Gradient Descent

He He[1]

CDS, NYU

Feb 9, 2021

[1]Slides based on Lecture 2b from David Rosenberg's course material.

Review: ERM

Our Setup from Statistical Learning Theory

The Spaces

- $\mathcal{X}$: input space
- $\mathcal{Y}$: outcome space
- $\mathcal{A}$: action space

Prediction Function (or "decision function")

A **prediction function** (or **decision function**) gets input $x \in \mathcal{X}$ and produces an action $a \in \mathcal{A}$:

$$\begin{aligned} f: \quad \mathcal{X} & \to \mathcal{A} \\ x & \mapsto f(x) \end{aligned}$$

Loss Function

A **loss function** evaluates an action in the context of the outcome y.

$$\begin{aligned} \ell: \quad \mathcal{A} \times \mathcal{Y} & \to \mathrm{R} \\ (a, y) & \mapsto \ell(a, y) \end{aligned}$$

Risk and the Bayes Prediction Function

Definition

The **risk** of a prediction function $f : \mathcal{X} \to \mathcal{A}$ is

$$R(f) = \mathbb{E}\ell(f(x), y).$$

In words, it's the **expected loss** of f on a new exampe (x, y) drawn randomly from $P_{\mathcal{X}\times\mathcal{Y}}$.

Definition

A **Bayes prediction function** $f^* : \mathcal{X} \to \mathcal{A}$ is a function that achieves the *minimal risk* among all possible functions:

$$f^* \in \arg\min_f R(f),$$

where the minimum is taken over all functions from $\mathcal{X}$ to $\mathcal{A}$.

- The risk of a Bayes prediction function is called the **Bayes risk**.

The Empirical Risk

Let $\mathcal{D}_n = ((x_1, y_1), \ldots, (x_n, y_n))$ be drawn i.i.d. from $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.

Definition

The **empirical risk** of $f : \mathcal{X} \to \mathcal{A}$ with respect to $\mathcal{D}_n$ is

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i).$$

- But we saw that the **unconstrained** empirical risk minimizer overfits.
  - i.e. if we minize $\hat{R}_n(f)$ over **all functions**, we overfit.

Constrained Empirical Risk Minimization

Definition

A **hypothesis space** $\mathcal{F}$ is a set of functions mapping $\mathcal{X} \to \mathcal{A}$.

- It is the collection of prediction functions we are choosing from.

- **Empirical risk minimizer** (ERM) in $\mathcal{F}$ is

$$\hat{f}_n \in \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

- From now on "ERM" always means "constrained ERM".

- So we should always specify the hypothesis space when we're doing ERM.

Example: Linear Least Squares Regression

Setup

- Input space $\mathcal{X} = \mathbf{R}^d$
- Output space $\mathcal{Y} = \mathbf{R}$
- Action space $\mathcal{Y} = \mathbf{R}$
- Loss: $\ell(\hat{y}, y) = (y - \hat{y})^2$
- **Hypothesis space:** $\mathcal{F} = \left\{ f : \mathbf{R}^d \to \mathbf{R} \mid f(x) = w^T x,\, w \in \mathbf{R}^d \right\}$

- Given data set $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,
  - Let's find the ERM $\hat{f} \in \mathcal{F}$.

Example: Linear Least Squares Regression

Objective Function: Empirical Risk

The function we want to minimize is the empirical risk:

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

where $w \in \mathbf{R}^d$ parameterizes the hypothesis space $\mathcal{F}$.

- Now, we have ended up with an optimization problem:

$$\min_{w \in \mathbf{R}^d} \hat{R}_n(w).$$

Gradient Descent

Unconstrained Optimization

Setting

Objective function $f : \mathrm{R}^d \to \mathrm{R}$ is *differentiable.*
Want to find

$$x^* = \arg\min_{x \in \mathrm{R}^d} f(x)$$

The Gradient

- Let $f : \mathbf{R}^d \to \mathbf{R}$ be differentiable at $x_0 \in \mathbf{R}^d$.
- The **gradient** of f at the point x_0, denoted $\nabla_x f(x_0)$, is the direction to move in for the **fastest increase** in $f(x)$, when starting from x_0.

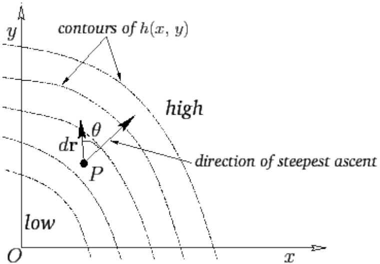


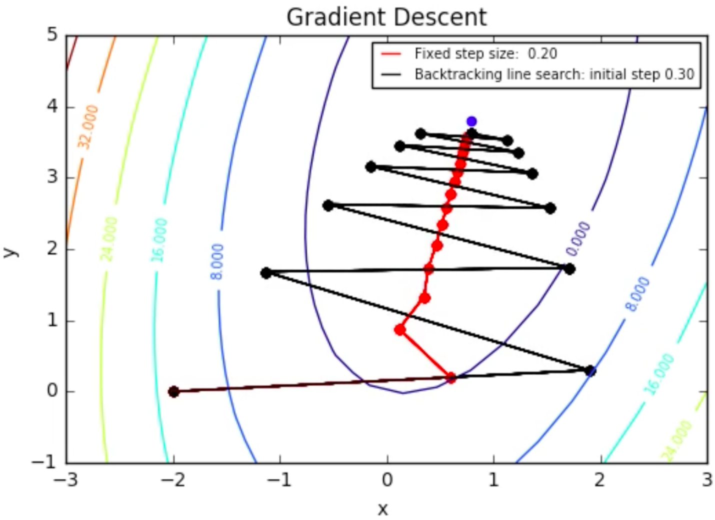
Figure A.111 from Newtonian Dynamics, by Richard Fitzpatrick.

Gradient Descent

- Initialize $x = 0$
- repeat
  - $x \leftarrow x - \underbrace{\eta}_{\text{step size}} \nabla f(x)$
- until stopping criterion satisfied

Choosing the step size is the key in gradient descent.

Gradient Descent Path



Gradient Descent: Step Size

- A fixed step size will work, eventually, as long as it's small enough (roughly - details to come)
  - Too fast, may diverge
  - In practice, try several fixed step sizes
- Intuition on when to take big steps and when to take small steps?

Convergence Theorem for Fixed Step Size

Theorem

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and ∇f is ***Lipschitz continuous*** *with constant $L > 0$, i.e.*

$$\|\nabla f(x) - \nabla f(x')\| \leqslant L\|x - x'\|$$

for any $x, x' \in \mathbb{R}^d$. Then gradient descent with fixed step size $\eta \leqslant 1/L$ ***converges****. In particular,*

$$f(x^{(k)}) - f(x^*) \leqslant \frac{\|x^{(0)} - x^*\|^2}{2\eta k}.$$

This says that gradient descent is guaranteed to converge and that it converges with rate $O(1/k)$.

Gradient Descent: When to Stop?

- Wait until $\|\nabla f(x)\|_2 \leqslant \varepsilon$, for some ε of your choosing.
  - (Recall $\nabla f(x) = 0$ at minimum.)
- For learning setting,
  - evalute performance on validation data as you go
  - stop when not improving, or getting worse