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

\[
\min_x f(x)
\]

\[
st \quad Ax = b
\]

If convex

**Augmented Lagrangian:**

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

**ALM:**

\[
x^{k+1} = \text{argmin}_x L_p(x, x^k)
\]

\[
\lambda^{k+1} = \lambda^k + \rho (Ax^{k+1} - b)
\]

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

\[ \min g(x, p) = f(x) + \left(\frac{p}{2}\right) \|Ax - b\|^2 \]

st. \quad Ax = b

Then dual ascent would give:

\[ x^{k+1} = \arg\min_x \quad g(x, p) \]

\[ x^{k+1} = x^k + \eta_k (Ax^k - b) \]

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

Notice this is a good choice of step.

Feasibility conditions for problem

\[ Ax^* - b = 0 \]

\[ \nabla f \bigg|_{x^*} + A^* x^* = 0 \]

primal \quad dual.
Now update gets
\[ 0 = \nabla_x L_\lambda (x^{k+1}, y^k) \]

because are minimized
\[ = \nabla_f |_{x^{k+1}} + A^T (y^* + \rho (Ax^{k+1} - b)) \]

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

so we have
\[ 0 = \nabla_f + A y^{k+1} \]

- so step yields dual-feasible pt.

Q: in my account of ALM, I updated
\[ \rho \] to make it bigger - why not here?
A: all that is required is sufficiently large $c$.

**ADM444:**

Imagine I have

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

st. $Ax + Bz = c$

$f, g$ convex.

- we form the augmented Lagrangian

$$L_p = f(x) + g(z) + \lambda^T (Ax + Bz - c)$$

$$+ \left(\frac{p}{2}\right) \left\|Ax + Bz - c\right\|_2^2$$
Now we do:

\[ x^{k+1} = \arg\min_x L_p(x, z^k, \lambda^k) \]
\[ z^{k+1} = \arg\min_z L_p(x^{k+1}, z, \lambda^k) \]
\[ x^{k+1} = \lambda^k + r (A x^{k+1} + B z^{k+1} - c) \]

(Notice we did not do ALM, because we minimize \( x \) and then \( z \)).

We can rescale the problem, sometimes more convenient.

Write \( r = A x + B z - c \).

\[ \lambda^T r + \frac{r^T r}{2} = \frac{p}{2} \| r + \frac{1}{\lambda} \lambda \|_2^2 - \frac{1}{2} \| \lambda \|_2^2 \]
\[ = \frac{p}{2} \| r + u \|_2^2 - \frac{p}{2} \| u \|_2^2 \]

Where \( u = \frac{\lambda}{2} \)
And \( \nu \) is often thought of as the scaled dual variable.

**Stopping:**

Notice that, at true soln, we have

\[
\begin{align*}
\text{Primal feasibility} & \quad \Rightarrow \quad A\mathbf{x}^* + B\mathbf{z}^* - \mathbf{c} = 0 \quad (A) \\
\Rightarrow \quad 0 & \in \partial f(\mathbf{x}^*) + A^T\mathbf{\lambda}^* \quad (B) \\
\Rightarrow \quad 0 & \in \partial g(\mathbf{z}^*) + B^T\mathbf{y}^* \quad (C) \\
\text{Dual feasibility} & \quad \Rightarrow \quad A\mathbf{x}^* + B\mathbf{z}^* - \mathbf{c} = 0
\end{align*}
\]
Notice that $z^{k+1}$ minimizes
\[ L_p(x^{k+1}, z^k, \lambda^k) , \quad \text{so} \]
\[ \begin{align*}
0 & \in \partial g \Bigg|_{z^{k+1}} + B^T \lambda^k + \rho B^T (Ax^{k+1} + Bz^{k+1} - c) \\
\text{[from the quadratic term in augmented Lagrangian]}
\end{align*} \]
\[ = \partial g \Bigg|_{z^{k+1}} + B^T \lambda^{k+1} \]
\[ \lambda^{k+1} \quad \text{[recall $\lambda$ update]} \]
so $C$ is always true.

→ we need to check $A$ and $B$

$A$ is size of residual:
check \[ \|r^k\| < \varepsilon \text{ stop} \]
as to \( \hat{B} \)

\[ x^{k+1} \text{ is argmin} \ L_p(x, z^k, \lambda^k) \]

which gives us

(Boyd p.18)

\[ \rho A^T B (z^{k+1} - z^k) \in \nabla f_{x^{k+1}} + A^T \lambda^{k+1} \]

so this motivates looking at

\[ \| s^{k+1} \| = \| \rho A^T B (z^{k+1} - z^k) \| \leq \varepsilon \]
Recall for ALM, $\lambda$ updated by

$$\lambda^{k+1} = \lambda \lambda^k$$  \quad \lambda \text{ usually } 2 \text{ or } 10$$

A better is available:

- increase $\lambda$ if $\mu$ is big compared to $s$

- decrease $\lambda$ if $s$ is big compared to $\mu$

- fix otherwise

$$\lambda^{k+1} = \begin{cases} 
\lambda^k & \|\lambda^k\| > \mu \|s^k\| \\
\lambda^k / \lambda^2 & \|s^k\| > \mu \|\lambda^k\| \\
\lambda^k & \text{otherwise}
\end{cases}$$

$$\mu > 1 \quad \tau > 1$$  \quad (2 \text{ is good})
Experience: ADMM gets to fair solutions fairly fast, but slow for tight optimization — you’d expect this from the subgradient step.

Example: Lasso

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

Same as

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

st. \ \ x - z = 0

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

$$g(z) = \lambda |x|_1$$
Augmented Lagrangian:

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

rescaled L.M.

So

**x-update:**

\[
(A^T A + \rho I) x^{k+1} = (A^T b + \rho (u^k - z^k))
\]

**z-update:**

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

notice this is separable across the components of z
Now consider a component of $z_{k+1}^i$.

must have

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

$$= \begin{cases} \lambda & \quad -\rho u_i - \rho (x_i - z_i) \\ 0 & \quad \text{if } x_i + u_i \geq \frac{\lambda}{\rho} \\ -\lambda & \quad \text{if } x_i + u_i < \frac{\lambda}{\rho} \end{cases}$$

**first case**

$$z_i = x_i + u_i - \frac{\lambda}{\rho}$$

**third case**

$$z_i = x_i + u_i + \frac{\lambda}{\rho}$$

**second case**

$$z_i = 0$$
soft thresholding operator

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

gives

\[ Z_i^{k+1} = S_{\lambda \rho} (x_i^{k+1} + u_i^k) \]

\underline{u - update}:

\[ u^{k+1} = u^k + x^{k+1} - z^{k+1} \]
Now imagine we have a lot of data. We want

$$\arg\min_{x} \|A x - b\|^2 + \lambda |x|,$$

could write as

$$\arg\min_{x} \sum_{i} f_{i}(x) + \lambda |x|,$$

where

$$f_{i}(x) = \|A_{i} x - b_{i}\|^2$$

is the $i$'th subset of data.

This is clunky w/ dual descent, but

**ADMM** is good.

$$\arg\min_{x} \sum_{i} f_{i}(x_{i}) + \lambda |x|,$$

$s.t.$

$$\nabla x_{i} - z = 0 \quad \text{for each } i.$$
Now we could introduce an augmented Lagrangian, get

\[ \min \sum_{i} f_i(x) + \sum_{i} u_i (x_i - z) + \frac{\rho}{2} \sum_{i} \| x_i - z \|^2 + \lambda \sum_{i} |z_i| \]

subject to \( x_i - z = 0 \).

This isn't separable, but we can do \( x_i \) updates, \( z \) update, \( u \) update.

**\( x_i \) update:**

\[ x_i^{k+1} = \text{argmin}_x \left[ f_i(x) + \sum_{i} u_i (x_i - z^k) + \frac{\rho}{2} \| x_i - z \|^2 \right] \]

**\( z \) update:**

\[ z^{k+1} = \text{argmin}_z \left[ \lambda \sum_{i} |z_i| + \sum_{i} u_i (x_i^{k+1} - z) + \frac{\rho}{2} \sum_{i} \| x_i - z \|^2 \right] \]

(Shrinkage will do this!)
ui update:

\[ u_i^{k+1} = u_i^k + (x_i - z) \]

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

Example: Sparse inverse covar selk.

1. \( a_i \sim N(0, \Sigma) \)

   where \( \Sigma \) is unknown, \( a_i \text{ IID} \).

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

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

$$a_i^T \Sigma^{-1} a_i$$

so the non-zero entries in $\Sigma^{-1}$ are pairwise interactions.

Q: Which entries of $\Sigma^{-1}$ are non-zero, based on evidence at $a_i$?
\[ A: \text{ we estimate } M = \Sigma^{-1} \]
\[ p(a_i | M) = \prod_i p(a_i | M) \]

we could minimize log likelihood.

\[ \min_M \quad -\sum_i \log p(a_i | M) \]
\[ = \sum_i \left[ a_i^T M a_i - \log \det(M) + \log(2\pi) \right] \]
\[ N \text{ Tr}(SM) \quad \text{ where } S \text{ is empirical covariance.} \]

\[ \min_M \quad \text{Tr}(SM) - \log \det(M) + \lambda \| \mathbf{M} \|_1 \]

\( \text{sparse, inducing 1-Norm.} \)
and $M \succ 0$

ADMM is good at this (Boyd notes)

$$M^{k+1} = \arg \min_M \left[ \text{Tr}(SM) - \log \det(M) + \left(\frac{\rho}{2}\right) \|M - Z^k + U^k\|_2^2 \right]$$

$$Z^{k+1} = \arg \min_Z \left[ \lambda \|Z\|_1 + \frac{\rho}{2} \|K^{k+1} - Z + U^k\|_2^2 \right]$$

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

This may not look great, BUT

$Z$ update can be done in closed form w/ shrinkage.

$M$ update can be done in closed form, too, with neat trick.
\[ S = M^{-1} + \rho(M - \overline{Z}^k + \overline{U}^k) = 0 \]

\[ \rho M - M^{-1} = \rho(\overline{Z}^k - \overline{U}^k) - S \]

\[ \text{unknown, Symmetric} \]

So we must deal with

\[ \rho M - M^{-1} = \Gamma \]

\[ = Q \Lambda Q^T \]

Write \( \overline{M} = Q^T MQ \)

Then \( \rho \overline{M} - \overline{M}^{-1} = \Lambda \)

Now it's easy!
Example: Consensus of regularization

\[
\begin{align*}
\min & \sum_i f_i(x_i) + g(Z) \\
\text{st.} & \quad x_i - Z = 0 \\
\end{align*}
\]

(we've seen this idea!)

unscaled form

\[
\begin{align*}
x_i^{k+1} &= \arg\min_x \left[ f_i(x) + g \lambda_i^k (x_i - Z) + \frac{\rho}{2} \|x_i - Z\|^2 \right] \\
Z^{k+1} &= \arg\min_Z \left[ g(Z) + \sum_i \left[ -\lambda_i^k (Z - \left(\frac{\cdot}{\rho}\right)^{1/2} x_i^{k+1} \right] \|x_i - Z\|^2 \right] \\
\lambda_i^{k+1} &= \lambda_i^k + \rho (x_i^{k+1} - Z^{k+1})
\end{align*}
\]

Rearrange \( Z \) step
\[ z^{k+1} = \arg \min_{z} \left[ g(z) + \left( \sqrt{\frac{\gamma}{2}} \right) \| z - x^{k+1} - \frac{1}{\epsilon} \tilde{y}^{k+1} \| \right] \]

This is a form of averaging.

If \( g(z) = \| z \|^2 \), we get a weighted average.

Generally, average w/ proximal step (later!)

(BOYD notes give scaled form)