

# IMPLICIT REGULARIZATION FOR OPTIMAL SPARSE RECOVERY

Tomas Vaškevičius<sup>1</sup>, Varun Kanade<sup>2</sup>, Patrick Rebeschini<sup>1</sup>

<sup>1</sup> Department of Statistics, <sup>2</sup> Department of Computer Science



## Problem Setting

Let  $\mathbf{w}^* \in \mathbb{R}^d$  be a  $k$ -sparse vector with  $\mathbf{k} \ll \mathbf{d}$ . We observe  $\mathbf{n} \ll \mathbf{d}$  data points  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, i \in \{1, \dots, n\}$  such that in matrix-vector notation the model reads

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \xi,$$

where  $\xi$  is a vector of independent  $\sigma^2$ -sub-Gaussian noise random variables. We want to find an estimator  $\hat{\mathbf{w}} \in \mathbb{R}^d$  with small **parameter estimation error**  $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2$ .

Classical approaches to solving the above problem add an explicit sparsity-inducing penalty term to the optimization objective. For example, the lasso is a solution to

$$\min_{\hat{\mathbf{w}} \in \mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 + \lambda \|\hat{\mathbf{w}}\|_1.$$

In this work, we investigate **implicit regularization** schemes for **gradient descent methods** applied to **unpenalized** least squares regression to solve the above problem.

## Reparameterization

The mean squared error is given by  $\mathcal{L}(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2/n$ . Performing gradient descent updates on  $\mathcal{L}(\mathbf{w})$  together with early-stopping induces a regularization effect similar to  $\ell_2$  penalization (ridge regression). This type of regularization does not induce sparsity and hence is unsuitable for solving our problem. Updates on  $\mathcal{L}(\mathbf{w})$  in this case read as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t) = \mathbf{w}_t - (2\eta)/n (\mathbf{X}^\top \mathbf{X}(\mathbf{w}_t - \mathbf{w}^*) - \mathbf{X}^\top \xi)$$

Instead, the key is to consider the following **reparametrization**. Let  $\odot$  denote a coordinate-wise multiplication for vectors. For  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$  let  $\mathbf{w} = \mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}$  and define the mean squared error objective on  $\mathbf{u}$  and  $\mathbf{v}$  as

$$\mathcal{L}(\mathbf{u}, \mathbf{v}) = \|\mathbf{X}(\mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}) - \mathbf{y}\|_2^2/n.$$

We show that using the above parametrization and applying gradient-based updates on  $(\mathbf{u}, \mathbf{v})$  instead of  $\mathbf{w}$  results in **sparsity-inducing implicit regularization effect**. For a constant learning rate  $\eta$ , the updates on  $\mathbf{u}$  and  $\mathbf{v}$  are given by

$$\begin{aligned} \mathbf{u}_{t+1} &= \mathbf{u}_t - \eta \frac{\partial \mathcal{L}(\mathbf{u}_t, \mathbf{v}_t)}{\partial \mathbf{u}_t} = \mathbf{u}_t \odot \left( \mathbf{1} - 4\eta \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} (\mathbf{w}_t - \mathbf{w}^*) - \frac{1}{n} \mathbf{X}^\top \xi \right) \right), \\ \mathbf{v}_{t+1} &= \mathbf{v}_t - \eta \frac{\partial \mathcal{L}(\mathbf{u}_t, \mathbf{v}_t)}{\partial \mathbf{v}_t} = \mathbf{v}_t \odot \left( \mathbf{1} + 4\eta \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} (\mathbf{w}_t - \mathbf{w}^*) - \frac{1}{n} \mathbf{X}^\top \xi \right) \right). \end{aligned}$$

The idea of the above parametrization comes from recent work on matrix factorization models, where low-rank constraints are imposed by letting  $\mathbf{W} = \mathbf{U}\mathbf{U}^\top$  [1, 2].

## Restricted Isometry Property

Our analysis is based on a standard assumption in the compressed sensing literature.

**Definition 1** (Restricted Isometry Property (RIP)). A  $n \times d$  matrix  $\mathbf{X}/\sqrt{n}$  satisfies the  $(\delta, k)$ -(RIP) if for any  $k$ -sparse vector  $\mathbf{w} \in \mathbb{R}^d$  we have

$$(1 - \delta) \|\mathbf{w}\|_2^2 \leq \|\mathbf{X}\mathbf{w}/\sqrt{n}\|_2^2 \leq (1 + \delta) \|\mathbf{w}\|_2^2.$$

Intuitively, the RIP assumption allows treating  $\mathbf{X}^\top \mathbf{X}/n$  as the identity matrix for sparse vectors. Various i.i.d. random ensembles (e.g., Gaussian or Rademacher) satisfy the RIP.

## Theorem 1 – Minimax Optimality

We define  $w_{\max}^* = \max_i |w_i^*|$  and  $w_{\min}^* = \min_{i: w_i^* \neq 0} |w_i^*|$ . We assume that  $\mathbf{X}/\sqrt{n}$  satisfies  $(\delta, k+1)$ -RIP with  $\delta = \tilde{O}(1/\sqrt{k})$ . Such a condition requires the dataset size  $n$  to scale quadratically with sparsity  $k$ , that is,  $\mathbf{n} = \Omega(k^2 \log(d/k))$ .

To prevent explosion, it is necessary to set the learning rate  $\eta \lesssim 1/w_{\max}^*$ . It is possible to estimate  $w_{\max}^*$  up to multiplicative constants at the computational cost of one gradient descent iteration, that is  $O(nd)$ . Hence, we let  $\eta \asymp 1/w_{\max}^*$ .

Set  $\mathbf{u}_0 = \mathbf{v}_0 = \alpha \mathbf{1}$ , where the initialization size  $\alpha$  satisfies  $0 < \alpha \leq \frac{\sigma^2 \wedge \sigma}{n((2d+1)\vee w_{\max}^*)^2} \wedge \frac{\sqrt{w_{\min}^*}}{2}$ . In particular,  $\alpha$  is a **polynomial function** in  $d^{-1}, n^{-1}, (w_{\max}^*)^{-1}, w_{\min}^*, \sigma$ , while the optimal stopping time is only affected **logarithmically** in  $\alpha^{-1}$ .

Then, after  $t = O(\frac{w_{\max}^* \sqrt{n}}{\sigma \sqrt{\log d}} \log \frac{1}{\alpha}) = \tilde{O}(\frac{w_{\max}^* \sqrt{n}}{\sigma})$  iterations we have  $\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \lesssim k \frac{\sigma^2 \log d}{n}$  with probability at least  $1 - 1/(8d^3)$ .

This rate is **minimax optimal** for sub-linear sparsity and cannot be improved in general.

## Key Proof Ideas

Our parametrization turns **additive** updates into **multiplicative** updates.

For every coordinate  $i$ ,  $\mathbf{u}_{t+1} \odot \mathbf{v}_{t+1} \preceq \mathbf{u}_t \odot \mathbf{v}_t$  hence for each  $i$   $\mathbf{u}_{t,i} \wedge \mathbf{v}_{t,i} \leq \alpha \approx 0$ .

Hence, for simplicity assume  $\mathbf{w}^* \succcurlyeq 0$  and use parametrization  $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{v}_t$ .

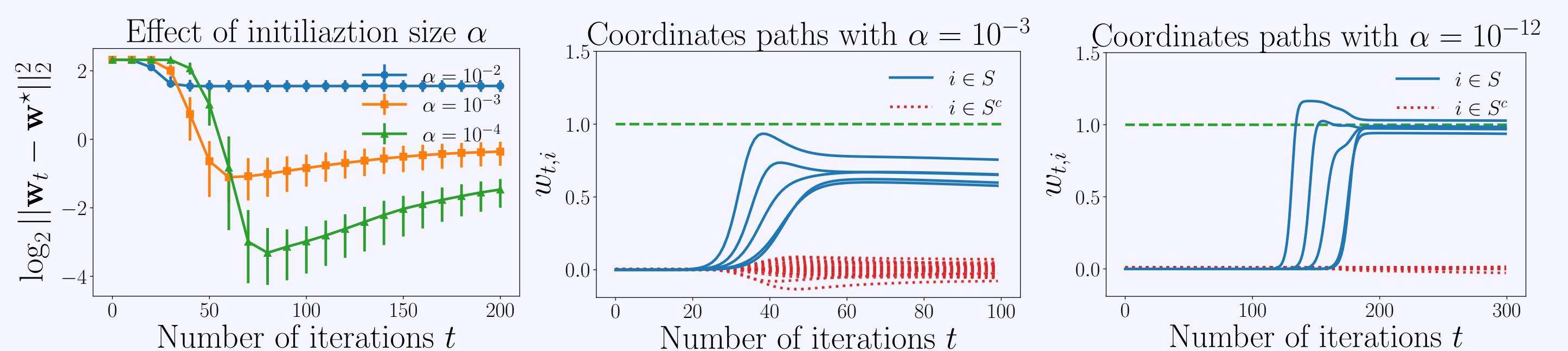
Assume  $\mathbf{X}^\top \mathbf{X}/n = \mathbf{I}$ . The updates become  $\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^* - \mathbf{X}^\top \xi/n))^2$ .

Then, the  $i^{th}$  coordinate converges in  $O(\eta^{-1} |\mathbf{w}_i^* + (\mathbf{X}^\top \xi)_i|^{-1} \log \alpha^{-1})$  iterations.

Hence, all coordinates converge **exponentially fast at different rates**.

## Necessity of Small Initialization

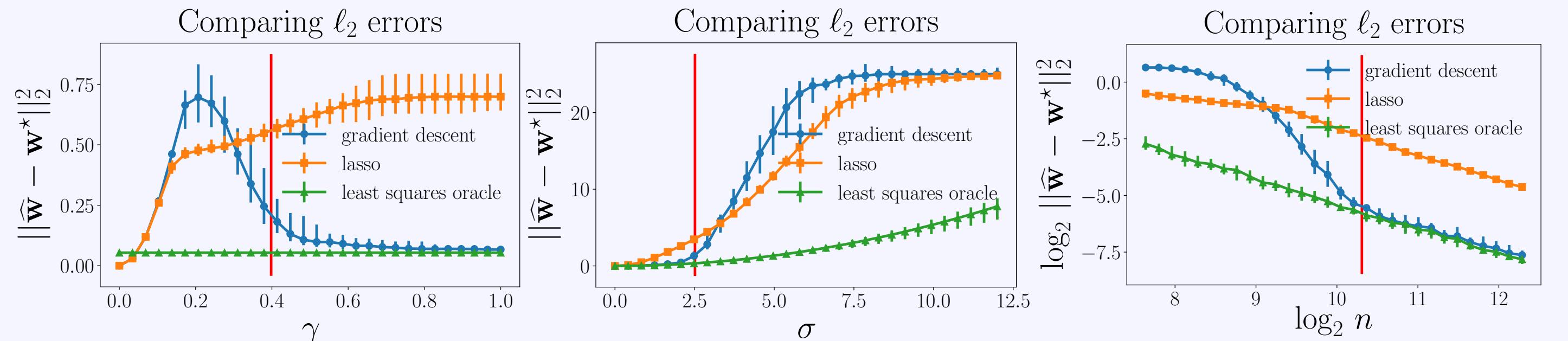
For any  $\varepsilon > 0$  and large enough  $t$  we have  $(1 + 2\varepsilon)^t \gg (1 + \varepsilon)^t$ . Hence with small enough  $\alpha$  we get the effect of **fitting coordinates one by one**.



In the plots above,  $S$  denotes the true support of  $\mathbf{w}^*$ . We let  $\mathbf{w}^* = \mathbf{1}_S$  (1 on  $S$ , 0 otherwise).

## Phase Transitions

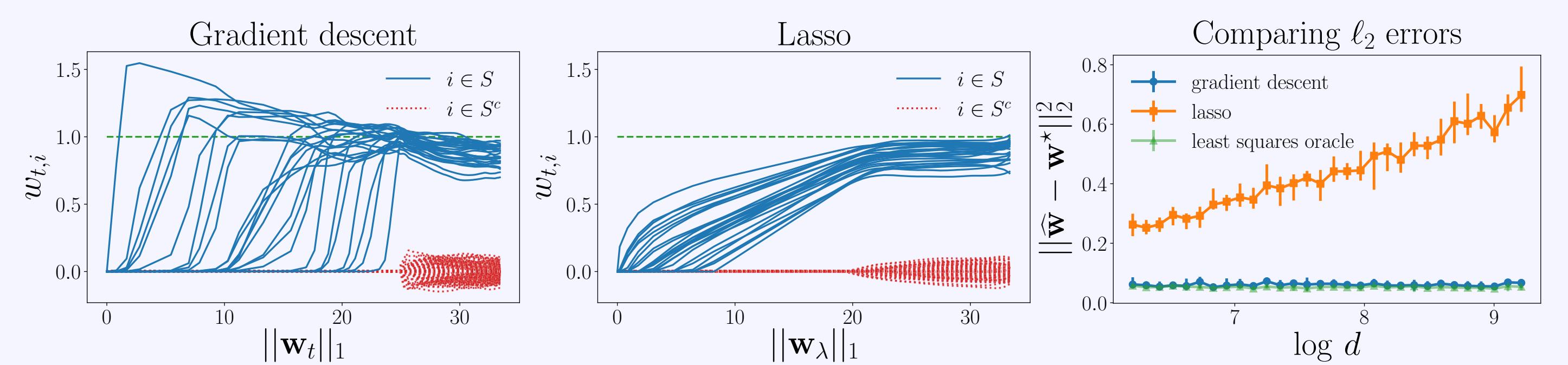
With the intuition above, as soon as  $w_{\min}^* - \|\frac{1}{n} \mathbf{X}^\top \xi\|_\infty > \|\frac{1}{n} \mathbf{X}^\top \xi\|_\infty$  all coordinates on the true support  $S$  grow exponentially at a faster rate than the coordinates on  $S^c$ .



Let  $\mathbf{w}^* = \gamma \mathbf{1}_S$ . Then,  $w_{\min}^* = \gamma$  and the red lines denote solutions to  $\gamma = 2 \cdot \mathbb{E}[\|\mathbf{X}^\top \xi\|_\infty/n]$ .

## Theorem 2 – Dimension Free Bounds

Consider the setting of Theorem 1. If in addition we have  $w_{\min}^* \gtrsim \|\mathbf{X}^\top \xi\|_\infty/n$  then after  $t = \tilde{O}(\frac{w_{\max}^* \sqrt{n}}{\sigma})$  iterations we have  $\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \lesssim k \frac{\sigma^2 \log k}{n}$  with probability at least  $1 - 1/(8k^3)$ .



## Theorem 3 – Computational Optimality

The coordinates  $i$  such that  $|w_i^*| \gtrsim w_{\max}^*$  converge in  $C \cdot \log \alpha^{-1}$  iterations after which the learning rate **for other coordinates can be safely increased**.

We can estimate  $\widehat{w}_{\max}^*$  such that  $w_{\max}^* \leq \widehat{w}_{\max}^* \leq 2w_{\max}^*$  in  $O(nd)$  time. For  $m = 2, 3, \dots$ , after every  $t = m \cdot C \cdot \log \alpha^{-1}$  iterations, **double the learning rate** for all  $i$  such that  $|w_{t,i}| \leq 2^{m-1} \widehat{w}_{\max}^*$ .

The resulting algorithm requires  $\tilde{O}(1)$  iterations, and hence,  $\tilde{O}(nd)$  time.

## References

- [1] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.
- [2] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.