# Unconstrained optimization: gradient descent, Newton, ...

Monday, September 30, 2024 9:36 AM

$\min_w f(w)$

- No constraints on $w$
- Assume $f$ is "smooth"

Necessary condition for optimality:

"Stationarity" $\nabla f(w) = 0$ } looks like root-finding!

Most optimization algo. framework

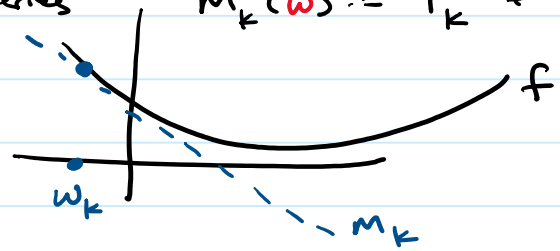
$w_{k+1} = \operatorname{argmin}_w m_k(w)$ ← model of $f$, easier to minimize than $f$ itself

most models involve Taylor series

ex:

- 1st order Taylor series

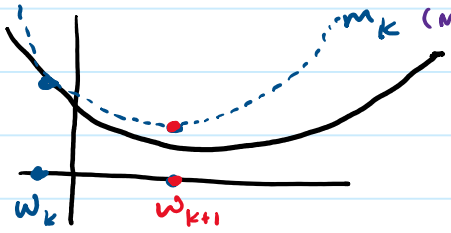
$m_k(w) := f_k + \langle \nabla f_k, w - w_k \rangle$ ($f_k = f(w_k)$, $\nabla f_k = \nabla f(w_k)$)




Not useful!

- 1st order Taylor series w/ quadratic penalty

$m_k(w) := f_k + \langle \nabla f_k, w - w_k \rangle + \underbrace{\frac{1}{2\eta} \|w - w_k\|_2^2}_{\text{penalty / regularizer}}$




Thm: If $f$ convex, $\nabla f$ is $L$-Lipschitz continuous (or $\nabla^2 f \preceq L \cdot I$) then if $\eta < \frac{2}{L}$ it'll converge

what is $\operatorname{argmin}_w m_k(w)$? Set $\nabla m_k = 0$

$0 = \nabla f_k + \frac{1}{\eta}(w - w_k)$

so

$w_{k+1} = w_k - \eta \nabla f_k$ GRADIENT DESCENT

# Unconstrained, page 2

Monday, September 30, 2024 9:48 AM

- 2nd order Taylor Series ($= \nabla^2 f(w_k)$)

$$m_k(w) = f_k + \langle \nabla f_k, w - w_k \rangle + \frac{1}{2} \langle w - w_k, \nabla^2 f_k (w - w_k) \rangle$$

So

$$w_{k+1} = \operatorname{argmin}_w m_k(w) = \dots$$

again, solve by setting $\nabla m_k = 0$

$$0 = \nabla f_k + \nabla^2 f_k (w - w_k)$$

$$w_{k+1} = w_k - (\nabla^2 f_k)^{-1} \cdot \nabla f_k \qquad \text{NEWTON'S METHOD}$$

Why don't we use Newton's method all the time?
(since it converges rapidly)

- plain Newton looks for $\nabla f(w) = 0$, a "stationary point", which is necessary for a global minimizer but (for nonconvex problems) not sufficient
- If $f: \mathbb{R}^d \to \mathbb{R}$, $\nabla^2 f_k$ is a $d \times d$ matrix
  - a lot of storage
  - inverting it costs $O(d^3)$

  These drawbacks are alleviated via "quasi-Newton" methods (BFGS)

  BFGS shares these drawbacks
- A bit complicated to extend to constrained problems

- What about "derivative free optimization" (DFO) / "zeroth order" methods?

ex: Nelder-Mead
Genetic algo.
Bayesian optimization
Trust-region model based

In large dimensions, these almost always are significantly inferior to gradient based methods (as long as gradient is available... cheaply which AutoDiff. helps with)