Notation

The standard problem $(P)$ is $\min\{f(x) : x \in \Omega\}$.
The linear approximation is $l_f(y;x) = f(x) + \nabla f(x)^T(y-x)$.
The projection operator is $\Pi_\Omega(x) = \text{argmin}_y\{\|y-x\| : y \in \Omega\}$.
The normal cone is $N_\Omega(\bar{x}) = \{n \in \mathbb{R}^n : n^T(y-\bar{x}) \leq 0, y \in \Omega\}$.

## Convexity

**Theorem 0.1.** *(Weierstrass) If $S$ is compact and $f$ is continuous on $S$, then $(P)$ has a global minimum.*

**Corollary 0.1.** *If $S$ is closed and $f$ is continuous on $S$ and $\lim_{\|x\|\to\infty, x\in S} f(x) = \infty$ then $(P)$ has a global minimum.*

**Proposition 0.1.** *$x^*$ is a local minimum of $(P)$ and $f \in \mathcal{C}^1(\mathbb{R}) \implies \nabla f(x^*) = 0$.*

**Proposition 0.2.** *$x^*$ is a local minimum of $(P)$ and $f \in \mathcal{C}^2(\mathbb{R}) \implies \nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \geq 0$.*

**Definition 0.1.** $f$ is $\beta$**-strongly convex** ($\beta > 0$) on $C$ if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\beta}{2}t(1-t)\|x-y\|^2$$

for all $x, y \in C$ and $t \in (0,1)$.

**Proposition 0.3.** *$f$ is $\beta$-strongly convex iff $f - \frac{\beta}{2}\|\cdot\|^2$ is strongly convex.*

**Proposition 0.4.** *For $\Omega \subseteq \mathbb{R}^n$ convex, $f \in \mathcal{C}^1(\Omega)$ and $\beta \in \mathbb{R}$, TFAE:*
*(a) $f - \frac{\beta\|\cdot\|^2}{2}$ is convex*
*(b) $\forall x, y \in \Omega$, $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|y-x\|^2$*
*(c)$\forall x, y \in \Omega$, $[\nabla f(y) - \nabla f(x)]^T(y-x) \geq \beta\|y-x\|^2$*

**Proposition 0.5.** *For $\Omega \subseteq \mathbb{R}^n$ convex, $f \in \mathcal{C}^1(\Omega)$ and $M \in \mathbb{R}$, TFAE:*
*(a) $\frac{M}{2}\|\cdot\| - f$ is convex*
*(b) $\forall x, y \in \Omega$, $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{M}{2}\|y-x\|^2$*
*(c)$\forall x, y \in \Omega$, $[\nabla f(y) - \nabla f(x)]^T(y-x) \leq M\|y-x\|^2$*

**Proposition 0.6.** *TFAE*
*(1) $f$ is convex on $C$*
*(2) $\{x, t \in C \times \mathbb{R} : f(x) \leq t\}$ is convex*
*(2) $\{x, t \in C \times \mathbb{R} : f(x) < t\}$ is convex*

**Proposition 0.7.** *For $\Omega \subseteq \mathbb{R}^n$ convex and $f \in \mathcal{C}^1(\Omega)$ the following are equivalent:*
*(a) $f$ is (strictly) convex on $\Omega$*
*(b) $f(y)(>) \geq f(x) + \nabla f(x)^T(y-x), \forall x, y \in \Omega$ ($x \neq y$)*
*(c) $[\nabla f(y) - \nabla f(x)]^T(y-x)(>) \geq 0, \forall x, y \in \Omega$ ($x \neq y$)*

**Proposition 0.8.** *Assume $\Omega \subseteq \mathbb{R}^n$ is convex, $f \in \mathcal{C}^1(\Omega)$ is convex on $\Omega$. TFAE for $\bar{x} \in \mathbb{R}^n$:*
*(a) $\bar{x}$ is a global minimum of $f$ on $\Omega$*
*(b) $\bar{x}$ is a local minimum of $f$ on $\Omega$*
*(c) $\nabla f(\bar{x})^T(x-\bar{x}) \geq 0, \forall x \in \Omega$*

*Remark* 0.1. If $\bar{x} \in \text{int}(\Omega)$ then $(c) \iff \nabla f(\bar{x}) = 0$.

**Proposition 0.9.** *If $\Omega \subseteq \mathbb{R}^n$ is convex, $f \in \mathcal{C}^1(\Omega)$ is strictly convex on $\Omega$ then $f$ has at most one global minimum.*

**Proposition 0.10.** *If $\Omega$ is convex, $f \in \mathcal{C}^1(\Omega)$, $\nabla f(\cdot)$ is $L-$Lipschitz continuous on $\Omega$, then*
*(1) $-\frac{L}{2}\|x-y\|^2 \leq f(y) - [f(x) + \nabla f(x)^T(y-x)] \leq \frac{L}{2}\|x-y\|^2$*
*(2) $-L\|x-y\|^2 \leq [\nabla f(y) - \nabla f(x)]^T(y-x) \leq L\|x-y\|^2$*

**Proposition 0.11. [**IMPORTANT**]** *If $\Omega \subseteq \mathbb{R}^n$ is closed and convex, and $f \in \mathcal{C}^1(\Omega)$ is $\beta-$strongly convex. Then, $f_* = \inf_x\{f(x) : x \in \Omega\}$ has a unique optimal solution $x^*$ and*

$$f(x) \geq f_* + \frac{\beta}{2}\|x-x^*\|^2, \forall x \in \Omega$$

## Projections

**Corollary 0.2.** *We have*
*(1) $\Pi_\Omega$ is well-defined*
*(2) $x^* = \Pi_\Omega(x) \iff \langle y - x^*, x - x^*\rangle \leq 0, \forall y \in \Omega$*
*(3) $\langle x_1 - x_2, \Pi_\Omega(x_1) - \Pi_\Omega(x_2)\rangle \geq \|\Pi_\Omega(x_1) - \Pi_\Omega(x_2)\|^2$*

*Remark* 0.2. f $f$ is convex, then $(P)$ is equivalent to $0 \in \nabla f(x^*) + N_\Omega(x^*)$. This follows from the fact that the optimality condition for the problem is

$$\nabla f(x^*)^T(y-x^*) \geq 0, \forall y \in \Omega \iff -\nabla f(x^*) \in N_\Omega(x^*).$$

**Proposition 0.12.** *Assume $\Omega \subseteq \mathbb{R}^n$ convex and $f \in \mathcal{C}^1(\Omega)$. Then,*
*(a) $\nabla^2 f(x) \geq 0, \forall x \in \Omega \implies f$ is convex on $\Omega$.*
*(b) $f$ is convex on $\Omega$ and int $\Omega \neq \emptyset \implies \nabla^2 f(x) \geq 0, \forall x \in \Omega$.*
*(c) $\nabla^2 f(x) > 0, \forall x \in \Omega \implies f$ is strictly convex on $\Omega$.*

**Corollary 0.3.** *Assume $\Omega \subseteq \mathbb{R}^n$ is convex, $f \in \mathcal{C}^2(\Omega)$. For $m, M \in \mathbb{R}$, we have*

$$\begin{aligned} & mI \leq \nabla^2 f(x) \leq MI \\ \iff & f(\cdot) - \frac{m}{2}\|\cdot\|^2 \text{ and } \frac{M}{2}\|\cdot\|^2 - f(\cdot) \text{ are convex} \\ \iff & \frac{m}{2}\|y-x\|^2 \leq f(y) - l_f(y;x) \leq \frac{M}{2}\|y-x\|^2 \\ \iff & \frac{m}{2}\|y-x\|^2 \leq [\nabla f(y) - \nabla f(x)]^T(y-x) \leq \frac{M}{2}\|y-x\|^2 \end{aligned}$$

## Algorithms

Steepest Descent

**Definition 0.2.** For a function $f \in \mathcal{C}^1(\mathbb{R}^n)$ which has $L-$Lipschitz continuous gradient, the **steepest descent with fixed step size** method is that for given $x_0 \in \mathbb{R}^n$ and $\theta \in (0,2)$, we update with

$$\begin{aligned} x_k &= x_{k-1} - \frac{\theta}{L}\nabla f(x_{k-1}) \\ k &\leftarrow k+1 \end{aligned}$$

**Proposition 0.13.** *Assume that $f(x_k) \geq \underline{f}$ in the above steepest descent method. Then for all $k > 1$ we have*

$$\min_{1\leq i\leq k} \|\nabla f(x_{i-1})\|^2 \leq \frac{f(x_0) - \underline{f}}{k}\left(\frac{2L}{\theta(2-\theta)}\right)$$

Projected Gradient

**Definition 0.3.** For $\Omega \subseteq \mathbb{R}^n$ convex, $f \in \mathcal{C}^1(\Omega)$ which has $L-$Lipschitz continuous gradient on $\Omega$, the **projected gradient** method is that for given $x_0 \in \mathbb{R}^n$ and $\theta \in (0, 2)$, we update with

$$x_k = \operatorname*{argmin}_{x\in\Omega}\left\{l_f(x; x_{k-1}) + \frac{L}{2\theta}\|x - x_{k-1}\|^2\right\}$$
$$k \leftarrow k+1$$

**Lemma 0.1.** *For all $k \geq 1$, under the projected gradient scheme, we have*

$$0 \in \nabla f(x_{k-1}) + N_\Omega(x_k) + \frac{L}{\theta}(x_k - x_{k-1})$$

**Lemma 0.2.** *Let $r_k = \frac{L}{\theta}(x_{k-1} - x_k)$ and $\bar{r}_k = r_k + \nabla f(x_k) - \nabla f(x_{k-1})$. Then $\bar{r}_k \in \nabla f(x_k) + N_\Omega(x_k)$ and $\|\bar{r}_k\| \leq L\left(\frac{1}{\theta} + 1\right)\|x_k - x_{k-1}\|$.*

**Proposition 0.14.** *Assume that $f(x_k) \geq \underline{f}$ for all $k \geq 0$. Then, for all $k \geq 1$ we have*

$$\min_{1\leq i\leq k}\|\bar{r}_i\|^2 \leq \frac{f(x_0) - \underline{f}}{k}\left(\frac{2L(\theta+1)^2}{\theta(2-\theta)}\right).$$

**Lemma 0.3.** *We have $f(x_{k-1}) - f(x_k) \geq \frac{L}{2}\left(\frac{2-\theta}{\theta}\right)\|x_k - x_{k-1}\|^2$.*

**Lemma 0.4.** *Given closed and convex $\Omega \subseteq \mathbb{R}^n$, a convex function $f \in \mathcal{C}^1(\Omega)$, which has $L-$Lipschitz continuous gradient, and the set of optimal solutions $\Omega^* \neq \emptyset$ for $(P)$, for every $k \geq 1$ and $x^* \in \Omega^*$ we have*

$$\|x_k - x^*\| \leq \|x_0 - x^*\|$$
$$f(x_k) - f_* \leq \frac{L}{2k}\|x_0 - x^*\|^2$$

*and hence if $x^* = P_{\Omega^*}(x_0)$ then $d_0 := \|x_0 - P_\Omega(x^*)\|$ and*

$$f(x_k) - f_* \leq \frac{Ld_0^2}{2k} \implies \min_{1\leq i\leq k}\|r_i\|^2 \sim \mathcal{O}(1/k^2)$$

*If in addition, $f$ is $\beta$ strongly convex, then*

$$f(x_k) - f^* \leq \frac{L}{2}\left(1 - \frac{\beta}{2}\right)^k d_0^2 \implies \|r_k\| \sim \mathcal{O}\left(\left(1 - \frac{\beta}{L}\right)^k\right)$$

Gradient-Type Methods

*Remark* 0.3. Assuming that $f$ is $L$-Lipschitz, and $x_{k+1} = x_k + \alpha_k d_k$, we need (using line minimization) $\alpha_k = -\frac{\nabla f(x_k)^T d_k}{L\|d_k\|^2} > 0$ which will imply $f(x_k) - f(x_{k+1}) \geq \frac{(\nabla f(x_k)^T d_k)^2}{2L\|d_k\|^2} > 0$.

*Remark* 0.4. Let $\epsilon_k = \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\|\|d_k\|}$. Then, $f(x_k) - f(x_{k+1}) \geq \frac{\epsilon_k^2\|\nabla f(x_k)\|^2}{2L}$ which implies

$$\min_{i\leq k-1}\|\nabla f(x_i)\|^2 \leq \frac{2L\left(f(x_0) - \underline{f}\right)}{\sum_{i=0}^{k-1}\epsilon_i^2}.$$

So if $\sum_{i=0}^\infty \epsilon_i^2 = \infty$ (e.g. $\epsilon_i \geq \underline{\epsilon}$ for all $i$), then $\lim_{k\to\infty}\min_{i\leq k}\|\nabla f(x_i)\|^2 = 0$. If $\epsilon_i \geq \epsilon$ for all $i$, then

$$\min_{i\leq k-1}\|\nabla f(x_i)\|^2 \leq \frac{2L\left(f(x_0) - \underline{f}\right)}{\epsilon^2 k}.$$

*Remark* 0.5. If $d_k = -D_k\nabla f(x_k)$ and $D_k$ is symmetric positive definite, then $\text{cond}(D_k) \leq \frac{1}{\epsilon} \implies \epsilon_k \geq \epsilon > 0$ and hence $\lim_{k\to\infty}\|\nabla f(x_k)\| = 0$.

*Remark* 0.6. $\lambda_{\min}(D)\|u\|^2 \leq u^T Du \leq \lambda_{\max}(D)\|u\|^2$ and $\|Du\| \leq \lambda_{\max}(D)\|u\|$. Hence $\epsilon_k \geq \frac{1}{\text{cond}(D_k)} \geq \epsilon$.

## Inexact Line Search

*Remark* 0.7. Assume now that $L$ is not known or does not exist and define $\phi_k(\alpha) = f(x_k + \alpha d_k) - f(x_k)$. We wish to choose $\alpha$ such that $\phi_k(\alpha) \leq \sigma\phi_k'(0)\cdot\alpha$ $(*)$.

- (a) Goldstein rule: For some constant $\tau \in (\sigma, 1)$, we require $\alpha_k$ to satisfy $\phi_k(\alpha) \geq \tau\phi_k'(0)\alpha$.
- (b) Wolfe-Powell (W-P) rule: For some constant $\tau \in (\sigma, 1)$, we require $\alpha_k$ to satisfy $\phi_k'(\alpha) \geq \tau\phi_k'(0)$.
- (c) Strong Wolfe-Powell rule: For some constant $\tau \in (\sigma, 1)$, we require $\alpha_k$ to satisfy $|\phi_k'(\alpha)| \leq -\tau\phi_k'(0)$.
- (d) Armijo's rule: Let $s > 0$ and $\beta \in (0, 1)$ be fixed constants. Choose $\alpha_k$ as the largest scalar from $\alpha \in \{s, s\beta, s\beta^2, ...\}$ such that (*) is satisfied. In other words, find $m$ such that

$$f(x^k + s\beta^m d^k) - f(x^k) = \phi(s\beta^m) \leq \sigma s\beta^m\nabla f(x^k)^T d^k$$

## Rates of Convergence

Consider the problem $(P)$ with $f \in \mathcal{C}^2(\mathbb{R}^n)$ and $H^k = \nabla^2 f(x^k)$.

Gradient Type Methods

These are of the form $x^{k+1} = x^k - \alpha^k D^k\nabla f(x^k)$.

**Proposition 0.15.** *For every $k \geq 0$, we have*

$$\frac{f(x^{k+1}) - f_*}{f(x^k) - f_*} \leq \left(\frac{M_k - m_k}{M_k + m_k}\right)^2 = \left(\frac{r_k - 1}{r_k + 1}\right)^2$$

*where $m_k = \lambda_{\min}((D^k)^{1/2}H^k(D^k)^{1/2})$, $M_k = \lambda_{\max}((D^k)^{1/2}H_k(D^k)^{1/2})$ and $r_k = M_k/m_k = cond(H_k) \geq 1$. If line minimization is used for $\alpha^k$ then*

$$\limsup_{k\to\infty}\frac{f(x^{k+1}) - f_*}{f(x^k) - f_*} \leq \limsup_{k\to\infty}\left(\frac{r_k - 1}{r_k + 1}\right)^2.$$

*Remark* 0.8. For the QP case with $f(x) = x^T Qx$, steepest descent with $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$ and $\alpha^k = \operatorname{argmin}_\alpha f(x^k + \alpha d^k)$ gives the above result with $m_k = \lambda_{\min}(Q), M_k = \lambda_{\max}(Q)$.

Local Convergence of Newton's Method

**Theorem 0.2.** *Assume $h \in \mathcal{C}^2(\mathbb{R}^n)$ and let $x^* \in \mathbb{R}^n$ be such that $h(x^*) = 0, h'(x^*)$ is non-singular. Then there exists $y > 0$ such that if $x_0 \in \bar{B}(x^*; y)$ then $\{x_k\}$ obtained as $x_{k+1} = x_k - [h'(x_k)]^{-1} h(x_k)$ is well-defined and*

$$\lim_{k\to\infty} x_k = x^* \text{ and } \limsup_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} < \infty.$$

## Conjugate Gradient Method

Classic CG Method

**Definition 0.4.** A set of directions $\{d_0, ..., d_k\} \subseteq \mathbb{R}^n$ are **$Q$-conjugate** if $d_i^T Q d_j = 0$ for every $0 \le i < j \le k$. Equivalently, $D_k^T Q D_k$ is diagonal.

**Algorithm 1.** *For $x_0 \in \mathbb{R}^n$, $f(x) = \frac{1}{2} x^T Q x - b^T x$, $Q > 0$ symmetric, let $d_0 = -g_0 = b - Qx_0$. For $k = 0, 1, 2, ...$ do*

$$x_{k+1} = x_k + \alpha_k d_k \text{ where } \alpha_k = -\frac{g_k^T d_k}{d_k^T Q d_k}.$$

*If $g_{k+1} = 0$, stop; else $d_{k+1} = -g_{k+1} + \beta_{k+1} d_k$ where $\beta_{k+1} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$. The algorithm terminates in at most $n$ steps and $f(x_{k+1})$ is minimized over $[d_0, ..., d_k]$.*

*Alternatively, $d_{k+1} = -g_{k+1} + \sum_{i=1}^k \beta_{ki} d_i$ where $\beta_{ki} = \frac{g_{k+1}^T Q d_i}{d_i^T Q d_i}$.*

**Lemma 0.5.** *If $d_0, ..., d_k$ are $Q$-conjugate and $g_{k+1} \notin [d_0, ..., d_k]$ then $d_{k+1}$ as above satisfies*

*(1) $d_{k+1}$ is $Q$-conjugate w.r.t. $\{d_0, ..., d_k\}$*

*(2) $[d_0, ..., d_{k+1}] = [d_0, ..., d_k, g_{k+1}]$*

**Theorem 0.3.** *Assume that $g_i \neq 0$, $i \in \{0, ..., h\}$. Then for all $i \in \{0, 1, ..., k\}$ we have*

*(i) $d_0, ..., d_i$ are $Q$-conjugate*

*(ii) $g_0, ..., g_i$ are orthogonal*

*(iii) $[d_0, ..., d_i] = [g_0, ..., g_i]$*

*(iv) $[d_0, ..., d_i] = [g_0, Qg^0, ..., Q^i g_0]$*

*(v) $\alpha_i = \|g_i\| / (d_i^T Q d_i)$ and $g_i^T d_i = -\|g_i\|^2$*

**Corollary 0.4.** *For every $k \ge 0$ and $P_k \in \mathcal{P}_k$, the set of degree $k$ polynomials with $P_k(0) = 1$, we have*

$$\frac{f(x_k) - f_*}{f(x_0) - f_*} \le \left( \max_{\lambda \in \sigma(Q)} |P_k(\lambda)| \right)^2.$$

**Corollary 0.5.** *For all $k \ge 0$, we have*

$$\frac{f(x_k) - f_*}{f(x_0) - f_*} \le 2 \left( \frac{\sqrt{r} - 1}{\sqrt{r} + 1} \right)^2$$

*where $r = M/m$ is the condition number of $Q$.*

General CG Methods

**Definition 0.5.** Consider $(P)$ where $f \in \mathcal{C}^1(\mathbb{R}^n)$ and $\Omega = \mathbb{R}^n$. The **CG framework**, given $x_0 \in \mathbb{R}^n$, is: For $k = 0, 1, ...$ do

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

where $\alpha_k > 0$ is the step size. Recall for convex quadratic, $\beta_k = \underbrace{\frac{\|g_{k+1}\|^2}{\|g_k\|^2}}_{(1)} = \underbrace{\frac{g_{k+1}^T (g_{k+1} - g_k)}{\|g_k\|^2}}_{(2)}$. Using (1) in the general case leads to the **Fletcher-Reeves** (FR) method while (2) leads to the **Polak-Ribière** (PR) method.

**Theorem 0.4.** *(PR) Assume that $f$ is such that for $0 < m \le M$,*

$$m\|u\|^2 \le u^T \nabla^2 f(x) u \le M \|u\|^2$$

*for all $x, u \in \mathbb{R}^n$. Then the PR-CG method with exact line search method converges to the unique global minimum.*

**Theorem 0.5.** *Assume that $f \in \mathcal{C}^2(\mathbb{R}^n)$ and $\{x : f(x) \le f(x_0)\}$ is bounded. Then there exists an accumulation point $\bar{x}$ of $\{x_k\}$ such that $\nabla f(\bar{x}) = 0$. If $f$ is convex then $\{\bar{x}_k\} \to \bar{x}$. he **Strong Wolfe-Powell inexact line search** is used in this scheme where $0 < \sigma < \tau < \frac{1}{2}$.*

## Nesterov's Method

**Theorem 0.6.** *The **Nesterov Method** has convergence $f(y_k) - f_* \le 4Ld_0^2/k^2$ for $f \in \mathcal{C}^1(\mathbb{R}^n)$ convex and $L$-Lipschitz. If in addition, $f$ is $\mu$-strongly convex, then $f(y_k) - f_* \le d_0^2 / \left[ \lambda \left( 1 + \sqrt{\frac{\mu}{2L}} \right)^{2(k-1)} \right]$.*

**Aside (for the exam)**. If $\underline{\phi} \le \min\{\phi(x)\}$ and $\phi$ is $\beta$-strongly convex, with $\bar{x} = \operatorname{argmin}_x \phi(x)$ then $\underline{\phi} + \frac{\beta}{2} \|x - \bar{x}\|^2 \le \phi(x)$.

**Aside (for the exam)**. If $f$ is $\mu$-strongly convex, then $\lambda f + \frac{1}{2} \|x - x_0\|^2$ is $(\lambda\mu + 1)$ strongly convex.

## Quasi-Newton Methods

Quasi-Newton Method's General Scheme

(0) Let $x^0 \in \mathbb{R}^n$ and $H_0 \in \mathbb{R}^{n \times n}$ symmetric and $H_0 > 0$ be given.

(1) For $k = 0, 1, 2, ...$ set $d_k = -H_k g_k$, $x_{k+1} = x_k + \alpha_k d_k$.

Update $H_k$ to obtain $H_{k+1} > 0$ and symmetric. Here, we want $H_k \sim [\nabla^2 f(x_k)]^{-1}$.

Secant Equation

$p_k = H_{k+1} q_k$

Rank-One Updates (SR1)

$H_{k+1} = H_k + a_k z_k z_k^T$ where $a_k \in \mathbb{R}$ and $z_k \in \mathbb{R}^n$. We want

$$p_k = H_{k+1} q_k = H_k q_k + a_k (z_k^T q_k) z_k$$

and so $z_k$ is proportional to $p_k - H_k q_k$. If we choose $z_k = p_k - H_k q_k$ then $1 = a_k (z_k^T q_k)$.

Rank-Two Updates

$H_{k+1} = H_k + +auu^T + bvv^T$ for $a \in \mathbb{R}$ and $u, v \in \mathbb{R}^n$. The secant equation implies that

$$p_k = H_{k+1}q_k = H_k q_k + a(u^T q_k)u + b(v^T q_k)v.$$

If we choose $u = p_k$ and $v = H_k q_k$ and enforce that $a(p_k^T q_k) = 1, b(q_k^T H_k q_k) = -1$, then we have the **Davidon-Fletcher-Powell** (DFP) method.

Sherman-Morrison Formula

**Proposition 0.16.** *Assume that* $A = B + USV^T$ *where* $S \in \mathbb{R}^{m\times m}, A, B \in \mathbb{R}^{n\times n}$ *non-singular and* $U, V \in \mathbb{R}^{n\times m}$. *If* $P = S^{-1} + V^T S^{-1} U$ *is non-singular then* $A^{-1} = B^{-1} - B^{-1}UP^{-1}V^T B^{-1}$.

Other Rank-Two Updates

We could try the following iteration scheme

$$x_{k+1} = x_k - \alpha_k B_k^{-1} g_k, B_k \approx \nabla^2 f(x_k)$$

We call this the **Broyden-Fletcher-Goldfarb-Shannon** (BFGS) update.

Broyden's Family of Algorithms

Let $\phi = \phi_k \in \mathbb{R}$. Then the method is defined as

$$\begin{aligned} H_{k+1}^{\phi} &= (1-\phi)H_{k+1}^{DFP} + \phi H_{k+1}^{BFGS} \\ &= \phi H_{k+1}^{DFP} + \phi v_k v_k^T \end{aligned}$$

where

$$v_k = (q_k^T H_k q_k)^{1/2} \left( \frac{p_k}{p_k^T q_k} - \frac{H_k q_k}{q_k^T H_k q_k} \right).$$

**Theorem 0.7.** *If* $H_k > 0, p_k^T q_k > 0, \phi \geq 0$ *then* $H_{k+1}^{\phi} > 0$.

**Theorem.** *If* $H_0 = I$ *then the iterates generated by Broyden's Quasi-Newton method, with the exact line search method, are identical to those generated by the conjugate gradient method.*

Convergence Result for General $f$

**Theorem 0.8.** *Let* $f : \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^2(\mathbb{R}^n)$ *and* $x_0 \in \mathbb{R}^n$ *be such that*

*(1)* $S = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ *is bounded and convex*

*(2)* $\nabla^2 f(x) > 0$ *for all* $x \in S$

*Let* $\{x_k\}$ *be a sequence generated by the Broyden Quasi-Newton method* $x_k = x_k - \alpha_k H_k^{\phi_k} g_k$ *where* $\phi_k \in [0,1]$ *and* $H_0 = I$ *and* $\alpha_k$ *is chosen by the W-P rule and* $\alpha_k = 1$ *is the first attempted step size. Then,* $\lim_{k\to\infty} x_k = x^*$ *superlinearly in the sense that*

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

*where* $x^*$ *is the unique global minimum of* $f$ *over* $S$.

## Miscellaneous

- Cauchy Schwartz: $|\langle u, v\rangle| \leq \|u\|\|v\|$.
- If the sublevel sets $\mathcal{N} = \{x : f(x) \leq f(x^0)\}$ are bounded and the step size of a gradient method is chosen to enforce a descent direction, then $\{x^k\}$ must have at least one limit point.
  - If in addition, $\nabla f$ is $L$-Lipschitz, $f$ is bounded below on $\mathcal{N}$, $x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$, and the "angle condition" in class holds – that is $\text{cond}(D_k) \leq \frac{1}{\epsilon} \implies \epsilon_k \geq \epsilon > 0$ – then $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$.
- $\{d^k\}$ is **gradient related** to $\{x^k\}$ if for any subsequence $\{x_k\}_{k\in\mathcal{K}}$ that converges to a non-stationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies
  $$\limsup_{k\to\infty, k\in\mathcal{K}} \nabla f(x^k)^T d^k < 0$$
  - The first order Taylor expansion of $x^{k+1}$ is
    $$f(x^{k+1}) = f(x^k) + \alpha^k \nabla f(x^k)^T d^k + o(\alpha^k)$$
  - If $d^k = -D^k \nabla f(x^k)$ and the eigenvalues are bounded in the sense that $c_1 \leq \lambda^k \leq c_2$ for positive $c_1, c_2$ and any eigenvalue $\lambda^k$ of $D^k$ then $\{d^k\}$ is gradient related.
    * If the eigenvalues of $D^k$ are bounded, then the "angle condition" in class holds. That is, $\text{cond}(D_k) \leq \frac{1}{\epsilon} \implies \epsilon_k \geq \epsilon > 0$ and hence $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$.
  - If $\{d^k\}$ is gradient related, and the minimization rule, or the limited minimization rule, Goldstein rule, or the Armijo rule is used, then all limit points of $\{x^k\}$ are stationary.
    * Limited minimization rule is $f(x^k + \alpha^k d^k) = \min_{\alpha\in[0,s]} f(x^k + \alpha d^k)$
    * For constant step size and $L$-Lipschitz $\nabla f(x)$, gradient related $\{d^k\}$ we require
      $$\epsilon \leq \alpha^k \leq (2-\epsilon)\bar{\alpha}^k, \bar{\alpha}^k = \frac{|\nabla f(x^k)^T d^k|}{L\|d^k\|^2}, \epsilon > 0$$
- The **conjugate gradient method** has the properties:
  - $\nabla f(x^{k+1})^T d^i = 0$ for $i = 0, 1, 2, ..., k$ and $x^{k+1}$ minimizes $f$ over $[d^0, d^1, ..., d^k]$
  - $[d^0, ..., d^k] = [g^0, ..., g^k]$ where $d^{k+1}$ is generated by applying Gram-Schmidt on $[d^0, ..., d^k]$ using $g^{k+1}$; if $d^{k+1} = 0$ then $g^{k+1} = 0$ (from the fact that $\Delta g^{k+1} = \alpha^k Q d^k \implies g^k = g^{k+1}$ and $g^{k+1} = 0$).
    * Note that Gram-Schmidt implies $g^{k+1} \perp [d^0, ..., d^k]$

## Constrained Optimization

**Definition 0.6.** We say that $x \in \mathbb{R}^n$ is a **regular point** of (ECP) if $\nabla h_1(x), ..., \nabla h_m(x)$ are linearly independent.

ECP Conditions

**Theorem 0.9.** *(Lagrange Multiplier Theorem - First order necessary optimality conditions) If $x^*$ is a regular local minimum of (ECP), then $\exists! \lambda^* \in \mathbb{R}^m$ such that $\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0$.*

**Theorem 0.10.** *(Second Order Necessary Conditions) If $x^*$ is a regular local minimum of (ECP), then there exists a unique $\lambda^* \in \mathbb{R}^m$ such that*

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0$$
$$d^T \left(\nabla^2 f(x^*) + \nabla^2 h(x^*)\lambda^*\right) d \geq 0$$

*for all $d \in V(x^*)$ where $V(x^*) = \{d \in \mathbb{R}^n : \nabla h(x^*)^T d = 0\}$.*

**Theorem 0.11.** *(Second Order Necessary Conditions) Assume that $f, h \in \mathcal{C}^2$ and $x^*$ is a regular local minimum of (ECP). Then there exists $\lambda^* \in \mathbb{R}^m$ such that*

$$\nabla_x L(x^*, \lambda^*) = 0$$
$$d^T \nabla^2_{xx} L(x^*, \lambda^*) d \geq 0$$

*for all $d \in V(x^*)$.*

**Theorem 0.12.** *(Second Order Sufficient Conditions) Assume that $f, h \in \mathcal{C}^2$ and $(x^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^m$ is such that*

$$\nabla_x L(x^*, \lambda^*) = 0, h(x^*) = 0,$$
$$d^T \nabla^2_{xx} L(x^*, \lambda^*) d > 0, \forall 0 \neq d \in V(x^*).$$

*Then $x^*$ is a strictly local minimum of ECP. In fact, there exists $\gamma > 0, \epsilon > 0$ such that*

$$f(x) \geq f(x^*) + \frac{\gamma}{2}\|x - x^*\|, \forall x \in \bar{B}(x^*, \epsilon) \text{ s.t. } h(x) = 0.$$

**Lemma 0.6.** *Let $P, Q$ be $n \times n$ symmetric matrices such that $Q \geq 0$ and $d^T P d > 0$ for every $d \neq 0$ such that $d^T Q d = 0$. Then $\exists \bar{c} \in \mathbb{R}$ such that*

$$P + cQ > 0, \forall c \geq \bar{c}.$$

**Theorem 0.13.** *Let $(x^*, \lambda^*)$ be a regular local minimum and Lagrange multiplier for (ECP) satisfying the 2nd order sufficiency condition. Then $\exists \delta > 0$ such that $\forall u \in \bar{B}(0, \delta)$ there exists a pair of regular local minimum and Lagrange multipliers $p(u) = (x(u), \lambda(u))$ for $(ECP)_u$ which is continuously differentiable, $(x(0), \lambda(0)) = (x^*, \lambda^*)$ and*

$$\nabla p(u) = -\lambda(u), p(u) = f(x(u)).$$

*where $(ECP)_u$ is the problem$\min_{h(x)=u}\{f(x)\}$. Note that $\nabla p(0) = -\lambda^*$.*

ICP Conditions

**Definition 0.7.** We say $x \in \mathbb{R}^n$ is **regular** if $\begin{cases} \nabla h_i(x), & i = 1, ..., m \\ \nabla g_j(x), & j \in A(x) \end{cases}$ are linearly independent.

**Theorem 0.14.** *(KKT Necessary Optimality Conditions) Let $x^*$ be a regular local minimum of (NLP). Then $\exists!(\lambda^*, \mu^*) \in \mathbb{R}^m \times \mathbb{R}^r$ such that*

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0,$$
$$h(x^*) = 0, g(x^*) \leq 0$$
$$\mu^* \geq 0, \mu_j = 0, \forall j \notin A(x^*).$$

*If, in addition, $f, g, h \in \mathcal{C}^2$ then*

$$d^T \nabla^2_{xx} L(x^*, \lambda^*) d \geq 0$$

*for every $d \in V(x^*)$ where*

$$V(x^*) = \left\{ d \in \mathbb{R}^n : {\nabla h(x^*)^T d = 0 \atop \nabla g_j(x^*)^T d = 0, j \in A(x^*)} \right\}.$$

**Theorem 0.15.** *(Second Order Sufficient Conditions) Assume $f, g, h \in \mathcal{C}^2$ and $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r$ satisfying*

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0$$
$$h(x^*) = 0, g(x^*) \leq 0$$
$$\mu^* \geq 0$$
$$\mu^*_j = 0, j \notin A(x^*)$$
$$d^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d > 0$$

*for all*

$$d \neq 0$$
$$\nabla h(x^*)^T d = 0$$
$$g_j(x^*)^T d = 0, j \in A(x^*).$$

*Also assume that $\mu_j > 0$ for $j \in A(x^*)$. Then $x^*$ is a strict local minimum.*

**Proposition 0.17.** *(Mangasarian-Fromovitz CQ) If $\nabla h_i(x^*) = 0$ and are linearly independent for $i = 1, 2, ..., m$ and $\exists d \in \mathbb{R}^m$ such that $\nabla h(x^*)^T d = 0, \nabla g_j(x^*)^T d < 0$ for $j \in A(x^*)$ then the first order necessary conditions are satisfied.*

**Proposition 0.18.** *(Slater CQ) If $h$ is affine, $g_j$ is convex, and $\exists \bar{x}$ such that $g_j(\bar{x}) < 0$ for all $j \in A(x^*)$, then the previous proposition holds.*

**Proposition 0.19.** *(Linear/Concave CQ) If $h$ is affine and $g$ is concave, the first order necessary conditions hold without the regularity condition.*

**Proposition 0.20.** *(General sufficiency condition) For the problem $(ICP)$ assume that $(x^*, \lambda^*, \mu^*)$ is such that $x^*$ is feasible and*

$$x^* \in \operatorname*{argmin}_{x \in X} L(x, \lambda^*, \mu^*)$$

*with $\mu^* \geq 0$ and $(\mu^*)^T g(x^*) = 0$ where the second condition is equivalent to $\mu_j = 0$ for $j \notin A(x^*)$. Then $x^*$ is a global minimum.*

Augmented Lagrangian

**Definition 0.8.** For $c > 0$, the **augmented Lagrangian function** is defined as

$$L_c(x, \lambda) = f(x) + \lambda^T h(x) + \frac{c}{2}\|h(x)\|^2.$$

**Proposition 0.21.** *Assume that $X = \mathbb{R}^n$ and $(x^*, \lambda^*)$ is a pair satisfying the 2nd order sufficiency condition, i.e.,*

$$\nabla_x L(x^*, \lambda^*) = 0, h(x^*) = 0$$
$$d^T \nabla^2_{xx} L(x^*, \lambda^*) d > 0 \text{ for every } d \text{ s.t. } \nabla h(x^*)^T d = 0.$$

*Then $x^*$ is a strict local minimum of $L_c(\cdot, \lambda^*)$ for every $c$ sufficiently large.*

General Approach (Penalty)

For $\{c_k\} \subseteq \mathbb{R}_{++}$ and $\{\lambda_k\} \subseteq \mathbb{R}^n$, find $x_k \in \operatorname{argmin}_{x \in X} L_{c_k}(\cdot, \lambda_k)$.

**Proposition 0.22.** *(Quadratic Penalty Method) Assume that $f, h$ are continuous, $X$ is closed and (ECP) is feasible. Suppose $\{\lambda_k\}$ is bounded and $c_k \to \infty$. Then every limit point of $\{x_k\}$ is a global minimum of (ECP). Notationally, we may write $v^k = c_k$.*

**Proposition 0.23.** *Assume that $X = \mathbb{R}^n$ and $f, g \in \mathcal{C}^1(\mathbb{R}^n)$. Assume also that*

$$\|\nabla_x L_{c_k}(x_k, \lambda_k)\| \leq \epsilon_k$$

*where $\{\lambda_k\}$ is bounded, $\epsilon_k \to 0$ and $c_k \to \infty$. Assume also $x_k \overset{k \in K}{\to} x^*$ where $x^*$ is a regular point. Then there exists $\lambda^* \in \mathbb{R}^n$ such that*

$$\lambda_k + c_k h(x_k) \to \lambda^*$$
$$\nabla f(x^*) + \nabla h(x^*) \lambda^* = 0$$
$$h(x^*) = 0.$$

Hessian Ill-Conditioning

We have

$$Q_k = \nabla^2_{xx} L_{c_k}(x_k, \lambda_k) = \nabla^2_{xx} L(x_k, \bar{\lambda}_k) + c_k \nabla h(x_k) \nabla h(x_k)^T$$

where $\bar{\lambda}_k = \lambda_k + c_k h(x_k)$ and as $k \to \infty$,

$$\nabla^2_{xx} L(x_k, \bar{\lambda}_k) \to \nabla^2_{xx} L(x^*, \lambda^*)$$
$$\nabla h(x_k) \nabla h(x_k)^T \to \nabla h(x^*) \nabla h(x^*)^T$$

and in the limit the matrix $Q_k$ will have $m$ eigenvalues tending to $\infty$ and $n - m$ eigenvalues which are bounded. So $\text{cond}(Q_k) \to \infty$.

Augmented Lagrangian Methods

*Remark* 0.9. Define $\{c_k\} \subseteq \mathbb{R}_{++}$ and $\{\lambda_k\} \subseteq \mathbb{R}^m$ and $x_k \in \operatorname{argmin}_{x \in X} L_{c_k}(x, \lambda_k)$. A previous proposition suggests the update $\lambda_{k+1} = \lambda_k + c_k h(x_k)$, which is called the **method of multipliers**.

**Proposition 0.24.** *Assume $x^*$ is a regular local minimum of (ECP) which satisfies the 2nd order sufficiency condition. Let $\bar{c} \geq 0$ be such that $\nabla^2 L_{\bar{c}}(x^*, \lambda^*) > 0$. Then $\exists \delta, \epsilon, M > 0$ such that*

*(a) For all $(\lambda_k, c_k)$ satisfying*

$$\|\lambda_k - \lambda^*\| \leq \delta c_k, c_k \geq \bar{c} \qquad (*)$$

*the problem*

$$\min_x L_{c_k}(x, \lambda_k)$$
$$\text{s.t. } \|x - x^*\| < \epsilon$$

*has a unique global minimum $x_k$. Moreover,*

$$\|x_k - x^*\| \leq \frac{M}{c_k} \|\lambda_k - \lambda^*\|$$

*(b) For all $(\lambda_k, c_k)$ satisfying $(*)$,*

$$\|\lambda_{k+1} - \lambda^*\| \leq \frac{M}{c_k} \|\lambda_k - \lambda^*\|$$

*where $\lambda_{k+1} = \lambda_k + c_k h(x_k)$.*

General Algorithms

A general algorithm is as follows:

(0) Let $\lambda_0 \in \mathbb{R}^m$ and $c_{-1} > 0$ be given and set $\epsilon_0 = \infty$ and $k = 0$.

(1) Set $c = c_{k-1}$.

(2) Compute $x \in \operatorname{argmin} L_c(\cdot, \lambda_k)$.
If $\|h(x)\| > \frac{1}{4}\epsilon_k$, set $c = 10c$ and go to (2).
Else, go to (3).

(3) Set $c_k = c, x_k = x, \lambda_{k+1} = \lambda_k + c_k h(x_k), \epsilon_{k+1} = \|h(x_k)\|$ and $k \hookleftarrow k + 1$. Go to (1).

** Note that we may replace $\frac{1}{4}$ with any constant less than 1, and 10 with any constant greater than 1.

**Proposition 0.25.** *If the global method does not loop in (2), then every accumulation point $x^*$ of $\{x_k\}$ which is regular satisfies $\nabla_x L(x^*, \lambda^*) = 0, h(x^*) = 0$ for some $\lambda^* \in \mathbb{R}^m$. Moreover, $\lambda^*$ is an accumulation point of $\{\lambda_k\}$.*

*Remark* 0.10. Consider the dual function $d_c(\lambda) = \min_{\|x - x^*\| \leq \epsilon} L_c(x, \lambda)$. For 2nd order sufficient solutions, we have the following dual relationship:

$$\sup_{\lambda \in \mathbb{R}^m} d_c(\lambda) = f^* = \min \ f(x) \text{ s.t. } h(x) = 0, \|x - x^*\| \leq \epsilon$$

In the (ICP) formulation,

$$L_c(x, \mu) = f(x) + \mu^T g^+(x, \mu, c) + \frac{c}{2} \|g^+(x, \mu, c)\|$$

where $g^+(x, \mu, c) = \max(g(x), -\frac{\mu}{2})$. We update with $\mu_{k+1} = \max(0, \mu_k + c_k g(x_k))$ in the global method.

## Barrier Methods

Under the $(ICP)$ framework, let $\mathcal{F} = \{x \in X : g(x) \leq 0\}, \mathcal{F}^0 = \{x \in X : g(x) < 0\}$ with the assumption that (1) $\mathcal{F}^0 \neq \emptyset$, (2) $\mathcal{F} \subseteq \text{cl}(\mathcal{F}^0)$.

Barrier Function

This is a function $\psi : \mathbb{R}^p_{++} \mapsto \mathbb{R}$ continuous such that $\psi(y(x)) \to \infty$ as $x \to \text{bd}(\mathbb{R}^p_{++})$.

Barrier Subproblem

For $\mu > 0$, the subproblem is $\min_{x\in\mathcal{F}^0}\{f(x)+\mu B(x)\}$ where $B(x) = \psi(-g(x))$.

Approach

For $\{\mu_k\} \subseteq \mathbb{R}_{++}$ such that $\mu_k \downarrow 0$, compute $x_k \in \operatorname{argmin}_{x\in\mathcal{F}^0} f(x) + \mu_k B(x)$.

**Theorem 0.16.** *Every accumulation point of $\{x_k\}$ is an optimal solution of (ICP).*

**Theorem 0.17.** *Assume that $\{x_k\}$ is a sequence of stationary points of $\min_{x\in\mathcal{F}^0} \phi_{\mu_k}(x)$ for some $\{\mu_k\} \downarrow 0$ and that $x_k \overset{k\in K}{\to} \bar{x}$ where $\bar{x}$ is a regular point of (ICP). Then*

$$\lambda_i^k = -\frac{\mu_k}{g_i(x_k)} \to \bar{\lambda}_i, i = 1, ..., p$$

*for some $\bar{\lambda} \in \mathbb{R}^p$. Moreover, $(\bar{x}, \bar{\lambda})$ satisfies the necessary optimality conditions of $(ICP)$.*

**Lemma 0.7.** *If $u_k$ satisfies $B^k u_k = b_k$ and $B^k \to B$ which is full column rank. Then $u_k \to u$ for some $u$.*

## Interior Point Methods

**See in-depth notes.**

Algorithm

(0) Let $(x_0, \mu_0) \in X^0 \times \mathbb{R}_{++}$ be such that $\delta_{\mu_0}(x_0) \le \delta$ and set $k \leftarrow 0$.

(1) Write $\mu_k > \frac{\epsilon}{n}\left(1+\frac{\delta}{\sqrt{n}}\right)^{-1}$ and do:

$\mu_{k+1} = \mu_k\left(1+\frac{\gamma}{\sqrt{n}}\right)^{-1}$ where $\gamma$ is chosen to satisfy $\delta_{\mu^+}(x) \le \sqrt{\delta}$

$x_{k+1} = x_k + \Delta x_k$ where $\Delta x_k = \Delta x(x_k, \mu_{k+1})$

Set $k \leftarrow k+1$.

(2) Output $x_k$.

**Proposition 0.26.** *The algorithm terminates in $\mathcal{O}\left(\sqrt{n}\log\frac{n\mu_0}{\epsilon}\right)$ iterations with $x \in X^0$ such that $c^T x - v^* \le \epsilon$.*

## Duality

Consider the framework to be $(ICP)$ :

$$\begin{aligned} (ICP)\ \min\ & f(x) \\ \text{s.t. } & g(x) \le 0 \\ & x \in X \end{aligned}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}^r$. For $(x, \mu) \in \mathbb{R}^n \times \mathbb{R}^r$, we define the **Lagrangian function**

$$L(x, \mu) = f(x) + \mu^T g(x).$$

**Definition 0.9.** We say $\mu^*$ is a **geometric multiplier** for (ICP) if

$$\mu^* \ge 0 \text{ and } f_* = \inf_{x\in X} L(x, \mu^*).$$

**Proposition 0.27.** *Let $\mu^*$ be a geometric multiplier. Then, $x^*$ is a global minimum of (ICP) if and only if*

$$\begin{aligned} & x^* \in \operatorname*{argmin}_{x\in X} L(x, \mu^*) \\ & g(x^*) \le 0 \\ & (\mu^*)^T g(x^*) = 0. \end{aligned}$$

*Remark* 0.11. If $f, g_j$ are convex for $j = 1, 2, ..., r$ and $X = \mathbb{R}^n$ then $L(\cdot, \mu^*)$ is convex and the above is reduced to: $x^*$ is a global minimum of (ICP) if and only if $\nabla L(x^*, \mu^*) = 0$ if and only if

$$\nabla f(x^*) + \sum_{j=1}^r \mu_j^* \nabla g_j(x^*) = 0.$$

ICP Duality

Let us define $q : \mathbb{R}^r \mapsto [-\infty, \infty)$ as $q(\mu) = \inf_{x\in X} L(x, \mu)$. The **dual problem** is

$$\begin{aligned} q^* = \sup_{\mu}\ & q(\mu) \\ \text{s.t. } & \mu \ge 0. \end{aligned}$$

**Proposition 0.28.** *(ICP Weak Duality) For every $\mu \ge 0$ and $x \in X$ such that $g(x) \le 0$ we have $f(x) \ge q(\mu)$ and hence $f^* \ge q^*$.*

**Proposition 0.29.** *Let $\mu^* \in \mathbb{R}^r$ be given. Then $\mu^*$ is a geometric multiplier if and only if $f^* = q^*$ and $\mu^*$ is a dual optimal solution.*

NLP Duality

For the (NLP) problem, define

$$\begin{aligned} L(x, \mu, \lambda) &= f(x) + \mu^T g(x) + \lambda^T h(x) \\ q(\mu, \lambda) &= \inf_{x\in X} L(x, \mu, \lambda) \end{aligned}$$

which are respectively the Lagrangian and dual function for (NLP).

**Proposition 0.30.** *(NLP Weak Duality) If $x$ if feasible for (NLP) and $(\mu, \lambda) \in \mathbb{R}_+^r \times \mathbb{R}^m$ then $f(x) \ge q(\mu, \lambda)$ and hence $f_* \ge q_*, f_* \ge q(\mu, \lambda)$, $f(x) \ge q_*$ where $q_* = \sup_{\mu\ge 0} q(\mu, \lambda)$.*

**Definition 0.10.** The pair $(\mu^*, \lambda^*) \in \mathbb{R}^r \times \mathbb{R}^m$ is a **geometric multiplier** (G.M.) if $\mu^* \ge 0$ and $f_* = q(\mu^*) = q_*$.

**Proposition 0.31.** *Let $(\mu^*, \lambda^*) \in \mathbb{R}^r \times \mathbb{R}^m$ be given such that $\mu^* \ge 0$. Then, $(\mu^*, \lambda^*)$ is a G.M. if and only if $(\mu^*, \lambda^*)$ is a dual optimal solution and $f_* = q_*$.*

**Proposition 0.32.** *A pair $(x^*, (\mu^*, \lambda^*))$ is an optimal solution-G.M. pair if and only if*

$$\begin{aligned} & x \text{ is feasible} \\ & x^* \in \operatorname*{argmin}_{x\in X} L(x, \mu^*, \lambda^*) \\ & \mu^* \ge 0 \\ & g(x^*) \le 0 \\ & (\mu^*)^T g(x^*) = 0. \end{aligned}$$

**Fact 0.1.** *For $x \in X$ and $\mu \geq 0$ we have*

$$q(\mu, \lambda) \leq L(x, \mu, \lambda) \leq f(x).$$

**Fact 0.2.** *For $x \in X$ and $\mu \geq 0$ we have*

$$\sup_{\substack{\mu \geq 0 \\ \lambda \in \mathbb{R}^m}} L(x, \mu, \lambda) = \begin{cases} f(x), & \text{if } g(x) \leq 0, h(x) = 0 \\ \infty, & \text{otherwise} \end{cases}.$$

**Proposition 0.33.** *(Saddle Point) A pair $(x^*, (\mu^*, \lambda^*))$ is an optimal solution-G.M. pair if and only if*

$$x^* \in X, \mu \geq 0$$
$$L(x,^* \mu, \lambda) \leq L(x^*, \mu^*, \lambda^*) \leq L(x, \mu^*, \lambda^*), \forall (\mu, \lambda) \in \mathbb{R}_+^r \times \mathbb{R}^m, \quad \forall x \in X$$

Existence of G.M.'s

Here, let us consider the (NLP) problem

$$\begin{aligned} f_* = \inf\ & f(x) \\ \text{s.t. } & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in X. \end{aligned}$$

**Proposition 0.34.** *Assume that:*

* $f_* \in \mathbb{R}$
* *h, g are affine*
* $f : \mathbb{R}^n \mapsto \mathbb{R}$ *is convex*
* *X is polyhedral*

*Then (NLP) has a G.M. and as a consequence $f_* = q_*$.*

**Proposition 0.35.** *Assume that:*

* $f_* \in \mathbb{R}$
* *h, g are affine*
* $f : \mathbb{R}^n \mapsto \mathbb{R}$ *is convex quadratic*
* *X is polyhedral*

*Then (NLP) has an optimal solution-G.M. pair.*

General Case

Consider the general problem

$$\begin{aligned} f_* = \inf\ & f(x) \\ \text{s.t. } & Ax \leq b \\ & g(x) \leq 0 \\ & x \in X \end{aligned}$$

**Proposition 0.36.** *Assume that:*

* $f_* \in \mathbb{R}$
* $X = C \cap P$ *where P is polyhedral, C is convex*
* $f : \mathbb{R}^n \mapsto \mathbb{R}, g_j : C \mapsto \mathbb{R}$ *are convex*
* $\exists \bar{x}$ *such that* $g(\bar{x}) < 0$, $A\bar{x} \leq b$, *and* $\bar{x} \in ri(C) \cap P$

*Then (NLP) has a G.M. pair and as a consequence $f_* = q_*$.*

## Augmented Lagrangian Methods vs. Duality

Consider the problem

$$\begin{aligned} f_* = \inf\ & f(x), & f : \mathbb{R}^n \mapsto \mathbb{R} \\ \text{s.t. } & Ax = b, & A \text{ is } m \times n \\ & x \in X, & X \subseteq \mathbb{R}^n \end{aligned}$$

the **value function** is $v(u) = \inf_{Ax-b=u} \{f(x)\}$ where clearly, $v(0) = f_*$.

**Proposition 0.37.** *If X is convex and f is convex on X then $v(\cdot)$ is convex.*

**Definition 0.11.** Define

$$v_\rho(u) = \begin{aligned} \inf\ & f(x) + \frac{\rho}{2}\|Ax - b\|^2 \\ \text{s.t. } & Ax - b = u \\ & x \in X \end{aligned}$$

**Proposition 0.38.** *If X is convex and f is convex on X then $v_\rho(\cdot)$ is $\rho$-strongly convex.*

**Proposition 0.39.** *Assume that X is convex compact and f is convex on X. Then:*

*(1) $d_\rho(\cdot)$ is concave and differentiable everywhere*

*(2) $\nabla d_\rho(\cdot)$ is $\frac{1}{\rho}$-Lipschitz continuous*

*(3) $\nabla d_\rho(\lambda) = -u_\rho(\lambda)$ where $u_\rho(\lambda) = \operatorname{argmin}_{u \in \mathbb{R}^m} v_\rho(u) + \lambda^T u$.*

*where*

$$d_\rho(\lambda) = L_\rho(x, \lambda) = \inf_{u \in \mathbb{R}^m} v_\rho(u) - \lambda^T u = \inf_{u \in \mathbb{R}^m} v(u) - \lambda^T u + \frac{\rho}{2}\|u\|^2.$$

*Remark* 0.12. Recall the augmented Lagrangian method:

(0) $\lambda_0 \in \mathbb{R}^m$ is given; set $k \leftarrow 1$.

(1) Set $x_k = \operatorname{argmin}_{x \in X} L_\rho(x, \lambda_{k-1})$

(2) Set $\lambda_k = \lambda_{k-1} + \rho(b - Ax_k)$

(3) Set $k \leftarrow k + 1$ and go to (1).

Note that in step (2) we have

$$\lambda_k = \lambda_{k-1} + \rho \nabla d(\lambda_{k-1}) = \lambda_{k-1} + \frac{1}{L_\rho} \nabla d(\lambda_{k-1})$$

so this is steepest ascent on $d(\lambda_{k-1})$. Note that this step can be then replaced with

$$\lambda_k = \lambda_{k-1} + \frac{\theta}{L_\rho} \nabla d(\lambda_{k-1}) = \lambda_{k-1} + \theta\rho(b - Ax_k), \theta \in (0, 2)$$