1 Proximal Gradient Method (PGM)

We consider problems of the form

\[ \min f_0(x) + f_1(x) \quad (P) \]

where the functions \( f_1 \) and \( f_0 \) are smooth and non-smooth, respectively. Both functions are assumed to be convex.

\[ x^{r+1} = \text{prox}_{\alpha^r f_0}(x^r - \alpha^r \nabla f_1(x^r)) \]

plugging in the definition for the \text{prox} operator, we get

\[ = \arg \min_x \alpha^r f_0(x) + \frac{1}{2} \|x - x^r + \alpha^r \nabla f_1(x^r)\|^2_2 \]

\[ = \arg \min_x \alpha^r f_0(x) + \frac{1}{2} \|x - x^r\|^2_2 + \alpha^r \langle \nabla f_1(x^r), x - x^r \rangle + \frac{1}{2} (\alpha^r)^2 \|\nabla f_1(x^r)\|^2_2 \]

however, the last term is a constant, therefore

\[ = \arg \min_x f_0(x) + f_1(x^r) + \langle \nabla f_1(x^r), x - x^r \rangle + \frac{1}{2} (\alpha^r)^2 \|x - x^r\|^2_2 \]

\[ = \arg \min_x \hat{f}(x; x^r). \]

Observe that \( \hat{f} \) involves the first-order approximation to \( f_1 \) as \( f_1(x^r) + \langle \nabla f_1(x^r), x - x^r \rangle \).

To ensure that the method converges to an optimal solution of \( P \), the parameters \( \alpha^r \) must be chosen sufficiently small. In particular, given the
Lipschitz constant $L$ for the function $f = f_0 + f_1$, the choice $\alpha^r \leq \frac{1}{L}$ leads to a sequence of upper-bounding function $\hat{f}(x, x^r) \geq f(x)$, $\forall x$. Indeed, under this setting, one can show that $f(x^{r+1}) \leq f(x')$ for $r \geq 0$, since $f(x^{r+1}) \leq \hat{f}(x^{r+1}, x^r) \leq \hat{f}(x^r; x^r) = f(x^r)$.

From a computational perspective, selecting $\alpha^r$ too small would lead $\hat{f}(x, x^r)$ to have a high curvature, therefore its minimizer will be very close to $x^r$. On the other hand, selecting it too large may reduce the curvature of $\hat{f}(x, x^r)$ arbitrarily and therefore $\hat{f}(\cdot)$ may no longer be an upper-bound of $f(\cdot)$.

**Special Cases** We begin with the case where $f_0 = 0$, therefore problem $P$ is a smooth and convex. In this setting, we have

$$x^{r+1} = \arg \min_x f_1(x^r) + \langle f_1(x^r), x - x^r \rangle + \frac{1}{2\alpha^r} \| x - x^r \|_2^2.$$ 

The above function is minimized by taking its derivative with respect to $x$, which gives the following optimality condition:

$$\nabla f_1(x^r) + \frac{1}{\alpha^r} (x - x^r) = 0.$$ 

Rewriting the above condition as

$$x^{r+1} = x^r - \alpha^r \nabla f_1(x^r),$$

we get the same update rule as we had for the gradient descent method.

Another special case where $f_1 = 0$ leads to the proximal point algorithm, which is left out of the scope of this lecture.

If we consider optimization problem:

$$\min_x f_1(x)$$

subject to: $x \in X$.

where $f_1(x)$ is smooth function. We can easily solve this problem with the gradient decent and projection. On the other hand, we can rewrite this problem in the form of $(P)$ with $f_0(x) = I_X$ and solve it with proximal
Figure 1: The exact gradient $\nabla f_1(x)$ and approximal gradient $\tilde{\nabla} f(x)$ in the projection problem.

Gradient:

$$\tilde{\nabla} f(x) = x - \text{prox}_{f_0}(x - \nabla f_1(x))$$

$$= x - \text{proj}_{\mathcal{X}}(x - \nabla f_1(x))$$

Figure 1 shows the relationship between the exact gradient $\nabla f_1(x)$ and approximal gradient $\tilde{\nabla} f(x)$ in the projection problem.

**Example 1** Successive Projection as Proximal Gradient Method

We consider the problem

$$x^* = \arg\min_x \text{dist}^2(x, \mathcal{X}_2)$$

subject to: $x \in \mathcal{X}_1$.

Clearly, the above problem has an objective function value of 0, when $x^* \in \mathcal{X}_1 \cap \mathcal{X}_2$. To show the equivalence of the successive projection method with the PGM, we rewrite the above problem as

$$\min I_{\mathcal{X}_1}(x) + \frac{1}{2}\text{dist}^2(x, \mathcal{X}_2),$$

where $I_{\mathcal{X}_1}$ is the non-smooth indicator function defined as

$$I_{\mathcal{X}_1}(x) = \begin{cases} 
0 & \text{if } x \in \mathcal{X}_1 \\
\infty & \text{otherwise.}
\end{cases}$$

The function $\text{dist}^2(x, \mathcal{X}_2)$ requires further clarification. We measure the dis-
tance of the point $x$ to the set $X_2$ as

$$\text{dist}^2(x, X_2) = \min \|z - x\|_2^2 \quad \text{subject to: } z \in X_2.$$  

Using the Danskin's Theorem, important conclusions can be drawn for the above problem. In particular, observe that for any $z$, the problem minimizes a strictly convex function, therefore can be interpreted as the minimum of (infinitely many) strictly convex functions. For such functions, Danskin’s Theorem states that the gradient at an arbitrary point $z$ is based on a single function. Noting that the minimizer of the above problem can be given as $z^* = \text{proj}_{X_2}(x)$, we get

$$\tilde{\nabla} \left( \frac{1}{2} \text{dist}(x, X_2) \right) = \tilde{\nabla} \left( \frac{1}{2} \|z^* - x\|_2^2 \right) = x - z^*.$$

We are now ready to iterate the proximal gradient method. Setting $\alpha = 1$

Figure 2: Successive projection between two sets $X_1$ and $X_2$ for all $r$, we observe that

$$x^{r+1} = \arg \min_x I_{X_1}(x) + \langle x^r - \text{proj}_{X_2}(x^r), x - x^r \rangle + \frac{1}{2}\|x - x^r\|_2^2$$

$$= \arg \min_x I_{X_1}(x) + \frac{1}{2}\|x - x^r + x^r - \text{proj}_{X_2}(x^r)\|_2^2$$

$$= \arg \min_x \frac{1}{2}\|x - \text{proj}_{X_2}(x^r)\|_2^2 \quad \text{subject to: } x \in X_1$$

$$= \text{proj}_{X_1} \left( \text{proj}_{X_2}(x^r) \right),$$

which shows that the successive projection method is indeed a special case of PGM.
Example 2  Least Absolute Shrinkage and Selection Operator (LASSO) Problem

Consider the problem

$$\min \frac{1}{2} \|Ax - b\|^2_2 + \lambda \|x\|_1.$$  

With respect to our previous notation, we have $f_0(x) = \lambda \|x\|_1$ and $f_1(x) = \frac{1}{2} \|Ax - b\|^2_2$.

For our discussion, we first define the shrinkage operator $S$ as

$$S_\lambda(z) = \arg \min_x \frac{1}{2} \|x - z\|^2_2 + \lambda \|x\|_1$$

$$= \min_x \frac{1}{2} \sum_i (x_i - z_i)^2 + \lambda \sum_i |x_i|.$$  

The optimal solution $x^*$ for the above problem can be easily characterized by considering the subgradient of its objective function. Below, we give the optimal solution as a function of $z$:

$$x_i^* = \begin{cases} 
  z_i - \lambda & z_i > \lambda \\
  z_i + \lambda & z_i < -\lambda \\
  0 & -\lambda \leq z_i \leq \lambda.
\end{cases}$$  

This is indeed what the shrinkage operator $S$ performs when we write $x^* = S_\lambda(z)$. In the literature, this operation is also referred to as *soft thresholding*.

![Figure 3: soft thresholding operation using shrinkage operator.](image)
Equipped with this new piece of information, we iterate the PGM as follows:

\[
x^{r+1} = \arg \min_x \lambda \|x\|_1 + \langle A^\top (Ax - b), x - x^r \rangle + \frac{1}{2\alpha^r} \|x - x^r\|_2^2
\]

\[
= \arg \min_x \lambda \|x\|_1 + \frac{1}{2\alpha^r} \|x - x^r + \alpha^r A^\top (Ax - b)\|_2^2
\]

\[
= \arg \min_x \lambda \alpha^r \|x\|_1 + \frac{1}{2} \|x - x^r - \alpha^r A^\top (Ax - b)\|_2^2
\]

\[
= S_{\lambda \alpha^r} (x^r - \alpha^r A^\top (Ax - b)).
\]

We note that the above argument can be generalized for arbitrary \( f_1 \), in which case we will have \( x^{r+1} = S_{\lambda \alpha^r} (x^r - \nabla f_1(x^r)) \).

**Example 3  Nuclear Norm Regularizer**

Consider the problem

\[
\min f_1(X) + \lambda \|X\|_*.
\]

As in previous examples, we iterate the PGM as follows:

\[
X^{r+1} = \arg \min_X \lambda \|X\|_* + \langle \nabla f_1(X^r), X - X^r \rangle + \frac{1}{2\alpha^r} \|X - X^r\|_F^2
\]

\[
= \arg \min_X \lambda \alpha^r \|X\|_* + \frac{1}{2} \|X - X^r + \alpha^r \nabla f_1(X^r)\|_F^2
\]

\[
= \arg \min_X \frac{1}{2} \|X - X^r + \alpha^r \nabla f_1(X^r)\|_F^2 + \alpha^r \lambda \|X\|_*.
\]

The above minimization can be written in compact form as

\[
\min_Z \frac{1}{2} \|Z - B\|_F^2 + \mu \|Z\|_*,
\]

where \( B = X^r - \alpha^r \nabla f_1(X^r) \) and \( \mu = \alpha^r \lambda \). Moreover, the solution for it follows a similar pattern as we have done in Example 2 with the shrinkage operator. In particular, we perform singular value decomposition on the matrix \( B \) as \( B = U \Sigma V^\top \). The optimal solution \( Z^* \) is then given by \( Z^* = \)
\[ \text{UDV}^t \text{ such that } \]

\[
D_{ii} = \begin{cases} 
\Sigma_{ii} - \mu & \Sigma_{ii} \geq \mu \\
0 & \text{otherwise.}
\end{cases}
\]

Observe that we again have a similar soft thresholding pattern in the optimal solution of the problem.