Definition 41 [Subgradient]: *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function defined on a convex set $\mathcal{D}$. A vector $\mathbf{h} \in \Re^n$ is said to be a subgradient of f at the point $\mathbf{x} \in \mathcal{D}$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{y} \in \mathcal{D}$. The set of all such vectors is called the subdifferential of f at $\mathbf{x}$.

Theorem 76 *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function defined on a convex set $\mathcal{D}$. A point $\mathbf{x} \in \mathcal{D}$ corresponds to a minimum if and only if*

$$\nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \geq 0$$

for all $\mathbf{y} \in \mathcal{D}$.

If $\nabla f(\mathbf{x})$ is nonzero, it defines a supporting hyperplane to $\mathcal{D}$ at the point $\mathbf{x}$. Theorem 77 implies that for a differentiable convex function defined on an open set, every critical point must be a point of (global) minimum.

Theorem 77 *Let $f : \mathcal{D} \rightarrow \Re$ be differentiable and convex on an open convex domain $\mathcal{D} \subseteq \Re^n$. Then $\mathbf{x}$ is a critical point of f if and only if it is a (global) minimum.*

Theorem 78 *Let $f : \mathcal{D} \rightarrow \Re$ with $\mathcal{D} \subseteq \Re^n$ be differentiable on the convex set $\mathcal{D}$. Then,*

1. *f is convex on $\mathcal{D}$ if and only if is its gradient ∇f is monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T(\mathbf{x} - \mathbf{y}) \geq 0 \tag{4.53}$$

2. *f is strictly convex on $\mathcal{D}$ if and only if is its gradient ∇f is strictly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$ with $\mathbf{x} \neq \mathbf{y}$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T(\mathbf{x} - \mathbf{y}) > 0 \tag{4.54}$$

3. *f is uniformly or strongly convex on $\mathcal{D}$ if and only if is its gradient ∇f is uniformly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T(\mathbf{x} - \mathbf{y}) \geq c||\mathbf{x} - \mathbf{y}||^2 \tag{4.55}$$

for some constant $c > 0$.

Definition 41 [Subgradient]: *Let $f : \mathcal{D} \to \Re$ be a convex function defined on a convex set $\mathcal{D}$. A vector $\mathbf{h} \in \Re^n$ is said to be a subgradient of f at the point $\mathbf{x} \in \mathcal{D}$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{y} \in \mathcal{D}$. The set of all such vectors is called the subdifferential of f at $\mathbf{x}$.

Theorem 76 *Let $f : \mathcal{D} \to \Re$ be a convex function defined on a convex set $\mathcal{D}$. A point $\mathbf{x} \in \mathcal{D}$ corresponds to a minimum if and only if*

$$\nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \geq 0 \qquad (*)$$

for all $\mathbf{y} \in \mathcal{D}$.

Proof: Suppose $x \in D$ satisfies $(*)$. If $y \in X$ then ($\because f$ is convex)

$f(y) \geq f(x) + \nabla^T f(x)(y-x) \geq f(x) \Rightarrow x$ is a global min point of f.

Suppose $x \in D$ is a global min point but $(*)$ does not hold.

i.e $\exists\, y \in D$ s.t $\nabla^T f\,(y-x) < 0$

Consider $z(t) = ty + (1-t)x$ $(t \in [0,1])$. Now the set $(z(t), t \in [0,1])$ is a feasible set. We claim that for a small $t > 0$,

$$\frac{d}{dt} f(z(t)) = \nabla^T f(x)(y-x) < 0 \Rightarrow f(z(t)) < f_0(x)$$

$\therefore$ A contradiction

Necessity: Suppose f is uniformly convex on $\mathcal{D}$. Then from theorem 75, we know that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) - \frac{1}{2}c||\mathbf{y} + \mathbf{x}||^2$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x} - \mathbf{y}) - \frac{1}{2}c||\mathbf{x} + \mathbf{y}||^2$$

Adding the two inequalities, we get (4.55). If f is convex, the inequalities hold with $c = 0$, yielding (4.54). If f is strictly convex, the inequalities will be strict, yielding (4.54).

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,

$$\phi(1) - \phi(0) = \phi'(t) \tag{4.56}$$

Letting $\mathbf{z} = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$, (4.56) translates to

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla^T f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \tag{4.57}$$

Also, by definition of monotonicity of ∇f, (from (4.53)),

$$(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) = \frac{1}{t}(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x}))^T (\mathbf{z} - \mathbf{x}) \geq 0 \tag{4.58}$$

Combining (4.57) with (4.58), we get,

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \\ &\geq \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \end{aligned} \tag{4.59}$$

By theorem 75, this inequality proves that f is convex. Strict convexity can be similarly proved by using the strict inequality in (4.58) inherited from strict monotonicity, and letting the strict inequality follow through to (4.59). For the case of strong convexity, from (4.55), we have

$$\begin{aligned} &\phi'(t) - \phi'(0) = (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) \\ &= \frac{1}{t}(\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{z} - \mathbf{x}) \geq \frac{1}{t}c||\mathbf{z} - \mathbf{x}||^2 = ct||\mathbf{y} - \mathbf{x}||^2 \end{aligned} \tag{4.60}$$

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 [\phi'(t) - \phi'(0)]dt \geq \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2 \tag{4.61}$$

which translates to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2$$

Back to optimization with constraints

Convex f & g_i

$$\min f(x) \quad \text{s.t. } g_i(x) \leq 0$$

option 1

$$F(x) = f(x) + \sum_i I_{g_i}(x)$$

option 2

$$F(x) = f(x) + \max_i \min_{u: g_i(u) \leq 0} \|x - u\|_2$$

$$\min_x F(x)$$

Either obtain solution by setting $g_F(x) = 0$ & solving for x OR applying a descent algorithm

Convex f, g_i

$$\min f(x) \quad \text{s.t. } g_i(x) \le 0$$

option 1

$$F(x) = f(x) + \underbrace{\sum_i I_{g_i}(x)}_{\text{subgradient = normal cone (on bndry)}}$$

option 2

$$F(x) = f(x) + \max_i \underbrace{\min_{u: g_i(u) \le 0} \|x-u\|_2}_{(x - P_{g_i}(x)) \,/\, \|x - P_{g_i}(x)\|}$$

$\min_x F(x)$

Either obtain solution by setting $g_F(x) = 0$ & solving for x OR applying a descent algorithm.

option 3

$$F_t(x) = f(x) + \left(-\frac{1}{t}\right) \sum_i \log(-g_i(x))$$

$$x^*(t) = \underset{x}{\operatorname{argmin}} \; F_t(x)$$

Barrier method

It turns out that analysing Barrier method (or analysing convergence of prox/projected gradient descent) becomes meaningful if we understand conditions for optimality for constrained opt. ..

option 4

$$F_k(x) = f_{Q_k}(x) + \sum_i I_{g_i}(x)$$

$$x^{k+1} = \underset{x}{\operatorname{argmin}} \; F_k(x)$$

Projected gradient method

Recall: $f_{Q_k}(x)$ = Quadratic approx to f around x^k

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

$$\therefore x^{k+1} = \underset{x}{\operatorname{argmin}} \; \frac{1}{2t}\left\|x - (x^k - t\nabla f(x^k))\right\|^2 + \sum_i I_{g_i}(x)$$

$$= \underset{x: g_i(x) \le 0}{\operatorname{argmin}} \|x - \hat{x}^{k+1}\|^2$$

$$= P_{C_1 \cap C_2 \ldots \cap C_m}(\hat{x}^{k+1})$$

More generally, the 4th option: called projected gradient descent

Proximal gradient descent.

$$\min f(x) + r(x)$$

A = features, x = feature wts in Lasso

f is differentiable $\frac{1}{2}\|Ax-y\|^2$

r is not differentiable eg: $\lambda\|x\|_1$

Iteratively solve: $x^{(0)}$

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}}\ f(x^{(k)}) + \nabla^T f(x^{(k)})(x - x^k) + \frac{1}{2t}\|x - x^k\|^2 + r(x)$$

$$\boxed{= \underset{x}{\operatorname{argmin}}\ \frac{1}{2t}\|x - (x^k - t\nabla f(x^k))\|^2 + r(x)}$$

$$= \underset{x}{\operatorname{argmin}}\ \frac{1}{2t}\|x - \hat{x}^{(k+1)}\|^2 + r(x)$$

$$\hat{x}^{(k+1)} = x^k - t\nabla f(x^k)$$

For our problem: $x^{(k+1)} = \min_x \|x-y\|^2 \ldots \|x\|_1$

H/W: complete & reduce to known problem

Recall: $\min_x \|y-x\|^2 + \lambda\|x\|_1$ had a closed form optimal soln:

$$x^*_i = \begin{cases} y_i + \lambda & \text{if } y_i < -\lambda \\ -y_i + \lambda & \text{if } y_i > \lambda \\ 0 & \text{o/w} \end{cases}$$

Obtained by setting a subgradient to 0.

Projected gradient descent is prox gradient descent when $r(x) = I_\Omega(x)$

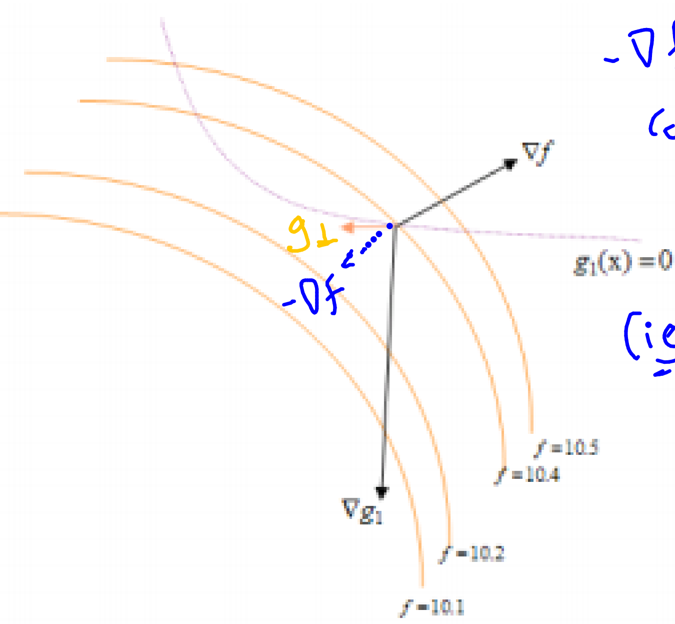
For the above problem: $f(x) = \frac{1}{2}\|y - Ax\|^2 \quad r(x) = \lambda\|x\|_1$

$$x^{k+1} = \underset{x}{\operatorname{argmin}}\ \|x - \hat{x}^{k+1}\|^2 + \lambda\|x\|_1 \qquad (\hat{x}^{(k+1)} = x^k - tA^TA(y - x^k))$$

$$= \begin{cases} \hat{x}_i^{(k+1)} + \lambda & \text{if } \hat{x}_i^{(k+1)} < -\lambda \\ -\hat{x}_i^{(k+1)} + \lambda & \text{if } \hat{x}_i^{(k+1)} > \lambda \\ 0 & \text{otherwise} \end{cases}$$

NECESSARY CONDITIONS FOR CONSTRAINED OPTIMALITY (pages 284-287 of ...
http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf)

min $f(x)$
s.t $g_1(x)=0$



$-\nabla f$ has a small component $\perp$ to $\nabla g_1(x)$ ($g_\perp$) (ie $g_1(x)=0$ remains)

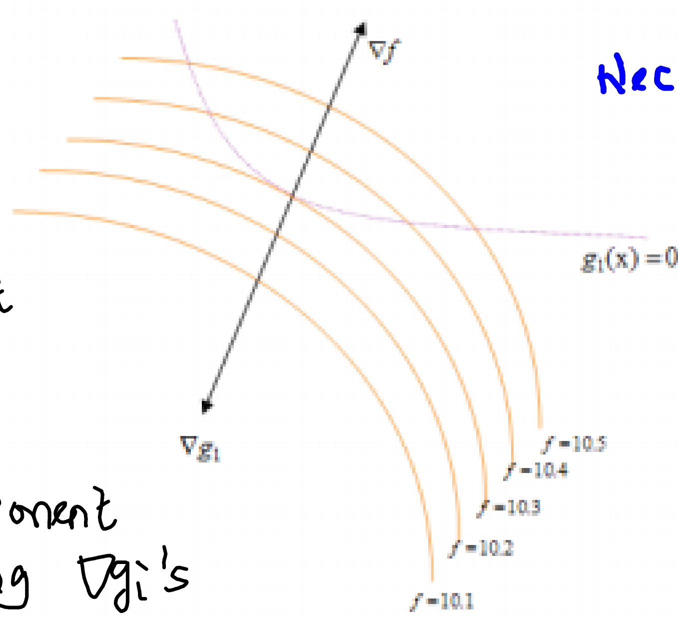
min $f(x)$
s.t $g_i(x)=0$ $i=1..m$

Figure 4.39: At any non-optimal and non-saddle point of the equality constrained problem, the gradient of the constraint will not be parallel to that of the function.

Necessary condition for multiple constraints

$\nabla f + \sum_i \lambda_i \nabla g_i = 0$ at x pt of min/max

ie. ∇f has no component $\perp$ to subspace containing ∇g_i's



Necessary condition

$\nabla f + \lambda \nabla g = 0$

at x pt of minimum/max

Figure 4.40: At the equality constrained optimum, the gradient of the constraint must be parallel to that of the function.

In the case of

$\min f(x)$

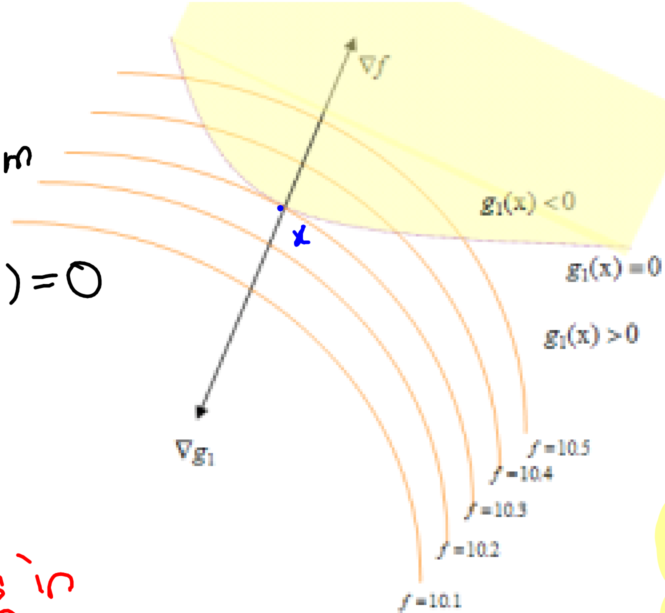
s.t $g_i(x) \leq 0 \quad i = 1..m$

$\nabla f(x) + \sum_i \lambda_i \nabla g_i(x) = 0$

$\lambda_i g_i(x) = 0$

$\lambda_i \geq 0$

(See Recession cones in Bertsekas)



We do not mind reducing f & g_1 simultaneously

$\therefore -\nabla f(x)$ should not have component $\perp$ to ∇g_1 & also should not have component along $-\nabla g_1(x)$ (since $g_1(x) < 0$ is o.k)

New additional condition

Figure 4.41: At the inequality constrained optimum, the gradient of the constraint must be parallel to that of the function.

$\nabla f(x) + \lambda \nabla g_1(x) = 0 \quad \lambda_1 \geq 0$

& $\lambda g_1(x) = 0$

This discussion of necessary conditions for "local" optimality holds also for non-convex, differentiable f & g_i's

We have invoked:

$$\left.\begin{aligned} &\min f(x) \\ &\text{s.t } g_i(x) \leq 0 \quad i = 1..m \\ &h_j(x) = 0 \quad j = 1..p \end{aligned}\right\} \; (*)$$

$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$

Necessary condition for optimality of $(*)$ at x^* is

$(**)$ $\left\{ \begin{aligned} &\nabla_x L(x^*, \lambda, \nu) = 0 \text{ for some } \lambda_i \geq 0 \text{ \& } \nu_j\text{'s (no constraints on } \nu_j\text{'s)} \\ &\text{\& } \lambda_i g_i(x^*) = 0 \\ &g_i(x^*) \leq 0 \quad h_j(x^*) = 0 \end{aligned} \right.$

$\left\{ \begin{aligned} &\underline{\underline{Q1}} \quad L^*(\lambda, \nu) = \min_x L(x, \lambda, \nu) \\ &\lambda^*, \nu^* = \operatorname*{argmax}_{\lambda \geq 0, \nu} L^*(\lambda, \nu) \end{aligned} \right.$

Could we say (when) that λ^*, ν^* that maximize the dual fn $L^d(\lambda, \nu)$ are precisely the λ's & ν's that satisfy the necessary conditions ✱Δ ?

Q2 If any λ^*, ν^*, x^* satisfy ✱Δ, should the duality gap be 0 ?

Q3 Do answers to Q1 & Q2 require convexity of f & g_i's & affiness of h_j's

Boyd uses slightly different notations in the following slides (WE) (BOYD)

(WE)

$$\min_x \; f(x)$$
$$\text{s.t.} \quad g_i(x) \le 0 \quad i = 1..m$$
$$h_j(x) = 0 \quad j = 1..p$$

(BOYD)

$$\min_x \; f_0(x)$$
$$\text{s.t.} \quad f_i(x) \le 0 \quad i = 1..m$$
$$h_j(x) = 0 \quad j = 1..p$$

Lagrange function

$$L(x,\lambda,\nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

Dual function

$$L^*(\lambda,\nu) = \min_x L(x,\lambda,\nu)$$

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu)$$

∴ Dual optimization prob is always a concave max/convex minimization problem

H/W: Prove

Lagrange dual function

$\max_{\lambda \succeq 0, \nu} g(\lambda,\nu) = \min_{\lambda \succeq 0, \nu} -g(\lambda,\nu)$

convex function in λ & ν

Lagrange dual function: $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$
\begin{aligned}
g(\lambda,\nu) &= \inf_{x\in\mathcal{D}} L(x,\lambda,\nu) \\
&= \inf_{x\in\mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right)
\end{aligned}
$$

$= \Omega\left(f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_i h_i(x)\right)$

g is concave, can be $-\infty$ for some λ, ν

lower bound property: if $\lambda \succeq 0$, then $g(\lambda,\nu) \le p^\star$

($\lambda \succeq 0$ by default means $\lambda \in \mathbb{R}^n_+$)

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \ge L(\tilde{x},\lambda,\nu) \ge \inf_{x\in\mathcal{D}} L(x,\lambda,\nu) = g(\lambda,\nu)$$

minimizing over all feasible $\tilde{x}$ gives $p^\star \ge g(\lambda,\nu)$

Least-norm solution of linear equations

$$
\begin{array}{ll}
\text{minimize} & x^T x \\
\text{subject to} & Ax = b
\end{array}
$$

dual function

- Lagrangian is $L(x,\nu) = x^T x + \nu^T(Ax - b)$
- to minimize L over x, set gradient equal to zero:

Using necessary condition.

$$\nabla_x L(x,\nu) = 2x + A^T\nu = 0 \quad \Longrightarrow \quad x = -(1/2)A^T\nu$$

- plug in in L to obtain g:

$$g(\nu) = L((-1/2)A^T\nu, \nu) = -\frac{1}{4}\nu^T A A^T \nu - b^T\nu$$

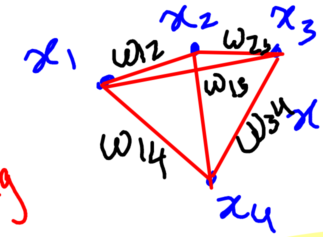
a concave function of ν

Quadratic in ν

lower bound property: $p^\star \ge -(1/4)\nu^T A A^T \nu - b^T \nu$ for all ν

$W = \begin{bmatrix} w_{11} & w_{12} & \\ w_{21} & \ddots & \\ & & w_{44} \end{bmatrix}$

Two-way partitioning



$x_i \in \{-1, +1\}$

Ising Models

$$
\begin{array}{ll}
\text{minimize} & \underbrace{x^T W x}_{\text{energy level of config}} \\
\text{subject to} & \underbrace{x_i^2 = 1}_{\text{Non convex domain}}, \quad i = 1, \ldots, n
\end{array}
$$

- a nonconvex problem; feasible set contains 2^n discrete points
- interpretation: partition $\{1, \ldots, n\}$ in two sets; W_{ij} is cost of assigning i, j to the same set; $-W_{ij}$ is cost of assigning to different sets

dual function

$$
\begin{aligned}
g(\nu) = \inf_x (x^T W x + \sum_i \nu_i (x_i^2 - 1)) \quad &= \quad \inf_x x^T (W + \mathbf{diag}(\nu)) x - \mathbf{1}^T \nu \\
&= \quad \begin{cases} -\mathbf{1}^T \nu & W + \mathbf{diag}(\nu) \succeq 0 \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}
$$

p.s.d constraint

lower bound property: $p^\star \geq -\mathbf{1}^T \nu$ if $W + \mathbf{diag}(\nu) \succeq 0$

example: $\nu = -\lambda_{\min}(W) \mathbf{1}$ gives bound $p^\star \geq n \lambda_{\min}(W)$

$W + \begin{bmatrix} -\lambda_{min} & 0 \\ 0 & -\lambda_{min} \end{bmatrix} \succeq 0$ substitute

Lagrange dual and conjugate function

$$
\begin{array}{ll}
\text{minimize} & f_0(x) \\
\text{subject to} & Ax \preceq b, \quad Cx = d
\end{array}
$$

dual function

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_{x \in \mathbf{dom}\, f_0} \left(f_0(x) + (A^T \lambda + C^T \nu)^T x - b^T \lambda - d^T \nu \right) \\
&= -f_0^*(-A^T \lambda - C^T \nu) - b^T \lambda - d^T \nu
\end{aligned}
$$

- recall definition of conjugate $f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$
- simplifies derivation of dual if conjugate of f_0 is kown

example: entropy maximization

$$
f_0(x) = \sum_{i=1}^n x_i \log x_i, \qquad f_0^*(y) = \sum_{i=1}^n e^{y_i - 1}
$$

Reminds you of Entropy classifier Logistic regression

The dual problem

Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on $p^\star$, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted $d^\star$
- λ, ν are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \mathbf{dom}\, g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \mathbf{dom}\, g$ explicit

example: standard form LP and its dual (page 5–5) (also seen for conic Prog CP)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0 \end{array} \quad \equiv \quad \begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \succeq 0 \end{array}$$

Recap: If both primal & dual are feasible & one of them is strictly feasible $\Rightarrow$ zero duality (strong duality)

Weak and strong duality

weak duality: $d^\star \le p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

  for example, solving the SDP

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array}$$

  gives a lower bound for the two-way partitioning problem on page 5–7

strong duality: $d^\star = p^\star$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

Slater's constraint qualification

strong duality holds for a convex problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

if it is strictly feasible, *i.e.*,

$$\exists x \in \mathbf{int}\, \mathcal{D}: \qquad f_i(x) < 0, \quad i = 1, \ldots, m, \qquad Ax = b$$

- also guarantees that the dual optimum is attained (if $p^\star > -\infty$)
- can be sharpened: *e.g.*, can replace $\mathbf{int}\, \mathcal{D}$ with $\mathbf{relint}\, \mathcal{D}$ (interior relative to affine hull); linear inequalities do not need to hold with strict inequality, . . .
- there exist many other types of constraint qualifications

Inequality form LP

primal problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

dual function

$$g(\lambda) = \inf_x \left((c + A^T\lambda)^T x - b^T\lambda \right) = \begin{cases} -b^T\lambda & A^T\lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

dual problem

$$\begin{array}{ll} \text{maximize} & -b^T\lambda \\ \text{subject to} & A^T\lambda + c = 0, \quad \lambda \succeq 0 \end{array}$$

- from Slater's condition: $p^\star = d^\star$ if $A\tilde{x} \prec b$ for some $\tilde{x}$
- in fact, $p^\star = d^\star$ except when primal and dual are infeasible

Complementary slackness

assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal

$$\begin{aligned} f_0(x^\star) = g(\lambda^\star, \nu^\star) &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\ &\le f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star) \\ &\le f_0(x^\star) \end{aligned}$$

hence, the two inequalities hold with equality

- $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (known as complementary slackness):

$$\lambda_i^\star > 0 \Longrightarrow f_i(x^\star) = 0, \qquad f_i(x^\star) < 0 \Longrightarrow \lambda_i^\star = 0$$

Karush-Kuhn-Tucker (KKT) conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

1. primal constraints: $f_i(x) \le 0$, $i = 1, \ldots, m$, $h_i(x) = 0$, $i = 1, \ldots, p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1, \ldots, m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + \sum_{i=1}^p \nu_i \nabla h_i(x) = 0$$

from page 5–17: if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

KKT conditions for convex problem

if $\tilde{x}$, $\tilde{\lambda}$, $\tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from complementary slackness: $f_0(\tilde{x}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- from 4th condition (and convexity): $g(\tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$

hence, $f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu})$

if **Slater's condition** is satisfied:

x is optimal if and only if there exist λ, ν that satisfy KKT conditions

- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(x) = 0$ for unconstrained problem

example: water-filling (assume $\alpha_i > 0$)

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log(x_i + \alpha_i) \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \end{array}$$

x is optimal iff $x \succeq 0$, $\mathbf{1}^T x = 1$, and there exist $\lambda \in \mathbf{R}^n$, $\nu \in \mathbf{R}$ such that

$$\lambda \succeq 0, \qquad \lambda_i x_i = 0, \qquad \frac{1}{x_i + \alpha_i} + \lambda_i = \nu$$

- if $\nu < 1/\alpha_i$: $\lambda_i = 0$ and $x_i = 1/\nu - \alpha_i$
- if $\nu \geq 1/\alpha_i$: $\lambda_i = \nu - 1/\alpha_i$ and $x_i = 0$
- determine ν from $\mathbf{1}^T x = \sum_{i=1}^n \max\{0, 1/\nu - \alpha_i\} = 1$

interpretation

- n patches; level of patch i is at height α_i
- flood area with unit amount of water
- resulting level is $1/\nu^\star$