Chapter 10 Newton's Method

- Algorithm $\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f\left(\mathbf{x}_t\right)$, a more general form will be arbitary H, $\mathbf{x}_{t+1} = \mathbf{x}_t - H\left(\mathbf{x}_t\right)\nabla f\left(\mathbf{x}_t\right)$.
- Lemma 10.1 A nondegenerate quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top M\mathbf{x} - \mathbf{q}^\top\mathbf{x} + c$, with M invertible and semmetric matrix, then Newton's method gives optimal solution of $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ in one step $\mathbf{x}_1 = \mathbf{x}^\star$.
  - Proof $\mathbf{x}_0 - \nabla^2 f(\mathbf{x}_0)^{-1}\nabla f\left(\mathbf{x}_0\right) = \mathbf{x}_0 - M^{-1}\left(M\mathbf{x}_0 - \mathbf{q}\right) = M^{-1}\mathbf{q} = \mathbf{x}^\star$
- Lemma 10.2 Newton's method is Affine invariant. Proof is straightforward.
- Lemma 10.3 If $\nabla^2 f\left(\mathbf{x}_t\right) \succ \mathbf{0}$, then $\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\operatorname{argmin}} f\left(\mathbf{x}_t\right) + \nabla f(\mathbf{x}_t)^\top\left(\mathbf{x} - \mathbf{x}_t\right) + \frac{1}{2}\left(\mathbf{x} - \mathbf{x}_t\right)^\top\nabla^2 f\left(\mathbf{x}_t\right)\left(\mathbf{x} - \mathbf{x}_t\right)$.

Super Linear (quadratic) Local Convergence ($\mathcal{O}(\log\log(1/\varepsilon))$

- Theorem 10.4 $f \in C^2$ with critical point $\mathbf{x}^\star$. Assume $\exists U(\mathbf{x}^\star) \subseteq \mathbf{dom}(f)$ s.t. (i) *bounded hessian* $\|\nabla^2 f(\mathbf{x})^{-1}\| \leq \frac{1}{\mu}, \quad \forall \mathbf{x} \in X$ (ii) *Lipschitz continuous Hessian* $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq B\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X$, then $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{B}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$.
  - Proof
    - $\mathbf{x}' - \mathbf{x}^\star = \mathbf{x} - \mathbf{x}^\star + H(\mathbf{x})^{-1}\int_0^1 H\left(\mathbf{x} + t\left(\mathbf{x}^\star - \mathbf{x}\right)\right)\left(\mathbf{x}^\star - \mathbf{x}\right)\mathrm{d}t$
    - also $\mathbf{x} - \mathbf{x}^\star = H(\mathbf{x})^{-1}H(\mathbf{x})\left(\mathbf{x} - \mathbf{x}^\star\right) = H(\mathbf{x})^{-1}\int_0^1 -H(\mathbf{x})\left(\mathbf{x}^\star - \mathbf{x}\right)\mathrm{d}t$, we subtract the above to the first equation, and get
    - $\mathbf{x}' - \mathbf{x}^\star = H(\mathbf{x})^{-1}\int_0^1 \left(H\left(\mathbf{x} + t\left(\mathbf{x}^\star - \mathbf{x}\right)\right) - H(\mathbf{x})\right)\left(\mathbf{x}^\star - \mathbf{x}\right)\mathrm{d}t$
    - Then by some inequality w.r.t. norm operation, $\|\mathbf{x}' - \mathbf{x}^\star\| \leq \|H(\mathbf{x})^{-1}\| \cdot \|\int_0^1 \left(H\left(\mathbf{x} + t\left(\mathbf{x}^\star - \mathbf{x}\right)\right) - H(\mathbf{x})\right)\left(\mathbf{x}^\star - \mathbf{x}\right)\mathrm{d}t\|$
      $\leq \|H(\mathbf{x})^{-1}\| \cdot \int_0^1 \|\left(H\left(\mathbf{x} + t\left(\mathbf{x}^\star - \mathbf{x}\right)\right) - H(\mathbf{x})\right)\left(\mathbf{x}^\star - \mathbf{x}\right)\|\mathrm{d}t$
      $\leq \|\left(\mathbf{x}^\star - \mathbf{x}\right)\| \cdot \|H(\mathbf{x})^{-1}\| \cdot \int_0^1 \|\left(H\left(\mathbf{x} + t\left(\mathbf{x}^\star - \mathbf{x}\right)\right) - H(\mathbf{x})\right)\|\mathrm{d}t = \frac{B}{2\mu}\|\left(\mathbf{x}^\star - \mathbf{x}\right)\|^2$
  - (*My own strange derivation, not successful*) By Lipschitz, using Lemma B, taking $y = x_{t+1}$ and $x = x_t$, we get $\|\nabla f(x_{t+1})\|^{-1} \leq \frac{B}{2\mu^2}\|\nabla f(x_t)\|^2$, this gives a similar relation w.r.t. t, but not on $\|x_t - x^\star\|$.
- Corollary 10.5 (Exercise 64) If starting points satisfies $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq \frac{\mu}{B}$, then $\|\mathbf{x}_T - \mathbf{x}^\star\| \leq \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T - 1}$
  - Proof
    - $\log\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \log\frac{B}{2\mu} + 2\log\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|$ or $\log\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| + \log\frac{B}{2\mu} \leq 2(\log\|\mathbf{x}_t - \mathbf{x}^\star\| + \log\frac{B}{2\mu})$
    - When the starting condition holds, $\log\|\mathbf{x}_0 - \mathbf{x}^\star\| + \log\frac{B}{2\mu} \leq -\log 2 \leq 0$, so every t, $\log\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| + \log\frac{B}{2\mu} \leq 0$,
    - so $\log\|\mathbf{x}_T - \mathbf{x}^\star\| + \log\frac{B}{2\mu} \leq -2^T\log 2$, $\|\mathbf{x}_T - \mathbf{x}^\star\| \leq \frac{2\mu}{B}(\frac{1}{2})^{2^T} = \frac{\mu}{B}(\frac{1}{2})^{2^T-1}$
- Lemma 10.6 (Exercise 65) If starting point satisfies $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq \frac{\mu}{B}$, then $\frac{\|\nabla^2 f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}^\star)\|}{\|\nabla^2 f(\mathbf{x}^\star)\|} \leq \left(\frac{1}{2}\right)^{2^t - 1}$.
  - Proof just use the two assumption.

Global analysis for strongly-convex smooth objectives

- Lemma A If f has L_2-Lipschitz Hessian w.r.t. some norm, then
  $|f(\mathbf{y}) - f\left(\mathbf{x}\right) - \nabla f(\mathbf{x})^\top\left(\mathbf{y} - \mathbf{x}\right) - \frac{1}{2}\left(\mathbf{y} - \mathbf{x}\right)^\top\nabla^2 f\left(\mathbf{x}\right)\left(\mathbf{y} - \mathbf{x}\right)| \leq \frac{L_2}{6}\|\mathbf{y} - \mathbf{x}\|_a^3$
  - Proof
    - L_2-Lipschitz Hessian means $\|H(\mathbf{y}) - H(\mathbf{y})\|_{a,a*} \leq \|\mathbf{y} - \mathbf{x}\|_{a,a*}$, where matrix norm is $\|A\|_{a,a*} = \mathbf{sup}_{x\neq 0}\frac{\|Ax\|_{a*}}{\|x\|_a}$, and $\|\cdot\|_{a*}$ is the dual norm.
    - Define $g(t \in [0,1]) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, and $\mathbf{z}_t := \mathbf{x} + t(\mathbf{y} - \mathbf{x})$, $g'(t) = \nabla f(z)^\top(\mathbf{y} - \mathbf{x})$,
      $g''(t) := (\mathbf{y} - \mathbf{x})^\top H(\mathbf{z}_t)(\mathbf{y} - \mathbf{x})$
    - Then $|g''(t) - g''(0)| = |(\mathbf{y} - \mathbf{x})^\top(H(\mathbf{z}_t) - H(\mathbf{x}))(\mathbf{y} - \mathbf{x})| \leq \|(H(\mathbf{z}_t) - H(\mathbf{x}))(\mathbf{y} - \mathbf{x})\|_{a*}\|\mathbf{y} - \mathbf{x}\|_a$
    - **RHS** $\leq \|H(\mathbf{z}_t) - H(\mathbf{x})\|_{a,a*}\|\mathbf{y} - \mathbf{x}\|_a^2$, by Lipschitz $\leq L_2 t\|\mathbf{y} - \mathbf{x}\|_a^3$.
    - Then $g'(t) - g'(0) = \int_0^t g''(t)d\tau \leq \int_0^t g''(0) + \tau\|\mathbf{y} - \mathbf{x}\|_a^3 d\tau = g''(0)t + \frac{L_2}{2}\|\mathbf{y} - \mathbf{x}\|_a^3 t^2$,
      - also $g'(t) - g'(0) \geq \int_0^t g''(0) - \tau\|\mathbf{y} - \mathbf{x}\|_a^3 d\tau = g''(0)t - \frac{L_2}{2}\|\mathbf{y} - \mathbf{x}\|_a^3 t^2$
    - Then $g(t) - g(0) = \int_0^t g'(\tau)d\tau \leq \int_0^t g'(0) + g''(0)\tau + \frac{L_2}{2}\|\mathbf{y} - \mathbf{x}\|_a^3\tau^2 d\tau = g'(0)t + \frac{1}{2}g''(0)t^2 + \frac{L_2}{6}\|\mathbf{y} - \mathbf{x}\|_a^3 t^3$
      - $g(t) - g(0) \geq \int_0^t g'(0) + g''(0)\tau - \frac{L_2}{2}\|\mathbf{y} - \mathbf{x}\|_a^3\tau^2 d\tau = g'(0)t + \frac{1}{2}g''(0)t^2 - \frac{L_2}{6}\|\mathbf{y} - \mathbf{x}\|_a^3 t^3$
    - Setting $t = 1$, and we get $|f(\mathbf{x}) - f\left(\mathbf{x}_t\right) - \nabla f(\mathbf{x}_t)^\top\left(\mathbf{x} - \mathbf{x}_t\right) - \frac{1}{2}\left(\mathbf{x} - \mathbf{x}_t\right)^\top\nabla^2 f\left(\mathbf{x}_t\right)\left(\mathbf{x} - \mathbf{x}_t\right)| \leq \frac{L_2}{6}\|\mathbf{x} - \mathbf{x}_t\|_a^3$
- Lemma B If f has L_2-Lipschitz Hessian w.r.t. some norm, then $\|\nabla f(\mathbf{y}) - \nabla f\left(\mathbf{x}\right) - \nabla^2 f\left(\mathbf{x}\right)\left(\mathbf{y} - \mathbf{x}\right)\|_{a*} \leq \frac{L_2}{2}\|\mathbf{y} - \mathbf{x}\|_a^2$
  - Proof
    - Similar to Lemma A, but define $\mathbf{k}(t) := \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, then $\partial_t\mathbf{k} = \nabla^2 f_t\left(\mathbf{y} - \mathbf{x}\right)$,
    - So we have $\|\partial_t\mathbf{k}(t) - \partial_t\mathbf{k}(0)\|_{a*} \leq L_2 t\|\mathbf{y} - \mathbf{x}\|_a^2$.
    - By triangular ineq, $\|\int\|_{a*} \leq \int\|\|_{a*}$, so we get similar result.

- Lemma C Assume f μ-convex and has L-Lipschitz continuous gradient, then Newton's method $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$ enjoys global linear convergence $f(\mathbf{x}_t) - f^* \leq \left(1 - \frac{\mu^2}{L^2}\right)^t (f(\mathbf{x}_0) - f^*)$ given $\gamma = \frac{\mu}{L}$.
  - Proof
    - By smoothness $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \left\langle -\gamma H_t^{-1} \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t) \right\rangle + \frac{L\gamma^2}{2} \| H_t^{-1} \nabla f(\mathbf{x}_t) \|^2$,
    - By strong convexity and smoothness $\frac{1}{L} \leq \| H_t^{-1} \| \leq \frac{1}{\mu}$, and by the fact of H_t beging symmetric and so is its inverse, $\| H_t^{-1} \nabla f(\mathbf{x}_t) \|^2 = (H_t^{-1/2} \nabla f(\mathbf{x}_t))^\top H_t^{-1} (H_t^{-1/2} \nabla f(\mathbf{x}_t)) \leq (H_t^{-1/2} \nabla f(\mathbf{x}_t))^\top \frac{1}{\mu} (H_t^{-1/2} \nabla f(\mathbf{x}_t))$ $= \langle H_t^{-1} \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t) \rangle / \mu$,
    - Then RHS $\leq \left(-\gamma + \frac{L\gamma^2}{2\mu}\right) \left\langle H_t^{-1} \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t) \right\rangle$, when $\gamma = \frac{\mu}{L}$ this term is minimized, and this term is RHS $= -\frac{\mu}{2L} \left\langle H_t^{-1} \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t) \right\rangle$
    - Again, by spectual lower bound of H_t^{-1}, $\left\langle H_t^{-1} \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t) \right\rangle \geq \frac{1}{L} \| \nabla f(\mathbf{x}_t) \|^2$, so RHS $\leq -\frac{\mu}{2L} \cdot \frac{1}{L} \| \nabla f(\mathbf{x}_t) \|^2$
    - Due to strong convexity, $f(z) - f(x^\star) \leq \frac{1}{2\mu} \| \nabla f(z) \|_*^2$, so $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{\mu^2}{L^2} (f(\mathbf{x}_t) - f^*)$

Methods to overcoming the local nature of Newton method

- *Newton method with line-search* select γ_t s.t. $f(\mathbf{x}_{t+1}) < f(\mathbf{x}_t)$ with sufficient decrease, $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$.
- *Dampled Newton method* $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1+\lambda_f(\mathbf{x}_t)} \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$ where $\lambda_f(\mathbf{x}) = \| \left[\nabla^2 f(\mathbf{x})\right]^{-1/2} \nabla f(\mathbf{x}) \|$,
  - we can show that $\lambda_f(\mathbf{x})/2 = f(x) - \mathbf{min}_y \{ f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2} (y - x)^\top \nabla^2 f(x) (y - x) \}$ is approximately twice the decrease of value using normal Newton's method.
  - If previous step have a lot decrease in f, which is appproximated by λ, then we compensate in the next iteration to have less.
  - This is guaranteed to converge.
- *Regularization approach* regularize the Hessian and adjust γ_t, $\mathbf{x}_{t+1} = \mathbf{x}_t - \left[\gamma_t I + \nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)$.
  - When γ is large, this is approximately GD.
- *Trust-region approach* $\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\mathbf{argmin}}\, f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t) (\mathbf{x} - \mathbf{x}_t)$, s.t. $\mathbf{s.t.}\ \| \mathbf{x} - \mathbf{x}_t \| \leq \Delta_k$, not to move too far away. similar to regularization approach.

Cubic Regularization (Nesterov & Polyak, 2006)

- motivation GD can be viewed as iteratively minimizing the quadratic upper bound function $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L_1}{2} \| \mathbf{y} - \mathbf{x} \|^2$.
- ALgorithm (*Subproblem*) $\mathbf{x}_{t+1} \in \underset{\mathbf{x}}{\mathbf{argmin}}\, \hat{f}(\mathbf{x}, \mathbf{x}_t)$, where $\hat{f}(\mathbf{x}, \mathbf{x}_t) := f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t) (\mathbf{x} - \mathbf{x}_t) + \frac{M}{6} \| \mathbf{x} - \mathbf{x}_t \|_2^3$
- Subproblem can be reduced to a convex problem, and can also be solved directly by GD to global optimal.
- No invertion of Hessian is needed.

- Lemma D (Graded assignment 4.2) Given two problem $u(\mathbf{h}) = \mathbf{g}^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top H \mathbf{h} + \frac{\mathbf{M}}{6} \| \mathbf{h} \|^3$ and $v(r) = -\frac{1}{2} \mathbf{g}^\top \left(H + \frac{Mr}{2} I_d\right)^{-1} \mathbf{g} - \frac{M}{12} r^3$, where $\mathcal{D} = \{ r \geq 0 \mid H + \frac{Mr}{2} I_d \succ 0 \}$, then $\inf_{\mathbf{h} \in \mathbb{R}^d} u(\mathbf{h}) = \sup_{r \in \mathcal{D}} v(r)$
  - Proof
    - First we prove $\inf_{\mathbf{h}} u(\mathbf{h}) \geqslant \sup_r v(r)$
      - First by adding and subtracting the same term $\frac{1}{2} \mathbf{h}^\top (\frac{1}{2} M r I_d) \mathbf{h}$, we have
        - $u(\mathbf{h}) = \mathbf{g}^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top H \mathbf{h} + \frac{1}{2} \mathbf{h}^\top (\frac{1}{2} M r I_d) \mathbf{h} - \frac{1}{2} \mathbf{h}^\top (\frac{1}{2} M r I_d) \mathbf{h} + \frac{\mathbf{M}}{6} \| \mathbf{h} \|^3$ $= \underline{\mathbf{g}^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \left(H + \frac{1}{2} M r I_d\right) \mathbf{h}} + \frac{\mathbf{M}}{6} \| \mathbf{h} \|^3 - \frac{Mr}{4} \| \mathbf{h} \|^2$
      - Now we write the underlined quadratic form into canonical form, $\mathbf{g}^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \left(H + \frac{1}{2} M r I_d\right) \mathbf{h} = \frac{1}{2} (\mathbf{h} - h(r))^\top \left(H + \frac{1}{2} M r I_d\right) (\mathbf{h} - h(r)) - \frac{1}{2} \mathbf{g}^\top \left(H + \frac{Mr}{2} I_d\right)^{-1} \mathbf{g}$, where $h(r) = -\left(H + \frac{Mr}{2} I_d\right)^{-1} g$
      - Plug this in and we get $u(\mathbf{h}) = \frac{1}{2} (\mathbf{h} - h(r))^\top \left(H + \frac{1}{2} M r I_d\right) (\mathbf{h} - h(r)) + \underbrace{\left[-\frac{1}{2} \mathbf{g}^\top \left(H + \frac{Mr}{2} I_d\right)^{-1} \mathbf{g} - \frac{M}{12} r^3 \right]}_{v(r)} + \frac{\mathbf{M}}{6} \| \mathbf{h} \|^3 - \frac{Mr}{4} \| \mathbf{h} \|^2 + \frac{M}{12} r^3$ $= v(r) + \frac{1}{2} (\mathbf{h} - h(r))^\top \left(H + \frac{1}{2} M r I_d\right) (\mathbf{h} - h(r)) + \frac{M}{12} (\| \mathbf{h} \| - r)^2 (r + 2 \| \mathbf{h} \|)$
      - By definition, the last two term are both non-negative, so we prove $\inf_{\mathbf{h}} u(\mathbf{h}) \geqslant \sup_r v(r)$.
    - Then we prove when $\mathbf{h} = h(r)$, equality holds.
      - In this circumstance, the second term is zero, so $u(h(r)) - v(r) = \frac{M}{12} (\| h(r) \| - r)^2 (r + 2 \| h(r) \|)$

    - Then we consider the analytical expression of $v(r)$, we expend w.r.t eigen values of H, $H = Q\mathbf{diag}\{\lambda_i\}_{i=1}^d Q^\top$, where $Q = [\mathbf{q}_1, \dots, \mathbf{q}_d]$ is an orthonormal matrix. Then $\frac{1}{2}\mathbf{g}^\top \left(\mathbf{H} + \frac{Mr}{2} I_d\right)^{-1} \mathbf{g} = \frac{1}{2}\mathbf{g}^\top Q \mathbf{diag}\left\{\frac{1}{Mr/2+\lambda_i}\right\}_{i=1}^d Q^\top \mathbf{g} = \frac{1}{2}\sum_{i=1}^d \frac{(\mathbf{q}_i^\top \mathbf{g})^2}{Mr/2+\lambda_i}$.
    - Then we have the derivatives of $v(r)$,
        - (i) $v(r) = -\frac{1}{2}\sum_{i=1}^d \frac{(\mathbf{q}_i^\top \mathbf{g})^2}{Mr/2+\lambda_i} - \frac{M}{12} r^3$,
        - (ii) $v'(r) = \frac{M}{2}\frac{1}{2}\sum_{i=1}^d \frac{(\mathbf{q}_i^\top \mathbf{g})^2}{(Mr/2+\lambda_i)^2} - \frac{M}{4} r^2 = \frac{M}{2}\frac{1}{2}\mathbf{g}^\top \left(H + \frac{Mr}{2} I_d\right)^{-2} \mathbf{g} - \frac{M}{4} r^2 = \frac{M}{4}\left[\left\|\left(H + \frac{Mr}{2} I_d\right)^{-1} \mathbf{g}\right\|^2 - r^2\right]$ $= \frac{M}{4}(\|h(r)\|^2 - r^2)$
        - (iii) $v''(r) = -\frac{M^2}{8}\sum_{i=1}^d \frac{(\mathbf{q}_i^\top \mathbf{g})^2}{(Mr/2+\lambda_i)^3} - \frac{M}{2} r$
    - By definition $H + \frac{Mr}{2} I_d \succ \mathbf{0}$, so that $Mr/2 + \boldsymbol{\lambda}_i > \mathbf{0}$, and since $\|\mathbf{g}\| \neq \mathbf{0}$ (otherwise trivial), this means $v''(r) < \mathbf{0}$, strictly concave, local maximual is global.
    - The optimial condition holds when $\|h\left(r^*\right)\| = r^*$, when this happens $u(h(r^*)) - v(r^*) = u(h(r^*)) - \sup_r v(r) = \frac{M}{12}(\|h(r^*)\| - r^*)^2(r^* + 2\|h(r^*)\|) = \mathbf{0}$. QED
  - Remark $v(r)$ is a convex program (while $u(\mathbf{h})$ is not), and can be solved by GD.

- Key Facts 1 Second order $\nabla^2 f\left(\mathbf{x}_t\right) + \frac{M}{2}\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\| \cdot I \succeq 0$.
  - Proof
    - $\mathbf{x}_{t+1} - \mathbf{x}_t = \mathbf{h} = h(r^\star)$ and $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 = \|h(r^\star)\|_2 = r^\star$, so $\nabla^2 f\left(\mathbf{x}_t\right) + \frac{M}{2}\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\| = H + \frac{Mr^\star}{2} I \succeq 0$ by definition of domain $\mathcal{D}$.

- Key Facts 2 First order $\left\|\nabla f\left(\mathbf{x}_{t+1}\right)\right\| \leq \frac{L_2+M}{2}\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\|^2$.
  - Proof 2
    - By the first order optimality condition we have $\nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{M}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 \cdot (\mathbf{x}_{t+1} - \mathbf{x}_t) = 0$
    - By Lemma B, $\|\nabla f(\mathbf{x}_{t+1}) - (f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t))\|_2 \leq \frac{L_2}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$,
    - By triangle inequality, we prove the result, $\|\nabla f(\mathbf{x}_{t+1})\|_2 \leq \|\nabla f(\mathbf{x}_{t+1}) - (f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t))\|_2 + \|f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)\|_2$.

- Key Facts 3 Zero order $f\left(\mathbf{x}_t\right) - f\left(\mathbf{x}_{t+1}\right) \geq \frac{M}{12}\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\|^3$ if $M \geq L_2$
  - Proof
    - By L_2-Lipschitz, $f(x) - \left(f\left(\mathbf{x}_t\right) + \nabla f(\mathbf{x}_t)^\top\left(\mathbf{x} - \mathbf{x}_t\right) + \frac{1}{2}\left(\mathbf{x} - \mathbf{x}_t\right)^\top \nabla^2 f\left(\mathbf{x}_t\right)\left(\mathbf{x} - \mathbf{x}_t\right)\right) \leq \frac{L_2}{6}\left\|\mathbf{x} - \mathbf{x}_t\right\|_2^3$
    - so taking $\mathbf{x} = \mathbf{x}_{t+1}$, we have
      - $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \geq -\frac{L_2}{6} r^3 - \mathbf{g}^\top \mathbf{h} - \frac{1}{2}\mathbf{h}^\top H \mathbf{h} = \frac{M - L_2}{6} r^3 - u(\mathbf{h}) = \frac{M-L_2}{6} r^3 - v(r)$
    - Since $M > L_2$, we have $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \geq -v(r) = \frac{1}{2}\mathbf{g}^\top \left(H + \frac{Mr}{2} I_d\right)^{-1} \mathbf{g} + \frac{M}{12} r^3 \geq \frac{M}{12} r^3$

- Implication 1 If $\mathbf{x}^\star$ is limiting point, then $\nabla f\left(\mathbf{x}^*\right) = 0, \nabla^2 f\left(\mathbf{x}^*\right) \succeq 0$, since $r = 0$, by Fact 1, $\nabla^2 f\left(\mathbf{x}^\star\right) \succeq 0$, by Fact 2, $\left\|\nabla f\left(\mathbf{x}^\star\right)\right\| = 0$.

- Implication 2 Convergence rate of $\min_{1 \leq i \leq t}\left\|\nabla f\left(\mathbf{x}_i\right)\right\| = O\left(\frac{1}{t^{2/3}}\right)$, since $f\left(\mathbf{x}_t\right) - f\left(\mathbf{x}_{t+1}\right) \geq \frac{M}{12}\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\|^3 \geq C \cdot \|\nabla f(\mathbf{x}_{t+1})\|_2^{3/2}$
  - then do sumation, $f\left(\mathbf{x}_1\right) - f\left(\mathbf{x}^\star\right) \geq C\sum_{i=1}^t \|\nabla f(\mathbf{x}_i)\|_2^{3/2} \geq C \cdot t[\min_i \|\nabla f(\mathbf{x}_i)\|_2]^{3/2}$

- Implication 3 If f convex, $f\left(\mathbf{x}_t\right) - f^* = O\left(\frac{1}{t^2}\right)$.
  - $f(\mathbf{x}_t) - f^\star \leq \|\nabla f(\mathbf{x}_t)\| \cdot \|\mathbf{x}_t - \mathbf{x}^\star\| \leq \|\nabla f(\mathbf{x}_t)\| \cdot \|\mathbf{x}_0 - \mathbf{x}^\star\|$ (*unproved)*, then $f\left(\mathbf{x}_t\right) - f\left(\mathbf{x}_{t+1}\right) \geq C \cdot \|\nabla f(\mathbf{x}_{t+1})\|_2^{3/2} \geq C'(\Delta f_{t+1})^{3/2}$. This gives $\Delta_t - \Delta_{t+1} \geq C\Delta_{t+1}^{3/2}$.
  - This means $\frac{1}{\sqrt{\Delta_{t+1}}} - \frac{1}{\sqrt{\Delta_t}} \geq C\frac{\Delta_t}{\sqrt{\Delta_{t+1}}(\sqrt{\Delta_t} + \sqrt{\Delta_{t+1}})} = \frac{C}{\sqrt{\frac{\Delta_{t+1}}{\Delta_t}}(1 + \sqrt{\frac{\Delta_{t+1}}{\Delta_t}})}$
  - Since $0 \leq \sqrt{\frac{\Delta_{t+1}}{\Delta_t}} \leq 1$, $\sqrt{\frac{\Delta_{t+1}}{\Delta_t}}(1 + \sqrt{\frac{\Delta_{t+1}}{\Delta_t}}) \leq 2$, so $\frac{1}{\sqrt{\Delta_{t+1}}} - \frac{1}{\sqrt{\Delta_t}} \geq C/2$ and $\frac{1}{\sqrt{\Delta_t}} - \frac{1}{\sqrt{\Delta_0}} \geq Ct/2$
  - $\Delta_t = O(1/t^2)$