CPSC 8420 Advanced Machine Learning
Week 6: Gradient Descent

Dr. Kai Liu

September 24, 2020

Motivation

Gradient Descent is one of the most widely used methods in machine learning to optimize solution.

- Most problems in machine learning they don't have a closed solution like Least Squares or PCA, thus an iterative updating algorithm may be more applicable.
- Obtaining a closed solution maybe computationally demanding with operations such as matrix inversion or SVD, *etc*.
- Gradient Descent works well not only for convex problem, but also non-convex one. Though global minimum can not be guaranteed, most of times, local minimum is acceptable.

Learning Outcomes

Our goal for today's lecture is to understand:

- Why Gradient Descent decreases the objective
- Convergence rate of GD in various conditions
- Variations of GD

Gradient Descent

Differentiable unconstrained minimization

$$\min_{x} f(x), \quad s.t. \quad x \in \mathbb{R}^n \tag{1}$$

where f (objective or cost function) is differentiable

Iterative descent algorithms

Start with a point x^0, and construct a sequence $\{x^t\}$ such that:

$$f(x^{t+1}) < f(x^t), \quad t = 0, 1, 2 \ldots \tag{2}$$

In each iteration, search in descent direction

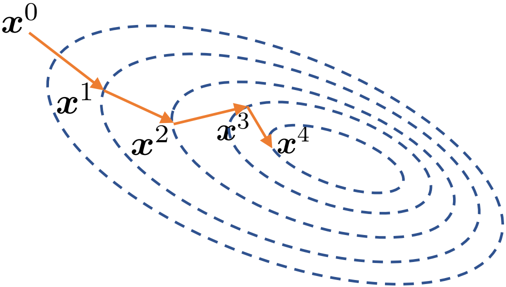
$$x^{t+1} = x^t + \eta_t d^t, \tag{3}$$

where d^t is the descent direction at x^t; $\eta_t > 0$ is the stepsize.

Gradient descent (GD)

One of the most important examples is **Gradient descent (GD)**:

$$x^{t+1} = x^t - \eta_t \nabla f(x^t). \tag{4}$$



It can be traced to Augustin Louis Cauchy.

Gradient descent (GD)

One of the most important examples is **Gradient descent (GD)**:

$$x^{t+1} = x^t - \eta_t \nabla f(x^t). \tag{5}$$

descent direction: $d^t = -\nabla f(x^t)$, which can be justified from Taylor Expansion:

$$f(x^{t+1}) \approx f(x^t) + \langle x^{t+1} - x^t, \nabla f(x^t) \rangle + \frac{1}{2} \langle x^{t+1} - x^t, \nabla^2 f(x^t)(x^{t+1} - x^t) \rangle. \tag{6}$$

Convergence of Gradient Descent

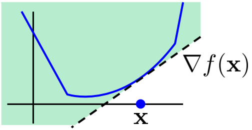
Strongly Convex Strongly Smooth Function

We say a continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is α-strongly convex (SC) and β-strongly smooth (SS) if for every $x, x^+ \in \mathbb{R}^p$, we have:

$$\frac{\alpha}{2}\|x^+ - x\|_2^2 \leq f(x^+) - f(x) - \langle \nabla f(x), x^+ - x \rangle \leq \frac{\beta}{2}\|x^+ - x\|_2^2, \quad (7)$$

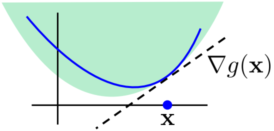
α, β can be taken as the smallest and largest singular value of $\nabla^2 f(x)$.

Convex, SC and SS



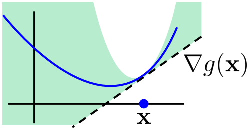
$f : \mathbb{R}^d \to \mathbb{R}$

CONVEX FUNCTION



$g : \mathbb{R}^d \to \mathbb{R}$

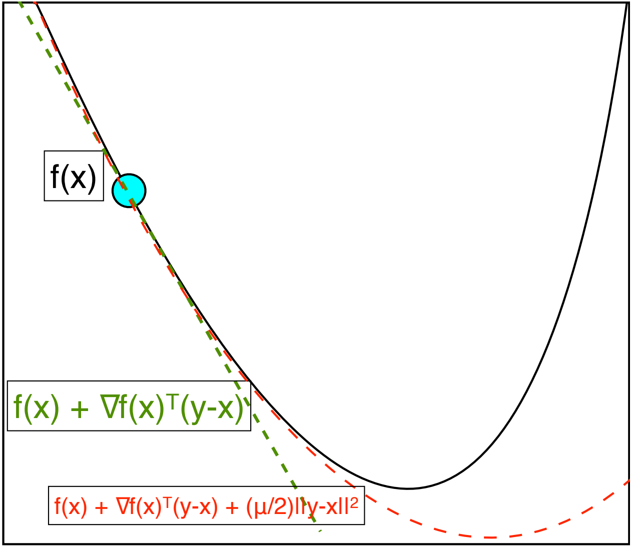
STRONGLY CONVEX FUNCTION



$g : \mathbb{R}^d \to \mathbb{R}$

STRONGLY SMOOTH CONVEX FUNCTION

Convex, SC and SS



Strongly Convex Strongly Smooth Function

$$\frac{\alpha}{2}\|x^+ - x\|_2^2 \leq f(x^+) - f(x) - \langle \nabla f(x), x^+ - x \rangle \leq \frac{\beta}{2}\|x^+ - x\|_2^2, \quad (8)$$

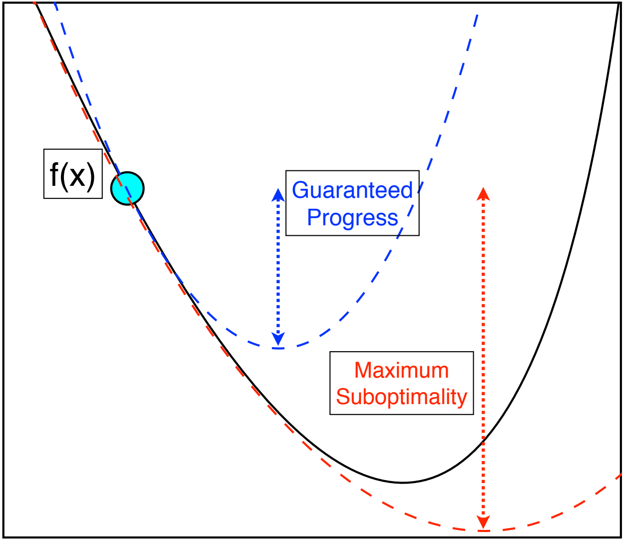
based on which we will have:

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \\ f(x^+) &\geq f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|^2. \end{aligned} \quad (9)$$

Replacing x^+ with x^*, then we will have:

$$\frac{1}{2\beta}\|\nabla f(x)\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2 \quad (10)$$

Progress within each Update



Linear Convergence Rate

$$
\begin{aligned}
f(x^+) - f(x^*) &\leq f(x) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \\
&\leq f(x) - f(x^*) - \frac{\alpha}{\beta}(f(x) - f(x^*)) \\
&= (1 - \frac{\alpha}{\beta})(f(x) - f(x^*)) \\
&\leq exp^{-\frac{\alpha}{\beta}}(f(x) - f(x^*))
\end{aligned} \tag{11}
$$

which implies $\frac{f(x^+)-f(x^*)}{f(x)-f(x^*)} = 1 - \frac{\alpha}{\beta}$, is the definition of linear convergence. $\frac{f(x^t)-f(x^*)}{f(x^0)-f(x^*)} \leq exp^{-\frac{\alpha}{\beta}t}$ Then to obtain ϵ-suboptimal result, we need $\mathcal{O}(\kappa log\frac{1}{\epsilon})$ iterations, where we define $\kappa = \frac{\beta}{\alpha}$ to be the **condition number**.

Revisiting Ridge Regression

Vanilla Least Squares:

$$\min_{\beta} \frac{1}{2}\|y - A\beta\|^2 \tag{12}$$

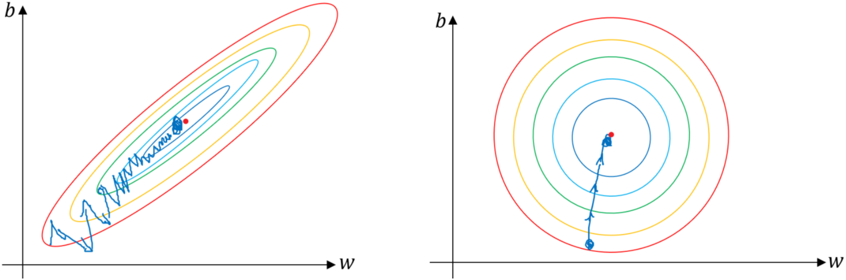
which is a convex problem with respect to β, and we can verify that $\nabla^2 f(\beta) = A^T A$, thus $\kappa = \frac{\sigma_{max}(A^T A)}{\sigma_{min}(A^T A)}$.

Vanilla Least Squares:

$$\min_{\beta} \frac{1}{2}\|y - A\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2 \tag{13}$$

which is also convex with respect to β, and we can verify that $\nabla^2 f(\beta) = A^T A + \lambda I$, thus the new $\tilde{\kappa} = \frac{\sigma_{max}(A^T A)+\lambda}{\sigma_{min}(A^T A)+\lambda} < \kappa$.

Geometry of κ to influence convergence rate



Non Strongly Convex

$$\begin{aligned} f(x^+) - f(x^*) &\leq f(x) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \\ &\leq \langle \nabla f(x), x - x^* \rangle - \frac{1}{2\beta}\|\nabla f(x)\|^2 \end{aligned} \tag{14}$$

on the other hand we have:

$$\begin{aligned} \|x^+ - x^*\|^2 &= \|x - \eta\nabla f(x) - x^*\|^2 \\ &= \|x - x^*\|^2 - 2\eta\langle \nabla f(x), x - x^* \rangle + \eta^2\|\nabla f(x)\|^2 \\ &= \|x - x^*\|^2 - 2\eta(\langle \nabla f(x), x - x^* \rangle - \frac{\eta}{2}\|\nabla f(x)\|^2), \end{aligned} \tag{15}$$

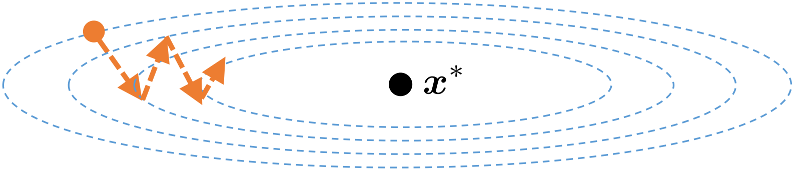
then we have
$\langle \nabla f(x), x - x^* \rangle - \frac{\eta}{2}\|\nabla f(x)\|^2 = \frac{1}{2\eta}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)$, and
therefore $f(x^+) - f(x^*) \leq \frac{1}{2\eta}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)$

Non Strongly Convex

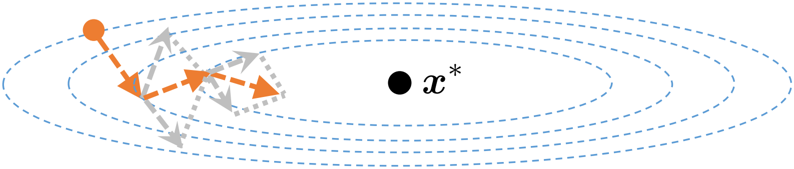
Summation the equation above from $k = 0$ to $k = T - 1$, we have: $\sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 - \|x_T - x^*\|^2}{2\eta} \leq \frac{\|x_0 - x^*\|^2}{2\eta}$, then $f(x_T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2T\eta}$, which is sub-linear convergence rate. Then to obtain ϵ-suboptimal result, we need $T = \mathcal{O}(\frac{1}{\epsilon})$ iterations.

Variations

Heavy-ball



gradient descent



heavy-ball method

Heavy-ball

B. Polyak

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t) + \underbrace{\theta_t(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})}_{\text{momentum term}}$$

- add inertia to the "ball" (i.e. include a momentum term) to mitigate zigzagging

Heavy-ball

iteration complexity: $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$

significant improvement over GD: $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ vs. $O(\kappa\log\frac{1}{\varepsilon})$

Nesterov Accelerate Gradient

Y. Nesterov

— *Nesterov '83*

$$\boldsymbol{x}^{t+1} = \boldsymbol{y}^t - \eta_t \nabla f(\boldsymbol{y}^t)$$
$$\boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{t}{t+3}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$$

Nesterov Accelerate Gradient

Iteration complexity: $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$, much faster than gradient descent: $\mathcal{O}(\frac{1}{\epsilon})$, alternates between gradient updates and proper extrapolation.
It is not a descent method (i.e. we may not have $f(x^{t+1}) \leq f(x^t)$).
It is one of the **most beautiful and mysterious** results in optimization.

Conclusion

	stepsize rule	convergence rate	iteration complexity
convex & smooth problems	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t^2}\right)$	$O\left(\frac{1}{\sqrt{\varepsilon}}\right)$
strongly convex & smooth problems	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^t\right)$	$O\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$