

# CMO: Coordinate Descent

Eklavya Sharma

**Objective:** Minimize  $f(x) = \frac{1}{2}x^T Qx - b^T x$ , where  $Q$  is symmetric and positive definite.  $Q$  also has a large size. So large that it's stored on secondary/network storage.

Let  $x^* = \operatorname{argmin}_x f(x)$ .

$$\nabla_f(x) = Qx - b = Q(x - x^*)$$

$$f(x^*) = -\frac{x^{*T} Q x^*}{2}$$

Let  $g(\alpha) = f(x^{(i)} + \alpha u)$  where  $u$  is a descent direction. i.e.  $\nabla_f(x^{(i)})^T u < 0$ .

$$g'(0) = \nabla_f(x^{(i)})^T u$$

$$g''(0) = u^T Q u$$

Now we'll use exact line search to find out the step size.

**Theorem 1.** Let  $\alpha^* = \operatorname{argmin}_\alpha g(\alpha)$ . Then

$$\alpha^* = -\frac{g'(0)}{g''(0)} > 0$$

$$g(\alpha^*) = f(x^{(i)}) - \frac{g'(0)^2}{2g''(0)} = f(x^{(i)}) - \frac{(\nabla_f(x^{(i)})^T u)^2}{2u^T Q u}$$

*Proof.* By Taylor series,

$$g(\alpha) = g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(0)$$

$$g'(\alpha) = g'(0) + \alpha g''(0)$$

By the necessary condition for local minimum,

$$g(\alpha^*) = 0 \implies \alpha^* = -\frac{g'(0)}{g''(0)}$$

□

**Theorem 2.**

$$f(x^{(i)}) - f(x^*) = \frac{\nabla_f(x^{(i)})^T Q^{-1} \nabla_f(x^{(i)})}{2}$$

*Proof sketch.* Let  $v = x^{(i)} - x^*$ . Replace  $x^{(i)}$  by  $x^* + v$  in  $f(x^{(i)}) - f(x^*)$ . The rest is algebraic manipulation.  $\square$

Let  $E(x) = f(x) - f(x^*)$ . Let  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$  be the eigenvalues of  $Q$ .

$$\begin{aligned}
\Delta &= \frac{E(x^{(i)}) - E(x^{(i+1)})}{E(x^{(i)})} \\
&= \frac{f(x^{(i)}) - f(x^{(i+1)})}{f(x^{(i)}) - f(x^*)} \\
&= \frac{g(0) - g(\alpha^*)}{f(x^{(i)}) - f(x^*)} \\
&= \frac{(\nabla_f(x^{(i)})^T u)^2}{2u^T Q u} \frac{2}{\nabla_f(x^{(i)})^T Q^{-1} \nabla_f(x^{(i)})} \\
&= \frac{(\nabla_f(x^{(i)})^T u)^2}{(u^T Q u) (\nabla_f(x^{(i)})^T Q^{-1} \nabla_f(x^{(i)}))} \\
&= \frac{(\nabla_f(x^{(i)})^T u)^2}{([\lambda_d, \lambda_1] \|u\|^2) \left( \left[ \frac{1}{\lambda_1}, \frac{1}{\lambda_d} \right] \|\nabla_f(x^{(i)})\|^2 \right)} \\
&\geq \frac{\lambda_d}{\lambda_1} \frac{(\nabla_f(x^{(i)})^T u)^2}{\|u\|^2 \|\nabla_f(x^{(i)})\|^2}
\end{aligned}$$

To prove linear convergence, we must come up with  $u$  such that  $\Delta$  is lower-bounded by a positive constant (because  $\frac{E(x^{(i+1)})}{E(x^{(i)})} = 1 - \Delta$ ).

Let  $e_j$  be the  $j^{\text{th}}$  column of the identity matrix. In coordinate-descent, we choose  $u$  to be  $e_j$  or  $-e_j$  for some  $j$ . This has the advantage of being computationally lightweight. For example,  $u^T Q u = Q_{j,j}$ , which takes  $O(1)$  time instead of  $O(d^2)$ .

Let  $g = \nabla_f(x^{(i)})$ . Let  $g_j$  be the  $j^{\text{th}}$  coordinate of  $g$ . For  $u$  to be a descent direction, we'll choose  $u = -\text{sgn}(g_j)e_j$ . Therefore,  $g^T u = -\text{sgn}(g_j)g_j^T e_j = -|g_j| < 0$ .

Also, we'll choose the  $j$  which has the highest value of  $|g_j|$ . Therefore,

$$\begin{aligned}
\|g\|^2 &= \sum_{k=1}^d |g_k|^2 \leq d|g_j|^2 \\
\Delta &\geq \frac{\lambda_d}{\lambda_1} \frac{(\nabla_f(x^{(i)})^T u)^2}{\|u\|^2 \|\nabla_f(x^{(i)})\|^2} = \frac{\lambda_d}{\lambda_1} \frac{|g_j|^2}{\|g\|^2} \geq \frac{\lambda_d}{d\lambda_1}
\end{aligned}$$