Optimization for Deep Learning

Lecture 3 Part II: Preconditioned Gradient Descent

Kun Yuan

Peking University

Preconditioned GD: motivation

- Another way to accelerate GD is to use preconditioning

- Consider an ill-conditioned quadratic problem

$$\min_x \quad x^T Q x + c^T x$$

where Q is a ill-conditioned matrix. GD is slow when solving the problem

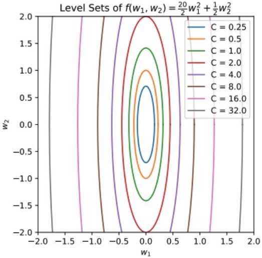


Figure: An ill-conditioned QP problem. (From Prof. Chris De Sa's lecture notes)

Preconditioned GD: motivation

- We now let $x = P^{\frac{1}{2}} w$ for some positive definite matrix P
- Since P is positive definite, x and w is an $1 - 1$ mapping
- If we choose $P = Q^{-1}$, we have $x^T Q x = w^T Q^{-\frac{1}{2}} Q Q^{\frac{1}{2}} w = \|w\|^2$

Preconditioned GD: motivation

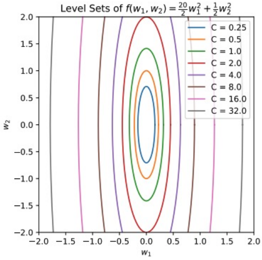
- With $x = P^{\frac{1}{2}} w$ and $P = Q^{-1}$, the ill-conditioned problem becomes

$$\min_{w} \quad \frac{1}{2}\|w\|^2 + c^T Q^{-\frac{1}{2}} w$$

  which is a benign problem. GD is fast to achieve $w^\star$.

- Once $w^\star$ is determined, we have $x^\star = P^{\frac{1}{2}} w^\star$.

Preconditioned GD: motivation



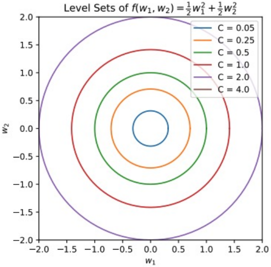


Figure: Left: an ill-conditioned QP problem. Right: a benign QP problem after transformation.(From Prof. Chris De Sa's lecture notes)

Preconditioned GD: derivation

- Consider a general ill-conditioned optimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x)$$

- We let $x = P^{\frac{1}{2}} w$ so that $g(w) = f(P^{\frac{1}{2}} w)$ is a nice function.

- Use gradient descent to minimize $g(w)$, i.e.,

$$w_{k+1} = w_k - \gamma \nabla g(w_k) = w_k - \gamma P^{\frac{1}{2}} \nabla f(P^{\frac{1}{2}} w_k)$$

- Left-multiplying $P^{\frac{1}{2}}$ to both sides, we achieve

$$\begin{aligned} P^{\frac{1}{2}} w_{k+1} &= P^{\frac{1}{2}} w_k - \gamma P \nabla f(P^{\frac{1}{2}} w_k) \\ \iff \quad x_{k+1} &= x_k - \gamma P \nabla f(x_k) \end{aligned}$$

  where P is called the preconditioning matrix.

Preconditioned GD: derivation

- The preconditioned GD algorithm

$$x_{k+1} = x_k - \gamma P \nabla f(x_k)$$

- It is critical to choose the preconditioning matrix P

- If $P = [\nabla^2 f(x_k)]^{-1}$, then preconditioned GD reduces to Newton's method

- It is critical to construct an efficient and effective P matrix

GD v.s. Newton's method

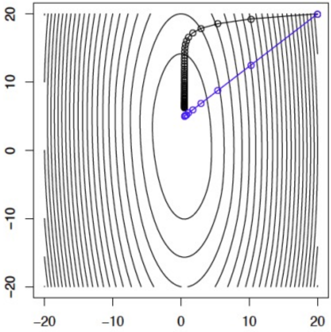


Figure: Convergence comparison between GD and Newton.