Implicit Regularization for Optimal Sparse Recovery

Varun Kanade[1], Patrick Rebeschini[2], Tomas Vaškevičius[2]

[1] Department of Computer Science, [2] Department of Statistics

UNIVERSITY OF OXFORD

Problem Setting

- Let $\mathbf{w}^\star \in \mathbb{R}^d$ be a k-sparse vector with $\boldsymbol{k \ll d}$. We observe $\boldsymbol{n \ll d}$ 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
$$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi,$$
where ξ is a vector of independent σ^2-sub-Gaussian noise random variables. We want to find an estimator $\widehat{\mathbf{w}} \in \mathbb{R}^d$ with small **parameter estimation error** $\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2^2$.
- 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
$$\min_{\widehat{\mathbf{w}} \in \mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\widehat{\mathbf{w}} - \mathbf{y}\|_2^2 + \lambda \|\widehat{\mathbf{w}}\|_1 .$$
- 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 ℓ_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
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t) = \mathbf{w}_t - (2\eta)/n\left(\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \mathbf{X}^\mathsf{T}\xi\right)$$
- Instead, the key is to consider the following **reparameterization**. 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 parameterization 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 η, the updates on $\mathbf{u}$ and $\mathbf{v}$ are given by
$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t, \mathbf{v}_t)}{\partial\mathbf{u}_t} = \mathbf{u}_t \odot \left(\mathbb{1} - 4\eta\left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right)\right),$$
$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t, \mathbf{v}_t)}{\partial\mathbf{v}_t} = \mathbf{v}_t \odot \left(\mathbb{1} + 4\eta\left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right)\right).$$
- The idea of the above parameterization comes from recent work on matrix factorization models, where low-rank constraints are imposed by letting $\mathbf{W} = \mathbf{U}\mathbf{U}^\mathsf{T}$ [1].

Restricted Isometry Property

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

Definition 1 (Restricted Isometry Property (RIP)). *A $n \times d$ matrix $\mathbf{X}/\sqrt{n}$ satisfies the (δ, k)-(RIP) if for any k-sparse vector $\mathbf{w} \in \mathbb{R}^d$ we have*
$$(1 - \delta)\|\mathbf{w}\|_2^2 \leq \left\|\mathbf{X}\mathbf{w}/\sqrt{n}\right\|_2^2 \leq (1 + \delta)\|\mathbf{w}\|_2^2 .$$

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

Theorem 1 – Minimax Optimality

Below $\lesssim$ denotes inequalities up to absolute multiplicative constants. Notation $a \asymp b$ means $a \lesssim b \lesssim a$. We also define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i: w^\star_i \neq 0} |w^\star_i|$. Finally, the notation $\widetilde{O}$ is used to hide logarithmic factors.

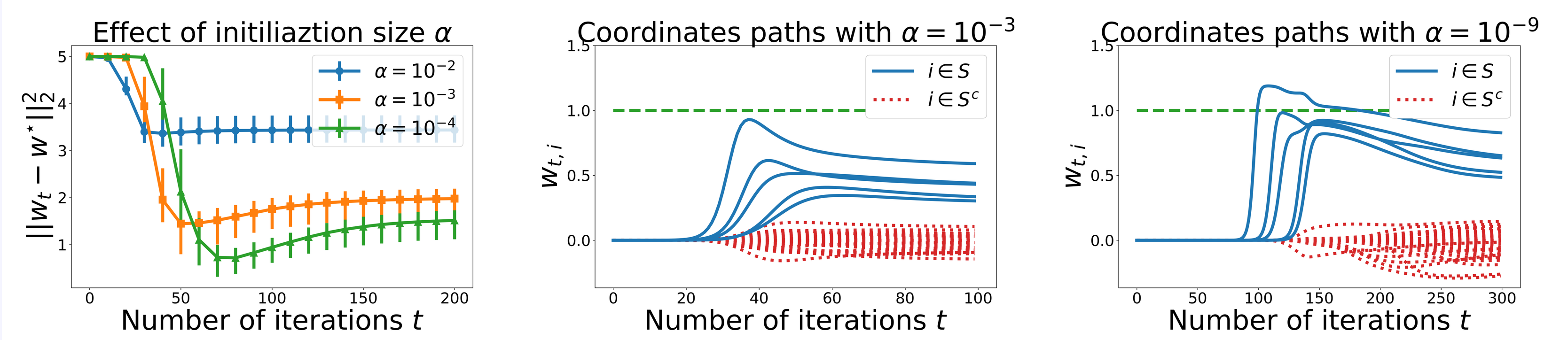
- We assume that $\mathbf{X}/\sqrt{n}$ satisfies $(\delta, k{+}1)$-RIP with $\boldsymbol{\delta = \widetilde{O}(1/\sqrt{k})}$. Such a condition requires dataset size n to scale quadratically with sparsity k, that is $\boldsymbol{n = \Omega(k^2 \log(d/k))}$.
- To prevent explosion, it is necessary to set the learning rate $\eta \lesssim 1/w^\star_{\max}$. It is possible to estimate $w^\star_{\max}$ up to multiplicative constants at the computational cost of one gradient descent iteration, that is $O(nd)$. Hence, we let $\boldsymbol{\eta \asymp 1/w^\star_{\max}}$.
- Set $\mathbf{u}_0 = \mathbf{v}_0 = \alpha$, where the initialization size α satisfies $0 < \alpha \leq \frac{\sigma^2 \wedge \sigma}{n((2d+1) \vee w^\star_{\max})^2} \wedge \frac{\sqrt{w^\star_{\min}}}{2}$. In particular, initialization size α is a **polynomial function** in $d^{-1}, n^{-1}, (w^\star_{\max})^{-1}, w^\star_{\min}, \sigma$, while the optimal stopping time (see below) is only affected **logarithmically** in α^{-1}.
- Then, after $t = O(\frac{w^\star_{\max}\sqrt{n}}{\sigma\sqrt{\log d}} \log\frac{1}{\alpha}) = \boldsymbol{\widetilde{O}(\frac{w^\star_{\max}\sqrt{n}}{\sigma})}$ iterations we have $\boldsymbol{\|\mathbf{w}_t - \mathbf{w}^\star\|_2^2 \lesssim k\frac{\sigma^2 \log d}{n}}$ with probability at least $1 - 1/(8d^3)$.
- The above rate is **minimax optimal** for sub-linear sparsity and cannot be improved in general.

Key Proof Ideas

- Our parameterization turns **additive** updates into **multiplicative** updates.
- For every coordinate i, $\mathbf{u}_{t+1} \odot \mathbf{v}_{t+1} \preccurlyeq \mathbf{u}_t \odot \mathbf{v}_t$ hence for each i $\boldsymbol{u_{t,i} \wedge v_{t,i} \leq \alpha \approx 0}$. Hence for simplicity assume $\mathbf{w}^\star \succcurlyeq 0$ and use parameterization $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$.
- Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$. The updates become $\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^\star - \mathbf{X}^\mathsf{T}\xi/n))^2$.
- Then, i-th coordinate converges in $O(\eta^{-1}\left|\boldsymbol{w^\star_i + (\mathbf{X}^\mathsf{T}\xi)_i/n}\right|^{-1} \log\alpha^{-1})$ iterations.
- Hence, all coordinates converge **exponentially fast at different rates**.

Necessity of Small Initialization

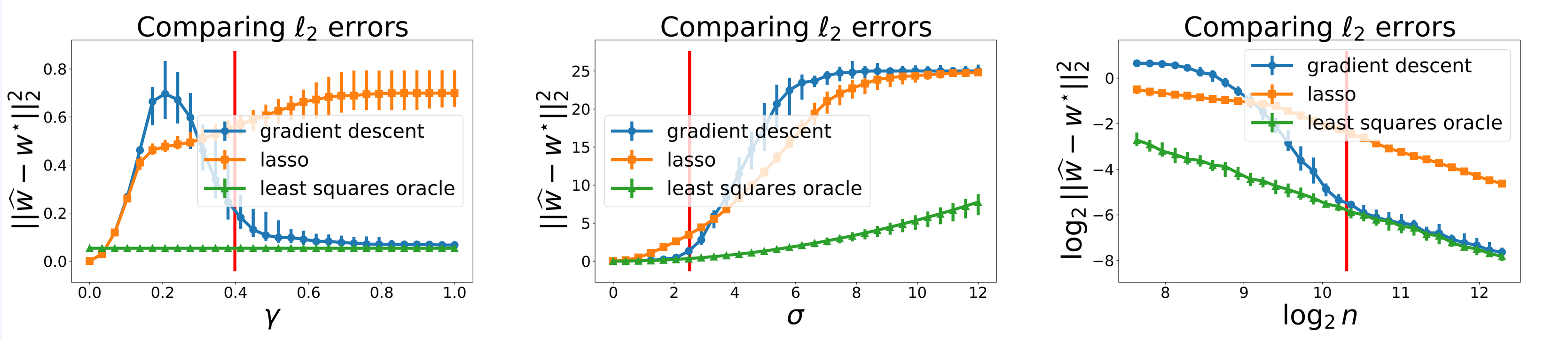
For any $\varepsilon > 0$ and large enough t we have $(1 + 2\varepsilon)^t \gg (1 + \varepsilon)^t$. Hence with small enough α we get the effect of **fitting coordinates one by one**.



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

Phase Transitions

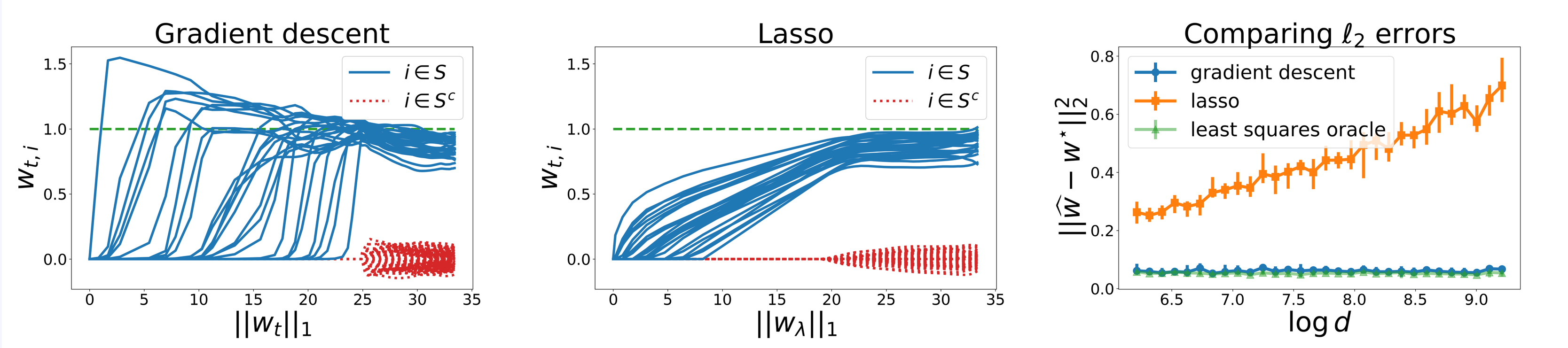
With the intuition above, as soon as $w^\star_{\min} - \|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\|_\infty > \|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\|_\infty$ all coordinates on the true support S grow exponentially at a faster rate than the coordinates on S^c.



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

Theorem 2 – Dimension Free Bounds

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



Theorem 3 – Computational Optimality

- The coordinates i such that $|w^\star_i| \gtrsim w^\star_{\max}$ converge in $O(\log\alpha^{-1})$ iterations after which the learning rate **remains unnecessarily small**. We can instead use different learning rates for different coordinates.
- We can compute $\hat{z}$ such that $w^\star_{\max} \leq \hat{z} \leq 2w^\star_{\max}$ in $O(nd)$ time. For $m = 2, 3, \dots$, after every $\boldsymbol{t = m\Omega(\log\alpha^{-1})}$ iterations, **double the learning rate** for all i such that $\left|w^\star_{t,i}\right| \leq 2^{-m-1}\hat{z}$.
- The resulting algorithm achieves the bounds of Theorems 1 and 2 in $\boldsymbol{\widetilde{O}(1)}$ iterations. Hence the total complexity of our algorithm is $\widetilde{O}(nd)$.

References

[1] 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.