Machine Learning (CSE 446): Regularization and Gradient Descent The "large d" regime.

Sham M Kakade

University of Washington
cse446-staff@cs.washington.edu

Least squares: What could go wrong?!

- The optimization problem:

$$\min_{\mathbf{w}} \frac{1}{N} \|Y - X\mathbf{w}\|^2$$

where Y is an N-vector and X is our $N \times d$ data matrix.

- The solution is the **least squares estimator**:

$$\mathbf{w}^{\text{least squares}} = (X^\top X)^{-1} X^\top Y$$

What if d is bigger than N? Even if not?

What could go wrong?

Suppose $d > N$:

What about $N > d$?

- What happens if features are very correlated?
  (e.g. 'rows/columns in our matrix are **co-linear**.)

A fix: Regularization

- **Regularize** the optimization problem:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|^2 =$$

$$\min_{\mathbf{w}} \frac{1}{N} \|Y - X^\top \mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2$$

- This particular case: “Ridge” Regression, Tikhonov regularization
- The solution is the **least squares estimator**:

$$\mathbf{w}^{\text{least squares}} = \left(\frac{1}{N} X^\top X + \lambda \mathbb{I} \right)^{-1} \left(\frac{1}{N} X^\top Y \right)$$

Why do we care about large d?

- Example: Suppose x is three dimensional, i.e. $x = (x[1], x[2], x[3])$. Define a new feature vector as follows:

$$\Phi(x) = \left(1, x[1], x[2], x[3], x[1]^2, x[2]^2, x[3]^2, x[1]x[2], x[1]x[3], x[2]x[3]\right).$$

  The first term is the bias term, the next three coordinates above are considered the "linear" terms, and the remaining terms are the quadratic terms.
- Now use $\Phi(x)$ instead of x in our regression problem.

Feature mappings give us more expressivity. They also "blow up" the dimensionality.

The “general” approach

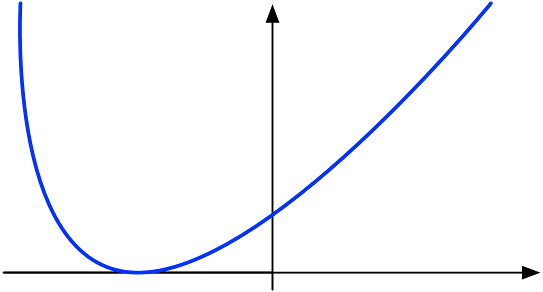
- The **regularized** optimization problem:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \mathbf{w} \cdot \mathbf{x}_i) + R(\mathbf{w})$$

- Penalty some w more than others.
  Example: $R(w) = \|w\|^2$

How do we find a solution quickly?

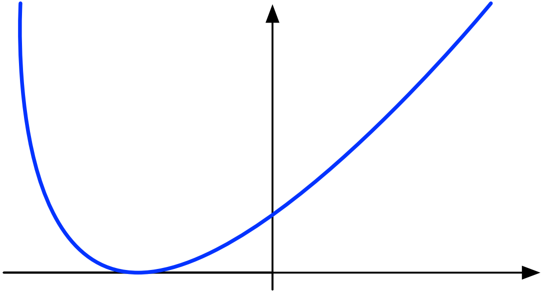
Remember: convexity



- A function $F(\cdot)$ is convex if for all $0 \leq t \leq 1$, w and w',

$$F((1-t)w+tw') \leq (1-t)F(w)+tF(w')$$

Gradient Descent



- Want to solve:

$$\min_{w} F(w)$$

- How should we update w?

Gradient Descent

Data: function $F : \mathbb{R}^d \to \mathbb{R}$, number of iterations K, step sizes $\eta^{(1)}, \ldots, \eta^{(K)}$
Result: $\mathbf{w} \in \mathbb{R}^d$
initialize: $\mathbf{w}^{(0)} = \mathbf{0}$;
for $k \in \{1, \ldots, K\}$ **do**
    $\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \eta^{(k)} \cdot \nabla F(\mathbf{w}^{(k-1)})$;
end
return $\mathbf{w}^{(K)}$;

Algorithm 1: GradientDescent

Gradient Descent: Convergence

- Letting $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} F(\mathbf{w})$ denote the global minimum
- Let $\mathbf{w}^{(k)}$ be our parameter after k updates.
- Thm: Suppose F is convex and "smooth". Using a **fixed step size** η, we have:

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \le O\left(\frac{1}{\cdot k}\right)$$

Gradient Descent: Simple example 1

- For $w \in \mathbb{R}$, $F(w) = \frac{1}{2}w^2$
- $w^* = \operatorname{argmin}_* F(w) = 0$
- $\frac{dF}{dw} = w$.
- The update:
$$w^{(k+1)} = w^{(k)} - \eta w^{(k)} = (1-\eta)w^{(k)}$$
- Always use $\eta > 0$ (for GD)
- For $\eta \geq 2$, $w^{(k)}$ does not converge.
(diverges for η strictly above 2).
- For $|\eta| < 1$, $w^{(k)}$ converges to 0 (quickly!).
- For $|\eta| = 1$, $w^{(1)} = 0$.
This convergence in one step is 'lucky', due to being in 1dim.

Gradient Descent: Simple example 2

- For $w \in \mathbb{R}^2$, $F(w) = \frac{1}{2} w^\top \mathrm{diag}(1,2) w = \frac{1}{2}\left(w_1^2 + 2w_2^2\right)$
- $\mathbf{w}^* = \mathrm{argmin}_{\mathbf{w}} F(\mathbf{w}) = 0$
- $\nabla F(w) = (w_1, 2w_2)^\top$.
- The update:
$$w^{(k+1)} = w^{(k)} - \eta \nabla F(w^{(k)})$$
- What happens here?

Gradient Descent: More formal statement

[noframenumbering]

- Letting $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} F(\mathbf{w})$ denote the global minimum
- Let $\mathbf{w}^{(k)}$ be our parameter after k updates.
- Thm: Suppose F is convex and "L-smooth". Using a **fixed step size** $\eta \leq \frac{1}{L}$, we have:

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq \frac{\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2}{\eta \cdot k}$$

- Smooth functions: for all w, w'

$$\|\nabla F(w) - \nabla F(w')\| \leq L\|w - w'\|$$

- Proof idea:
  1. If our gradient is large, we will make good progress decreasing our function value:

  2. If our gradient is small, we must have value near the optimal value: