Regularization

He He

Slides based on Lecture 2c from David Rosenberg's course material.

CDS, NYU

Feb 16, 2021

ℓ_2 and ℓ_1 Regularization

Complexity Penalty

Goal: balance between complexity of the hypothesis space $\mathcal{F}$ and the training loss

Complexity measure: $\Omega : \mathcal{F} \to [0, \infty)$, e.g. number of features

Penalized ERM (Tikhonov regularization)

For complexity measure $\Omega : \mathcal{F} \to [0, \infty)$ and fixed $\lambda \geqslant 0$,

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

As usual, find λ using validation data.

Number of features as complexity measure is hard to optimize—other measures?

Weight Shrinkage: Intuition

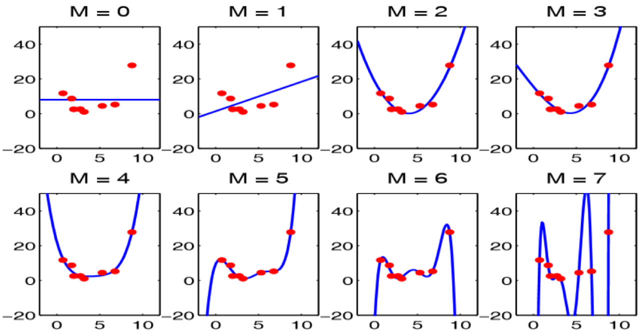
Consider linear regression on the following data, which line would you prefer? [draw]

Weight Shrinkage: Intuition

Consider linear regression on the following data, which line would you prefer? [draw]

- Prefer the line with smaller slope: small change in the input does not cause large change in the output
- If the estimated weights change by a small amount, it wouldn't cause huge change in the prediction (less sensitive to data)

Weight Shrinkage: Polynomial Regression



- Large weights are needed to "wiggle" the curve
- Want to regularize the weights to make them smaller, e.g.

$$\hat{y} = 0.001x^7 + 0.003x^3 + 1 \text{ vs } \hat{y} = 1000x^7 + 500x^3 + 1$$

(Adapated from Mark Schmidt's slide)

Linear Regression with L2 Regularization

- Consider linear models

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

- Square 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 square loss over $\mathcal{F}$:

$$\hat{w} = \operatorname*{arg\,min}_{w \in \mathrm{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2$$

- Can overfit when d is large compared to n, e.g. $d \gg n$ very common in NLP (e.g. a 1M features for 10K documents).

Linear Regression with L2 Regularization

Penalize "large" weights where size of weights is measured by ℓ_2 norm:

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

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

- Also known as **ridge regression**.
- We get back linear least square regression with $\lambda = 0$.
- ℓ_2 regularization can be used for other models too (e.g. neural networks)

How does ℓ_2 regularization induce "regularity"?

- Short answer: it controls "sensitivity" of the function.
- For $\hat{f}(x) = \hat{w}^T x$, $\hat{f}$ is **Lipschitz continuous** with Lipschitz constant $L = \|\hat{w}\|_2$.
- That is, when moving from x to $x + h$, $\hat{f}$ changes no more than $L\|h\|$.
- So ℓ_2 regularization controls the maximum rate of change of $\hat{f}$.
- Proof:

$$
\begin{aligned}
\left|\hat{f}(x+h) - \hat{f}(x)\right| &= |\hat{w}^T(x+h) - \hat{w}^T x| = \left|\hat{w}^T h\right| \\
&\leqslant \|\hat{w}\|_2 \|h\|_2 \quad \text{(Cauchy-Schwarz inequality)}
\end{aligned}
$$

- Note that other norms also provides a bound on L due to the equivalence of norms: $\exists C > 0$ s.t. $\|\hat{w}_2\|_2 \leqslant C\|\hat{w}_2\|_p$

Linear Regression vs Ridge Regression

Objective:

- Linear: $L(w) = \frac{1}{2}\|Xw - y\|_2^2$
- Ridge: $L(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$

Gradient:

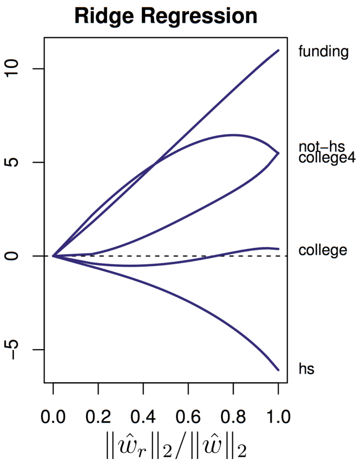
- Linear: $\nabla L(w) = X^T(Xw - y)$
- Ridge: $\nabla L(w) = X^T(Xw - y) + \lambda w$
  - Also known as **weight decay** in neural networks

Closed-form solution:

- Linear: $X^T X w = X^T y$
- Ridge: $(X^T X + \lambda I)w = X^T y$
  - $(X^T X + \lambda I)$ is always invertible

Ridge Regression: Regularization Path

Regulariztion path shows how the weights vary as we change the regularization strength



$$
\begin{aligned}
\hat{w}_r &= \underset{\|w\|_2^2 \le r^2}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left(w^T x_i - y_i\right)^2 \\
\hat{w} &= \hat{w}_\infty = \text{Unconstrained ERM}
\end{aligned}
$$

- For $r = 0$, $\|\hat{w}_r\|_2/\|\hat{w}\|_2 = 0$.
- For $r = \infty$, $\|\hat{w}_r\|_2/\|\hat{w}\|_2 = 1$

Modified from Hastie, Tibshirani, and Wainwright's *Statistical Learning with Sparsity*, Fig 2.1. About predicting crime in 50 US cities.

Lasso Regression

Penalize the ℓ_1 norm of the weights:

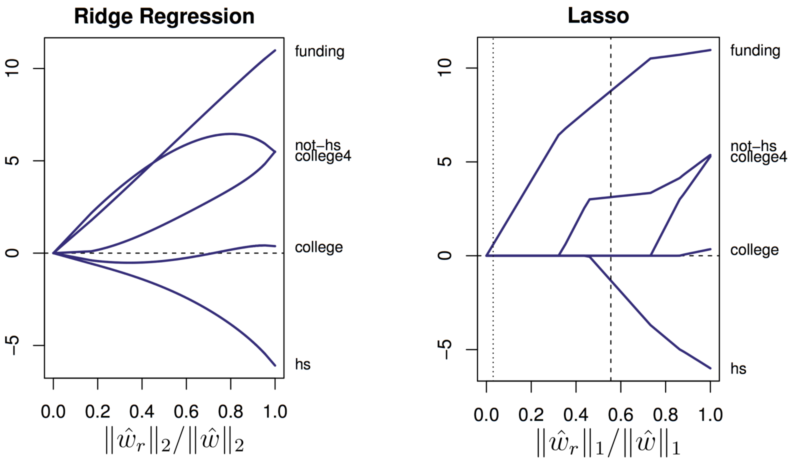
Lasso Regression (Tikhonov Form)

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

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

Ridge vs. Lasso: Regularization Paths

Lasso gives sparse weights:



Modified from Hastie, Tibshirani, and Wainwright's *Statistical Learning with Sparsity*, Fig 2.1. About predicting crime in 50 US cities.

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

Regularization and Sparsity

Constrained Empirical Risk Minimization

Constrained ERM (Ivanov regularization)

For complexity measure $\Omega : \mathcal{F} \to [0, \infty)$ and fixed $r \geqslant 0$,

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

$$\text{s.t. } \Omega(f) \leqslant r$$

Lasso Regression (Ivanov Form)

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

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

r has the same role as λ in penalized ERM (Tikhonov).

Ivanov vs Tikhonov Regularization

- Let $L : \mathcal{F} \to \mathrm{R}$ 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":
  - Any solution f^* you could get from Ivanov, can also get from Tikhonov.
  - Any solution f^* you could get from Tikhonov, can also get from Ivanov.
- Can get conditions for equivalence from Lagrangian duality theory.
- In practice, both approaches are effective.
- We will use whichever that is more convenient.

Ivanov vs Tikhonov Regularization (Details)

Ivanov and Tikhonov regularization are equivalent if:

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

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

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

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

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

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

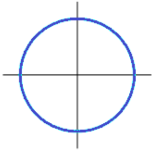
is also an Ivanov solution for some $r > 0$. That is, $\exists r > 0$ such that

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

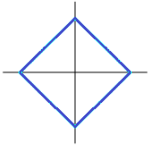
The ℓ_1 and ℓ_2 Norm Constraints

- 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 $\{(w_1, w_2) \in \mathrm{R}^2\}$.

- ℓ_2 contour:
  $w_1^2 + w_2^2 = r$



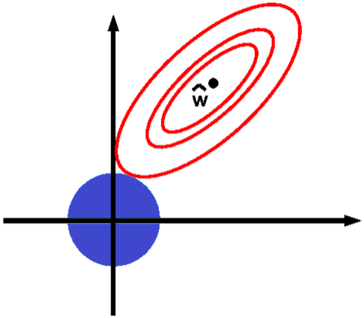
- ℓ_1 contour:
  $|w_1| + |w_2| = r$



Where are the "sparse" solutions?

The Famous Picture for ℓ_2 Regularization

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

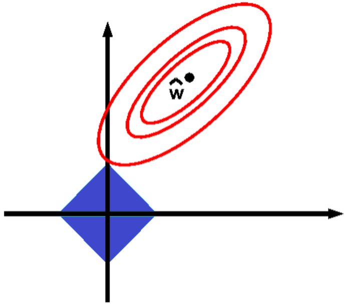


- Blue region: Area satisfying complexity constraint: $w_1^2 + w_2^2 \leqslant r$
- Red lines: contours of $\hat{R}_n(w) = \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$.

KPM Fig. 13.3

The Famous Picture for ℓ_1 Regularization

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



- Blue region: Area satisfying complexity constraint: $|w_1| + |w_2| \leqslant r$
- Red lines: contours of $\hat{R}_n(w) = \sum_{i=1}^n \left(w^T x_i - y_i\right)^2$.
- ℓ_1 solution tends to touch the corners.

KPM Fig. 13.3

Why does ℓ_1 gives sparse solution?

Geometric intuition: Euclidean projection onto a convex set encourages solutions at corners or edges.

- $\hat{w}$ in red/green regions are closest to corners in the ℓ_1 ball.

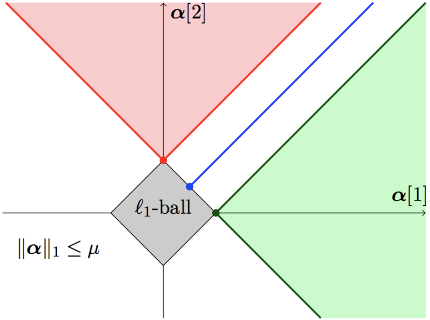


Fig from Mairal et al.'s Sparse Modeling for Image and Vision Processing Fig 1.6

Why does ℓ_1 gives sparse solution?

Geometric intuition: Euclidean projection onto a convex set encourages solutions at corners or edges.

- ℓ_2 ball encourages solution in any direction equally.

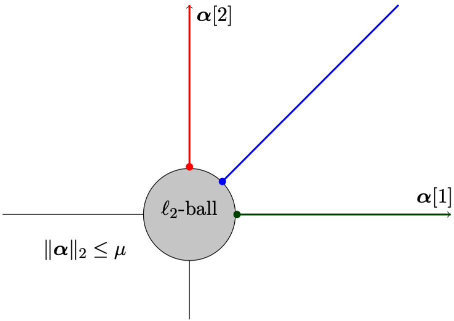


Fig from Mairal et al.'s Sparse Modeling for Image and Vision Processing Fig 1.6

Why does ℓ_1 gives sparse solution?

For ℓ_2 regularization,

- As w_i becomes smaller, there is less and less penalty
  - What is the ℓ_2 penalty for $w_i = 0.0001$?
- The gradient goes to zero as w_i moves towards zero

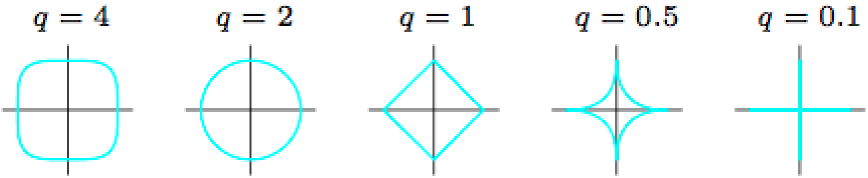
For ℓ_1 regularization,

- The function is non-smooth and the gradient stays the same as the weights becomes smaller
- Thus it pushes them to exactly zero even if the weights are already tiny

(More discussion in lecture)

The $(\ell_q)^q$ Constraint

- Generalize to ℓ_q : $(\|w\|_q)^q = |w_1|^q + |w_2|^q$.
- Contours of $\|w\|_q^q = |w_1|^q + |w_2|^q$:



- Note: $\|w\|_q$ is a norm if $q \geqslant 1$, but not for $q \in (0,1)$
- ℓ_q constraint when $q < 1$ is non-convex, so hard to optimize
- ℓ_0 ($\|w\|_0$) is defined as the number of non-zero weights, i.e. subset selection