d+1)} \right) + \log 6n(d+1)V(L+1) \right] \\ & \leq (d+2) [(s+1) \log(2V^2(L+1)(3n(d+1))) + \log(6n(d+1)V(L+1))]. \end{aligned}$$

444 \square

445 We are now fully prepared to prove the two main results.

Theorem A.9. Let \hat{f} be a Star_d estimator and \mathcal{H} be the set defined in 7 for $\mathcal{F} = \mathcal{F}(L, \mathbf{p}, s)$. The following expectation bound on excess loss holds:

$$\mathbb{E} \mathcal{E}_\Delta(\hat{f}) \leq 2(F' + V(L+1)) \cdot \left[\frac{K(C)}{n} + M(C) \cdot \frac{c_{A.8} d s \log(VLnd)}{n} \right],$$

where $K(C)$, $M(C)$ defined in (10) for constants

$$C = \min \left\{ \frac{c_{A.2}}{4F'}, \frac{c_{A.2}}{4V(L+1)(2+c_{A.2})} \right\}, \quad F' = \sup_{\mathcal{F}} |Y - f|_\infty.$$

Proof. By using Theorem A.4 and inequality 3 we have

$$\mathbb{E} \mathcal{E}_\Delta(\hat{f}) \leq (2F' + V(L+1)(2+c_{A.2})/2) \cdot \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i h(\mathbf{X}_i) - Ch(\mathbf{X}_i)^2 \right\},$$

446 where $C = \min \left\{ \frac{c_{A.2}}{4F'}, \frac{c_{A.2}}{4V(L+1)(2+c_{A.2})} \right\}$, $F' = \sup_{\mathcal{F}} |Y - f|_\infty$ almost surely.

By using Lemma A.6 and corollary A.8 we get desired result

$$\mathbb{E}_\sigma \sup_{\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\sigma_i \xi_i h(\mathbf{X}_i) - Ch(\mathbf{X}_i)^2 \right\} \leq \frac{K(C)}{n} + M(C) \cdot \frac{c_{A.8} d s \log(VLnd)}{n}.$$

447

□

Theorem A.10. Let \hat{f} be a Star_d estimator and let \mathcal{H} be the set defined in 7 for $\mathcal{F} = \mathcal{F}(L, \mathbf{p}, s)$. Assume for \mathcal{H} the lower isometry bound in Definition 3.2 holds with $\eta_{lib} = c_{A.2}/4$ and some $\delta_{lib} < 1$. Let $\xi_i = Y_i - f^*(\mathbf{X}_i)$. Define

$$A := \sup_{h \in \mathcal{H}} \frac{\mathbb{E} h^4}{(\mathbb{E} h^2)^2} \text{ and } B := \sup_{\mathbf{X}, Y} \mathbb{E} \xi^4.$$

Then there exist 3 absolute constants $c'_{A.10}, c_{\tilde{A}.10}, c_{A.10} > 0$ (that only depend on $c_{A.2}$), such that

$$\mathbb{P} \left(\mathcal{E}_\Delta(\hat{f}) > 4D \right) \leq 4(\delta_{lib} + \delta)$$

as long as $n > \frac{16(1-c'_{A.10})^2 A}{c_{\tilde{A}.10}^2} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.10}/4)$, where

$$K := \left(\sqrt{\sum_{i=1}^n \xi_i^2 / n} + 2c_{\tilde{A}.10} \right), \quad M := \sup_{h \in \mathcal{H} \setminus \{0\}} \frac{\sum_{i=1}^n h(\mathbf{X}_i)^2 \xi_i^2}{c_{\tilde{A}.10} \sum_{i=1}^n h(\mathbf{X}_i)^2},$$

$$D := \max \left(\frac{K}{n} + M \cdot \frac{c_{A.8} d s \log(VLnd) + \log \frac{1}{\delta}}{n}, \frac{32\sqrt{AB}}{c'_{A.10}} \frac{1}{n} \right)$$

448 and $c_{A.8}$ is some independent constant.

Proof. By using Theorem A.5 for any $u > \frac{32\sqrt{AB}}{c'_{A.5}} \frac{1}{n}$ we have

$$\mathbb{P} \left(\mathcal{E}_\Delta(\hat{f}) > 4u \right) \leq 4\delta_{lib} + 4\mathbb{P} \left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \xi_i h(\mathbf{X}_i) - c_{\tilde{A}.5} h(\mathbf{X}_i)^2 > u \right)$$

449 as long as $n > \frac{16(1-c'_{A.5})^2 A}{c_{\tilde{A}.5}^2} \vee n_0(\mathcal{H}, \delta_{lib}, c_{A.2}/4)$.

450 By using Lemmas A.6 and A.8 we have with probability no more than δ for any $C > 0$:

$$\sup_{\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n \sigma_i \xi_i h(\mathbf{X}_i) - \frac{C}{2} h(\mathbf{X}_i)^2 \right\} \geq \frac{K(C)}{2} \varepsilon + \frac{M(C)}{2} \cdot \frac{\log \mathcal{N}_2(\mathcal{H}, \varepsilon) + \log \frac{1}{\delta}}{n},$$

451 where $K(C)$, $M(C)$ defined in (10). Combining this inequality for $C = 2c_{\tilde{A}.10} = 2c_{\tilde{A}.5}$ and

452 $c'_{A.10} = c'_{A.5}$, $c_{A.10} = c_{A.2}$ we get the required result. □

Lemma A.11 (Lemma 9). *Let $V \subset \mathbb{R}^n$ be a finite set, $|V| = N$. Then, for any $C > 0$:*

$$\mathbb{E}_\sigma \max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \sigma_i \xi_i v(\mathbf{X}_i) - Cv(\mathbf{X}_i)^2 \right] \leq M \frac{\log N}{n}.$$

For any $\delta > 0$:

$$\mathbb{P}_\sigma \left(\max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \sigma_i \xi_i v(\mathbf{X}_i) - Cv(\mathbf{X}_i)^2 \right] > M \frac{\log N + \log \frac{1}{\delta}}{n} \right) \leq \delta,$$

where

$$M := \sup_{v \in V \setminus \{0\}} \frac{\sum_{i=1}^n v(\mathbf{X}_i)^2 \xi_i^2}{2C \sum_{i=1}^n v(\mathbf{X}_i)^2}.$$

B Result Tables

Here we additionally present tables with the results of numerical experiments. Particularly for runs with a small number of *epochs*. It can be observed that the SnapStar algorithm is quite good with a strong budget constraint. The results also include relatively large run for the FASHION MNIST dataset. At the moment, ClassicStar (new warm-up) takes 10 – 11th place in the leaderboard⁵ for this dataset. Full versions of the following tables can be found in the repository⁶.

Name	d	MSE	MAE	R^2	TRAIN MSE	TIME (sec)
Snap Star (shot warm-up)	5	10.881±0.575	2.229	0.869	1.976	7.8
Snap Star (new warm-up)	5	11.285±0.650	2.283	0.864	2.656	6.6
Snap Ensemble	5	11.862±0.616	2.306	0.858	2.629	6.6
Ensemble	5	12.568±0.878	2.399	0.849	4.220	6.8
Classic Star (no warm-up)	5	11.365±0.410	2.278	0.864	2.978	7.2
Classic Star (new warm-up)	5	12.157±0.822	2.353	0.854	3.320	6.2
Big NN	5	12.068±0.860	2.411	0.855	3.644	4.0
Snap Star (shot warm-up)	4	11.276±0.582	2.269	0.865	2.329	6.2
Snap Star (new warm-up)	4	11.598±0.729	2.292	0.861	2.739	5.0
Snap Ensemble	4	11.819±0.341	2.316	0.858	2.819	5.0
Ensemble	4	12.059±0.614	2.365	0.855	3.732	5.0
Classic Star (no warm-up)	4	11.608±0.722	2.286	0.861	3.198	6.2
Classic Star (new warm-up)	4	11.890±0.966	2.319	0.857	3.093	5.2
Big NN	4	12.556±0.904	2.383	0.849	3.746	4.0

Table 4: BOSTON HOUSE PRICING. Part of results at 30 epochs, $p = 0.1$, $lr = 0.01$

Name	d	MSE	MAE	R2	TRAIN MSE	TIME (sec)
Snap Star (shot warm-up)	5	76.31 ± 0.17	5.97	0.362	70.64	733
Snap Star (new warm-up)	5	76.21 ± 0.10	5.99	0.363	71.34	667
Snap Ensemble	5	76.42 ± 0.11	6.02	0.361	70.03	543
Ensemble	5	76.34 ± 0.07	6.05	0.361	72.05	711
Classic Star (no warm-up)	5	76.57 ± 0.15	6.07	0.36	73.62	783
Classic Star (new warm-up)	5	76.06 ± 0.10	6.00	0.364	72.59	807
Big NN	5	77.04 ± 0.21	6.02	0.356	75.62	436
Snap Star (shot warm-up)	4	76.30 ± 0.12	5.99	0.362	71.04	632
Snap Star (new warm-up)	4	76.14 ± 0.11	6.01	0.363	71.78	565
Snap Ensemble	4	76.46 ± 0.12	6.02	0.360	70.37	452
Ensemble	4	76.40 ± 0.08	6.05	0.361	72.08	593
Classic Star (no warm-up)	4	76.51 ± 0.04	6.04	0.36	73.76	652
Classic Star (new warm-up)	4	76.01 ± 0.10	6.01	0.364	72.69	676
Big NN	4	77.06 ± 0.18	6.03	0.355	75.63	375
Snap Star (shot warm-up)	3	76.39 ± 0.32	5.98	0.361	71.62	530
Snap Star (new warm-up)	3	76.10 ± 0.07	6.00	0.363	72.38	463
Snap Ensemble	3	76.53 ± 0.14	6.02	0.360	70.77	362
Ensemble	3	76.43 ± 0.09	6.05	0.361	72.12	473
Classic Star (no warm-up)	3	76.51 ± 0.12	6.04	0.360	74.00	522
Classic Star (new warm-up)	3	76.16 ± 0.13	6.01	0.363	72.81	546
Big NN	3	76.80 ± 0.23	6.03	0.358	75.60	315

Table 5: MILLIION SONG. Part of results at 10 epochs

⁵<https://paperswithcode.com/sota/image-classification-on-fashion-mnist>

⁶A link on a GitHub repository will be provided in the final version.

Name	d	accuracy	entropy	TIME (sec)
Snap Star (shot warm-up)	3	0.900±0.002	0.284±0.008	340.333
Snap Star (new warm-up)	3	0.898±0.002	0.285±0.008	313.0
Snap Ensemble	3	0.897±0.003	0.290±0.009	272.667
Ensemble	3	0.887±0.001	0.310±0.005	272.667
Classic Star (no warm-up)	3	0.893±0.002	0.298±0.007	339.667
Classic Star (new warm-up)	3	0.893±0.002	0.297±0.007	285.667
Big NN	3	0.890±0.010	0.299±0.022	214.333
Snap Star (shot warm-up)	2	0.894±0.007	0.294±0.020	248.667
Snap Star (new warm-up)	2	0.892±0.001	0.294±0.006	230.333
Snap Ensemble	2	0.891±0.006	0.302±0.021	203.667
Ensemble	2	0.886±0.004	0.313±0.008	203.0
Classic Star (no warm-up)	2	0.889±0.003	0.304±0.009	249.0
Classic Star (new warm-up)	2	0.889±0.004	0.303±0.008	203.667
Big NN	2	0.892±0.003	0.304±0.007	165.333
Snap Star (shot warm-up)	1	0.891±0.002	0.299±0.006	159.0
Snap Star (new warm-up)	1	0.885±0.001	0.318±0.008	149.333
Snap Ensemble	1	0.889±0.001	0.304±0.007	136.0
Ensemble	1	0.886±0.005	0.314±0.011	136.333
Classic Star (no warm-up)	1	0.888±0.002	0.311±0.001	158.0
Classic Star (new warm-up)	1	0.891±0.002	0.302±0.005	122.333
Big NN	1	0.886±0.002	0.315±0.005	117.333

Table 6: FASHION MNIST. Part of results at 5 epochs, $lr = 0.001$

Name	d	accuracy	entropy	TIME (sec)
Snap Star (shot warm-up)	5	0.898	1.152	2588.0
Snap Star (new warm-up)	5	0.898	1.136	2369.0
Snap Ensemble	5	0.902	0.330	2036.0
Ensemble	5	0.918	0.229	2052.0
Classic Star (no warm-up)	5	0.922	0.229	2589.0
Classic Star (new warm-up)	5	0.923	0.228	2239.0
Big NN	5	0.910	0.481	1560.0

Table 7: FASHION MNIST. All of results at 25 epochs, $lr = 0.001$