CS 726 Cheat Sheet Midterm

I. Setting the Stage

1. $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$; $\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^d x_i^2}$; $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$;
$\|x\|_1 = \sum_{i=1}^d |x_i|$; $\|x\|_\infty = \max_i |x_i|$; $\|z\|_* = \sup_{x\in\mathbb{R}^d, \|x\|=1} \langle z, x\rangle$

Prop 1.1 $|\langle z, x\rangle| \le \|z\|_* \cdot \|x\|$

dual norm $\|\cdot\|_*$ $\quad \frac{1}{p}+\frac{1}{q}=1$, $\|\cdot\|_p$ dual $\|\cdot\|_q$ $p \ge 1$

$\forall p \ge 1\ \forall r > p$: $\|x\|_r \le \|x\|_p \le d^{1/p - 1/r}\|x\|_r$

2. **Def 2.1** $\mathcal{X}$ is convex if $x, y \in \mathcal{X}$ and $\lambda \in (0,1)$:

Obs Eq $\quad (1-\lambda)x + \lambda y = x + \lambda(y-x) \in \mathcal{X}$

Def 3.1 $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is lower semicontinuous at $\bar{x} \in \mathbb{R}^d$ if $\liminf_{x\to\bar{x}} f(x) = f(\bar{x})$ and lower semicontinuous on $\mathbb{R}^d$ if it is lower semicontinuous at every $\bar{x} \in \mathbb{R}^d$

Def 3.2 $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is Lipschitz-continuous on a set $\mathcal{X} \subseteq \mathbb{R}^d$ if there is $M < \infty$ s.t. $\forall x, y \in \mathcal{X}$, $|f(x) - f(y)| \le M\|x - y\|$

Def 3.3 $f: \mathbb{R}^d \to \mathbb{R}$ is L-smooth wrt $\|\cdot\|$ on $\mathcal{X}$ if $\exists L < \infty$ s.t. $\forall x, y \in \mathcal{X}$: $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|$

Def 3.4 $f: \mathbb{R}^d \to \mathbb{R}$ is (K, L)-weakly smooth for $K \in [1, 2]$ if $\exists L < \infty$ $\forall x, y \in \mathcal{X}$: $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|^{K-1}$

Def 3.6 $f: \mathbb{R}^d \to \mathbb{R}$ is convex if $\forall x, y \in \mathbb{R}^d\ \forall \lambda \in (0,1)$,
$f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y)$

Lemma 3.7 $f: \mathbb{R}^d \to \mathbb{R}$ is convex iff its epigraph, i.e.,
$\text{epi}(f) = \{(x, a) : x \in \mathbb{R}^d, a \in \mathbb{R}, f(x) \le a\}$ is convex

Lemma 3.8 $f: \mathbb{R}^d \to \mathbb{R}$ differentiable, then f convex iff $\forall x, y \in \mathbb{R}^d$,
$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle$

3.1. $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ continuously differentiable $(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x, y \in \mathbb{R}^d)$
$\|x - y\| \le \delta \Rightarrow \|\nabla^2 f(x) - \nabla^2 f(y)\|_* \le \varepsilon$ (twice)

operator norm for symmetric A: $\|A\|_2 = \sup_{x\in\mathbb{R}^d, \|x\|_2=1} \|Ax\|_2$

Thm 3.9 (Taylor) Let $f: \mathbb{R}^d \to \mathbb{R}$ continuously differentiable. $\forall x, y \in \mathbb{R}^d$:

(i) $f(y) = f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y - x\rangle\, dt$

(ii) $\exists t \in (0,1)$ s.t. $f(y) = f(x) + \langle \nabla f(x + t(y-x)), y - x\rangle$ (MVT)

If f twice continuously differentiable,

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

(iv) $\exists t \in (0,1)$ s.t. $f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle \nabla^2 f(x + t(y-x))(y-x), y-x\rangle$

Lemma 3.10 $f: \mathbb{R}^d \to \mathbb{R}$ L-smooth. $\forall x, y \in \mathbb{R}^d$, both hold: $f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$; $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y - x\|^2$

Lemma 3.11 $f: \mathbb{R}^d \to \mathbb{R}$ twice continuously differentiable function. If $\forall x \in \mathbb{R}^d$
$\|\nabla^2 f(x)\|_{p,q} = \sup_{y:\|y\|_p=1} \|\nabla^2 f(x)\, y\|_q \le L < \infty$, $\frac{1}{p}+\frac{1}{q}=1$, $p \ge 1$,
then f is L-smooth wrt $\|\cdot\|_p$

Lemma 3.12 $f: \mathbb{R}^d \to \mathbb{R}$ twice continuously differentiable, then f is L-smooth wrt $\|\cdot\|_2$ iff $-LI \preceq \nabla^2 f(x) \preceq LI$

Def 3.13 $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ strongly convex with modulus $\mu > 0$ if $\forall x, y \in \mathbb{R}^d$, $\lambda \in (0,1)$: $f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y) - \frac{\lambda(1-\lambda)\mu}{2}\|y - x\|^2$

Lemma 3.14 $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ continuously differentiable, then f is μ-strongly convex iff $\forall x, y \in \mathbb{R}^d$: $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$

Cor f twice continuously differentiable, then μ-strongly convex wrt $\|\cdot\|_2$ iff $\nabla^2 f(x) \succeq \mu I$

Def 3.15 $p \ge 2$, $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ p-uniformly convex with modulus $\mu > 0$ wrt. $\|\cdot\|$ if $\forall x, y \in \mathbb{R}^d$, $\lambda \in (0,1)$: (p,μ)-uniformly convex
$f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y) - \frac{\lambda(1-\lambda)\mu}{p}\|y - x\|^p$

Cor f is (p, μ)-uniformly convex, then $\forall x, y \in \mathbb{R}^d$
$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{p}\|y - x\|^p$

Def 3.17 $f: \mathbb{R}^d \to \mathbb{R}$ satisfy (r, μ)-sharp or Łojasiewicz inequality on $\mathcal{X}$ if $\mathcal{X}^* = \arg\min_{x\in\mathbb{R}^d} f(x)$ is nonempty and $\exists \mu > 0$, $r > 0$, $\forall x \in \mathcal{X}$: $f(x) - \min_y f(y) \ge \frac{\mu}{r}\,\text{dist}(x, \mathcal{X}^*)^r$ where $\text{dist}(x, \mathcal{X}^*) = \min_{x^* \in \mathcal{X}^*} \|x - x^*\|$

Cor μ-strongly convex $\Rightarrow$ $(2, \mu)$-sharp

Def 3.18 $f: \mathbb{R}^d \to \mathbb{R}$ differentiable satisfy PŁ condition on $\mathcal{X}$ if $\mathcal{X}^*$ is nonempty and $\exists \mu, r > 0$ s.t. $\forall x \in \mathcal{X}$:
$f(x) - \min_y f(y) \le \frac{\mu}{r}\|\nabla f(x)\|^r$

Local and Global Solns

Def 4.1 $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ differentiable, $\mathcal{X}$ closed and convex. $x \in \mathcal{X}$
stationary point for (P) if $\forall y \in \mathcal{X}$, $\langle \nabla f(x), y - x\rangle \ge 0$
iff $\exists x \in \mathcal{X}$ s.t. $\nabla f(x) = 0$

Def 4.2 Given (P) and $x^* \in \mathcal{X}$,

- x^* local minimum if $\exists$ nbhd N_{x^*} $\forall x \in N_{x^*} \cap \mathcal{X}$, $f(x) \ge f(x^*)$
- x^* global soln if $\forall x \in \mathcal{X}$, $f(x) \ge f(x^*)$

Thm 4.3 (1st-order Necessity Optimality Conditions) f continuously differentiable and $\mathcal{X}$ closed and convex; if $x^* \in \mathcal{X}$ global soln, x^* must be a stationary point;
If $x^* \in \mathcal{X}$ local soln, $\exists N$ of x^* s.t. $\forall y \in N \cap \mathcal{X}$, $\langle \nabla f(x^*), y - x^*\rangle \ge 0$

Thm 4.4 (2nd-order Necessary and Sufficient Conditions for Unconstrained) f twice continuously differentiable, $\mathcal{X} \equiv \mathbb{R}^d$:

- If $x^* \in \mathbb{R}^d$ local soln, then $\|\nabla f(x^*)\|_2 = 0$ and $\nabla^2 f(x^*) \succeq 0$
- If $\exists x^* \in \mathbb{R}^d$ s.t. $\|\nabla f(x^*)\|_2 = 0$ and $\nabla^2 f(x^*) \succ 0$, then x^* strict local minimizer

Thm 4.5 (Optimality Conditions for Convex) f convex, $\mathcal{X}$ closed and convex:

(i) Every local soln is also global

(ii) set of solns is convex

(iii) If f differentiable, then x^* global soln iff x^* stationary point

Thm 4.6 (Optimality Conditions for Convex Prob Strongly Convex Obj)
f μ-strongly convex and continuous on its domain; $\mathcal{X}$ closed, convex, and non-empty intersection w/ effective domain of f, then soln to (P) is attained and unique.

II. Basic Descent Methods $\quad x_{k+1} = x_k + a_k z_k$

Def 1.1 $z \in \mathbb{R}^d$ is a descent direction of $f: \mathbb{R}^d \to \mathbb{R}$ at x if
$\exists t > 0$ s.t. $f(x + t'z) \le f(x)$ $\forall 0 < t' \le t$

Prop 1.2 $f: \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable (in nbhd of $x \in \mathbb{R}^d$), then $z \in \mathbb{R}^d$ s.t. $\langle \nabla f(x), z\rangle < 0$ is a descent direction for f at x

Gradient Descent: $\nabla f(x_k) \ne 0$, $x_{k+1} = x_k - \lambda_k \nabla f(x_k)$

Lemma 2.1 $f: \mathbb{R}^d \to \mathbb{R}$ L-smooth wrt $\|\cdot\|_2$ and $x_0 \in \mathbb{R}^d$, consider GD with step size $\lambda_k = \lambda \in (0, \frac{2}{L}]$. For all $k \ge 0$, $f(x_{k+1}) - f(x_k) \le -\frac{\lambda}{2}\|\nabla f(x_k)\|_2^2$
Further if $f(x) \ge f^* > -\infty$, $\forall x \in \mathbb{R}^d$, $k \ge 0$, $\min_{i\le k}\|\nabla f(x_i)\|_2^2 \le \frac{2(f(x_0) - f^*)}{\lambda(k+1)}$
so $\forall \varepsilon > 0$, GD takes at most $k = \left\lfloor \frac{2(f(x_0) - f^*)}{\lambda\varepsilon^2} \right\rfloor$ iterations to construct x_i, $i \in [1, k]$ s.t. $\|\nabla f(x_i)\|_2 \le \varepsilon$

Lemma 2.2 $f: \mathbb{R}^d \to \mathbb{R}$ convex and L-smooth wrt $\|\cdot\|_2$ and $x_0 \in \mathbb{R}^d$, consider GD with step size $\lambda_k = \lambda \in (0, 1/L]$. Then $\forall k \ge 0$,
$f(x_{k+1}) - f(x^*) \le \frac{\|x_0 - x^*\|_2^2}{2\lambda(k+1)}$. So $\forall \varepsilon > 0$, GD after at most $k = \left\lceil \frac{\|x_0 - x^*\|_2^2}{2\lambda\varepsilon} \right\rceil$ iterations we have $f(x_k) - f(x^*) \le \varepsilon$

Lemma 2.3 $f: \mathbb{R}^d \to \mathbb{R}$ μ-strongly convex and L-smooth wrt $\|\cdot\|_2$ and $x_0 \in \mathbb{R}^d$, GD w/ step size $\lambda_k = \lambda \in (0, 1/L]$. $\forall k \ge 0$,
$\|x_k - x^*\|_2^2 \le (1 - \lambda\mu)^k \|x_0 - x^*\|_2^2$
So $\forall \varepsilon > 0$, GD after at most $k = \max\left\{0, \left\lceil \frac{2}{\lambda\mu} \log\left(\frac{\|x_0 - x^*\|_2}{\varepsilon}\right)\right\rceil\right\}$ iterations we have $\|x_k - x^*\|_2 \le \varepsilon$

Cor 2.4 $f:\mathbb{R}^d\to\mathbb{R}$ μ-strongly convex and L-smooth wrt $\|\cdot\|_2$ and $x_0\in\mathbb{R}^d$, consider GD w/ step size $\lambda_k=\lambda\in(0,1/L]$. Then,

- After at most $k=\max\{0,\lceil\frac{2}{\lambda\mu}\log(\frac{\|x_0-x^*\|_2}{\varepsilon})\rceil\}$ iterations we have $f(x_k)-f(x^*)\le\varepsilon$
- After at most $k=\max\{0,\lceil\frac{2}{\lambda\mu}\log(\frac{L\|x_0-x^*\|_2}{\varepsilon})\rceil\}$ iterations we have $\|\nabla f(x_k)\|_2\le\varepsilon$

II. Nesterov Acceleration

1. Smooth and Convex Setting

Algo 1 Nesterov-Smooth(x_0, f, L, K)

Init: $a_0=A_0=\frac{1}{L}$, $v_0=x_0-a_0\nabla f(x_0)$, $y_0=x_0-\frac{1}{L}\nabla f(x_0)$
for $k=1$ to k do
  $a_k=\frac{k+2}{L}$, $A_k=a_k+A_{k-1}$
  $x_k=\frac{A_{k-1}}{A_k}y_{k-1}+\frac{a_k}{A_k}v_{k-1}$
  $v_k=v_{k-1}-a_k\nabla f(x_k)$
  $y_k=x_k-\frac{1}{L}\nabla f(x_k)$
end for
return y_k

$f(y_k)-f(x^*)\le G_k\le\frac{2L\|x^*-x_0\|_2^2}{(k+1)(k+4)}$ choosing $a_k=\frac{k+2}{2L}$

2. Smooth and Strongly Convex Optimization

2.1 Restart

Algo 2 Nesterov-Restart(x_0, f, L, μ, R)

Init: $\bar{x}_0=x_0$
for $r=1$ to R do
  $\bar{x}_r=$ Nesterov-Smooth$(\bar{x}_{r-1}, f, L, \lceil\sqrt{\frac{8L}{\mu}}\rceil)$
end for
return $\bar{x}_R$

Motivation: Lemma 3.14 w/ x^*

Lemma 2.1 Let f smooth, convex, $(2,\mu)$-sharp, $\mathcal{X}^*=\operatorname{argmin}_{x\in\mathbb{R}^d}f(x)$. Given $x_0\in\mathbb{R}^d$, consider running Algo 2 for $R\ge1$ iterations. Then given $\varepsilon>0$, if $R\ge2\log_2(\frac{\operatorname{dist}(x_0,\mathcal{X}^*)}{\varepsilon})$, we have $\operatorname{dist}(\bar{x}_R,\mathcal{X}^*)\le\varepsilon$.

The total number of operations is $O(\sqrt{\frac{L}{\mu}}\log_2(\frac{\operatorname{dist}(x_0,\mathcal{X}^*)}{\varepsilon}))$

2.2 Direct Algorithm

Algo 3 Nesterov-Smooth-Strongly-Convex(x_0, f, L, μ, K)

Init: $a_0=A_0=\frac{1}{L-\mu}$, $v_0=\frac{x_0+\mu a_0x_0-a_0\nabla f(x_0)}{1+\mu A_0}$, $y_0=x_0-\frac{1}{L}\nabla f(x_0)$
for $k=1$ to K do
  $a_k=$ positive soln to $\frac{a_k^2}{(A_{k-1}+a_k)(1+\mu(A_{k-1}+a_k))}=\frac{1}{L}$
  $A_k=a_k+A_{k-1}$
  $x_k=\frac{A_{k-1}}{A_k(1-\mu/L)}y_{k-1}+(1-\frac{A_{k-1}}{A_k(1-\mu/L)})v_{k-1}$
  $v_k=\frac{(1+\mu A_{k-1})v_{k-1}+a_k(\mu x_k-\nabla f(x_k))}{1+\mu A_k}$
  $y_k=x_k-\frac{1}{L}\nabla f(x_k)$
end for
return y_k

$f(y_k)-f(x^*)\le G_k\le\frac{\|x^*-x_0\|_2^2}{2A_k}$; $f(y_k)-f(x^*)<(1-\sqrt{\frac{\mu}{L}})^k\frac{(L-\mu)\|x^*-x_0\|_2^2}{2}$

$k\ge\sqrt{\frac{L}{\mu}}\log(\frac{(L-\mu)\|x^*-x_0\|_2^2}{2\varepsilon})$ then $f(y_k)-f(x^*)\le\varepsilon$

IV. Appendix

Inequalities

triangle: $\|x+y\|\le\|x\|+\|y\|$ (L^2, L^p by Minkowski)
reverse triangle: $\|x-y\|\ge|\|x\|-\|y\||$
Jensen: $f(\sum_{i=1}^k a_ix_i)\le\sum_{i=1}^k a_if(x_i)$ given f convex, $\lambda_i\ge0$ s.t. $\sum_{i=1}^k a_i=1$ (Cauchy-Schwarz) $p=q=2$
Hölder: $\|fg\|_1\le\|f\|_p\|g\|_q$ given $p,q\in[1,\infty]$, $\frac{1}{p}+\frac{1}{q}=1$
Young: $ab\le\frac{a^p}{p}+\frac{b^q}{q}$ (equal iff $a^p=b^q$) given $a,b\ge0$, $p>1, q>1$, $\frac{1}{p}+\frac{1}{q}=1$

Directional Derivative: $\nabla_v f(x)=\nabla f(x)\cdot v=\lim_{t\to0}\frac{f(x+tv)-f(x)}{t}$
Complete Square: $\|u\|_2^2+\|v\|_2^2-2\langle u,v\rangle=\|u-v\|_2^2\le n\|u-v\|_\infty^2$
$\|u\|_2^2+\|v\|_2^2-\langle u,v\rangle\le$

HW Pf holds, $\forall x\in\mathbb{R}^d$: $\frac{1}{2L}\|\nabla f(x)\|_2^2\le f(x)-f(x^*)\le\frac{L}{2}\|x^*-x\|_2^2$

2.6 f L-smooth convex fn, $L\in(0,\infty)$
(a) $\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|_2^2\le f(y)-f(x)-\langle\nabla f(x),y-x\rangle$
(b) $f(x_{k+1})\le f(x_k)-\frac{1}{2L}(\|\nabla f(x_{k+1})\|_2^2+\|\nabla f(x_k)\|_2^2)$
(c) $\|\nabla f(x_{k+1})\|_2\le\|\nabla f(x_k)\|_2$
(d) $\|\nabla f(x_k)\|_2^2\le\frac{L(f(x_0)-f(x_k))}{k}$

1.3 $\mathcal{X},\mathcal{Y}$ convex, $\mathcal{X}\cap\mathcal{Y}$ must convex, $\mathcal{X}\cup\mathcal{Y}$ may not
1.5. f Conv$_x\to f(a^Tx)$ convex; $f(x)=\|x\|$ convex; $f(x)=\frac{1}{2}\|x\|_2^2$ convex
1.7. $x^TAx\ge\lambda_{\min}\|x\|_2^2$ $x^TAx\le\lambda_{\max}\|x\|_2^2$
3.1. f differentiable, $x^*\in\operatorname{argmin}_{x\in\mathbb{R}^d}f(x)$
(i) if f L-smooth, $\frac{1}{2L}\|\nabla f(x)\|_2^2\le f(x)-f(x^*)\le\frac{L}{2}\|x-x^*\|_2^2$
(ii) iff f L-smooth and convex, $\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|_2^2\le f(y)-f(x)-\langle\nabla f(x),y-x\rangle$
(iii) if f L-smooth and μ-strongly convex,
$\langle\nabla f(x)-\nabla f(y),x-y\rangle\ge\frac{\mu L}{\mu+L}\|x-y\|_2^2+\frac{1}{\mu+L}\|\nabla f(x)-\nabla f(y)\|_2^2$
3.4. Equivalent Nesterov with $x_k=y_{k-1}+\frac{a_k}{A_k}(\frac{A_{k-1}}{a_{k-1}}-1)(y_{k-1}-y_{k-2})$

Taylor Thm in Exam Review $f(y)=f(x)+\langle\nabla f(x),y-x\rangle+\int_0^1\langle\nabla f(x+t(y-x)),y-x\rangle dt$

Additions:

- By Taylor, μ-str cvx wrt $\|\cdot\|_2\iff\nabla^2f(x)\succeq\mu I\ \forall x$
- Smooth + non-smooth $\not\Rightarrow$ smooth cvx + str cvx $\not\Rightarrow$ str cvx
- Bregman Divergence: $D_\psi(x,y)=\psi(x)-\psi(y)-\langle\nabla\psi(y),x-y\rangle$ (1st order approx error) $\forall x,y,z\in\mathbb{R}^d$
  3-pt identity: $D_\psi(x,y)+D_\psi(z,y)+\langle\nabla\psi(z)-\nabla\psi(y),x-z\rangle=D_\psi(x,z)$
  $h(x)=\langle z,x\rangle+D_\psi(x,b)$; $y=\operatorname{argmin}_{x\in\mathcal{X}}h(x)$; $\forall x\in\mathcal{X}$
  $h(x)\ge\langle z,y\rangle+D_\psi(y,b)+D_\psi(x,y)$

CS726 Cheat Sheet Final

IV. Constrained, Projection-Based Optimization

Def 1.1 (Normal Cone) $\mathcal{X}$ closed convex. $\forall x \in \mathcal{X}$,
$N_{\mathcal{X}}(x)$ defined $= \{z \in \mathbb{R}^d : \langle z, y - x\rangle \le 0, \ \forall y \in \mathcal{X}\}$
Local solution: $(x^* \in \mathcal{X}) \ -\nabla f(x^*) \in N_{\mathcal{X}}(x^*), \ \forall y \in \mathcal{X}$
(Stationary) $\langle -\nabla f(x^*), y - x^*\rangle \le 0 \Leftrightarrow \langle \nabla f(x^*), y - x^*\rangle \ge 0$

Thm 1.2 Given (P), $x^* \in \mathcal{X}$ local sol $\Rightarrow -\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$;
f convex $\Rightarrow -\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$ iff x^* global sol
f str. cvx. $\Rightarrow \exists x^*$ s.t. $-\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$ unique sol

2. Def Euclidean Projection $P_{\mathcal{X}}(x) = \operatorname{argmin}_{y \in \mathcal{X}} \|y - x\|_2$

Lemma 2.1 $\mathcal{X}$ closed cvx, $\forall x \in \mathbb{R}^d$, $P_{\mathcal{X}}(x)$ exists & unique
$-(P_{\mathcal{X}}(x) - x) \in N_{\mathcal{X}}(x)$, $\forall y \in \mathcal{X}: \langle P_{\mathcal{X}}(x) - x, y - P_{\mathcal{X}}(x)\rangle \ge 0$

Example 2.2 $\mathcal{X} = \{x \in \mathbb{R}^d : x \ge 0\}$, $P_{\mathcal{X}}(x) = \max\{x, 0\}$ element-wise

Example 2.3 $\mathcal{X} = \{x \in \mathbb{R}^d : a \le x \le b\}$, $P_{\mathcal{X}}(x) = \max\{a, \min\{x, b\}\}$

Example 2.4 $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$, $P_{\mathcal{X}}(x) = \begin{cases} x & x \in \mathcal{X} \\ x/\|x\|_2 & x \notin \mathcal{X} \end{cases}$

Example 2.5 ℓ_2 ball, prob simplex: compute $P_{\mathcal{X}}(x)$ in $O(d \log d)$

Lemma 2.6 $\mathcal{X}$ closed cvx, $\langle P_{\mathcal{X}}(x) - y, x - y\rangle \ge 0 \ \forall y \in \mathcal{X}$
equality iff $y = P_{\mathcal{X}}(x)$

Lemma 2.7 $\mathcal{X}$ closed cvx, $P_{\mathcal{X}}$ is nonexpansive $\forall x, y \in \mathbb{R}^d$
$\|P_{\mathcal{X}}(x) - P_{\mathcal{X}}(y)\|_2 \le \|x - y\|_2$

3. PGD
$$x_{k+1} = \operatorname*{argmin}_{y \in \mathcal{X}} \{f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \tfrac{L}{2}\|y - x_k\|_2^2\}$$
$$= \operatorname*{argmin}_{y \in \mathcal{X}} \{\tfrac{L}{2}\|y - x_k + \tfrac{1}{L}\nabla f(x_k)\|_2^2\}$$
$$= P_{\mathcal{X}}(x_k - \tfrac{1}{L}\nabla f(x_k))$$
$$x_{k+1} = x_k - x_k + P_{\mathcal{X}}(x_k - \tfrac{1}{L}\nabla f(x_k)) = x_k - \tfrac{1}{L}\underbrace{L(x_k - P_{\mathcal{X}}(x_k - \tfrac{1}{L}\nabla f(x_k)))}$$

Def 3.1 $\mathcal{X}$ closed cvx, $\eta > 0$, gradient mapping defined by
$G_\eta(x) = \eta(x - P_{\mathcal{X}}(x - \tfrac{1}{\eta}\nabla f(x)))$
PGD: $x_{k+1} = x_k - \tfrac{1}{\eta} G_\eta(x_k)$

Lemma 3.2 $\mathcal{X}$ closed cvx, min L-smooth f, $\bar{x} = P_{\mathcal{X}}(x - \tfrac{1}{\eta}\nabla f(x))$
$(G_\eta(x) = \eta(x - \bar{x}))$. If for some $\varepsilon > 0$, $\|G_\eta(x)\|_2 \le \varepsilon$:
$-\nabla f(\bar{x}) \in N_{\mathcal{X}}(\bar{x}) + B(0, \varepsilon(\tfrac{L}{\eta} + 1))$
$\Leftrightarrow (\forall y \in \mathcal{X}): \langle \nabla f(\bar{x}), y - \bar{x}\rangle \ge -\varepsilon(\tfrac{L}{\eta} + 1)\|y - \bar{x}\|_2$
$\Rightarrow (\forall y \in \mathcal{X} \cap B(\bar{x}, 1)): \langle \nabla f(\bar{x}), y - \bar{x}\rangle \ge -\varepsilon(\tfrac{L}{\eta} + 1)$
$G_\eta(x) = 0 \Rightarrow x = \bar{x}$ and $-\nabla f(x) \in N_{\mathcal{X}}$

Lemma 3.3 (Descent Lemma PGD) f L-smooth, $\mathcal{X}$ closed cvx,
$\bar{x} = P_{\mathcal{X}}(x - \tfrac{1}{\eta}\nabla f(x))$ $(x \in \mathcal{X})$, then $G_\eta(x) = \eta(x - \bar{x})$.
If $\eta \ge L$, then $f(\bar{x}) \le f(x) - \frac{1}{2\eta}\|G_\eta(x)\|_2^2$

Thm 3.4 (Convergence Nonconvex f PGD) f L-smooth, $\mathcal{X}$ closed cvx
$x_0 \in \mathcal{X}$, $\eta \ge L$, $f_* > -\infty$. $\min_{0 \le i \le k} \|G_\eta(x_i)\|_2^2 \le \frac{2\eta(f(x_0) - f_*)}{k+1}$
$\Rightarrow \forall \varepsilon > 0$, $k \ge \lceil \frac{2\eta(f(x_0) - f_*)((L/\eta) + 1)^2}{\varepsilon^2} \rceil$, $\exists \hat{x} \in \{x_1, \dots, x_{k+1}\}$
s.t. $-\nabla f(\hat{x}) \in N_{\mathcal{X}}(\hat{x}) + B(0, \varepsilon)$

3.2 (Convergence convex f PGD) $f(x_{k+1}) - f(x^*) \le f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x^* - x_{k+1}\rangle$
$\langle \nabla f(x_k), x^* - x_{k+1}\rangle \ge \frac{\eta}{2}(\|x_{k+1} - x_k\|_2^2 - \|x^* - x_k\|_2^2 + \|x^* - x_{k+1}\|_2^2)$ $(x^* - x_{k+1}$ expand$)$
$f(x_{k+1}) - f(x^*) \le \frac{\eta\|x^* - x_1\|_2^2}{2(k+1)}$, $\forall \varepsilon > 0$, $f(x_k) - f(x^*) \le \varepsilon$
after $k = \lceil \frac{\eta\|x_0 - x^*\|_2^2}{2\varepsilon} \rceil$

V. Projection-Free (Frank-Wolfe) Method for Constrained Convex Opt

produce sparse sol.

2. Basic FW $v_k = \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x_k), x\rangle$; $x_{k+1} = \frac{A_{k-1}}{A_k} x_k + \frac{a_k}{A_k} v_k$

(Analysis) $f(x^*) \ge \min_{x \in \mathcal{X}} \{\frac{1}{A_k}\sum a_i(f(x_i) + \langle \nabla f(x_i), x - x_i\rangle)\}$ $(U_k = f(x_{k+1}))$
$\ge \frac{1}{A_k}\sum a_i f(x_i) + \frac{1}{A_k}\sum a_i \min_{x \in \mathcal{X}}\langle \nabla f(x_i), x - x_i\rangle$
$= \frac{1}{A_k}\sum a_i f(x_i) + \frac{1}{A_k}\sum a_i \langle \nabla f(x_i), v_i - x_i\rangle = L_k$
$A_k G_k - A_{k-1}G_{k-1} \le \frac{a_k^2 L}{2A_k}\|v_k - x_k\|_2^2 \le \frac{a_k^2 L}{2A_k} D^2$
$A_0 G_0 \le \frac{a_0^2 L}{2A_0} D^2$
$f(x_{k+1}) - f(x^*) \le G_k \le \frac{LD^2}{2A_k}\sum \frac{a_i^2}{A_i} \le \frac{2LD^2}{k+2}$ $\frac{a_i^2}{A_i}$ choose $a_i a_{i+1}$, $A_i = \frac{(i+1)(i+2)}{2}$

Lemma 3.1 f L-smooth, $\mathcal{X}$ closed cvx, alg access the feasible set $\mathcal{X}$ only via linear min oracle requires at least $\min\{\frac{d}{2}, \frac{LD^2}{16\varepsilon}\}$ iters to construct $\hat{x}$ s.t. $f(\hat{x}) - \min_{x \in \mathcal{X}} f(x) \le \varepsilon$ for $\varepsilon > 0$, Lower bound applies even if f str. cvx.

VI. Nonsmooth Convex Optimization

Def 2.1 f is subdifferentiable at x if $\exists g_x \in \mathbb{R}^d$ s.t. $\forall y \in \mathbb{R}^d$,
$f(y) \ge f(x) + \langle g_x, y - x\rangle$, g_x: subgrad of f at x.
Set of all subgrads of f at x called subdifferential
Set of f at x $(\partial f(x))$. ∂ linear for every x.

Fact 2.2 Every cvx lower semicont. fn is subdiff'ble on int dom f.

Thm 2.4 f cvx lower semicont. fn. Then f M-Lip. cont. wrt $\|\cdot\|$ on int dom f iff $(\forall x \in \text{int dom } f)(\forall g_x \in \partial f(x)): \|g_x\|_* \le M$

Projected Subgrad Descent f M-Lip. cont. wrt $\|\cdot\|_2$, not descent method
$$x_{k+1} = \operatorname*{argmin}_{y \in \mathcal{X}} \{a_k\langle g_k, y\rangle + \tfrac{1}{2}\|y - x_k\|_2^2\}$$

(Analysis) $f(x^*) \ge L_k := \frac{1}{A_k}\sum a_i(f(x_i) + \langle g_{x_i}, x^* - x_i\rangle)$ $g_{x_i} \in \partial f(x_i)$
$x_k^{out} = \frac{1}{A_k}\sum a_i x_i$ $U_k = \frac{1}{A_k}\sum a_i f(x_i) \ge f(x_k^{out})$ (Jensen)
$f(x_k^{out}) - f(x^*) \le G_k = \frac{1}{A_k}\sum a_i\langle g_{x_i}, x^* - x_i\rangle$
$h_k(y) := a_k\langle g_k, y\rangle + \frac{1}{2}\|y - x_k\|_2^2$, $x_{k+1} = \operatorname{argmin}_{y \in \mathcal{X}} h_k(x)$
h_k quadratic $\Rightarrow h_k(y) = h_k(x) + \langle \nabla h_k(x), y - x\rangle + \frac{1}{2}\|y - x\|_2^2$
(bound) $h(x^*) \ge h(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|_2^2$ ≥ 0 for x_{k+1}
$-a_k\langle g_{x_k}, x^* - x_k\rangle \le -\frac{1}{2}\|x^* - x_{k+1}\|_2^2 + \frac{1}{2}\|x^* - x_k\|_2^2 + \frac{a_k^2 M^2}{2}$
$A_k G_k = -\sum a_k\langle g_{x_k}, x^* - x_k\rangle \le \frac{1}{2}\|x^* - x_0\|_2^2 + \sum \frac{a_k^2 M^2}{2}$
$f(x_k^{out}) - f(x^*) \le \frac{\frac{1}{2}\|x^* - x_0\|_2^2 + \sum a_k^2 M^2}{A_k}$

Choose $a_k = C$, $f(x_k^{out}) - f(x^*) \le \frac{M\|x^* - x_0\|_2}{\sqrt{k+1}}$
Choose $a_k = \frac{D}{M\sqrt{k+1}}$, $f(x_k^{out}) - f(x^*) \le \frac{MD}{\sqrt{k+1}}$
Choose $a_k = \frac{D}{M\sqrt{k+1}}$, $f(x_k^{out}) - f(x^*) = O(\frac{MD\log(k+1)}{\sqrt{k+1}})$
Choose $a_k = \frac{1}{\sqrt{k+1}}$, $\dots = O(\frac{(\|x^* - x_0\|_2^2 + M^2)\log(k+1)}{\sqrt{k+1}})$

VII. Stochastic Convex Optimization

- f is M-Lip cont
- $\forall x \in \mathcal{X}$, estimate subgrad $\tilde{g}(x, \xi)$ $\xi \sim D$ i.i.d. rate $\frac{\log k}{k}$

properties: $\mathbb{E}_{\xi \sim D}[\tilde{g}(x, \xi)] = g_x$ $\exists g_x \in \partial f(x)$ (unbiased estimate)
$\mathbb{E}_{\xi \sim D}[\|\tilde{g}(x, \xi) - g_x\|_*^2] \le \sigma^2 < \infty$ (bounded variance)

S-PSubGD (Stochastic Projected Subgrad Descent)
$x_{k+1} = \operatorname{argmin}_{y \in \mathcal{X}} \{a_k\langle \tilde{g}(x_k, \xi_k), y\rangle + \frac{1}{2}\|y - x_k\|_2^2\}$
$x_k^{out} = \frac{1}{A_k}\sum a_i x_i$, $U_k = \frac{1}{A_k}\sum a_i f(x_i) \ge f(x_k^{out})$,
$L_k = \frac{1}{A_k}\sum a_i(f(x_i) + \langle g_{x_i}, x^* - x_i\rangle)$, $g_{x_i} \in \partial f(x_i)$
$\Rightarrow f(x_k^{out}) - f(x^*) \le G_k = -\frac{1}{A_k}\sum a_i\langle g_{x_i}, x^* - x_i\rangle$
$\Rightarrow -a_k\langle g_{x_k}, x^* - x_k\rangle = a_k\langle \tilde{g}(x_k, \xi_k), x_k - x^*\rangle + a_k\langle g_{x_k} - \tilde{g}(x_k, \xi_k), x_k - x^*\rangle$
$\Leftrightarrow a_k\langle \tilde{g}(x_k, \xi_k), x_k - x^*\rangle \le \underbrace{a_k\langle \tilde{g}(x_k, \xi_k) - g_{x_k}, x_k - x_{k+1}\rangle}_{=0 \text{ under } \mathbb{E}} + \underbrace{a_k\langle g_{x_k} - \tilde{g}(x_k, \xi_k), x_{k+1} - x^*\rangle}_{\le -\mathbb{E}[\cdot] \le a_k^2\sigma^2} + \frac{1}{2}\|x^* - x_k\|_2^2 - \frac{1}{2}\|x^* - x_{k+1}\|_2^2 + \frac{a_k^2 M^2}{2}$ (mgf M)
$\Rightarrow \mathbb{E}_{\{\xi_0, \dots, x_k\} \sim D^{k+1}}[f(x_k^{out}) - f(x^*)] \le \underbrace{\frac{\|x^* - x_0\|_2^2 + a_0^2(M^2 + \sigma^2)}{2A_k}}_{\mathbb{E}[G_k]}$

VIII. Second Order Methods

Basic Newton: $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$

Local quadratic convergence of Newton's method:
$\{x_k\}$ seq converges to x^*

1) Q-linear, if $\exists r \in (0,1)$ s.t. $\|x_{k+1} - x^*\| \le r \|x_k - x^*\|$ $\forall k$ large

2) Q-quad, if $\exists$ const $M>0$ s.t. $\|x_{k+1} - x^*\| \le M \|x_k - x^*\|^2$ $\forall k$ large

Thm 3.5 Suppose f twice cont diff and $\nabla^2 f(x)$ Lip-cont. in nbhd of x^*, s.t. $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$. x_k evolve basic Newton:

(i) x_0 "sufficiently close" to x^*, $\{x_k\}$ converges to x^*,

(ii) rate of convergence of $\{x_k\}$ quadratic

(iii) sequence of gradient norms $\{\|\nabla f(x_k)\|\}$ converges quadratically to 0.

Global converge $f(x_k) - f(x^*) \le \varepsilon$ after $C(L_H, m, M, x_0, x^*) + \log\log(\frac{1}{\varepsilon})$

(Damped) $x_{k+1} = x_k - a_k \nabla^2 f(x_k)^{-1} \nabla f(x_k)$ a_k: line search

For $\nabla^2 f(x_k) \succ 0$, consider Quasi-Newton:

$x_{k+1} = x_k - a_k B_k^{-1} \nabla f(x_k)$ where $B_k \succeq \delta I$ for some $\delta > 0$

$a_k = -B_k^{-1} \nabla f(x_k)$ satisfies Weak Wolfe conditions: $\exists 0 < c_1 < c_2 < 1$:

(WW1) $f(x_k + a_k d_k) \le f(x_k) + c_1 d_k \langle \nabla f(x_k), d_k \rangle$

(WW2) $\langle \nabla f(x_k + a_k d_k), d_k \rangle \ge c_2 \langle \nabla f(x_k), d_k \rangle$

(WW1) $f(x_k + a_k p_k) \le f(x_k) + c_1 a_k \nabla f_k^T p_k$

(WW2) $\nabla f(x_k + a_k p_k)^T p_k \ge c_2 \nabla f_k^T p_k$

$\Rightarrow f(x_{k+1}) \le f(x_k) - c_1 \frac{1-c_2}{L} \frac{\langle \nabla f(x_k), p_k \rangle^2}{\|p_k\|^2}$ f L-smooth, p_k direction

Thm 3.6 f twice cont. diff. $x_{k+1} = x_k + a_k d_k$, a_k satisfies WW $c_1 \le \frac{1}{2}$.
If the sequence $\{x_k\}$ converges to x^* s.t. $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$ and if satisfies $\lim_{k\to\infty} \frac{\|\nabla f(x_k) + \nabla^2 f(x_k) d_k\|}{\|d_k\|} = 0$ (super linear) then

(i) $a_k = 1$ is admissible $\forall k > k_0$ and (reverse: 3.7) (QN $a_k = 1$)

(ii) if $a_k = 1$ $\forall k > k_0$, $\{x_k\}$ converges to x^* superlinearly

Proof of Thm 3.5 Goal: if $\|x_0 - x^*\|$ suff. small, then

1) $\exists M > 0$ constant s.t. $\forall k$: $\|x_{k+1} - x^*\| \le M \|x_k - x^*\|^2$

2) $\exists M' > 0$ also const s.t. $\forall k$: $\|\nabla f(x_{k+1})\| \le M \|\nabla f(x_k)\|^2$

1) $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$; x_k enbhd of x^*, $\nabla^2 f$ L_H-Lip

$$\|x_{k+1} - x^*\| = \|x_k - x^* - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)\|$$

$$= \|(\nabla^2 f(x_k))^{-1} (\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k))\|$$

$$= \|(\nabla^2 f(x_k))^{-1} (\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \underbrace{\nabla f(x^*)}_{=0}))\|$$

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

$$\|x_{k+1} - x^*\| \le \|(\nabla^2 f(x_k))^{-1}\| \cdot \|\int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k)))(x_k - x^*) dt\|$$

$$\le \|(\nabla^2 f(x_k))^{-1}\| \cdot \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k))\| \|x^* - x_k\| dt$$

$$\le \|(\nabla^2 f(x_k))^{-1}\| \cdot \frac{L_H}{2} \|x^* - x_k\|^2 \quad (\le L_H \cdot t \|x^* - x_k\|)$$

$$\le 2 \|\nabla^2 f(x_k))^{-1}\| \cdot \frac{L_H}{2} \|x^* - x_k\|^2 \le M \|x^* - x_k\|^2 \quad (= L_H \|(\nabla^2 f(x^*))^{-1}\|)$$

2) $\nabla f(x_k) = \nabla^2 f(x_k) (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$

$x_{k+1} - x_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$

$\nabla^2 f(x_k)(x_{k+1} - x_k) = -\nabla f(x_k)$

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

$$= \|\int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k) dt - \nabla^2 f(x_k)(y_{k+1} - y_k)\|$$

$$\le \int_0^1 \|\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x_k)\| \cdot \|x_{k+1} - x_k\| dt$$

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

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

$$= M' \|\nabla f(x_k)\|^2 \quad M' = 2 L_H \|(\nabla^2 f(x^*))^{-1}\| \text{ const}$$

secant equation: $B_{k+1} s_k = y_k$

$s_k = x_{k+1} - x_k = a_k p_k$

$y_k = \nabla f_{k+1} - \nabla f_k$

curvature: $s_k^T y_k > 0$