SDSC6015 - Semester A, 2025

Stochastic Optimization for Machine Learning

6. Proximal Algorithms

1 Motivation

We are interesting to solve the following composite model:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \{F(\boldsymbol{x}) \equiv f(\boldsymbol{x}) + g(\boldsymbol{x})\}$$

with the following assumptions

- $g : \mathbb{R}^n \to (-\infty, \infty]$ is proper closed and convex.
- $f : \mathbb{R}^n \to (-\infty, \infty]$ is proper (convex) and closed, $\mathrm{dom}(f)$ is convex, $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$, and f is L_f-smooth over $\mathrm{int}(\mathrm{dom}(f))$.
- The optimal set of this problem is nonempty and denoted by $\mathcal{X}^\star$. The optimal value of the problem is denoted by $F^\star$.

The above setting covers some special cases:

Examples

- **Smooth unconstrained minimization**

- **Convex constrained smooth minimization**

- L_1**-regularized minimization**

To motivate proximal algorithms, consider

$$\min_{\boldsymbol{x} \in C} f(\boldsymbol{x})$$

for some $C \subset \mathbb{R}^n$. Using projected gradient method, we can solve the above problem by

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= P_C(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)) \\
&= \arg\min_{\boldsymbol{x} \in C} \left\{ f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2\alpha_k} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 \right\}.
\end{aligned}$$

One natural idea to generalize the above step to our setting is

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + g(\boldsymbol{x}) + \frac{1}{2\alpha_k} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 \right\} \\
&= \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \alpha_k g(\boldsymbol{x}) + \frac{1}{2} \|\boldsymbol{x} - (\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k))\|^2 \right\} \\
&=: \operatorname{prox}_{\alpha_k g}(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k))
\end{aligned}$$

where the operator $\operatorname{prox}(\cdot)$ is called the proximal operator.

2 The Proximal Operator

Definition 1. *Given a function $f : \mathbb{R}^n \to (-\infty, \infty]$, the* **proximal mapping** *of f is the operator given by*

$$\mathit{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ f(\boldsymbol{u}) + \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|^2 \right\}.$$

Note that $\operatorname{prox}_f(\boldsymbol{x})$ could be empty, a singleton, and a set of multiple vectors.

Examples

$$g_1(x) = 0, \qquad g_2(x) = \begin{cases} 0, & x \neq 0 \\ -\lambda, & x = 0 \end{cases}, \qquad g_3(x) = \begin{cases} 0, & x \neq 0 \\ \lambda, & x = 0 \end{cases},$$

where $\lambda > 0$.

Theorem 1. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper closed and convex function. Then, $\mathrm{prox}_f(\boldsymbol{x})$ is a singleton for any $\boldsymbol{x} \in \mathbb{R}^n$.*

From now on, we will write $\mathrm{prox}_f(\boldsymbol{x}) = \boldsymbol{y}$ instead of $\mathrm{prox}_f(\boldsymbol{x}) = \{\boldsymbol{y}\}$ for the ease of simplicity.

2.1 Examples

- (Constant) $f(\boldsymbol{x}) = c$ for some $c \in \mathbb{R}$.

- (Affine) $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{c} + b$ for some $\boldsymbol{c} \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

- (Convex quadratic) $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c$ for some $\boldsymbol{A} \in \mathbb{S}^n_+$, $\boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Below are some one-dimensional examples:

- $f(x) = \begin{cases} \mu x & x \geq 0 \\ \infty & x < 0 \end{cases}$ for some $\mu \in \mathbb{R}$.

- $f(x) = \lambda|x|$ for some $\lambda \in \mathbb{R}_+$.

- $f(x) = \begin{cases} \lambda x^3 & x \geq 0 \\ \infty & x < 0 \end{cases}$ for some $\lambda \in \mathbb{R}_+$.

- $f(x) = \begin{cases} -\lambda \log x & x > 0 \\ \infty & x \leq 0 \end{cases}$ for some $\lambda \in \mathbb{R}_+$.

- $f(x) = \delta_{[0,\eta]\cap\mathbb{R}}(x)$ for some $\eta \in [0, \infty]$.

3 Computing Prox

There are some important results that are useful for us to compute the prox.

Theorem 2. *Suppose that* $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \cdots \times \mathbb{R}^{n_m} \to (-\infty, \infty]$ *is given by*

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \sum_{i=1}^{m} f_i(\boldsymbol{x}_i)$$

for any $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$*, for* $i \in [m]$*. Then for any* $\boldsymbol{x}_1 \in \mathbb{R}^{n_1}, \boldsymbol{x}_2 \in \mathbb{R}^{n_2}, \ldots, \boldsymbol{x}_m \in \mathbb{R}^{n_m}$*,*

$$prox_f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = prox_{f_1}(\boldsymbol{x}_1) \times prox_{f_2}(\boldsymbol{x}_2) \times \cdots \times prox_{f_m}(\boldsymbol{x}_m).$$

Example Suppose $g(\boldsymbol{x}) = \lambda \|\boldsymbol{x}\|_1$.

Example Suppose

$$g(\boldsymbol{x}) = \begin{cases} -\lambda \sum_{j=1}^{n} \log x_j, & \boldsymbol{x} \geq \mathbf{0}, \\ \infty, & \text{otherwise}, \end{cases}$$

for some $\lambda > 0$.

Theorem 3 (scaling and translation)**.** *Let* $g : \mathbb{R}^n \to (-\infty, \infty]$ *be a proper function. Let* $\lambda \neq 0$ *and* $\mathbf{a} \in \mathbb{R}^n$*. Define* $f(\boldsymbol{x}) = g(\lambda \boldsymbol{x} + \boldsymbol{a})$*. Then,*

$$prox_f(\boldsymbol{x}) = \frac{1}{\lambda}\left[prox_{\lambda^2 g}(\lambda \boldsymbol{x} + \boldsymbol{a}) - \boldsymbol{a}\right].$$

Theorem 4 (prox of $\lambda g(\cdot/\lambda)$)**.** *Let* $g : \mathbb{R}^n \to (-\infty, \infty]$ *be proper, and let* $\lambda \neq 0$*. Define* $f(\boldsymbol{x}) = \lambda g(\boldsymbol{x}/\lambda)$*. Then*

$$prox_f(\boldsymbol{x}) = \lambda prox_{g/\lambda}(\boldsymbol{x}/\lambda).$$

Theorem 5 (quadratic perturbation)**.** *Let* $g : \mathbb{R}^n \to (-\infty, \infty]$ *be proper, and let* $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \frac{c}{2}\|\boldsymbol{x}\|^2 + \boldsymbol{a}^\top \boldsymbol{x} + \gamma$*, where* $c > 0$*,* $\boldsymbol{a} \in \mathbb{R}^n$*, and* $\gamma \in \mathbb{R}$*. Then*

$$prox_f(\boldsymbol{x}) = prox_{\frac{1}{c+1}g}\left(\frac{\boldsymbol{x} - \boldsymbol{a}}{c+1}\right).$$

Theorem 6 (affine composition). *Let $g : \mathbb{R}^m \to (-\infty, \infty]$ be a proper closed convex function and let $f(\boldsymbol{x}) = g(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ where $\boldsymbol{A}\boldsymbol{A}^\top = \alpha \boldsymbol{I}$ for some constant $\alpha > 0$. Then for any $\boldsymbol{x} \in \mathbb{R}^n$,*

$$prox_f(\boldsymbol{x}) = \boldsymbol{x} + \frac{1}{\alpha}\boldsymbol{A}^\top(prox_{\alpha g}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}).$$

Example Let $g : \mathbb{R}^d \to (-\infty, \infty]$ be proper closed and convex, and let $f : \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to (-\infty, \infty]$ be defined as

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, \boldsymbol{x}_m) = g(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_m).$$

4 Proximal of Indicators & Orthogonal Projections

Consider the special case where g is an indicator function, *i.e.* $g(\boldsymbol{x}) = \delta_C(\boldsymbol{x})$, where C is a nonempty set.

Theorem 7. *Let $C \subseteq \mathbb{R}^n$ be nonempty. Then $\text{prox}_{\delta_C}(\boldsymbol{x}) = P_C(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^n$. If C is closed and convex, then $P_C(\boldsymbol{x})$ is a singleton for any $\boldsymbol{x} \in \mathbb{R}^n$.*

Previous examples:

1. If $C = \mathbb{R}^n_+$, $P_C = \max\{\boldsymbol{x}, \boldsymbol{0}\} = [x]_+$

2. If $C = \text{Box}[\boldsymbol{\ell}, \boldsymbol{u}] = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}\}$, $P_C(\boldsymbol{x}) = (\min\{\max\{x_i, \ell_i\}, u_i\})_{i=1}^n$

3. If $C = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ ($\boldsymbol{A}$ has full row rank), $P_C(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$

4. If $C = B_{\|\cdot\|_2}[\boldsymbol{c}, r]$, $P_C(\boldsymbol{x}) = \boldsymbol{c} + \dfrac{r}{\max\{\|\boldsymbol{x} - \boldsymbol{c}\|_2, r\}}(\boldsymbol{x} - \boldsymbol{c})$

5. If $C = \{\boldsymbol{x} \mid \boldsymbol{a}^\top \boldsymbol{x} \leq \alpha\}$, $\boldsymbol{x} - \dfrac{[\boldsymbol{a}^\top \boldsymbol{x} - \alpha]_+}{\|\boldsymbol{a}\|^2}\boldsymbol{a}$

Theorem 8. *Let $C \subseteq \mathbb{R}^n$ be given by*

$$C = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^\top \boldsymbol{x} = b,\ \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}\},$$

where $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{a} \neq \mathbf{0}$, $b \in \mathbb{R}$, $\boldsymbol{\ell} \in [-\infty, \infty)^n$, $\boldsymbol{u} \in (-\infty, \infty]^n$. Assume that $C \neq \emptyset$. Then

$$P_C(\boldsymbol{x}) = P_{Box[\boldsymbol{\ell},\boldsymbol{u}]}(\boldsymbol{x} - \mu^\star \boldsymbol{a}),$$

where $Box[\boldsymbol{\ell}, \boldsymbol{u}] = \{\boldsymbol{y} \in \mathbb{R}^n \mid \ell_i \leq y_i \leq u_i,\ \forall i \in [n]\}$, and $\mu^\star$ is a solution of the equation

$$\boldsymbol{a}^\top P_{Box[\boldsymbol{\ell},\boldsymbol{u}]}(\boldsymbol{x} - \mu \boldsymbol{a}) = b.$$

Corollary 1. *For any $\boldsymbol{x} \in \mathbb{R}^n$,*

$$P_{\Delta_n}(\boldsymbol{x}) = [\boldsymbol{x} - \mu^\star \mathbf{e}]_+,$$

where $\mu^\star$ is a root of the equation

$$\mathbf{e}^\top [\boldsymbol{x} - \mu^\star \mathbf{e}]_+ - 1 = 0.$$

Theorem 9. *Let*

$$C_\alpha = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \le \alpha\},$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ *is proper closed and convex, and* $\alpha \in \mathbb{R}$*. Assume that there exists* $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ *for which* $f(\hat{\boldsymbol{x}}) \le \alpha$*. Then*

$$P_C(\boldsymbol{x}) = \begin{cases} P_{dom(f)}(\boldsymbol{x}), & f(P_{dom(f)}(\boldsymbol{x})) \le \alpha, \\ prox_{\lambda^\star f}(\boldsymbol{x}) & otherwise, \end{cases}$$

where $\lambda^\star$ *is any positive root of the equation*

$$\varphi(\lambda) \equiv f(prox_{\lambda f}(\boldsymbol{x})) - \alpha = 0.$$

Also, φ *is non-increasing.*

Example Let $C = B_{\|\cdot\|_1}[\mathbf{0}, \alpha] = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_1 \le \alpha\}$, where $\alpha > 0$.

There are a lot more results....

5 Proximal Gradient Method

Recall that

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x}-\boldsymbol{x}_k) + g(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2 \right\} \\
&= \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \alpha_k g(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k))\|^2 \right\} \\
&= \operatorname{prox}_{\alpha_k g}\left(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)\right) \\
&= \operatorname{prox}_{\frac{1}{L_k} g}\left(\boldsymbol{x}_k - \frac{1}{L_k}\nabla f(\boldsymbol{x}_k)\right),
\end{aligned}$$

where we denote $\alpha_k = 1/L_k$ to be the stepsize. To simply the notation, we denote

$$T_L^{f,g}(\boldsymbol{x}) = \operatorname{prox}_{\frac{1}{L}g}\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right).$$

We will omit the superscripts f, g and write $T_L(\cdot)$ instead of $T_L^{f,g}(\cdot)$.

The Proximal Gradient Method

Initialization: $\boldsymbol{x}_0 \in \operatorname{int}(\operatorname{dom}(f))$

General step: For $k = 0, 1, 2, \ldots$

(a) pick $L_k > 0$.

(b) set $\boldsymbol{x}_{k+1} = T_{L_k}(\boldsymbol{x}_k) = \operatorname{prox}_{\frac{1}{L_k}g}\left(\boldsymbol{x}_k - \frac{1}{L_k}\nabla f(\boldsymbol{x}_k)\right)$.

Stepsize strategies

Proposition 1. *For any* $\boldsymbol{x} \in int(dom(f))$ *for which*

$$f(T_L(\boldsymbol{x})) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (T_L(\boldsymbol{x}) - \boldsymbol{x}) + \frac{L}{2}\|T_L(\boldsymbol{x}) - \boldsymbol{x}\|^2,$$

it holds that

$$F(\boldsymbol{x}) - F(T_L(\boldsymbol{x})) \geq \frac{1}{2L}\|L(\boldsymbol{x} - T_L(\boldsymbol{x}))\|^2.$$

Note that the above condition holds when $L = L_f$. This observation motivates the following stepsize strategies:

1. **Constant.** $L_k = L_f$ for all k.

2. **Backtracking.** With parameters (s, η) where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At the kth iteration, set $L_k = L_{k-1}$. Then, while

   $$f(T_{L_k}(\boldsymbol{x}_k)) > f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (T_{L_k}(\boldsymbol{x}_k) - \boldsymbol{x}_k) + \frac{L_k}{2}\|T_{L_k}(\boldsymbol{x}_k) - \boldsymbol{x}_k\|^2,$$

   we set $L_k := \eta L_k$.

Theorem 10. *Let* $\{\boldsymbol{x}_k\}_{k\geq 0}$ *be the sequence generated by the proximal gradient method with either a constant stepsize in which* $L_k \equiv L_f$ *for all* $k \geq 0$ *or the backtracking procedure. Then for any* $\boldsymbol{x}^\star \in \mathcal{X}^\star$ *and* $k \geq 0$,

$$F(\boldsymbol{x}_k) - F^\star \leq \frac{\alpha L_f \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2k}, \tag{1}$$

where $\alpha = 1$ *in the constant stepsize setting and* $\alpha = \max\{\eta, s/L_f\}$ *if backtracking rule is employed.*

6 Proximal Point Method

Consider the problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} g(\boldsymbol{x}),$$

where $g : \mathbb{R}^n \to (-\infty, \infty]$ is a proper closed and convex function. This is a special case with $f \equiv 0$. The associated algorithm is called the **proximal point method**.

The Proximal Point Method

Initialization: $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $c > 0$.

General step: For $k = 0, 1, 2, \ldots$

$$\boldsymbol{x}_{k+1} = \operatorname{prox}_{cg}(\boldsymbol{x}_k).$$

Theorem 11. *Let $g : \mathbb{R}^n \to (-\infty, \infty]$ be a proper closed and convex function. Assume that the problem $\min_{\boldsymbol{x} \in \mathbb{R}^n} g(\boldsymbol{x})$ has a nonempty optimal set $\mathcal{X}^\star$, and let the optimal value be given by $g^\star$. Let $\{\boldsymbol{x}_k\}_{k \geq 0}$ be the sequence generated by the proximal point method with parameter $c > 0$. Then,*

(a) $g(\boldsymbol{x}_k) - g^\star \leq \dfrac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2ck}$ *for any $\boldsymbol{x}^\star \in \mathcal{X}^\star$ and $k \geq 0$;*

(b) the sequence $\{\boldsymbol{x}_k\}_{k \geq 0}$ converges to some point in $\mathcal{X}^\star$.

7 The Augmented Lagrangian Method and ADMM

7.1 The Augmented Lagrangian Method

We consider the problem

$$H^{\star} = \min \left\{ H(\boldsymbol{x}, \boldsymbol{z}) \equiv h_1(\boldsymbol{x}) + h_2(\boldsymbol{z}) \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c} \right\},$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{m \times p}$, and $\boldsymbol{c} \in \mathbb{R}^m$. Here we assume that h_1 and h_2 are proper closed and convex functions. The Lagrangian of the above problem is

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) &= h_1(\boldsymbol{x}) + h_2(\boldsymbol{z}) + \boldsymbol{y}^{\top}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}) \\ &= h_1(\boldsymbol{x}) + \boldsymbol{y}^{\top}\boldsymbol{A}\boldsymbol{x} + h_2(\boldsymbol{z}) + \boldsymbol{y}^{\top}\boldsymbol{B}\boldsymbol{z} - \boldsymbol{y}^{\top}\boldsymbol{c}. \end{aligned}$$

The dual function $g(\boldsymbol{y})$ is given by

We apply the proximal point method (the convergence is known) to the above problem with parameter $\rho > 0$:

The solution of the proximal step is (using Fermat's optimality and conjugate subgradient theorem) $\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$ where

$$\begin{aligned}
\boldsymbol{x}_{k+1} &\in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{y}_{k+1} + h_1(\boldsymbol{x}), \\
\boldsymbol{z}_{k+1} &\in \operatorname*{arg\,min}_{\boldsymbol{z}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{B}^\top \boldsymbol{y}_{k+1} + h_2(\boldsymbol{z}).
\end{aligned}$$

The Augmented Lagrangian Method

Initialization: pick $\boldsymbol{y}_0 \in \mathbb{R}^m$, $\rho > 0$.

General step: For $k = 0, 1, 2, \ldots$, compute

$$\begin{aligned}
(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1}) &\in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^p} \left\{ h_1(\boldsymbol{x}) + h_2(\boldsymbol{z}) + \frac{\rho}{2}\left\| \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c} + \frac{1}{\rho}\boldsymbol{y}_k \right\|^2 \right\} \\
\boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c}).
\end{aligned}$$

7.2 Alternating Direction Method of Multipliers (ADMM)

In general, the step in the augmented Lagrangian method

$$(\boldsymbol{x}^\star, \boldsymbol{z}^\star) = \underset{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^p}{\arg\min} \left\{ h_1(\boldsymbol{x}) + h_2(\boldsymbol{z}) + \frac{\rho}{2}\left\| \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c} + \frac{1}{\rho}\boldsymbol{y}_k \right\|^2 \right\}$$

is not easy. The ADMM replaces the above update using alternating minimization method.

<u>**ADMM**</u>

Initialization: pick $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\boldsymbol{z}_0 \in \mathbb{R}^p$, $\boldsymbol{y}_0 \in \mathbb{R}^m$, $\rho > 0$.

General step: For $k = 0, 1, 2, \ldots$, compute

$$\begin{aligned}
\boldsymbol{x}_{k+1} &\in \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min} \left\{ h_1(\boldsymbol{x}) + \frac{\rho}{2}\left\| \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{c} + \frac{1}{\rho}\boldsymbol{y}_k \right\|^2 \right\} \\
\boldsymbol{z}_{k+1} &\in \underset{\boldsymbol{z}\in\mathbb{R}^p}{\arg\min} \left\{ h_2(\boldsymbol{z}) + \frac{\rho}{2}\left\| \boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c} + \frac{1}{\rho}\boldsymbol{y}_k \right\|^2 \right\} \\
\boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c}).
\end{aligned}$$

7.3 Special Case: $f_1(\boldsymbol{x}) + f_2(\boldsymbol{A}\boldsymbol{x})$

Consider

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ f_1(\boldsymbol{x}) + f_2(\boldsymbol{A}\boldsymbol{x}) \right\}.$$

.

ADMM - $f_1(\boldsymbol{x}) + f_2(\boldsymbol{A}\boldsymbol{x})$

Initialization: pick $\boldsymbol{x}_0, \boldsymbol{w}_0, \boldsymbol{y}_0^2 \in \mathbb{R}^n,\ \boldsymbol{z}_0, \boldsymbol{y}_0^1 \in \mathbb{R}^m,\ \rho > 0.$

General step: For $k = 0, 1, 2, \ldots,$ compute

$$
\begin{aligned}
\boldsymbol{x}_{k+1} &= (\boldsymbol{I} + \boldsymbol{A}^\top \boldsymbol{A})^{-1} \left(\boldsymbol{A}^\top \left[\boldsymbol{z}_k - \frac{1}{\rho} \boldsymbol{y}_k^1 \right] + \boldsymbol{w}_k - \frac{1}{\rho} \boldsymbol{y}_k^2 \right) \\
\boldsymbol{z}_{k+1} &= \operatorname{prox}_{\frac{1}{\rho} f_2} \left(\boldsymbol{A}\boldsymbol{x}_{k+1} + \frac{1}{\rho} \boldsymbol{y}_k^1 \right) \\
\boldsymbol{w}_{k+1} &= \operatorname{prox}_{\frac{1}{\rho} f_1} \left(\boldsymbol{x}_{k+1} + \frac{1}{\rho} \boldsymbol{y}_k^2 \right) \\
\boldsymbol{y}_{k+1}^1 &= \boldsymbol{y}_{k,1} + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{z}_{k+1}) \\
\boldsymbol{y}_{k+1}^2 &= \boldsymbol{y}_{k,2} + \rho(\boldsymbol{x}_{k+1} - \boldsymbol{w}_{k+1})
\end{aligned}
$$

Applications

- L_1-regularized least squares
$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2+\lambda\|\boldsymbol{x}\|_1.$$

- Robust regression
$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_1.$$

- Basis pursuit
$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{\|\boldsymbol{x}\|_1 \mid \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}\}$$

.