Option 1: Generalized Gradient Descent

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

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

- Illustration on Lasso: $\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}} ||A\mathbf{x} - \mathbf{y}||^2 + \|\mathbf{x}\|_1$. You can successively use $\mathbf{z} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k)$.
-
-
-
-

Illustration on Lasso

Iterative Soft Thresholding Algorithm for Solving Lasso

Proximal Subgradient Descent for Lasso

- Let $\varepsilon(\mathbf{w}) = \|\phi\mathbf{w} - \mathbf{y}\|_2^2$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{w}^{(0)}$
  - Let $\widehat{\mathbf{w}}^{(\mathbf{k+1})}$ be a next gradient descent iterate for $\varepsilon(\mathbf{w}^k)$
  - Compute $\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\textbf{argmin}}||\mathbf{w} - \widehat{\mathbf{w}}^{(\mathbf{k+1})}||_\mathbf{2}^\mathbf{2} + \lambda\mathbf{t}||\mathbf{w}||_\mathbf{1}$ 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. ...
    3. ...
  - Set $k = k + 1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{w}^k$ w.r.t $\mathbf{w}^{(k-1)}$)

Iterative Soft Thresholding Algorithm (Proximal Subgradient Descent) for Lasso

- Let $\varepsilon(\mathbf{w}) = \|\phi\mathbf{w} - \mathbf{y}\|_2^2$
- **Iterative Soft Thresholding Algorithm:**
  Initialization: Find starting point $\mathbf{w}^{(0)}$
  - Let $\widehat{\mathbf{w}}^{(\mathbf{k+1})}$ be a next iterate for $\varepsilon(\mathbf{w}^k)$ computed using using any (gradient) descent algorithm
  - Compute $\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{argmin}||\mathbf{w} - \widehat{\mathbf{w}}^{(\mathbf{k+1})}||_\mathbf{2}^\mathbf{2} + \lambda\mathbf{t}||\mathbf{w}||_\mathbf{1}$ by:
    1. If $\widehat{w}_i^{(k+1)} > \lambda t/2$, then $w_i^{(k+1)} = -\lambda t/2 + \widehat{w}_i^{(k+1)}$
    2. If $\widehat{w}_i^{(k+1)} < -\lambda t/2$, then $w_i^{(k+1)} = \lambda t/2 + \widehat{w}_i^{(k+1)}$
    3. 0 otherwise.

  Basically we translated inequalities for w into inequalities for \hat{w}

  - Set $k = k + 1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{w}^k$ w.r.t $\mathbf{w}^{(k-1)}$)

Option 1: Generalized Gradient Descent

- Recall

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

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

[7] 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

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

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** $prox_t(\mathbf{z})$ **can be solved exactly**[7] then

[7] 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

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

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** $prox_t(\mathbf{z})$ **can be solved exactly**[7] then convergence result (and proof) is similar to that for gradient descent

Just use a convenient step size t^k = 1/L

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

[7] 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]

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 $O(1/\sqrt{k})$.

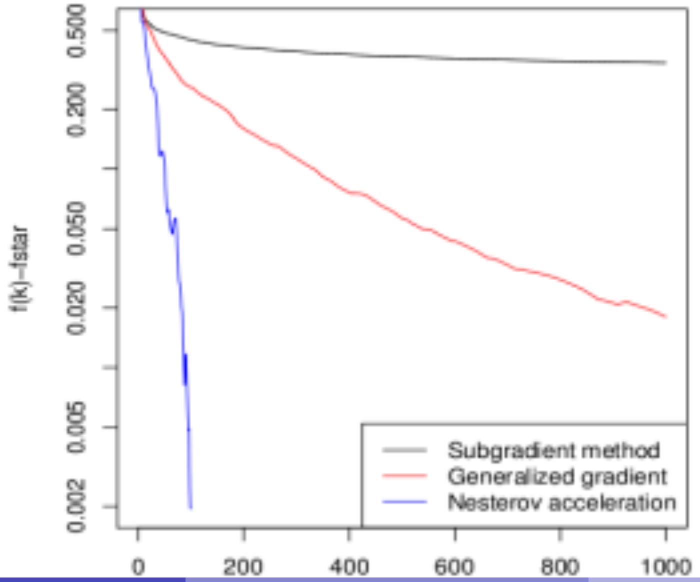
Convergence Rate: Generalized Gradient Descent vs. Subgradient Descent

- 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 $prox_t(x)$ can be solved exactly then convergence result (and proof) is similar to that for gradient descent

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

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 $k = 1, 2, 3, \ldots$

y has replaced your gradient descent update

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

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

Or Equivalently,

real iterate at k-1

unrestricted iterate at k-1

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

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

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

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

Algorithm: (Nesterov) Accelerated Generalized Gradient Descent

Convergence of O(1/k^2)

Initialize $\mathbf{x}_u^{(0)}, \mathbf{x}^{(0)} \in \Re^n$

Initialize $k = 1$

repeat

1. $\theta_k = 2/(k+1)$
2. $\mathbf{y} = (1-\theta_k)\mathbf{x}^{(k-1)} + \theta_k \mathbf{x}_u^{(k-1)}$.
3. Choose a step size $t^k > 0$ using exact or backtracking ray search. often t^k = O(1/k
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 $||\mathbf{x}^k - \mathbf{x}^{k-1}|| \le \epsilon$ or $f(\mathbf{x}^k) > f(\mathbf{x}^{k-1})$) is satisfied[a]

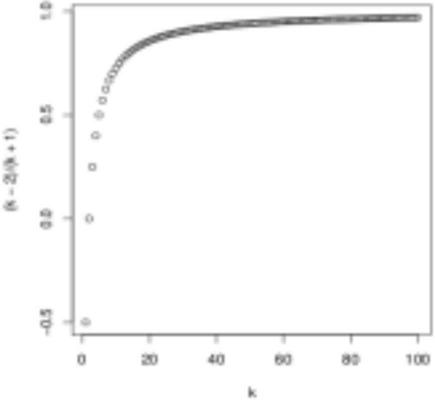
[a] Better criteria can be found using Lagrange duality theory, etc.

Figure 11: The gradient descent algorithm.

(Nesterov) Accelerated Generalized Gradient Descent

initially no momentum

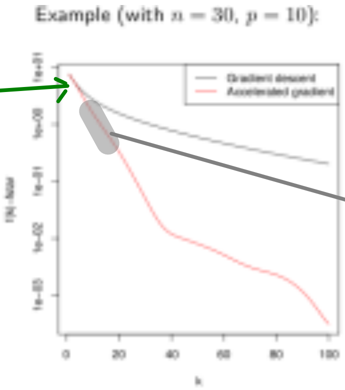
1. First step $k = 1$ is just usual generalized gradient update: $\mathbf{x}^{(1)} = \text{prox}_{t^1}(\mathbf{x}^{(0)} - t^1 \nabla f(\mathbf{x}^{(0)}))$
2. Thereafter, the method carries some "momentum" from previous iterations
3. $c(\mathbf{x}) = 0$ gives accelerated gradient method
4. The method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

Initial behaviours are similar for the two



Momentum helps you accelerate only after some time

Figure 13: 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:

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

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

$$f(\mathbf{x}^k) \leq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}^k - \mathbf{y}) + \frac{L}{2}||\mathbf{x}^k - \mathbf{y}||^2 \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}^k - \mathbf{y}) + \frac{1}{2t}||\mathbf{x}^k - \mathbf{y}||^2 \quad (48)$$

- 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

$$f(\mathbf{x}^k) \leq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}^k - \mathbf{y}) + \frac{L}{2}||\mathbf{x}^k - \mathbf{y}||^2 \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}^k - \mathbf{y}) + \frac{1}{2t}||\mathbf{x}^k - \mathbf{y}||^2 \quad (48)$$

- 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 $\mathbf{z}$,

$$c(\mathbf{x}^k) \leq c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y} + t\nabla f(\mathbf{y}))^T(\mathbf{z} - \mathbf{x}^k) = c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) + \nabla f(\mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) \quad (49)$$

Adding inequalities (48) and (49) 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 $\mathbf{z}$,

$$c(\mathbf{x}^k) \leq c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y} + t\nabla f(\mathbf{y}))^T(\mathbf{z} - \mathbf{x}^k) = c(\mathbf{z}) + \frac{1}{t}(\mathbf{x}^k - \mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) + \nabla f(\mathbf{y})^T(\mathbf{z} - \mathbf{x}^k) \quad (49)$$

Adding inequalities (48) and (49) 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

$$f(\mathbf{x}^k)-f(\mathbf{x}^*)-(1-\theta)(f(\mathbf{x}^{(k-1)})-f(\mathbf{x}^*)) \leq \frac{1}{t}(\mathbf{x}^k-\mathbf{y})^T((1-\theta)\mathbf{x}^{(k-1)}+\theta\mathbf{x}^*-\mathbf{x}^k)+\frac{1}{2t}||\mathbf{x}^k-\mathbf{y}||^2 \tag{50}$$

- 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 (50) we have

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

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

Adding these two inequalities together, we get

$$f(\mathbf{x}^k)-f(\mathbf{x}^*)-(1-\theta)(f(\mathbf{x}^{(k-1)})-f(\mathbf{x}^*)) \leq \frac{1}{t}(\mathbf{x}^k-\mathbf{y})^T((1-\theta)\mathbf{x}^{(k-1)}+\theta\mathbf{x}^*-\mathbf{x}^k)+\frac{1}{2t}||\mathbf{x}^k-\mathbf{y}||^2 \tag{50}$$

- 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 (50) 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.)

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

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

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

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

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

Hence we conclude

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

Homework:
Understand and appreciate importance of choices on \theta_k etc

Generalized Gradient Descent and its Special Cases

Recall

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

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: sum of indicators on constraints g_i(x) <= 0)

Generalized Gradient Descent and its Special Cases

Recall

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

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)

Accelerated ISTA ==> FISTA

Case 1: Projection Methods

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 $\mathbf{x}_u^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}$ be the working set that reduces $f(\mathbf{x})$ without constraints (unbounded)
- To find the constrained working set, 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

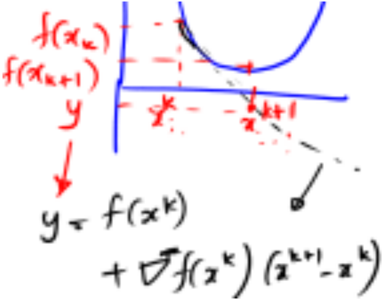
- We can find $\Delta \mathbf{x}$ as the change in $\mathbf{x}$ along some steepest descent direction of f without constraints
- Thus, let $\mathbf{x}_u^{k+1} = \mathbf{x}^k + \Delta \mathbf{x}$ be the working set that reduces $f(\mathbf{x})$ without constraints (unbounded)
- To find the constrained working set, we project $\mathbf{x}_u^{k+1}$ onto $\mathcal{C}$ to get the projected point $\mathbf{x}_p^{k+1}$ by solving:

$$\underline{\mathbf{x}_p^{(k+1)}} = \underline{P_{\mathcal{C}}(\mathbf{x}_u^{(k+1)})} = \text{argmin} \left\| \mathbf{x}_u^{(k+1)} - \mathbf{z} \right\|_2^2 + I_{\mathcal{C}}(\mathbf{z}) = \underline{\underset{\mathbf{z} \in \mathcal{C}}{\text{argmin}} \left\| \mathbf{x}_u^{(k+1)} - \mathbf{z} \right\|_2^2}$$

- 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: 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
  $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \leq 0$
  (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

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

(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? [Leaving it as homework]

Recall: For subgradient descent, we could give no such guarantee!

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 $\mathbf{x}_u^k = \mathbf{x}_p^{k-1} - t^k \nabla f(\mathbf{x}_p^{k-1})$. Use your unconstrained update
3. Set $\mathbf{x}_p^k = \underset{\mathbf{z} \in \mathcal{C}}{\operatorname{argmin}} \left\|\mathbf{x}_u^k - \mathbf{z}\right\|_2^2$. Project the unconstrained update onto the constraints
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. Next

Figure 15: The projected gradient descent algorithm.

successive iterates are almost coinciding

(more stringent is) that function value is consistently increasing over several projection iterations

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:$\mathbf{x}_p^{k+1} = P_{\mathcal{C}}(\mathbf{x}_u^{k+1}) = \text{argmin}_{\mathbf{z}\in\mathcal{C}} \left\|\mathbf{x}_u^{k+1} - \mathbf{z}\right\|_2^2$

$$\|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 = \|\mathbf{x}^* - \mathbf{x}^k\|^2 - 2t\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) + t^2|\nabla f(\mathbf{x}^k)|^2 \tag{51}$$

- 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 (51)

Homework

$$\|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 \leq \|\mathbf{x}^* - \mathbf{x}^k\|^2 - 2t\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) + t^2 l^2 \tag{52}$$

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

- Further, based on (52)

$$2t\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) \leq \|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 + t^2 l^2 \tag{53}$$

- As per definition of convexity:

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(x^*) \leq \frac{1}{K}\sum_{k=1}^{K}\Big(f(\mathbf{x}^k) - f(\mathbf{x}^*)\Big) \leq \frac{1}{K}\sum_{k=1}^{K}\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) \tag{54}$$

- Substituting for $\nabla f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*)$ from (53) into (54), we get (55):

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{2tK}\sum_{k=1}^{K}\left(\|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^* - \mathbf{x}_u^{k+1}\|^2 + t^2 l^2\right) \tag{55}$$

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:

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{2tK}\left(\|\mathbf{x}^* - \mathbf{x}^0\|^2 - \|\mathbf{x}^* - \mathbf{x}_u^{K+1}\|^2\right) + \frac{tl^2}{2} \tag{56}$$

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

$$f\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{2tK}\left(d^2\right) + \frac{tl^2}{2} \leq \frac{\mathbf{d}l}{\sqrt{K}} \tag{57}$$

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

Demystifying the Projection Step

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

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

Needs more tools (Lagrange

- 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 $p = 1, 2, \infty$

Your assignment 1 is primarily the first constraint
(and possibly also third)

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 (58) with B as an $n \times n$ matrix:

$$\begin{aligned} \min_{\mathbf{z}\in\Re^n} \quad & \tfrac{1}{2}(\mathbf{z}-\mathbf{x})^T B(\mathbf{z}-\mathbf{x}) \\ \text{subject to} \quad & A^T\mathbf{z} = \mathbf{b} \end{aligned} \tag{58}$$

For projected gradient descent, $B =$ I (identity matrix)

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 (58) with B as an $n \times n$ matrix:

$$\begin{aligned} \min_{\mathbf{z}\in\Re^n} \quad & \frac{1}{2}(\mathbf{z}-\mathbf{x})^T B(\mathbf{z}-\mathbf{x}) \\ \text{subject to} \quad & A^T\mathbf{z} = \mathbf{b} \end{aligned} \tag{58}$$

For projected gradient descent, $B = I$. Further, if $n = 2$ and $m = 1$, the minimization problem (58) amounts to finding a point $\mathbf{y}^*$ on a line $a_{11}z_1 + a_{12}z_2 = b$ that is closest to $\mathbf{x}$.

Expect y* to lie on the line/plane such x--y* is perpendicular to the line/plane

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

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

Constraint that should disappear is multiplied with a penalty lambd

$$\min_{\mathbf{z}\in\Re^n,\lambda\in\Re^m} \frac{1}{2}(\mathbf{z}-\mathbf{x})^T B(\mathbf{z}-\mathbf{x}) + \lambda^T(A^T\mathbf{z}-\mathbf{b}) \tag{59}$$

The function $L(\mathbf{z},\lambda)$ is called the lagrangian and involves the lagrange multiplier $\lambda\in\Re^m$.
- A sufficient condition for optimality of $L(\mathbf{z},\lambda)$ at a point $L(\mathbf{z}^*,\lambda^*)$ is that $\nabla L(\mathbf{z}^*,\lambda^*) = 0$ and $\nabla^2 L(\mathbf{z}^*,\lambda^*) \succ 0$. For this specific problem:

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

and

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

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

- The point $(\mathbf{z}^*, \lambda^*)$ must therefore satisfy, $A^T\mathbf{z}^* = \mathbf{b}$ and $A\lambda^* = -B\mathbf{z}^* + \frac{1}{2}(B + B^T)\mathbf{x}$.
- Recap: If B is taken to be the identity matrix, $n = 2$ and $m = 1$, the minimization problem (58) amounts to finding a point $\mathbf{y}^*$ on a line $a_{11}z_1 + a_{12}z_2 = b$ that is closest to $\mathbf{x}$.
- From geometry, the point on a line closest to $\mathbf{x}$ is the point of intersection $\mathbf{p}^*$ of a perpendicular (or least possible[8] obtuse angle) from the origin to the line. However, the solution for the minimum of (59), for these conditions coincides with $\mathbf{p}^*$ and is given by:

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

That is, for $n = 2$ and $m = 1$, the solution to (59) is the same as the solution to (58)

- For general n and m,

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

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

Elaboration on the Geometry of the Project
Right angle FOR Affine Set/Unbounded sets
Least possible obtuse angle FOR
Polyhedron/Bounded Sets

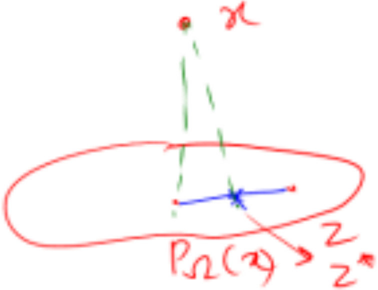
- **Claim:** If $P_{\mathcal{C}}(\mathbf{x})$ is a projection of $\mathbf{x}$, then

$$\left(\mathbf{z} - P_{\mathcal{C}}(\mathbf{x})\right)^\top \left(\mathbf{x} - P_{\mathcal{C}}(\mathbf{x})\right) \le 0, \; \forall\, \mathbf{z} \in \mathcal{C}$$

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

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

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



- Since $P_{\mathcal{C}}(x) = \text{argmin}_{z \in \mathcal{C}} \|x - z\|_2^2$,
  $\left\|x - P_{\mathcal{C}}(x)\right\|^2 \leq \|x - z^*\|^2$

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

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

- 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 $\left\langle z - P_{\mathcal{C}}(x), x - P_{\mathcal{C}}(x) \right\rangle \leq 0$

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

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

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