Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 4: Gradient Descent for Unconstrained Optimization

Unconstrained Optimization

We consider the following form of unconstrained optimization problem

$$\min_{x} f(x) \tag{1}$$

where $x \in \mathbb{R}^d$ is a parameter to be optimized.

Gradient Descent

Algorithm 1: Gradient Descent

Input: $f(\cdot)$, x_0, and $\eta_1, \eta_2, \ldots$

Output: x_T

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

  Let $x_t = x_{t-1} - \eta_t \nabla f(x_{t-1})$.

Return: x_T

Alternative Formulation of Gradient Descent

It can be easily checked that gradient descent solves the following optimization problem at each step t:

$$x_t = \arg\min_x f_t(x)$$

$$f_t(x) = \left[f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + \frac{1}{2\eta_t} \|x - x_{t-1}\|_2^2 \right]. \qquad (2)$$

If $f(x)$ is L smooth and $\eta_t \leq 1/L$, the optimization problem of (2) is an upper bound of $f(x)$ that equals $f(x)$ at $x = x_{t-1}$.
GD successively optimizes an upper bound of $f(x)$.

Smooth and Strong Convexity

We will first present properties of smooth and strongly convex functions. Recall that an L-smooth function satisfies the condition

$$f(x) \leq f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{L}{2}\|x - x_0\|_2^2.$$

An λ-strongly convex function satisfies the condition

$$f(x) \geq f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{\lambda}{2}\|x - x_0\|_2^2.$$

Gradient Evaluation and Sub-optimality (smooth)

Proposition

If $f(x)$ is L-smooth, then

$$f(x) - \min_x f(x) \geq \frac{1}{2L} \|\nabla f(x)\|_2^2. \tag{3}$$

This means primal suboptimality upper-bounds squared gradient

Relationship of optimization and gradient:

- large gradient: has not optimized well
- small gradient: not clear

Proof

Let $\delta = -\eta \nabla f(x)$, and use the L-smoothness definition

$$f(x+\delta) \leq f(x) + \nabla f(x)^\top \delta + \frac{L}{2}\|\delta\|_2^2,$$

we obtain

$$f(x - \eta \nabla f(x)) \leq f(x) - (\eta - 0.5\eta^2 L)\nabla f(x)^\top \nabla f(x). \tag{4}$$

Since $f(x - \eta \nabla f(x)) \geq \min_x f(x)$, we obtain the result with $\eta = 1/L$.

Gradient Evaluation and Suboptimality (strong-convex)

Proposition

If $f(x)$ is λ-strongly convex, then

$$f(x) - \min_x f(x) \leq \frac{1}{2\lambda} \|\nabla f(x)\|_2^2. \tag{5}$$

This result means when gradient is small, then the optimization problem is approximately solved.

Proof

Let x_* be the unique solution. We have from the definition of strong convexity that

$$\begin{aligned} f(x_*) \geq & f(x) + \nabla f(x)^\top (x_* - x) + \frac{\lambda}{2}\|x_* - x\|_2^2 \\ = & f(x) - \frac{1}{2\lambda}\|\nabla f(x)\|_2^2 + \frac{1}{2\lambda}\|\nabla f(x) + \lambda(x_* - x)\|_2^2 \\ \geq & f(x) - \frac{1}{2\lambda}\|\nabla f(x)\|_2^2. \end{aligned}$$

This implies the desired bound.

Convergence: smooth and strongly convex

Theorem

If $f(x)$ is L-smooth and λ-strongly convex, then (1) has a unique solution x_*. Given any fixed learning rate $\eta_t = \eta \leq 1/L$, we have

$$[f(x_t) - f(x_*)] \leq (1 - \eta\lambda)^t [f(x_0) - f(x_*)].$$

The rate of convergence for smooth and strongly convex functions in Theorem 1 is linear convergence.

Interpretation

If we set $\eta = 1/L$, then the number of iterations t to achieve an ϵ-suboptimal solution

$$f(x_t) \leq f(x_*) + \epsilon$$

is

$$t = O\left(\frac{1}{\eta\lambda}\log(1/\epsilon)\right).$$

With $\eta = 1/L$, this is

$$t = O\left(\frac{L}{\lambda}\log(1/\epsilon)\right).$$

Convergence of Parameters

For smooth and strongly convex problems, we can also obtain the convergence of parameters.

Theorem

If $f(x)$ is L-smooth and λ-strongly convex, then (1) has a unique solution x_*. Given any fixed learning rate $\eta_t = \eta \leq 1/L$, we have

$$\|x_t - x_*\|_2^2 \leq (1 - \eta\lambda)^t \|x_0 - x_*\|_2^2.$$

Proof

$$
\begin{aligned}
&\|x_t - x_*\|_2^2 \\
=&\|x_{t-1} - x_* - \eta \nabla f(x_{t-1})\|_2^2 \\
=&\|x_{t-1} - x_*\|^2 - 2\eta \nabla f(x_{t-1})^\top (x_{t-1} - x_*) + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq&\|x_{t-1} - x_*\|^2 + 2\eta \left[f(x_*) - f(x_{t-1}) - \frac{\lambda}{2}\|x_{t-1} - x_*\|_2^2 \right] + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq&\|x_{t-1} - x_*\|^2 + 2\eta \left[f(x_*) - f(x_{t-1}) - \frac{\lambda}{2}\|x_{t-1} - x_*\|_2^2 \right] \\
&\qquad + \eta^2 2L[f(x_{t-1}) - f(x_*)] \\
=&(1 - \eta\lambda)\|x_{t-1} - x_*\|^2 + 2\eta(1 - \eta L)\left[f(x_*) - f(x_{t-1})\right] \\
\leq&(1 - \eta\lambda)\|x_{t-1} - x_*\|^2.
\end{aligned}
$$

Using induction on t, we obtain the desired bound.

Smooth Functions

For a function that is convex and smooth but not strongly-convex, its solution may not be finite, as shown in Figure 1.

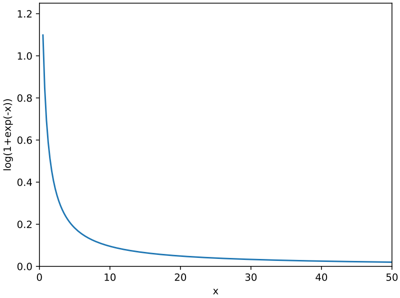

Figure: Solution of Smooth Convex Function

Convergence for Smooth Functions

Theorem

If $f(x)$ is L-smooth and convex, then given an arbitrary finite $\bar{x}$, and a fixed learning rate $\eta_t = \eta \leq 1/L$, we have

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t-1}) \leq f(\bar{x}) + \frac{\|x_0 - \bar{x}\|_2^2}{2T\eta}.$$

Proof

The technique in this proof starts with a similar argument as that of Theorem 2 with $\lambda = 0$:

$$
\begin{aligned}
\|x_t - \bar{x}\|_2^2 =& \|x_{t-1} - \bar{x} - \eta \nabla f(x_{t-1})\|_2^2 \\
=& \|x_{t-1} - \bar{x}\|^2 - 2\eta \nabla f(x_{t-1})^\top (x_{t-1} - \bar{x}) + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq& \|x_{t-1} - \bar{x}\|^2 + 2\eta \left[f(\bar{x}) - f(x_{t-1})\right] + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq& \|x_{t-1} - \bar{x}\|^2 + 2\eta \left[f(\bar{x}) - f(x_{t-1})\right] + 2\eta [f(x_{t-1}) - f(x_t)] \\
=& \|x_{t-1} - \bar{x}\|^2 + 2\eta \left[f(\bar{x}) - f(x_t)\right].
\end{aligned}
$$

Now by summing $t = 1$ to T, we obtain

$$
\|x_T - \bar{x}\|_2^2 \leq \|x_0 - \bar{x}\|_2^2 + 2\eta \sum_{t=1}^{T} \left[f(\bar{x}) - f(x_t)\right].
$$

This implies the desired bound.

Interpretation

To obtain an average sub-optimality of ϵ, we need

$$T = \frac{\|x_0 - \bar{x}\|_2^2}{2\eta} \cdot \frac{1}{\epsilon}$$

steps, which is sublinear.
Since $\bar{x}$ is arbitrary, we may optimize the bound and obtain

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t-1}) \leq \inf_{\bar{x}} \left[f(\bar{x}) + \frac{\|x_0 - \bar{x}\|_2^2}{2T\eta} \right]$$

Example

Consider regression problem, where we have training data (x_i, y_i), where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We assume a linear model

$$y_i = x_i^\top w + \text{noise},$$

and we solve the problem for w using ridge regression:

$$\hat{w} = \arg\min_w f(w), \qquad f(w) = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top w - y_i)^2 + \lambda_0\|w\|_2^2\right].$$

With $\eta = 1/L$, the convergence of gradient descent to ϵ-optimality is

$$O\left(\frac{\lambda_{\max} + \lambda_0}{\lambda_{\min} + \lambda_0}\log\frac{1}{\epsilon}\right)$$

if $\lambda_0 > 0$. When $\lambda_0 = 0$, then the number of steps is

$$O\left((\lambda_{\max} + \lambda_0)\frac{1}{\epsilon}\right).$$