Large Scale Optimization for Machine Learning:
Lecture 19

Lecturer: Meisam Razaviyayn Scribe: Wei Ran

Oct. 31, 2017

1 Proximal Gradient Method (PGM)

We consider problems of the form

$$\min \ f_0(\mathbf{x}) + f_1(\mathbf{x}) \qquad (\mathcal{P})$$

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

$$\begin{aligned}
\mathbf{x}^{r+1} &= \mathbf{prox}_{\alpha^r f_0}(\mathbf{x}^r - \alpha^r \nabla f_1(\mathbf{x}^r)) \\
&\quad \textit{plugging in the definition for the } \mathbf{prox} \textit{ operator, we get} \\
&= \arg\min_{\mathbf{x}} \ \alpha^r f_0(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^r + \alpha^r \nabla f_1(\mathbf{x}^r)\|_2^2 \\
&= \arg\min_{\mathbf{x}} \ \alpha^r f_0(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^r\|_2^2 + \alpha^r \langle \nabla f_1(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{1}{2}(\alpha^r)^2 \|\nabla f_1(\mathbf{x}^r)\|_2^2 \\
&\quad \textit{however, the last term is a constant, therefore} \\
&= \arg\min_{\mathbf{x}} \ f_0(\mathbf{x}) + f_1(\mathbf{x}^r) + \langle \nabla f_1(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{1}{2\alpha^r}\|\mathbf{x} - \mathbf{x}^r\|_2^2 \\
&= \arg\min_{\mathbf{x}} \ \hat{f}(\mathbf{x}; \mathbf{x}^r).
\end{aligned}$$

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

To ensure that the method convergences to an optimal solution of $\mathcal{P}$, the parameters α^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 $\hat{f}(\mathbf{x}, \mathbf{x}^r) \geq f(\mathbf{x}),\ \forall \mathbf{x}$. Indeed, under this setting, one can show that $f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r)$ for $r \geq 0$, since $f(\mathbf{x}^{r+1}) \leq \hat{f}(\mathbf{x}^{r+1}; \mathbf{x}^r) \leq \hat{f}(\mathbf{x}^r; \mathbf{x}^r) = f(\mathbf{x}^r)$.

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

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

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

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

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

Rewriting the above condition as

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

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:

$$\begin{aligned} &\min_{\mathbf{x}} \ f_1(\mathbf{x}) \\ &\text{subject to:} \ \ \mathbf{x} \in \mathcal{X}. \end{aligned}$$

where $f_1(\mathbf{x})$ 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 $(\mathcal{P})$ with $f_0(\mathbf{x}) = \mathcal{I}_{\mathcal{X}}$ and solve it with proximal

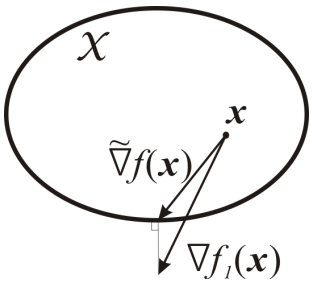

Figure 1: The exact gradient $\nabla f_1(\mathbf{x})$ and approximal gradient $\widetilde{\nabla} f(\mathbf{x})$ in the projection problem.

gradient:

$$\begin{aligned}\widetilde{\nabla} f(\mathbf{x}) &= \mathbf{x} - \mathbf{prox}_{f_0}(\mathbf{x} - \nabla f_1(\mathbf{x})) \\ &= \mathbf{x} - \mathbf{proj}_{\mathcal{X}}(\mathbf{x} - \nabla f_1(\mathbf{x}))\end{aligned}$$

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

Example 1 Successive Projection as Proximal Gradient Method

We consider the problem

$$\begin{aligned}\mathbf{x}^* = \quad & \arg\min_{\mathbf{x}} \ \mathrm{dist}^2(\mathbf{x}, \mathcal{X}_2) \\ \text{subject to:} \quad & \mathbf{x} \in \mathcal{X}_1.\end{aligned}$$

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

$$\min \ \mathcal{I}_{\mathcal{X}_1}(\mathbf{x}) + \frac{1}{2}\mathrm{dist}^2(\mathbf{x}, \mathcal{X}_2),$$

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

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

The function $\mathrm{dist}^2(\mathbf{x}, \mathcal{X}_2)$ requires further clarification. We measure the dis-

tance of the point $\mathbf{x}$ to the set $\mathcal{X}_2$ as

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

Using the Danskin's Theorem, important conclusions can be drawn for the above problem. In particular, observe that for any $\mathbf{z}$, 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 $\mathbf{z}$ is based on a *single* function. Noting that the minimizer of the above problem can be given as $\mathbf{z}^* = \mathbf{proj}_{\mathcal{X}_2}(\mathbf{x})$, we get

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

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

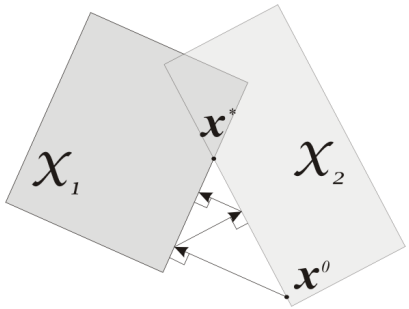

Figure 2: Successive projection between two sets $\mathcal{X}_1$ and $\mathcal{X}_2$

for all r, we observe that

$$\begin{aligned}
\mathbf{x}^{r+1} &= \arg\min_{\mathbf{x}} \ \mathcal{I}_{\mathcal{X}_1}(\mathbf{x}) + \langle \mathbf{x}^r - \mathbf{proj}_{\mathcal{X}_2}(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^r\|_2^2 \\
&= \arg\min_{\mathbf{x}} \ \mathcal{I}_{\mathcal{X}_1}(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^r + \mathbf{x}^r - \mathbf{proj}_{\mathcal{X}_\in}(\mathbf{x}^r)\|_2^2 \\
&= \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{proj}_{\mathcal{X}_2}(\mathbf{x}^r)\|_2^2 \quad \text{subject to: } \mathbf{x} \in \mathcal{X}_1 \\
&= \mathbf{proj}_{\mathcal{X}_1}\left(\mathbf{proj}_{\mathcal{X}_2}(\mathbf{x}^r)\right),
\end{aligned}$$

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

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

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

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

$$\begin{aligned} S_\lambda(\mathbf{z}) &= \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2+\lambda\|\mathbf{x}\|_1 \\ &= \min_{\mathbf{x}} \ \frac{1}{2}\sum_i (x_i-z_i)^2+\lambda\sum_i |x_i|. \end{aligned}$$

The optimal solution $\mathbf{x}^*$ 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 $\mathbf{z}$:

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

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

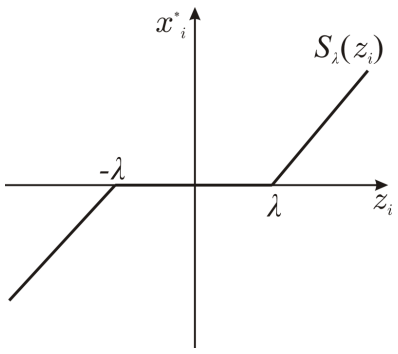

Figure 3: *soft thresholding* operation using shrinkage operator.

Equipped with this new piece of information, we iterate the PGM as follows:

$$\begin{aligned}
\mathbf{x}^{r+1} &= \arg\min_{\mathbf{x}} \ \lambda\|\mathbf{x}\|_1 + \langle \mathbf{A}^\intercal(\mathbf{A}\mathbf{x}^r - \mathbf{b}), \mathbf{x} - \mathbf{x}^r\rangle + \frac{1}{2\alpha^r}\|\mathbf{x} - \mathbf{x}^r\|_2^2 \\
&= \arg\min_{\mathbf{x}} \ \lambda\|\mathbf{x}\|_1 + \frac{1}{2\alpha^r}\|\mathbf{x} - \mathbf{x}^r + \alpha^r\mathbf{A}^\intercal(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2 \\
&= \arg\min_{\mathbf{x}} \ \lambda\alpha^r\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x} - (\mathbf{x}^r - \alpha^r\mathbf{A}^\intercal(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2 \\
&= S_{\lambda\alpha^r}(\mathbf{x}^r - \alpha^r\mathbf{A}^\intercal(\mathbf{A}\mathbf{x} - \mathbf{b})).
\end{aligned}$$

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

Example 3 Nuclear Norm Regularizer

Consider the problem

$$\min \ f_1(\mathbf{X}) + \lambda\|\mathbf{X}\|_*.$$

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

$$\begin{aligned}
\mathbf{X}^{r+1} &= \arg\min_{\mathbf{X}} \ \lambda\|\mathbf{X}\|_* + \langle \nabla f_1(\mathbf{X}^r), \mathbf{X} - \mathbf{X}^r\rangle + \frac{1}{2\alpha^r}\|\mathbf{X} - \mathbf{X}^r\|_2^2 \\
&= \arg\min_{\mathbf{X}} \ \lambda\|\mathbf{X}\|_* + \frac{1}{2\alpha^r}\|\mathbf{X} - \mathbf{X}^r + \alpha^r\nabla f_1(\mathbf{X}^r)\|_F^2 \\
&= \arg\min_{\mathbf{X}} \ \frac{1}{2}\|\mathbf{X} - \mathbf{X}^r + \alpha^r\nabla f_1(\mathbf{X}^r)\|_F^2 + \alpha^r\lambda\|\mathbf{X}\|_*.
\end{aligned}$$

The above minimization can be written in compact form as

$$\min_{\mathbf{Z}} \ \frac{1}{2}\|\mathbf{Z} - \mathbf{B}\|_F^2 + \mu\|\mathbf{Z}\|_*,$$

where $\mathbf{B} = \mathbf{X}^r - \alpha^r\nabla f_1(\mathbf{X}^r)$ 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 $\mathbf{B}$ as $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$. The optimal solution $\mathbf{Z}^*$ is then given by $\mathbf{Z}^* =$

$\mathbf{UDV}^\intercal$ 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.