1. Property of Smooth Functions

$f \in C^1(\mathbb{R}^d)$ $\quad \|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$, $x, y \in \mathbb{R}^d$

Prove $f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|^2$

Proof. let $g(t) = f(x+t(y-x))$, $t \in [0,1]$

$g(0) = f(x)$, $g(1) = f(y)$

$g'(t) = \langle \nabla f(x+t(y-x)), (y-x) \rangle$ by chain rule.

Thus $g \in C[0,1]$ and $g' \in C(0,1)$.

We need to prove

$$g(1) - g(0) \le g'(0) + \frac{L}{2}\|y-x\|^2$$

In fact $g(1) - g(0) = \int_0^1 g'(t)\,dt = \int_0^1 \langle \nabla f(x+t(y-x)), y-x \rangle\, dt$

$$g(1) - g(0) - g'(0) = \int_0^1 \langle \nabla f(x_t) - \nabla f(x), y-x \rangle\, dt, \quad x_t = x+t(y-x)$$

$$\le \int_0^1 \|\nabla f(x_t) - \nabla f(x)\|\,\|y-x\|\, dt \quad \text{(Cauchy inequality)}$$

$$\le L\int_0^1 t\|y-x\|^2\, dt = \frac{L\|y-x\|^2}{2}$$

We are done. #

2 Convergence of GF under KL condition

$\inf_x f(x) = 0$, $\|\nabla f(x)\|^2 \geq \mu f(x)^\alpha$ $(x_t)_{t\geq 0}$ be GF solution

(a) $\alpha > 1$,

$$f(x_t) \leq \frac{1}{(f(x_0)^{1-\alpha} + \mu(\alpha-1)t)^{\frac{1}{\alpha-1}}} \sim t^{-\frac{1}{\alpha-1}}$$

Proof. If there exist $0 \leq s \leq t$ s.t. $x_s = 0$

Then from $\dot{x}_t = -\nabla f(x_t)$, $x_t \equiv 0$ for $t \geq s$ we are done.

Otherwise, $x_s > 0$, $0 \leq s \leq t$

We have $\frac{df(x_t)}{dt} = -\|\nabla f(x_t)\|^2$

$$f(x_t)^{1-\alpha} - f(x_0)^{1-\alpha} = \int_0^t (f(x_r)^{1-\alpha})'_r \, dr$$

$$= \int_0^t (\alpha-1) f(x_r)^{-\alpha} \|\nabla f(x_r)\|^2 dt$$

$$\geq \int_0^t \mu(\alpha-1) dt = \mu(\alpha-1)t$$

Therefore $f(x_t)^{1-\alpha} \geq f(x_0)^{1-\alpha} + \mu(\alpha-1)t$

$$f(x_t) \leq \frac{1}{(f(x_0)^{1-\alpha} + \mu(\alpha-1)t)^{\frac{1}{\alpha-1}}} \quad (\text{As } \alpha > 1)$$

#

(b) $\alpha < 1$, $f(x_t) \le \left(f(x_0)^{1-\alpha} - \mu(1-\alpha)t\right)^{\frac{1}{1-\alpha}}$, $\forall t < \frac{f(x_0)^{1-\alpha}}{\mu(1-\alpha)}$

<u>Proof.</u> If there exists $0 \le s \le t$, $f(x_s) = 0$

then $\dot{x}_t = -\nabla f(x_t)$ implies $f(x_t) = 0$

with $t < \frac{f(x_0)^{1-\alpha}}{\mu(1-\alpha)}$ we are done. Otherwise $f(x_s) > 0$, $0 \le s \le t$,

we have $f(x_t)^{1-\alpha} - f(x_0)^{1-\alpha}$

$$= \int_0^t \left(f(x_r)^{1-\alpha}\right)_r' \, dt$$

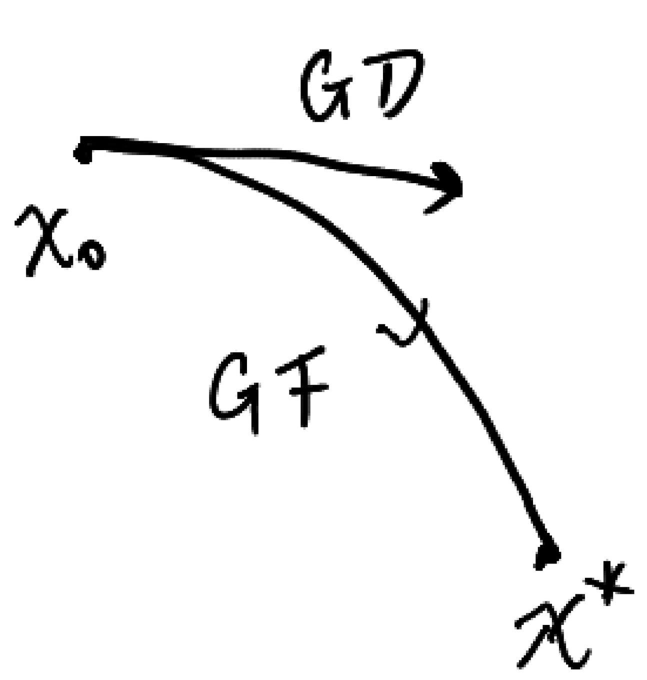
$$= -\int_0^t (1-\alpha) f(x_r)^{-\alpha} \|\nabla f(x_r)\|^2 dt$$

$$\le -\int_0^t (1-\alpha)\mu \, dt = -\mu(1-\alpha)t$$

$$f(x_t)^{1-\alpha} \le f(x_0)^{1-\alpha} - \mu(1-\alpha)t$$

$$f(x_t) \le \left(f(x_0)^{1-\alpha} - \mu(1-\alpha)t\right)^{\frac{1}{1-\alpha}} \quad \#$$

(c) Because finite-time convergence of GF means after a finite time T, going on the trajectory of GF will reach minima



But finite-step GD goes on the <u>tangent line</u> of trajectory of GF, will probably reach different point after T.

From another perspective, suppose learning rate is $\{\eta_t\}_0^{T_1}$

To let GF approximate GD, $\eta_t \ll 1$

but $T \approx \sum_{t=0}^{T_1} \eta_t$ This implies $T_1 \to \infty$

3 Implicit bias of GD for linear regression

$$\hat{R}(\beta) = \frac{1}{2}\|X\beta - y\|_2^2$$

$X \in \mathbb{R}^{n\times d}$ $\beta \in \mathbb{R}^d$ $y \in \mathbb{R}^n$ $d > n$, $\min_\beta \hat{R}(\beta) = 0$

Optimize with GD

$$\beta_0 = 0$$
$$\beta_{t+1} = \beta_t - \eta \nabla \hat{R}(\beta_t)$$

Prove $\lim_{t\to\infty} \beta_t = \bar{\beta}$, $\bar{\beta} := \arg\min_\beta \|\beta\|_2$ s.t. $X\beta = y$.

<u>Proof</u>. $\hat{R}(\beta) = \frac{1}{2}(\beta^T X^T - y^T)(X\beta - y)$

$$= \frac{1}{2}\beta^T X^T X \beta - (X^T y)^T \beta + \frac{1}{2} y^T y$$

$$\nabla_\beta \hat{R}(\beta) = X^T X \beta - X^T y$$

Dynamics is $\beta_{t+1} = \beta_t - \eta(X^T X \beta_t - X^T y)$

$$\beta_{t+1} = (I_d - \eta X^T X)\beta_t + \eta X^T y$$

$$\beta_{t+1} - \bar{\beta} = (I_d - \eta X^T X)(\beta_t - \bar{\beta})$$

$$\beta_t - \bar{\beta} = -(I_d - \eta X^T X)^t \bar{\beta}$$

We analyze spectrum of $I_d - \eta X^T X$.

Consider SVD decomposition of X

$$X = T \Sigma S, \quad T \in O(n), \ S \in O(d), \quad \Sigma = \begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \ 0 \end{pmatrix}, \quad \sigma_i > 0, \ r \le n$$

$$X^T X = S^T \Sigma^T T^T T \Sigma S = S^T \Sigma^T \Sigma S$$

$$= S^T \begin{pmatrix} \sigma_1^2 & & & \\ & \ddots \ \sigma_r^2 & & \\ & & 0 & \\ & & & \ddots \ 0 \end{pmatrix} S$$

$$I_d - \gamma X^T X = S^T I_d S - \gamma S^T \begin{pmatrix} \sigma_1^2 & & & \\ & \ddots \ \sigma_r^2 & & \\ & & 0 & \\ & & & \ddots \ 0 \end{pmatrix} S$$

$$= S^T \begin{pmatrix} 1-\gamma\sigma_1^2 & & & \\ & \ddots \ 1-\gamma\sigma_r^2 & & \\ & & 0 & \\ & & & \ddots \ 0 \end{pmatrix} S$$

$$\beta_t - \bar{\beta} = - S^T \begin{pmatrix} (1-\gamma\sigma_1^2)^t & & & \\ & \ddots \ (1-\gamma\sigma_r^2)^t & & \\ & & 1 & \\ & & & \ddots \ 1 \end{pmatrix} S \bar{\beta} \qquad \text{For } 0 < \gamma < \min_i \frac{1}{\sigma_i^2},$$

$$\lim_{t \to \infty} \beta_t - \bar{\beta} = - S^T \begin{pmatrix} 0 & \\ & I_{d-r} \end{pmatrix} S \bar{\beta}$$

It is well known $\bar{\beta}$ admits the form $\bar{\beta} = S^T \begin{pmatrix} \frac{1}{\sigma_1} & & & \\ & \ddots \ \frac{1}{\sigma_r} & & \\ & & 0 & \\ & & & \ddots \ 0 \end{pmatrix} T^T y$ from linear algebra

Thus

$$-S^T \begin{pmatrix} 0 & \\ & I_{d-r} \end{pmatrix} S \bar{\beta} = - \underset{d\times d}{S^T} \underset{d \times d}{\begin{pmatrix} 0 & \\ & I_{d-r} \end{pmatrix}} \underset{d \times n}{\begin{pmatrix} \frac{1}{\sigma_1} & & & \\ & \ddots \ \frac{1}{\sigma_r} & & \\ & & 0 & \\ & & & \ddots \ 0 \end{pmatrix}} \underset{n \times 1}{T^T y}$$

$$= - S^T 0 \, T^T y = 0$$

We are done #

4 The dynamic behavior of HB method

(a) $f(x) = \frac{1}{2}hx^2 \quad \nabla f(x) = hx$

Dynamics is $\ddot{x}_t = -\gamma \dot{x}_t - h x_t$ This is second-order constant coefficient linear ODE.

characteristic equation

$$\lambda^2 + \gamma\lambda + h = 0$$

Case 1. $\gamma^2 > 4h$

$$x_t = A_1 e^{\frac{-\gamma+\sqrt{\gamma^2-4h}}{2}t} + A_2 e^{\frac{-\gamma-\sqrt{\gamma^2-4h}}{2}t}$$

for $\gamma>0, h>0$. $x_t \to 0$ with rate $O(e^{\frac{-\gamma+\sqrt{\gamma^2-4h}}{2}t})$

Other (γ, h) leads to diverges $(x_t \to \infty)$

Case 2. $\gamma^2 = 4h$

$$x_t = (A_1 t + A_2) e^{\frac{-\gamma}{2}t}$$

for $\gamma>0$, $x_t \to 0$ with rate $O(t e^{-\frac{\gamma}{2}t})$, not monotonic

for $\gamma \le 0$ diverges $(x_t \to \infty)$

Case 3. $\gamma^2 < 4h$

$$x_t = A_1 e^{-\frac{\gamma}{2}t} \cos\left(\frac{\sqrt{4h-\gamma^2}}{2}t\right) + A_2 e^{-\frac{\gamma}{2}t} \sin\left(\frac{\sqrt{4h-\gamma^2}}{2}t\right)$$

for $\gamma>0$, $x_t \to 0$ with rate $O(e^{-\frac{\gamma}{2}t})$, not monotonic

for $\gamma \le 0$ diverges

where above, A_1, A_2 are constants depending on initializations.

(b) See last page of this PDF

5 Convergence of HB method for convex problems

$$\ddot{x}_t = -\dot{x}_t - \nabla f(x_t), \quad f \text{ convex}, \quad x^* = \min_x f(x)$$

$$V(t) = f(x_t) - f(x^*) + \frac{1}{2}\|x_t - x^* + \dot{x}_t\|^2$$

(a) $\dot{V}(t) \le f(x^*) - f(x_t)$

Proof.

$$\begin{aligned} \dot{V}(t) &= \langle \nabla f(x_t), \dot{x}_t \rangle + \langle x_t - x^* + \dot{x}_t, \dot{x}_t + \ddot{x}_t \rangle \\ &= \langle \nabla f(x_t), \dot{x}_t \rangle + \langle -\nabla f(x_t), x_t - x^* + \dot{x}_t \rangle \\ &= \langle \nabla f(x_t), x^* - x_t \rangle \\ &\le f(x^*) - f(x_t) \quad \text{by convexity of } f \quad \# \end{aligned}$$

(b) $\int_0^T (f(x_t) - f(x^*))\,dt \le V(0)$

Proof. Because of (a),

$$\int_0^T (f(x_t) - f(x^*))\,dt \le -\int_0^T \dot{V}(t)\,dt = V(0) - V(T)$$

$$\le V(0) \quad \left(\text{as } V(T) = f(x_T) - \min_x f(x) + \frac{1}{2}\|x_T - x^* + \dot{x}_T\|^2 \ge 0\right)$$

(c) $\bar{x}_T := \frac{1}{T}\int_0^T x_t\,dt$ Show $f(\bar{x}_T) - f(x^*) \le \frac{V(0)}{T}$

Proof $f(\bar{x}_T) = f\left(\frac{1}{T}\int_0^T x_t\,dt\right) \le \frac{1}{T}\int_0^T f(x_t)\,dt$ (*)

$$\le \frac{1}{T}\left(\int_0^T f(x^*)\,dt + V(0)\right) \quad (\text{because of (b)})$$

i.e. $f(\bar{x}_T) - f(x^*) \le \frac{V(0)}{T}$ #

((*): Random vector X realized by uniformly pick $t \in [0, T]$, and take x_t
$\mathbb{E}X = \frac{1}{T}\int_0^T x_t\,dt$ by Jensen inequality $f(\mathbb{E}X) \le \mathbb{E}f(X)$.)