- Consider the objective

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

- Indicator function for $g_i(x)$

$C_i = \{x \mid g_i(x) \leq 0\}$

$$I_{g_i}(x) = \begin{cases} 0, & \text{if } g_i(x) \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

  - We have shown that this is convex

- We will use subgradient descent to solve this optimization

Option 1: Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Let $C_i = \{x \mid g_i(x) \leq 0\}$
- We take

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

- Consider the subgradient of F:

Subgradient descent possible

$$g_F(x) = g_f(x) + \sum_i g_{I_{C_i}}(x)$$

- Recall that $g_{I_{C_i}}(x)$ is $d \in \mathbf{R}^n$ s.t. $d^\top x \geq d^\top y$, $\forall y \in C_i$
- $g_{I_{C_i}}(x) = 0$ if x is in the interior of C_i, and has other solutions if x is on the boundary

Option 1: More General

- Consider the following sum of a differentiable function $f(x)$ and a nondifferentiable function $c(x)$
- We take

$$\min_x F(x) = \min_x f(x) + c(x)$$

- Like gradient descent, consider the first order approximation for $f(x)$ around x^k leaving $c(x)$ alone:

$c(x) \equiv 0$ gives you gradient descent

$$\min_x f(x^k) + \nabla^T f(x^k)(x - x^k) + \frac{1}{2t}||x - x^k||^2 + c(x)$$

- Adding $\frac{t}{2}\|\nabla f(x^k)\|^2$ to the objective (without any loss) to complete squares & dropping $f(x^k)$ from the objective

$$x^{k+1} = \operatorname{argmin}_x \frac{1}{2t}||x - (\underbrace{x^k - t\nabla f(x^k)}_{\text{grad descent step w/o } c})||^2 + c(x)$$

- In general, such a step is called a *proximal* step

$$x^{k+1} = prox_t\left(||x^k - t\nabla f(x^k))||^2 + c(x)\right)$$

Option 1: Generalized Gradient Descent

- Interesting because in many settings, $prox_t(x)$ can be computed efficiently

$$prox_t(z) = \operatorname*{argmin}_x \frac{1}{2t}||x - z||^2 + c(x)$$

- Illustration on Lasso[1] $\min_x \underbrace{\frac{1}{2}||Ax - y||_2^2}_{f} + \lambda \underbrace{||x||_1}_{c}$
- x^{k+1}
- $= \operatorname*{argmin}_x \frac{1}{2t}||x - (x^k - t\nabla f(x^k))||^2 + c(x)$
- $= \operatorname*{argmin}_x \frac{1}{2t}||x - (x^k - tA^T(Ax^k - y))||^2 + \lambda||x||_1$

$= \operatorname*{argmin}_x \frac{1}{2t}||x - z^k||^2 + \lambda||x||_1 = \operatorname*{argmin}_x ||x - z^k||^2 + 2t\lambda||x||_1$

[1] How did we come up with the iterative algo for Lasso on page 8 of http://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture23a.pdf?

Illustration on Lasso[2]

$$\arg\min_x \frac{1}{2}\|x - z^k\|_2^2 + \lambda t \|x\|_1 = x^{k+1}$$

$$\partial\left(\|x\|_1\right) = \partial\left(\max_{s_i \in [\{-1,-1\}^n]} s^T x\right)$$

$$= \left\{ \begin{bmatrix} -1 & \text{if } x_i < 0 \\ +1 & \text{if } x_i > 0 \\ \theta \in [-1,1] & \text{if } x_i = 0 \end{bmatrix} \right\}$$

$$\partial\left(\frac{1}{2}\|x - z^k\|_2^2\right) = \left(x - z^k\right)$$

$$\Rightarrow x^{k+1} = \begin{cases} -\lambda t + z_i^k & \text{if } z_i^k > \lambda t \\ 0 & \text{if } -\lambda t \le z_i^k \le \lambda t \\ \lambda t + z_i^k & \text{if } z_i^k < -\lambda t \end{cases} \Bigg\} \quad *$$

[2] Justification of the iterative algo for Lasso on page 8 of http://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture23a.pdf

Illustration on Lasso[3]

Overall algo:

start with an $x^{(0)}$, $k=0$

$$\boxed{\begin{array}{l}\text{Compute } z^{(k)} = x^{(k)} - t^{(k)}\left(A^T(Ax^k - y)\right) \\ \quad \& \ x^{(k+1)} \text{ using } * \end{array}}$$

until duality gap $\frac{1}{2}\|Ax^{k+1} - y\|_2^2 + \lambda\|x^k\|_1 - \frac{1}{2}\|u^k - y\|_2^2 \leq \epsilon$

For KKT conditions,

see http://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture25a.pdf

[3] Justification of the iterative algo for Lasso on page 8 of
http://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture23a.pdf

Option 1: Generalized Gradient Descent

- Recall

$$prox_t(z) = \operatorname*{argmin}_x \frac{1}{2t}||x - z||^2 + c(x)$$

- Gradient Descent: $c(x) = 0$
- Projected Gradient Descent: $c(x) = \sum_i g_{I_{C_i}}(x)$
- Proximal Minimization: $f(x) = 0$
- Convergence: If $f(x)$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(x)$ is convex and $prox_t(x)$ can be solved exactly then convergence result (and proof) is similar to that for gradient descent

& $t^{(k)} < 1/L$

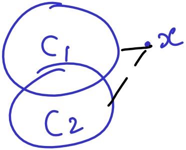
$$f(x^k) - f(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}\left(f(x^i) - f(x^*)\right) \leq \frac{\left\|x^{(0)} - x^*\right\|^2}{2tk}$$

Eg: Projected Gradient Descent

$g_I = \max_i dist(x, C_i)$

- Let

$$dist(x, C_i) = \min_{u \in C_i} ||x - u||^2$$



- We define

$$D(x) = \max_i dist(x, C_i)$$

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

  - If C_i is closed and convex, a unique minimizer $P_{C_i}(x)$ exists (projection of x on C_i)
  - $dist(x, C_i) = 0$ if $x \in C_i$

- Recall discussion on subgradient descent for this problem in class notes[4]

[4] http://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture22a.pdf

Projected gradient descent

$x^{(0)}$

$z^{(k)} = x^{(k)} - t\nabla f(x^{(k)})$

Compute $x^{(k+1)}$ by appling alternating (subgradient descent-based) projections with $C_i = \{x \mid g_i(x) \leq 0\}$

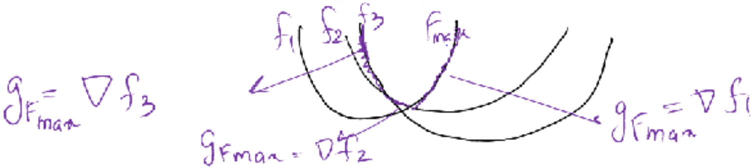
Refer to page 17 of http://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture22a.pdf
for details of the subgradient ascent algorithm

- We get the subgradient of $D(x)$ as

$$g_D(x) = \nabla dist(x, C_i) \text{ if } D(x) = dist(x, C_i)$$

- For illustration, consider

$$g_{F_{max}}(x) = \nabla f_i(x) \text{ if } f_i(x) = \max_j f_j(x)$$



  - If f_i gives maximum value at a point, $g_{F_{max}}$ will be ∇f_i at that point
  - At the points of intersection of f_i and f_j, we will get some convex combination of ∇f_i and ∇f_j