Algorithmic Foundations of Learning

Lecture 14
Least Squares Regression.
Implicit Bias and Implicit Regularization

Patrick Rebeschini

Department of Statistics
University of Oxford

Explicit and Implicit Regularization

Empirical Risk: $R(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, w\rangle - Y_i)^2 = \frac{1}{n}\|\mathbf{x}w - Y\|_2^2.$

Explicit Regularization for sparse recovery:

1. Lasso estimator $W^{p1} = \operatorname{argmin}_{w\in\mathbb{R}^d} R(w) + 2\lambda\|w\|_1$
2. Tune regularization parameter: $\lambda = \|\nabla R(w^\star)\|_\infty = \sigma\frac{\|\mathbf{x}^\top\xi\|_\infty}{n}$
3. Run a gradient descent method (e.g. ISTA) $(W_t)_{t\geq 0}$ to approximate W^{p1}

$$\boxed{\|W_t - w^\star\|_2 \leq \underbrace{\|W_t - W^{p1}\|_2}_{\text{optimization error}} + \underbrace{\|W^{p1} - w^\star\|_2}_{\text{statistics error}}}$$

Implicit Regularization for least square regression:

1. Gradient descent $(W_t)_{t\geq 0}$ designed to find a minimizer of R
2. Tune parameters: $W_0^\star$, $\eta^\star$, and $t^\star$ to minimize

$$\boxed{\|W_{t^\star} - w^\star\|_2}$$

Quite surprising we can do this!

Empirical risk minimization: type of regularizations

ERM paradigm:

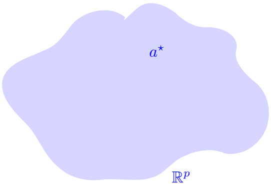
- Consider the *empirical risk* $R(a) = \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i, a), Y_i)$
- Compute $A^\star \in \text{argmin } R(a)$?

As $n < \infty$, we need to **regularize**. Depending on the problem (i.e. on $\mathcal{P}, \ell, f$):

Explicit regularization

Choose class $\mathcal{A}$
Compute $A^\star_{\mathcal{A}} \in \arg\min_{a \in \mathcal{A}} R(a)$

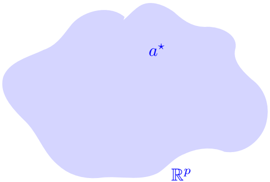
Statistics / Computation



Implicit regularization

Choose and tune algorithm **aimed at** computing $A^\star \in \arg\min_{a \in \mathbb{R}^p} R(a)$

Statistics $+$ Computation



Setup

- Assumption: the unknown parameter lies in the span of the data, i.e.

$$w^\star = \mathbf{x}^\top \omega = \sum_{i=1}^{n} \omega_i x_i$$

- Empirical (or sample) covariance matrix:

$$\mathbf{c} := \frac{\mathbf{x}^\top \mathbf{x}}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \in \mathbb{R}^{d \times d}$$

- $\mathbf{c}$ is symmetric positive semi-definite, then

$$\mathbf{c} = \mathbf{u} \boldsymbol{\mu} \mathbf{u}^\top$$

$\mathbf{u}^\top = \mathbf{u}^{-1}$ and $\boldsymbol{\mu} := \operatorname{diag}(\mu_1, \dots, \mu_r, \underbrace{0, \dots, 0}_{d-r})$ $0 < \mu_r \leq \dots \leq \mu_1$

- $r \leq d$ is the rank of the matrix
- Pseudoinverse $\mathbf{c}^+ = \mathbf{u} \boldsymbol{\mu}^+ \mathbf{u}^\top$ with $\boldsymbol{\mu}^+ := \operatorname{diag}\left(\frac{1}{\mu_1}, \dots, \frac{1}{\mu_r}, \underbrace{0, \dots, 0}_{d-r}\right)$

Least Square Regression: with and without Regularization

- **Unregularized problem** $\min\{R(w)\}$**:**

$$\nabla R(w) = \frac{2}{n}\mathbf{x}^\top(\mathbf{x}w - Y) = 0 \qquad \longrightarrow \qquad \mathbf{c}W^\star = \frac{\mathbf{x}^\top Y}{n}$$

- **If $\mathbf{c}$ is invertible**, the unique solution given by

$$\boxed{W^\star = \mathbf{c}^{-1}\frac{\mathbf{x}^\top Y}{n} = w^\star + \sigma\mathbf{c}^{-1}\frac{\mathbf{x}^\top \xi}{n}}$$

- **If $\mathbf{c}$ is not invertible**, infinitely many solutions. Least squares solution:

$$\boxed{W^\star_{\text{l.s.}} = \mathbf{c}^{+}\frac{\mathbf{x}^\top Y}{n} = \operatorname{argmin}\left\{\|w\|_2 : w \in \operatorname*{argmin}_{w\in\mathbb{R}^d} R(w)\right\} = \boldsymbol{\pi} w^\star + \sigma\mathbf{c}^{+}\frac{\mathbf{x}^\top \xi}{n}}$$

$$\mathbf{c}^{+}\frac{\mathbf{x}^\top \mathbf{x}}{n} = \mathbf{c}^{+}\mathbf{c} = \mathbf{u}\boldsymbol{\mu}^{+}\boldsymbol{\mu}\mathbf{u}^\top = \mathbf{u}\,\mathrm{diag}(1,\ldots,1,\underbrace{0,\ldots,0}_{d-r})\mathbf{u}^\top = \mathbf{u}_{1:r}\mathbf{u}_{1:r}^\top = \boldsymbol{\pi}$$

$\boldsymbol{\pi}$ is the orthogonal projection operator onto the range of $\mathbf{c}$

- **Ridge regression** $\min\{R(w) + \lambda\|w\|_2^2\}$**:** $\boxed{W^\star_{ridge} = (\mathbf{c} + \lambda I)^{-1}\frac{\mathbf{x}^\top Y}{n}}$

Gradient Descent

- Gradient Descent:

$$\boxed{W_{t+1} = W_t - \frac{\eta}{2}\nabla R(W_t) = (I - \eta\mathbf{c})\,W_t + \eta\frac{\mathbf{x}^\top Y}{n}}$$

- If $W_0 = 0$:

$$W_t = \left(\sum_{k=0}^{t-1}(I-\eta\mathbf{c})^k\right)\eta\frac{\mathbf{x}^\top Y}{n} = \underbrace{\mathrm{Inv}_t(\eta\mathbf{c})\eta\mathbf{c}w^\star}_{\mathbf{E}W_t} + \underbrace{\sigma\mathrm{Inv}_t(\eta\mathbf{c})\eta\frac{\mathbf{x}^\top\xi}{n}}_{W_t - \mathbf{E}W_t}$$

To run GD no need to compute $\mathbf{c}$, which costs $O(d^2)$

Gradient Descent (Proposition 14.2)

$$\boxed{W_t = \underbrace{\sum_{i=1}^{r}(1-(1-\eta\mu_i)^t)u_iu_i^\top w^\star}_{\mathbf{E}W_t} + \underbrace{\sigma\sum_{i=1}^{r}\frac{1-(1-\eta\mu_i)^t}{\mu_i}u_iu_i^\top\frac{\mathbf{x}^\top\xi}{n}}_{W_t-\mathbf{E}W_t}}$$

Proof of Proposition 14.2

- As $\mathbf{u}\mathbf{u}^\top = \mathbf{u}^\top\mathbf{u} = I$, $\mathrm{Inv}_t(\eta\mathbf{c}) = \sum_{k=0}^{t-1}(\mathbf{u}(I-\eta\boldsymbol{\mu})\mathbf{u}^\top)^k = \mathbf{u}\sum_{k=0}^{t-1}(I-\eta\boldsymbol{\mu})^k\mathbf{u}^\top$.
- Using that $\sum_{k=0}^{t-1} x^k = \frac{1-x^t}{1-x}$ for any $x \in \mathbb{R}\setminus\{1\}$ and $\sum_{k=0}^{t-1} 1 = t$, we obtain

$$\begin{aligned}
\mathrm{Inv}_t(\eta\mathbf{c}) &= \mathbf{u}\,\mathrm{diag}\left(\frac{1-(1-\eta\mu_1)^t}{\eta\mu_1},\dots,\frac{1-(1-\eta\mu_r)^t}{\eta\mu_r},t,\dots,t\right)\mathbf{u}^\top \\
&= \mathbf{u}\,\mathrm{diag}\left(\frac{1-(1-\eta\mu_1)^t}{\eta\mu_1},\dots,\frac{1-(1-\eta\mu_r)^t}{\eta\mu_r},0,\dots,0\right)\mathbf{u}^\top + \mathbf{u}\,\mathrm{diag}(0,\dots,0,t,\dots,t)\mathbf{u}^\top \\
&= \mathbf{u}_{1:r}\,\mathrm{diag}\left(\frac{1-(1-\eta\mu_1)^t}{\eta\mu_1},\dots,\frac{1-(1-\eta\mu_r)^t}{\eta\mu_r}\right)\mathbf{u}_{1:r}^\top + t\mathbf{u}_{r+1:d}\mathbf{u}_{r+1:d}^\top \\
&= \mathbf{u}_{1:r}\,\mathrm{diag}\left(1-(1-\eta\mu_1)^t,\dots,1-(1-\eta\mu_r)^t\right)\mathbf{u}_{1:r}^\top\mathbf{u}_{1:r}\,\mathrm{diag}\left(\frac{1}{\eta\mu_1},\dots,\frac{1}{\eta\mu_r}\right)\mathbf{u}_{1:r}^\top + t(I-\boldsymbol{\pi}) \\
&= \mathbf{u}(I-(I-\eta\boldsymbol{\mu})^t)\mathbf{u}^\top(\eta\mathbf{c})^+ + t(I-\boldsymbol{\pi}) \\
&= (I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)(\eta\mathbf{c})^+ + t(I-\boldsymbol{\pi}).
\end{aligned}$$

- By the properties of the pseudoinverse, we have $(I-\boldsymbol{\pi})\mathbf{x}^\top = 0$. If fact, for a generic matrix $\mathbf{m}$ it can be shown that $(\mathbf{m}^\top\mathbf{m})^+\mathbf{m}^\top = \mathbf{m}^+, \mathbf{m}^+\mathbf{m}\mathbf{m}^\top = \mathbf{m}^\top$. As $\boldsymbol{\pi} = \boldsymbol{c}^+\boldsymbol{c}$ by (14.2) and $\boldsymbol{c} = \mathbf{x}^\top\mathbf{x}/n$ by definition, by two properties above:

$$(I-\boldsymbol{\pi})\mathbf{x}^\top = (I-(\mathbf{x}^\top\mathbf{x})^+\mathbf{x}^\top\mathbf{x})\mathbf{x}^\top = (I-\mathbf{x}^+\mathbf{x})\mathbf{x}^\top = \mathbf{x}^\top - \mathbf{x}^+\mathbf{x}\mathbf{x}^\top = \mathbf{x}^\top - \mathbf{x}^\top = 0.$$

- So, using that $\mathbf{c} = \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top$ we find $\mathrm{Inv}_t(\eta\mathbf{c})\eta\mathbf{c} = (I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)$, and

$$W_t - \mathbf{E}W_t = \sigma\,\mathrm{Inv}_t(\eta\mathbf{c})\eta\frac{\mathbf{x}^\top\xi}{n} = \sigma(I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)\mathbf{c}^+\frac{\mathbf{x}^\top\xi}{n}.$$

Implicit Bias

Implicit Bias (Proposition 14.3)

$$\lim_{t\to\infty} W_t = \underbrace{\boldsymbol{\pi} w^\star}_{\lim_{t\to\infty} \mathbf{E} W_t} + \underbrace{\sigma \mathbf{c}^+ \frac{\mathbf{x}^\top \xi}{n}}_{\lim_{t\to\infty} (W_t - \mathbf{E} W_t)} = W^\star_{\text{l.s.}}$$

with rate given by

$$\|W_t - W^\star_{\text{l.s.}}\|_2 \leq (1-\eta\mu_r)^t \|w^\star\|_2 + \frac{\sigma}{\sqrt{n}} \frac{(1-\eta\mu_1)^t}{\mu_r} \left\| \frac{\mathbf{x}^\top \xi}{\sqrt{n}} \right\|_2$$

Where does implicit bias come from?

$$x_{s+1} = \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ f(x_s) + \nabla f(x_s)^\top (y - x_s) + \frac{1}{2\eta_s} \|y - x_s\|_2^2 \right\}$$

Implicit Regularization

Implicit Regularization (Theorem 14.5)

$$\|W_t - w^\star\|_2 \le \underbrace{\|\mathbf{E}W_t - \boldsymbol{\pi} w^\star\|_2}_{\text{bias error}} + \underbrace{\|W_t - \mathbf{E}W_t\|_2}_{\text{concentration error}} + \underbrace{\|w^\star - \boldsymbol{\pi} w^\star\|_2}_{\text{approximation error}}$$

Let $\eta^\star \le \frac{1}{\mu_1}$, $t^\star \ge \frac{1}{\log(1/(1-\eta\mu_r))} \log\left(\frac{\|w^\star\|_2}{\sigma} \frac{\sqrt{n}}{\tilde{c}}\right)$ for a given $c \in (0,1)$. Then,

$$\mathbf{P}\left(\|W_{t^\star} - w^\star\|_2 \le 2\sigma \frac{\tilde{c}}{\sqrt{n}} + \|w^\star - \boldsymbol{\pi} w^\star\|_2\right) \ge 1 - \delta$$

with $\tilde{c} = \frac{1}{\mu_r}\sqrt{\sum_{i=1}^r \mu_i + c\sum_{i=1}^r \frac{\mu_i^2}{\mu_1}}$ and $\delta = e^{-\frac{c^2}{8}\sum_{i=1}^r (\mu_i/\mu_1)^2}$

GD solves the problem optimally (stats and computation) if:

- Eigenvalues $\{\mu_1, \ldots, \mu_r\}$ are upper and lower bounded by univ. constants
- Signal-to-noise ratio $\frac{\|w^\star\|_2}{\sigma}$ is upper bounded by a universal constant

Proof of Theorem 14.5 (Part I)

- Bias term: from Proposition 14.2, using that $\boldsymbol{\pi} = \sum_{i=1}^r u_i u_i^\top$, we have

$$
\begin{aligned}
\|\mathbf{E}W_t - \boldsymbol{\pi} w^\star\|_2 &= \left\| \sum_{i=1}^r (1 - (1 - \eta\mu_i)^t) u_i u_i^\top w^\star - \sum_{i=1}^r u_i u_i^\top w^\star \right\|_2 \\
&= \left\| - \sum_{i=1}^r (1 - \eta\mu_i)^t u_i u_i^\top w^\star \right\|_2 \\
&\leq \left\| - \sum_{i=1}^r (1 - \eta\mu_i)^t u_i u_i^\top \right\| \|w^\star\|_2 \leq (1 - \eta\mu_r)^t \|w^\star\|_2
\end{aligned}
$$

- Concentration term:

$$
\begin{aligned}
\|W_t - \mathbf{E}W_t\|_2 &= \left\| \sigma \sum_{i=1}^r \frac{1 - (1 - \eta\mu_i)^t}{\mu_i} u_i u_i^\top \frac{\mathbf{x}^\top \xi}{n} \right\|_2 \\
&\leq \sigma \left\| \sum_{i=1}^r \frac{1 - (1 - \eta\mu_i)^t}{\mu_i} u_i u_i^\top \right\| \frac{\|\mathbf{x}^\top \xi\|_2}{n} \\
&\leq \frac{\sigma}{\sqrt{n}} \frac{1 - (1 - \eta\mu_1)^t}{\mu_r} \frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}.
\end{aligned}
$$

Proof of Theorem 14.5 (Part II)

- The random vector $V := \frac{\mathbf{x}^\top \xi}{\sqrt{n}}$ is Gaussian with mean 0 and covariance matrix $\mathbf{c}$

- We will now show that $\|V\|_2^2 = (\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}})^2$ has the same distribution as $\sum_{i=1}^r \mu_i Z_i^2$, where $Z_1, \ldots, Z_r$ are i.i.d. standard Gaussian random variables.

- Let $\mathbf{c}^{1/2} = \mathbf{u}\boldsymbol{\mu}^{1/2}\mathbf{u}^\top$ be the square root of the matrix $\mathbf{c}$, with $\boldsymbol{\mu}^{1/2} = \mathrm{diag}(\sqrt{\mu_1}, \ldots, \sqrt{\mu_r}, 0, \ldots, 0)$. Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be a Gaussian random vector with mean 0 and covariance I. Then, the random vector V has the same distribution as the random vector $T = \mathbf{c}^{1/2}\mathbf{u}Z$. In fact, T is Gaussian being a linear combination of a Gaussian vector and its variance is given by

$$\mathbf{E}TT^\top = \mathbf{E}[\mathbf{c}^{1/2}\mathbf{u}ZZ^\top\mathbf{u}^\top\mathbf{c}^{1/2}] = \mathbf{c}^{1/2}\mathbf{u}\mathbf{E}[ZZ^\top]\mathbf{u}^\top\mathbf{c}^{1/2} = \mathbf{c}^{1/2}\mathbf{u}\mathbf{u}^\top\mathbf{c}^{1/2} = \mathbf{c}.$$

- Then, as $\mathbf{c} = \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top$, we find

$$\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}\right)^2 = \|V\|_2^2 = V^\top V \sim T^\top T = Z^\top \mathbf{u}^\top \mathbf{c}\mathbf{u}Z$$

$$= Z^\top \mathbf{u}^\top \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top \mathbf{u}Z = Z^\top \boldsymbol{\mu} Z = \sum_{i=1}^r \mu_i Z_i^2$$

Proof of Theorem 14.5 (Part III)

- In particular, $\mathbf{E}\left[\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}\right)^2\right] = \mathbf{E}[\|V\|_2^2] = \sum_{i=1}^r \mu_i \mathbf{E}[Z_i^2] = \sum_{i=1}^r \mu_i$.
- From **Problem 3.3** in the Problem Sheets, recall that each Z_i^2 is sub-exponential with parameters $\nu^2 = 4$ and $c = 4$, namely:

$$\mathbf{E}e^{t(Z_i^2-1)} \leq e^{\nu^2 t^2/2} \qquad \text{for any } t \in (-1/c, 1/c).$$

- By Chernoff's bound we have, for any $\varepsilon, t > 0$,

$$\mathbf{P}(\|V\|_2^2 - \mathbf{E}[\|V\|_2^2] \geq \varepsilon) \leq e^{-t\varepsilon}\mathbf{E}e^{t(\|V\|_2^2 - \mathbf{E}[\|V\|_2^2])} = e^{-t\varepsilon}\mathbf{E}e^{t\sum_{i=1}^r \mu_i (Z_i^2-1)}$$
$$= e^{-t\varepsilon}\prod_{i=1}^r \mathbf{E}e^{t\mu_i(Z_i^2-1)}.$$

If $t\mu_1 < 1/4$, then the previous result yields

$$\mathbf{P}(\|V\|_2^2 - \mathbf{E}[\|V\|_2^2] \geq \varepsilon) \leq e^{-t\varepsilon}\prod_{i=1}^r e^{2t^2\mu_i^2} = e^{-t\varepsilon + 2t^2\sum_{i=1}^r \mu_i^2}.$$

The smallest upper bound is obtained by choosing $t = \frac{\varepsilon}{4\sum_{i=1}^r \mu_i^2}$ and yields

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}} \geq \sqrt{\sum_{i=1}^r \mu_i + \varepsilon}\right) = \mathbf{P}\left(\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}\right)^2 - \sum_{i=1}^r \mu_i \geq \varepsilon\right) \leq e^{-\varepsilon^2/(8\sum_{i=1}^r \mu_i^2)}.$$

Proof of Theorem 14.5 (Part IV)

- Choosing $\varepsilon = c\sum_{i=1}^{r}\mu_i^2/\mu_1$, where c is any positive constant strictly less than 1,

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top\xi\|_2}{\sqrt{n}} < \sqrt{\sum_{i=1}^{r}\mu_i + c\sum_{i=1}^{r}\frac{\mu_i^2}{\mu_1}}\right) \geq 1 - e^{-\frac{c^2}{8}\sum_{i=1}^{r}(\mu_i/\mu_1)^2}.$$

- Hence, so far we proved that for any $c \in (0,1)$ we have

$$\mathbf{P}\Big(\|W_t - w^\star\|_2 \leq (1-\eta\mu_r)^t\|w^\star\|_2 + \frac{\sigma}{\sqrt{n}}\tilde{c} + \|w^\star - \boldsymbol{\pi}w^\star\|_2\Big) \geq 1-\delta,$$

with $\tilde{c} = \frac{1}{\mu_r}\sqrt{\sum_{i=1}^{r}\mu_i + c\sum_{i=1}^{r}\frac{\mu_i^2}{\mu_1}}$ and $\delta = e^{-\frac{c^2}{8}\sum_{i=1}^{r}(\mu_i/\mu_1)^2}$.

- Choosing $t^\star$ such that $(1-\eta\mu_r)^{t^\star}\|w^\star\|_2 = \frac{\sigma}{\sqrt{n}}\tilde{c}$ yields the final result.