ℓ_1 and ℓ_2 Regularization

David Rosenberg

New York University

October 29, 2016

Tikhonov and Ivanov Regularization

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 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 complexity **at most** 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: $r = 0, 1.2, 2.6, 5.4, \ldots$ gives nested spaces:

$$\mathcal{F}_0 \subset \mathcal{F}_{1.2} \subset \mathcal{F}_{2.6} \subset \mathcal{F}_{5.4} \subset \cdots \subset \mathcal{F}$$

Constrained Empirical Risk Minimization

Constrained ERM (Ivanov regularization)

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

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

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

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

Penalized Empirical Risk Minimization

Penalized ERM (Tikhonov regularization)

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

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

- Choose λ using validation data or cross-validation.
- (Ridge regression formulation in Homework #1 was of this form.)

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 convenient because it's *unconstrained* minimization.

Proof of equivalence based on Lagrangian duality – a topic of Lecture 3.

Ivanov vs Tikhonov Regularization (Details)

Ivanov and Tikhonov regularization are equivalent if:

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

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

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

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

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

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

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

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

ℓ_1 and ℓ_2 Regularization

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

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

- 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

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

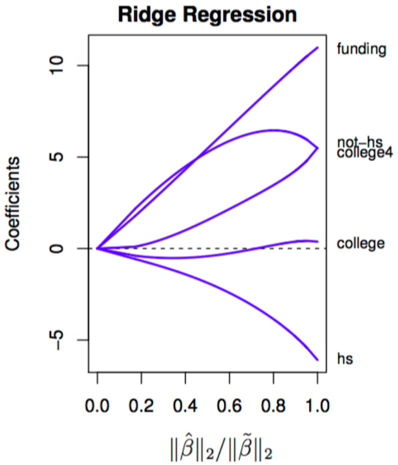
where $\|w\|_2^2 = w_1^2 + \cdots + w_d^2$ is the square of the ℓ_2-norm.

Ridge Regression (Ivanov Form)

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

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

Ridge Regression: Regularization Path



$\tilde{\beta}$ is unregularized solution; $\hat{\beta}$ is the ridge solution.

Plot from Hastie, Tibshirani, and Wainwright's Statistical Learning with Sparsity, Figure 2.1

Lasso Regression: Workhorse (2) of Modern Data Science

Lasso Regression (Tikhonov Form)

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

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

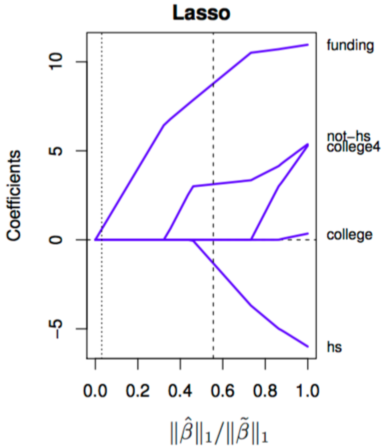
where $\|w\|_1 = |w_1| + \cdots + |w_d|$ is the ℓ_1-norm.

Lasso Regression (Ivanov Form)

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

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

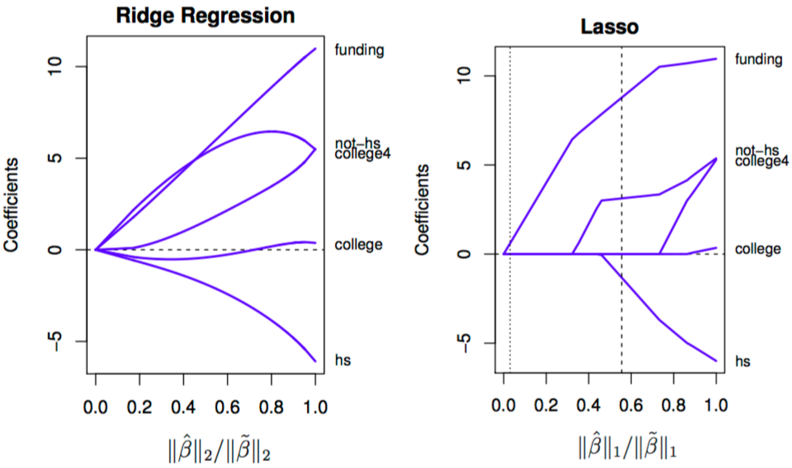
Lasso Regression: Regularization Path



$\tilde{\beta}$ is unregularized solution; $\hat{\beta}$ is the lasso solution.

Plot from Hastie, Tibshirani, and Wainwright's Statistical Learning with Sparsity, Figure 2.1

Ridge vs. Lasso: Regularization Paths



Plot from Hastie, Tibshirani, and Wainwright's Statistical Learning with Sparsity, Figure 2.1

Lasso Gives Feature Sparsity: So What?

Coefficient are 0 $\implies$ don't need those features. What's the gain?

- 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

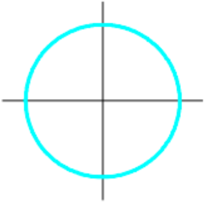
Ivanov and Tikhonov Equivalent?

- For ridge regression and lasso regression,
  - the Ivanov and Tikhonov formulations are equivalent
  - [We may prove this in homework assignment 3.]
- We will use whichever form is most convenient.

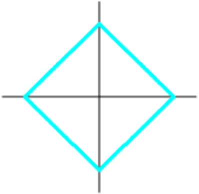
The ℓ_1 and ℓ_2 Norm Constraints

- For visualization, restrict to 2-dimensional input space
- $\mathcal{F} = \{f(x) = w_1 x_1 + w_2 x_2\}$ (linear hypothesis space)
- Represent $\mathcal{F}$ by $\left\{(w_1, w_2) \in \mathbf{R}^2\right\}$.

- ℓ_2 contour:
  $w_1^2 + w_2^2 = r$



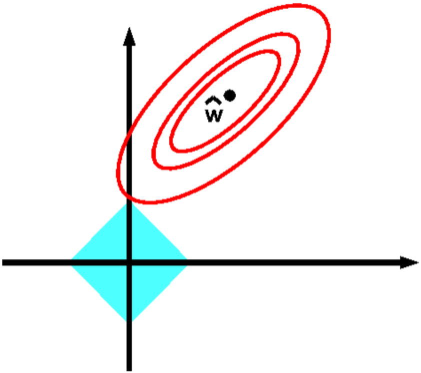
- ℓ_1 contour:
  $|w_1| + |w_2| = r$



Where are the "sparse" solutions?

The Famous Picture for ℓ_1 Regularization

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

KPM Fig. 13.3

The Empirical Risk for Square Loss

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

$$\hat{R}_n(w) = \frac{1}{n}\|Xw - y\|^2$$

- $\hat{R}_n$ 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 square", we can show for any $w \in \mathbf{R}^d$:

$$\hat{R}_n(w) = \frac{1}{n}(w - \hat{w})^T X^T X (w - \hat{w}) + \hat{R}_n(\hat{w})$$

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

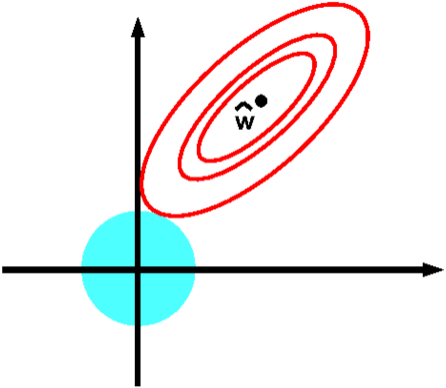
$$\left\{ w \mid \hat{R}_n(w) = c + \hat{R}_n(w) \right\} = \left\{ w \mid (w - \hat{w})^T X^T X (w - \hat{w}) = c \right\},$$

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

- We'll derive this in homework #2.

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 $w_1^2 + w_2^2 \leqslant r$

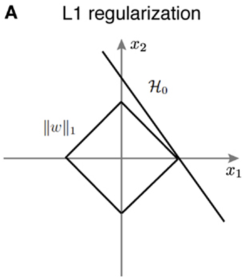


- Red lines: contours of $\hat{R}_n(w) = \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$.
- Blue region: Area satisfying complexity constraint: $w_1^2 + w_2^2 \leqslant r$

KPM Fig. 13.3

The Quora Picture

- From Quora: "Why is L1 regularization supposed to lead to sparsity than L2?"



- Doesn't seem like this figure represents the situation well...
- But maybe sometimes it does?

Figure from https://www.quora.com/Why-is-L1-regularization-supposed-to-lead-to-sparsity-than-L2.

Finding the Lasso Solution

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

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

- Do you see why $a^{+} \geqslant 0$ and $a^{-} \geqslant 0$?
- How do you write a in terms of a^{+} and a^{-}?
- How do you write $|a|$ in terms of a^{+} and 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:

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

$$\text{subject to } w_i^+ \geqslant 0 \text{ for all } i$$

$$w_i^- \geqslant 0 \text{ for all } i$$

- Objective is **differentiable** (in fact, **convex and quadratic**)
- $2d$ variables vs d variables
- $2d$ constraints vs no constraints
- A "**quadratic program**": a convex quadratic objective with linear constraints.
    - Could plug this into a generic QP solver.

Projected SGD

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

$$\text{subject to } w_i^+ \geqslant 0 \text{ for all } i$$

$$w_i^- \geqslant 0 \text{ for all } i,$$

where 1 represents a column vector of 1's in $\mathbf{R}^d$.

- 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 .

Coordinate Descent Method

- **Goal:** Minimize $L(w) = L(w_1, \ldots, w_d)$ over $w = (w_1, \ldots, w_d) \in \mathbf{R}^d$.
- In gradient descent or SGD,
  - each step potentially changes all entries of w.
- In each step of **coordinate descent**,
  - we adjust only a single w_i.
- In each step, solve

$$w_i^{\text{new}} = \underset{w_i}{\arg\min}\, L(w_1, \ldots, w_{i-1}, \mathbf{w_i}, w_{i+1}, \ldots, w_d)$$

- Solving this argmin may itself be an iterative process.
- Coordinate descent is great when
  - it's easy or easier to minimize w.r.t. one coordinate at a time

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 $j \in \{1, \ldots, d\}$
  - $w_j^{\text{new}} \leftarrow \arg\min_{w_j} L(w_1^{(t)}, \ldots, w_{j-1}^{(t)}, \mathbf{w_j}, w_{j+1}^{(t)}, \ldots, w_d^{(t)})$
  - $w^{(t+1)} \leftarrow w^{(t)}$
  - $w_j^{(t+1)} \leftarrow w_j^{\text{new}}$
  - $t \leftarrow t + 1$

- Random coordinate choice $\implies$ **stochastic coordinate descent**
- Cyclic coordinate choice $\implies$ **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

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

Then

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

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

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

Coordinate Descent: When does it work?

- Suppose we're minimizing $f : \mathbf{R}^d \to \mathbf{R}$.
- Sufficient conditions:
  1. f is continuously differentiable and
  2. f is strictly convex in each coordinate
- But lasso objective

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

  is not differentiable...
- Luckily there are weaker conditions...

Coordinate Descent: The Separability Condition

Theorem

[a] *If the objective f has the following structure*

$$f(w_1,\ldots,w_d) = g(w_1,\ldots,g_d) + \sum_{j=1}^{d} h_j(x_j),$$

where

- $g : \mathbf{R}^d \to \mathbf{R}$ *is differentiable and convex, and*
- *each* $h_j : \mathbf{R} \to \mathbf{R}$ *is convex (but not necessarily differentiable)*

then the coordinate descent algorithm converges to the global minimum.

[a] Tseng 1988: "Coordinate ascent for maximizing nondifferentiable concave functions", Technical Report LIDS-P

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 $j \in \{1, \ldots, 2d\}$
    - $\tilde{w}_j \leftarrow \tilde{w}_j + \max\left\{ -\tilde{w}_j, -\nabla_j L(\tilde{w}) \right\}$