Gradient Descent!

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

The gradient of f at w is

$$\nabla f(w) = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \end{bmatrix}$$

Or... $\nabla f(w)$ is the unique vector s.t. $\forall u \in \mathbb{R}^d$

$$u^T \nabla f(w) = \lim_{\alpha \to 0} \frac{f(w+\alpha u) - f(w)}{\alpha}$$

$$= \frac{\partial}{\partial \alpha} f(w+\alpha u)\Big|_{\alpha = 0}$$

e.g. if $f(w) = w^T A w$ for matrix A. Then:

$$\frac{\partial}{\partial \alpha} f(w+\alpha u) = \frac{\partial}{\partial \alpha} (w+\alpha u)^T A (w+\alpha u)$$

$$= \frac{\partial}{\partial \alpha} w^T A w + \alpha \cdot u^T A w + \alpha \cdot w^T A u + \alpha^2 u^T A u$$

$$= u^T A w + w^T A u + 2\alpha u^T A u$$

$$@\ \alpha = 0 \quad = u^T A w + w^T A u = u^T \underbrace{(A w + A^T w)}_{\nabla f(w)}$$

Hessian: $\nabla^2 f \leftrightarrow H$

$$\nabla^2 f(w) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \cdots & \\ \vdots & & \end{bmatrix} \in \mathbb{R}^{d \times d}$$

or: the unique symmetric matrix s.t.

$$u^T \nabla^2 f(w) u = \frac{\partial^2}{\partial \alpha^2} f(w + \alpha u)\Big|_{\alpha = 0}.$$

if $f: \mathbb{R}^1 \to \mathbb{R}$, then $\nabla f: \mathbb{R}^1 \to \mathbb{R}^1$ & $\nabla^2 f: \mathbb{R}^1 \to \mathbb{R}^{1 \times 1}$.

A general recipe: initialize $w_0 \in \mathbb{R}^d$

loop until converged:
- select some "step" $s \in \mathbb{R}^d$
- $w \leftarrow w + s$
- "converged" if $\|s\| \leq \delta$ (δ tolerance parameter)

Conclusion

$$\ell(w+s) \approx \ell(w) + s^T \nabla \ell(w)$$

$$\ell(w+s) \approx \ell(w) + s^T \nabla \ell(w) + \frac{1}{2} s^T \nabla^2 \ell(w) s$$

Gradient descent: idea look at $\ell(w+s)$ as a first order approximation

direction of steepest descent is $-\nabla \ell(w)$

$$w \leftarrow w - \underbrace{\alpha}_{\text{step size, learning rate}} \nabla \ell(w) \quad \text{(gradient descent)}$$

$$\ell(w - \alpha \nabla \ell(w)) = \ell(w) - \alpha \nabla \ell(w)^T \nabla \ell(w) + \frac{1}{2} (\alpha \nabla \ell(w))^T \underbrace{\nabla \ell(w + t\alpha \nabla \ell(w))}_{\text{bounded}} (\alpha \nabla \ell(w))$$

$$= \ell(w) - \alpha \|\nabla \ell(w)\|^2 + \alpha^2 \sigma(\|\nabla \ell(w)\|^2)$$

$$= \ell(w) - \underbrace{(\alpha - \sigma(\alpha^2))}_{\geq \frac{\alpha}{2}} \|\nabla \ell(w)\|^2$$

$\alpha - C\alpha^2$

pick $\alpha \leq \frac{1}{2C}$

$$\ell(w - \alpha \nabla \ell(w)) \leq \ell(w) - \frac{\alpha}{2} \|\nabla \ell(w)\|^2$$

if we didn't converge, $\|s\| \geq \delta \Rightarrow \|\alpha \nabla \ell(w)\| \geq \delta$

$$\ell(w - \alpha \nabla \ell(w)) \leq \ell(w) - \frac{\delta^2}{2\alpha}$$

✓ GD converges

Newton's Method

$$\ell(\omega + s) \approx \ell(\omega) + s^T \nabla \ell(\omega) + \frac{1}{2} s^T H(\omega) s$$

($H(\omega) = \nabla^2 \ell(\omega)$)

$$0 = \nabla \ell(\omega) + H(\omega) s \Rightarrow s = -(H(\omega))^{-1} \nabla \ell(\omega)$$

Newton's s

$$\omega \leftarrow \omega - (H(\omega))^{-1} \nabla \ell(\omega) \quad \text{(Newton's method)}$$

e.g. $\omega_{t+1} = \omega_t - (H(\omega_t))^{-1} \nabla \ell(\omega_t)$