ADMM. Boyd 2011.

Dual Ascent : primal problem. $\min f(x)$ convex.
s.t. $Ax=b$.

$\downarrow$

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

$\downarrow$

dual function : $g(y) = -f^*(-A^Ty) - b^Ty = \inf_x L(x,y)$.

$\downarrow$

algorithm : $x^{k+1} = \operatorname{argmin} L(x, y^k)$ ← recover "optimal" x from "optimal" dual variable

$y^{k+1} = y^k + \alpha^k (Ax^{k+1} - b)$ ← approximate "optimal" g(y) with estimated gradient as $\nabla g(y) = Ax^+ - b$. [assuming ~~$\nabla_y \inf_x L(x,y) = \inf_x \nabla_y L(x,y)$~~

$\exists\, x^+$ s.t. $\inf L(x,y) = L(x^+, y)$

then $\nabla_y \inf L(x,y) = \nabla_y L(x^+, y)$

Note: ① need L to be bdd below for most y.

② $\not\exists$ $g(y^k) \uparrow$ with $k \uparrow$.

③ if f nondifferentiable, it's called dual subgradient method. [$Ax^+ - b$ is a subgradient]

Dual Decomposition : decompose x into disjoint variable groups, then use dual ascent.
⟹ groups can be updated parallelely.

Augmented Lagrangian & the Method of Multiplier.

augmented problem : $\min f(x) + \frac{\rho}{2}\|Ax-b\|_2^2$
s.t. $Ax=b$. ⟹ $L_\rho(x,y) = f(x) + y^T(Ax-b) + \frac{\rho}{2}\|Ax-b\|_2^2$

① robust (why?).
② no longer need convexity of f(x) for alg to converge.

method of multiplier : $x^{k+1} = \operatorname{argmin}_x L_\rho(x, y^k)$

$y^{k+1} = y^k + \rho(Ax^{k+1} - b)$ $\alpha_k = \rho$ makes (x^{k+1}, y^{k+1}) dual feasible ~~minimize~~ ???
the definition is weird here

Alternating Direction Method of Multipliers :
(try to use dual decomposition in Augmented. Lagrangian with method of multiplier)
good computing property — good convergence property.

$\min f(x) + g(z)$ ⟹ (both convex)
s.t. $Ax + Bz = c$ ⟹ $L_\rho(x,z,y) = f(x) + g(z) + y^T(Ax+Bz-c) + \frac{\rho}{2}\|Ax+Bz-c\|_2^2$

algorithm :
$$x^{k+1} = \operatorname{argmin}_x L_\rho(x, z^k, y^k)$$
$$z^{k+1} = \operatorname{argmin}_z L_\rho(x^{k+1}, z, y^k)$$
$$y^{k+1} = y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

Scaled Form of ADMM.

$$\begin{cases} \text{residual } \overset{r(x,z)}{r} = Ax+Bz-c \\ \text{scaled dual variable } u = \frac{1}{\rho}y \end{cases} \Rightarrow \quad L_\rho(x,z,y) = f(x)+g(z) + \rho u^T r + \frac{\rho}{2}\|r\|^2$$

$$= f(x)+g(z)+\frac{\rho}{2}\|r+u\|^2 - \frac{\rho}{2}\|u\|^2$$

Original ~~a~~ problem: $\min f(x)+g(z)+\frac{\rho}{2}\|r\|^2$ s.t. $r=0$. and scale dual variable with ρ.

algorithm:

$$x^{k+1} = \underset{x}{\mathrm{argmin}}\left[f(x)+\frac{\rho}{2}\|Ax+Bz^k-c+u^k\|_2^2\right]$$

$$z^{k+1} = \underset{z}{\mathrm{argmin}}\left[g(z)+\frac{\rho}{2}\|Ax^{k+1}+Bz-c+u^k\|_2^2\right]$$

$$u^{k+1} = u^k + \underbrace{Ax^{k+1}+Bz^{k+1}-c}_{r^k}$$

r^k [is the approximation of $\nabla g(u)$]

Convergence: [1] Theoretical: assumption ① f, g closed, proper & convex

② $L_0(x,z,y) = f(x)+g(z)+y^T(Ax+Bz-c)$ have a saddle point.

$\Downarrow$

strong duality

theoretical convergence

[2] make sure the algorithm converge to optimal point:

KKT condition:

$$Ax^*+Bz^*-c=0$$
$$0=\partial f(x^*)+A^Ty^*$$
$$0=\partial g(z^*)+B^Ty^*.$$

$\downarrow$

two residuals small: dual residual: $s^{k+1} = \rho A^TB(z^{k+1}-z^k)$

primal residual: $r^{k+1} = Ax^{k+1}+Bz^{k+1}-c$ $\Rightarrow 0$.

[3] stopping criteria:

developed under $f(x^k)+g(z^k)-p^* \leq -y^{k^T}\underline{r}^k+(x^k-x^*)^T\underline{s}^k$.

$$\begin{cases} \|r^k\|_2 \leq \varepsilon^{pri} \\ \|s^k\|_2 \leq \varepsilon^{dual} \end{cases}$$

refer to page 19 for detailed suggestion.

[4]. let ρ increases with k allows for faster convergence.

notice: ~~let~~ stop the increase after some iterations s.t. theoretical convergence holds.

Notes: ① x- & z- updates are indeed proximal gradient method when $A, B = I$.

② for $\min_x f(x) + \|Ax - v\|_2^2$ update, quadratic term improves the conditioning of the function, strong convexity !!! wow.
thus improves the behavior of gradient descent method.
(refer to ch9 of cvbook_boyd).

③ other 2 ways to speed up: { Early stopping [theoretically justified !!! ???]
Warm start

What are the exact problems ADMM can solve:

generally speak: if obj func $f(x) = f_1(x) + f_2(x)$ & it's hard to optimize simultaneously
then add an "equivalent term" z. $\min f(x) = f_1(x) + f_2(z)$.
s.t. $x - z = 0$.
⟹ decompose using ADMM ⟹ separately update f_1 & f_2.

application: ① Closest points in 2 sets. [or common point].

② pd cone constrain: set $g(z)$ as the indicator function of condition $z \succeq 0$.
useful.
⟹ z is updated with projection on $\{z \succeq 0\}$.
projection on S_+: eigen decomp.

③ l_1 penalized function: core idea: put l_1 norm in z-update step.
z: $\|\cdot\|_1 + \|\ \|_2^2$ can be solved with soft-thresholding
even if $\|\ \|_2^2$ is complicated, we can use proximal gradient [which is basically using another ADMM in this step].

④ group lasso (even with overlapping group). solve in a collect-distribute way.
[a special case for consensus & sharing].