Oct 30, 2023

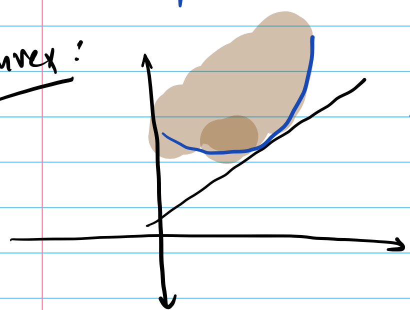
Risk $R(\theta) = E[\ell(Y, \theta)]$

Optimization

$$\nabla R(\theta_0) = 0$$

Assumption $R(.)$ is strongly convex

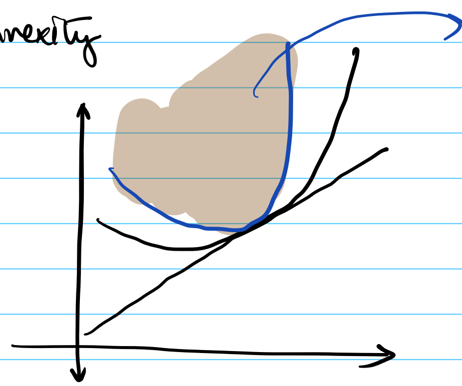
Convex:



$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle$$

inner product

Strong Convexity

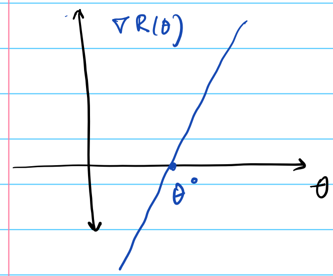


Is also above the parabola.

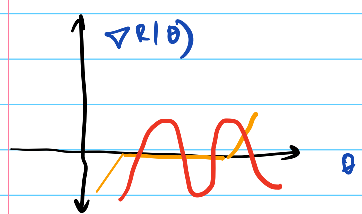
(rules out flat pts.)

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\alpha}{2} \| x - x^* \|^2$$

We want $\nabla R(\theta)$ to look like this.



Strong Convexity rules out cases like :—

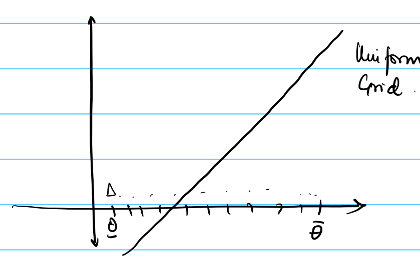


1. Grid Search. $[\underline{\theta}, \bar{\theta}]$

$\theta_i = \underline{\theta} + i\Delta$

$\min_{\theta_i} \| \nabla R(\theta_i) \|$, N steps

$\Delta = \frac{\bar{\theta} - \underline{\theta}}{N}$ (Accuracy)



$\varepsilon \geq |\theta^* - \theta_0|$

$N \sim \frac{1}{\varepsilon}$

You do not have a closed form.

2. Bisection Algorithm

$a_0 = \underline{\theta}$, $b_0 = \bar{\theta}$

$c_1 = \dfrac{\bar{\theta} + \underline{\theta}}{2}$

$\text{sign}(\nabla R(a_0)) \neq \text{sign}(\nabla R(b_0))$

If $\text{sign}(\nabla R(b_0)) \neq \text{sign}(\nabla R(c_1))$

$a_1 = c_1, \; b_1 = b_0$

$a_k, \; b_k, \quad c_{k+1} = \dfrac{a_k + b_k}{2}$

If $\text{sign}(\nabla R(c_{k+1})) \neq \text{sign}(\nabla R(b_k))$

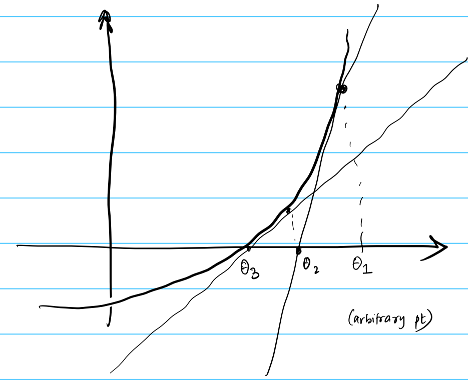
$a_{k+1} = c_{k+1}, \quad b_{k+1} = b_k \qquad [a_k, b_k]$

$|b_k - a_k| = \varepsilon = \dfrac{|\bar{\theta} - \underline{\theta}|}{2^k} \qquad N \sim \log \dfrac{1}{\varepsilon}$

3. Newton Algorithm

$$H_\theta = \left(\frac{\partial^2 R(\theta)}{\partial\theta_i\, \partial\theta_j}\right)_{ij}$$

In one dimension, this would just be double derivative.



Iteration k : θ_k

$\nabla R(\theta_k) + H(\theta_k)(\theta - \theta_k) = 0$ ∴

$\theta_{k+1} = \theta_k - H(\theta_k)^{-1} \nabla R(\theta_k)$

$$\underbrace{\Delta R(\theta_k)}_{R(\theta_{k+1}) - R(\theta_k)} \approx \underbrace{R'(\theta_k)}_{\nabla R(\theta_k)} \cdot (\theta_{k+1} - \theta_k) + \underbrace{\frac{1}{2} R''(\theta_k)}_{H(\theta_k)} \cdot (\theta_{k+1} - \theta_k)$$

$R(\theta_{k+1}) - R(\theta_k) \approx$

$0 =$

(same distⁿ of the R.V.)

$R(\theta) = E[\ell(Y, \theta)]$ we can't keep ~~along~~ the R.V. ?

A sample: $\{Y_i\}_{i=1}^{N}$ which has a cutoff → 1) data collection is expensive; 2) most of the data is historical; there is only so much data there.

Analogy Principle

a set of N numbers

We take the sample & create a new distribution "replicating" sampling scheme $\{y_i\}_{i=1}^{N}$

New r.v.

$P(Z = y_i) = \frac{1}{N}$ (Z takes value with equal probability)

empirical risk:

$$\hat{R}(\theta) = E_N[\ell(z, \theta)]$$
$$= \frac{1}{N} \sum_{i=1}^{N} \ell(y, \theta)$$

hats are for empirical stuff