(DRAFT 0.1) ℓ_1 and ℓ_2 Regularization

David Rosenberg

New York University

February 4, 2015

Hypothesis Spaces

- We've spoken vaguely about "bigger" and "smaller" hypothesis spaces
- In practice, convenient to work with a **nested sequence** of spaces:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_n \cdots \subset \mathcal{F}$$

Decision Trees

- $\mathcal{F} = \{\text{all decision trees}\}$
- $\mathcal{F}_n = \{\text{all decision trees of depth } \leqslant n\}$

Complexity Measures for Decision Functions

- Number of variables / features
- Depth of a decision tree
- Degree of a polynomial
- A measure of smoothness:

$$f \mapsto \int \left\{ f''(t) \right\}^2 dt$$

- How about for linear models?
    - ℓ_0 complexity: number of non-zero coefficients
    - ℓ_1 "lasso" complexity: $\sum_{i=1}^{d} |w_i|$, for coefficients $w_1, \ldots, w_d$
    - ℓ_2 "ridge" complexity: $\sum_{i=1}^{d} w_i^2$ for coefficients $w_1, \ldots, w_d$

Nested Hypothesis Spaces from Complexity Measure

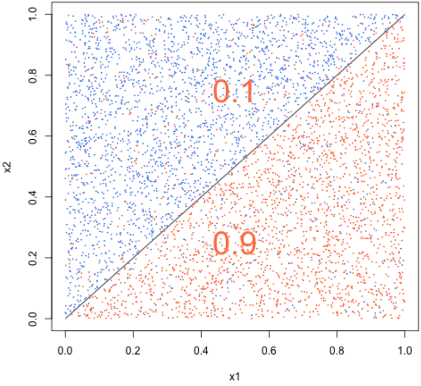
- Hypothesis space: $\mathcal{F}$
- Complexity measure $\Omega : \mathcal{F} \to \mathbf{R}^{\geqslant 0}$
- Consider all functions in $\mathcal{F}$ *with maximum complexity r:*

$$\mathcal{F}_r = \{f \in \mathcal{F} \mid \Omega(f) \leqslant r\}$$

  - If Ω is a norm on $\mathcal{F}$, this is a **ball of radius** r in $\mathcal{F}$.
- Increasing complexities: $r = 0, 1, 2, 5, \ldots$ gives nested spaces:

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_5 \subset \cdots \subset \mathcal{F}$$

Excess Risk Decomposition, Nested Space, and Trees

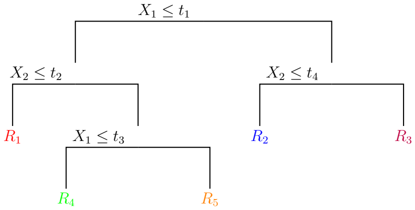


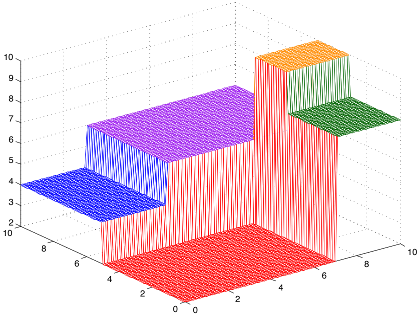
$$\mathcal{Y} = \{\text{blue}, \text{orange}\}$$
$$P_{\mathcal{X}} = \text{Uniform}([0,1]^2)$$
$$\mathbb{P}(\text{orange} \mid x_1 > x_2) = .9$$
$$\mathbb{P}(\text{orange} \mid x_1 < x_2) = .1$$

Bayes Error Rate $= 0.1$

Regression Trees

- Partition space on one variable at a time

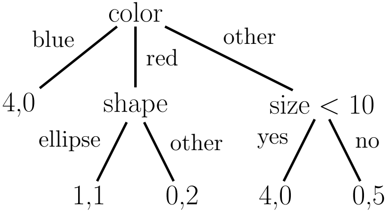




KPM Figure 16.1

Classification Trees

- Classification Tree
- $4,0$ in the leaf node means 4 successes, 0 failures



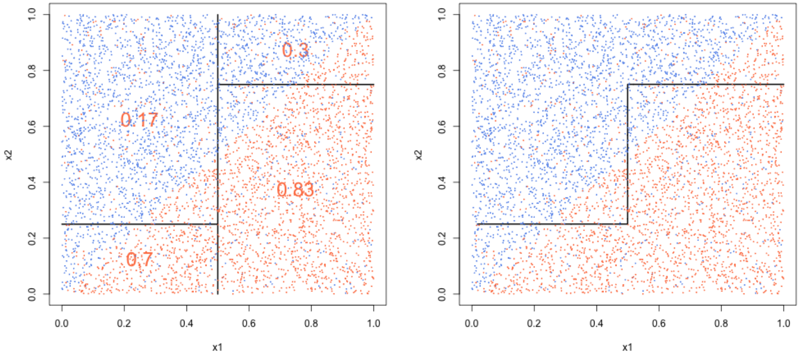
- Depth of the tree is one measure of complexity

KPM Figure 16.2

Hypothesis Space: Decision Tree

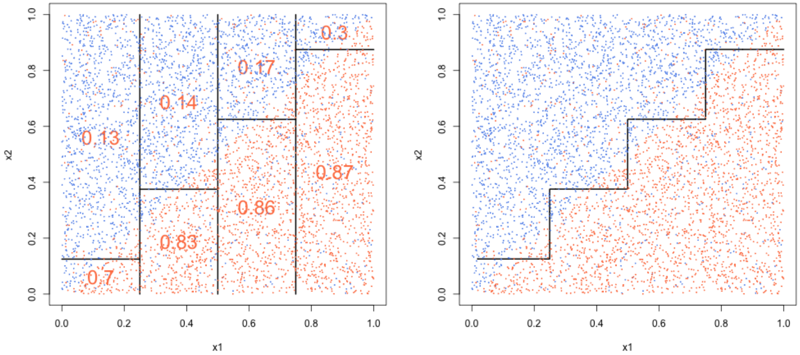
- $\mathcal{F} = \left\{ \text{all decision tree classifiers on } [0,1]^2 \right\}$
- $\mathcal{F}_d = \left\{ \text{all decision tree classifiers on } [0,1]^2 \text{ with DEPTH} \leqslant d \right\}$
- We'll consider
$$\mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_4 \cdots \subset \mathcal{F}_{15}$$
- Bayes error rate $= 0.1$

Theoretical Best in $\mathcal{F}_2$



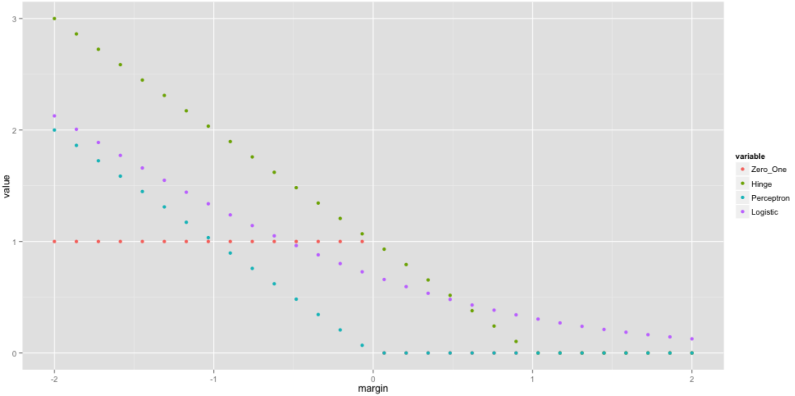
- Risk Minimizer (e.g. assuming **infinite training data**)
- Risk = P(error) = 0.2
- Approximation Error = 0.2 - 0.1 = 0.1

Theoretical Best in $\mathcal{F}_3$



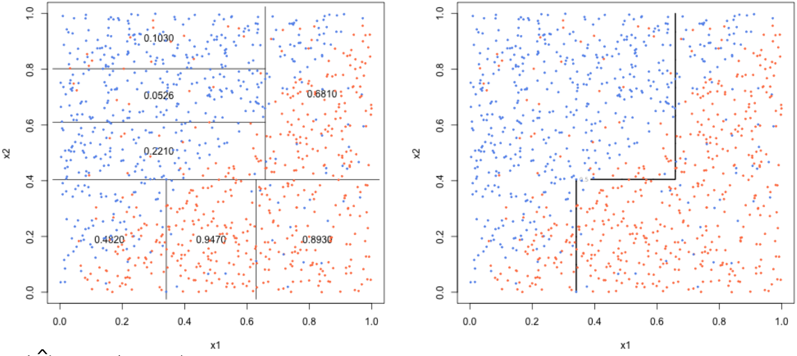
- Risk Minimizer (e.g. assuming **infinite training data**)
- Risk = P(error) = 0.15
- Approximation Error = 0.15 - 0.1 = 0.05

Theoretical Best in $\mathcal{F}_4$



- Risk Minimizer (e.g. assuming **infinite training data**)
- Risk = P(error) = 0.125
- Approximation Error = 0.125 - 0.1 = 0.025

Decision Tree in $\mathcal{F}_3$ Estimated From Sample ($n = 1024$)

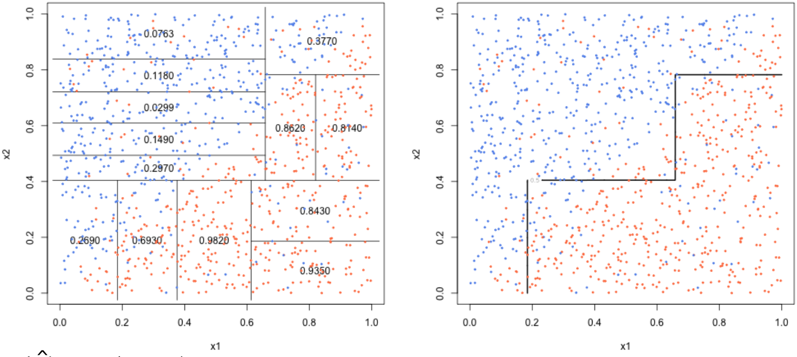


$$R(\hat{f}) = \mathbb{P}(\text{error}) = 0.176 \pm .004$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.176 \pm .004}_{R(\hat{f})} - \underbrace{0.150}_{\min_{f \in \mathcal{F}_3} R(f)}$$

$$= .026 \pm .004$$

Decision Tree in $\mathcal{F}_4$ Estimated From Sample ($n = 1024$)

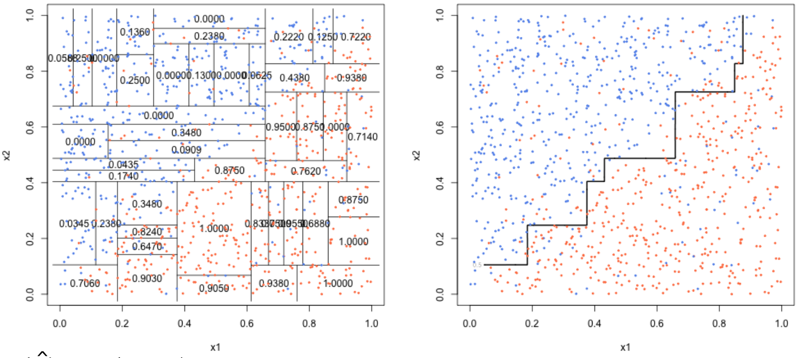


$$R(\hat{f}) = \mathbb{P}(\text{error}) = 0.144 \pm .005$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.144 \pm .005}_{R(\hat{f})} - \underbrace{0.125}_{\min_{f \in \mathcal{F}_3} R(f)}$$

$$= .019 \pm .005$$

Decision Tree in $\mathcal{F}_6$ Estimated From Sample ($n = 1024$)

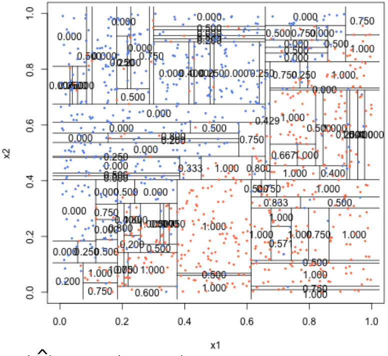


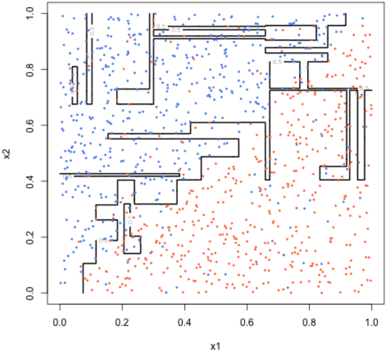
$$R(\hat{f}) = \mathbb{P}(\text{error}) = 0.148 \pm .007$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.148 \pm .007}_{R(\hat{f})} - \underbrace{0.106}_{\min_{f \in \mathcal{F}_3} R(f)}$$

$$= .042 \pm .008$$

Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 1024$)



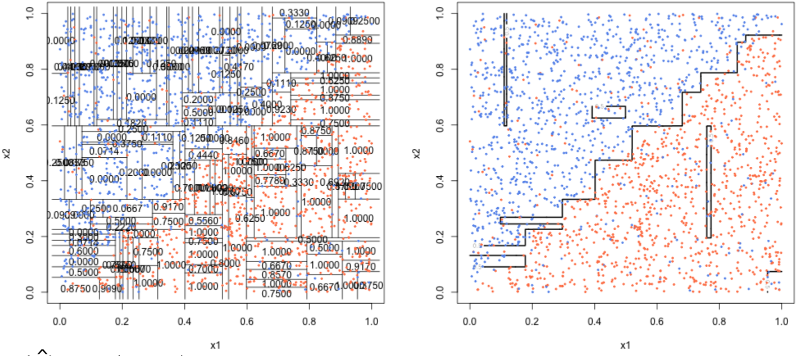


$$R(\hat{f}) = \mathbb{P}(\text{error}) = 0.162 \pm .009$$

$$\text{Estimation Error} + \text{Optimization Error} = \underbrace{0.162 \pm .009}_{R(\hat{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)}$$

$$= .061 \pm .009$$

Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 2048$)

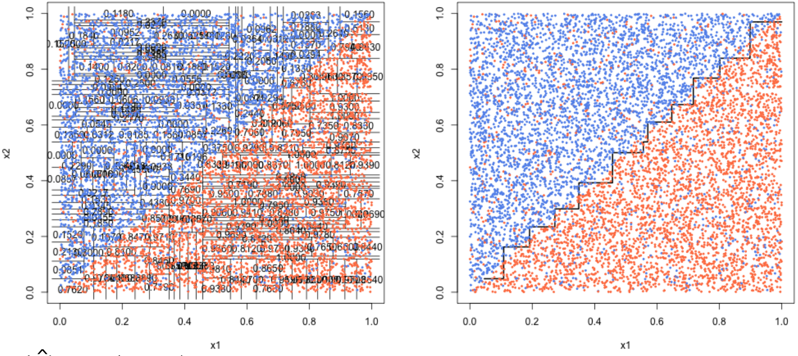


$$R(\hat{f}) = \mathbb{P}(\text{error}) = 0.146 \pm .006$$

$$\text{Estimation Error} + \text{Optimization Error} = \underbrace{0.146 \pm .006}_{R(\hat{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)}$$

$$= .045 \pm .006$$

Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 8192$)

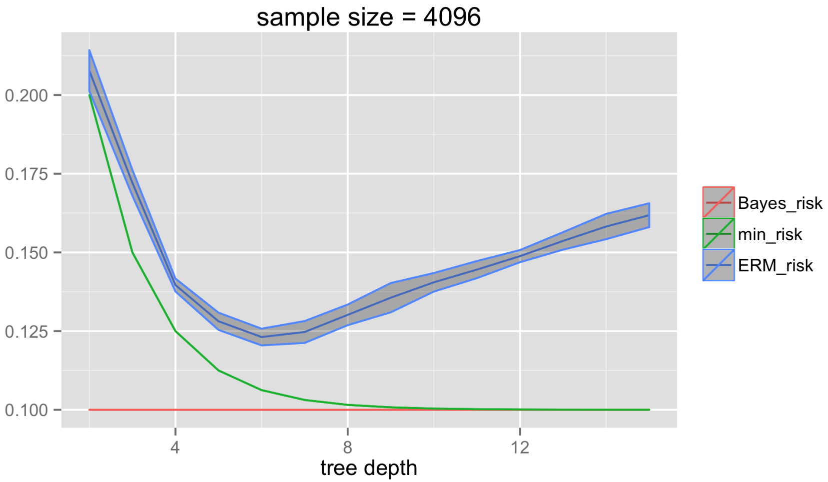


$R(\hat{f}) = \mathbb{P}(\text{error}) = 0.121 \pm .002$

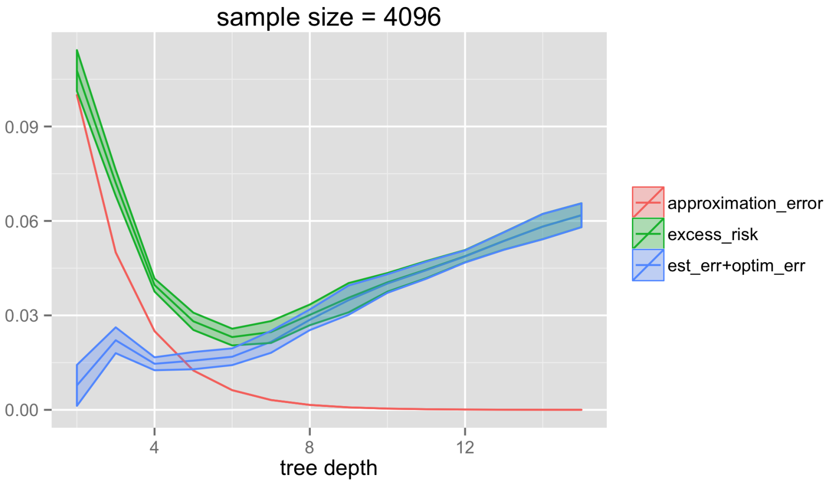
$$\text{Estimation Error+Optimization Error} = \underbrace{0.121 \pm .002}_{R(\hat{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)}$$

$$= .019 \pm .002$$

Risk Summary



Excess Risk Decomposition



Constrained Empirical Risk Minimization

Constrained ERM (Ivanov regularization)

For complexity measure $\Omega : \mathcal{F} \to \mathbf{R}^{\geqslant 0}$ and fixed $r \geqslant 0$,

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$
$$\text{s.t. } \Omega(f) \leqslant r$$

- Choose r using validation data or cross-validation.
- Each r corresponds to a different hypothesis spaces. Could also write:

$$\min_{f \in \mathcal{F}_r} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Penalized Empirical Risk Minimization

Penalized ERM (Tikhonov regularization)

For complexity measure $\Omega : \mathcal{F} \to \mathbf{R}^{\geqslant 0}$ and fixed $\lambda \geqslant 0$,

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f)$$

- Choose λ using validation data or cross-validation.

Ivanov vs Tikhonov Regularization

- Let $L : \mathcal{F} \to \mathbf{R}$ be any performance measure of f
  - e.g. $L(f)$ could be the empirical risk of f
- For many L and Ω, Ivanov and Tikhonov are "equivalent".
- What does this mean?
  - Any solution you could get from Ivanov, can also get from Tikhonov.
  - Any solution you could get from Tikhonov, can also get from Ivanov.
- In practice, both approaches are effective.
- Tikhonov often more convenient because it's an *unconstrained* minimization.

Ivanov vs Tikhonov Regularization

Ivanov and Tikhonov regularization are equivalent if:

1. For any choice of $r > 0$, the Ivanov solution

$$f_r^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} L(f) \text{ s.t. } \Omega(f) \leqslant r$$

   is also a Tikhonov solution for some $\lambda > 0$. That is, $\exists \lambda > 0$ such that

$$f_r^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} L(f) + \lambda \Omega(f).$$

2. Conversely, for any choice of $\lambda > 0$, the Tikhonov solution:

$$f_\lambda^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} L(f) + \lambda \Omega(f)$$

   is also an Ivanov solution for some $r > 0$. That is, $\exists r > 0$ such that

$$f_\lambda^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} L(f) \text{ s.t. } \Omega(f) \leqslant r$$

Linear Least Squares Regression

- Consider linear models

$$\mathcal{F} = \left\{ f : \mathbf{R}^d \to \mathbf{R} \mid f(x) = w^T x \text{ for } w \in \mathbf{R}^d \right\}$$

- Loss: $\ell(\hat{y}, y) = \frac{1}{2}(y - \hat{y})^2$
- Training data $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Linear least squares regression is ERM for ℓ over $\mathcal{F}$:

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2$$

.

- Can **overfit** when d is large compared to n.
- e.g.: $d \gg n$ very common in Natural Language Processing problems (e.g. a 1M features for 10K documents).

Ridge Regression: Workhorse of Modern Data Science

Ridge Regression (Tikhonov Form)

The ridge regression solution for regularization parameter $\lambda \geqslant 0$ is

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2^2,$$

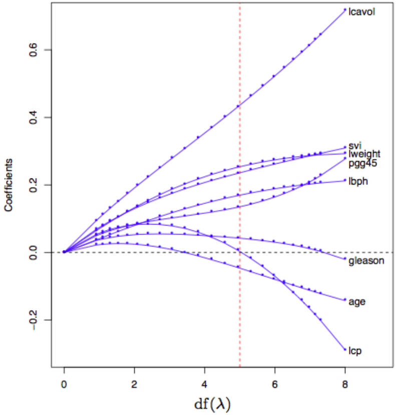
where $\|w\|_2^2 = w_1^2 + \cdots + w_d^2$ is the square of the ℓ_2-norm.

Ridge Regression (Ivanov Form)

The ridge regression solution for complexity parameter $r \geqslant 0$ is

$$\hat{w} = \underset{\|w\|_2^2 \leqslant r}{\arg\min} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2.$$

Ridge Regression: Regularization Path



$\text{df}(\lambda = \infty) = 0 \qquad \text{df}(\lambda = 0) = \text{input dimension}$

Plot from Hastie et al.'s ESL, 2nd edition, Fig. 3.8

Lasso Regression: Workhorse (2) of Modern Data Science

Lasso Regression (Tikhonov Form)

The lasso regression solution for regularization parameter $\lambda \geqslant 0$ is

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_1,$$

where $\|w\|_1 = |w_1| + \cdots + |w_d|$ is the ℓ_1-norm.

Lasso Regression (Ivanov Form)

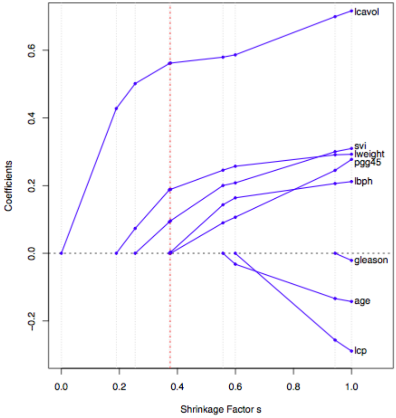
The lasso regression solution for complexity parameter $r \geqslant 0$ is

$$\hat{w} = \underset{\|w\|_1 \leqslant r}{\arg\min} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2.$$

Lasso Gives Feature Sparsity: So What?

- Time/expense to compute/buy features
- Memory to store features (e.g. real-time deployment)
- Identifies the important features
- Better prediction? sometimes
- As a feature-selection step for training a slower non-linear model

Lasso Regression: Regularization Path



Shrinkage Factor $s = r/|\hat{w}|_1$, where $\hat{w}$ is the ERM (the unpenalized fit).

Plot from Hastie et al.'s ESL, 2nd edition, Fig. 3.10

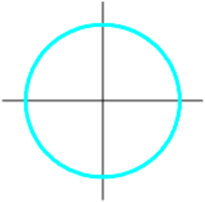
Ivanov and Tikhonov Equivalent?

- For ridge regression and lasso regression,
  - the Ivanov and Tikhonov formulations are equivalent
  - [Can prove this in a homework assignment.]
- We will use whichever form is most convenient.

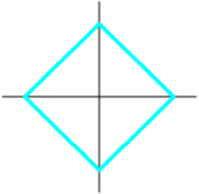
The ℓ_1 and ℓ_2 Norm Constraints

- For visualization, restrict to 2-dimensional input space
- $\mathcal{F} = \{f(x) = w_1x_1 + w_2x_2\}$ (linear hypothesis space)
- Represent $\mathcal{F}$ by $\left\{(w_1, w_2) \in \mathbf{R}^2\right\}$.

- ℓ_2 contour: $w_1^2 + w_2^2 = r^2$



- ℓ_1 contour: $|w_1| + |w_2| = r$



Where are the "sparse" solutions?

The Empirical Risk for Square Loss

- Denote the empirical risk of $f(x) = w^T x$ by

$$\hat{R}_n(w) = \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 = \|Xw - y\|^2$$

- $\hat{R}_n$ is minimized by $\hat{w} = \left(X^T X\right)^{-1} X^T y$, the OLS solution.
- What does $\hat{R}_n$ look like around $\hat{w}$?

The Empirical Risk for Square Loss

- By completing the quadratic form[1], we can show for any $w \in \mathbf{R}^d$:

$$\hat{R}_n(w) = R_{\text{ERM}} + (w - \hat{w})^T X^T X (w - \hat{w})$$

where $R_{\text{ERM}} = \hat{R}_n(\hat{w})$ is the optimal empirical risk.

- Set of w with $\hat{R}_n(w)$ exceeding R_{ERM} by $c > 0$ is

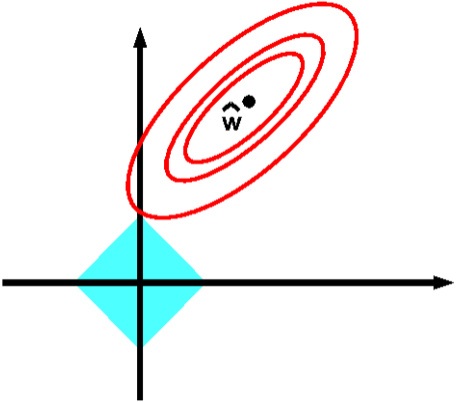
$$\left\{ w \mid \hat{R}_n(w) = c + R_{\text{ERM}} \right\} = \left\{ w \mid (w - \hat{w})^T X^T X (w - \hat{w}) = c \right\},$$

which is an **ellipsoid centered at $\hat{w}$.**

[1] Plug into this easily verifiable identity
$\theta^T M \theta + 2b^T \theta = (\theta + M^{-1}b)^T M(\theta + M^{-1}b) - b^T M^{-1} b$. This actually proves the OLS solution is optimal, without calculus.

The Famous Picture for ℓ_1 Regularization

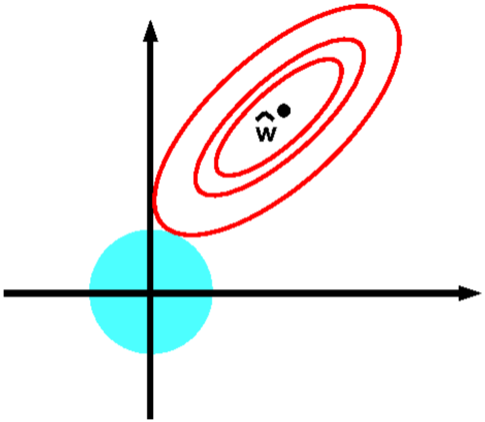
- $f_r^* = \arg\min_{w \in \mathbf{R}^2} \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$ subject to $|w_1| + |w_2| \leqslant r$



- Red lines: contours of $\hat{R}_n(w) = \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$.
- Blue region: Area satisfying complexity constraint: $|w_1| + |w_2| \leqslant r$

The Famous Picture for ℓ_2 Regularization

- $f_r^* = \arg\min_{w \in \mathbf{R}^2} \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$ subject to $w_1^2 + w_2^2 \leqslant r$



- Red lines: contours of $\hat{R}_n(w) = \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$.
- Blue region: Area satisfying complexity constraint: $w_1^2 + w_2^2 \leqslant r$

How to find the Lasso solution?

- How to solve the Lasso?

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda |w|_1$$

- $|w|_1$ is not differentiable!

Splitting a Number into Positive and Negative Parts

- Consider any number $a \in \mathbf{R}$.
- Let the **positive part** of a be

$$a^+ = a1(a \geqslant 0).$$

- Let the **negative part** of a be

$$a^- = -a1(a \leqslant 0).$$

- Note: $a^+ \geqslant 0$ and $a^- \geqslant 0$.
- So

$$a = a^+ - a^-$$

- and

$$|a| = a^+ + a^-.$$

How to find the Lasso solution?

- The Lasso problem

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda |w|_1$$

- Replace each w_i by $w_i^+ - w_i^-$.
- Write $w^+ = \left(w_1^+, \ldots, w_d^+\right)$ and $w^- = \left(w_1^-, \ldots, w_d^-\right)$.

The Lasso as a Quadratic Program

- Substituting $w = w^+ - w^-$ and $|w| = w^+ + w^-$, Lasso problem is:

$$
\begin{aligned}
\min_{w^+, w^- \in \mathbf{R}^d} & \sum_{i=1}^{n} \left(\left(w^+ - w^- \right)^T x_i - y_i \right)^2 + \lambda \left(w^+ + w^- \right) \\
\text{subject to } & w_i^+ \geqslant 0 \text{ for all } i \\
& w_i^- \geqslant 0 \text{ for all } i
\end{aligned}
$$

- Objective is differentiable (in fact, **quadratic**)
- $2d$ variables vs d variables
- $2d$ constraints vs no constraints
- (Can show this is a "**quadratic program**". Will definite this later.)

Projected SGD

$$\min_{w^+, w^- \in \mathbf{R}^d} \sum_{i=1}^{n} \left(\left(w^+ - w^- \right)^T x_i - y_i \right)^2 + \lambda \left(w^+ + w^- \right)$$

$$\begin{aligned} \text{subject to } & w_i^+ \geqslant 0 \text{ for all } i \\ & w_i^- \geqslant 0 \text{ for all } i \end{aligned}$$

- Solution:
    - Take a stochastic gradient step
    - "Project" w^+ and w^- into the constraint set
        - In other words, any component of w^+ or w^- is negative, make it 0 .
- Note: Sparsity pattern may change frequently as we iterate

Coordinate Descent Method

Coordinate Descent Method

Goal: Minimize $L(w) = L(w_1, \ldots w_d)$ over $w = (w_1, \ldots, w_d) \in \mathbf{R}^d$.

- **Initialize** $w^{(0)} = 0$
- **while** not converged:
    - Choose a coordinate $j \in \{1, \ldots, d\}$
    - $w_j^{\text{new}} \leftarrow \arg\min_{w_j} L(w_1^{(t)}, \ldots, w_{j-1}^{(t)}, \mathbf{w_j}, w_{j+1}^{(t)}, \ldots, w_d^{(t)})$
    - $w^{(t+1)} \leftarrow w^{(t)}$
    - $w_j^{(t+1)} \leftarrow w_j^{\text{new}}$
    - $t \leftarrow t+1$

- For when it's easier to minimize w.r.t. one coordinate at a time
- Random coordinate choice $\Longrightarrow$ **stochastic coordinate descent**
- Cyclic coordinate choice $\Longrightarrow$ **cyclic coordinate descent**

Coordinate Descent Method for Lasso

- Why mention coordinate descent for Lasso?
- In Lasso, the coordinate minimization has a **closed form solution**!

Coordinate Descent Method for Lasso

Closed Form Coordinate Minimization for Lasso

$$\hat{w}_j = \arg\min_{w_j \in \mathbf{R}} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda |w|_1$$

Then

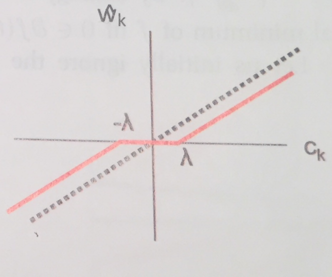
$$\hat{w}_j(c_j) = \begin{cases} (c_j + \lambda)/a_j & \text{if } c_j < -\lambda \\ 0 & \text{if } c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & \text{if } c_j > \lambda \end{cases}$$

$$a_j = 2\sum_{i=1}^{n} x_{ij}^2 A \qquad c_j = 2\sum_{i=1}^{n} x_{ij}(y_i - w_{-j}^T x_{i,-j})$$

where w_{-j} is w without component j and similarly for $x_{i,-j}$.

The Coordinate Minimizer for Lasso

$$\hat{w}_j(c_j) = \begin{cases} (c_j + \lambda)/a_j & \text{if } c_j < -\lambda \\ 0 & \text{if } c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & \text{if } c_j > \lambda \end{cases}$$



KPM Figure 13.5

Coordinate Descent Method – Variation

- Suppose there's no closed form? (e.g. logistic regression)
- Do we really need to fully solve each inner minimization problem?
- A single projected gradient step is enough for ℓ_1 regularization!
  - Shalev-Shwartz & Tewari's "Stochastic Methods..." (2011)

Stochastic Coordinate Descent for Lasso – Variation

- Let $\tilde{w} = (w^+, w^-) \in \mathbf{R}^{2d}$ and

$$L(\tilde{w}) = \sum_{i=1}^{n} \left(\left(w^+ - w^- \right)^T x_i - y_i \right)^2 + \lambda \left(w^+ + w^- \right)$$

Stochastic Coordinate Descent for Lasso - Variation

Goal: Minimize $L(\tilde{w})$ s.t. $w_i^+, w_i^- \geqslant 0$ for all i.

- **Initialize** $\tilde{w}^{(0)} = 0$
  - **while** not converged:
    - Randomly choose a coordinate $j \in \{1, \ldots, 2d\}$
    - $\tilde{w}_j \leftarrow \tilde{w}_j + \max\left\{-\tilde{w}_j, -\nabla_j L(\tilde{w})\right\}$

The $(\ell_q)^q$ Norm Constraint

- Generalize to ℓ_q norm: $(\|w\|_q)^q = |w_1|^q + |w_2|^q$.
- $\mathcal{F} = \{f(x) = w_1x_1 + w_2x_2\}$.
- Contours of $\|w\|_q^q = |w_1|^q + |w_2|^q$: