CSE6243: Advanced Machine Learning **Fall 2024**

Lecture 5: Optimization: Gradient Descent and Density Parametrization I

Lecturer: Bo Dai *Scribes: Aditya Shukla, Pranav Malireddy*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

5.1 Recap

In the previous class, we discussed the following topics:

- Convex sets (examples and convex operations)
- Convex functions (conditions, Jensen's inequality, 1st and 2nd conditions, operational convexity)

These concepts help us recognize and solve convex optimization problems.

5.2 New Content

5.2.1 Gradient Descent

We will focus on unconstrained optimization for this lecture. Consider the following optimization problem:

$$\min_{x \in \Omega} f(x)$$

where $f(x)$ is the objective function and Ω is the domain.

Algorithm: Gradient Descent

Input: Initial point x_0.
For $t = 1, 2, 3, \ldots$:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

where η is the learning rate, and the gradient $\nabla f(x)$ must be differentiable.

5.2.2 Assumptions and Definitions

Assumption: $f(x)$ is convex and L-smooth.
Definition: L-smoothness implies:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \forall x, y$$

5.2.3 Proposition (L-Smoothness Properties)

Given that $f(x)$ is convex and L-smooth, the following properties hold:

(a) $f(x)$ is convex and L-smooth $\Rightarrow (b)$

(b) $0 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{L}{2}||x-y||^2 \Rightarrow (c)$

(c) $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}||\nabla f(x) - \nabla f(y)||^2 \Rightarrow (d)$

(d) $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}||\nabla f(x) - \nabla f(y)||^2 \Rightarrow (a)$

5.2.3.1 Observation 1

Gradient descent update rule for step t

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

The quadratic approximation of f(x) around x_t gives

$$x_{t+1} = \min_x f(x_t) + \nabla f(x_t)^T(x - x_t) + \frac{1}{2\eta}||x - x_t||^2$$

Setting the gradient of the quadratic approximation to 0 to find the update

$$\nabla L(x) = \nabla f(x_t) + \frac{1}{\eta}(x - x_t) = 0$$

Simplifying this gives the final update rule, similar to the first equation

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

5.2.3.2 Observation 2

We begin by evaluating the change in the objective function f after updating from x_t to x_{t+1}:

$$f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^T(-\eta \nabla f(x_t)) + \frac{L}{2}||-\eta \nabla f(x_t)||^2$$

Here, we use the first-order approximation of $f(x)$ at x_t and include L to bound the second-order term.

Using statement (b) from above, we simplify the gradient and norm terms and get:

$$f(x_{t+1}) - f(x_t) \leq -\eta||\nabla f(x_t)||^2 + \frac{L\eta^2}{2}||\nabla f(x)||^2$$

This expression shows that the function value decreases proportionally to $\eta||\nabla f(x_t)||^2$, but there's a correction term involving the Lipschitz constant and the step size squared.

Now, factoring terms and combining them gives:

$$f(x_{t+1}) - f(x_t) \leq -\eta(1 - \frac{L\eta}{2})||\nabla f(x_t)||^2$$

Finally, if we set $\eta = \frac{1}{L}$, the term $1 - L\eta = 0$, leading to:

$$(\eta = \frac{1}{L} \Rightarrow 1 - L\eta = 0)$$

5.2.3.3 Observation 3

We begin by introducing a bound on the distance between x_{t+1} and the optimal solution x_*:

$$||x_{t+1} - x_*|| \leq ||x_t - x_*||$$

This inequality suggests that the distance between the updated point x_{t+1} and the optimal solution is no greater than the distance between the current point x_t and x_*.

Next, squaring both sides:

$$||x_{t+1} - x_*||^2 \leq ||x_t - \frac{1}{2}\nabla f(x_t) - x_*||^2$$

Here, we subtract the gradient term $\nabla f(x_t)$, scaled by $\frac{1}{2}$, from x_t, capturing how the function behaves after taking a gradient step.

Expanding the terms on the right-hand side:

$$||x_{t+1} - x_*||^2 \leq ||x_t - x_*||^2 + \frac{1}{L^2}||\nabla f(x_t)||^2 - \frac{2}{L}\nabla f(x_t)^T(x_t - x_*)$$

The expression now includes three terms: the squared distance between x_t and x_*, a correction based on the squared norm of the gradient, and a term involving the inner product of the gradient and the distance between x_t and x_*.

Given the condition:

$$\nabla f(x_t) = 0, \quad \nabla f(x)^T(x - x_t) \geq \frac{1}{L}||x - x_*||^2$$

This condition ensures that the gradient at x_t vanishes at the optimal point x_*, and it gives a lower bound on the inner product of the gradient with the difference between x and x_t.

Finally, we conclude:

$$||x_{t+1} - x_*||^2 \leq -\frac{2}{L} \cdot \frac{1}{L}||x_t - x_*||^2 \leq ||x_t - x_*||^2$$

This shows that the distance to the optimal point x_* decreases with each step, given that the gradient is well-behaved and the step size is chosen appropriately.

5.2.3.4 Observation 4

$$||\nabla f(x_t)||^2 \geq \frac{(f(x_t) - f(x_*))}{||x_t - x_*||}$$

We bound the function difference using the gradient and the distance from the optimal solution:

$$f(x_t) - f(x_*) \leq \nabla f(x_t)^T (x_t - x_*) \leq ||\nabla f(x_t)||||x_t - x_*||$$

Using Observation 2, we get:

$$f(x_t) - f(x_*) \leq \frac{1}{2}||\nabla f(x_t)||^2$$

This allows us to bound the function difference in terms of the gradient:

$$f(x_t) - f(x_*) \leq -\frac{1}{2L}\left[\frac{(f(x_t) - f(x_*))}{||(x_t - x_*)||}\right]^2$$

$$f(x_t) - f(x_*) \leq -\frac{1}{2L}\left[\frac{(f(x_t) - f(x_*))}{||(x_0 - x_*)||}\right]^2$$

Now, we define the difference between consecutive function values:

$$f(x_{t+1}) - f(x_t) = \epsilon_{t+1}, \quad (f(x_t) - f(x_*)) = \epsilon_t$$

Introducing β:

$$\beta = \frac{1}{2L} \cdot \frac{1}{||x_0 - x_*||^2}$$

We then establish the following relationship:

$$\frac{(\epsilon_{t+1} - \epsilon_t)}{\epsilon_{t+1} \cdot \epsilon_t} \leq \frac{-\beta\epsilon_t^2}{\epsilon_{t+1} \cdot \epsilon_t}$$

Which leads to the inequality:

$$\Rightarrow \frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t+1}} \leq \beta \cdot \frac{\epsilon_t}{\epsilon_{t+1}} \leq -\beta$$

$$\Rightarrow \frac{1}{\epsilon_t} + \beta \leq \frac{1}{\epsilon_{t+1}}$$

$$\frac{1}{\epsilon_t} \geq \frac{1}{\epsilon_{t-1}}$$

$$\Rightarrow \frac{1}{\epsilon_{t-1}} + \beta$$

$$\frac{1}{\epsilon_0} + \beta_t \leq \frac{1}{\epsilon_t} \Rightarrow \beta_t \leq \frac{1}{\epsilon_t}$$

$$\Rightarrow \beta_t \leq \frac{1}{f(x_t) - f(x_*)}$$

5.2.3.5 Theorem 1:

We begin with a bound on the function value at iteration t:

$$f(x_t) - f(x_*) \leq \frac{1}{\beta_t}$$

This inequality gives us an upper bound on the difference between the function value at iteration t and the optimal function value, in terms of β_t.

Next, we define the gradient of the function as the average gradient over n samples:

$$\nabla f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

We can sample R terms from the n samples:

Sample R from n:

$$\nabla f(x) = \frac{1}{R} \sum_{j=1}^{R} \nabla f_j(x)$$

This approximation reduces the computational cost by estimating the gradient using only a subset R of the full sample set.

5.2.4 Density Parameterization:

We start by introducing the base measure function, denoted as:

$$h(x) \rightarrow \text{ base measure}$$

The base measure $h(x)$ represents a fundamental component of the density function that helps in parameterizing the distribution.

Next, we introduce the sufficient statistic, which is crucial for describing the data:

$$T(x) \rightarrow \text{ sufficient statistician partition function}$$

Here, $T(x)$ represents the sufficient statistic, and it's used in conjunction with the partition function to express the density. The partition function normalizes the distribution and is key in probability calculations.

5.2.4.1 Exponential Family:

The probability density function of the exponential family is given by:

$$p(x) = h(x)\ exp(w^T T(x) - A(w))$$

In this expression: - $h(x)$ is the base measure (as discussed before). - $T(x)$ is the sufficient statistic. - w is the parameter of the distribution. - $A(w)$ is the log-partition function, which normalizes the distribution.

The log-partition function is defined as:

$$A(w) = \log \int \exp(w^T T(x))\, dx$$

This function ensures that the distribution integrates to 1, which is a necessary property for any probability density function.

Now, let's check the normalization condition:

$$\int p(x)\, dx = \int h(x)\ exp(w^T T(x)) * \ exp(-A(w))\, dx$$

This step shows that the density integrates to 1 by using the log-partition function $A(w)$ to cancel out terms.

We can factorize the expression as:

$$\int p(x)\, dx = [\int h(x)\ exp(w^T T(x)\, dx)]\ exp(A(w))$$

Finally, since the integral of the probability density must equal 1:

$$\int p(x)\, dx = 1$$