Iterative Soft Thresholding Algorithm (Proximal Subgradient Descent) for Lasso

min f(x) + c(x)

- 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 $prox_{\|\mathbf{x}\|_1}\left(\mathbf{z}^{(k+1)}\right) = \mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\mathbf{argmin}} \underline{\frac{1}{2t}}\|\mathbf{x} - \mathbf{z}^{(k+1)}\|_2^2 + \lambda\|\mathbf{x}\|_1$ as follows:
    1. If $z_i^{(k+1)} > \lambda t$, then $x_i^{(k+1)} = -\underline{\lambda t +} z_i^{(k+1)}$
    2. If $z_i^{(k+1)} < -\lambda t$, then $x_i^{(k+1)} = \underline{\lambda t + z_i}^{(k+1)}$
    3. 0 otherwise.
  - 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

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

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

we have already derived this first entry in the table

Since x^3 is differentiable, the penalization of \lambda on z is much softer

Tables for the Proximal Operator

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

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

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 $A \in S^n_+, \mathbf{b} \in \Re^n$)	$(A + I)^{-1}(\mathbf{z} - \mathbf{b})$

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: $\lambda\lvert x\rvert$	$[\lvert z\rvert - \lambda]_+ sign(z)$
$\lambda x \quad x \geq 0$ $\infty \quad x < 0$	$[z-\lambda]_+$
$\lambda x^3 \quad x \geq 0$ $\infty \quad x < 0$	$\frac{-1+\sqrt{1+12\lambda[z]_+}}{6\lambda}$
$-\lambda \log x \quad x > 0$ $\infty \quad x \leq 0$	$\frac{z+\sqrt{z^2+4\lambda}}{2}$
$\lambda x \quad 0 \leq x \leq \alpha$ $\infty \quad \text{otherwise}$	$min\{max\{z-\lambda, 0\}, \alpha\}$

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: $c(\mathbf{x}) = \sum_{i=1}^{n} c_i(\mathbf{x}_i)$	???
$c(\lambda\mathbf{x} + \mathbf{a})$	??
$\lambda c\left(\frac{1}{\lambda}\mathbf{x}\right)$	??
$c(\mathbf{x}) + \mathbf{a}^T\mathbf{x} + \frac{\beta}{2}\lVert\mathbf{x}\rVert^2 + \gamma$	??
$c(A\mathbf{x} + \mathbf{b})$	??
$c(\lVert\mathbf{x}\rVert)$	??

Can we recover the prox of the composition of function as a composition of prox operations

Calculus for the Proximal Operator: See https://archive.siam.org/books/mo25/mo25_ch6.pdf

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

$c(\mathbf{x}) =$	For $t=1$, $prox_c(\mathbf{z}) =$
Sum over components: $c(\mathbf{x}) = \sum_{i=1}^{n} c_i(\mathbf{x}_i)$ where $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$	Product over components: $prox_c(\mathbf{z}) = \prod_{i=1}^{n} prox_{c_i}(\mathbf{z}_i)$ where $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]$
$c(\lambda\mathbf{x}+\mathbf{a})$ where $\lambda \neq 0$ and c is proper	$\frac{1}{\lambda}\left[prox_{\lambda^2 c}(\lambda\mathbf{z}+\mathbf{a}) - \mathbf{a}\right]$
$\lambda c\left(\frac{1}{\lambda}\mathbf{x}\right)$ where $\lambda \neq 0$ and c is proper	$\lambda prox_{c/\lambda}\left(\frac{1}{\lambda}\mathbf{z}\right)$
$c(\mathbf{x}) + \mathbf{a}^T\mathbf{x} + \frac{\beta}{2}\|\mathbf{x}\|^2 + \gamma$ where $\beta > 0$, $\gamma \in \Re$, c is proper	$prox_{\frac{1}{\beta+1}c}\left(\frac{\mathbf{z}-\mathbf{a}}{\gamma+1}\right)$
$c(A\mathbf{x}+\mathbf{b})$ where c is proper closed and convex, $\mathbf{b} \in \Re^n$, $AA^T = \alpha I$, $\alpha > 0$	$\mathbf{z} + \frac{1}{\alpha}A^T\left(prox_{\alpha c}(A\mathbf{z}+\mathbf{b}) - A\mathbf{z} - \mathbf{b}\right)$
$c(\|\mathbf{x}\|)$ where $\mathbf{b} \in \Re^n$, $AA^T = \alpha I$, $\alpha > 0$	$prox_c(\|\mathbf{z}\|)\frac{\mathbf{z}}{\|\mathbf{z}\|}$ $\quad \mathbf{z} \neq 0$ $\{\mathbf{u} \mid \|\mathbf{u}\| = prox_c(0)\}$ $\quad \mathbf{z} = 0$

Generalized Gradient Descent and its Special Cases

Recall

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

It's special cases are:

1. Gradient Descent ⇒ c(x) = constant

Generalized Gradient Descent and its Special Cases

Recall

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

It's special cases are:

1. Gradient Descent $\Rightarrow c(\mathbf{x}) = 0$
2. Projected Gradient Descent $\Rightarrow$ c(x) = Indicator function of the constraint function g(x) <=0

Generalized Gradient Descent and its Special Cases

Recall

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

It's special cases are:

1. Gradient Descent ⇒ $c(\mathbf{x}) = 0$
2. Projected Gradient Descent ⇒ $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$ (Example:

Generalized Gradient Descent and its Special Cases

Recall

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

It's special cases are:

1. Gradient Descent ⇒ $c(\mathbf{x}) = 0$
2. Projected Gradient Descent ⇒ $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$ (Example: $= \sum_i \underline{I_{g_i}(\mathbf{x})}$)
3. Alternating Projection/Proximal Minimization: $f(\mathbf{x}) = 0$
4. Alternating Direction Method of Multipliers
5. Special Cases for Specific Objectives
   - LASSO: (Fast) Iterative Shrinkage Thresholding Algorithm (ISTA/FISTA)

and c(x) = sum
of indicators

Case 2: Projection Methods

Demystifying the Projection Step

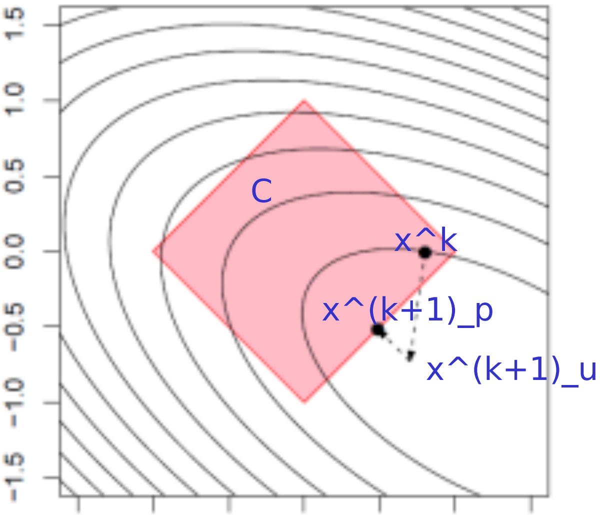
$$\begin{aligned}\mathbf{x}_p^{(k+1)} &= prox_{I_{\mathcal{C}}}\left(\mathbf{x}_u^{(k+1)}\right) = \underset{\mathbf{x}}{\operatorname{argmin}} \left\|\mathbf{x}_u^{(k+1)} - \mathbf{x}\right\|_2^2 + \underline{I_{\mathcal{C}}(\mathbf{x})} \\ &= \underset{\underline{\mathbf{x}\in\mathcal{C}}}{\operatorname{argmin}} \left\|\mathbf{x}_u^{(k+1)} - \mathbf{x}\right\|_2^2 = \text{Proj\_C(z)}\end{aligned}$$

this term dominates

Demystifying the Projection Step

$$\begin{aligned}
\mathbf{x}_p^{(k+1)} &= \mathit{prox}_{I_{\mathcal{C}}}\left(\mathbf{x}_u^{(k+1)}\right) &&= \operatorname*{argmin}_{\mathbf{x}} \left\|\mathbf{x}_u^{(k+1)} - \mathbf{x}\right\|_2^2 + I_{\mathcal{C}}(\mathbf{x}) \\
&= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{C}} \left\|\mathbf{x}_u^{(k+1)} - \mathbf{x}\right\|_2^2 &&= P_{\mathcal{C}}\left(\mathbf{x}_u^{(k+1)}\right)
\end{aligned}$$

Projected Gradient Descent: Illustrated



Algorithm: Projected Gradient Descent (We use $\mathbf{x}_u^k$ instead of $\mathbf{z}^k$)

```
Find a starting point $\mathbf{x}_p^0 \in \mathcal{C}$.
Set $k = 1$
repeat
    1. Choose a step size $t^k \propto 1/\sqrt{k}$.
    2. Set $\mathbf{x}_u^k = \mathbf{x}_p^{k-1} - t^k \nabla f(\mathbf{x}_p^{k-1})$.
    3. Set $\boxed{\mathbf{x}_p^k = \underset{\mathbf{z} \in \mathcal{C}}{\operatorname{argmin}} \left|\left|\mathbf{x}_u^k - \mathbf{z}\right|\right|_2^2.}$
    4. Set $k = k + 1$.
until stopping criterion (such as $||\mathbf{x}_p^k - \mathbf{x}_p^{k-1}|| \le \epsilon$ or $f(\mathbf{x}_p^k) > f(\mathbf{x}_p^{k-1})$) is satisfied[a]
```

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

Figure 12: The projected gradient descent algorithm.

Table of Orthogonal Projections: See https://archive.siam.org/books/mo25/mo25_ch6.pdf

$$P_C(\mathbf{z}) = prox_{I_C}(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}}\ \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + I_C(\mathbf{x}) = \underset{\mathbf{x}\in C}{\text{argmin}}\ \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2$$

Observation: All expressions are about dropping perpendicular from z to the constant set

Set $C =$	For $t = 1$, $P_C(\mathbf{z}) =$	Assumptions
$\Re^n_+$	$[\mathbf{z}]_+$	
$\text{Box}[\mathbf{l},\mathbf{u}]$	$P_C(\mathbf{z})_i = min\{max\{z_i, l_i\}, u_i\}$	$l_i \leq u_i$
$\text{Ball}[\mathbf{c}, r]$	$\mathbf{c} + \frac{r}{max\{\|\mathbf{z}-\mathbf{c}\|_2, r\}}(\mathbf{z}-\mathbf{c})$ if max $\doteq$ r, P(z) = z	$\|.\|_2$ ball, centre $\mathbf{c} \in \Re^n$ & radius $r > 0$
$\{\mathbf{x} \| A\mathbf{x} = \mathbf{b}\}$	$\mathbf{z} - A^T(AA^T)^{-1}(A\mathbf{z}-\mathbf{b})$ derived on subsequent slides	$A \in \Re^{m\times n}$, $\mathbf{b} \in \Re^m$, A is full row rank
$\{\mathbf{x} \| \mathbf{a}^T\mathbf{x} \leq b\}$	$\mathbf{z} - \frac{[\mathbf{a}^T\mathbf{x}-b]_+}{\|\mathbf{a}\|^2}$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
Δ_n	$[\mathbf{z}-\mu^*\mathbf{e}]_+$ where $\mu^* \in \Re$ satisfies $\mathbf{e}^T[\mathbf{z}-\mu^*\mathbf{e}]_+ = 1$	
$\underline{H_{\mathbf{a},b}} \cap \text{Box}[\mathbf{l},\mathbf{u}]$ hyperplane	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a})$ where $\mu^* \in \Re$ satisfies $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a}) = b$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$H^-{}_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$ halfspace	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z})$ $\quad \mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) \leq b$ $P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a})$ $\quad \mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) > b$ where $\lambda^* \in \Re$ satisfies $\quad \mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a}) = b$ & $\lambda^* > 0$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$B_{\|.\|_1}[0, \alpha]$	$\mathbf{z}$ $\quad \|z\|_1 \leq \alpha$ $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z})$ $\quad \|z\|_1 > \alpha$ where $\lambda^* > 0$, & $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z}) = \alpha$	$\alpha > 0$

Easy to Project Sets $\mathcal{C}$ (with closed form solutions)

- Solution set of a linear system $\mathcal{C} = \{\mathbf{x} \in \Re^n : A^T\mathbf{x} = \mathbf{b}\}$
- Affine images $\mathcal{C} = \{A\mathbf{x} + \mathbf{b} : \mathbf{x} \in \Re^n\}$
- Nonnegative orthant $\mathcal{C} = \{\mathbf{x} \in \Re^n : \mathbf{x} \succeq 0\}$. It may be hard to project on arbitrary polyhedron.
- Norm balls $\mathcal{C} = \{\mathbf{x} \in \Re^n : \|\mathbf{x}\|_p \leq 1\}$, for $p = 1, 2, \infty$

Projected Gradient Descent for Affine Constraint Set $\mathcal{C}$

Solution set of a linear system $\mathcal{C} = \{\mathbf{x} \in \Re^n : A^T\mathbf{x} = \mathbf{b}\}$

$$\mathbf{x}_p^{(k+1)} = P_{\mathcal{C}}(\mathbf{x}_u^{(k+1)}) = \arg\min_{A^T\mathbf{z}=\mathbf{b}} \frac{1}{2}\left\|\mathbf{x}_u^{(k+1)} - \mathbf{z}\right\|_2^2$$

For $\mathbf{z}, \mathbf{x} \in \Re^n$, A as an $n \times m$ matrix, $\mathbf{b}$ is a vector of size m, consider the slightly more general problem (50) with B as an $n \times n$ matrix:

$$\begin{aligned} \min_{\mathbf{x}\in\Re^n} \quad & \frac{1}{2}(\mathbf{x}-\mathbf{z})^T B(\mathbf{x}-\mathbf{z}) \\ \text{subject to} \quad & A^T\mathbf{x} = \mathbf{b} \end{aligned} \tag{50}$$

For projected gradient descent, $B = I$

Projected Gradient Descent for Affine Constraint Set $\mathcal{C}$

Solution set of a linear system $\mathcal{C} = \{\mathbf{x} \in \Re^n : A^T\mathbf{x} = \mathbf{b}\}$

$$\mathbf{x}_p^{(k+1)} = P_{\mathcal{C}}(\mathbf{x}_u^{(k+1)}) = \arg\min_{A^T\mathbf{z}=\mathbf{b}} \frac{1}{2}\left\|\mathbf{x}_u^{(k+1)} - \mathbf{z}\right\|_2^2$$

For $\mathbf{z}, \mathbf{x} \in \Re^n$, A as an $n \times m$ matrix, $\mathbf{b}$ is a vector of size m, consider the slightly more general problem (50) with B as an $n \times n$ matrix:

$$\begin{aligned} \min_{\mathbf{x}\in\Re^n} \quad & \frac{1}{2}(\mathbf{x}-\mathbf{z})^T B(\mathbf{x}-\mathbf{z}) \\ \text{subject to} \quad & A^T\mathbf{x} = \mathbf{b} \end{aligned} \tag{50}$$

For projected gradient descent, $B = I$. Further, if $n = 2$ and $m = 1$, the minimization problem (50) amounts to finding a point $\mathbf{x}^*$ on a line $a_{11}x_1 + a_{12}x_2 = b$ that is closest to $\mathbf{z}$.

Projected Gradient Descent for Affine Constraint Set $\mathcal{C}$

- Consider minimization of the modified objective function $L(\mathbf{z},\lambda) = \frac{1}{2}(\mathbf{x}-\mathbf{z})^T B(\mathbf{x}-\mathbf{z}) + \lambda^T(A^T\mathbf{z}-\mathbf{b})$.

$$\min_{\mathbf{x}\in\Re^n,\lambda\in\Re^m} \quad \frac{1}{2}(\mathbf{x}-\mathbf{z})^T B(\mathbf{x}-\mathbf{z}) + \lambda^T(A^T\mathbf{x}-\mathbf{b}) \tag{51}$$

The function $L(\mathbf{x},\lambda)$ is called the lagrangian and involves the lagrange multiplier $\lambda\in\Re^m$.

- A sufficient condition for optimality of $L(\mathbf{x},\lambda)$ at a point $L(\mathbf{x}^*,\lambda^*)$ is that $\nabla L(\mathbf{x}^*,\lambda^*) = 0$ and $\nabla^2 L(\mathbf{x}^*,\lambda^*) \succ 0$. For this specific problem:

$$\nabla L(\mathbf{x}^*,\lambda^*) = \begin{bmatrix} B\mathbf{x}^* - \frac{1}{2}(B+B^T)\mathbf{z} + A\lambda^* \\ A^T\mathbf{x}^* - \mathbf{b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$\nabla^2 L(\mathbf{x}^*,\lambda^*) = \begin{bmatrix} B & A \\ A^T & 0 \end{bmatrix} \succ 0$$

Projected Gradient Descent for Affine Constraint Set $\mathcal{C}$

- The point $(\mathbf{x}^*, \lambda^*)$ must therefore satisfy, $A^T\mathbf{x}^* = \mathbf{b}$ and $A\lambda^* = -B\mathbf{x}^* + \frac{1}{2}(B + B^T)\mathbf{z}$.
- Recap: If B is taken to be the identity matrix, $n = 2$ and $m = 1$, the minimization problem (50) amounts to finding a point $\mathbf{x}^*$ on a line $a_{11}x_1 + a_{12}x_2 = b$ that is closest to $\mathbf{z}$.
- From geometry, the point on a line closest to $\mathbf{z}$ is the point of intersection $\mathbf{p}^*$ of a perpendicular (or least possible[10] obtuse angle) from $\mathbf{z}$ to the line. However, the solution for the minimum of (51), for these conditions coincide with $\mathbf{p}^*$ and are given by:

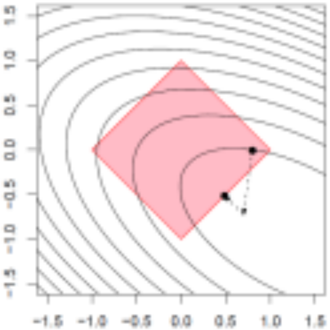
$$x_1^* = z_1 - \frac{a_{11}(a_{11}z_1 + a_{12}z_2 - b}{(a_{11})^2 + (a_{12})^2} \quad x_2^* = z_2 - \frac{a_{12}(a_{11}z_1 + a_{12}z_2 - b)}{(a_{11})^2 + (a_{12})^2}$$

  That is, for $n = 2$ and $m = 1$, the solution to (51) is the same as the solution to (50)
- For general n and m, with $\mathbf{z} \equiv \mathbf{x}_u^{(k+1)}$,

$$\mathbf{x}^* = \mathbf{x}_p^{(k+1)} = P_{\mathcal{C}}(\mathbf{x}_u^{(k+1)}) = \arg\min_{A^T\mathbf{x}=\mathbf{b}} \frac{1}{2}\left\|\mathbf{x}_u^{(k+1)} - \mathbf{x}\right\|_2^2 = \mathbf{x}_u^{(k+1)} - A(A^TA)^{-1}(A^T\mathbf{x}_u^{(k+1)} - \mathbf{b}$$

[10] See following slides for some elaboration on geometry of the projection

Projected Gradient Descent: Illustrated and Summarized



- Illustration of Projected Gradient Descent on Quadratic Objective with bounded affine (Polyhedral) constraint set
- The line joining point of projection $\mathbf{x}_p^k = P_C(\mathbf{x}_u^k)$ to $\mathbf{x}_u^k$ forms least possible obtuse angle[a] with line joining $\mathbf{x}_p^k = P_C(\mathbf{x}_u^k)$ to any point $\mathbf{y} \in C$.

[a] See following slides for some elaboration on geometry of the projecti

Table of Orthogonal Projections: See https://archive.siam.org/books/mo25/mo25_ch6.pdf

$$P_C(\mathbf{z}) = prox_{I_C}(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}}\ \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + I_C(\mathbf{x}) = \underset{\mathbf{x}\in C}{\text{argmin}}\ \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2$$

Set $C =$	For $t = 1$, $P_C(\mathbf{z}) =$		Assumptions
$\Re^n_+$	$[\mathbf{z}]_+$		
$\text{Box}[\mathbf{l},\mathbf{u}]$	$P_C(\mathbf{z})_i = min\{max\{z_i, l_i\}, u_i\}$		$l_i \leq u_i$
$\text{Ball}[\mathbf{c}, r]$	$\mathbf{c} + \frac{r}{max\{\|\mathbf{z}-\mathbf{c}\|_2, r\}}(\mathbf{z}-\mathbf{c})$		$\|.\|_2$ ball, centre $\mathbf{c} \in \Re^n$ & radius $r > 0$
$\{\mathbf{x}\|A\mathbf{x} = \mathbf{b}\}$	$\mathbf{z} - A^T(AA^T)^{-1}(A\mathbf{z}-\mathbf{b})$		$A \in \Re^{m\times n}$, $\mathbf{b} \in \Re^m$, A is full row rank
$\{\mathbf{x}\|\mathbf{a}^T\mathbf{x} \leq b\}$	$\mathbf{z} - \frac{[\mathbf{a}^T\mathbf{x}-b]_+}{\|\mathbf{a}\|^2}$		$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
Δ_n	$[\mathbf{z}-\mu^*\mathbf{e}]_+$ where $\mu^* \in \Re$ satisfies $\mathbf{e}^T[\mathbf{z}-\mu^*\mathbf{e}]_+ = 1$		
$H_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a})$ where $\mu^* \in \Re$ satisfies $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a}) = b$		$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$H^-{}_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z})$ $P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a})$ where $\lambda^* \in \Re$ satisfies	$\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) \leq b$ $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) > b$ $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a}) = b$ & $\lambda^* > 0$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$B_{\|.\|_1}[0, \alpha]$	$\mathbf{z}$ $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z})$ where $\lambda^* > 0$, & $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z}) = \alpha$	$\|z\|_1 \leq \alpha$ $\|z\|_1 > \alpha$	$\alpha > 0$

Elaboration on the Geometry of the Projected Gradient Descent

Right angle FOR Affine Set/Unbounded sets

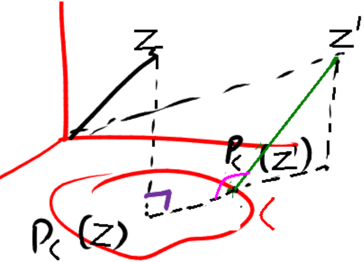
Least possible obtuse angle FOR Polyhedron/Bounded Sets

- **Claim:** If $P_{\mathcal{C}}(\mathbf{z})$ is a projection of $\mathbf{z}$, then

$$\left(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})\right)^T \left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right) \leq 0, \ \forall \mathbf{y} \in \mathcal{C}$$

- That is, the angle between $\left(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})\right)$ and $\left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right)$ is obtuse (or right-angled for the projected point), $\forall \mathbf{y} \in \mathcal{C}$

Proof of this claim is on following slides for your reading



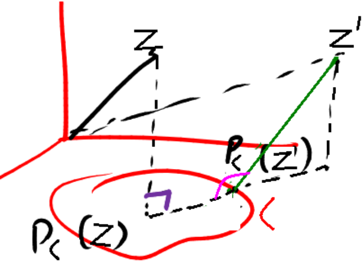
- **Claim:** If $P_{\mathcal{C}}(\mathbf{z})$ is a projection of $\mathbf{z}$, then

$$\left(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})\right)^T \left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right) \leq 0, \forall\, \mathbf{y} \in \mathcal{C}$$

- That is, the angle between $\left(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})\right)$ and $\left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right)$ is obtuse (or right-angled for the projected point), $\forall \mathbf{y} \in \mathcal{C}$

For the more general $prox_C$ operator,

$$\left(\mathbf{y} - prox_C(\mathbf{z})\right)^T \left(\mathbf{z} - prox_C(\mathbf{z})\right) \leq 0, \forall\, \mathbf{y} \qquad (52)$$



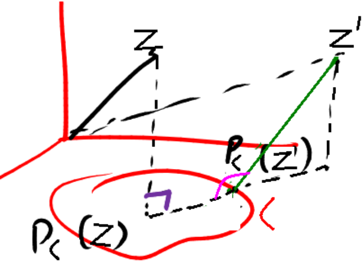
- **Claim:** If $P_{\mathcal{C}}(\mathbf{z})$ is a projection of $\mathbf{z}$, then

$$\left(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})\right)^T \left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right) \leq 0, \,\forall\, \mathbf{y} \in \mathcal{C}$$

- That is, the angle between $\left(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})\right)$ and $\left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right)$ is obtuse (or right-angled for the projected point), $\forall \mathbf{y} \in \mathcal{C}$

For the more general $prox_C$ operator,

$$\left(\mathbf{y} - prox_C(\mathbf{z})\right)^T \left(\mathbf{z} - prox_C(\mathbf{z})\right) \leq 0, \,\forall\, \mathbf{y} \qquad (52)$$



In fact, the conditions in (53), (54) and (55) can be proved to be equivalent[a] (when c is assumed to be convex)

$$\left(\mathbf{y} - \mathbf{z}^*\right)^T \left(\mathbf{z} - \mathbf{z}^*\right) \leq 0, \,\forall\, \mathbf{y} \qquad (53)$$

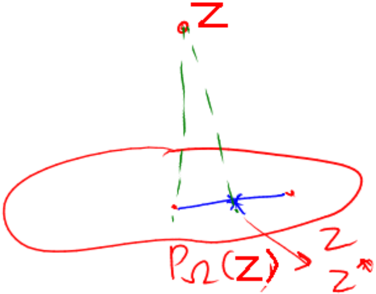
$$\mathbf{z}^* = prox_c(\mathbf{z}) \qquad (54)$$

$$\mathbf{z} - \mathbf{z}^* \in \partial c(\mathbf{z}^*) \qquad (55)$$

[a] Theorem 6.39 of https://archive.siam.org/books/mo25/mo25_ch6.pdf

Proof for $\langle \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \rangle \leq 0$

- To be more general, let us consider an inner product $\langle \mathbf{a}, \mathbf{b} \rangle$ instead of $\mathbf{a}^T\mathbf{b}$
- Let $\underline{\mathbf{y}^* = (1-\alpha)P_{\mathcal{C}}(\mathbf{z}) + \alpha\mathbf{y},}$ for some $\alpha \in (0,1)$, and $\mathbf{y} \in \mathcal{C}$
  $\implies \mathbf{y}^* = P_{\mathcal{C}}(\mathbf{z}) + \alpha(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z}))$, $\mathbf{y}^* \in \mathcal{C}$

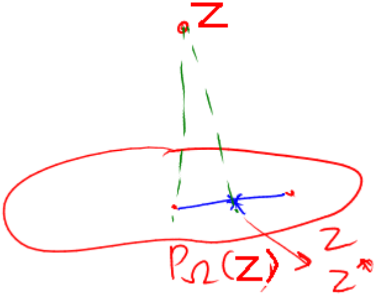


$$P_{\mathcal{C}}(\mathbf{z}) = \text{argmin}_{\mathbf{y}\in\mathcal{C}} \|\mathbf{z} - \mathbf{y}\|_2^2$$

$\Rightarrow$

Proof for $\langle \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \rangle \leq 0$

- To be more general, let us consider an inner product $\langle \mathbf{a}, \mathbf{b} \rangle$ instead of $\mathbf{a}^T\mathbf{b}$
- Let $\mathbf{y}^* = (1-\alpha)P_{\mathcal{C}}(\mathbf{z}) + \alpha\mathbf{y}$, for some $\alpha \in (0,1)$, and $\mathbf{y} \in \mathcal{C}$
  $\implies \mathbf{y}^* = P_{\mathcal{C}}(\mathbf{z}) + \alpha(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z}))$, $\mathbf{y}^* \in \mathcal{C}$



$$P_{\mathcal{C}}(\mathbf{z}) = \text{argmin}_{\mathbf{y}\in\mathcal{C}}\|\mathbf{z} - \mathbf{y}\|_2^2$$
$$\Rightarrow \underline{\left\|\mathbf{z} - P_{\mathcal{C}}(\mathbf{z})\right\|^2 \leq \|\mathbf{z} - \mathbf{y}^*\|^2}$$

Proof for $\left\langle \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \right\rangle \leq 0$

$$
\begin{aligned}
&\|\mathbf{z} - \mathbf{y}^*\|^2 \\
&= \left\| \mathbf{z} - \left(P_{\mathcal{C}}(\mathbf{z}) + \alpha(\mathbf{y} - P_{\mathcal{C}}(\mathbf{z})) \right) \right\|^2 \\
&= \left\| \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \right\|^2 + \alpha^2 \left\| \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}) \right\|^2 - 2\alpha \left\langle \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}) \right\rangle \\
&\geq \left\| \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \right\|^2
\end{aligned}
$$

$$
\implies \left\langle \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}) \right\rangle \leq \frac{\alpha}{2} \left\| \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}) \right\|^2, \forall \alpha \in (0,1)
$$

- Thus, the LHS can either be 0 or a negative value. Any positive value of the LHS will lead to a contradiction for some small $\alpha \to 0$
- Hence, we proved that $\left\langle \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \right\rangle \leq 0$

Proof for $\langle \mathbf{y} - P_{\mathcal{C}}(\mathbf{z}), \mathbf{z} - P_{\mathcal{C}}(\mathbf{z}) \rangle \leq 0$

- We can also prove that if $\langle \mathbf{z} - \mathbf{z}^*, \mathbf{y} - \mathbf{z}^* \rangle \leq 0$, $\forall \mathbf{y} \in \mathcal{C}$ s.t. $\mathbf{y} \neq \mathbf{z}^*$, and $\mathbf{z}^* \in \mathcal{C}$, then

$$\mathbf{z}^* = P_{\mathcal{C}}(\mathbf{z}) = \operatorname*{argmin}_{\bar{\mathbf{y}} \in \mathcal{C}} \|\mathbf{z} - \bar{\mathbf{y}}\|_2^2$$

- Consider $\|\mathbf{z} - \mathbf{y}\|^2 - \|\mathbf{z} - \mathbf{z}^*\|^2$
  $= \left\|\mathbf{z} - \mathbf{z}^* + (\mathbf{z}^* - \mathbf{y})\right\|^2 - \|\mathbf{z} - \mathbf{z}^*\|^2$
  $= \|\mathbf{z} - \mathbf{z}^*\|^2 + \|\mathbf{y} - \mathbf{z}^*\|^2 - 2\langle \mathbf{z} - \mathbf{z}^*, \mathbf{y} - \mathbf{z}^* \rangle - \|\mathbf{z} - \mathbf{z}^*\|^2$
  $= \|\mathbf{y} - \mathbf{z}^*\|^2 - 2\langle \mathbf{z} - \mathbf{z}^*, \mathbf{y} - \mathbf{z}^* \rangle$
  > 0
- $\implies \|\mathbf{z} - \mathbf{y}\|^2 > \|\mathbf{z} - \mathbf{z}^*\|^2$, $\forall \mathbf{y} \in \mathcal{C}$ s.t. $\mathbf{y} \neq \mathbf{z}^*$
- This proves that $\mathbf{z}^* = P_{\mathcal{C}}(\mathbf{z})$

Case 1: Projected (Gradient) Descent

- We can find $\Delta\mathbf{x}$ as the change in $\mathbf{x}$ along some steepest descent direction of f without constraints
- Thus, let $\mathbf{x}_u^{k+1} = \mathbf{z}^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}$ iterate reduces $f(\mathbf{x})$ without constraints
- To find the proximal update when $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$, we

Case 1: Projected (Gradient) Descent

- We can find $\Delta\mathbf{x}$ as the change in $\mathbf{x}$ along some steepest descent direction of f without constraints
- Thus, let $\mathbf{x}_u^{k+1} = \mathbf{z}^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}$ iterate reduces $f(\mathbf{x})$ without constraints
- To find the proximal update when $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$, we project $\mathbf{x}_u^{k+1}$ onto $\mathcal{C}$ to get the projected point $\mathbf{x}_p^{k+1}$ by solving:

Case 1: Projected (Gradient) Descent

- We can find $\Delta\mathbf{x}$ as the change in $\mathbf{x}$ along some steepest descent direction of f without constraints
- Thus, let $\mathbf{x}_u^{k+1} = \mathbf{z}^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}$ iterate reduces $f(\mathbf{x})$ without constraints
- To find the proximal update when $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$, we project $\mathbf{x}_u^{k+1}$ onto $\mathcal{C}$ to get the projected point $\mathbf{x}_p^{k+1}$ by solving:

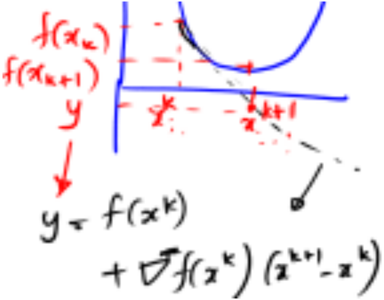
$$\mathbf{x}_p^{(k+1)} = P_{\mathcal{C}}\left(\mathbf{z}^{k+1}\right) = \operatorname*{argmin}_{\mathbf{x}} \left\|\mathbf{z}^{(k+1)} - \mathbf{x}\right\|_2^2 + I_{\mathcal{C}}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{C}} \left\|\mathbf{z}^{(k+1)} - \mathbf{x}\right\|_2^2 = prox_{I_{\mathcal{C}}}(\mathbf{z})$$

- Thus, the projected point $\mathbf{x}_p^{(k+1)}$ is the point in $\mathcal{C}$ that is the closest to the unbounded optimal point $\mathbf{x}_u^{(k+1)}$ if $\mathcal{C}$ is a non-empty closed convex set

Recall a necessary condition for descent direction: Dot product of update step with gradient <= 0

Recall: Descent direction for a convex function

- For a descent in a convex function f, we must have
  $f(\mathbf{x}^{k+1}) \geq$ Value at $\mathbf{x}^{k+1}$ obtained by linear interpolation from $\mathbf{x}^k$



- *ie.* $f(\mathbf{x}^{k+1}) \geq f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k)$
- Thus, for $\Delta\mathbf{x}^k$ to be a descent direction, it is necessary that
  $\underline{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \leq 0}$
  (where $\Delta\mathbf{x}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$)

Question: Descent Direction and Projected Gradient Descent

- We want that the point obtained after the projection of $\mathbf{x}_u^{k+1}$ be a descent from $\mathbf{x}_p^k$ for the function f

$$\underline{\nabla f(\mathbf{x}^k) \cdot \Delta \mathbf{x}_p \leq 0}$$

(where $\Delta \mathbf{x}_p^{(k+1)} = P_{\mathcal{C}}(\mathbf{x}_u^{k+1}) - \mathbf{x}_p^k = \mathbf{x}_p^{(k+1)} - \mathbf{x}_p^k$)

- Are we guaranteed this?

Algorithm: Projected Gradient Descent

```
Find a starting point $\mathbf{x}_p^0 \in \mathcal{C}$.
Set $k = 1$
repeat
    1. Choose a step size $t^k \propto 1/\sqrt{k}$.
    2. Set $\mathbf{x}_u^k = \mathbf{x}_p^{k-1} - t^k \nabla f(\mathbf{x}_p^{k-1})$.
    3. Set $\mathbf{x}_p^k = \underset{\mathbf{z} \in \mathcal{C}}{\operatorname{argmin}} \left\| \mathbf{x}_u^k - \mathbf{z} \right\|_2^2$.
    4. Set $k = k + 1$.
until stopping criterion (such as $||\mathbf{x}_p^k - \mathbf{x}_p^{k-1}|| \leq \epsilon$ or $f(\mathbf{x}_p^k) > f(\mathbf{x}_p^{k-1})$) is satisfied[a]
```

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

Figure 13: The projected gradient descent algorithm.

Option 1: Generalized Gradient Descent

- Recall

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

1. Gradient Descent ⇒

[11] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Option 1: Generalized Gradient Descent

- Recall

$$\mathit{prox}_c(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + c(\mathbf{x})$$

1. Gradient Descent $\Rightarrow c(\mathbf{x}) = 0$
2. Projected Gradient Descent $\Rightarrow$

[11] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Option 1: Generalized Gradient Descent

- Recall

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

1. Gradient Descent ⇒ $c(\mathbf{x}) = 0$
2. Projected Gradient Descent ⇒ $c(\mathbf{x}) = \sum_i I_{C_i}(\mathbf{x})$
3. Proximal Minimization ⇒ $f(\mathbf{x}) = 0$

We will discuss these specific cases after a short discussion on convergence

[11] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Option 1: Generalized Gradient Descent

- Recall

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

1. Gradient Descent ⇒ $c(\mathbf{x}) = 0$
2. Projected Gradient Descent ⇒ $c(\mathbf{x}) = \sum_i I_{C_i}(\mathbf{x})$
3. Proximal Minimization ⇒ $f(\mathbf{x}) = 0$

We will discuss these specific cases after a short discussion on convergence

- Convergence: If $f(\mathbf{x})$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(\mathbf{x})$ **is convex and** $prox_c(\mathrm{z})$ **can be solved exactly**[11] then

[11] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Option 1: Generalized Gradient Descent

- Recall

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

1. Gradient Descent ⇒ $c(\mathbf{x}) = 0$
2. Projected Gradient Descent ⇒ $c(\mathbf{x}) = \sum_i I_{C_i}(\mathbf{x})$
3. Proximal Minimization ⇒ $f(\mathbf{x}) = 0$

We will discuss these specific cases after a short discussion on convergence

- Convergence: If $f(\mathbf{x})$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(\mathbf{x})$ **is convex and** $prox_c(\mathbf{z})$ **can be solved exactly**[11] then convergence result (and proof) is similar to that for gradient descent

Recall sublinear
rate of convergence

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

[11] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization[Bertsekas and Tseng '94]

Summary results for Generalized Gradient Descent: (Details at https://archive.siam.org/books/mo25/mo25_ch10.pdf

For one of three backtracking procedures **B1, B2** and **B3**

- With no convexity assumption: Convergence can be proved using **B1** (Theorem 10.15)
- With convexity of f: $O(1/k)$ rate of convergence using **B2** (Theorem 10.21)
- With strong convexity of f: Linear rate of convergence using **B2** (Theorem 10.29)
- Assuming upper bound on norm of gradient ∇f (that is, Lipschitz continuitu of f), we get weaker $O(1/\sqrt{k})$ convergence rate (Extra optional slides at the end)

! Recommended optional reading

Algorithms: Generalized Gradient Descent

Goal: $\mathbf{x}^* = \text{argmin}_{\mathbf{x}}\, f(\mathbf{x}) + c(\mathbf{x})$

```
Find a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$
repeat
    1. Set $\Delta\mathbf{x}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$.
    2. Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search to obtain $\widehat{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$
    3. Obtain $\mathbf{x}^{(k+1)} = prox_c\left(\widehat{\mathbf{x}}^{(k+1)}\right)$.
    4. Set $k = k + 1$.
until stopping criterion (such as $||\mathbf{x}^{(k+1)} - \mathbf{x}^k||_2 \leq \epsilon$) is satisfied
```

The steepest descent method can be thought of as changing the coordinate system in a particular way and then applying the gradient descent method in the changed coordinate system.

Convergence Rate: Generalized Gradient Descent vs. Subgradient Descent

- Recap: For Subgraident Descent: The subgradient method has convergence rate $O(1/\sqrt{k})$; to get $f(\mathbf{x}_{best}^{(k)}) - f(\mathbf{x}^*) \leq \epsilon$, we need $O(1/\sqrt{\epsilon^2})$ iterations. This is actually the best we can do; e.g., we can't do better than $O(1/\sqrt{k})$.

Convergence Rate: Generalized Gradient Descent vs. Subgradient Descent

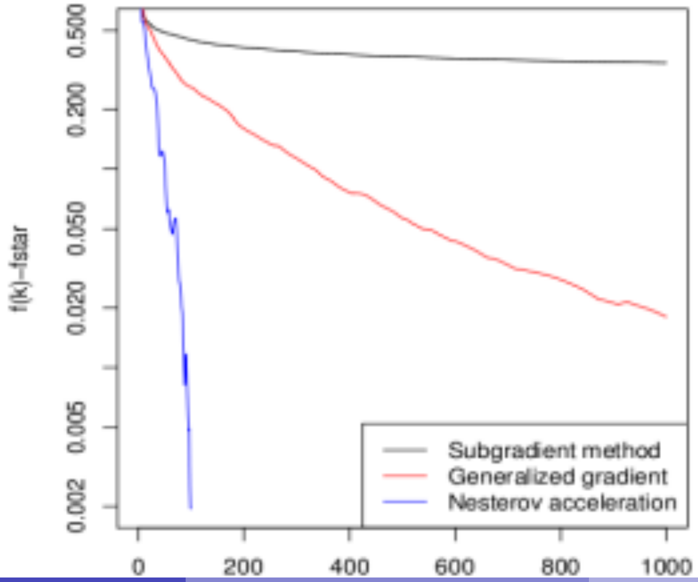
- Recap: For Subgraident Descent: The subgradient method has convergence rate $O(1/\sqrt{k})$; to get $f(\mathbf{x}_{best}^{(k)}) - f(\mathbf{x}^*) \leq \epsilon$, we need $O(1/\sqrt{\epsilon^2})$ iterations. This is actually the best we can do; e.g., we can't do better than $O(1/\sqrt{k})$.
- For generalized Gradient Descent: If $f(x)$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(x)$ is convex and $prox_c(x)$ can be solved exactly then convergence result (and proof) is similar to that for gradient descent

We appreciate that we do better than subgradient descent by making use of differentiability of f(x) part of the objective

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

Better convergence ($O(1/k)$) because of assuming (i) Differentiability of $f(\mathrm{x})$ and (ii) Lipschitz continuity of $\nabla f(\mathrm{x})$.
Can we do even better without strong convexity (which is not possible for $c(\mathrm{x})$)?

(Nesterov) Accelerated Generalized Gradient Descent



(Nesterov) Accelerated Generalized Gradient Descent

The problem is:

$$\min_{x \in \mathbb{R}^n} f(\mathbf{x}) + c(\mathbf{x})$$

where $f(\mathbf{x})$ is convex and differentiable, $c(\mathbf{x})$ is convex and not necessarily differentiable.

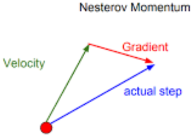
- Initialize $\mathbf{x}_u^{(0)} \in \mathbb{R}^n$
- repeat for $k = 1, 2, 3, \ldots$

$$\mathbf{y} = \mathbf{x}^{(k-1)} + \frac{k-2}{k+1}(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)})$$

$$\mathbf{x}^{(k)} = \text{prox}_{t^k}(\mathbf{y} - t^k \nabla f(\mathbf{y}))$$

Steps for implementing

Or Equivalently,



$$\boxed{\mathbf{y} = (1-\theta_k)\mathbf{x}^{(k-1)} + \theta_k \mathbf{x}_u^{(k-1)}}$$

y is update after adding velocity

$$\boxed{\mathbf{x}^k = \text{prox}_{t^k}(\mathbf{y} - t^k \nabla f(\mathbf{y}))}$$

compute prox on update after adding velocity

$$\boxed{\mathbf{x}_u^{(k)} = \mathbf{x}^{(k-1)} + \frac{1}{\theta_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})}$$

where $\theta_k = 2/(k+1)$.

Algorithm: (Nesterov) Accelerated Generalized Gradient Descent

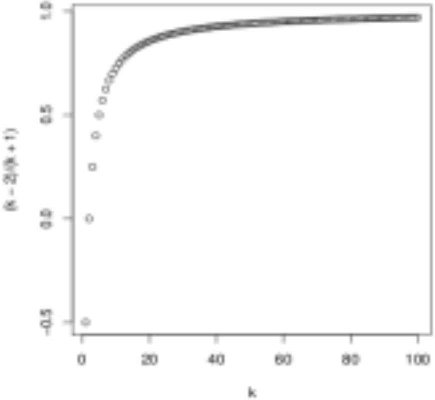
```
**Initialize** $\mathbf{x}_u^{(0)}, \mathbf{x}^{(0)} \in \Re^n$
Initialize $k = 1$
**repeat**
    1. $\theta_k = 2/(k+1)$
    2. $\mathbf{y} = \underline{(1-\theta_k)}\mathbf{x}^{(k-1)} + \underline{\theta_k \mathbf{x}_u}^{(k-1)}$.
    3. Choose a step size $t^k > 0$ using exact or backtracking ray search.
    4. $\mathbf{x}^k = \text{prox}_{t^k}(\mathbf{y} - t^k \nabla f(\mathbf{y}))$
    5. $\mathbf{x}_u^{(k)} = \mathbf{x}^{(k-1)} + \frac{1}{\theta_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})$
    6. 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 15: The gradient descent algorithm.

(Nesterov) Accelerated Generalized Gradient Descent

1. First step $k = 1$ is just usual generalized gradient update: $\mathbf{x}^{(1)} = \text{prox}_{t^1}(\mathbf{x}^{(0)} - t^1 \nabla f(\mathbf{x}^{(0)}))$
2. Thereafter, the method carries some "momentum" from previous iterations
3. $c(\mathbf{x}) = 0$ gives accelerated gradient method
4. The method accelerates more towards the end of iterations



(Nesterov) Accelerated Generalized Gradient Descent

Examples showing the performance of accelerated gradient descent compared with usual gradient descent.

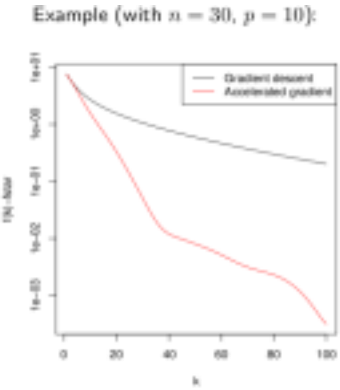

Example (with $n = 30$, $p = 10$):

Figure 17: Example 1: Performance of accelerated gradient descent compared with usual gradient descent

(Nesterov) Accelerated Generalized Gradient Descent: Convergence

Minimize $f(\mathbf{x}) = f(\mathbf{x}) + c(\mathbf{x})$ assuming that:
f is convex, differentiable, ∇f is Lipschitz with constant $L > 0$, and
c is convex, the prox function can be evaluated.

Theorem

Accelerated generalized gradient method with fixed step size $t \leq 1/L$ satisfies:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \frac{2||x^{(0)} - x^*||^2}{\underline{t(k+1)^2}}$$

Accelerated generalized gradient method can achieve the optimal $O(1/k^2)$ rate for first-order method, or equivalently, if we want to get $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \epsilon$, we only need $O(1/\sqrt{\epsilon})$ iterations. Now we prove this theorem.

(Nesterov) Accelerated Generalized Gradient Descent: Proof

Proof:

First we bound both the convex functions $f(\mathbf{x}^k)$ and $c(\mathbf{x}^k)$.

- Since $t \leq 1/L$ and ∇f is Lipschitz with constant $L > 0$, we have

$$f(\mathbf{x}^k) \leq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}^k - \mathbf{y}) + \frac{L}{2}||\mathbf{x}^k - \mathbf{y}||^2 \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}^k - \mathbf{y}) + \frac{1}{2t}||\mathbf{x}^k - \mathbf{y}||^2 \quad (56)$$

- In $\mathbf{x}^k = \text{prox}_t(\mathbf{y} - t\nabla f(\mathbf{y}))$, let $\mathbf{h} = \mathbf{x}^k$ and $\mathbf{w} = \mathbf{y} - t\nabla f(\mathbf{y})$. Then

$$\mathbf{h} = \text{prox}_t(\mathbf{w}) = \arg\min_{\mathbf{h}} \frac{1}{2t}||\mathbf{w} - \mathbf{h}||^2 + c(\mathbf{h})$$

- For this, we must have

$$0 \in \partial(\frac{1}{2t}||\mathbf{w} - \mathbf{h}||^2 + c(\mathbf{h})) = -\frac{1}{t}(\mathbf{w} - \mathbf{h}) + \partial c(\mathbf{h}) \quad \Rightarrow \quad -\frac{1}{t}(\mathbf{w} - \mathbf{h}) \in \partial c(\mathbf{h})$$

- According to the definition of subgradient, we have for all $\mathbf{z}$,

(Nesterov) Accelerated Generalized Gradient Descent: Proof

Proof:

First we bound both the convex functions $f(\mathbf{x}^k)$ and $c(\mathbf{x}^k)$.

- Since $t \leq 1/L$ and ∇f is Lipschitz with constant $L > 0$, we have

$$f(\mathbf{x}^k) \leq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}^k - \mathbf{y}) + \frac{L}{2}||\mathbf{x}^k - \mathbf{y}||^2 \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}^k - \mathbf{y}) + \frac{1}{2t}||\mathbf{x}^k - \mathbf{y}||^2 \quad (56)$$

- In $\mathbf{x}^k = \text{prox}_t(\mathbf{y} - t\nabla f(\mathbf{y}))$, let $\mathbf{h} = \mathbf{x}^k$ and $\mathbf{w} = \mathbf{y} - t\nabla f(\mathbf{y})$. Then

$$\mathbf{h} = \text{prox}_t(\mathbf{w}) = \arg\min_{\mathbf{h}} \frac{1}{2t}||\mathbf{w} - \mathbf{h}||^2 + c(\mathbf{h})$$

- For this, we must have

$$0 \in \partial(\frac{1}{2t}||\mathbf{w} - \mathbf{h}||^2 + c(\mathbf{h})) = -\frac{1}{t}(\mathbf{w} - \mathbf{h}) + \partial c(\mathbf{h}) \quad \Rightarrow \quad -\frac{1}{t}(\mathbf{w} - \mathbf{h}) \in \partial c(\mathbf{h})$$

- According to the definition of subgradient, we have for all $\mathbf{z}$,

$$c(\mathbf{z}) \geq c(\mathbf{h}) - \frac{1}{t}(\mathbf{h} - \mathbf{w})^T(\mathbf{z} - \mathbf{h}) \quad \Rightarrow \quad c(\mathbf{h}) \leq c(\mathbf{z}) + \frac{1}{t}(\mathbf{h} - \mathbf{w})^T(\mathbf{z} - \mathbf{h})$$

  for all $\mathbf{z}, \mathbf{w}$ and $\mathbf{h} = \text{prox}_t(\mathbf{w})$.

(Nesterov) Accelerated Generalized Gradient Descent: Proof (contd.)

Substituting back for both $\mathbf{h}$ and $\mathbf{w}$ in the above inequality we get for all $\mathbf{z}$,

$$c(\mathbf{x}^k) \leq c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y} + t\nabla f(\mathbf{y}))^T(\mathbf{z} - \mathbf{x}^k) = c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) + \nabla f(\mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) \quad (57)$$

Adding inequalities (56) and (57) we get for all $\mathbf{z}$,

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

Since f is convex,

(Nesterov) Accelerated Generalized Gradient Descent: Proof (contd.)

Substituting back for both $\mathbf{h}$ and $\mathbf{w}$ in the above inequality we get for all $\mathbf{z}$,

$$c(\mathbf{x}^k) \leq c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y} + t\nabla f(\mathbf{y}))^T(\mathbf{z} - \mathbf{x}^k) = c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) + \nabla f(\mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) \quad (57)$$

Adding inequalities (56) and (57) we get for all $\mathbf{z}$,

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

Since f is convex, using $f(\mathbf{z}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{z} - \mathbf{y})$, we further get

$$f(\mathbf{x}^k) \leq f(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) + \frac{1}{2t}||\mathbf{x}^k - \mathbf{y}||^2$$

Now take $\mathbf{z} = \mathbf{x}^{(k-1)}$, multiply both sides by $(1-\theta)$ and for $\mathbf{z} = \mathbf{x}^*$ multiply both sides by θ,

$$(1-\theta)f(\mathbf{x}^k) \leq (1-\theta)f(\mathbf{x}^{(k-1)}) + \frac{1-\theta}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{x}^{(k-1)} - \mathbf{x}^k) + \frac{1-\theta}{2t}||\mathbf{x}^k - \mathbf{y}||^2$$

$$\theta f(\mathbf{x}^k) \leq \theta f(\mathbf{x}^*) + \frac{\theta}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{x}^* - \mathbf{x}^k) + \frac{\theta}{2t}||\mathbf{x}^k - \mathbf{y}||^2$$

(Nesterov) Accelerated Generalized Gradient Descent: Proof (contd.)

Adding these two inequalities together, we get

$$f(\mathbf{x}^k)-f(\mathbf{x}^*)-(1-\theta)(f(\mathbf{x}^{(k-1)})-f(\mathbf{x}^*)) \leq \frac{1}{t}(\mathbf{x}^k-\mathbf{y})^T((1-\theta)\mathbf{x}^{(k-1)}+\theta\mathbf{x}^*-\mathbf{x}^k)+\frac{1}{2t}||\mathbf{x}^k-\mathbf{y}||^2 \tag{58}$$

- Using $\mathbf{x}_u^k = \mathbf{x}^{(k-1)} + \frac{1}{\theta}(\mathbf{x}^k - \mathbf{x}^{(k-1)})$ and $\mathbf{y} = (1-\theta)\mathbf{x}^{(k-1)} + \theta\mathbf{x}_u^{(k-1)}$, we have $(1-\theta)\mathbf{x}^{(k-1)} + \theta\mathbf{x}^* - \mathbf{x}^k = \theta(\mathbf{x}^* - \mathbf{x}_u^k)$ and using this again in the second equation, $\mathbf{x}^k - \mathbf{y} = \theta(\mathbf{x}_u^k - \mathbf{x}_u^{(k-1)})$
- Substituting these equations into the RHS of inequality (58) we have

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) - (1-\theta)(f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)) \quad \leq \frac{\theta}{2t}\underline{(\mathbf{x}_u^k - \mathbf{x}_u^{(k-1)})}^T[2\theta(\mathbf{x}^* - \mathbf{x}_u^k) + \theta(\mathbf{x}_u^k - \mathbf{x}_u^{(k-1)})]$$

(Nesterov) Accelerated Generalized Gradient Descent: Proof (contd.)

Adding these two inequalities together, we get

$$f(\mathbf{x}^k)-f(\mathbf{x}^*)-(1-\theta)(f(\mathbf{x}^{(k-1)})-f(\mathbf{x}^*)) \leq \frac{1}{t}(\mathbf{x}^k-\mathbf{y})^T((1-\theta)\mathbf{x}^{(k-1)}+\theta\mathbf{x}^*-\mathbf{x}^k)+\frac{1}{2t}||\mathbf{x}^k-\mathbf{y}||^2 \tag{58}$$

- Using $\mathbf{x}_u^k = \mathbf{x}^{(k-1)} + \frac{1}{\theta}(\mathbf{x}^k - \mathbf{x}^{(k-1)})$ and $\mathbf{y} = (1-\theta)\mathbf{x}^{(k-1)} + \theta\mathbf{x}_u^{(k-1)}$, we have $(1-\theta)\mathbf{x}^{(k-1)} + \theta\mathbf{x}^* - \mathbf{x}^k = \theta(\mathbf{x}^* - \mathbf{x}_u^k)$ and using this again in the second equation, $\mathbf{x}^k - \mathbf{y} = \theta(\mathbf{x}_u^k - \mathbf{x}_u^{(k-1)})$
- Substituting these equations into the RHS of inequality (58) we have

$$f(\mathbf{x}^k)-f(\mathbf{x}^*)-(1-\theta)(f(\mathbf{x}^{(k-1)})-f(\mathbf{x}^*)) \leq \frac{\theta}{2t}\underline{(\mathbf{x}_u^k-\mathbf{x}_u^{(k-1)})}^T[2\theta(\mathbf{x}^*-\mathbf{x}_u^k)+\theta(\mathbf{x}_u^k-\mathbf{x}_u^{(k-1)})]$$

$$= \frac{\theta^2}{2t}\underline{(\mathbf{x}^*-\mathbf{x}_u^{(k-1)})-(\mathbf{x}^*-\mathbf{x}_u^{(k-1)})]}^T[(\mathbf{x}^*-\mathbf{x}_u^k)+(\mathbf{x}^*-\mathbf{x}_u^{(k-1)})]$$

$$= dfrac\theta^2 2t(||\mathbf{x}_u^{(k-1)}-\mathbf{x}^*||^2-||\mathbf{x}_u^k-\mathbf{x}^*||^2)$$

(Nesterov) Accelerated Generalized Gradient Descent: Proof (contd.)

$$\frac{t}{\theta_k^2}\left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)\right) + \frac{1}{2}||\mathbf{x}_u^{(k)} - \mathbf{x}^*||^2 \le \frac{t(1-\theta_k)}{\theta_k^2}\left(f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)\right) + \frac{1}{2}||\mathbf{x}_u^{(k-1)} - \mathbf{x}^*||^2$$

Since $\theta = 2/(k+1)$, using $\frac{1-\theta_k}{\theta_k^2} \le \frac{1}{\theta_{k-1}^2}$, we have

$$\frac{t}{\theta_k^2}\left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)\right) + \frac{1}{2}||\mathbf{x}_u^{(k)} - x^*||^2 \le \frac{t}{\theta_{k-1}^2}\left(f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)\right) + \frac{1}{2}||\mathbf{x}_u^{(k-1)} - \mathbf{x}^*||^2$$

Iterating this inequality and using $\theta_1 = 1$ we get

$$\frac{t}{\theta_k^2}\left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)\right) + \frac{1}{2}||\mathbf{x}_u^{(k)} - \mathbf{x}^*||^2 \le \frac{t(1-\theta_1)}{\theta_1^2}\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right) + \frac{1}{2}||\mathbf{x}_u^{(0)} - \mathbf{x}^*||^2 \le \frac{1}{2}||\mathbf{x}^{(0)} - \mathbf{x}^*||^2$$

Hence we conclude

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \le \frac{\theta_k^2}{2t}||\mathbf{x}^{(0)} - \mathbf{x}^*||^2 = \frac{2||\mathbf{x}^{(0)} - \mathbf{x}^*||^2}{t(k+1)^2}$$

Generalized Gradient Descent and its Special Cases

Recall

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

It's special cases are:

1. Gradient Descent: $c(\mathbf{x}) = 0$
2. Projected Gradient Descent: $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$ (Example:

Generalized Gradient Descent and its Special Cases

Recall

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

It's special cases are:

1. Gradient Descent: $c(\mathbf{x}) = 0$
2. Projected Gradient Descent: $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$ (Example: $= \sum_i \underline{I_{g_i}(\mathbf{x})}$)
3. Alternating Projection/Proximal Minimization: $f(\mathbf{x}) = 0$
4. Alternating Direction Method of Multipliers
5. Special Cases for Specific Objectives
    - LASSO: (Fast) Iterative Shrinkage Thresholding Algorithm (ISTA/FISTA)

Convergence of Projected Gradient Descent (even under weaker assumptions)

Convergence of Projected Gradient Descent: Weaker assumptions

- Recall: Assuming Lipschitz continuity on gradient ∇f and convexity of f and assuming bounded iterates and assuming convexity of $\mathcal{C}$ (and therefore of $I_{\mathcal{C}}$) we obtained $O(1/k)$ convergence rate for (Generalized and hence for) Projected Gradient Descent
- Assuming upper bound on norm of gradient ∇f (that is, Lipschitz continuitu of f), we get weaker $O(1/\sqrt{k})$ convergence rate for Projected Gradient Descent

Convergence of Projected Gradient Descent: Weaker assumptions

- Recall: Assuming Lipschitz continuity on gradient ∇f and convexity of f and assuming bounded iterates and assuming convexity of $\mathcal{C}$ (and therefore of $I_{\mathcal{C}}$) we obtained $O(1/k)$ convergence rate for (Generalized and hence for) Projected Gradient Descent
- Assuming upper bound on norm of gradient ∇f (that is, Lipschitz continuitu of f), we get weaker $O(1/\sqrt{k})$ convergence rate for Projected Gradient Descent
- **Proof:** To project $\mathbf{x}_u^{k+1} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k)$ onto the non-empty closed convex set $\mathcal{C}$ to get the projected point $\mathbf{x}_p^{k+1}$, we solve:$\mathbf{x}_p^{k+1} = P_{\mathcal{C}}(\mathbf{x}_u^{k+1}) = \text{argmin}_{\mathbf{z}\in\mathcal{C}} \left\|\mathbf{x}_u^{k+1} - \mathbf{z}\right\|_2^2$

$$\|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 = \|\mathbf{x}^* - \mathbf{x}^k\|^2 - 2t\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) + t^2|\nabla f(\mathbf{x}^k)|^2 \tag{59}$$

- If: (i) $\mathbf{d}$ is diameter of $\mathcal{C}$, *i.e.*, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\|\mathbf{x} - \mathbf{y}\| \leq \mathbf{d}$ (ii) l is upper bound on norm of gradients, *i.e.*, $\|\nabla f(\mathbf{x})\| \leq l$ and (iv) step size $t = \frac{\mathbf{d}}{l\sqrt{K}}$, then substituting for l into (59)

$$\|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 \leq \|\mathbf{x}^* - \mathbf{x}^k\|^2 - 2t\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) + t^2 l^2 \tag{60}$$

Convergence of Proj. Grad. Descent: Weaker assumptions (contd.)

- Further, based on (60)

$$2t\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) \leq \|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 + t^2 l^2 \tag{61}$$

- As per definition of convexity:

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(x^*) \leq \frac{1}{K}\sum_{k=1}^{K}\left(f(\mathbf{x}^k) - f(\mathbf{x}^*)\right) \leq \frac{1}{K}\sum_{k=1}^{K}\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) \tag{62}$$

- Substituting for $\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*)$ from (61) into (62), we get (63):

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{2tK}\sum_{k=1}^{K}\left(\|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 + t^2 l^2\right) \tag{63}$$

Convergence of Proj. Grad. Descent: Weaker assumptions (contd.)

- Expanding the summation over $\|\mathbf{x}^* - \mathbf{x}^k\|^2$, all terms get canceled except for the first and last:

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{2tK}\left(\|\mathbf{x}^* - \mathbf{x}^0\|^2 - \|\mathbf{x}^* - \mathbf{x}_u^{K+1}\|^2\right) + \frac{tl^2}{2} \tag{64}$$

- Since $\mathbf{d}$ is diameter of $\mathcal{C}$, *i.e.*, $\|\mathbf{x}^* - \mathbf{x}^0\|^2 \leq \mathbf{d}^2$ and since $-\|\mathbf{x}^* - \mathbf{x}_u^{K+1}\|^2 \leq 0$,

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{2tK}\left(d^2\right) + \frac{tl^2}{2} \leq \frac{\mathbf{d}l}{\sqrt{K}} \tag{65}$$

- Therefore, if $t = \frac{\mathbf{d}}{l\sqrt{K}}$, $f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) \leq \min_{x \in \mathcal{C}} f(\mathbf{x}) + \frac{\mathbf{d}l}{\sqrt{K}}$

Convergence of Proj. Grad. Descent: Weaker assumptions (contd.)

- To get solution that is ϵ approximate with $\epsilon = \frac{\mathbf{d}\mathbf{g}}{\sqrt{K}}$, you need number of gradient iterations that is $K = \left(\frac{\mathbf{d}g}{\epsilon}\right)^2 = O\left(\frac{1}{\epsilon}\right)^2$

Extra and Optional: Alternative Projection Method

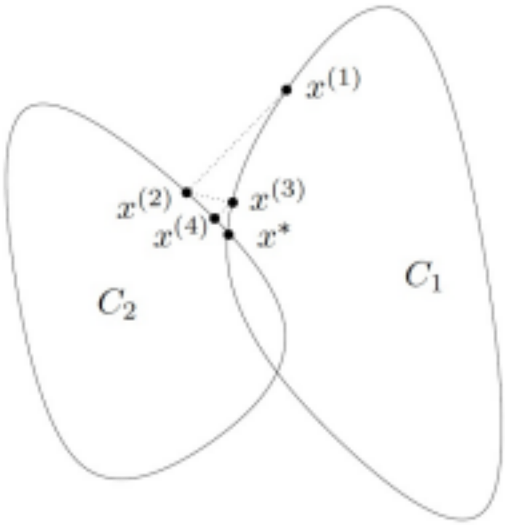


Figure 19: Alternating Projection from Boyd Notes

Example: Subgradients and Alternating Projections

Problem: Given m closed convex sets $C_1, C_2, \ldots, C_m$, we want to find $\mathbf{x}^* \in \bigcap_i^m C_i$.
First, we define

$$f(x) = \max_{i=1,\ldots,m} \text{dist}(x, C_i)$$

where

$$\text{dist}(x, C) = \min_{u \in C} \|x - u\|$$

is the closest we can get to x if we have to stay in the set C.
Also,

$$f(\mathbf{x}^*) = 0 \iff \mathbf{x}^* \in \bigcap_i^m C_i$$

Therefore, the optimization problem is to minimize

$$\min_{x \in \mathbb{R}^n} f(x)$$

which, when equal to 0 is the point we are looking for.

Example: Subgradients and Alternating Projections (contd.)

Since C is closed and convex, there is a unique point $u^* = P_C(x)$. This unique point is the projection of x onto C, and it minimizes $\|x - u\|$ over $u \in C$. We can thus write

$$\mathrm{dist}(x, C) = \|x - P_C(x)\|$$

Finding subgradient of f_i

We want to calculate the subgradient of f because if we can do so, we can apply subgradient methods and obtain an algorithm to solve our problem.

First, we consider $f_i(x)$ of C_i. It turns out that $f_i(x)$ is differentiable. For each i, if we take a point not in C_i, i.e $x \notin C_i$ and $\|x - P_C(x)\| \neq 0$, it turns out that

$$\frac{x - P_C(x)}{\|x - P_C(x)\|} \tag{66}$$

is a subgradient of $f_i(x)$. We obtain this by just taking the projected point and finding the gradient without the chain rule.

Show that (123) is a subgradient of f_i at x.

Example: Subgradients and Alternating Projections (contd.)

Finding subgradients of f:

Using a rule we learnt from earlier on in the course, if

$$f(x) = \max_{i=1,\ldots,m} f_i(x)$$

then,

$$\partial f(x) = \operatorname{conv}\left(\bigcup_{j: f_j(x) = f(x)} \partial f_j(x) \right)$$

What this means is that the subgradient of $f(x)$ is equal to the convex hull of the union of all maximal $f_j(x)$'s, and take the respective subdifferentials.

If $f_i(x) = f(x) \neq 0$ (when it is 0, we are done), then

$$\frac{x - P_C(x)}{\|x - P_C(x)\|} \in \partial f(x)$$

This gives us a prescription for finding the subgradients.

Example: Subgradients and Alternating Projections

Subgradient descent:

We will use a particular stepsize, known as the Polyak stepsize, because this particular choice will give us a famous algorithm that is a special case of the subgradient method. For the purpose of illustration, the Polyak stepsize is

$$t_k = f(\mathbf{x}^{(k-1)})$$

and the subgradient descent update rule is

$$\begin{aligned}
\mathbf{x}^{(k)} &= \mathbf{x}^{(k-1)} - t_k \partial f(\mathbf{x}^{(k-1)}) \\
&= \mathbf{x}^{(k-1)} - f(\mathbf{x}^{(k-1)}) \frac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|} \text{ where } \mathbf{x}^{(k-1)} \text{ is farthest from } C_i \\
&= \mathbf{x}^{(k-1)} - \mathbf{x}^{(k-1)} + P_{C_i}(x) \\
&= P_{C_i}(x)
\end{aligned}$$

So the update rule is just to take $\mathbf{x}^{(k-1)}$ and project it to the set it is farthest from.

This is also known as the alternating projections algorithm. By using the subgradient method, we can now use what we know about subgradients to say things about the *alternating*

Extra and Optional: Nesterov's Theorem

Theorem

Nesterov's Theorem: For any $k \leq n - 1$ and starting point $\mathbf{x}^{(0)}$, there is a function in the problem class such that any nonsmooth first-order method satisfies

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \geq \frac{RG}{2(1 + \sqrt{k+1})}$$

Proof.

Let $k = n - 1$ and $\mathbf{x}^{(0)} = 0$.

$$f(x) = \max_{i=1\ldots n} x_i + \frac{1}{2}||x||^2$$

The optimal $\mathbf{x}^*$ here $= (-1/n, \ldots, -1/n)$, with the optimal function value $f(\mathbf{x}^*) = -\frac{1}{2n}$. If $R = \frac{1}{\sqrt{n}}$, then f is Lipschitz with $G = 1 + \frac{1}{\sqrt{n}}$.

Claim: At any iteration i from 1 to n, all of the elements of x from x_{i+1} to x_n are 0. To show this, let us assume we have some oracle that gives us $g = e_j + x$, where j is the smallest index of x's maximum value x_j, e_j is the basis vector.