Motivation

$$\min_{x\in\mathbb{R}^n} \Big[f(x) = \sum_{i=1}^{m} \ell(\langle a_i, x\rangle) \Big]$$ - Generalized linear models

$\ell:\mathbb{R}\to\mathbb{R}$ convex smooth loss.

$$\langle f'(x), h\rangle = \sum_{i=1}^{m} \ell'(\langle a_i, x\rangle)\langle a_i, h\rangle$$

$$\langle f''(x)h, h\rangle = \sum_{i=1}^{m} \ell''(\langle a_i, x\rangle)\langle a_i, h\rangle^2 \ \le$$

Assume $\ell''(t) \le L_1 \ \forall t$

$$\le \ L_1\langle Bh, h\rangle = L_1\|h\|_B^2 \le \underbrace{L_1 \cdot \|B\|}_{} \cdot \|h\|_2^2$$

Lipschitz const. of the gradient w.r.t. $\|\cdot\|_2$

$$\underline{B := \sum_{i=1}^{m} a_i a_i^T}$$

Gradient Method: $f(x_k) - f^* \le \varepsilon$

$$K = O\left(\frac{L_1\|x_0 - x^*\|_B^2}{\varepsilon}\right) \le O\left(\frac{\overbrace{L_1\|B\|}\cdot\|x_0 - x^*\|_2^2}{\varepsilon}\right)$$

$$|D^3 f(x)[h,h,u]| = \Big| \sum_{i=1}^{m} \ell'''(\langle a_i, x\rangle)\cdot\langle a_i, h\rangle^2\cdot\langle a_i, u\rangle \Big| \le$$

$$\le \underbrace{\max_{1\le i\le m} |\langle a_i, u\rangle| \cdot}_{} \sum_{i=1}^{m} |\ell'''(\langle a_i, x\rangle)|\cdot\langle a_i, h\rangle^2$$

$$\max_{1\le i\le m} |\langle a_i, u\rangle| \le \sqrt{\sum_{i=1}^{m}\langle a_i, u\rangle^2} = \langle Bu, u\rangle^{1/2} = \|u\|_B$$

Assume $\ell'''(t) \leq L_2 \ \forall t$. We set:

$$|D^3 f(x)[h,h,u]| \leq L_2 \cdot \|u\|_B \cdot \|h\|_B^2$$

The Hessian is Lipschitz w.r.t. $\|\cdot\|_B$.

<u>Cubic Newton</u>:

$$K = O\left(\left[\frac{L_2 \cdot D^3}{\varepsilon}\right]^{\frac{1}{2}}\right)$$

$$D \geq \|x_0 - x^*\|_B$$

Diameter of the initial sublevel set.

<u>Example</u> Logistic Loss

$$\ell(t) = \ln(1+e^t)$$

$$L_1 = \frac{1}{4} \qquad L_2 = \frac{1}{6\sqrt{3}}$$

$$\boxed{|\ell'''(t)| \leq \ell''(t)}$$

<u>Example</u> Exponential Loss

$$\ell(t) = e^t \qquad L_1 = L_2 = +\infty$$

$$|D^3 f(x)[h,h,u]| \leq \|u\| \sum_{i=1}^{m} |\ell'''(\langle a_i, x\rangle)| \langle a_i, h\rangle^2 \leq$$

$$\leq \|u\| \underbrace{\sum_{i=1}^{m} \ell''(\langle a_i, x\rangle) \langle a_i, h\rangle^2}$$

$$= \|u\| \cdot \langle f''(x)h, h\rangle .$$

Quasi-Self-Concordant Functions

Def. We say $f: \mathbb{R}^n \to \mathbb{R}$ (strictly convex, differentiable)

is quasi-self-concordant with $M \geq 0$:

$$|D^3 f(x)[h, h, u]| \leq M \cdot \|h\|_x^2 \cdot \|u\|$$

Main Properties h - fixed direction, $x, y \in \mathbb{R}^n$

$$g(t) = \ln \|h\|^2_{x+t(y-x)}$$

$$|g'(t)| = \frac{|D^3 f(x+t(y-x))[h, h, y-x]|}{\|h\|^2_{x+t(y-x)}} \leq$$

$$\leq M\|y-x\| \qquad \forall t \in [0,1]$$

$$\left| \ln \frac{\|h\|_y^2}{\|h\|_x^2} \right| = |g(1) - g(0)| = \left|\int_0^1 g'(t)\,dt\right| \leq M\|y-x\|$$

Lemma $\forall x, y \in \mathbb{R}^n$

$$f''(x) \cdot e^{-M\|y-x\|} \preceq f''(y) \preceq f''(x) \cdot e^{M\|y-x\|}$$

Fix any $h \in \mathbb{R}^n$: $\|h\| \leq 1$

Taylor's Theorem

$$\langle f'(y) - f'(x) - f''(x)(y-x), h \rangle =$$

$$= \int_0^1 (1-\tau) D^3 f(x + \tau(y-x)) [y-x, y-x, h] \, d\tau \leq$$

$$\leq M \cdot \|h\| \cdot \int_0^1 (1-\tau) \underbrace{D^2 f(x+\tau(y-x)) [y-x]^2} \, d\tau \leq$$

$$\leq M \cdot \underbrace{\|h\|}_{\leq 1} \cdot \|y-x\|_x^2 \boxed{\int_0^1 (1-\tau) e^{\tau M \|y-x\|} \, d\tau} \equiv \varphi(M\|y-x\|)$$

<u>Lemma</u> $\forall x, y \in \mathbb{R}^n$

$$\|f'(y) - f'(x) - f''(x)(y-x)\|_* \leq M \|y-x\|_x^2 \cdot \varphi(M\|y-x\|)$$

$$\varphi(t) = \frac{e^t - 1 - t}{t^2} > 0 \text{ , convex, monotone } \nearrow$$

Gradient Regularization of Newton's Method

Cubic Newton:

solution of a univariate subproblem

$$x^+ = x - \left(f''(x) + \frac{Hr^*}{2} B \right)^{-1} f'(x)$$

$$r^* \approx \sqrt{\frac{\|f'(x^+)\|_*}{H}}.$$

Idea: to replace $\|f'(x^+)\|_* \mapsto \|f'(x)\|_*$

Consider iterations of the form:

$$x^+ = x - \left(f''(x) + H \cdot \|f'(x)\|_*^{\alpha} B \right)^{-1} f'(x)$$

$0 \leq \alpha \leq 1$

$\alpha = 0$: constant regularization.

$\alpha = \frac{1}{2}$: similar to the Cubic Regularization,

$\boxed{\alpha = 1} \Rightarrow$ the method has local quadratic convergence!

Gradient Regularization with $\alpha = 1$

We fix $H > 0$, $x \in \mathbb{R}^n$ - current point.

Solve the linear system to find x^+:

$$f'(x) + f''(x)(x^+ - x) + H\|f'(x)\|_* B(x^+ - x) = 0$$

$$\langle \cdot , x^+ - x \rangle :$$

$$\underbrace{\|x^+ - x\|_x^2}_{\geq 0} + \underbrace{H\|f'(x)\|_* \cdot \|x^+ - x\|^2}_{\geq 0} = \langle f'(x), x - x^+ \rangle$$
$$\leq \|f'(x)\|_* \cdot \|x - x^+\|$$

Lemma

- $\|x^+ - x\| \leq \frac{1}{H}$ $\Rightarrow$ implicit "trust-region"
- $\|x^+ - x\|_x^2 \leq \|f'(x)\|_* \cdot \|x^+ - x\|$.

Let's choose $\boxed{H := M}$

$$\varphi(M\|x^+ - x\|) \leq \varphi\left(\frac{M}{H}\right) = \varphi(1) = \underbrace{e - 2}_{\rho} = 0.71828\ldots$$

Progress of one step $\boxed{H := M}$

$$\|f'(x^+) - \underbrace{f'(x) - f''(x)(x^+ - x)}_{M\|f'(x)\|_* B(x^+ - x)}\|_* \leq M\|f'(x)\| \cdot \|x^+ - x\|\rho$$

- $\|x^+ - x\| = r$
- $\|f'(x)\|_* = g$
- $\|f'(x^+)\|_* = g^+$

$$\|f'(x^+) + MgB(x^+ - x)\|_* \leq \rho M g r$$

Taking square:

$$g_+^2 + (Mg)^2 r^2 + 2Mg\langle f'(x^+), x^+ - x\rangle \leq \rho^2 M^2 g^2 r^2$$

$$\Rightarrow \quad \boxed{\langle f'(x^+), x - x^+\rangle \geq \frac{1}{2Mg} g_+^2}$$

Theorem For one step, we have:

$$f(x) - f(x^+) \geq \langle f'(x^+), x - x^+\rangle \geq \frac{1}{2M} \cdot \underbrace{\left[\frac{\|f'(x^+)\|_*}{\|f'(x)\|_*}\right]^2}_{\text{"approximation error"}} \cdot \|f'(x)\|_*$$

For GM: $f_k - f_{k+1} \gtrsim \|f'(x^+)\|^2$

Cubic Newton: $f_k - f_{k+1} \gtrsim \|f'(x^+)\|^{3/2}$

$f_k - f_{k+1} \gtrsim \|f'(x^+)\|$

Without appr. error:

convexity

$$f_k - f_{k+1} \geq \frac{1}{2M} \cdot \|f'(x)\|_*^2 \geq \frac{1}{2MD} f_k$$

$$f_{k+1} \leq \left(1 - \frac{1}{2MD}\right) f_k \Rightarrow \text{linear rate!}$$

To solve the problem: $f(x_k) - f^* \leq \varepsilon$

$$K = O\left(MD \cdot \ln \frac{f(x_0) - f^*}{\varepsilon}\right).$$