Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 5: Beyond Gradient Descent: Quadratic Objective Function

Quadratic Optimization

In this lecture, we consider the following quadratic optimization problem

$$\min_x Q(x), \qquad Q(x) = \frac{1}{2} x^\top A x - b^\top x, \tag{1}$$

In this problem, we assume that A is a $d \times d$ symmetric positive definite matrix. b and x are d dimensional vectors.

Ridge Regression

In machine learning, the ridge regression problem below can be regarded as quadratic optimization:

$$\min_w \left[\frac{1}{n} \sum_{i=1}^{n} (x_i^\top w - y_i)^2 + \lambda w^\top w \right].$$

Optimal Solution

The solution of (1) is given by the linear equation

$$x_* = A^{-1} b.$$

The system can be solved by gradient descent, with gradient

$$\nabla_x Q(x) = Ax - b.$$

Using $b = Ax_*$, we can write

$$\nabla Q(x) = A(x - x_*).$$

Gradient Descent

We can apply the gradient descent method, which leads to the following iterative algorithm:

$$x_t = x_{t-1} - \eta(Ax_{t-1} - b).$$

It can be checked that for this problem, we have

$$x_t - x_* = (I - \eta A)(x_{t-1} - x_*).$$

Therefore

$$x_t - x_* = (I - \eta A)^t(x_0 - x_*).$$

Convergence Rate

We have convergence result:

$$\|x_t - x_*\|_2 \leq \rho^t \|x_0 - x_*\|_2.$$

If we let $\eta = 1/L$, then $\rho = 1 - \lambda/L = 1 - 1/\kappa$, where $\kappa = L/\lambda$ is the condition number.
The optimal choice is at

$$\eta = 2/(\lambda + L),$$

and the corresponding optimal convergence rate is

$$\rho = \frac{L - \lambda}{L + \lambda} = \frac{\kappa - 1}{\kappa + 1},$$

Conjugate Gradient

Algorithm 1: Conjugate Gradient

Input: A, b, and x_0

Output: x_T

Let $r_0 = b - Ax_0$

Let $p_0 = r_0$

for $t = 1, 2, \ldots, T$ **do**

  Let $\alpha_{t-1} = r_{t-1}^\top r_{t-1} / p_{t-1}^\top A p_{t-1}$

  Let $x_t = x_{t-1} + \alpha_{t-1} p_{t-1}$

  Let $r_t = r_{t-1} - \alpha_{t-1} A p_{t-1}$

  Let $\beta_{t-1} = r_t^\top r_t / r_{t-1}^\top r_{t-1}$

  Let $p_t = r_t + \beta_{t-1} p_{t-1}$

Return: x_T

Properties

It can be shown that

$$r_t = b - Ax_t = -\nabla Q(x_t).$$

The update direction of x_t is not along gradient r_t but aggregated gradient:

$$p_t = -\nabla Q(x_t) + \beta_{t-1} p_{t-1}$$

By expanding p_{t-1} recursively, we obtain

$$p_t = -\nabla Q(x_t) - \beta_{t-1} \nabla Q(x_{t-1}) - \beta_{t-1}\beta_{t-2} \nabla Q(x_{t-2}) - \cdots$$

This means that the search direction p_t is a weighted average of historical gradient as well as the current gradient.

Convergence

Theorem

We have the following convergence result for CG:

$$\|x_t - x_*\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \|x_0 - x_*\|_A,$$

where κ is the condition number of (1).

Conjugate Gradient: pros and cons

Pros:

- is an "optimal" (for linear systems, among all such methods) first order method that employs first order gradient information, and converges faster than regular gradient descent.
- It also automatically sets the parameters α_t and β_t, which is an advantage for general linear systems.

Cons:

- However, for machine learning, which has a finite sum structure over data, such as ridge regression, we often need to use stochastic versions of first order method. In such case, it is difficult to generalize the computation of α_t and β_t as in CG.
- Also CG does not handle non-smooth regularizer such as Lasso well.

Heavy Ball Method

The *Heavy Ball Method*, by Polyak, is very similar to CG, and can be stated in the following recursive equations

$$\begin{aligned} x_t =& x_{t-1} + \alpha_t p_{t-1} \\ p_t =& -\nabla Q(x_t) + \beta_t p_{t-1}. \end{aligned} \tag{2}$$

Set $\alpha_t = \alpha$ and $\beta_t = \beta$.

Alternative Form

It can be verified that by eliminating p_t, the heavy-ball method with constant α and β in (2) can be rewritten as

$$x_t = x_{t-1} - \alpha \nabla Q(x_{t-1}) + \beta(x_{t-1} - x_{t-2}).$$

This formulation can be regarded as a discretization of the following differential equation

$$\frac{d^2x(t)}{dt^2} = -\alpha \nabla f(x(t)) - (1-\beta)\frac{dx(t)}{dt},$$

Convergence

Theorem

We have the following convergence result for the heavy-ball method. Consider $\alpha > 0$ *and* $\beta \in [0, 1]$ *such that*

$$\beta \geq \max((1 - \sqrt{\alpha L})^2, (1 - \sqrt{\alpha \lambda})^2),$$

then there exists a constant $C(\alpha, \beta)$ *that depends on* α *and* β *such that*

$$\overline{\lim_{t \to \infty}} \|x_t - x_*\|_2^{2/t} \leq \beta.$$

Proof

$$x_t = x_{t-1} - \alpha A(x_{t-1} - x_*) + \beta(x_{t-1} - x_{t-2}).$$

That is,

$$x_t - x_* = [(1+\beta)I - \alpha A](x_{t-1} - x_*) - \beta I(x_{t-2} - x_*).$$

This means

$$\begin{bmatrix} x_{t+1} - x_* \\ x_t - x_* \end{bmatrix} = M \begin{bmatrix} x_t - x_* \\ x_{t-1} - x_* \end{bmatrix},$$

where

$$M = \begin{bmatrix} (1+\beta)I - \alpha A & -\beta I \\ I & 0 \end{bmatrix}.$$

Proof

If we let

$$z_t = \begin{bmatrix} x_{t+1} - x_* \\ x_t - x_* \end{bmatrix},$$

then we have

$$z_t = M z_{t-1} = \cdots M^t z_0.$$

Proof

First, let U be the orthogonal matrix such that $U^\top AU$ is a diagonal matrix Λ, and let

$$V = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix},$$

then

$$V^\top MV = \begin{bmatrix} (1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix}.$$

This matrix can be rearranged to block diagonal form with 2×2 block matrices of

$$M_2(\gamma) = \begin{bmatrix} 1+\beta-\alpha\gamma & -\beta \\ 1 & 0 \end{bmatrix},$$

with γ being an eigenvalue of A.

Proof

Eigenvalues of M are the eigenvalues of $M_2(\gamma)$ with γ being an eigenvalue of A. It is easy to check that eigenvalues of $M(\gamma)$ are

$$\lambda_1 = \frac{1}{2}\left(-\sqrt{(\alpha\gamma-\beta-1)^2-4\beta}-\alpha\gamma+\beta+1\right),$$
$$\lambda_2 = \frac{1}{2}\left(\sqrt{(\alpha\gamma-\beta-1)^2-4\beta}-\alpha\gamma+\beta+1\right).$$

The condition of the theorem implies that for $\gamma \in [\lambda, L]$, we have

$$\beta \geq (1-\sqrt{\alpha\gamma})^2.$$

This means that $\sqrt{(\alpha\gamma-\beta-1)^2-4\beta}$ is an imaginary number, and $|\lambda_1| = \sqrt{\beta}$ and $|\lambda_2| = \sqrt{\beta}$. It follows that the spectral radius of M is $\sqrt{\beta}$.

Summary

For quadratic objective functions, one can improve gradient descent using the following family of first order methods:

$$\begin{aligned} x_t &= x_{t-1} + \alpha_t p_{t-1} \\ p_t &= -\nabla Q(x_t) + \beta_t p_{t-1}. \end{aligned}$$

CG, which sets α and β automatically, is the optimal method among all first order methods. However, CG does not generalize to stochastic setting well.

Heavy-ball method, requiring manual settings of α and β, was stated for general nonlinear optimization by Polyak. It is also widely used in deep learning. However, only asymptotic convergence can be proved for general functions.

Next time, we will discuss Nesterov's acceleration, which is a similar method, for which global convergence results can be obtained for convex functions.