Alternating Descent method of multipliers (ADMM)

- recall dual ascent
  (rather than go down the primal, we can go up the dual)
- recall we can recover soln to primal from soln to dual, if strong duality holds
- recall method of multipliers

Consider

$$\min \quad f(x)$$
$$\text{st} \quad Ax = b$$

f convex

Augmented Lagrangian:

$$L_\rho(x, \lambda) = f(x) + \lambda^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|^2$$

ALM:

$$x^{k+1} = \underset{x}{\operatorname{argmin}}\ L_\rho(x, \lambda^k)$$
$$\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$$

You should see this as a variant of dual ascent, working with a modified objective

Consider

$$\min \; g(x,\rho) = f(x) + \left(\frac{\rho}{2}\right) \|Ax - b\|^2$$
$$\text{s.t.} \quad Ax = b$$

Then dual ascent would give:

$$x^{k+1} = \operatorname{argmin}_x \; g(x,\rho)$$

$$\lambda^{k+1} = \lambda^k + \eta_k (Ax^{k+1} - b)$$

If we choose $\eta_k = \rho$, then we get ALM.

Notice this is a good choice of step.

Feasibility conds for problem

$$Ax^* - b = 0 \quad , \quad \nabla f\big|_{x^*} + A^T \lambda^* = 0$$

primal   dual.

Now update gets

$$0 = \nabla_x L_\rho (x^{k+1}, y^k)$$

because we minimized

$$= \nabla_x f\Big|_{x^{k+1}} + A^T\left(y^k + \rho(Ax^{k+1} - b)\right)$$

~~assuming $\eta = \rho$~~

but if $\eta_k = \rho$, $y^{k+1} = y^k + \rho(Ax^{k+1} - b)$

So we have

$$0 = \nabla_x f + A^T y^{k+1}$$

– so step yields dual-feasible pt.

<u>Q:</u> in my account of ALM, I updated ρ to make it bigger – why not here?

A: all that is required is sufficiently large ρ.

<u>ADMM</u>:

imagine I have

$$\min \quad f(x) + g(z)$$
$$\text{s.t.} \quad Ax + Bz = c$$

f, g convex.

- we form the augmented Lagrangian.

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

Now we do:

$$x^{k+1} = \underset{x}{\operatorname{argmin}}\ L_\rho(x, z^k, \lambda^k)$$

$$z^{k+1} = \underset{z}{\operatorname{argmin}}\ L_\rho(x^{k+1}, z, \lambda^k)$$

$$\lambda^{k+1} = \lambda^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right)$$

(Notice we did not do ALM, because we min in x, then z)

We can rescale the problem, sometimes more convenient.

Write $r = Ax + Bz - c$.

$$\lambda^T r + \frac{\rho}{2} r^T r = \frac{\rho}{2}\left\| r + \frac{1}{\rho}\lambda \right\|^2 - \frac{1}{2\rho}\|\lambda\|^2$$

$$= \frac{\rho}{2}\|r + u\|^2 - \frac{\rho}{2}\|u\|^2$$

where

$$u = \frac{1}{\rho}\lambda$$

And u is often thought of as the scaled dual variable.

Stopping:

Notice that, at true soln, we have

Primal feasibility $\rightarrow$

$$Ax^* + Bz^* - c = 0 \qquad \text{(A)}$$

$$0 \in \partial f\big|_{x^*} + \cancel{A}\, A^T \lambda^* \qquad \text{(B)}$$

$$0 \in \partial g\big|_{z^*} + B^T \cancel{y}\, \lambda^* \qquad \text{(C)}$$

Dual feasibility

Notice that z^{k+1} minimizes

$L_\rho(x^{k+1}, z, \lambda^k)$, so

$$0 \in \partial g_{z^{k+1}} + B^T \lambda^k + \rho B^T (Ax^{k+1} + Bz^{k+1} - c)$$

from the quad term in augmented Lagrangian

$$= \partial g_{z^{k+1}} + B^T \lambda^{k+1}$$

recall λ update.

<u>so</u> C is <u>always</u> true.

→ we need to check A and B

A is size of residual ∴

check

$$\|r^k\| \leq \varepsilon_{stop}$$

as to B :

x^{k+1} is $\mathrm{argmin}\ L_\rho(x, z^k, \lambda^k)$

which gives us

(Boyd p.18)

$$\rho A^T B(z^{k+1} - z^k) \in \partial f_{x^{k+1}} + A^T \lambda^{k+1}$$

so this motivates looking at

$$\|s^{k+1}\| = \|\rho A^T B(z^{k+1} - z^k)\| \leq \varepsilon_{\text{small}}$$

Recall for ALM, ρ updated by

$$\rho^{k+1} = \gamma \rho^k, \quad \gamma \text{ usually } 2 \text{ or } 10$$

A better is available:

- increase ρ if r is big compared to s
- Decrease ρ if s is big compared to r
- fix otherwise

$$\rho^{k+1} = \begin{cases} \tau \rho^k & \|r^k\| > \mu \|s^k\| \\ \rho^k / \tau & \|s^k\| > \mu \|r^k\| \\ \rho^k & \text{otherwise} \end{cases}$$

$\mu > 1$, $\tau > 1$ (2 is good)

Experience: ADMM gets to fair sol'n's fairly fast, but slow for tight optimization – you'd expect this from the subgradient step

Example: Lasso

$$\min \quad \frac{1}{2}\|Ax-b\| + \lambda|x|_1$$

same as

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

$$\text{s.t.} \quad x - z = 0$$

$$f(x) = \left(\frac{1}{2}\right)(Ax-b)^T(Ax-b)$$

$$g(z) = \lambda|x|_1$$

Augmented Lagrangian:

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

rescaled L.M.

so

x-update:

$$(A^TA + \rho I)\, x^{k+1} = (A^Tb + \rho(u^k - z^k))$$

z-update:

$$z^{k+1} = \underset{z}{\operatorname{argmin}} \left[\lambda|z| + \frac{\rho}{2}\|x^{k+1} - z\|^2 + \rho u^T(x-z) \right]$$

notice this is separable across the components of z

Now consider a component of

z_k^{k+1}

must have

$$0 \in \partial \left[\lambda |z_i| + \rho u_i (x_i^{k+1} - z_i) + \frac{\rho}{2} \| x_i^{k+1} - z_i \|^2 \right]$$

$$= \begin{cases} \lambda \\ 0 \\ -\lambda \end{cases} - \rho u_i - \rho (x_i^{k+1} - z_i)$$

eg. first case

$$z_i = x_i + u_i - \frac{\lambda}{\rho} \qquad \text{if } x_i + u_i > \frac{\lambda}{\rho}$$

~~t~~hird case

$$z_i = x_i + u_i + \frac{\lambda}{\rho} \qquad \text{if } x_i + u_i < \frac{\lambda}{\rho}$$

Second $\qquad z_i = 0$

Soft thresholding operator

$$S_k(a) = \begin{cases} a - k & a > k \\ 0 & |a| \leq k \\ a + k & a < -k \end{cases}$$

gives

$$z_i^{k+1} = S_{\lambda/\rho}\left(x_i^{k+1} + u_i^{k}\right)$$

<u>u - update</u> :

$$u^{k+1} = u^{k} + x^{k+1} - z^{k+1}$$

Now imagine we have a lot of data.

want

$$\underset{x}{\operatorname{argmin}} \ \|Ax - b\|^2 + \lambda|x|.$$

could write as

$$\operatorname{argmin} \sum_i f_i(x) + \lambda|x|,$$

where $f_i(x) = \|A_i x - b\|^2$

A_i ← i'th subset of data.

This is clunky w/ dual descent, But ADMM is good.

$$\operatorname{argmin} \sum_i f_i(x_i) + \lambda|z|,$$

st

$$x_i - z = 0 \quad \text{for each } i.$$

Now we could ~~do~~ introduce an augmented Lagrangian, get

$$\min. \quad \sum_i f_i(x_i) + \sum_i u_i^T (x_i - z) + \frac{\rho}{2} \sum_i \|(x_i - z)\|^2 + \lambda |z|_1$$

$$\text{st} \quad x_i - z = 0$$

This ISNT separable, but we can do x_i updates, ~~z~~ z update, u update.

x_i update:

$$x_i^{k+1} = \underset{x}{\operatorname{argmin}} \left[f_i(x) + u_i^{k\,T} (x_i - z^k) + \frac{\rho}{2} \|x_i - z\|^2 \right]$$

z update:

$$z^{k+1} = \underset{z}{\operatorname{argmin}} \left[\lambda |z|_1 + \sum_i u_i^{k\,T} (x_i^{k+1} - z) + \frac{\rho}{2} \sum_i \|x_i - z\|^2 \right]$$

(Shrinkage will do this!)

<u>u_i update:</u>

$$u_i^{k+1} = u_i^k + (x_i - z)$$

Notice that this pattern applies to SVMs, variety of others, including group lasso. (See Boyd notes).

<u>Example</u>: <u>Sparse inverse covar seln.</u>

- I have $a_i \sim N(0, \Sigma)$

  where Σ is unknown, a_i IID.

  BUT Σ^{-1} is known to be sparse

- You can think of this as a graphical model – one node per component of a, and an edge between nodes if they interact

- recall $p(a_i) = \frac{1}{2\pi |\Sigma|} e^{-\frac{a^T \Sigma^{-1} a}{2}}$

so if you think about the energy of this graphical model, it is

$$\frac{a^T \Sigma^{-1} a}{2}$$

So the non-zero entries in Σ^{-1} are pairwise interactions.

- Q: which entries of Σ^{-1} are non-zero, based on evidence a_i?

A: we estimate $M = \Sigma^{-1}$,

$$p(a \mid M) = \prod_i p(a_i \mid M)$$

we could min − loglikelihood.

$$\min_M \; -\sum_i \log p(a_i \mid M)$$

$$= \sum_i \left[a_i^T \frac{M}{2} a_i - \log\det(M) + \log(2\pi) \right]$$

$N\,Tr(SM)$ where S is empirical covariance.

ignore

so

get

$$\min_M \; Tr(SM) - \log\det(M) + \lambda \|M\|_1$$

sparsity inducing Norm.

and $M \succ 0$

ADMM is good at this (BOYD notes)

$$M^{k+1} = \underset{M}{\operatorname{argmin}} \left[\operatorname{Tr}(SM) - \log\det(M) + (\rho/2) \| M - Z^k + U^k \|^2 \right]$$

$$Z^{k+1} = \underset{Z}{\operatorname{argmin}} \left[\lambda |Z|_1 + \frac{\rho}{2} \| M^{k+1} - Z + U^k \|^2 \right]$$

$$U^{k+1} = U^k + M^{k+1} - Z^{k+1}$$

This may not look great, BUT

Z update cane be done in closed form w/ shrinkage.

M update can be done in closed form, too, with neat trick.

$$S - M^{-1} + \rho(M - Z^k + U^k) = 0$$

So $\rho M - M^{-1} = \rho(Z^k - U^k) - S$

unknown

known, symmetric

So we must deal w/

$$\rho M - M^{-1} = \Gamma$$

known

$$= Q \Lambda Q^T$$

write $\bar{M} = Q^T M Q$

diag

then $\rho \bar{M} - \bar{M}^{-1} = \Lambda$

now its easy!

Example: Consensus w/ regularization

$$\min \sum_i f_i(x_i) + g(z)$$

$$\text{st.} \quad x_i - z = 0$$

(we've seen this idea!)

unscaled form

$$x_i^{k+1} = \arg\min_x \left[f_i(x) + \lambda_i^{kT}(x_i - z^k) + \frac{\rho}{2}\|x_i - z^k\|^2 \right]$$

$$z^{k+1} = \arg\min_z \left[g(z) + \sum_i \left[-\lambda_i^{kT} z + \left(\frac{\rho}{2}\right)\|x_i^{k+1} - z\|^2 \right] \right]$$

$$\lambda_i^{k+1} = \lambda_i^k + \rho\left(x_i^{k+1} - z^{k+1}\right)$$

Rearrange z step

$$z^{k+1} = \underset{z}{\operatorname{argmin}} \Big[g(z) + \underbrace{\left(N\rho/2\right) \| z - \bar{x}^{k+1} - \tfrac{1}{\rho} \bar{y}^{k+1} \|^2}_{\text{this is a form of averaging.}} \Big]$$

if $g(z) = \alpha \|z\|^2$, we get a weighted average

generally, average w/ proximat step
(later!)

(BOYD notes give scaled form).