$A \succeq B \Leftrightarrow A-B$ is P.S.D

Gradient Descent $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

α^k : stepsize/learning rate >0

Instead of deterministic $\nabla f(x_k)$, stochastic z_k can be used s.t. $\mathbb{E}[z_k] = \nabla f(x_k)$

Convex f^n A f^n $f: \mathbb{R}^d \to \mathbb{R}$ is convex if for any $\vec{x}, \vec{y} \in \mathbb{R}^d, \lambda \in [0,1]$

$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

Structured Optimisation $\min_{x \in C} f(x)$

s.t $g_i(x) \leq 0 \quad i = 1, 2, \ldots, m$

→ For a convex f, First Order optimality: $\nabla f(x^*) = 0$

→ For convex opt all local minima $\Rightarrow$ global minima

Oracle Complexity No. of oracle calls for given precision

For $f \in \mathcal{F}$ (class of f^n) & any algo A

$N(\varepsilon, A, f) = \min\{ n \mid f(x_n) - f(x^*) \leq \varepsilon \}$

Uniform Oracle Complexity $N(\varepsilon, A, \mathcal{F}) = \sup_{f \in \mathcal{F}} N(\varepsilon, A, f)$

Performance analysis of zeroth order oracle Let f be L-lipschitz in ∞ norm. Consider gridpoints $\frac{2\varepsilon}{L}$ apart

Then $f(x_{Algo}) - f(x^*) \leq f(x_{nearest\, x^*}) - f(x^*) \leq L \| x_{near} - x^* \|_\infty = L\varepsilon/L$

GD for Quadratic Objective $f(x) = \frac{x^T Q x}{2} - b^T x$ $\quad Q$: P.S.D & symmetric

$\lambda_{max}(Q) = M$

$\lambda_{min}(Q) = m$

Conditional Number $(Q) = \frac{M}{m} = \chi$

L2 Operator Norm (used in upper bd Mat $\vec{v}$ product) $|||R|||_2 = \max \frac{\|R\vec{x}\|_2}{\|\vec{x}\|_2}, \quad R\vec{v} \leq |||R|||_2 \|\vec{v}\|_2$

$x_{k+1} - x_k = -\alpha_k (Q x_k - Q x^*)$

$x_{k+1} - x^* = (I - \alpha_k Q)(x_k - x^*)$

$\|x_{k+1} - x^*\|_2 \leq |||I - \alpha_k Q|||_2 \, \|x_k - x^*\|_2$

for $\alpha_k = \frac{2}{M+m}$ $\quad \|x_{n+1} - x^*\|_2 \leq \left(\frac{\chi - 1}{\chi + 1}\right)^{n+1} \|x_0 - x^*\|_2$ → Exponential / Geometric / Linear Convergence.

For const step size above α_k gives tightest upper bound

$\frac{2}{M+m} = \min_{\alpha} \max_{\lambda \in [m, M]} |1 - \alpha\lambda|$

$N(\varepsilon) \geq \chi \log\left(\frac{\|x_0 - x^*\|_2}{\varepsilon}\right)$

↳ Initialisation, Precision, Problem structure.

If χ large: Skewed Sublevel set (oscillations)

χ small: circular.

$\left(\frac{\chi-1}{\chi+1}\right)^n \leq \left(1 - \frac{1}{\chi}\right)^n \leq e^{-\frac{n}{\chi}}$

Here $\succeq 0$ denotes semidefinite partial ordering Lowner Order

Smoothness A differentiable f^n f is L smooth

$\|\nabla f(\vec{x}) - \nabla f(\vec{y})\|_2 \leq L \|\vec{x} - \vec{y}\|_2$

$\Leftrightarrow |f(\vec{y}) - f(\vec{x}) - \nabla f(\vec{x})^T (\vec{y} - \vec{x})| \leq \frac{L}{2} \|\vec{y} - \vec{x}\|^2$

$\Leftrightarrow -LI \preceq \nabla^2 f(\vec{x}) \preceq LI$ (if f twice differentiable)

Strong Convexity $\Rightarrow$ Convexity.

Strong Convexity A diff. f^n f is μ strongly convex if

$f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x}) + \frac{\mu}{2} \|\vec{y} - \vec{x}\|_2^2$

$\Leftrightarrow \nabla^2 f(x) \succeq \mu I$ (if twice differentiable)

$\Leftrightarrow \|\nabla f(\vec{x}) - \nabla f(\vec{y})\|_2 \geq \mu \|\vec{x} - \vec{y}\|_2$

$\Leftrightarrow f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex

Strong Convexity through Regularisation When objective f^n F is convex then $F + \frac{\mu}{2}\|\cdot\|_2^2$ is μ strongly convex.

In ML models for convex F,

- If $f: \mathbb{R}^d \to \mathbb{R}$ is μ strongly convex, then it has a unique minimizer.

Lojasiewicz's Inequality If f is differentiable & μ-strongly convex with unique minimizer η_*, we have $\|\nabla f(\theta)\|_2^2 \geq 2\mu (f(\theta) - f(\eta_*))$

Proof Take $\min_y$ on both sides of S.C. $\quad f(y) \geq g(y) \; \forall y \Rightarrow \min_y f(y) \geq \min_y g(y)$

Convergence of GD. Solving $\min_{x \in \mathbb{R}^d} f(x)$

Assumption: f has a global lower bound $f(x) \geq \bar{f}$

① f is M smooth (possibly nonconvex)

After T iterations of GD with $\alpha_k = \frac{1}{M}$, we have

a) $\min_{0 \leq k \leq T-1} \|\nabla f(x_k)\| \leq \sqrt{\frac{2M(f(x_0) - \bar{f})}{T}}$

b) $\lim_{T \to \infty} \|\nabla f(x^T)\| = 0$

Proof: $f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{M}{2}\|x_{k+1} - x_k\|^2$

$f(x_{k+1}) \leq f(x_k) + \frac{M}{2}\left[\left(\alpha - \frac{1}{M}\right)^2 - \frac{1}{M^2}\right] \|\nabla f(x_k)\|_2^2 \quad x_{k+1} = x_k - \alpha \nabla f(x_k)$

Descent Lemma f is M smooth & GD with $\alpha = \frac{1}{M}$ then $f(x_{k+1}) \leq f(x_k) - \frac{1}{2M}\|\nabla f(x_k)\|_2^2$

Add descent lemma $k = 0, 1, \ldots, T-1$ $\quad \sum_{k=0}^{T-1} \|\nabla f(x_k)\|_2^2 \leq 2M(f(x_0) - f(x_T))$

Oracle Complexity (OC) $O\left(\frac{1}{\varepsilon^2}\right) \quad$ RoC $= O\left(\frac{1}{\sqrt{T}}\right)$

If χ large: Skewed Sublevel set drawn as a flat ellipse; χ small drawn as a circle.

Differential Calculus

- Quadratic Forms:
Assuming $A=A^T$ with
$f(x)=\frac{x^TAx}{2}-b^Tx$
$\nabla f(x)=Ax-b$, $\nabla^2 f(x)=A$

if A is not symmetric
$\nabla f(x)=\frac{1}{2}(A+A^T)x-b$
$\nabla^2 f(x)=\frac{1}{2}(A+A^T)$

$X\in\mathbb{R}^{n\times d}$, $y\in\mathbb{R}^n$
$F(\theta)=\frac{1}{2n}\|y-X\theta\|^2$
$\nabla F(\theta)=\frac{1}{n}X^T(X\theta-y)$ $\nabla^2F(\theta)=\frac{X^TX}{n}$

classmate
Date
Page

2 f^n is convex & L smooth

Cocoercivity: If f is convex L smooth f^n on $\mathbb{R}^d$, then $\forall x,y\in\mathbb{R}^d$

$$\frac{1}{L}\|\nabla f(x)-\nabla f(y)\|_2^2\le(\nabla f(x)-\nabla f(y))^T(x-y)$$

Moreover we have $f(y)\ge f(x)+\nabla f(x)^T(y-x)+\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|_2^2$

Proof $f(x)+\nabla f(x)^T(z-x)\le f(z)\le f(y)+\nabla f(y)^T(z-y)+\frac{L}{2}\|z-y\|_2^2$

Find z by minimizing the difference b/w LHS & RHS. $x\to y$ $y\to x$ add

$f(x_{k+1})\le f(x_k)-\frac{1}{2L}\|\nabla f(x)\|_2^2$ By descent lemma

$f(x_{k+1})\le f(x^*)-\nabla f(x_k)^T(x^*-x_k)-\frac{1}{2M}\|\nabla f(x_k)\|_2^2$ Convexity of f

$=f(x^*)+\frac{M}{2}\left[\|x_k-x^*\|_2^2-\|\underbrace{x_k-\frac{\nabla f(x^*)}{M}}_{x_{k+1}}-x^*\|_2^2\right]$

$Tf(x_T)\le\sum_{k=1}^T f(x_T)\le\frac{L}{2}\left[\|x_0-x^*\|_2^2-\|x_T-x^*\|_2^2\right]+Tf(x^*)$ Telescopic sums

$$\boxed{f(x_T)-f(x^*)\le\frac{L}{2T}\|x_0-x^*\|_2^2}$$

$OC=O(\frac{1}{\epsilon})$ $RoC=O(1/T)$

3 f is L smooth & μ strongly convex

Use Lojasiewicz's inequality in descent lemma.

$\alpha_k=\frac{1}{M}$
$f(x_{k+1})\le f(x_k)-\frac{2m}{2M}(f(x_k)-f(x^*))$
$f(x_{k+1})-f(x^*)\le(1-\frac{m}{M})(f(x_k)-f(x^*))$
$f(x_T)-f(x^*)\le(1-\frac{m}{M})^T(f(x^0)-f(x^*))$
$OC=O(\log\frac{1}{\epsilon})\frac{M}{m}$ RoC: Exponential.

For $\alpha_k=\frac{2}{M+m}$ It we have $\|x_T-x^*\|_2\le\left(\frac{1-\frac{m}{M}}{1+\frac{m}{M}}\right)^T\|x_0-x^*\|_2$

$f(x)=f(x_0)+\langle\nabla f(x_0),x-x_0\rangle+o(\|x-x_0\|)$

$f(x)=f(x_0)+\langle\nabla f(x_0),x-x_0\rangle+\frac{1}{2}(x-x_0)^T\nabla^2f(x_0)(x-x_0)+o(\|x-x_0\|^2)$

If 2x cont Diff $f(x)=f(x_0)+\langle\nabla f(x_0),x-x_0\rangle+\frac{1}{2}\int_0^1(x-x_0)^T\nabla^2f((1-t)x_0+tx)(x-x_0)dt$

Constrained Optimization

Def^n Let $\Omega\subset\mathbb{R}^n$ be a closed convex set. At any $x\in\Omega$

$N_\Omega(x)=\{d\in\mathbb{R}^n: d^T(y-x)\le 0\ \forall y\in\Omega\}$

$\min_{x\in\Omega}f(x)$ Ω: closed & convex

Theorem
a) If f is continously differentiable. If $x^*\in\Omega$ is a local sol^n of above problem, then $-\nabla f(x^*)\in N_\Omega(x^*)$.
b) If f is also convex, then $-\nabla f(x^*)\in N_\Omega(x^*)\Rightarrow x^*$ is local (global) minima
c) Furthermore if f is strongly convex x^* is unique satisfying $-\nabla f(x^*)\in N_\Omega(x^*)$

Taylor's Theorem Given a continuously differentiable f^n $f:\mathbb{R}^n\to\mathbb{R}$, and given $x,t\in\mathbb{R}^n$

Integral Form $f(x+t)=f(x)+\int_0^1\nabla f(x+\gamma t)^Tt\,d\gamma$

Mean-Value Form. $f(x+p)=f(x)+\nabla f(x+\gamma p)^Tp$, for some $\gamma\in(0,1)$

add & subtract $\nabla f(x)p$

If f is twice continuously differentiable,

$\nabla f(x+t)=\nabla f(x)+\int_0^1\nabla^2f(x+\gamma t)\vec{t}\,d\gamma$

$f(x+t)=f(x)+\nabla f(x)^Tt+\frac{t^T}{2}\nabla^2f(x+\gamma t)t$ for some $\gamma\in(0,1)$

$\to f(x+t)=f(x)+\nabla f(x)^Tt+o(\|t\|)$

Euclidean Projection

$P_\Omega(x)=\text{argmin}\{\|z-x\|:z\in\Omega\}$

$\hookrightarrow$ sol^n of minimize$_z$ $\frac{1}{2}\|z-x\|^2$ subject to $z\in\Omega$ — Strongly convex thus unique minimizer

$-\nabla f(P_\Omega(x))\in N_c(P_\Omega(x))\Rightarrow-(P_\Omega(x)-x)\in N_c(P_\Omega(x))$

Minimum Principle $\langle x-P_\Omega(x),y-P_\Omega(x)\rangle\le 0\ \forall y\in\Omega$
(uniquely characterises projection)

Contraction Property $\|P_\Omega(x)-P_\Omega(y)\|_2\le\|x-y\|_2$ for closed & cvx Ω

Proof $\|x-y\|_2^2=\|(x-P_\Omega(x))-(y-P_\Omega(y))+(P_\Omega(x)-P_\Omega(y))\|_2^2$

Projected Gradient Algorithm. $\min f(x)$ subject to $x\in\Omega$

DANGER: One may be tempted to write $\nabla f(x^*)=0$, but should not do while working with constrained optimisation

$\hookrightarrow x_{k+1}=P_\Omega(x_k-\alpha_k\nabla f(x_k))$

x^* is optimal sol^n iff $x^*=P_\Omega(x^*-\alpha\nabla f(x^*))$

$-\nabla f(x^*)\in N_c(x^*)\Rightarrow\langle-\nabla f(x^*),z-x^*\rangle\le 0\ \forall z\in C$

$\langle x^*-\alpha\nabla f(x^*)-x^*,z-x^*\rangle\le 0$

classmate

Date ____

Page ____

$\min_{x \in C} f(x)$

1 Suppose C is closed & convex & f is M smooth & m S.C.

we run PGD $x_{k+1} = P_C(x_k - \alpha \nabla f(x_k))$ with $\alpha = \frac{2}{M+m}$

then we have $\|x_k - x^*\|_2 \leq \left(\frac{1-\frac{m}{M}}{1+\frac{m}{M}}\right)^k \|x_0 - x^*\|_2$

Identical to unconstrained setup. Projection not hurting

Proof

$\|x_{k+1} - x^*\|_2 = \|P_C(x_k - \alpha\nabla f(x_k)) - P_C(x^* - \alpha \nabla f(x^*))\|_2$

Contraction Property $\forall x, y \in C$ $\leq \|(x - \alpha\nabla f(x)) - (y - \alpha \nabla f(y))\|_2 \leq \left(\frac{1-\frac{m}{M}}{1+\frac{m}{M}}\right)\|x-y\|$

Proof: Taylor.

2 C closed cvx, f Msmooth & cvx PGD with $\alpha = \frac{1}{M}$

$f(x_T) - f(x^*) \leq \frac{M}{2T}\|x_0 - x^*\|^2$

Projected Descent Lemma for f Msmooth & cvx

$f(T(x)) - f(y) \leq \langle g_C(x), x-y\rangle - \frac{1}{2M}\|g_C(x)\|_2^2$

$T(x) = P_C(x - \frac{1}{M}\nabla f(x))$ $g_C(x) = M(x - T(x))$ $\forall x, y \in C$

~~add~~ subtract add $f(x)$ use ~~convexity &~~ M smoothness & convexity respectively

$f(T(x)) - f(y) \leq \frac{M}{2}(\|x-y\|^2 - \|T(x) - y\|^2)$

$x = x_k$ $y = x_k$ $\Rightarrow f(x_{k+1}) \leq f(x_k)$

$x = x_k$ $y = x^*$ $\Rightarrow$ add

avoids projection

Frank Wolfe / Conditional Gradient Ω: closed, cvx & bounded

Frank Wolfe Direction

$\vec{d}_k = \underset{y\in\Omega}{\text{argmin}}\, f(x_k) + \langle \nabla f(x_k), y - x_k\rangle = \underset{y\in\Omega}{\text{argmin}} \langle \nabla f(x_k), y\rangle$

$\vec{x}_{k+1} = (1-\alpha_k)\vec{x}_k + \alpha_k \vec{d}_k$

Suppose Ω closed, cvx, bounded has diameter D.

$f(\cdot)$ is M smooth & convex, $\alpha_k = \frac{2}{k+2}$

FW yields $f(x_k) - f(x^*) \leq \frac{2MD^2}{k+2}$ for $k = 1, 2, \ldots$

$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{M}{2}\|x_{k+1} - x_k\|^2$

Using $x_{k+1} = (1-\alpha_k)x_k + \alpha_k d_k$

$\leq f(x_k) + \alpha_k\langle \nabla f(x_k), d_k - x_k\rangle + \frac{M}{2}\alpha_k^2\|d_k - x_k\|^2$

$\leq f(x_k) + \alpha_k \underbrace{\langle \nabla f(x_k), x^* - x_k\rangle}_{\text{convexity } \leq f(x^*) - f(x_k)} + \frac{M}{2}\alpha_k^2 D^2$

then $\alpha_k = \frac{2}{k+2}$

$f(x_{k+1}) - f(x^*) \leq \frac{k}{k+2}(f(x_k) - f(x^*)) + \frac{M}{2}\left(\frac{k}{k+2}\right)^2 D^2$ induction

$\leq \frac{2MD^2}{k+3}$

Epigraph $\text{epi}(f) = \{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq t\}$

$f: \mathbb{R}^d \to \mathbb{R}$ is convex $\Leftrightarrow$ epi(f) is a convex set

If f is convex, define subgradient at x as g which satisfies

Sub-Gradient $f(y) \geq f(x) + \langle g, y - x\rangle$

sub differential $g \in \partial f(x)$ set of all subgradient at x

f is differentiable at $x \overset{\text{cvx}}{\Leftrightarrow} \partial f(x) = \{\nabla f(x)\}$

$f(x+td) = f(x) + t\langle f(x), d\rangle + o(t)$

x^* is global minima $\Leftrightarrow \vec{0} \in \partial f(x^*)$

defines the hyperplane supporting epi(f) at x

Indicator Function (defined on relative interior)

$I_C(x) = \begin{cases} 0 & x \in C \\ +\infty & x \notin C \end{cases}$

convex

$\partial I_C(x) = N_C(x)$ $\forall x \in C$

Calculus for SubDifferentials

1. $\partial(\lambda f(x)) = \lambda\partial(f(x))$
2. $\underbrace{\partial f_1(x) + \partial f_2(x)}_{}$ $\subseteq$ $\partial(f_1 + f_2)(x)$

Minkowski sum $A + B = \{a + b \mid a \in A, b \in B\}$

Remark $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$ is true when $\text{int}(\text{dom}(f_1) \cap \text{dom}(f_2)) \neq \phi$

Connection to constrained optimisation. Ω: closed convex.

$\min_{x\in\Omega} f(x) \Leftrightarrow \min_{x \in \mathbb{R}^d} f(x) + I_\Omega(x)$

x^* is sol^n of $\min_{x\in\mathbb{R}^d} f(x)$ iff $\vec{0} \in \partial f(x^*)$

$0 \in \partial f(x^*) + \partial I_C(x^*)$ interior cond^n satisfied.

if f diff. $0 \in \nabla f(x^*) + \partial I_C(x^*) \Rightarrow -\nabla f(x^*) \in \partial I_C(x^*)$

else $0 \in \partial f(x^*) + N_C(x^*)$

Danskin's Theorem Given f^n $\phi: \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ & a compact set $Z \subseteq \mathbb{R}^m$ consider

$f(x) = \max_{z\in Z}\phi(x,z)$ where $\phi(x,z)$ is cvx in x $\forall z$

Also $\phi(x,z)$ is continuous on dom$(f) \times Z$

$Z^*(x) = \{z^* \in Z \mid f(x) = \phi(x, z^*)\}$. Then

$\partial f(x) = \text{conv}\left(\bigcup_{z^*\in Z^*} \partial\phi(x, z^*)\right)$. Moreover if ϕ is differentiable $\partial f(x) = \text{conv}(\nabla_x \phi(x, z^*), z^* \in Z^*)$

consider $f(x) = g(x) + h(x)$ diff convex

Algorithms: Sub-Gradient Method

$\min_{x \in \mathbb{R}^d} f(x)$ → non differentiable & non smooth

Assumption: $\|g\|_2 \le G$ for any $g \in \partial f(x)$ for any x.

$x_{k+1} = x_k - \alpha_k g_k$

Theorem. Running for T steps.

$$f\left(\frac{\sum_{k=1}^T \alpha_k x_k}{\sum_{k=1}^T \alpha_k}\right) - f(x^*) \le \frac{\|x_1 - x^*\| + G^2 \sum_{k=1}^T \alpha_k^2}{2 \sum_{k=1}^T \alpha_k}$$

Proof: $\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha_k \underbrace{\langle g_k, x_k - x^* \rangle}_{} + (\alpha_k)^2 G^2$

Jensen $\le 2\alpha_k (f(x^*) - f(x_k))$

Lower Bound by Nesterov: $\min_{x \in \mathbb{R}^d} f(x)$ f is convex & G-Lipschitz ($\Rightarrow$ all subgradients bounded by G)

Let $\|x_0 - x^*\|_2 \le D$ *argmin.

First order algo: Algo receives $g_k \in \partial f(x_k)$, algo generates next iterate $x_{k+1} \in x_0 + \text{span}\{g_0, g_1, \dots, g_k\}$

Theorem: For any such 1^{st} order algo, with init x_0, $\exists$ a f^n (convex & Lipschitz), such that for all iterations $k \le d-1$

$$f(x_k) - f(x^*) \ge DG / 2(1 + \sqrt{k+1})$$

l_2 ball radius $D = \frac{1}{\sqrt{d}}$

f^n $f(x) = \max_{j \in [d]} x_j + \frac{1}{2}\|x\|_2^2$

For a symmetric f^n minimum is attained when all coordinates same

Proximal Methods: Composite F^n $f(x) = \underset{\text{differentiable}}{g(x)} + \underset{\text{convex}}{h(x)}$

Proximal Operator: $\text{prox}_h(x) = \underset{y \in \mathbb{R}^d}{\text{argmin}} \underbrace{\left(h(y) + \frac{1}{2}\|y - x\|^2 \right)}_{\text{strongly convex}}$

$0 \in \partial h(\text{prox}_h(x)) + [\text{prox}_h(x) - x]$

if $h(x) = I_\Omega(x)$, $\text{prox}_h(x) = P_\Omega(x)$

Ex 1: For $h(x) = \lambda \|x\|_1$

↳ soft thresholding

$$\text{prox}_h(x) = \begin{cases} x_i - \lambda & x_i > \lambda \\ 0 & |x_i| \le \lambda \\ x_i + \lambda & x_i < -\lambda \end{cases}$$

Ex2: $h(y) = (f(x) + \langle \nabla f(x), y - x \rangle)\alpha$

$\text{Prox}_h(x) = x - \alpha \nabla f(x)$

Pure Proximal $x_{k+1} = \text{Prox}_{\alpha f}(x_k)$

Lemma: $x^* = \text{prox}_{\alpha f}(x^*) \Leftrightarrow 0 \in \partial f(x^*)$

Proximal Gradient Descent $x_{k+1} = \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla g(x_k))$

Theorem: Suppose g is differentiable M-smooth & m-strongly convex. We used PGD with $\alpha_k = \frac{2}{M+m}$, then

$$\|x_k - x^*\|_2 \le \left(\frac{1 - \frac{m}{M}}{1 + \frac{m}{M}} \right)^k \|x_0 - x^*\|_2$$

from monotonicity → $\langle \partial f(x) - \partial f(y), x - y \rangle > 0$

We get of subgradient that Firmly NonExpansive $\langle y - x, \text{prox}_h(y) - \text{prox}_h(x) \rangle \ge \|\text{prox}_h(x) - \text{prox}_h(y)\|^2$

$\|\text{prox}_h(x) - \text{prox}_h(y)\|_2 \le \|y - x\|_2$

$T(x) = \text{prox}_{\alpha h}(S(x))$

$S(x) = x - \alpha \nabla g(x)$

$$\begin{aligned} \|x_{k+1} - x^*\|_2 &= \|T(x_k) - T(x^*)\| \\ &= \|\text{prox}_{\alpha h}(S(x_k)) - \text{prox}_{\alpha h}(S(x^*))\| \\ &\le \|S(x_k) - S(x^*)\|_2 \\ &\le \frac{1 - m/M}{1 + m/M} \|x_k - x^*\|_2 \end{aligned}$$

Theorem: PGD for smooth functions. g is cvx & M smooth

Run PGD with $\alpha_k = \frac{1}{M}$ $f(x_T) - f(x^*) \le \frac{M}{2T}\|x_0 - x^*\|^2$

$f(x) = g(x) + h(x)$
$x^* \in \text{argmin} f$

$T(x) = \text{prox}_{\alpha_k h}\left(x - \frac{\nabla g(x)}{M}\right)$

$g_{prox}(x) := M(x - T(x))$

$x_{k+1} = x_k - \frac{1}{M} g_{prox}(x_k)$

Descent Lemma

$$f(T(x)) - f(y) \le \langle g_{prox}(x), x - y \rangle - \frac{1}{2M}\|g_{prox}(x)\|^2$$

SGD $x \to \boxed{\text{S.S.}} \to G(x, \xi)$ s.t. $E_\xi[G(x, \xi)] = \nabla f(x)$

Variance Control Boundedness Assumption $E_\xi \|G(x, \xi)\|^2 \le B^2$

Theorem: Suppose f is convex. We observe stochastic sub gradients that are (i) Unbiased $E_\xi[G(x; \xi)] \in \partial f(x)$ (ii) B^2 bounded $E_\xi \|G(x; \xi)\|^2 \le B^2$ (iii) Ind. across rounds

Proof Sketch: $f(x_k) - f(x^*) \le \langle \nabla f(x_k), x_k - x^* \rangle$

$\nabla f(x_k) = \frac{x_k - x_{k+1}}{\alpha_k} - \omega_k$

$2\langle a, b \rangle = \|a + b\|^2 - \|a\|^2 - \|b\|^2$

$$E[f(\bar{x}^T)] - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2\sum_{k=1}^T \alpha_k} + B^2 \frac{\sum_{k=1}^T \alpha_k^2}{2\sum_{k=1}^T \alpha_k}$$

μ-PL cond^n $\langle \nabla f(y), y - x^* \rangle \ge \mu \|y - x^*\|^2$

Relaxing Boundedness Condition: $E_\xi \|G(x; \xi)\|^2 \le A^2 \|x - x^*\|^2 + B^2$

Theorem: If f satisfies μ-PL condition & stochastic gradients are A-B bounded. Denote $\Delta x_k = E\|x_k - x^*\|^2$

We have $\Delta x_{k+1} \le (1 - 2\alpha_k \mu + \alpha_k^2 A^2)\Delta x_k + \alpha_k^2 B^2$

if $A \ne 0$ with step size $\alpha_k = \frac{\mu}{A^2}$ $\Delta x_{k+1} \le \left(1 - \frac{\mu^2}{A^2}\right)\Delta x_k + \left(\frac{\mu B}{A^2}\right)^2$

$$E[\|x_k - x^*\|^2] \le \left(1 - \frac{\mu^2}{A^2}\right)^k \|x_0 - x^*\|^2 + \frac{\left(\frac{\mu B}{A^2}\right)^2}{\mu^2 / A^2}$$

Proof Sketch: $\|x_{k+1} - x^*\|^2 = \|x_k - \alpha_k G(x_k, \xi_k) - x^*\|^2$

(Pure) $f: \mathbb{R}^d \to \mathbb{R}$ is twice differentiable.

Newton's Method.

$$x_{l+1} = x_l - [\nabla^2 f(x_l)]^{-1} \nabla f(x_l)$$

Local Convergence of Newton Method.

Assumption Ⓐ x^* unique with $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq mI$ $(m>0)$

Ⓑ $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \le L\|x-y\|_2$

Theorem Under Ⓐ&Ⓑ. Given any x_0 satisfying, $\|x_0 - x^*\| \le \frac{m}{3L}$ then Newton's method is well defined & we get.

$$\|x_{l+1} - x^*\|_2 \le \frac{3L}{2m}\|x_l - x^*\|^2 \quad \forall l \ge 1$$

$$(\Rightarrow \|x_l - x^*\| \le \frac{m}{3L} \; ; \; l = 1, 2, \cdots)$$

Proof Sketch

$$x_{l+1} - x^* = x_l - x^* - (\nabla^2 f(x_l))^{-1} \underbrace{(\nabla f(x_l) - \nabla f(x^*))}$$

Use Fund. Theorem of calculus & Jensen on operator norm.

Doubly Exponential/Quadratic Convergence

Self Concordant A convex f^n $f: \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|f'''(x)| \le 2(f''(x))^{3/2} \quad \forall x \in \text{dom} f.$$

f^n on $\mathbb{R}^n$ $f: \mathbb{R}^n \to \mathbb{R}$ is self-concordant if it is self-concordant along every line in its domain.

i.e. $\tilde{f}(t) = f(x+tv)$ is SC f^n of t $\forall x \in \text{dom} f$ & $\forall v$

SC Calculus.

1. Scaling f is SC $\Rightarrow$ af is SC for $a \ge 1$
2. Sum of SC f^n is SC
3. Composition with affine f^n $f: \mathbb{R}^n \to \mathbb{R}$ is S.C., and $A \in \mathbb{R}^{n \times m}$ $b \in \mathbb{R}^n$, then $f(Ax+b)$ is S.C.

Backtracking Line Search

given a descent direction Δx for f at $x \in \text{dom} f$, $\alpha \in (0, 0.5)$, $\beta \in (0,1)$. $t := 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, $t := \beta t$

Newton's Method

Newton Decrement: $\lambda(x) = [\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)]^{\frac{1}{2}}$

Variational Characterisation $= \sup\limits_{v \neq 0} \dfrac{-v^T \nabla f(x)}{(v^T \nabla^2 f(x) v)^{1/2}}$

given starting pt $x \in \text{dom} f$, tolerance $\epsilon > 0$

repeat

1. Compute Newton step & Decrement $\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$, $\lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$
2. Stopping Criteria: quit if $\lambda^2/2 \le \epsilon$
3. Line Search: Choose step size t by backtracking line search.
4. Update $x := x + t\Delta x_{nt}$

For self concordant f^n f $\quad f(x) - \inf_x f(x) \le \lambda^2(x)$ if $\lambda(x) \le 0.68$.

Newton's Method for S.C. f^n with Backtracking line search.

Repeat

1. Compute Newton step $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$
2. Compute Newton decrement $\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{1/2}$
3. Stopping Criteria $\lambda^2(x)/2 \le \epsilon$.
4. line Search: select 't' by Backtracking Line Search.
5. Update $x := x + t\Delta x_{nt}$

Global algo as we don't require suitable starting point

Convergence Guarantees Assumption: f is strictly convex & S.C.

Theorem: $\exists\, \eta$ and $\gamma\,(>0)$ with $0 < \eta \le 1/4$ that depends on Backtracking line search param α & β s.t.

i) if $\lambda(x_k) > \eta$ then $f(x_{k+1}) - f(x_k) \le -\gamma$

ii) if $\lambda(x_k) \le \eta$ then Backtracking L.S. selects $t=1$ and

$$2\lambda(x_{k+1}) \le (2\lambda(x_k))^2 \quad [\lambda(x_l) \le \eta \Rightarrow \lambda(x_{l+1}) \le \eta]$$

$$\#\text{steps} \le \frac{f(x_0) - p^*}{\gamma} + \ln \ln \frac{1}{\epsilon}$$

Dual Problem

minimize $f(x)$ subject to $Ax = b$

Lagrangian: $L(x,y) = f(x) + y^T(Ax-b)$

dual f^n: $g(y) = \inf_x L(x,y)$

dualproblem maximize $g(y)$

recover $x^* = \operatorname{argmin}_x L(x,y^*)$

Dual Ascent Method

$x_{k+1} = \operatorname{argmin}_x L(x, y_k)$ // x-minimization

$y_{k+1} = y_k + \alpha_k (Ax_{k+1} - b)$ // dual update

Alternating direction method of multipliers.

ADMM problem form (with f, g convex)

minimize $f(x) + g(z)$

subject to $Ax + Bz = C$

$L_\rho(x,y,z) = f(x) + g(z) + y^T(Ax+Bz-c) + \frac{\rho}{2}\|Ax+Bz-c\|^2$

ADMM $x_{k+1} = \operatorname{argmin}_x L_\rho(x, z_k, y_k)$ // x-minimization

$z_{k+1} = \operatorname{argmin}_z L_\rho(x_{k+1}, z, y_k)$ // z-minimization

$y_{k+1} = y_k + \rho(Ax_{k+1} + Bz_{k+1} - c)$ // Dual Update.

ADMM with scaled dual variable.

Combine linear & quadratic terms in Augmented Lagrangian

$L_\rho(x,z,y) = f(x) + g(z) + y^T(Ax+By-c) + \frac{\rho}{2}\|Ax+Bz-c\|^2$

$= f(x) + g(z) + \frac{\rho}{2}\|Ax+Bz-c+u\|^2 + \text{const.}$

$u = \frac{y}{\rho}$

$x_{k+1} = \operatorname{argmin}_x \left(f(x) + \frac{\rho}{2}\|Ax+Bz_k-c+u_k\|^2 \right)$

$z_{k+1} = \operatorname{argmin}_z \left(g(z) + \frac{\rho}{2}\|Ax_{k+1}+Bz-c+u_k\|^2 \right)$

$u_{k+1} = u_k + (Ax_{k+1} + Bz_{k+1} - c)$

Calculus

$f(x) = \frac{x^T A x}{2} - b^T x$ $\quad F(\theta) = \frac{1}{2n}\|y - X\theta\|^2$

$\nabla f(x) = \frac{(A+A^T)x}{2} - b$ $\quad \nabla F(\theta) = \frac{X^T(X\theta - y)}{n}$

Smoothness.

$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x-y\|_2 \Leftrightarrow -LI \le \nabla^2 f(x) \le LI$

$\Leftrightarrow |f(y) - f(x) - \nabla f(x)^T(y-x)| \le \frac{L}{2}\|y-x\|^2$

Strong Convexity

$\nabla^2 f(x) \succeq \mu I \Leftrightarrow f(x) - \frac{\mu}{2}\|x\|^2$ is convex

$\Leftrightarrow \|\nabla f(\vec{x}) - \nabla f(\vec{y})\|_2 \ge \mu\|x-y\|_2$

$\Leftrightarrow f(\vec{y}) \ge f(\vec{x}) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$

Convergence of GD $x_{k+1} = x_k - \alpha\nabla f(x)$

① f is M smooth

Descent Lemma, $\alpha_k = 1/M$

$f(x_{k+1}) \le f(x_k) - \frac{1}{2M}\|\nabla f(x_k)\|^2$

$\min_{0 \le k \le T-1} \|\nabla f(x_k)\| \le \sqrt{\frac{2M}{T}(f(x_0) - \bar{f})}$

② f is cvx + M smooth.

Co-coercivity $f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$

$\Rightarrow \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \le (\nabla f(x) - \nabla f(y))^T(x-y)$

$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$

$\le f(x^*) - \nabla f(x_k)^T(x^* - x_k) - \frac{1}{2M}\|\nabla f(x_k)\|^2$

$= f(x^*) + \frac{M}{2}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right]$

$f(x_T) - f(x^*) \le \frac{L}{2T}\|x_0 - x^*\|^2$

③ f is M smooth + μ strongly convex

Lojaseiwicz's Inequality. for μ strongly cvx f

$\|\nabla f(\theta)\|^2 \ge 2\mu(f(\theta) - f(\operatorname{argmin} f))$

Use LI in descent lemma.

$\alpha_k = \frac{1}{M}$ $\quad f(x_T) - f(x^*) \le \left(1 - \frac{m}{M}\right)^T (f(x_0) - f(x^*))$

Normal Cone Let Ω be a closed cvx set. $x \in \Omega$

$N_\Omega(x) = \{d \in \mathbb{R}^n : d^T(y-x) \le 0 \ \forall y \in \Omega\}$

Theorem If x^* is sol^n to constrained problem then we have $-\nabla f(x^*) \in N_C(x^*)$

Euclidean Projection $P_\Omega(x) = \operatorname{argmin}\{\|z-x\| : z \in \Omega\}$

sol^n of $\min_{z \in \Omega} \frac{1}{2}\|z-x\|^2$ $\quad -\nabla f(P_\Omega(x)) \in N(P_\Omega(x))$

$\Rightarrow \langle x - P_\Omega(x), y - P_\Omega(x)\rangle \le 0 \ \forall y \in \Omega.$

Contraction $\|P_\Omega(x) - P_\Omega(y)\|_2 \le \|x-y\|_2$

Projected GD $\min_x f(x)$ s.t. $x \in \Omega$

$x_{k+1} = P_\Omega(x_k - \alpha_k\nabla f(x_k))$

x^* is fixed point. $-\nabla f(x^*) \in N_C(x^*)$

$\Rightarrow \langle x^* - \alpha\nabla f(x^*) - x^*, z - x^*\rangle \le 0$

1. f $m+M$ $\quad \alpha = \frac{2}{M+m}$ $\quad \|x_k - x^*\|_2 \le \left(\frac{1 - \frac{m}{M}}{1 + \frac{m}{M}}\right)^k \|x_0 - x^*\|_2$

$\|x_{k+1} - x^*\|_2 = \|P_C(x_k - \alpha\nabla f(x_k)) - P_C(x^* - \alpha\nabla f(x^*))\|$

$\le \|(x - \alpha\nabla f(x)) - (y - \alpha\nabla f(y))\|_2 \le \frac{1 - \frac{m}{M}}{1 + m/M}\|x - y\|$

2. f M smooth & cvx PGD with $\alpha = 1/M$

$f(x_T) - f(x^*) \le M/2T\,\|x_0 - x^*\|^2$

Projected Descent Lemma

$f(T(x)) - f(y) \le \langle g_C(x), x-y\rangle - \frac{1}{2M}\|g_C(x)\|_2^2$

$T(x) = P_C(x - \frac{1}{M}\nabla f(x))$ $\quad g_C(x) = M(x - T(x))$

$f(T(x)) - f(y) \le \frac{M}{2}(\|x-y\|^2 - \|T(x) - y\|^2)$

$x, y \quad x_k, x_k \quad x_k, x^*$

Frank Wolfe. $d_k = \operatorname{argmin}\langle\nabla f(x_k), y\rangle$

$x_{k+1} = (1-\alpha)x_k + \alpha_k d_k$

- For M smooth & cvx f, $\alpha_k = \frac{2}{k+2}$

FW $\Rightarrow f(x_k) - f(x^*) \le \frac{2MD^2}{k+2}$

Epigraph $\mathrm{epi}(f) = \{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \le t\}$

Sub-Gradient $f(y) \ge f(x) + \langle g, y-x\rangle$

$g \in \partial f(x)$ [Subdifferential]

f cvx diff $\Rightarrow \partial f(x) = \{\nabla f(x)\}$

$I_C(x)$ $\quad \partial I_C(x) = N_C(x) \ \forall x \in C.$

$\partial f_1(x) + \partial f_2(x) \subseteq \partial(f_1 + f_2)(x)$

$=$ when $\mathrm{int}(\mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)) \ne \phi$

Constrained Opt $\min_{x \in \Omega} f(x) \Leftrightarrow \min f(x) + I_\Omega(x)$

$0 \in \partial f(x^*) + N_C(x^*)$

Danskin's Theorem f^n $\phi : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ &

$f(x) = \max_{z \in Z} \phi(x,z)$ $\quad \phi(x,z)$ is cvx in $x \ \forall z.$

Also $\phi(x,z)$ is continuous.

$Z^*(x) = \{z^* \in Z \mid f(x) = \phi(x, z^*)\}$ Then

$\partial f(x) = \mathrm{conv}\left(\bigcup_{z^* \in Z} \partial\phi(x, z^*)\right)$

Subgradient Method $\|g\|_2 \le G$, $g \in \partial f(x)$ $\quad \min f(x)$

$$f\left(\frac{\sum_{k=1}^T \alpha_k x_k}{\sum_{k=1}^T \alpha_k}\right) - f(x^*) \le \frac{\|x_1 - x^*\| + G^2\sum_{k=1}^T \alpha_k^2}{2\sum_{k=1}^T \alpha}$$

Proof $\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha_k\underbrace{\langle g_k, x_k - x^*\rangle}_{} + \alpha_k^2 G^2$

$\le 2\alpha_k(f(x^*) - f(x_k))$

Nesterov Lower bound

$f(x_k) - f(x^*) \ge DG/2(1 + \sqrt{k+1})$ $\quad f^n = \max_j x_j + \frac{1}{2}\|x\|^2$

Proximal Methods. $f(x) = g(x) + h(x)$

g: differentiable. h: cvx

Pure proximal: $x_{k+1} = \mathrm{prox}_{\alpha f}(x_k)$

$\mathrm{prox}_h(x) = \operatorname{argmin}_y\left(h(y) + \frac{1}{2}\|y-x\|^2\right)$

Proximal GD $x_{k+1} = \mathrm{prox}_{\alpha_k h}(x_k - \alpha_k\nabla g(x_k))$

Fixed pt $\quad x \in z + \lambda\partial f(z) \Leftrightarrow 0 \in \partial_z f(z) + \frac{1}{2\lambda}\|z - x\|^2$

$z \in (I + \lambda\partial f)^{-1}x$ iff $x = \mathrm{prox}_{\lambda f}(x)$

$x^* = \operatorname{argmin} f(x) + g(x)$

$0 \in \nabla f(x^*) + \partial g(x^*)$

$(I - \lambda\nabla f)x^* \in x^* + \lambda\partial g(x^*)$

$x^* = (I + \lambda\partial g)^{-1}(I - \lambda\nabla f)(x^*)$

$\therefore x^* = \mathrm{prox}_{\lambda g}(x^* - \lambda\nabla f(x^*))$

Theorem Suppose g diff. $M+m$. PGD $\alpha_k = 2/M+m$.

$\|x_k - x^*\|_2 \le \left(\frac{1 - \frac{m}{M}}{1 + m/M}\right)^k \|x_0 - x^*\|_2$

Proof: $0 \in \partial h(\mathrm{prox}_h(x)) + [\mathrm{prox}_h(x) - x]$ $\quad \langle\partial f(x) - \partial f(y), x - y\rangle \ge 0$

$\langle y - x, \mathrm{prox}_h(y) - \mathrm{prox}_h(x)\rangle \ge \|\mathrm{prox}_h(x) - \mathrm{prox}_h(y)\|^2$

$T(x) = \mathrm{prox}_{\alpha h}(S(x))$ $\quad S(x) = x - \alpha\nabla g(x)$

$\|x_{k+1} - x^*\|_2 = \|T(x_k) - T(x^*)\| = |\mathrm{prox}_{\alpha h} S(x_k) - \mathrm{prox}_{\alpha h} S(x^*)|$

$\le \|S(x_k) - S(x^*)\|_2 \le \frac{1 - m/M}{1 + m/M}\|x_k - x^*\|_2$

Theorem: PGD for smooth f^n g is cvx & M smooth

$\alpha_k = \frac{1}{M}$ $\quad f(x_T) - f(x^*) \le \frac{M}{2T}\|x_0 - x^*\|^2$

$T(x) = \mathrm{prox}_{\alpha_k h}\left(x - \frac{\nabla g(x)}{M}\right)$ $\quad g_{prox}(x) = M(x - T(x))$

$x_{k+1} = x_k - \frac{1}{M}g_{prox}(x_k)$

Same descent lemma $C \to$ prox.

Theorem: f cvx ① $E_\xi[G(x;\xi)] \in \partial f(x)$

(ii) $E[\|G(x;\xi)\|^2] \le B^2$

$E[f(\bar{x}^T)] - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2\sum_k \alpha_k} + B^2\frac{\sum_k \alpha_k^2}{2\sum_k \alpha_k}.$

μ-PL cond^n $\Rightarrow f(x_k) - f(x^*) \le \langle\nabla f(x_k), x_k - x^*\rangle$

$\nabla f(x_k) = \frac{x_k - x_{k+1}}{\alpha_k} - w_k$

$2\langle a, b\rangle = (a+b)^2 - a^2 - b^2$

μ-PL cond^n $\langle\nabla f(y), y - x^*\rangle \ge \mu\|y - x^*\|^2$

$\mathbb{E}_\xi\|G(x;\xi)\|^2 \le A^2\|x - x^*\|^2 + B$

Theorem If f satisfies μ-PL & gradient AB bounded

$\Delta x_k = E\|x_k - x^*\|^2$

$\Delta x_{k+1} \le (1 - 2\alpha_k\mu + \alpha_k^2 A^2)\Delta x_k + \alpha_k^2 B^2$

$\frac{\mu}{A^2}$ $\quad 1/2\mu k$

Polyak's

$x_{k+1} = x_k - \alpha_k\nabla f(x_k) + \beta_k(x_k - x_{k-1})$

Nesterov

$x_{k+1} = x_k - \alpha_k\nabla f(x_k + \beta_k(x_k - x_{k-1})) + \beta_k(x_k - x_{k-1})$

$1+\beta \quad 1 - \alpha Q \quad -\beta I - \alpha Q \quad Q = U\Lambda U^T \quad \pi$ $\quad$ 1 odd $j = \frac{i+1}{2}$

$I \quad 0$ $\quad$ 1 even $j = n + \frac{i}{2}$

$\alpha_k = \frac{1}{M}$ $\quad \beta_k = \frac{\sqrt{\frac{M}{m}} - 1}{\sqrt{\frac{M}{m}} + 1}$

Optimality $\frac{x^T A x}{2} - e_1^T x \to$ Tridiagonal $-1\ 2\ -1$

$x_i^* = 1 - \frac{i}{d+1}$

$\|x_k - x^*\|^2 \ge \sum_{j=k+1}^d x_j^{*2}$

$\ge \frac{1}{8}\|x_0 - x^*\|^2$ $\quad k < d/2$

Newton's Method.

$f(x) = f(x_l) + \langle \nabla f(x_l), x - x_l \rangle + \frac{1}{2}(x - x_l)^T \nabla^2 f(x_l)(x - x_l) + o(\|x - x_l\|^2)$

$x_{l+1} = x_l - [\nabla^2 f(x_l)]^{-1} \nabla f(x_l)$

Local Convergence of Newton

(A) x^* unique with $\nabla^2 f(x^*) \succeq mI$

(B) $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \le L\|x - y\|$

__Theorem__ Under A+B given x_0 s.t. $\|x_0 - x^*\| \le m/3L$

i) $\|x_{l+1} - x^*\| \le \frac{3L}{2m}\|x_l - x^*\|^2$

$x_{l+1} - x^* = x_l - x^* - \nabla^2 f(x_l)^{-1} \int_0^1 \nabla^2 f(x^* + t(x_l - x^*))(x_l - x^*)\,dt$

__Self Concordant__

$f^n: R \to R \quad |f'''(x)| \le 2(f''(x))^{3/2}$

$f^n: f: R^n \to R$ is S.C. if $\tilde{f}(t) = f(x + tv)$ is SC f^n of t $\forall x \in \mathrm{dom} f$ & $\forall v$

· Sum of SC f^n is SC.

Composition. $f: R^n \to R$ is S.C. $A \in R^{n \times m}$ $b \in R^n$ $f(Ax + b)$ is S.C.

Newton's $\lambda(x) = [\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)]^{1/2}$

$= \sup_{v \ne 0} \frac{-v^T \nabla f(x)}{(v^T \nabla^2 f(x) v)^{1/2}}$

$\nabla f(x^+) - \nabla f(x) = \int_0^1$

wlog $\nabla^2 f(x) = I$ $\quad \lambda(x^+) \le \frac{1}{1 - \lambda(x)}\|\nabla f(x^+)\|$ $\quad x^+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$

$\nabla f(x^+) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(x^+ - x))^T (x^+ - x)\,dt$

$\| \int_0^1 \frac{1}{(1 - t\lambda(x))^2} - 1 \, dt \| \nabla f(x)$

ADMM $\min f(x) + g(z)$

s.t. $Ax + Bz = c$

$L_\rho(x, y, z) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2$

$u = y/\rho$

$L_\rho(x, z, u) = f(x) + g(z) + \frac{\rho}{2}\|Ax + Bz - c + u\|^2$

$x_{k+1} = \arg\min_x \left(f(x) + \frac{\rho}{2}\|Ax + Bz_k - c + u_k\|^2\right)$

$z_{k+1} = \arg\min_z \left(g(z) + \frac{\rho}{2}\|Ax_{k+1} + Bz - c + u_k\|^2\right)$

$u_{k+1} = u_k + (Ax_{k+1} + Bz_{k+1} - c)$

$g = I_c(x > 0$

$x_{k+1} = \arg\min_x f(x) + \frac{\rho}{2}\|x - z_k + u_k\|_2^2$

$z_{k+1} = (x_{k+1} + u_k)_+ \quad u_{k+1} = u_k + x_{k+1} - z_{k+1}$

$P + \rho I \quad A^T \quad x \qquad q - \rho(z_k - u^k) = 0$

$A \quad 0 \quad \gamma \qquad b$

__ADAGRAD__ $\alpha_k = \frac{R}{\sqrt{\sum_{s=1}^k \|\nabla f(x_s)\|^2}}$

$\bar{x}^T = \frac{1}{T}\sum_{k=1}^T x_k$

$f(\bar{x}^T) - f(x^*) \le \frac{1}{T}\sum_{k=1}^T \nabla f(x_k)^T (x_k - x^*)$

$x = x_k \quad x_{k+1} = P_\Omega[x_k - \alpha_k \nabla f(x_k)]$

$\langle \nabla f(x_k), x_{k+1} - x^* \rangle \le \frac{2}{2\alpha_k}\langle x_k - x_{k+1}, x_{k+1} - x^* \rangle$

$2ab = (a+b)^2 - a^2 - b^2$

$\sum_{t=1}^T \frac{a_t}{\sqrt{\sum_{s=1}^t a_s}} \le 2\sqrt{\sum_{t=1}^T a_t}$

$f(x_k) - f(x^*) \le \langle \nabla f(x_k), x_k - x^* \rangle - \frac{1}{2M}\|\nabla f(x_k)\|^2$

Young's inequality

$ab \le \frac{\lambda a^2}{2} + \frac{b^2}{2\lambda}$

$\alpha \in 0, 0.5 \quad \beta\ 0, 1 \quad t = 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$

$t := \beta t$

$i_L(x) = \begin{cases} 0 & x \in L \\ \infty \end{cases}$

$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \frac{mM}{m+M}\|x - y\|^2 + \frac{1}{M+m}\|\nabla f(x) - \nabla f(y)\|^2$

$f(x+t) = f(x) + \nabla f(x + \gamma t)^T t = f(x) + \nabla f(x)^T t + o(\|t\|)$

$= f(x) + \nabla f(x)^T t + \frac{t^T}{2}\nabla^2 f(x + \gamma t)t$

$\gamma \in (0, 1)$

$f(x)$

$= f(x_0) + \int_0^1 \nabla f(x_0 + t(x - x_0))^T (x - x_0)\,dt$

$= f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}\int_0^1 (x - x_0)^T \nabla^2 f(x_0 + t(x - x_0))(x - x_0)\,dt.$