Constrained Optimization in $\Re$: Recap

Global Extrema on Closed Intervals

Recall the extreme value theorem. A consequence is that:

- if either of c or d lies in (a, b), then it is a critical number of f;
- else each of c and d must lie on one of the boundaries of $[a, b]$.

This gives us a procedure for finding the maximum and minimum of a continuous function f on a closed bounded interval $\mathcal{I}$:

Procedure

[Finding extreme values on closed, bounded intervals]:

1. *Find the critical points in int$(\mathcal{I})$.*
2. *Compute the values of f at the critical points and at the endpoints of the interval.*
3. *Select the least and greatest of the computed values.*

Global Extrema on Closed Intervals (contd)

- To compute the maximum and minimum values of $f(x) = 4x^3 - 8x^2 + 5x$ on the interval $[0, 1]$,

Global Extrema on Closed Intervals (contd)

- To compute the maximum and minimum values of $f(x) = 4x^3 - 8x^2 + 5x$ on the interval $[0, 1]$,
  - We first compute $f'(x) = 12x^2 - 16x + 5$ which is 0 at $x = \frac{1}{2}, \frac{5}{6}$.
  - Values at the critical points are $f(\frac{1}{2}) = 1$, $f(\frac{5}{6}) = \frac{25}{27}$.
  - The values at the end points are $f(0) = 0$ and $f(1) = 1$.
  - Therefore, the minimum value is $f(0) = 0$ and the maximum value is $f(1) = f(\frac{1}{2}) = 1$.

Global Extrema on Closed Intervals (contd)

- To compute the maximum and minimum values of $f(x) = 4x^3 - 8x^2 + 5x$ on the interval $[0, 1]$,
  - We first compute $f'(x) = 12x^2 - 16x + 5$ which is 0 at $x = \frac{1}{2}, \frac{5}{6}$.
  - Values at the critical points are $f(\frac{1}{2}) = 1$, $f(\frac{5}{6}) = \frac{25}{27}$.
  - The values at the end points are $f(0) = 0$ and $f(1) = 1$.
  - Therefore, the minimum value is $f(0) = 0$ and the maximum value is $f(1) = f(\frac{1}{2}) = 1$.
- In this context, it is relevant to discuss the one-sided derivatives of a function at the endpoints of the closed interval on which it is defined.

Global Extrema on Closed Intervals (contd)

Definition

[One-sided derivatives at endpoints]: *Let f be defined on a closed bounded interval $[a, b]$. The (right-sided) derivative of f at $x = a$ is defined as*

$$f'(a) = \lim_{h \to 0^+} \frac{f(a+h) - f(a)}{h}$$

Similarly, the (left-sided) derivative of f at $x = b$ is defined as

$$f'(b) = \lim_{h \to 0^-} \frac{f(b+h) - f(b)}{h}$$

Essentially, each of the one-sided derivatives defines one-sided slopes at the endpoints.

Global Extrema on Closed Intervals (contd)

Based on these definitions, the following result can be derived.

Claim

If f is continuous on $[a, b]$ and $f'(a)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at a.

- *If $f(a)$ is the maximum value of f on $[a, b]$, then $f'(a) \leq 0$ or $f'(a) = -\infty$.*
- *If $f(a)$ is the minimum value of f on $[a, b]$, then $f'(a) \geq 0$ or $f'(a) = \infty$.*

If f is continuous on $[a, b]$ and $f'(b)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at b

Global Extrema on Closed Intervals (contd)

Based on these definitions, the following result can be derived.

Claim

If f is continuous on $[a,b]$ and $f'(a)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at a.

- *If $f(a)$ is the maximum value of f on $[a,b]$, then $f'(a) \leq 0$ or $f'(a) = -\infty$.*
- *If $f(a)$ is the minimum value of f on $[a,b]$, then $f'(a) \geq 0$ or $f'(a) = \infty$.*

If f is continuous on $[a,b]$ and $f'(b)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at b

- *If $f(b)$ is the maximum value of f on $[a,b]$, then $f'(b) \geq 0$ or $f'(b) = \infty$.*
- *If $f(b)$ is the minimum value of f on $[a,b]$, then $f'(b) \leq 0$ or $f'(b) = -\infty$.*

Global Extrema on Closed Intervals (contd)

The following result gives a useful procedure for finding **extrema on closed intervals**.

Claim

If f is continuous on $[a, b]$ and $f''(x)$ exists for all $x \in (a, b)$. Then,

- *If $f''(x) \leq 0,\ \forall x \in (a, b)$, then the minimum value of f on $[a, b]$ is either $f(a)$ or $f(b)$. If, in addition, f has a critical point $c \in (a, b)$, then $f(c)$ is the maximum value of f on $[a, b]$.*
- *If $f''(x) \geq 0,\ \forall x \in (a, b)$, then the maximum value of f on $[a, b]$ is either $f(a)$ or $f(b)$. If, in addition, f has a critical point $c \in (a, b)$, then $f(c)$ is the minimum value of f on $[a, b]$.*

Global Extrema on Open Intervals

The next result is very useful for finding **extrema on open intervals.**

Claim

Let $\mathcal{I}$ be an open interval and let $f'(x)$ exist $\forall x \in \mathcal{I}$.

- *If $f'(x) \geq 0, \ \forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global minimum value of f on $\mathcal{I}$.*
- *If $f'(x) \leq 0, \ \forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global maximum value of f on $\mathcal{I}$.*

For example, let $f(x) = \frac{2}{3}x - \sec x$ and
$\mathcal{I} = (\frac{-\pi}{2}, \frac{\pi}{2})$.

Global Extrema on Open Intervals

The next result is very useful for finding **extrema on open intervals.**

Claim

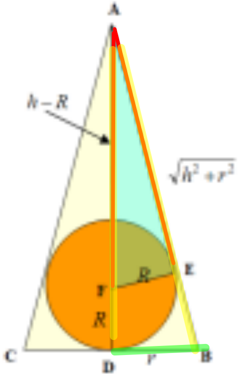
Let $\mathcal{I}$ be an open interval and let $f'(x)$ exist $\forall x \in \mathcal{I}$.

- *If $f'(x) \geq 0, \ \forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global minimum value of f on $\mathcal{I}$.*
- *If $f'(x) \leq 0, \ \forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global maximum value of f on $\mathcal{I}$.*

For example, let $f(x) = \frac{2}{3}x - \sec x$ and $\mathcal{I} = (\frac{-\pi}{2}, \frac{\pi}{2})$. $f'(x) = \frac{2}{3} - \sec x \tan x = \frac{2}{3} - \frac{\sin x}{\cos^2 x} = 0 \Rightarrow x = \frac{\pi}{6}$. Further, $f''(x) = -\sec x(\tan^2 x + \sec^2 x) < 0$ on $(\frac{-\pi}{2}, \frac{\pi}{2})$. Therefore, f attains the maximum value $f(\frac{\pi}{6}) = \frac{\pi}{9} - \frac{2}{\sqrt{3}}$ on $\mathcal{I}$.

Global Extrema on Open Intervals (contd)

As another example, let us find the dimensions of the cone with minimum volume that can contain a sphere with radius R. Let h be the height of the cone and r the radius of its base. The objective to be minimized is the volume $f(r,h) = \frac{1}{3}\pi r^2 h$. The constraint betwen r and h is shown in Figure 10. The traingle AEF is similar to traingle ADB and therefore, $\frac{h-R}{R} = \frac{\sqrt{h^2+r^2}}{r}$.



Global Extrema on Open Intervals (contd)

Our first step is to reduce the volume formula to involve only one of[8] r^2 or h.
The algebra involved will be the simplest if we solved for h.
The constraint gives us $r^2 = \frac{R^2 h}{h-2R}$. Substituting this expression for r^2 into the volume formula, we get $g(h) = \frac{\pi R^2}{3}\frac{h^2}{(h-2R)}$ with the domain given by $\mathcal{D} = \left\{h|2R < h < \infty\right\}$.
Note that $\mathcal{D}$ is an open interval.
$g' = \frac{\pi R^2}{3}\frac{2h(h-2R)-h^2}{(h-2R)^2} = \frac{\pi R^2}{3}\frac{h(h-4R)}{(h-2R)^2}$ which is 0 in its domain $\mathcal{D}$ if and only if $h = 4R$.
$g'' = \frac{\pi R^2}{3}\frac{2(h-2R)^3-2h(h-4R)(h-2R)^2}{(h-2R)^4} = \frac{\pi R^2}{3}\frac{2(h^2-4Rh+4R^2-h^2+4Rh)}{(h-2R)^3} = \frac{\pi R^2}{3}\frac{8R^2}{(h-2R)^3}$, which is greater than 0 in $\mathcal{D}$.
Therefore, g (and consequently f) has a unique minimum at $h = 4R$ and correspondingly, $r^2 = \frac{R^2 h}{h-2R} = 2R^2$.

[8] Since r appears in the volume formula only in terms of r^2.

Constrained Optimization and Subgradient Descent

Constrained Optimization

- Consider the objective

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

- Recall: Indicator function for $g_i(x)$

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

  - We have shown that this is convex if each $g_i(\mathbf{x})$ is convex.
- Option 1: Subgradient descent on f(x) + I_g(x)

Constrained Optimization

- Consider the objective

$$\min f(\mathbf{x})$$

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

- Recall: Indicator function for $g_i(x)$

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

  - We have shown that this is convex if each $g_i(\mathbf{x})$ is convex.
- Option 1: Use subgradient descent to minimize $f(\mathbf{x}) + \sum_i I_{g_i}(\mathbf{x})$
- Option 2:

Constrained Optimization

- Consider the objective

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

- Recall: Indicator function for $g_i(x)$

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

  - We have shown that this is convex if each $g_i(\mathbf{x})$ is convex.
- Option 1: Use subgradient descent to minimize $f(\mathbf{x}) + \sum_i I_{g_i}(\mathbf{x})$
- Option 2: Barrier Method (approximate $I_{g_i}(\mathbf{x})$ using some differentiable and non-decreasing function such as $-(1/t)\log -u$), Augmented Lagrangian, ADMM, *etc.*

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

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

- Recap a subgradient of F:

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

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

- Recap a subgradient of F: $\mathbf{h}_F(\mathbf{x}) = \mathbf{h}_f(\mathbf{x}) + \sum_i \mathbf{h}_{I_{C_i}}(\mathbf{x})$. Recall that

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

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

- Recap a subgradient of F: $\mathbf{h}_F(\mathbf{x}) = \mathbf{h}_f(\mathbf{x}) + \sum_i \mathbf{h}_{I_{C_i}}(\mathbf{x})$. Recall that
  - $\mathbf{h}_f(\mathbf{x}) = \nabla f(\mathbf{x})$ if $f(\mathbf{x})$ is differentiable. Also, $-\nabla f(\mathbf{x})$ at $\mathbf{x}^k$ optimizes

Let us treat the gradient of f at x^k as that vector which minimized the second order quadratic expansion of f around x^k

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \le 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

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

- Recap a subgradient of F: $\mathbf{h}_F(\mathbf{x}) = \mathbf{h}_f(\mathbf{x}) + \sum_i \mathbf{h}_{I_{C_i}}(\mathbf{x})$. Recall that
  - $\mathbf{h}_f(\mathbf{x}) = \nabla f(\mathbf{x})$ if $f(\mathbf{x})$ is differentiable. Also, $-\nabla f(\mathbf{x})$ at $\mathbf{x}^k$ optimizes the first order approximation for $f(\mathbf{x})$ around $\mathbf{x}^k$: $-\nabla f(\mathbf{x}) = \underset{\mathbf{h}}{\operatorname{argmin}}\, f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)\mathbf{h} + \frac{1}{2}||\mathbf{h}||^2$:
    Variations on the form of $\frac{1}{2}||\mathbf{h}||^2$ lead to Mirror Descent etc. replacing with entropic regularizer
  - $\mathbf{h}_{I_{C_i}}(x)$ is $\mathbf{d} \in \mathbf{R}^n$ s.t. $\mathbf{d}^T\mathbf{x} \ge \mathbf{d}^T\mathbf{y}$, $\forall \mathbf{y} \in C_i$. Also, $\mathbf{h}_{I_{C_i}}(\mathbf{x}) = 0$ if $\mathbf{x}$ is in the interior of C_i, and has other solutions if $\mathbf{x}$ is on the boundary:
    Analysis for convex g_i's leads to KKT conditions and Dual Ascent etc.

Option 1: Generalized Gradient Descent

- Consider the problem of minimizing the following sum of a differentiable function $f(\mathbf{x})$ and a (possibly) nondifferentiable function $c(\mathbf{x})$ (an example being $\sum_i I_{C_i}(\mathbf{x})$)

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + c(\mathbf{x})$$

- As in gradient descent, consider the first order approximation for $f(\mathbf{x})$ around $\mathbf{x}^k$ leaving $c(\mathbf{x})$ alone to obtain the next iterate $\mathbf{x}^{k+1}$:

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2t}||\mathbf{x} - \mathbf{x}^k||^2 + c(\mathbf{x})$$

Option 1: Generalized Gradient Descent

- Consider the problem of minimizing the following sum of a differentiable function $f(\mathbf{x})$ and a (possibly) nondifferentiable function $c(\mathbf{x})$ (an example being $\sum_i I_{C_i}(\mathbf{x})$)

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + c(\mathbf{x})$$

- As in gradient descent, consider the first order approximation for $f(\mathbf{x})$ around $\mathbf{x}^k$ leaving $c(\mathbf{x})$ alone to obtain the next iterate $\mathbf{x}^{k+1}$:

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

- Deleting $f(\mathbf{x}^k)$ from the objective and adding $\frac{t}{2}||\nabla f(\mathbf{x}^k)||^2$ to the objective (without any loss) to complete squares, we obtain $\mathbf{x}^{k+1}$ as:

Option 1: Generalized Gradient Descent

- Consider the problem of minimizing the following sum of a differentiable function $f(\mathbf{x})$ and a (possibly) nondifferentiable function $c(\mathbf{x})$ (an example being $\sum_i I_{C_i}(\mathbf{x})$)

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + c(\mathbf{x})$$

- As in gradient descent, consider the first order approximation for $f(\mathbf{x})$ around $\mathbf{x}^k$ leaving $c(\mathbf{x})$ alone to obtain the next iterate $\mathbf{x}^{k+1}$:

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

- Deleting $f(\mathbf{x}^k)$ from the objective and adding $\frac{t}{2}||\nabla f(\mathbf{x}^k)||^2$ to the objective (without any loss) to complete squares, we obtain $\mathbf{x}^{k+1}$ as:
$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}} \frac{1}{2t}||\mathbf{x} - (\mathbf{x}^k - t\nabla f(\mathbf{x}^k))||^2 + c(\mathbf{x})$

(point closest to the unregulated gradient descent update with a later regulation using c(x))

- In general, such a step is called a *proximal step with respect to $c(\mathbf{x})$*

$$\mathbf{x}^{k+1} = prox_c\left(\mathbf{x}^k - t\nabla f(\mathbf{x}^k)\right) = \operatorname{argmin}_{\mathbf{x}} \frac{1}{2t}||\mathbf{x} - (\mathbf{x}^k - t\nabla f(\mathbf{x}^k))||^2 + c(\mathbf{x})$$

this unregulated descent will be often referred to as z

PROX gives you the point closes to the unregulated (wrt to c(x)) update when we also bring in the effect of (minimizing) c(x)

Basically we have phased out the subgradient descent update into two phases

(a) unregulated update (such as gradient descent) for f(x) alone

(b) course correction based on c(x)

Algorithm: The Generalized Gradient Descent

$$\min_{\mathbf{x}} f(\mathbf{x}) + c(\mathbf{x})$$

Find a starting point $\mathbf{x}_p^0$. $=$
Set $k = 1$
repeat

1. Choose a step size $t^k \propto 1/\sqrt{k}$ or using exact or backtracking ray search or .
2. Set $\mathbf{z}^k = \mathbf{x}^{k-1} - t^k \nabla f(\mathbf{x}^{k-1})$.
3. Set $\mathbf{x}^k = prox_c\left(\mathbf{z}^k\right)$.
4. Set $k = k + 1$.

until stopping criterion (such as $||\mathbf{x}^k - \mathbf{x}^{k-1}|| \leq \epsilon$ or $f(\mathbf{x}^k) > f(\mathbf{x}^{k-1})$) is satisfied[a]

[a] Better criteria can be found using Lagrange duality theory, etc.

Figure 11: The generalized gradient descent algorithm.

Option 1: Generalized Gradient Descent

- Interesting because in many settings, $prox_c(\mathbf{z})$ can be computed efficiently

$$prox_c(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}}\ \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + c(\mathbf{x})$$

- Theorem: If c is a proper convex[9] function with a closed epigraph then (for $t > 0$) it has a unique value of $prox_c(\mathbf{z})$. *Hint: The quadratic term introduces strong convexity* $\Rightarrow$ *strict convexity*. A strictly convex function has a unique minimizer

it is finite value < +inf atleast at one point and is not -inf everywhere else

Option 1: Generalized Gradient Descent

- Interesting because in many settings, $prox_c(\mathbf{z})$ can be computed efficiently

$$prox_c(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}} \frac{1}{2t}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

- Theorem: If c is a proper convex[9] function with a closed epigraph then (for $t > 0$) it has a unique value of $prox_c(\mathbf{z})$. *Hint: The quadratic term introduces strong convexity ⇒ strict convexity*. For non-convex c, the solution set is non-empty under similar conditions.

For $x \in \Re$, $c(\mathbf{x}) =$	For $z \in \Re$ & $t = 1$, $prox_c(z) =$
Simplified Lasso: $\lambda\|x\|_1$	??
μx $x \geq 0$; ∞ $x < 0$	??
$\mu\lambda x^3$ $x \geq 0$; ∞ $x < 0$	??
$-\lambda \log x$ $x > 0$; ∞ $x \leq 0$	?? Inspired by or inspires barrier function
$\delta_{[0,\eta]\cap\Re}$	??

Option 1: Generalized Gradient Descent

- Interesting because in many settings, $prox_c(\mathbf{z})$ can be computed efficiently

$$prox_c(\mathbf{z}) = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + c(\mathbf{x})$$

- Theorem: If c is a proper convex[9] function with a closed epigraph then (for $t > 0$) it has a unique value of $prox_c(\mathbf{z})$. *Hint: The quadratic term introduces strong convexity $\Rightarrow$ strict convexity*. For non-convex c, the solution set is non-empty under similar conditions.

For $x \in \Re$, $c(\mathbf{x}) =$	For $z \in \Re$ & $t = 1$, $prox_c(z) =$
Simplified Lasso: $\lambda\|x\|_1$	??
$\begin{cases}\mu x & x \geq 0 \\ \infty & x < 0\end{cases}$	??
$\begin{cases}\mu\lambda x^3 & x \geq 0 \\ \infty & x < 0\end{cases}$	??
$\begin{cases}-\lambda \log x & x > 0 \\ \infty & x \leq 0\end{cases}$	??
$\delta_{[0,\eta]\cap\Re}$	??

$c(\mathbf{x}) =$	For $t = 1$, $prox_c(\mathbf{z}) =$
Constant: c	??
Affine: $\mathbf{a}^T\mathbf{x} + b$	??
Convex quadratic: $\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ (where $A \in S^n_+, \mathbf{b} \in \Re^n$)	??
Sum over components: $c(\mathbf{x}) = \sum_{i=1}^{n} c_i(\mathbf{x}_i)$	???
$c(\lambda\mathbf{x} + \mathbf{a})$	??
$\lambda c\left(\frac{1}{\lambda}\mathbf{x}\right)$	??
$c(\mathbf{x}) + \mathbf{a}^T\mathbf{x} + \frac{\beta}{2}\|\mathbf{x}\|^2 + \gamma$	??
$c(A\mathbf{x} + \mathbf{b})$	??
$c(\|\mathbf{x}\|)$	??

calculus

Iterative Soft Thresholding Algorithm for Solving Lasso

Proximal Subgradient Descent for Lasso

- Let $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2$, $c(\mathbf{x}) = \|\mathbf{x}\|_1$ and $F(\mathbf{x}) = f(\mathbf{x}) + c(\mathbf{x})$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{x}^{(0)}$
  - Let $\widehat{\mathbf{x}}^{(k+1)} \equiv \mathbf{z}^{(k+1)}$ be a next gradient descent iterate for $f(\mathbf{x}^k)$
  - Compute $\mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\mathbf{argmin}} \frac{1}{2}\|\mathbf{x} - \mathbf{z}^{(k+1)}\|_2^2 + \lambda t\|\mathbf{x}\|_1$ by setting subgradient of this objective to $\mathbf{0}$. This results in (see https://www.cse.iitb.ac.in/~cs709/notes/enotes/lassoElaboration.pdf)
    1. ...
    2. ... Vector x^(k+1) is obtained by componentwise minimization
    3. ...
  - Set $k = k + 1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{x}^k$ w.r.t $\mathbf{x}^{(k-1)}$)

prox step

Iterative Soft Thresholding Algorithm (Proximal Subgradient Descent) for Lasso

- Let $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2$, $c(\mathbf{x}) = \|\mathbf{x}\|_1$ and $F(\mathbf{x}) = f(\mathbf{x}) + c(\mathbf{x})$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{x}^{(0)}$
  - Let $\mathbf{z}^{(k+1)}$ be a next gradient descent iterate for $f(\mathbf{x}^k)$
  - Compute $prox_{\|\mathbf{x}\|_1}\left(\mathbf{z}^{(k+1)}\right) = \mathbf{x}^{(k+1)} =$

Iterative Soft Thresholding Algorithm (Proximal Subgradient Descent) for Lasso

- Let $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2$, $c(\mathbf{x}) = \|\mathbf{x}\|_1$ and $F(\mathbf{x}) = f(\mathbf{x}) + c(\mathbf{x})$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{x}^{(0)}$
  - Let $\mathbf{z}^{(k+1)}$ be a next gradient descent iterate for $f(\mathbf{x}^k)$
  - Compute $prox_{\|\mathbf{x}\|_1}\left(\mathbf{z}^{(k+1)}\right) = \mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\textbf{argmin}} \frac{1}{2t}\|\mathbf{x} - \mathbf{z}^{(k+1)}\|_2^2 + \lambda\|\mathbf{x}\|_1$ as follows:

Iterative Soft Thresholding Algorithm (Proximal Subgradient Descent) for Lasso

- Let $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2$, $c(\mathbf{x}) = \|\mathbf{x}\|_1$ and $F(\mathbf{x}) = f(\mathbf{x}) + c(\mathbf{x})$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{x}^{(0)}$
  - Let $\mathbf{z}^{(k+1)}$ be a next gradient descent iterate for $f(\mathbf{x}^k)$
  - Compute $prox_{\|\mathbf{x}\|_1}\left(\mathbf{z}^{(k+1)}\right) = \mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\textbf{argmin}} \frac{1}{2t}\|\mathbf{x} - \mathbf{z}^{(k+1)}\|_2^2 + \lambda\|\mathbf{x}\|_1$ as follows:
    1. If $z_i^{(k+1)} > \lambda t$, then $x_i^{(k+1)} = -\lambda t + z_i^{(k+1)}$
    2. If $z_i^{(k+1)} < -\lambda t$, then $x_i^{(k+1)} = \lambda t + z_i^{(k+1)}$
    3. 0 otherwise.

    If unregulated z was gretater than lambda t reduce it by that amount
  - Set $k = k + 1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{x}^k$ w.r.t $\mathbf{x}^{(k-1)}$)

Tables for the Proximal Operator

$$prox_c(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}} \frac{1}{2t}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

For $x \in \Re$, $c(x) =$	For $z \in \Re$ & $t = 1$, $prox_c(z) =$
Simplified Lasso: $\lambda\|x\|$	$[\|x\| - \lambda]_+ sign(x)$
$\mu x \quad x \geq 0$ $\infty \quad x < 0$	$[x - \mu]_+$
$\mu\lambda x^3 \quad x \geq 0$ $\infty \quad x < 0$	$\frac{-1 + \sqrt{1 + 12\lambda[x]_+}}{6\lambda}$
$-\lambda \log x \quad x > 0$ $\infty \quad x \leq 0$	$\frac{x + \sqrt{x^2 + 4\lambda}}{2}$
$\delta_{[0,\eta]\cap\Re}$	$min\{max\{x, 0\}, \eta\}$

Tables for the Proximal Operator

$$prox_c(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}} \frac{1}{2t}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

For $x \in \Re$, $c(x) =$	For $z \in \Re$ & $t = 1$, $prox_c(z) =$
Simplified Lasso: $\lambda\|x\|$	$[\|x\| - \lambda]_+ sign(x)$
$\mu x \quad x \geq 0$; $\infty \quad x < 0$	$[x - \mu]_+$
$\mu\lambda x^3 \quad x \geq 0$; $\infty \quad x < 0$	$\frac{-1 + \sqrt{1 + 12\lambda[x]_+}}{6\lambda}$
$-\lambda \log x \quad x > 0$; $\infty \quad x \leq 0$	$\frac{x + \sqrt{x^2 + 4\lambda}}{2}$
$\delta_{[0,\eta]\cap\Re}$	$min\{max\{x, 0\}, \eta\}$

For $x \in \Re$, $c(\mathbf{x}) =$	For $z \in \Re$ & $t = 1$, $prox_c(\mathbf{z}) =$
Constant: c	$\mathbf{z}$
Affine: $\mathbf{a}^T\mathbf{x} + b$	$\mathbf{z} - \mathbf{a}$
Convex quadratic: $\frac{1}{2}\mathbf{x}^TA\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ (where $A \in S^n_+, \mathbf{b} \in \Re^n$)	$(A + I)^{-1}(\mathbf{z} - \mathbf{b})$

Tables for the Proximal Operator

$$prox_c(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}} \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + c(\mathbf{x})$$

For $x \in \Re$, $c(x) =$	For $z \in \Re$ & $t = 1$, $prox_c(z) =$
Simplified Lasso: $\lambda\|x\|$	$[\|x\| - \lambda]_+ sign(x)$
$\mu x \quad x \geq 0$ $\infty \quad x < 0$	$[x - \mu]_+$
$\mu\lambda x^3 \quad x \geq 0$ $\infty \quad x < 0$	$\frac{-1+\sqrt{1+12\lambda[x]_+}}{6\lambda}$
$-\lambda \log x \quad x > 0$ $\infty \quad x \leq 0$	$\frac{x+\sqrt{x^2+4\lambda}}{2}$
$\delta_{[0,\eta]\cap\Re}$	$min\{max\{x, 0\}, \eta\}$

For $x \in \Re$, $c(\mathbf{x}) =$	For $z \in \Re$ & $t = 1$, $prox_c(\mathbf{z}) =$
Constant: c	$\mathbf{z}$
Affine: $\mathbf{a}^T\mathbf{x} + b$	$\mathbf{z} - \mathbf{a}$
Convex quadratic: $\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ (where $A \in S^n_+, \mathbf{b} \in \Re^n$)	$(A + I)^{-1}(\mathbf{z} - \mathbf{b})$
Sum over components: $c(\mathbf{x}) = \sum_{i=1}^{n} c_i(\mathbf{x}_i)$	???
$c(\lambda\mathbf{x} + \mathbf{a})$	??
$\lambda c\left(\frac{1}{\lambda}\mathbf{x}\right)$	??
$c(\mathbf{x}) + \mathbf{a}^T\mathbf{x} + \frac{\beta}{2}\|\|\mathbf{x}\|\|^2 + \gamma$	??
$c(A\mathbf{x} + \mathbf{b})$	??
$c(\|\|\mathbf{x}\|\|)$	??