Chris Junchi Li

Optimization for Statistical Machine Learning, Part I

Springer

Contents

1 Gradient and Projected Gradient Descent Algorithms 1
1.1 Basics of Gradient Descent Strategy 1
1.1.1 Geometric Conditions of the Objective Functions 2
1.1.2 Choosing the Step Sizes 3
1.2 Theory of Gradient Descent Algoirthms 4
1.2.1 Convergence under Q-Lipschitz Condition 5
1.2.2 Convergence under Q-Lipschitz and α-Strong Convexity . . . 6
1.2.3 Convergence under L-Smoothness Condition 7
1.2.4 Convergence under L-Smoothness and α-Strong Convexity . 8
1.3 Constrained Optimization and Projected Gradient Descent 9
1.3.1 Projected Gradient Descent Method 9
1.3.2 Geometry of the Projection Step 9
1.3.3 The Impact of the Projection Step on Backtracking Line Search 10
1.3.4 Theory of Projected Gradient Descent 11
1.4 Machine Learning Applications 12
1.4.1 Fitting Sparse Generalized Linear Model 12
1.4.2 Application to Matrix Sensing and Matrix Completion 12
1.4.3 Application to Graphical Models 12

2 Proximal Gradient Method and Proximal Operator 13
2.1 Basics of Projected Gradient Descent Algorithm 14
2.1.1 Proximal Gradient Algorithm 14
2.1.2 Proximal Operator and Generalized Gradient 15
2.1.3 Geometric Conditions and Step Size Selection 16
2.1.4 Choosing the Step Size 16
2.2 Theory of Proximal Gradient Descent Algorithm 17
2.3 An Operator Splitting Framework for Proximal Methods 19
2.3.1 Forward-Backward Splitting of the Proximal Operator 20
2.3.2 Basic Concepts and Properties of Operators 21
2.3.3 Properties of Forward and Backward Operators 21

2.4 A Unified Operator Theory for Proving Algorithmic Convergence . . 25

3 Peaceman-Rachford Operator Splitting Algorithm Family 31
3.1 Relaxed Peaceman-Rachford Operator Splitting Algorithm 32
3.2 Convergence Rates of Relaxed Peaceman-Rachford Algorithm 33
3.3 ADMM 37
3.4 A Unified ADMM by Majorization and Minimization 41
3.4.1 Standard ADMM 41
3.4.2 Proximal ADMM 42
3.4.3 Linearized ADMM 42
3.4.4 Linearized ADMM with Parallel Splitting 42
3.4.5 Proximal Linearized ADMM Parallel Splitting 42
3.4.6 Majorant functions 43
3.4.7 Unified Gauss-Seidel ADMM 44
3.4.8 Unified Jacobian ADMM 45

4 Stochastic Gradient Descent 47
4.1 Basics of Stochastic Gradient Descent 47
4.1.1 Geometric Conditions of the Objective Functions 48
4.2 Theory of Stochastic Gradient Descent 49
4.2.1 Convergence under stochastic Q-Lipschitz conditions 51
4.2.2 Convergence under stochastic Q-Lipschitz conditions and α-strong convexity 52
4.2.3 Convergence under stochastic (L,σ)-smoothness condition . 53
4.2.4 Convergence under stochastic (L,σ)-smoothness and strongly convexity conditions 54

5 Stochastic Proximal Gradient Descent 57
5.1 Theory of Stochastic Proximal and Projected Gradient Descent 58
5.1.1 Analysis of stochastic projected gradient descent 59
5.1.2 Analysis of stochastic proximal gradient descent 59

6 SVRG and SDCA 61
6.1 SVRG 61
6.2 Stochastic Dual Coordinate Ascent 64

7 Structured Optimization 65
7.1 Computational Complexity 65

Chapter 1
Gradient and Projected Gradient Descent Algorithms

We start with an introduction of the gradient descent methods.

We consider an unconstrained optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \tag{1.0.1}$$

where $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ is convex and may not be differentiable. We always consider the setting that there exists x^*, such that

$$f(x^*) = \min_{x\in\mathbb{R}^d} f(x).$$

1.1 Basics of Gradient Descent Strategy

The gradient based algorithm iterates the following equations:

(sub-)Gradient Method

For $t = 0, 1, \ldots, k$, iterates

$$x^{t+1} = x^t - \eta_t g(x^t) \tag{1.1.1}$$

until converges

Here $g(x^t) \in \partial f(x^t)$ is a subgradient. When f is differentiable, we denote $g(x^t) = \nabla f(x^t)$, thus the updating equation becomes

$$x^{t+1} = x^t - \eta_t \nabla f(x^t). \tag{1.1.2}$$

When f is twice differentiable, such an updating step can be viewed as minimizing a local quadratic surrogate of the function f at x^t. More specifically, we have

$$x^{t+1} = \operatorname*{argmin}_x \Big\{ f(x^t) + \nabla f(x^t)^\top (x - x^t) + \frac{1}{2\eta_t} \|x - x^t\|_2^2 \Big\}.$$

1.1.1 Geometric Conditions of the Objective Functions

The main challenge of developing a gradient based methods is to choose a right stepsize η_t. To understand the theory of choosing the stepsize, we consider the following geometric conditions.

Condition 1.1 (Q-Lipschitz) *This condition requires that f is Q-Lipschitz:*

$$|f(y) - f(x)| \le Q\|y - x\|_2.$$

Assumption 1.1 is equivalent to the condition that $\forall g \in \partial f(x)$, $\|g\|_2 \le Q$.

Condition 1.2 (L-Smoothness) *We assume f is differential and its gradient function ∇f is L-Lipschitz:*

$$\big|\nabla f(y) - \nabla f(x)\big|_2 \le L\|y - x\|_2^2.$$

The above condition is equivalent to the fact f can be upper bounded by a quadratic function at any point, i.e.,

$$\forall y, \quad f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2.$$

Condition 1.3 (α-Strong Convexity) *This condition requires*

$$\forall y, \quad f(y) \ge f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|_2^2.$$

To understand these conditions. First, the Q-Lipschitz condition and L-smoothness condition impose Lipschitz conditions on the original function and its gradient function separately. They do not imply each other.

Example 1.1 (Q-Lipschitz vs. L-Smoothness). To understand why Q-Lipschitz condition and L-smoothness condition do not imply each other. we consider XXXXX.

Since L-smoothness upper bound the original function at any point by a quadratic function. This property secures that according to the gradient step, the objective function can achieve a significant decay in terms of the gradient magnitude. More specifically, we have the following important theorem:

Proposition 1.1. *Under the L-smoothness condition, if choosing $\eta_t \le 1/L$, we have*

$$f(x^{t+1}) \le f(x^t) - \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2. \tag{1.1.3}$$

Proof. By the L-smoothness condition, we have

$$\forall y, \quad f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|_2^2.$$

Setting $y = x^{t+1}$, $x = x^t$, and using the fact that

$$y - x = x^{t+1} - x^t = -\eta_t \nabla f(x^t),$$

we have

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) - \eta_t \nabla f(x)^\top \nabla f(x^t) + \frac{L}{2}\eta_k^2 \|\nabla f(x^t)\|_2^2 \\ &\leq f(x^t) - \eta_t \|\nabla f(x^t)\|_2^2 + \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2 \\ &= f(x^t) - \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2, \end{aligned}$$

where we have used the fact that $\eta_t \leq 1/L$.

1.1.2 Choosing the Step Sizes

The art of the gradient based algorithms is about choosing the right step sizes. In particular, the optimization algorithms run under two different modes: adaptive mode and non-adaptive mode.

Definition 1.1 (Adaptive mode). The choice of step size does not depend on the optimization accuracy (measured by $\varepsilon := f(x^t) - f(x^*)$).

Definition 1.2 (Non-adaptive mode). The choice of step size depends on the optimization accuracy.

Under different conditions, we list the step size rule in the following:

Rule 1: Constant step size. We simply choose step size $\eta_0 = \eta_1 = \ldots = \eta$.
Rule 2: Backtracking line search.
Start from η_0, we iterate over $\eta_t = \beta\eta_t$ until

$$f(x^{t+1}) \leq f(x^t) - \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^t.$$

It is easy to show that with constant number of iterations, the above line search algorithm will terminate and its output satisfies

$$\eta_t \geq \min\{\eta_0, \frac{\beta}{L}\}.$$

Rule 3: Deminishing step size

1.2 Theory of Gradient Descent Algoirthms

we now present the main theorem of different gradient based optimization optimizations.

Theorem 1.4. *Let* $R := \|x^0 - x^*\|_2$ *and* $f(x^k_{\text{best}}) := \min_{1\le t\le k} f(x_t)$. *Then*

Option (a). *Under Condition 1.1 and Condition 1.3 with* $\alpha = 0$, *we have*

$$f(x^k_{\text{best}}) - f(x^*) \le \frac{R^2 + Q^2 \sum_{t=1}^k \eta_t^2}{2\sum_{t=1}^k \eta_t}. \tag{1.2.1}$$

It is easy to see that if we choose $\eta_t = 1/t$, *then* $f(x^k_{\text{best}}) - f(x^*) = O(1/\log k)$. *A very slow rate!*
Similarly, if we choose $\eta_t = R/(Q\sqrt{k})$, *we have*

$$f(x^k_{\text{best}}) - f(x^*) \le \frac{RQ}{\sqrt{k}}. \tag{1.2.2}$$

However, such a choice of step size depends on the optimization accuracy, thus is non-adaptive.

Option (b). *Under Condition 1.1 and Condition 1.3 with* $\alpha > 0$, *choosing*

$$\eta_t = \frac{2}{\alpha(t+1)},$$

we have

$$f(x^k_{\text{best}}) - f(x^*) \le \frac{2Q^2}{\alpha(k+1)}. \tag{1.2.3}$$

The most striking property of this result is that it does not depend on R. This is due to the combination of the strong convexity and choice of an adaptive step size.

Option (c). *Under Condition 1.2 and Condition 1.3 with* $\alpha = 0$, *choosing* $\eta_t = \eta \le 1/L$ *or by backtracking line search, we have*

$$f(x^k_{\text{best}}) - f(x^*) \le \frac{R^2}{2\eta_{\min} k}, \tag{1.2.4}$$

where $\eta_{\min} \ge \min\{\eta_0, \beta/L\}$.

Option (d). *Under Condition 1.2 and Condition 1.3 with* $\alpha > 0$, *choosing* $\eta_t = \eta \le 1/L$ *or by backtracking line search, we have*

$$f(x_{\text{best}}^k) - f(x^*) \le (1-\alpha\eta_{\min})^k \frac{R^2}{2\eta_{\min}}. \tag{1.2.5}$$

$$\|x^{t+1} - x^*\|_2 \le \sqrt{1-\alpha\eta_{\min}}\|x^t - x^*\|_2. \tag{1.2.6}$$

Before we proving the results, we need to first understand what convexity brings to us. In the proof section, we always assume the objective function f is differentiable. For nondifferentiable settings, simply replacing the gradient $\nabla f(x^t)$ by a subgradient $g(x^t) \in \partial f(x^t)$ would make the proof goes through.

First, by Condition 1.3, we have

$$f(x^*) \ge f(x^t) + \nabla f(x^t)^\top (x^* - x^k) + \frac{\alpha}{2}\|x^t - x^*\|_2^2. \tag{1.2.7}$$

This implies that

$$\begin{aligned}
&f(x^t) - f(x^*) \\
&\le \nabla f(x^t)^\top (x^k - x^*) - \frac{\alpha}{2}\|x^t - x^*\|_2^2 \\
&= \frac{1}{\eta_t}(x^t - x^{t+1})^\top (x^t - x^*) - \frac{\alpha}{2}\|x^t - x^*\|_2^2 \\
&= \frac{1}{2\eta_t}\Big((1-\alpha\eta_t)\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big) + \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2, \qquad (1.2.8)
\end{aligned}$$

where (1.2.8) follows from the relationship that

$$a^\top b = \frac{\|a\|_2^2 + \|b^2\| - \|a-b\|_2^2}{2}.$$

1.2.1 Convergence under Q-Lipschitz Condition

We aim to prove option (a) where we only have Condition 1.1 and Condition 1.3 with $\alpha = 0$.

Proof. By (1.2.8), we have

$$\begin{aligned}
&\overbrace{f(x^t) - f(x^*)}^{a_t} \qquad (1.2.9)\\
&\le \frac{1}{2\eta_t}\Big(\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big) + \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2 \qquad (1.2.10)\\
&\le \frac{1}{2\eta_t}\Big(\underbrace{\|x^t - x^*\|_2^2}_{b_t} - \underbrace{\|x^{t+1} - x^*\|_2^2}_{b_{t+1}}\Big) + \frac{\eta_t}{2}Q^2. \qquad (1.2.11)
\end{aligned}$$

Therefore, we have

$$\eta_t a_t \le \frac{1}{2}(b_t - b_{t+1}) + \frac{\eta_t^2}{2}Q^2.$$

Summing over t on both sides, we get

$$\sum_{t=1}^{k} \eta_t a_t \le \frac{b_1}{2} + \frac{\sum_{t=1}^{k} \eta_t^2}{2} Q^2 \le \frac{R^2}{2} + \frac{\eta_t^2}{2} Q^2.$$

Furthermore, since $f(x_{\text{best}}^k) - f(x^*) \le f(x^t) - f(x^*)$ for all $t = 1, \dots, k$, we have

$$\sum_{t=1}^{k} \eta_t \left(f(x_{\text{best}}^k) - f(x^*) \right) \le \frac{R^2}{2} + \frac{\eta_t^2}{2} Q^2.$$

Therefore,

$$f(x_{\text{best}}^k) - f(x^*) \le \frac{R^2 + Q^2 \sum_{t=1}^{k} \eta_t^2}{2 \sum_{t=1}^{k} \eta_t^2}.$$

This proves (1.2.1).

To show (1.2.2), we set $\eta_t = \eta$, and get

$$f(x_{\text{best}}^k) - f(x^*) \le \frac{R^2}{2k\eta} + \frac{Q^2 \eta}{2}.$$

Optimizing the right hand side gives the desired result $\eta = R/(Q\sqrt{k})$.

1.2.2 Convergence under Q-Lipschitz and α-Strong Convexity

We aim to prove option (b) where we have Condition 1.1 and Condition 1.3 with $\alpha > 0$.

Proof. Again, by (1.2.8), we have

$$\overbrace{f(x^t) - f(x^*)}^{a_t} \tag{1.2.12}$$

$$\le \frac{1}{2\eta_t} \left((1 - \alpha\eta_t) \|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2 \right) + \frac{\eta_t}{2} \|\nabla f(x^t)\|_2^2 \tag{1.2.13}$$

$$\le \frac{1}{2\eta_t} \left((1 - \alpha\eta_t) \underbrace{\|x^t - x^*\|_2^2}_{b_t} - \underbrace{\|x^{t+1} - x^*\|_2^2}_{b_{t+1}} \right) + \frac{\eta_t}{2} Q^2. \tag{1.2.14}$$

Compared with the previous case where $\alpha = 0$, the only difference here is that we have an extra factor $0 < 1 - \alpha\eta_t < 1$. Such a contractive factor gives us the opportunity for a 'more cleaned' telescoping. Intuitively, if we can get recursion

$$a_t \le C_1 \cdot \left((t-1) b_t + (t+1) b_{t+1} \right) + \frac{C_2}{t}.$$

Multiplying t on both size and summing over t, we get

$$\sum_{t=1}^{k} t a_t \le \underbrace{C \cdot \sum_{t=1}^{k} \big((t-1)t b_t + t(t+1) b_{t+1}\big)}_{=0} + C_2 k.$$

This gives us $\sum_{t=1}^{k} t a_t \le C_2 k$, which further leads to a rate of $O(1/k)$.

To implement the above insight, the key trick is to choose the step size

$$\eta_t = \frac{2}{\alpha(t+1)},$$

we have

$$1 - \alpha\eta_t = 1 - \frac{2}{t+1} = \frac{t-1}{t+1}.$$

Plugging η_t into the upper bound of $f(x^t) - f(x^*)$, we have

$$f(x^t) - f(x^*) \le \frac{\alpha}{4}\Big((t-1)b_t - (t+1)b_{t+1}\Big) + \frac{Q^2}{\alpha(t+1)}.$$

Multiplying t on both size and summing over t, we get

$$\sum_{t=1}^{k} t\big(f(x^t) - f(x^*)\big) \le 0 + \frac{kQ^2}{\alpha}.$$

Furthermore, since $f(x^k_{\text{best}}) - f(x^*) \le f(x^t) - f(x^*)$ for all $t = 1, \ldots, k$, we have

$$\big(\sum_{t=1}^{k} t\big)\big(f(x^k_{\text{best}}) - f(x^*)\big) \le \frac{kQ^2}{\alpha}.$$

Thus,

$$f(x^k_{\text{best}}) - f(x^*) \le \frac{2kQ^2}{\alpha k(k+1)} = \frac{2Q^2}{\alpha(k+1)}.$$

This proves the result in (1.2.3).

1.2.3 Convergence under L-Smoothness Condition

We aim to prove option (c) where we have Condition 1.2 and Condition 1.3 with $\alpha = 0$.

Proof. By (1.2.8), we have

$$\begin{aligned} & f(x^t) - f(x^*) \\ & \le \frac{1}{2\eta_t}\Big((1-\alpha\eta_t)\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big) + \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2. \end{aligned} \tag{1.2.15}$$

This time, f is L-smooth but instead of Q-Lipschitz, by (1.1.3) in Proposition 1.1, we have

$$f(x^{t+1}) \leq f(x^t) - \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2.$$

This implies that

$$\frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2 \leq f(x^t) - f(x^{t+1}). \tag{1.2.16}$$

Plugging this into (1.2.15), we have

$$\begin{aligned} f(x^{t+1}) - f(x^*) &\leq \frac{1}{2\eta_t}\Big((1-\alpha\eta_t)\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big) \\ &\leq \frac{1}{2\eta_{\min}}\Big((1-\alpha\eta_{\min})\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big). \end{aligned} \tag{1.2.17}$$

Since we have $\alpha = 0$, we have

$$f(x^{t+1}) - f(x^*) \leq \frac{1}{2\eta_{\min}}\Big(\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big).$$

Summing over t from 0 to $k-1$, we have

$$\begin{aligned} \sum_{t=0}^{k-1} f(x^{t+1}) - f(x^*) &\leq \frac{1}{2\eta_{\min}}\Big(\|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2\Big) \\ &\leq \frac{1}{2\eta_{\min}}\|x^0 - x^*\|_2^2 = \frac{R^2}{2\eta_{\min}}. \end{aligned}$$

Since $f(x^t)$ is monotone decreasing, thus $f(x^k_{\text{best}}) = f(x^k)$. We have

$$f(x^k_{\text{best}}) - f(x^*) \leq \frac{R^2}{2\eta_{\min}k}.$$

This proves the result in (1.2.4).

1.2.4 Convergence under L-Smoothness and α-Strong Convexity

We aim to prove option (d) where we have Condition 1.1 and Condition 1.3 with $\alpha > 0$.

Proof. First, by (1.2.17), since $f(x^{t+1}) - f(x^*) \geq 0$, we have

$$(1-\alpha\eta_{\min})\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2 \geq 0.$$

Therefore, $\|x^{t+1} - x^*\|_2 \leq \sqrt{1-\alpha\eta_{\min}}\|x^t - x^*\|_2$. This proves (1.2.6).

Again, by (1.2.17), we have

$$\begin{aligned} f(x^{t+1}) - f(x^*) &\leq \frac{1}{2\eta_{\min}}\Big((1-\alpha\eta_{\min})\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big) \\ &\leq \frac{1}{2\eta_{\min}}(1-\alpha\eta_{\min})\|x^t - x^*\|_2^2 \\ &\leq \frac{1}{2\eta_{\min}}(1-\alpha\eta_{\min})^{k+1}\|x^0 - x^*\|_2^2 \\ &\leq \frac{1}{2\eta_{\min}}(1-\alpha\eta_{\min})^{k+1}R^2. \end{aligned}$$

We then have $f(x_{\text{best}}^k) - f(x^*) \leq (1-\alpha\eta_{\min})^k \frac{R^2}{2\eta_{\min}}$, which proves (1.2.6).

1.3 Constrained Optimization and Projected Gradient Descent

We now consider a constrained optimization problem

$$\min_{x\in C} f(x) \tag{1.3.1}$$

where $C \subset \mathbb{R}^d$ is a convex set and $f : \mathbb{R}^d \to \mathbb{R}$ is convex and may not be differentiable. We always consider the setting that there exists x^*, such that

$$f(x^*) = \min_{x\in C} f(x).$$

1.3.1 Projected Gradient Descent Method

Projected gradient algorithm

For $t = 0, 1, \ldots,$, we iterates over the following two steps:

Step 1: $z^{t+1} = x^t - \eta_t \nabla f(x^t)$;
Step 2: $x^{t+1} = \min_{x\in C} \|x - z^{t+1}\|_2^2 := P_C(z^{t+1})$.

When f is nondiffernetiable at x^t, we replace $f(x^t)$ by its subgradient at x^t.

1.3.2 Geometry of the Projection Step

To add: The figures

There is very nice convex geometry of the projection step. In particular, we have the following proposition.

Proposition 1.2. *We have following properties: For any $x \in C$*

$$(x - x^{t+1})^\top (x^{t+1} - z^{t+1}) \geq 0, \tag{1.3.2}$$

$$\|x^{t+1} - x\|_2 \leq \|z^{t+1} - x\|_2. \tag{1.3.3}$$

1.3.3 The Impact of the Projection Step on Backtracking Line Search

Compared with the unconstrained settings, the only thing changed is an extra projection step. In fact, this step only affects very few things as we now listed.

First, we need to modify the backtracking line search. Recall that in the unconstrained setting: Start from η_0, we iterate over $\eta_t = \beta\eta_t$ until

$$f(x^{t+1}) \leq f(x^t) - \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2.$$

For the projected gradient method, we modify the line search stopping criterion to be

$$f(x^{t+1}) \leq f(x^t) + \nabla f(x^t)^\top (x^{t+1} - x^t) + \frac{1}{2\eta_t}\|x^{t+1} - x^t\|_2^2. \tag{1.3.4}$$

Remark 1.1. Without the projection step, we have

$$x^{t+1} - x^t = -\eta_t \nabla f(x^t).$$

This coincides with the unconstrained setting.

It is easy to show that with constant number of iterations, the above line search algorithm will terminate and its output satisfies

$$\eta_t \geq \min\Big\{\eta_0, \frac{\beta}{L}\Big\}.$$

It is obvious that $\eta_k = \eta \leq 1/L$ satisfies the above inequality. To see this, by L-smoothness, we have

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \nabla f(x^t)^\top (x^{t+1} - x^t) + \frac{L}{2}\|x^{t+1} - x^t\|_2^2 \\ &\leq f(x^t) + \nabla f(x^t)^\top (x^{t+1} - x^t) + \frac{1}{2\eta_t}\|x^{t+1} - x^t\|_2^2. \end{aligned}$$

The line search criterion in (1.3.4) implies an important property of a one-step decay $f(x^{t+1}) - f(x^t)$, as summarized by the following proposition:

Proposition 1.3. *If f is L-smooth and $\eta_t \leq 1/L$, we have*

$$f(x^{t+1}) - f(x^t) \leq \frac{1}{2\eta_t}\Big[\|x^{t+1} - z^{t+1}\|_2^2 - \|z^{t+1} - x^t\|_2^2\Big].$$

Proof. By (1.3.4), we have

$$\begin{aligned}
&f(x^{t+1}) - f(x^t) \\
&\leq \nabla f(x^t)^\top (x^{t+1} - x^t) + \frac{1}{2\eta_t}\|x^{t+1} - x^t\|_2^2 \\
&= -\frac{1}{\eta_t}(z^{t+1} - x^t)^\top (x^{t+1} - x^t) + \frac{1}{2\eta_t}\|x^{t+1} - x^t\|_2^2 \\
&= \frac{1}{2\eta_t}\Big[\|x^{t+1} - z^{t+1}\|_2^2 - \|z^{t+1} - x^t\|_2^2\Big].
\end{aligned}$$

This finishes the proof.

1.3.4 Theory of Projected Gradient Descent

The theoretical results and proofs of the projected gradient method are almost the same as the proof for the gradient descent in the unconstrained settings. Here we mainly illustrate the difference.

First, by the convexity Condition 1.3, we have

$$f(x^*) \geq f(x^t) + \nabla f(x^t)^\top (x^* - x^k) + \frac{\alpha}{2}\|x^t - x^*\|_2^2. \tag{1.3.5}$$

This implies that

$$\begin{aligned}
&f(x^t) - f(x^*) \\
&\leq \nabla f(x^t)^\top (x^k - x^*) - \frac{\alpha}{2}\|x^t - x^*\|_2^2 \\
&= \frac{1}{\eta_t}(x^t - z^{t+1})^\top (x^t - x^*) - \frac{\alpha}{2}\|x^t - x^*\|_2^2 \\
&= \frac{1}{2\eta_t}\Big((1 - \alpha\eta_t)\|x^t - x^*\|_2^2 - \|z^{t+1} - x^*\|_2^2 + \|x^t - z^{t+1}\|_2^2\Big) \qquad (1.3.6) \\
&\leq \frac{1}{2\eta_t}\Big((1 - \alpha\eta_t)\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big) + \frac{\eta_t}{2}\|\nabla f(x^t)\|_2^2. \qquad (1.3.7)
\end{aligned}$$

This result is exactly the same as (1.2.8) as for the unconstrained cases.

If f is L-smooth and $\eta_t \leq 1/L$, by Proposition 1.3, we have

$$f(x^{t+1}) - f(x^t) \leq \frac{1}{2\eta_t}\Big[\|x^{t+1} - z^{t+1}\|_2^2 - \|z^{t+1} - x^t\|_2^2\Big]. \tag{1.3.8}$$

Summing (1.3.8) and (1.3.6), we obtain

$$
\begin{aligned}
& f(x^{t+1}) - f(x^*) \\
& \leq \frac{1}{2\eta_t}\Big((1-\alpha\eta_t)\|x^t - x^*\|_2^2 \underbrace{-\|z^{t+1} - x^*\|_2^2 + \|x^{t+1} - z^{t+1}\|_2^2}_{\leq -\|x^{t+1}-x^*\|_2^2 \text{ by geometry.}}\Big) \\
& \leq \frac{1}{2\eta_t}\Big((1-\alpha\eta_t)\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\Big). \qquad (1.3.9)
\end{aligned}
$$

This result is exactly the same as (1.2.17) as for the unconstrained cases.

Given the above theoretical interfaces, the remaining proofs follow exactly the same way as the unconstrained cases.

1.4 Machine Learning Applications

1.4.1 Fitting Sparse Generalized Linear Model

Many machine learning problems can be solved by projected gradient descent method.

1.4.2 Application to Matrix Sensing and Matrix Completion

1.4.3 Application to Graphical Models

Chapter 2
Proximal Gradient Method and Proximal Operator

In this lecture, we introduce the proximal gradient descent algorithm and its analysis. Essentially, we aim to show that the theory of the proximal gradient descent algorithm is almost identical to that of the vanilla gradient descent algorithm. However, the specific proximal gradient descent algorithm is not our major goal in this lecture. The main purpose is to use proximal gradient descent algorithm to illustrate an operator based framework, which provides insight on designing new optimization algorithms.

We consider an unconstrained optimization problem with a composite objective function

$$\min_{x\in\mathbb{R}^d} f(x) := \min_{x\in\mathbb{R}^d} g(x)+h(x) \tag{2.0.1}$$

where $x \in \mathbb{R}^d$ and $h, g : \mathbb{R}^d \to \mathbb{R}$ are convex and may not be differentiable. We always consider the setting that there exists x^*, such that

$$f(x^*) = \min_{x\in\mathbb{R}^d} f(x).$$

It is easy to see that the constrained optimization with L-smooth objective is a special case of this setting. To see this, we consider a constrained optimization problem $\min_{x\in C} g(x)$. It can be equivalently represented as

$$\min_{x\in\mathbb{R}^d} g(x)+h_C(x),$$

where

$$h_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C. \end{cases}$$

2.1 Basics of Projected Gradient Descent Algorithm

We consider the objective function in (2.0.1). The proximal gradient descent algorithm iterates over two steps: (i) a forward gradient step and (ii) a backward proximal step. Essentially, we can view proximal gradient descent as a more generalized version of the projected gradient descent.

2.1.1 Proximal Gradient Algorithm

The proximal gradient descent algorithm runs as following.

Proximal Gradient Descent Algorithm

For $t = 0, 1, \ldots, k$, iterates

$$z^{t+1} = x^t - \eta_t \nabla g(x^t), \tag{2.1.1}$$

$$x^{t+1} = \operatorname*{argmin}_y \frac{1}{2\eta_t}\|y - z^{t+1}\|_2^2 + h(y). \tag{2.1.2}$$

until converge.

What is the intuition of the proximal gradient algorithm? It is easy to see that we can equivalently represent the above two-step algorithm into a one step optimization problem:

$$x^{t+1} = \operatorname*{argmin}_y \underbrace{g(x^t) + \nabla g(x^t)^\top (y - x^t) + \frac{1}{2\eta_t}\|y - x^t\|_2^2}_{\text{A quadratic approximation of g(x) at } x^t} + h(y). \tag{2.1.3}$$

We see the idea is the same as the vanilla gradient descent case, with the only difference that we put a quadratic surrogate on the function g but instead of the function h. The following proposition reveals the equivalence of these formulations.

Proposition 2.1. *The optimization problem in* (2.1.3) *is equivalent to that in* (2.1.2).

Proof. The optimization problem in (2.1.3) can be represented as

$$\begin{aligned} x^{t+1} &= \operatorname*{argmin}_y g(x^t) + \nabla g(x^t)^\top (y - x^t) + \frac{1}{2\eta_t}\|y - x^t\|_2^2 + h(y) \\ &= \operatorname*{argmin}_y \frac{1}{2\eta_t}\|y - \underbrace{\left(x^t - \eta_t \nabla g(x^t)^\top (y - x^t)\right)}_{z^{t+1}}\|_2^2 + h(y). \end{aligned}$$

This finishes the proof.

2.1.2 Proximal Operator and Generalized Gradient

The reason that the above algorithm is called proximal gradient descent is because (2.1.2) can be viewed as a 'proximal' operator, which is defined as

Definition 2.1 (Proximal Operator). Given any function $h : \mathbb{R}^d \to \mathbb{R}$, we define the proximal operator as

$$\text{Prox}_h(x) = \operatorname*{argmin}_y \frac{1}{2}\|x-y\|_2^2 + h(y).$$

As explained before, the proximal operator can be viewed as a 'generalized' projection operator.

The theoretical analysis of the proximal gradient algorithm is similar to that of the vanilla gradient descent algorithm. To get think link, we need to first define a notion of generalized gradient

Definition 2.2 (Generalized gradient). For the proximal gradient descent algorithm, we define the 'generalized' gradient as

$$G_t := G_{\eta_t}(x^t) = \frac{1}{\eta_t}(x^t - x^{t+1}).$$

It is obvious that

$$x^{t+1} = x^t - \eta_t G_t.$$

So G_t servers very similar purpose as the gradient in the vanilla gradient descent algorithm $\nabla g(x^t)$ (without worrying about the term $h(x)$). However, we need to be more careful here. The vanila gradient $\nabla g(x^t)$ only depends on x^t but G_t depends on both x^t and x^{t+1} due to the effect of h. To see this, since x^{t+1} minimizes (2.1.2), we have

$$\begin{aligned} 0 &\in \partial h(x^{t+1}) + \frac{1}{\eta_t}(x^{t+1} - z^{t+1}) \\ &= \partial h(x^{t+1}) + \underbrace{\frac{1}{\eta_t}(x^{t+1} - x^t)}_{G_t} + \nabla g(x^t). \end{aligned}$$

Therefore, we have

$$G_t - \nabla g(x^t) \in \partial h(x^{t+1}).$$

We then have

$$\begin{aligned} x^{t+1} &= x^t - \eta_t G_t \\ &= x^t - \eta_t \big(\underbrace{\nabla g(x^t)}_{\in \partial g(x^t)} + \underbrace{G_t - \nabla g(x^t)}_{\in \partial h(x^{t+1})}\big). \end{aligned}$$

2.1.3 Geometric Conditions and Step Size Selection

To analyze proximal gradient descent algorithm, we need to impose conditions on g and h. Unlike the vanilla setting where g can be either Q-Lipschitz or L-smooth, here we only assume g to be L-smooh (Think about why?). However, we only assume convexity on h (which does not have to be smooth).

To formalize, we start with some definitions.

Definition 2.3 (Proper function). A function f is proper if there exists x, $f(x) < \infty$.

Intuitively, a function is proper if its function values are not always infinity

Definition 2.4 (Closed function). A function f is closed if the sub-level set $\{x : f(x) \leq c\}$ is closed for any c.

The following proposition characterizes the uniqueness of the solution of the proximal operator.

Proposition 2.2. *If h is proper, convex, and closed, the solution of the proximal operator*

$$\text{Prox}_h(x) = \underset{y}{\text{argmin}} \frac{1}{2}\|x-y\|_2^2 + h(y)$$

is uniqiue.

Proof. Since $h(x)$ is convex and the first part of the objective function $\frac{1}{2}\|x-y\|_2^2$ is strongly convex, we have the whole objective function $\frac{1}{2}\|x-y\|_2^2 + h(x)$ is strongly convex. To prove the uniqueness result, all we need is to prove the sublevel set

$$A_c := \left\{ y : \ \frac{1}{2}\|x-y\|_2^2 + h(x) \leq c \right\}$$

needs to fill in.

2.1.4 Choosing the Step Size

As we explained before, the proximal gradient descent method can be viewed as a generalization of the projected gradient method. Thus it is reasonable that they have identical backtracking line search criterion.
More specifically, recall that in the projected gradient descent problem (2.2): Start from $\eta_t = \eta^0$, we iterate over $\eta_t = \beta\eta_t$ until

$$g\left(x^{t+1}\right) \leq g\left(x^t\right) + \nabla g\left(x^t\right)^\top \left(x^{t+1} - x^t\right) + \frac{1}{2\eta_t}\left\|x^{t+1} - x^t\right\|_2^2 \tag{2.7}$$

For the more general proximal gradient descent algorithm, we use exactly the same step size rule as described in (2.7). It is easy to show that with constant number of iterations, the above line search algorithm will terminate and its output satisfies

$$\eta_t \geq \min\left\{\eta^0, \frac{\beta}{L}\right\}$$

It is obvious that $\eta_k = \eta \leq 1/L$ satisfies (2.7). To see this, by L-smoothness, we have

$$\begin{aligned} g\left(x^{t+1}\right) &\leq g\left(x^t\right) + \nabla g\left(x^t\right)^\top \left(x^{t+1} - x^t\right) + \frac{L}{2}\left\|x^{t+1} - x^t\right\|_2^2 \\ &\leq g\left(x^t\right) + \nabla g\left(x^t\right)^\top \left(x^{t+1} - x^t\right) + \frac{1}{2\eta_t}\left\|x^{t+1} - x^t\right\|_2^2 \end{aligned}$$

The line search criterion in (2.7) implies an important property of a onestep decay $g\left(x^{t+1}\right) - g\left(x^t\right)$, as summarized by the following proposition:

Proposition 2.7. If f is L-smooth and $\eta_t \leq 1/L$, we have

$$g\left(x^{t+1}\right) - g\left(x^t\right) \leq \frac{1}{2\eta_t}\left[\left\|x^{t+1} - z^{t+1}\right\|_2^2 - \left\|z^{t+1} - x^t\right\|_2^2\right]$$

Proof. By (2.7), we have

$$\begin{aligned} &g\left(x^{t+1}\right) - g\left(x^t\right) \\ &\leq \nabla g\left(x^t\right)^\top \left(x^{t+1} - x^t\right) + \frac{1}{2\eta_t}\left\|x^{t+1} - x^t\right\|_2^2 \\ &= -\frac{1}{\eta_t}\left(z^{t+1} - x^t\right)^\top \left(x^{t+1} - x^t\right) + \frac{1}{2\eta_t}\left\|x^{t+1} - x^t\right\|_2^2 \\ &= \frac{1}{2\eta_t}\left[\left\|x^{t+1} - z^{t+1}\right\|_2^2 - \left\|z^{t+1} - x^t\right\|_2^2\right] \end{aligned}$$

This finishes the proof.

2.2 Theory of Proximal Gradient Descent Algorithm

We now present the theory of the proximal gradient descent algorithm. It is almost identical to that of the projected gradient descent algorithm. Here we mainly illustrate the difference.

Theorem 2.8 (Convergence of Proximal Gradient Gradient Algorithm). Let $f(x) = g(x) + h(x)$ where h is proper, convex, and closed. Let $R := \|x^0 - x^*\|_2$, we have

Option (a). Let $g(x)$ satisfy Condition 1.2 and Condition 1.3 with $\alpha = 0$, choosing $\eta_t = \eta \leq 1/L$ or by backtracking line search, we have

$$f\left(x^k\right) - f\left(x^*\right) \leq \frac{R^2}{2\eta_{\min} k}$$

where $\eta_{\min} \geq \min\left\{\eta^0, \beta/L\right\}$.
Option (b). Let $g(x)$ satisfy Condition 1.2 and Condition 1.3 with $\alpha > 0$, choosing $\eta_t = \eta \leq 1/L$ or by backtracking line search, we have

$$f\left(x^k\right) - f\left(x^*\right) \leq \left(1 - \alpha\eta_{\min}\right)^k \frac{R^2}{2\eta_{\min}}$$
$$\left\|x^{t+1} - x^*\right\|_2 \leq \sqrt{1 - \alpha\eta_{\min}} \left\|x^t - x^*\right\|_2$$

Proof. The proof follows exactly the same procedure as that for the projected gradient method.
Step 1: Analyzing what α-strong convexity of g brings us.
First, since g satisfies the convexity Condition 1.3, we have

$$g\left(x^*\right) \geq g\left(x^t\right) + \nabla g\left(x^t\right)^\top \left(x^* - x^k\right) + \frac{\alpha}{2}\left\|x^t - x^*\right\|_2^2$$

This implies that

$$\begin{aligned}
& g\left(x^t\right) - g\left(x^*\right) \\
& \leq \nabla g\left(x^t\right)^\top \left(x^k - x^*\right) - \frac{\alpha}{2}\left\|x^t - x^*\right\|_2^2 \\
& = \frac{1}{\eta_t}\left(x^t - z^{t+1}\right)^\top \left(x^t - x^*\right) - \frac{\alpha}{2}\left\|x^t - x^*\right\|_2^2 \\
& = \frac{1}{2\eta_t}\left(\left(1 - \alpha\eta_t\right)\left\|x^t - x^*\right\|_2^2 - \left\|z^{t+1} - x^*\right\|_2^2 + \left\|x^t - z^{t+1}\right\|_2^2\right)
\end{aligned} \tag{2.8}$$

This result is exactly the same as (1.12) as for the unconstrained cases.
Step 2: Analyzing what L-smoothness of g brings us.
Second, since g is also L-smooth and $\eta_t \leq 1/L$, by Proposition 2.7, we have

$$g\left(x^{t+1}\right) - g\left(x^t\right) \leq \frac{1}{2\eta_t}\left[\left\|x^{t+1} - z^{t+1}\right\|_2^2 - \left\|z^{t+1} - x^t\right\|_2^2\right] \tag{2.9}$$

Step 3: Analyzing what convexity of h brings us.
By (2.6), we know that $G_t - \nabla g\left(x^t\right) \in \partial h\left(x^t\right)$. Therefore, we have

$$h\left(x^{t+1}\right) \leq h\left(x^*\right) + \left(G_t - \nabla g\left(x^t\right)\right)^\top \left(x^{t+1} - x^*\right) \tag{2.10}$$

Summing over (2.8), (2.9), and (2.10), we get

$$f\left(x^{t+1}\right) - f\left(x^*\right)$$

$$
\begin{aligned}
&\leq \frac{1}{2\eta_t}\left((1-\alpha\eta_t)\left\|x^t-x^*\right\|_2^2-\left\|z^{t+1}-x^*\right\|_2^2+\left\|x^t-z^{t+1}\right\|_2^2\right)\\
&\quad+\frac{1}{2\eta_t}\left[\left\|x^{t+1}-z^{t+1}\right\|_2^2-\left\|z^{t+1}-x^t\right\|_2^2\right]\\
&\quad-\frac{1}{\eta_t}\left(x^{t+1}-x^t\right)^\top\left(x^{t+1}-x^*\right)+\frac{1}{\eta_t}\left(z^{t+1}-x^t\right)^\top\left(x^{t+1}-x^*\right)\\
&\leq \frac{1}{2\eta_t}\left((1-\alpha\eta_t)\left\|x^t-x^*\right\|_2^2-\left\|z^{t+1}-x^*\right\|_2^2+\left\|x^t-z^{t+1}\right\|_2^2\right)\\
&\quad+\frac{1}{2\eta_t}\left[\left\|x^{t+1}-z^{t+1}\right\|_2^2-\left\|z^{t+1}-x^t\right\|_2^2\right]\\
&\quad-\frac{1}{\eta_t}\left(x^{t+1}-x^t\right)^\top\left(x^{t+1}-x^*\right)+\frac{1}{\eta_t}\left(z^{t+1}-x^*\right)^\top\left(x^{t+1}-x^*\right)\\
&\quad-\frac{1}{\eta_t}\left(x^t-x^*\right)^\top\left(x^{t+1}-x^*\right)\\
&= \frac{1}{2\eta_t}\left[(1-\alpha\eta_t)\left\|x^t-x^*\right\|_2^2-\left\|z^{t+1}-x^*\right\|_2^2+\left\|x^t-z^{t+1}\right\|_2^2\right]\\
&\quad+\frac{1}{2\eta_t}\left[\left\|x^{t+1}-z^{t+1}\right\|_2^2-\left\|z^{t+1}-x^t\right\|_2^2\right]\\
&\quad-\frac{1}{2\eta_t}\left[\left\|x^{t+1}-x^t\right\|_2^2+\left\|x^{t+1}-x^*\right\|_2^2-\left\|x^t-x^*\right\|_2^2\right]\\
&\quad+\frac{1}{2\eta_t}\left[\left\|z^{t+1}-x^*|_2^2+\right\|x^{t+1}-x^*\right\|_2^2-\left\|z^{t+1}-x^{t+1}\right\|_2^2\right]\\
&\quad-\frac{1}{2\eta_t}\left[\left\|x^t-x^*|_2^2+\right\|x^{t+1}-x^*\right\|_2^2-\left\|x^t-x^{t+1}\right\|_2^2\right]\\
&= \frac{1}{2\eta_t}\left[(1-\alpha\eta_t)\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^*\right\|_2^2\right] \qquad (2.11)
\end{aligned}
$$

This result is exactly the same as (1.21) as for the unconstrained cases. Given this theoretical interfaces, the remaining proofs follow exactly the same way as the unconstrained cases.

2.3 An Operator Splitting Framework for Proximal Methods

Recall the proximal gradient descent for the unconstrained optimization problem $\min_{x\in\mathbb{R}^d} g(x)+h(x)$, where $g(x)$ is L-smooth, and $h(x)$ is convex. Proximal gradient descent takes the form

$$z^{t+1} = x^t - \eta_t \nabla g(x^t) \qquad (2.3.1)$$

$$x^{t+1} = \mathrm{Prox}_{\eta_t h}(z^{t+1}), \qquad (2.3.2)$$

where the proximal operator $\mathrm{Prox}_{\eta h}:\mathbb{R}^d\to\mathbb{R}^d$ is defined as

$$\mathrm{Prox}_{\eta h}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \frac{1}{2\eta}\|y-x\|_2^2 + h(y).$$

2.3.1 Forward-Backward Splitting of the Proximal Operator

The proximal step can be equivalently expressed with first order optimality condition

$$0 \in \frac{1}{\eta}(x^{t+1}-z^{t+1}) + \partial h(x^{t+1}), \text{ or equivalently}$$

$$z^{t+1} \in x^{t+1} + \eta\partial h(x^{t+1}) = (I+\eta\partial h)(x^{t+1}).$$

With an abuse of notation, we can write the proximal step using forward/backward operators. Precise definition of these operators will be introduced later.

$$x^{t+1} = (I+\eta\partial h)^{-1} z^{t+1} = \underbrace{(I+\eta\partial h)^{-1}}_{\text{backward operator}} \underbrace{(I-\eta\nabla g)}_{\text{forward operator}} (x^t). \tag{2.3.3}$$

Equation (2.3.3) is called **forward-backward splitting**, which is equivalent to the proximal gradient steps (2.3.1) and (2.3.2).

Remark. The notation of forward-backward operator is related to the forward/backward Euler schemes for solving differential equations. For convex function $f(x), x\in\mathbb{R}^d$, consider the ODE

$$\frac{d}{ds}x(s) = -\nabla f\big(x(s)\big),$$

which is the continuous counterpart of gradient descent for solving $\min_x f(x)$. The Forward/Backward Euler scheme for solving this ODE are Forward Euler scheme:

$$\frac{x^{t+1}-x^t}{\eta} = -\nabla f(x^t) \;\Leftrightarrow\; x^{t+1} = \underbrace{(1-\eta\nabla f)}_{\text{forward operator}}(x^t), \tag{2.3.4}$$

and Backward Euler scheme:

$$\frac{x^{t+1}-x^t}{\eta} = \nabla f(x^{t+1}) \;\Leftrightarrow\; x^{t+1} = \underbrace{(1+\eta\nabla f)^{-1}}_{\text{backward operator}}(x^t), \tag{2.3.5}$$

where $\{x^t\}_{t=1,2\ldots}$ is the discretized solution for $x(s)$ with discretization step as η.

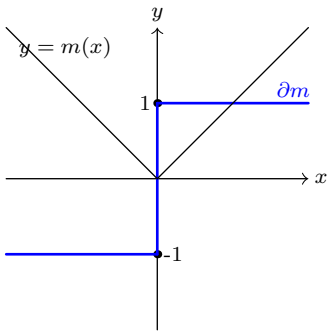

Fig. 2.1 Illustration of subgradient for ℓ_1 function

2.3.2 Basic Concepts and Properties of Operators

2.3.3 Properties of Forward and Backward Operators

Here we will formally define the notion of operator examine some properties that's essential to the convergence of the a wide family of operator splitting algorithm, as a high level abstraction for

Definition 2.5 (Operator). An operator T on $\mathbb{R}^d$ is a subset of $\mathbb{R}^d \times \mathbb{R}^d$. We define

$$T(x) := \{y \mid (x,y) \in T\},$$

$$\text{dom}T := \{x \mid \exists y : (x,y) \in T\}.$$

Example. The subgradient ∂m of ℓ_1 function $m(x) = \|x\|_1, x \in \mathbb{R}^d$ where the subgradient is defined as

$$\partial m = \{(x,y) | x,y \in \mathbb{R}^d, \text{ and } m(z) \geq m(x) + y^\top (z-x)\ \forall z \in \mathbb{R}^d\}.$$

Figure 2.3.3 below is an illustration of ∂m on $\mathbb{R} \times \mathbb{R}$.

Definition 2.6 (Inverse of Operator). For any operator T, we can define the inverse of T as

$$T^{-1} := \{(x,y) \mid (y,x) \in T\}.$$

If operator T is single-valued, i.e. $T(x)$ is a singleton set $\{y\}$ for all $x \in \text{dom}T$, we will denote $T(x)$ as y with slight abuse of notation, and will call the operator T as a function.

Proposition 2.3. *For any convex function $f : \mathbb{R}^d \to \mathbb{R}$, the operator $(I + \eta \partial f)^{-1}$ is actually an $\mathbb{R}^d \to \mathbb{R}^d$ function. This operator is called the resolvent of the operator ∂f.*

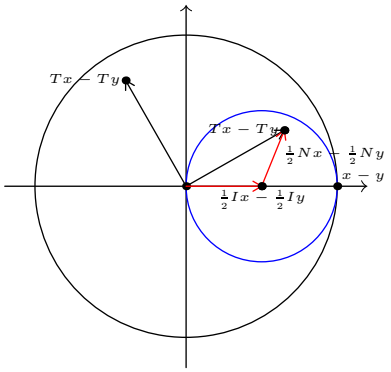

Fig. 2.2 Illustration of Non-Expansiveness (NE) and Firmly Non-Expansiveness (FNE)

Proof. For any $(x,y) \in (I+\eta\partial f)^{-1}$, we have

$$(y,x) \in (I+\eta\partial f) \Rightarrow x = (I+\eta\partial f)(y) = y+\eta\partial f(y).$$

Since f is convex, (x,y) in the above equation satisfies the optimality condition for $\min_{z\in\mathbb{R}^d} \frac{1}{\eta}\|z-x\|_2^2 + h(z)$. Since the objective function is strongly convex, and thus has a unique optimizer, we know that

$$y = \underset{z\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{\eta}\|z-x\|_2^2 + h(z)$$

is unique, i.e. $(I+\eta\partial f)^{-1}(x)$ is a single-valued mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Next we introduce the three most important properties for our analysis: contraction, non-expansiveness and firmly non-expansiveness.

Definition 2.7 (Contraction, Non-Expansiveness (NE) and Firmly Non-Expansiveness (FNE)). Let T be an operator as a subset of $\mathbb{R}^d \times \mathbb{R}^d$,

- Contraction. $\|y-y'\|_2 \le C\|x-x'\|_2$, $C<1$ for all $(x,y),(x',y') \in T$
- Non-Expansiveness (NE). $\|y-y'\|_2 \le \|x-x'\|_2$ for all $(x,y),(x',y') \in T$.
- Firmly Non-Expansiveness (FNE). $\|y-y'\|_2 \le \langle y-y', x-x'\rangle$ for all $(x,y),(x',y') \in T$.

Figure 2.3.3 illustrates the notions of NE and FNE. For any NE operator T, $Tx-Ty$ should fall into the black circle, while for FNE operator T, $Tx-Ty$ should fall into the blue circle. Lemma 2.1 below aims to explain this geometric intuition where it claims that FNE operator T can be decomposed as $\frac{1}{2}(I+N)$ where N is an NE operator.

Remark. In the operator splitting formulation of proximal gradient descent algorithm,

$$x^{t+1} = \underbrace{(I+\eta\partial h)^{-1}}_{\text{backward operator}} \underbrace{(I-\eta\nabla g)}_{\text{forward operator}} (x^t),$$

we hope both the forward and backward operator to be FNE to prove convergence results.

By Lemma 2.2, the forward splitting operator $I-\eta\nabla f$ is FNE if and only if $\eta\nabla f$ is FNE. In Lemma 2.3 and Theorem 2.4, we'll further show that $I-\eta\nabla f$ is FNE if and only if the convex function f is differentiable and L-smooth with $\eta \leq 1/L$.

The backward splitting operator $(I+\eta\partial h)^{-1}$ is FNE if and only if h is convex, which will be proved in Lemma [TODO].

Lemma 2.1. *An operator F is FNE if and only if there exists a NE operator N such that $F=(I+N)/2$.*

Proof. Denote $N=2F-I$. For any $(x,y),(x',y')\in F$, let $z=2y-x$ and $z'=2y'-x$, then we have $(x,z),(x',z')\in N$. Note that

$$\begin{aligned}\langle y-y',x-x'\rangle &= \frac{1}{2}\|x-x'\|_2^2+\frac{1}{2}\langle z-z',x-x'\rangle,\\ \|y-y'\|_2^2 &= \|\frac{1}{2}(x+z)-\frac{1}{2}(x'+z')\|_2^2\\ &= \frac{1}{4}\|x-x'\|_2^2+\frac{1}{4}\|z-z'\|_2^2+\frac{1}{2}\langle z-z',x-x'\rangle.\end{aligned}$$

We immediately have

$$\langle y-y',x-x'\rangle \geq \|y-y'\|_2^2 \Leftrightarrow \|z-z'\|_2^2 \leq \|x-x'\|_2^2,$$

which means F is FNE $\Leftrightarrow$ $2F-I$ is NE.

Lemma 2.2. *F is FNE if and only if $I-F$ is FNE.*

Proof. Using Lemma 2.1, we can argue that F is FNE $\Leftrightarrow$ $2F-I$ is NE $\Leftrightarrow$ $I-2F$ is NE $\Leftrightarrow$ $(I+I-2F)/2=I-F$ is FNE.

Lemma 2.3. *Let $f:\mathbb{R}^d\to\mathbb{R}$ be a convex and differentiable function. The following statements are equivalent.*

(1). NE of operator ∇f: $\|\nabla f(x)-\nabla f(y)\|_2\leq\|x-y\|_2$ for all $x,y\in\mathbb{R}^d$.
(2). $f(x)\geq f(y)+\nabla f(y)^\top(x-y)+\frac{1}{2}\|\nabla f(x)-\nabla f(y)\|_2^2$ for all $x,y\in\mathbb{R}^d$.
(3). FNE of operator ∇f: $\langle\nabla f(x)-\nabla f(y),x-y\rangle\geq\|\nabla f(x)-\nabla f(y)\|_2^2$ for all $x,y\in\mathbb{R}^d$.

Proof. (2) $\Rightarrow$ (3). By switching x and y in (2) we have

$$f(y)\geq f(x)+\nabla f(x)^\top(y-x)+\frac{1}{2}\|\nabla f(x)-\nabla f(y)\|_2^2.$$

Suming this up with (2) gives (3).

(3) $\Rightarrow$ (1). Using Cauchy-Schwarz inequality we have

$$\|\nabla f(x)-\nabla f(y)\|_2^2\leq\langle\nabla f(x)-\nabla f(y),x-y\rangle\leq\|\nabla f(x)-\nabla f(y)\|\|x-y\|,$$

which implies (1) when $\|\nabla f(x) - \nabla f(y)\|$ cancels out on both sides.

(1) $\Rightarrow$ (2). Fix any z' , denote $u(z) = f(z) - f(z') - \nabla f(z')^\top (z - z')$. Since f is convex, we have

$$\begin{cases} \nabla g(z) = \nabla f(z) - \nabla f(z'), \\ g(z) \geq 0. \end{cases}$$

By (1) we know that $\|\nabla g(z)\| \leq \|z - z'\|$, and hence

$$0 \leq g(z) \leq g(z') + \nabla g(z')^\top (z - z') + \frac{1}{2}\|z - z'\|_2^2.$$

Let's choose $z' = x$ and $z = x - \nabla f(x) + \nabla f(y)$ and replace them in the previous inequality, we have

$$\begin{aligned} 0 \leq & \underbrace{f(x) - f(y) - \nabla f(y)^\top (x - y)}_{g(z')} - \underbrace{\|\nabla f(x) - \nabla f(y)\|_2^2}_{\nabla g(z')^\top (z - z')} \\ & + \frac{1}{2}\underbrace{\|\nabla f(y) - \nabla f(x)\|_2^2}_{\|z - z'\|_2^2}, \\ & = f(x) - f(y) - \nabla f(y)^\top (x - y) - \frac{1}{2}\|\nabla f(x) - \nabla f(y)\|_2^2 \end{aligned}$$

which is exactly (2).

Equipped with the lemma we have, let's prove a theorem about forward splitting operator $T_F = I - \eta \nabla f$.

Theorem 2.4. *For any convex and differentiable function $f : \mathbb{R}^d \Rightarrow \mathbb{R}$,*

1. *$T_F = 1 - \eta \nabla f$ is FNE if and only if there exists $L \leq 1/\eta$ such that f is L-smooth;*
2. *if f is L-smooth with $L \leq 1/\eta$ and f α-strongly convex, then $\|T_F(x) - T_F(y)\|_2^2 \leq (1 - \alpha\eta)\|x - y\|_2^2$.*

Proof. Proof of (1)

$$\begin{aligned} T_F \text{ is FNE} & \iff \eta \nabla f \text{ is FNE (by Lemma 2.2)} \\ & \Leftrightarrow \eta \nabla f \text{is 1-Lipschitz (by Lemma 2.3)} \\ & \Leftrightarrow \text{ there exists } L \leq 1/\eta \text{ such that } f \text{ is } L\text{-smooth.} \end{aligned}$$

Proof of (2). If f is α strongly convex, we have

$$\begin{aligned} f(x) &\geq f(y) + \nabla f(y)^\top (x - y) + \frac{\alpha}{2}\|x - y\|^2, \\ f(y) &\geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2. \end{aligned}$$

Summing the above two equations up gives

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2.$$

By (1) we know T_F is FNE and thus $\eta\nabla f$ is FNE. We then cite Lemma 2.3 and claim that $\eta\nabla f$ is 1-Lipschitz, i.e.

$$\eta\|\nabla f(x)-\nabla f(y)\|_2 \leq \|x-y\|_2,$$

and thus

$$\langle \nabla f(y)-\nabla f(x), y-x\rangle \geq \alpha\eta\|\nabla f(x)-\nabla f(y)\|_2.$$

Now we know

$$\begin{aligned}
&\|T_F(y)-T_F(x)\|_2^2\\
&=\|y-x-\eta(\nabla f(y)-\nabla f(x))\|_2^2\\
&=\|y-x\|_2^2+\eta^2\|\nabla f(y)-\nabla f(x)\|_2^2-2\eta\langle\nabla f(y)-\nabla f(x),y-x\rangle\\
&\leq\|y-x\|_2^2+\eta^2\|\nabla f(y)-\nabla f(x)\|_2^2-2\alpha\eta^2\|\nabla f(y)-\nabla f(x)\|^2\\
&=\|y-x\|_2^2+(1-2\alpha)\eta^2\|\nabla f(y)-\nabla f(x)\|_2^2\\
&\leq\|y-x\|_2^2+(1-2\alpha)\eta\|y-x\|_2^2\\
&=(1+\eta-2\alpha\eta)\|y-x\|_2^2.
\end{aligned}$$

2.4 A Unified Operator Theory for Proving Algorithmic Convergence

Why we are interested in studying the operator theory? The main reason is that, combined with a fixed point argument, the operator theory provides a unified framework for us to establish the convergence theory of a large family of optimization algorithms (Though the operator theory is not very powerful at proving the exact rate of convergence). We will study the convergence of the iterative series

$$z^{t+1}=T\left(z^t\right), t\geq 1$$

where T belongs to a class of operator called 'averaged operator', which is a subclass of non-expansive operator and also includes the firmly nonexpansive operator as a special case.

Before we present this unified theory, let's start with the definition of averaged operator.

Definition 2.20 (Averaged operator). If an operator T can be written as $T=(1-\lambda)I+\lambda N$ where N is a nonexpansive operator and $0\leq\lambda\leq 1$, we call T an averaged operator.

Remark. A firmly nonexpansive operator must be an averaged operator (corresponding to choose $\lambda=1/2$ in the above definition).

Lemma 2.21. If T is nonexpansive, denote $T_\lambda=(1-\lambda)I+\lambda T$ for $\lambda\in(0,1]$. We can prove that

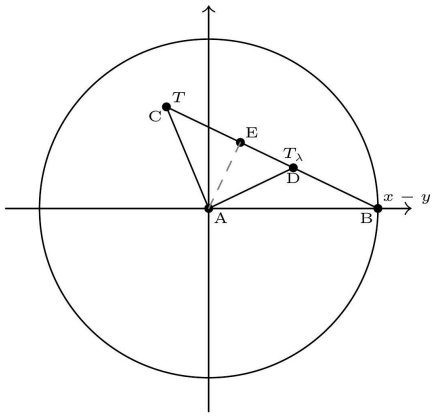

Fig. 2.3 Proof of Lemma 2.21

$$\|T_\lambda(x) - T_\lambda(y)\|_2^2 \leq \|x-y\|_2^2 - \frac{1-\lambda}{\lambda}\|(I-T_\lambda)(x) - (I-T_\lambda)(y)\|_2^2$$

Remark. This important lemma shows that averaging the non-expansive operator T can 'strengthen' the non-expansiveness. Instead of having $\|T(x)-T(y)\|_2^2 \leq \|x-y\|_2^2$ for T, the lemma shows that T_λ has a tighter contraction property than T with the negative term $-\frac{1-\lambda}{\lambda}\|(I-T_\lambda)(x)-(I-T_\lambda)(y)\|_2^2 = -(1-\lambda)\|(T-I)(x)-(T-I)(y)\|_2^2$ on the right-hand-side.

Proof. We will demonstrate the proof using Figure 2.21. In Figure 2.21, the point C represents $T(x)-T(y)$ and the point D represents $T_\lambda(x)-T_\lambda(y)$. Now we know that the length of the segments BD and CD represents these two quantities $|\mathrm{BD}| = \|(x-y)-(T\lambda(x)-T_\lambda(y))\|_2 = \|(I-T_\lambda)(x)-(I-T_\lambda)(y)\|_2$, $|\mathrm{CD}| = \frac{1-\lambda}{\lambda}|\mathrm{BD}|$, and thus

$$\frac{1-\lambda}{\lambda}\|(I-T_\lambda)(x)-(I-T_\lambda)(y)\|_2^2 = \mathrm{BD}\cdot\mathrm{DC}$$

Besides we have

$$|\mathrm{AD}| = \|T(x)-T(y)\|_2$$
$$|\mathrm{AB}| = \|x-y\|_2$$

We now only need to prove

$$\mathrm{AD}^2 + \mathrm{BD}\cdot\mathrm{DC} \leq \mathrm{AB}^2$$

Let E be on the line segment BC (or on its extension) such that $\mathrm{AE} \perp \mathrm{BC}$. Denote $h = |\mathrm{AE}|$. Now we have

$$\mathrm{BD}\cdot\mathrm{DC} = (\mathrm{BE}-\mathrm{DE})\cdot(\mathrm{DE}+\mathrm{EC}) = \mathrm{BE}\cdot\mathrm{EC} + (\mathrm{BE}-\mathrm{EC})\cdot\mathrm{DE} - \mathrm{DE}^2$$

Besides we have $\mathrm{AD}^2 = \mathrm{DE}^2 + h^2$ and $\mathrm{AB}^2 = \mathrm{BE}^2 + h^2$. We now have

$$\mathrm{AD}^2 + \mathrm{BD} \cdot \mathrm{DC} \leq \mathrm{AB}^2$$

which is equivalent to

$$h^2 + \mathrm{BE} \cdot \mathrm{EC} + (\mathrm{BE} - \mathrm{EC}) \cdot \mathrm{DE} \leq \mathrm{BE}^2 + h^2$$

This is further equivalent to

$$\begin{aligned}(\mathrm{BE} - \mathrm{DE})(\mathrm{BE} - \mathrm{EC}) \geq 0 \Leftrightarrow \mathrm{BD} \cdot (\mathrm{BE} - \mathrm{EC}) \geq 0 \Leftrightarrow |\mathrm{BE}| \geq |\mathrm{EC}| \\ \Leftrightarrow |\mathrm{AB}| \geq |\mathrm{AC}| \Leftrightarrow T \text{ is nonexpansive}\end{aligned}$$

based on which we conclude the proof of this lemma.
For any series $\{\lambda_t\}_{t\geq 1}$ where $\lambda_t \in (0,1], t \geq 1$, suppose the sequence $\{z^t\}_{t\geq 1}$ is generated by an iterative process where

$$z^{t+1} = T_{\lambda_t}\left(z^t\right), T_{\lambda_t} = (1-\lambda)I + \lambda T \tag{2.20}$$

Now let's present the main theorem of this section, based on which we will quickly establish the convergence of $\{z^t\}_{t\geq 1}$.

Theorem 2.22. Given a sequence $\{\lambda_t\}_{t\geq 1}$ where $\lambda_t \in (0,1], t \geq 1$ and a nonexpansive operator T, denote z^* as any fixed point of T such that $z^* = T(z^*)$. Denote $\tau_t = \lambda_t(1-\lambda_t)$. The sequence $\{z^t\}_{t\geq 1}$ is described in (2.20). Then we have

(1) The sequence $\left\{\|z^t - z^*\|_2^2\right\}_{t\geq 1}$ is non-decreasing.

(2) The fixed point residual is bounded by

$$\left\|T\left(z^k\right) - z^k\right\|_2^2 \leq \frac{\left\|z^0 - z^*\right\|_2^2}{\sum_{t=1}^k \tau_t}$$

Remark. The fixed point set of $T_{\lambda_t}, 0 < \lambda_t \leq 1$ is the same as the that of T because we have

$$T_{\lambda_t}(x^*) = x^* \Leftrightarrow (1-\lambda_t)x^* + \lambda_t T(x^*) = x^* \Leftrightarrow T(x^*) = x^*$$

Given Theorem 2.22, we can establish the convergence of sequence $\{z^t\}_{\geq 1}$ as in Proposition 2.23.

Proposition 2.23. Assuming $\sum_{t=1}^k \tau_t \to +\infty$ in Theorem 2.22 as $k \to \infty$, the sequence $\{z^t\}_{\geq 1}$ as described in equation (2.20) converges to some fixed point of T.

Proof. Based on (1) in Theorem 2.22, we know that

$$\left\|z^t - z^*\right\|_2^2 \leq \left\|z^0 - z^*\right\|_2^2$$

and thus there exists a subsequence of $\{z^t\}_{t\geq 1}$ converging to some $\widetilde{z}$. Let's denote the subsequence as $\{\widetilde{z}^t\}_{t\geq 1}$. Using the statement about fixed point residual (2) in Theorem 2.22, we know that

$$\left\|T\left(\widetilde{z}^k\right)-\widetilde{z}^k\right\|_2^2 \leq \frac{\left\|z^0-z^*\right\|_2^2}{\sum_{t=1}^k \tau_t} \to 0$$

because of the assumption $\sum_{t=1}^k \tau_t \to +\infty$. Let $k \to \infty$, we have $\|T(\widetilde{z})-\widetilde{z}\| = 0$ which means that $\widetilde{z}$ is a fixed point of T. Apply Theorem 2.22 again to the fixed $\widetilde{z}$, we know that $\|z^t-\widetilde{z}\|$ is non-decreasing and a subsequence of $\{z^t\}_{t\geq 1}$ converges to $\widetilde{z}$, and thus we conclude that the sequence $\{z^t\}_{t\geq 1}$ converges to $\widetilde{z}$.

Here we demonstrate some examples of using Theorem 2.22 to prove the convergence of optimization algorithms.

Example 1. Gradient Descent. Suppose function f is L-smooth, proper, closed and convex, and the step size $\eta \leq 1/L$, we know the forward operator $T_F = (I+\eta\nabla f)$ is firmly non-expansive, which means T_F is an averaged operator $T_F = (I+N)/2$ for some non-expansive N. Gradient descent iteration can be written as $x^{t+1} = T_F(x^t)$. Let $\lambda_t = 1/2$ in Theorem 2.22 and we conclude that the sequence $\{x^t\}_{t\geq 1}$ converges to some fixed point of N, which is also a fixed point of T_F, and a global optimizer of convex function f.

Example 2. Proximal point algorithm. The proximal point iteration can be written as $x^{t+1} = \text{prox}_{\eta f}(x^t)$. The proximal operator $\text{prox}_{\eta f}$ is firmly nonexpansive for convex function f. We can conclude the convergence using the same line of reasoning as Example 1.

Example 3. Proximal gradient descent. Consider function $f = g+h$ where g is L-smooth and h is proper, closed and convex. Suppose the step size $\eta \leq 1/L$. We have proven before that the forward operator $T_F = (I+\eta\nabla g)$ and backward operator $T_B = (1+\eta\partial h)^{-1}$ are both firmly non-expansive. The proximal gradient descent iteration is

$$x^{t+1} = T_B \circ T_F\left(x^t\right)$$

where $T_B \circ T_F$ is actually an averaged operator because

$$T_B = \frac{1}{2}(I+N_B), \quad T_F = \frac{1}{2}(I+N_F),$$
$$T_B \circ T_F = \frac{1}{4}I + \frac{3}{4}\frac{N_F+N_B+N_B\circ N_F}{3} = \frac{1}{4}I+\frac{3}{4}N$$

Choosing $\lambda_t = 3/4$ in Theorem 2.22 can give us the convergence result of $\{x^t\}_{t\geq 1}$. Now we present the proof of Theorem 2.22.

Proof of Theorem 2.22. (1) Using the definition of the iteration, we have

$$\begin{aligned}\left\|z^{t+1}-z^*\right\|_2^2 &= \left\|T_{\lambda_t}\left(z^t\right)-T_{\lambda_t}\left(z^*\right)\right\|_2^2 \\ &\leq \left\|z^t-z^*\right\|_2^2 - \frac{1-\lambda_t}{\lambda_t}\left\|\left(I-T_{\lambda_t}\right)\left(z^t\right)-\left(I-T_{\lambda_t}\right)\left(z^*\right)\right\|_2^2 \\ &= \left\|z^t-z^*\right\|_2^2 - \frac{1-\lambda_t}{\lambda_t}\left\|z^t-T_{\lambda_t}\left(z^t\right)\right\|_2^2 \end{aligned} \tag{2.21}$$

where (2.21) is based on Lemma 2.21 and equation (2.22) is based on the fact that z^* is also a fixed point for any T_{λ_t} with $\lambda_t \in (0,1]$. From equation (2.22) we can conclude that the series $\left\{\|z^t - z^*\|_{t\geq 1}\right.$ is non-decreasing.
(2) Since $z^{t+1} = (1-\lambda_t)z^t + \lambda_t T\left(z^t\right)$, we know that

$$z^{t+1} - z^t = \lambda_t \left(T\left(z^t\right) - z^t\right)$$

We then have

$$\begin{aligned}\tau_t \left\|T\left(z^t\right) - z^t\right\|_2^2 = \frac{1-\lambda_t}{\lambda_t}\left\|z^{t+1} - z^t\right\|_2^2 = \frac{1-\lambda_t}{\lambda_t}\left\|T_{\lambda_t}\left(z^t\right) - z^t\right\|_2^2 \\ \leq \left\|z^t - z^*\right\|_2^2 - \left\|z^{t+1} - z^*\right\|_2^2\end{aligned}$$

where the last inequality is from equation (2.22). Summing the above equation from 1 to k gives us

$$\sum_{t=0}^{k} \tau_t \left\|T\left(z^t\right) - z^t\right\|_2^2 \leq \left\|z^0 - z^*\right\|_2^2 - \left\|z^{k+1} - z^*\right\|_2^2 \tag{2.23}$$

We also notice that

$$\begin{aligned}\left\|T\left(z^{t+1}\right) - z^{t+1}\right\|_2 &\leq \left\|T\left(z^{t+1}\right) - \left[(1-\lambda_t)z^t + \lambda_t T\left(z^t\right)\right]\right\|_2 \\ &\leq \left\|T\left(z^{t+1}\right) - T\left(z^t\right)\right\|_2 + (1-\lambda_t)\left\|T\left(z^t\right) - z^t\right\|_2 \\ (T \text{ is non-expansive }) &\leq \left\|z^{t+1} - z^t\right\|_2 + (1-\lambda_t)\left\|T\left(z^t\right) - z^t\right\|_2 \\ &= \left\|T\left(z^t\right) - z^t\right\|_2 \end{aligned} \tag{2.24}$$

Combining (2.23) and (2.24) we have

$$\left(\sum_{t=0}^{k} \tau_t\right)\left\|T\left(z^k\right) - z^k\right\|_2^2 \leq \sum_{t=0}^{k} \tau_t \left\|T\left(z^t\right) - z^t\right\|_2^2 \leq \left\|z^0 - z^*\right\|_2^2$$

by which we directly prove (2) in Theorem 2.22.

Chapter 3
Peaceman-Rachford Operator Splitting Algorithm Family

In this lecture we continue under the operator splitting framework and discuss another class of operator splitting methods, Peaceman-Rachford operator splitting family. We will show how to directly apply the operator fixed point iteration convergence theorem (Theorem 2.22) to derive convergence results for the Peaceman-Rachford algorithm family. One important algorithm within the Peaceman-Rachford family is the Alternating Direction Method of Multipliers (ADMM), which will be discussed at the end of the lecture.

When applying operator splitting methods to the optimization problem

$$\min f(x) = \min g(x) + h(x)$$

where g and h are convex, closed and proper, there are some general guidelines to follow.

- Step 1. We define non-expansive operators T_g and T_h based on g and h respectively.
- Step 2. We check if there exits a mapping from the fixed point of $T_g \circ T_h$ into the zero set of $\partial g + \partial h$, i.e. we hope there exists a mapping M such that

$$\left\{M(z) \mid T_g \circ T_h(z) = z\right\} \subseteq \text{zer}(\partial g + \partial h)$$

For any relation R, the zero set of R is defined as zer $(R) = \{x \mid (x,0) \in R\}$.

- Step 3. We can run the fixed point iteration $z^{t+1} = (T_g \circ T_h)_{\lambda_t}(z^t)\, T$ times until the desired accuracy is achieved.
- Step 4. We calculate $M\left(z^T\right)$ as the approximation to the minimizer of the optimization problem.

We will demonstrate this strategy in the proof of Peaceman-Rachford algorithm later.

3.1 Relaxed Peaceman-Rachford Operator Splitting Algorithm

The Peaceman-Rachford operator is a composition of two 'reflection operator's.

Definition 3.1 (Reflection Operator). For any convex, closed and proper function f and $\eta > 0$, the reflection operator is defined as

$$\text{Refl}_{\eta f}(z) = \left(2\text{Prox}_{\eta f} - I\right)(z)$$

Since $\text{Prox}_{\eta f}$ is a firmly non-expansive operator, according to Lemma 2.15, we know that Refl_{nf} is a non-expansive operator.

Relaxed Peaceman-Rachford Algorithm

Given a series $\{\lambda_t\}_{t\geq 0}, \lambda_t \in (0,1], \eta > 0$ and initialization z^0, relaxed Peaceman-Rachford iteration is defined by

$$z^{t+1} = (T_{\text{PRS}})_{\lambda_t}\left(z^t\right) = (1-\lambda_t)z^t + \lambda_t T_{\text{PRS}}\left(z^t\right)$$

where $T_{\text{PRS}} = \text{Refl}_{\eta g} \circ \text{Refl}_{\eta h}$.

When $\lambda_t = 1/2$, the above algorithm is also called Douglas-Rachford Operator Splitting Algorithm.

To prove the convergence of the above algorithm, we will follow the general strategy presented at the beginning. Since Refl_{nf} and $\text{Refl}_{\eta f}$ are nonexpansive, we can use them as T_f and T_g in Step 1. Lemma 3.2 shows that the proximal operator can be used as the mapping from the fixed point of the Peaceman-Rachford algorithm into the set of optimizers, as suggested by Step 2 of our strategy.

Lemma 3.2. We can prove that

$$\text{zer}(\partial g + \partial h) = \left\{\text{Prox}_{\eta h}(z) \mid z \in \mathbb{R}^d, T_{\text{PRS}}(z) = z\right\}$$

Proof.

$$\begin{aligned}
&0 \in \partial g(x) + \partial h(x) \\
&\Leftrightarrow \exists y \text{ s.t. } x - y \in \eta\partial g(x), y - x \in \eta\partial h(x) && (3.1)\\
&\Leftrightarrow \exists y \text{ s.t. } x = \text{Prox}_{\eta g}\left[\text{Refl}_{\eta h}(y)\right], x = \text{Prox}_{\eta h}(y) && (3.2)\\
&\Leftrightarrow \exists y \text{ s.t. } 2x - \text{Refl}_{\eta h}(y) = 2\,\text{Prox}\left[\text{Refl}_{\eta h}\right](y) - \text{Refl}_{\eta h}(y) && (3.3)\\
&\text{and } 2x - \text{Refl}_{\eta h}(y) = 2\,\text{Prox}(y) - \left[2\,\text{Prox}_{\eta h}(y) - y\right] = y && (3.4)\\
&\text{and } x = \text{Prox}_{\eta h}(y) && (3.5)
\end{aligned}$$

$$\Leftrightarrow \exists y \text{ s.t. } y = T_{\text{PRS}}(y) \qquad (3.6)$$

Now let's justify the line of reasoning as follows.

- Equation (3.1). The $\Rightarrow$ in (3.1) is trivial and we can justify the $\Leftarrow$ in (3.1) as follows. If $0 \in \partial g(x) + h(x)$, we know there exists $a \in \partial g(x), b \in \partial h(x)$ such that $a + b = 0$. Denote $y = x - \eta a$. We have $x - y = \eta a \in \eta \partial g(x)$ and $y - x = -\eta b \in \eta \partial h(x)$.
- Equation (3.2). By definition $x = \text{Prox}_{\eta h}(y)$ is equivalent to $y - x \in \eta \partial h(x)$ in (3.1). We also know that

$$\begin{aligned} x - y \in \eta \partial g(x) &\Leftrightarrow \text{Refl}_{\eta h}(y) = 2x - y \in (I + \eta \partial g)(x) \\ &\Leftrightarrow x = \text{Prox}_{\eta g} \text{Refl}_{\eta h}(y) \end{aligned}$$

  which justifies (3.2).

- Equation (3.3), (3.4) and (3.5) directly follows from (3.2).

Theorem 3.3 (Convergence of Relaxed Peaceman-Rachford Algorithm). The sequence $\left\{\text{Prox}_{\eta h}\left(z^t\right)\right\}_{t \geq 0}$ converges to an optimizer x^* where

$$x^* \in \arg\min f(x) = \arg\min g(x) + h(x)$$

and $\{z^t\}_{t \geq 0}$ is defined by the Relaxed Peaceman-Rachford iteration.

Proof. Since T_{PRS} is a non-expansive operator, we conclude that the sequence $\{z^t\}_{t \geq 0}$ in Peaceman-Rachford algorithm converges to a fixed point z^* of T_{PRS} according to Theorem 2.22 and Proposition 2.23.

According to Lemma 3.2, we know that $\text{Prox}_{\eta h}(z^*) \in \text{zer}(\partial g + \partial h)$, and because g and h are convex, closed and proper, we have

$$\text{Prox}_{\eta h}(z^*) = x^* \in \arg\min f(x) = \arg\min g(x) + h(x).$$

The proximal operator $\text{Prox}_{\eta h}$ is continuous since h is a continuous function. And thus we know that $\text{Prox}_{\eta h}(z^t)$ converges to $\text{Prox}_{\eta h}(z^*) = x^*$.

3.2 Convergence Rates of Relaxed Peaceman-Rachford Algorithm

Lemma 3.4. Denote $x_h^t = z^t - \eta \nabla h\left(x_h^t\right)$ and $x_g^t = x_h^t - \eta \nabla h\left(x_h^t\right) - \eta \nabla g\left(x_g^t\right)$. We have the relationship

$$z^{t+1} - z^t = 2\lambda_t \left(x_g^t - x_h^t\right) = -2\lambda_t \eta \left[\nabla h\left(x_h^t\right) + \nabla g\left(x_g^t\right)\right]$$

Proof. We know from the definition of x_g^t and x_h^t that

$$\text{Refl}_{\eta h}\left(z^t\right) = (I + \eta \partial g) x_g^t = x_h^t - \eta \nabla\left(x_h^t\right)$$

$$= 2\,\text{Prox}_{\eta h}\left(z^t\right) - z^t = 2x_h^t - z^t$$

and $x_g^t + \eta \partial g\left(x_g^t\right) = x_h^t - \eta \nabla h\left(x_h^t\right) = 2x_h^t - z^t$, which implies that

$$x_g^t = \text{Prox}_{\eta g}\left(2x_h^t - z^t\right)$$

Therefore we have

$$\begin{aligned}\text{Refl}_{\eta g} \circ \text{Refl}_{\eta h}\left(z^t\right) &= \text{Refl}_{\eta g}\left(2x_h^t - z^t\right) \\ &= 2\text{Prox}_{\eta g}\left(2x_h^t - z^t\right) - \left(2x_h^t - z^t\right) = 2x_g^t - \left(2x_h^t - z^t\right)\end{aligned}$$

and thus

$$\begin{aligned}z^{t+1} - z^t &= \lambda_t\left(\text{Refl}_{\eta g} \circ \text{Refl}_{\eta h}\left(z^t\right) - z^t\right) = 2\lambda_t\left(x_g^t - x_h^t\right) \\ &= -2\lambda_t\left[\nabla h\left(x_h^t\right) + \nabla g\left(x_g^t\right)\right]\end{aligned}$$

Roadmap this time. Step 1. points with knwon derivatives

$$\begin{aligned}x_h^t &= \text{Prox}_{\eta h}\left(z^t\right) \\ x_g^t &= \text{Prox}_{\eta g}\left(\text{Refl}\left(z^t\right)\right)\end{aligned}$$

Step 2. Convexity

$$g\left(x_g^t\right) + h\left(x_h^t\right) - g\left(x^t\right) - h\left(x^*\right) \leq \ldots$$

Step 3. Regularity condition

$$\begin{aligned}g\left(x_h^t\right) - g\left(x_g^t\right) &\leq \ldots \\ h\left(x_g^t\right) - h\left(x_h^t\right) &\leq \ldots\end{aligned}$$

Step 4. Telescoping.
T_{PRS} is NE, we apply the averaged operator convergence theorem and Lemma 3.5. If x^* exists, then $\tau_t = \lambda_t\left(1 - \lambda_t\right)$ and
$-\left\|z^t - z^*\right\|_2^2$ is non-increasing.
$-\left\|T_{PRS}\left(z^t\right) - z^t\right\|_2^2$ non-increasing.
$-\sum_{t=1}^k \tau_t \left\|T_{PRS}\left(z^t\right) - z^t\right\|_2^2 \leq \left\|z^t - z^*\right\|_2^2$
$-\left\|T_{PRS}\left(z^k\right) - z^k\right\|_2^2 \leq \left(\left\|z^0 - z^*\right\|_2^2\right) / \left(\sum_{t=0}^k \tau_t\right)$
Main results for convergence rate

$$\bar{x}^k = \frac{\sum_{t=0}^k \lambda_t x^t}{\sum_{t=0}^k \lambda_t}$$

and $\tau_t = \lambda_t\left(1 - \lambda_t\right)$. x^t is chosen as x_g^t or x_h^t in different cases. $x^* = \text{Prox}_{\eta h}\left(z^*\right)$.
Case (a). g is Q-Lipschitz $x^t = x_h^t$; or h is Q-Lipschitz and $x^t = x_g^t$. Ergodic result

$$f\left(\bar{x}^*\right) = f\left(x^*\right) \leq \left[\frac{1}{4\eta}\left\|z^0 - x^*\right\|_2^2 + Q\left\|z^0 - z^*\right\|_2\right] \frac{1}{\sum_{t=0}^k \lambda_t}$$

although the PRS is not symmetric in g or h , this result is symmetric in g/h. Non-ergodic result

$$f\left(x^t\right)-f\left(x^*\right)\leq\frac{1}{2\eta}\left[\left\|z^*-z^*\right\|_2+\left\|z^*-x^*\right\|_2+Q\eta\right]\frac{\left\|z^0-z^*\right\|_2}{\sqrt{\sum_{t=0}^k\tau_t}}$$

$$f\left(x^k\right)-f\left(x^*\right)=o\left(\frac{1}{\sqrt{\sum_{t=[k/2]+1}^k\tau_t}}\right)$$

Case (b). No assumption of L smoothness. g is α-strongly convex, $x^t=x_g^t$. h is α-strongly convex, $x^t=x_h^t$.

$$\left\|x_{\text{best}}^k-x^*\right\|_2^2\leq\frac{\left\|z^0-z^*\right\|_2^2}{4\alpha\eta}\frac{1}{\sum_{t=0}^k\lambda_t}$$

By strongly convexity. we have $f(x)-f\left(x^*\right)\geq\frac{\alpha}{2}\left\|x-x^*\right\|_2^2$ which means the decision variable convergence is easier than function value convergence.

$$x_{\text{best}}^*=\arg\min_{x_1,x_2,\ldots x_k}\left\|x^t-x^*\right\|_2^2$$

$$\left\|\bar{x}^k-x^*\right\|_2^2\leq\frac{\left\|z^0-z^t\right\|_2^2}{4\alpha\eta}\frac{1}{\sum_{t=0}^k\lambda_t}$$

$$\left\|x^k-x^*\right\|_2^2\leq\frac{\left\|z^0-z^*\right\|_2^2}{2\alpha\eta}\frac{1}{\sqrt{\sum_{t=0}^k\tau_t}}$$

Case (c) g is L-smooth $x^t=x_h^t$ or h is L-smooth $x^t=x_g^t$.

$$f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\leq\frac{\eta L+1}{2}\left[1+\frac{1}{\inf t\geq 0\left(1-\lambda_t\right)/\lambda_t}\right]\frac{\left\|z^0-z^*\right\|_2^2}{4\eta\sum_{t=0}^k\lambda_t}$$

$$f\left(\bar{x}^k\right)-f\left(x^*\right)\leq\frac{\eta L+1}{2}\left[1+\frac{1}{\inf t\geq 0\left(1-\lambda_t\right)/\lambda_t}\right]$$

Case (c') h is L-smooth. $\lambda_t=1/2.\eta<\kappa/L$ where $\kappa\approx 1.24698$ is the positive root of $x^3+x^2-2x-1.x^t=x_g^t$ and we have

$$f\left(x^k\right)-f\left(x^*\right)\leq C\cdot\frac{1}{2\eta(k+1)}$$

where

$$C=\left\|x_h^0-x^*\right\|_2^2+\frac{1}{1+(\eta L)^2}\left[(\eta L)^3-2\eta L-1\right]\left\|z^0-z^*\right\|_2^2$$

Case (d) $1.g$ is L-smooth, α-strongly convex

$$C(\lambda) = 1 - \frac{\lambda}{2} \min\left\{ \frac{4\eta\alpha}{(1+\eta L)^2}, 1-\lambda \right\}$$

2. h is L-smooth and α-strongly convex

$$C(\lambda) = 1 - 4\eta\lambda\alpha/(1+\eta L)^2$$

3. One is L-smooth and the other is α-strongly convex

$$C(\lambda) = 1 - \frac{4\lambda}{3} \min\left\{ \eta\alpha, \frac{1}{\eta L}, 1-\lambda \right\}$$

Proof of Case (a).

$$S_f(x,y) = \max\left\{ \frac{\alpha_f}{2}\|x-y\|_2^2, \frac{1}{2L_f}\|\nabla f(x) - \nabla f(y)\|_2^2 \right\}$$

$L_f = \infty$ if f is not L-smooth.
Then

$$g(x) \geq g(y) + \langle x-y, \nabla g(y)\rangle + S_g(x,y)$$
$$h(x) \geq h(y) + \langle x-y, \nabla h(y)\rangle + S_h(x,y)$$

The above result is from the relationship we've known before in the proof of the equivalence of NE and FNE operator.

$$\|\nabla f(x) - \nabla f(y)\| \leq \|x-y\| \Leftrightarrow$$
$$f(x) \geq f(y) = \nabla f(y)^\top (x-y) + \frac{1}{2}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Lemma 3.6 (Upper Fundamental Inequality). df
For PPA

$$f\left(x^t\right) - f\left(x^*\right) + S_f\left(x^t, x^*\right) \leq \frac{1}{2\eta}\left[\left\|x^t - x^0\right\|_2^2 - \left\|x^{t+1} - x^0\right\|_2^2 - \|\ldots\|_2^2\right]$$

Lemma 3.7 (Lower Fundamental Inequality).

$$g\left(x_g^t\right) + h\left(x_h^t\right) - g\left(x^*\right) - h\left(x^*\right) \geq \frac{1}{\eta}\left\langle x_h^t - x_g^t, z^* - x^*\right\rangle + S_g\left(x_g^t, x^*\right) + S_h\left(x_h^t, x^*\right)$$
$$4\lambda_t\eta\left[g\left(x_g^t\right) + h\left(x_h^t\right) - g\left(x^*\right) - h\left(x^*\right) + S_g\left(x_g^t, x^*\right) + S_h\left(x_h^t, x^*\right)\right]$$
$$\leq \left\|z^t - x^*\right\|_2^2 - \left\|z^{t+1} - x^*\right\|_2^2 + (1 - 1/\lambda_t)\left\|z^{t+1} - z^t\right\|_2^2$$
$$= \left\|z^t - z^*\right\|_2^2 - \left\|z^{t+1} - z^*\right\|_2^2 + 2\left\langle z^t - z^{t+1}, z^* - x^*\right\rangle + (1 - 1/\lambda_t)\left\|z^{t+1} - z^t\right\|_2^2$$

We want to lower bound $g\left(x_g^t\right)+h\left(x_h^t\right)-g\left(x^*\right)-h\left(x^*\right)$ by terms of decision variables.

Proof.

$$g\left(x_g^t\right)-g\left(x^*\right)\geq\left\langle x_g^t-x^*,\nabla g\left(x^*\right)\right\rangle+S_g\left(x_g^t,x^*\right)$$
$$h\left(x_h^t\right)-h\left(x^*\right)\geq\left\langle x_h^t-x^*,\nabla h\left(x^*\right)\right\rangle+S_h\left(x_h^t,x^*\right)$$

If we sum the above two equations together and use the relationship $\nabla g\left(x^*\right)=-\nabla h\left(x^*\right)$ and $z^*=x^*+\eta\nabla h\left(x^*\right)$, we will get the result we want.

Lemma 3.8 (Trick). For $\lambda\in(0,1]$ and $z_t=\left(T_{PRS}\right)_\lambda\left(z^t\right)$. For some term $A^t=A\left(x_g^t,x_h^t,x^*\right),B^t=B\left(x_g^t,x_h^t,x^*\right)$. If for all $\lambda\in(0,1]$,

$$\lambda A^t\leq\left\|z^t-z^*\right\|_2^2-\left\|z_\lambda-z^*\right\|_2^2+\left(1-\frac{1}{\lambda}\right)\left\|z_\lambda-z^t\right\|_2^2$$
$$\lambda B^t\leq\left\|z^t-x^*\right\|_2^2-\left\|z_\lambda-x^*\right\|_2^2+\left(1-\frac{1}{\lambda}\right)\left\|z_\lambda-z^t\right\|_2^2$$

Then we have $\bar{A}^k$,

$$\bar{A}^k,A_{\text{best}}^k\leq\frac{\left\|z^0-z^*\right\|_2^2}{\sum_{t=0}^k\lambda_t}$$
$$A^k\leq2\left\|z^0-z^*\right\|_2\cdot\left\|z^k-T_{PRS}\left(z^k\right)\right\|_2^2$$

3.3 ADMM

For the optimization problem,

$$\min_x f(x)=\min_x g(x)+h(x)$$

we have the relaxed Peaceman-Rachford Splitting algorithm, i.e.

$$z^{t+1}=\left(T_{PRS}\right)_\lambda\left(z^t\right)$$
$$T_{PRS}=\text{Refl}_{\eta g}\circ\text{Refl}_{\eta h}$$

For the optimization problem with linear constaints

$$\begin{aligned}\min\nolimits_{x\in\mathbb{R}^{d_1},y\in\mathbb{R}^{d_2}}\ & f(x,y)=g(x)+h(y)\\ \text{such that }& Ax+By=b\end{aligned}$$

We can define the Lagrangian dual

$$L(x,y,w)=g(x)+h(y)-w^\top(Ax+By-b)$$
$$\min_{x,y}L(x,y,w)=-g^*\left(A^\top w\right)-h^*\left(B^\top w\right)+w^\top b$$

Let's further define

$$d_g(w) = g^*\left(A^\top w\right)$$
$$d_h(w) = h^*\left(B^\top w\right) - w^\top b$$

We can derive the relaxed ADMM algorithm when we apply the relaxed-PRS to

$$\min_w d_g(w) + d_h(w)$$

First of all, remember the chain rule

$$\partial d_g(w) = A\partial g^*\left(A^\top w\right)$$
$$\partial d_h(w) = \partial h^*\left(B^\top w\right) - b$$

Lemma 3.9. If f is closed and convex, we have

$$\underbrace{w \in \partial f(x)}_{(1)} \Leftrightarrow \underbrace{x \in \partial f^*(w)}_{(2)} \Leftrightarrow \underbrace{x^\top w = f(x) + f^*(w)}_{(3)}$$

Proof. (1) $\Leftrightarrow$ (3).

$$\begin{aligned} w \in \partial f(x) &\Leftrightarrow \forall y, f(y) - f(x) \geq w^\top (y - x) \\ &\Leftrightarrow \forall y, w^\top y - f(y) \leq w^\top x - f(x) \\ &\Leftrightarrow \langle w, x\rangle - f(x) = \sup_y w^\top y - f(y) = f^*(w) \end{aligned}$$

(2) $\Leftrightarrow$ (3). $f = f^{**}$.

Lemma 3.10 (Primal Representation Lemma).

$$\begin{aligned} & w = \mathrm{Prox}_{\eta d_g}(z) \\ \Leftrightarrow\ & x = \arg\min_y g(y) - \langle z, Ay\rangle + \frac{\eta}{2}\|Ay\|_2^2 \\ & \text{and } w = z - \eta Ax \end{aligned}$$

and also

$$\begin{aligned} & w = \mathrm{Prox}_{\eta_d}(z) \\ \Leftrightarrow\ & x = \arg\min_y h(y) - \langle z, Ay\rangle + \frac{\eta}{2}\|Ay - b\|_2^2 \\ & \text{and } w = z - \eta(Ax - b) \end{aligned}$$

Remark. This lemma uses primal variable representation to help with the dual calculation.

Proof.

$$
\begin{aligned}
& w = \text{Prox}_{\eta d_g}(z) \\
\Leftrightarrow & z \in (I + \eta \partial d_g) w \\
\Leftrightarrow & z \in w + \eta A \partial g^* \left(A^\top w \right) \\
\Leftrightarrow & \exists x \in \partial g^* \left(A^\top w \right), w = z - \eta A x \\
\Leftrightarrow & \exists x, s.t. A^\top w \in \partial g(x), w = z - \eta A x \\
\Leftrightarrow & \exists x, 0 \in \partial g(x) - A^\top w, w = z - \eta A x \\
\Leftrightarrow & \exists, 0 \in \partial g(x) - A^\top z + \eta A^\top A x, w = z - \eta A x \\
\Leftrightarrow & x = \arg\min_y g(y) - \langle z, Ay \rangle + \frac{\eta}{2} \|Ay\|_2^2, w = z - \eta A x
\end{aligned}
$$

Proposition 3.11. Let $\{z^t\}_{t \geq 0}$ be generated by relaxed PRS applied to $\min_w d_g(w) + d_h(w)$. Choose initial points $w_{d_h}^{(-1)} = z^0, x^{(-1)} = 0, y^{(-1)} = 0$ and $\lambda_{(-1)} = 1/2$. Then from $t = -1$, we have

$$
\begin{aligned}
y^{t+1} = & \arg\min_y h(y) - \left\langle w_{d_h}^t, Ax^t + By - b \right\rangle \\
& \frac{\eta}{2} \left\| Ax^t + By - b + (2\lambda_t - 1) \left(Ax^t + By^t - b \right) \right\|_2^2 \qquad (3.7) \\
w_{d_h}^{t+1} = & w_{d_h}^t - \eta \left(Ax^t + By^{t+1} - b \right) - \eta \left(2\lambda_t - 1 \right) \left(Ax^t + By^t - b \right) \qquad (3.8) \\
x^{t+1} = & \arg\min_x g(x) - \left\langle w_{d_h}^{t+1}, Ax + By^{t+1} - b \right\rangle + \frac{\eta}{2} \left\| Ax + By^{t+1} - b \right\|_2^2 \qquad (3.9)
\end{aligned}
$$

Remark. ADMM is not a primal-dual algorithm. The primal variable shows up only to calculate the dual, but not to help calculate primal function value.

Proof. Note that the following update equations

$$
\begin{aligned}
W_{d_h}^t &= \text{Prox}_{\eta d_h} \left(z^t \right) \\
W_{d_g}^t &= \text{Prox}_{n d_g} \left(2 w_{d_n}^t - z^t \right) \\
z^{t+1} &= z^t + 2\lambda_t \left(w_{d_g}^t - w_{d_h}^t \right)
\end{aligned}
$$

are equivalent to

$$
z^{t+1} = (\text{Refl} \circ \text{Refl})_{\lambda_t} \left(z^t \right)
$$

First of all we have

$$
2 w_{d_h}^t - z^t = 2 \text{Prox}_{\eta d_h} \left(z^t \right) - z^t = \text{Refl}_{\eta d_h} \left(z^t \right)
$$

We also have

$$
\begin{aligned}
& z^{t+1} - (1-\lambda_t) z^t + \lambda_t \left(z^t + 2 \left(w_{d_g}^t - w_{d_h}^t \right) \right) \\
&= (1-\lambda_t) z^t + \lambda_t \left(2 \operatorname{Prox}_{\eta d_g} \left(\operatorname{Refl}_{\eta d_h} \left(z^t \right) \right) + z^t - 2 w_{d_h}^t \right. \\
&= (1-\lambda_t) z^t + \lambda_t \operatorname{Ref}_{\eta_{d_g}} \left(\operatorname{Ref}_{\eta d_n} \left(z^t \right) \right) \\
&= \ldots
\end{aligned}
$$

Then we know

$$
\begin{aligned}
y^t &= \arg\min_y h(y) - \left\langle z^t, By - b \right\rangle + \frac{\eta}{2} \|By - b\|_2^2 \\
w_{d_h}^t &= z^t - \eta \left(By^t - b \right) \\
x^t &= \arg\min_x g(x) - \left\langle 2 w_{d_h}^t - z^t, Ax \right\rangle + \frac{\eta}{2} \|Ax\|_2^2 \\
w_{d_g}^t &= 2 w_{d_g}^t - z^t - \eta A x^t \\
z^{t+1} &= z^t + 2\lambda_t \left(w_{d_g}^t - w_{d_h}^t \right)
\end{aligned}
$$

and thus

$$
\begin{aligned}
z^{t+1} &= z^t + 2\lambda_t \left(w_{d_g}^t - w_{d_h}^t \right) \\
&= z^t + w_{d_g}^t - w_{d_h}^t + (2\lambda_t - 1) \left(w_{d_g}^t - w_{d_h}^t \right) \\
&= z^t + w_{d_g}^t - w_{d_h}^t + \eta (2\lambda_t - 1) \left(Ax^t + By^t - b \right) \\
&= w_{d_h}^t - \eta A x^t - \eta (2\lambda_t - 1) \left(Ax^t + By^t - b \right)
\end{aligned}
$$

We also have

$$
\begin{aligned}
y^{t+1} &= \arg\min_y h(y) - \left\langle z^{t+1}, By - b \right\rangle + \frac{\eta}{2} \|By - b\|_2^2 \\
&= \arg\min_y h(y) - \left\langle w_{d_h}^t - \eta A x^t - \eta (2\lambda_t - 1) \left(Ax^t + By^t - b \right), By - b \right\rangle \\
&+ \frac{\eta}{2} \|By - b\|_2^2
\end{aligned}
$$

which implies the first equation (3.7) in the proposition.
Besides, if we plug in the equation

$$
z^{t+1} = w_{d_h}^t - \eta A x^t - \eta (2\lambda_t - 1) \left(Ax^t + By^t - b \right)
$$

into

$$
w_{d_h}^t = z^t - \eta \left(By^t - b \right)
$$

we can have the second equation (3.8).

3.4 A Unified ADMM by Majorization and Minimization

Goal

$$\min f(x) = f(x_1, x_2, \ldots, x_n), \text{ such that } Ax = \sum_{i \geq 0} A_i x_i = b$$

There are two types in ADMM in general, Gauss-Seidel ADMM and Jacobian ADMM.

$$\text{Guass-Seidel } x_i^{k+1} = \arg\min_{x_i} \widehat{L}\left(x_1^{k+1}, \ldots x_{i-1}^{k+1}, x_{i+1}^{k}, \ldots, x_n^{k}\right)$$
$$\text{Jacobian } x_i^{k+1} = \arg\min_{x_i} \widehat{L}\left(x_1^{k}, \ldots x_{i-1}^{k}, x_{i+1}^{k}, \ldots x_n^{k}\right)$$

Define

$$L(x, \lambda, \beta) = f(x) + \langle \lambda, Ax - b \rangle + \frac{\beta}{2} \|Ax - b\|_2^2$$

Augmented Lagrangian Method (ALM) is the following iteration equations

$$r^k(x) = \frac{\beta^k}{2} \left\| Ax - b + \lambda^k / \beta^k \right\|_2^2$$
$$x^{k+1} = \arg\min_x L\left(x, \lambda^k, \beta^k\right) = \arg\min_x f(x) + r^k(x)$$
$$\lambda^{k+1} = \lambda^+ \beta^k \left(Ax^{k+1} - b\right)$$

There are two types of interpretations of ADMM, i.e. Alternating Minimization (AM) and Majorization Minimization (MM). Gauss-Seidel ADMM can be explained by AM followed by MM. Jacobian ADMM can be explained by MM followed by AM.

3.4.1 Standard ADMM

Standard ADMM is just the ALM above, which appeared in the 1960s.

$$x_1^{k+1} = \arg\min_{x_1} L\left(\left[x_1, x_2^k\right], \lambda^k, \beta^k\right) = \arg\min f_1(x_1) + r_1^k(x_1)$$
$$x_2^{k+1} = \arg\min_{x_2} L\left(\left[x_1^{k1+}, x_2\right], \lambda^k, \beta^k\right) = \arg\min f_2(x_2) + r_2^k(x_2)$$
$$r_1^k(x_1) = \frac{\beta^k}{2} \|A_1 x_1 + A_2 x_2 - b + \lambda_k / \beta_k\|_2^2$$
$$r_2^k(x_2) = \frac{\beta^k}{2} \|A_1 x_1^{k+1} + A_2 x_2 - b + \frac{\lambda^k}{\beta^k}$$

3.4.2 Proximal ADMM

Assume f_1 is L_1-smooth

$$\widehat{f}_1(x_1) = f\left(x_1^k\right) + \left\langle \nabla f\left(x_1^k, x_1 - x_1^k\right)\right\rangle + \frac{L_1}{2}\left\|x_1 - x_1^k\right\|_2^2$$

We have $\widehat{f}_1(x_1) \geq f_1(x_1)$. For the case $f_1 = g_1 + h_1$ and h_1 is L_1-smooth, we have

$$\widehat{f}_1(x_1) = g_1(x_1) + h_1\left(x_1^k\right) + \left\langle \nabla h_1\left(x_1^k\right), x_1 - x_1^k\right\rangle + \frac{L_1}{2}\left\|x_1 - x_1^k\right\|_2^2$$

3.4.3 Linearized ADMM

By linearizing the function $\widehat{r}_1^k$, we have

$$x_1^{k+1} = \arg\min_{x_1} f_1(x_1) + \widehat{r}_1^k(x_1)$$

where

$$\widehat{r}_1^k(x_1) = r_1^k\left(x_1^k\right) + \left\langle \nabla r_1^k\left(x_1^k\right), x_1 - x_1^k\right\rangle + \frac{\eta}{2}\left\|x_1 - x_1^k\right\|_2^2$$

and $\eta \geq \|A_1\|_2^2$. This formulation is called linearized ADMM.

3.4.4 Linearized ADMM with Parallel Splitting

The linearized ADMM with parallel splitting updates x_i in parallel, i.e.
$x_i^{k+1} = \arg\min_{x_i} f_i(x_i) + \left\langle A_i^\top\left(\beta^{(k)}\left(Ax^k - b\right) + \lambda^k\right), x_i\right\rangle + \frac{\beta^{(k)}\eta_i}{2}\left\|x_i - x_i^k\right\|_2^2$, for $\eta_i > n\|A_i\|_2^2$.

3.4.5 Proximal Linearized ADMM Parallel Splitting

We use $\widehat{f}_i$ to upper bound f_i and update x_i in parallel,

$$x_i^{k+1} = \arg\min_{x_i} \widehat{f}_i(x_i) + \left\langle A_i^\top\left(\beta^{(k)}\left(Ax^k - b\right) + \beta^k\right), x_i\right\rangle + \frac{\beta^{(k)}\eta_i}{2}\left\|x_i - x_i^k\right\|_2^2$$

Questions.

- What kind of majorant functions can be used in ADMMs?
- Is that possible to give a unified convergence analysis of ADMMs which use different majorant functions by using certain common properties of majorant functions?
- What is the connection between the convergence speed of ADMMs and the used majorant functions?
- How to choose the proper majorant functions for designing efficient ADMMs?

3.4.6 Majorant functions

Definition 3.12 (Lipschitz Continuity). Function f is $\{\mathbf{L}_i\}_{i=1}^n$-smooth if $\mathbf{L}_i \succeq 0$ and

$$\|f(x) - f(y) - \langle \nabla f(y), x-y\rangle\| \leq \frac{1}{2}\sum_{i=1} \|x_i - y_i\|_2^2$$

Definition 3.13 (Strong Convexity). A function $f : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \ldots \mathbb{R}^{p_n} \Rightarrow \mathbb{R}$ is called $\{\mathbf{P}_i\}_{i=1}^n$ strongly convex if there exists $\mathbf{P}_i \succeq 0$ such that the function $f(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_i\|_{\mathbf{P}_i}$ is strongly-convex, i.e.

$$f(x) \geq f(y) + \langle f(x), x-y\rangle + \frac{1}{-}\sum_{i=1} \|x_i - y_i\|_{\mathbf{P}_i}$$

Definition 3.14 (Majorant First-Order Surrogate). A function $\widehat{f} : \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_n} \to \mathbb{R}$ is a majorant first-order surrogate of $f : \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_n} \to \mathbb{R}$ near $\boldsymbol{\kappa} = [\boldsymbol{\kappa}_1; \cdots; \boldsymbol{\kappa}_n]$ with $\boldsymbol{\kappa}_i \in \mathbb{R}^{p_i}$ when the following conditions are satisfied:

- Majorization. $\widehat{f}$ is a majorant function of f, i.e., $\widehat{f}(\mathbf{x}) \geq f(\mathbf{x})$ for any $\mathbf{x}$. - Proximity. there exists $\boldsymbol{L}_i \succeq \mathbf{0}$ such that the approximation error $h(\mathbf{x}) := \widehat{f}(\mathbf{x}) - f(\mathbf{x})$ satisfies

$$|h(\mathbf{x})| \leq \frac{1}{2}\sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\kappa}_i\|_{L_i}^2 \tag{3.10}$$

- Separability. $\widehat{f}$ is separable w.r.t. $\mathbf{x}_i$'s; i.e., there exist $\widehat{f}_i$'s such that $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{f}_i(\mathbf{x}_i)$.

We denote by $\mathscr{S}_{\{\boldsymbol{L}_i, \boldsymbol{P}_i\}_{i=1}^n}(f, \boldsymbol{\kappa})$ the set of $\{\boldsymbol{P}_i\}_{i=1}^n$-strongly convex surrogates.

Lemma 3.15. If the approximation error $h(\mathbf{x}) = \widehat{f}(\mathbf{x}) - f(\mathbf{x})$ satisfies the following Smoothness assumption, i.e.,

$$h(\boldsymbol{x}) \text{ is } \{\boldsymbol{L}_i\}_{i=1}^n\text{-smooth, } h(\boldsymbol{\kappa}) = 0 \text{ and } \nabla h(\boldsymbol{\kappa}) = 0 \tag{3.11}$$

then the Proximity assumption in (3.10) holds.

Examples of majorant surrogates.

- If f is separable, we have $\widehat{f} = f$.
- Proximal surrogate. f is convex, $\mathbf{L} \succeq 0$. Let

$$\widehat{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \boldsymbol{\kappa}\|_{\mathbf{L}}^2$$

We have $\widehat{f} \in S_{\{\mathbf{L},\mathbf{L}\}}(f,k)$.

- Linearized Gradient Surrogate. Function f is convex and $\{\mathbf{L}_i\}_{i=1}^n$ smooth.

$$\widehat{f}(\mathbf{x}) = f(\boldsymbol{\kappa}) + \langle \nabla f(\boldsymbol{\kappa}), x - \boldsymbol{\kappa}\rangle + \frac{1}{2}\sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\kappa}_i\|_{\mathbf{L}_i}^2$$

We have $\widehat{f} \in S_{\{\mathbf{L}_i,\mathbf{L}_i\}}(f,\boldsymbol{\kappa})$

- Proximal Gradient Surrogate. $f = f_1 + f_2$ where f_1 is convex and $\{\mathbf{L}_i\}_{i=1}^n$ - smooth and f_2 is convex. Let

$$\widehat{f}(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\boldsymbol{\kappa}) + \langle \nabla f_2(\boldsymbol{\kappa}), \mathbf{x} - \boldsymbol{\kappa}\rangle + \frac{1}{2}\sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\kappa}_i\|_2^2$$

3.4.7 Unified Gauss-Seidel ADMM

If we only have two blocks, Guass-Seidel ADMM is usually faster than Jacobian ADMM. Here we consider the case for $n = 2$ where there're only two blocks.

Algorithm 1 A Unified Framework of Gauss-Seidel ADMMs For $k = 0, 1, 2, \cdots$ do

1. Compute a majorant first-order surrogate $\widehat{f}^k \in \mathscr{S}_{\{\boldsymbol{L}_i,\boldsymbol{P}_i\}_{i=1}^2}\left(f, \mathbf{x}^k\right)$ with $\widehat{f}^k(\mathbf{x}) = \widehat{f}_1^k(\mathbf{x}_1) + \widehat{f}_2^k(\mathbf{x}_2)$.
2. Update $\mathbf{x}_1$ by solving

$$\begin{aligned}\mathbf{x}_1^{k+1} &= \arg\min_{\mathbf{x}_1} \widehat{f}_1^k(\mathbf{x}_1) + \widehat{r}_1^k(\mathbf{x}_1) \\ &= \arg\min_{\mathbf{x}_1} \widehat{f}_1^k(\mathbf{x}_1) + r_1^k(\mathbf{x}_1) + \frac{\beta^k}{2}\left\|\mathbf{x}_1 - \mathbf{x}_1^k\right\|_{\mathbf{G}_1}^2\end{aligned}$$

3. Update $\mathbf{x}_2$ by solving

$$\begin{aligned}\mathbf{x}_2^{k+1} &= \arg\min_{\mathbf{x}_2} \widehat{f}_2^k(\mathbf{x}_2) + \widehat{r}_2^k(\mathbf{x}_2) \\ &= \arg\min_{\mathbf{x}_2} \widehat{f}_2^k(\mathbf{x}_2) + r_2^k(\mathbf{x}_2) + \frac{\beta^k}{2}\left\|\mathbf{x}_2 - \mathbf{x}_2^k\right\|_{\mathbf{G}_2}^2\end{aligned}$$

4. Update λ by $\lambda^{k+1} = \lambda^k + \beta^k\left(\mathbf{A}\mathbf{x}^{k+1} - \boldsymbol{b}\right)$.
5. Choose $\beta^{k+1} \geq \beta^k$.
   end

The following table shows that previous Gauss-Seidel ADMMs are special cases of Algorithm 1 with different $\widehat{f}_1$ and $\mathbf{G}_1$. In this table, $\eta_1 > \|\mathbf{A}_1\|_2^2$.

Table 3.1 Previous Gauss-Seidel ADMMs are special cases of Algorithm (1).

	$\widehat{f}_1^k(\mathbf{x}_1)$	$\mathbf{G}_1$
ADMM	$f_1(\mathbf{x}_1)$	$\mathbf{0}$
P-ADMM	Lipschitz Gradient Surrogate or Proximal Gradient Surrogate	$\mathbf{0}$
L-ADMM	$f_1(\mathbf{x}_1)$	$\eta_1 I - \mathbf{A}_1^\top \mathbf{A}_1$
PL-ADMM	Lipschitz Gradient Surrogate or Proximal Gradient Surrogate	$\eta_1 I - \mathbf{A}_1^\top \mathbf{A}_1$

Assume that there exists a KKT point (x^*, λ^*), i.e. $Ax^* = b$ and $-A^\top \lambda^* \in \partial f(x^*)$. Theorem 3.16. In Algorithm 1, assume that $\widehat{f}^k \in \mathscr{S}_{\{\boldsymbol{L}_i, \boldsymbol{P}_i\}_{i=1}^{Z_1}}\left(f, \mathbf{x}^k\right)$ with $\boldsymbol{P}_i \succeq \boldsymbol{L}_i \succeq \mathbf{0}, i = 1,2, \mathbf{G}_1 \succeq \mathbf{0}$ in (??), and $\mathbf{G}_2 \succ \mathbf{0}$ in (??). For any $K > 0$, let $\overline{\mathbf{x}}^K = \sum_{t=0}^K \gamma^t \mathbf{x}^{t+1}$ with $\gamma^k = \left(\beta^k\right)^{-1} / \sum_{k=0}^K \left(\beta^k\right)^{-1}$. Then

$$\begin{aligned}&f\left(\overline{\mathbf{x}}^K\right) - f(\mathbf{x}^*) + \left\langle \mathbf{A}^\top \lambda^*, \overline{\mathbf{x}}^K - \mathbf{x}^* \right\rangle + \frac{\beta^{(0)}\alpha}{2}\left\|\mathbf{A}\overline{\mathbf{x}}^K - \boldsymbol{b}\right\|^2 \\ \leq &\frac{\sum_{i=1}^2 \left\|\mathbf{x}_i^* - \mathbf{x}_i^0\right\|_{\mathbf{H}_i^0}^2 + \left\|\lambda^* - \lambda^0\right\|_{\mathbf{H}_3^0}^2}{2\sum_{k=0}^K \left(\beta^k\right)^{-1}}\end{aligned} \tag{3.12}$$

where $\alpha = \min\left\{\frac{1}{2}, \frac{\sigma_{\min}^2(\mathbf{G}_2)}{2\|\mathbf{A}_2\|_2^2}\right\}, \mathbf{H}_1^0 = \frac{1}{\beta^{(0)}}\boldsymbol{L}_1 + \mathbf{G}_1, \mathbf{H}_2^0 = \frac{1}{\beta^{(0)}}\boldsymbol{L}_2 + \mathbf{A}_2^\top \mathbf{A}_2 + \mathbf{G}_2$, and $\mathbf{H}_3^0 = \left(1/\beta^{(0)}\right)^2 \boldsymbol{I}$.

3.4.8 Unified Jacobian ADMM

Algorithm 2 A Unified Framework of Jacobian ADMMs
For $k = 0, 1, 2, \cdots$ do

1. Compute a majorant first-order surrogate $\widehat{f}^k \in \mathscr{S}_{\{\boldsymbol{L}_i, \boldsymbol{P}_i\}_{i=1}^n}\left(f, \mathbf{x}^k\right)$ with $\widehat{f}^k(\mathbf{x}) = \sum_{i=1}^n \widehat{f}_i^k(\mathbf{x}_i)$

2. Update $\mathbf{x}_i, i = 1,2,\ldots n$, in parallel by solving

$$\begin{aligned}\mathbf{x}_i^{k+1} &= \arg\min_{\mathbf{x}_i} \widehat{f}_i^k(\mathbf{x}_i) + \widehat{r}_i^k(\mathbf{x}_i) \\ &= \arg\min_{\mathbf{x}_i} \widehat{f}_i^k(\mathbf{x}_i) + \frac{\beta^k}{2}\left\|\mathbf{A}_i\mathbf{x}_i + \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \boldsymbol{b} + \frac{\lambda^k}{\beta^k}\right\|^2 \\ &+ \frac{\beta^k}{2}\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|_{\mathbf{G}_i}^2 + \beta^k c_i^k\end{aligned}$$

3. Update λ by $\lambda^{k+1} = \lambda^k + \beta^k\left(\mathbf{A}\mathbf{x}^{k+1} - \boldsymbol{b}\right)$.
4. Choose $\beta^{k+1} \geq \beta^k$.

end

Chapter 4
Stochastic Gradient Descent

We consider the same unconstrained optimization problem as the problem in Chapter 1 for gradient descent.

$$\min_{x\in\mathbb{R}^d} f(x) \tag{4.1}$$

where $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ is convex and may not be differentiable. There exists x^* such that

$$f(x^*) = \min_{x\in\mathbb{R}^d} f(x)$$

4.1 Basics of Stochastic Gradient Descent

The stochastic gradient descent algorithm is similar to gradient descent. In stead of accessing the accurate 'oracle' gradient information at every step x_t, we only have a stochastic but unbiased version of the gradient.

Stochastic (sub-)Gradient Descent Method
For $t = 1,\ldots,k$, iterates

$$x^{t+1} = x^t - \eta_t g\left(x^t, \xi^{t+1}\right) \tag{4.2}$$

until converges, where $\xi^{t+1}, t = 1,\ldots,k$ are independent random variables and the stochastic gradient is unbiased,

$$\mathbb{E}_\xi[g(x,\xi)] = \nabla f(x), \text{ or } \mathbb{E}_\xi[g(x,\xi)] \in \partial f(x)$$

for all $x \in \mathbb{R}^d$.

4.1.1 Geometric Conditions of the Objective Functions

Geometric conditions also have an huge impact on the choice of step size for stochastic gradient descent. Since we don't have direct access to the oracle,
the stochastic versions of geometric conditions are somewhat different from the deterministic case we discussed in Chapter 1.

Condition 4.1 (Stochastic Q-Lipschitz). This condition requires the square norm of the stochastic gradient is bounded in expectation, i.e.

$$\mathbb{E}_\xi \left[\|g(x,\xi)\|_2^2\right] \leq Q^2$$

Comparison with deterministic counterpart. Q-Lipschitz condition for convex function f in gradient descent says that for all $g \in \partial f, \|g\|_2^2 \leq Q^2$.

Condition 4.2 (Stochastic (L,σ)-smoothness). This condition requires that the underlying function f is L-smooth and the stochasticity of the gradient has bounded variance.

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2 \tag{4.3}$$

$$\mathbb{E}_\xi \left[\|g(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2, \text{ for all } x \tag{4.4}$$

Remark. Combining unbiasedness of $g(x,\xi)$ and equation (4.4), we have the bound on the second moment of $g(x,\xi)$ here

$$\begin{aligned}
\sigma^2 &\geq \mathbb{E}_\xi \left[\|g(x,\xi) - \nabla f(x)\|_2^2\right] \\
&= \mathbb{E}_\xi \left[\|g(x,\xi)\|_2^2 - 2\langle g(x,\xi), \nabla f(x)\rangle + \|\nabla f(x)\|_2^2\right] \\
&= \mathbb{E}_\xi \left[\|g(x,\xi)\|_2^2\right] - 2\left\langle \mathbb{E}_\xi[g(x,\xi)], \nabla f(x)\right\rangle + \|\nabla f(x)\|_2^2 \\
\text{(unbiasedness)} \quad &= \mathbb{E}_\xi \left[\|g(x,\xi)\|_2^2\right] - 2\langle \nabla f(x), \nabla f(x)\rangle + \|\nabla f(x)\|_2^2 \\
&= \mathbb{E}_\xi \left[\|g(x,\xi)\|_2^2\right] - \|\nabla f(x)\|_2^2
\end{aligned}$$

and thus we have

$$\mathbb{E}_\xi \left[\|g(x,\xi)\|_2^2\right] \leq \|\nabla f(x)\|_2^2 + \sigma^2 \tag{4.5}$$

which is very similar to the stochastic Q-Lipschitz condition.

Condition 4.3 (α-strongly Convexity). This condition is the same deterministic version as before,

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\alpha}{2}\|y-x\|_2^2, \text{ for all } x, y$$

where $\alpha \geq 0$ and the condition reduces to convexity when $\alpha = 0$.

4.2 Theory of Stochastic Gradient Descent

We will list the convergence rate of stochastic gradient descent under different geometric conditions and compare the results with gradient descent. Denote $R = \|x_1 - x^*\|$. Let k be the total number of iteration. For any sequence $\{x^t\}_{t=1,\ldots,k}$ produced by the stochastic gradient descent algorithm, we denote

$$\bar{x}^k = \frac{\sum_{t=1}^k \eta_t x^t}{\sum_{t=1}^k \eta_t}, \text{ and }, x_{\text{best}}^k = \underset{x^t,(t=1,\ldots,k)}{\operatorname{argmin}} f\left(x^t\right) \tag{4.6}$$

Option (a). Suppose we have stochastic Q-Lipschitz condition and α strongly convexity with $\alpha = 0$.
We can achieve the $1/\sqrt{k}$ rate for function value convergence, which is the same as the rate for the gradient descent under the same conditions, i.e.

$$\mathbb{E}\left[f\left(\bar{x}^k\right) - f\left(x^*\right)\right] \leq \frac{RQ}{\sqrt{k}}, \text{ and } \mathbb{E}\left[f\left(x_{\text{best}}^k\right) - f\left(x^*\right)\right] \leq \frac{RQ}{\sqrt{k}}$$

with non-adaptive constant step size $\eta_t = R/(Q\sqrt{k})$.
Since we do not have strongly convex $\alpha = 0$ here, the optimizer x^* may not be unique so we do not study the convergence rate of decision variables.
Option (b). Suppose we have stochastic Q-Lipschitz condition and α strongly convexity with $\alpha > 0$. We can have $1/k$ rate for decision variable convergence and $1/k$ rate for function value convergence, which are the same rates as the gradient descent under the same conditions, i.e.

$$\mathbb{E}\left[f\left(x_{\text{best}}^k\right) - f\left(x^*\right)\right] \leq \frac{2Q^2}{\alpha(k+1)}$$
$$\mathbb{E}\left[\left\|x^k - x^*\right\|_2^2\right] \leq \frac{8Q^2}{\alpha^2 k}$$

with adaptive diminishing step size $\eta_t = 2/(\alpha(t+1))$.
Remark. We will see later that the proof here for Option (a) and Option (b) is almost the same with the proof for gradient descent.
Option (c). Suppose we have stochastic L-smoothness and α-strongly convexity with $\alpha = 0$. The rate of function value convergence is $1/\sqrt{k}$, which is worse than the $1/k$ rate in gradient descent under the same conditions, i.e.

$$\mathbb{E}\left[f\left(\bar{x}^k\right) - f\left(x^*\right)\right] \leq \frac{R\sigma}{\sqrt{k}}, \text{ and } \mathbb{E}\left[f\left(x_{\text{best}}^k\right) - f\left(x^*\right)\right] \leq \frac{R\sigma}{\sqrt{k}}$$

which non-adaptive constant step size $\eta_t = R/(\sigma\sqrt{k})$.
Option (d). Suppose we have stochastic (L,σ)-smoothness, α-strongly convexity with $\alpha > 0$. Here we no longer have the linear convergence rate as we do in the gradient descent setting. Instead, we only have $1/k$ rate for
both decision variable and function value convergence, i.e.

$$\mathbb{E}\left[f\left(x_{\text{best}}^{k}\right)-f\left(x^{*}\right)\right] \leq \frac{4\sigma^{2}\alpha}{(k+1)+4\alpha/L}$$
$$\mathbb{E}\left[\left\|x^{k+1}-x^{*}\right\|_{2}^{2}\right] \leq \frac{4k\alpha\sigma^{2}}{t+2\alpha/L}$$

with adaptive diminishing step size $\eta_t = 2\alpha/(t+2\alpha/L)$.
Before we prove the results above, let's exploit convexity to get an interface based on which we can plug in other assumptions. Using convexity of f, we have

$$f\left(x^{*}\right) \geq f\left(x^{t}\right)+\nabla f\left(x^{t}\right)^{\top}\left(x^{*}-x^{t}\right)+\frac{\alpha}{2}\left\|x^{t}-x^{*}\right\|_{2}^{2}$$

where $\alpha \geq 0$. Let $\{\mathscr{F}_t\}_{1\leq t\leq k}$ be the filtration generated by the stochastic process $\{x^t\}_{1\leq t\leq k}$. By taking conditional expectations we have

$$\begin{aligned}
&\mathbb{E}\left[f\left(x^{t}\right)-f\left(x^{*}\right) \mid \mathscr{F}_{t}\right] \\
&\leq \mathbb{E}\left[-\frac{\alpha}{2}\left\|x^{t}-x^{*}\right\|_{2}^{2}+\left\langle\frac{x^{t+1}-x^{t}}{\eta_{t}}, x^{*}-x^{t}\right\rangle \middle| \mathscr{F}_{t}\right] && (4.7)\\
&=\mathbb{E}\left[-\frac{\alpha}{2}\left\|x^{t}-x^{*}\right\|_{2}^{2}+\frac{1}{2\eta_{t}}\left(\left\|x^{t}-x^{*}\right\|+\left\|x^{t+1}-x^{t}\right\|_{2}^{2}-\left\|x^{t+1}-x^{*}\right\|\right) \middle| \mathscr{F}_{t}\right] && (4.8)\\
&=\frac{1}{2\eta_{t}}\mathbb{E}\left[\left(1-\alpha\eta_{t}\right)\left\|x^{t}-x^{*}\right\|_{2}^{2}-\left\|x^{t+1}-x^{*}\right\|_{2}^{2} \mid \mathscr{F}_{t}\right]+\frac{1}{2\eta_{t}}\mathbb{E}\left[\left\|x^{t+1}-x^{t}\right\|_{2}^{2} \mid \mathscr{F}_{t}\right] && (4.9)
\end{aligned}$$

where (4.7) follows directly from the unbiasedness of stochastic gradient, i.e.

$$\nabla f\left(x^{t}\right)=\mathbb{E}\left[g\left(x^{t}, \xi\right)\right]=\mathbb{E}\left[\left(x^{t+1}-x^{t}\right)/\eta_{t}\right]$$

and (4.8) is a direct consequence of Pythagorean theorem

$$2\langle a, b\rangle=\frac{\|a\|_{2}^{2}+\|b\|_{2}^{2}-\|a-b\|_{2}^{2}}{2}$$

with $a = x^{t+1}-x^t$ and $b = x^*-x^t$.
Remark. Note that the key equation (4.9), which will be used as the interface for convergence rate analysis for $\alpha = 0$ and $\alpha > 0$ later, has the same structure as the key equation (1.12) in the analysis for gradient descent. Some of the techniques in gradient descent case can be seamlessly applied here to derive the convergence rates and figure out the proper choice of step sizes.

4.2.1 Convergence under stochastic Q-Lipschitz conditions

Here we prove Option (a) where we have stochastic Q-Lipschitz condition and α-strongly convexity with $\alpha = 0$. Starting from equation (4.9), the proof is almost the same as the gradient descent case.
The stochastic Q-Lipschitz condition gives us

$$\frac{1}{2\eta_t}\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|_2^2\right] = \frac{1}{2\eta_t}\mathbb{E}\left[\eta_t^2\left\|g\left(x^t,\xi^{t+1)}\right)\right\|_2^2\right] \leq \frac{\eta_t}{2}Q^2$$

Plug in this equation into the key interface (4.9) with $\alpha = 0$, we have

$$\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\mid\mathscr{F}_t\right] \leq \frac{1}{2\eta_t}\mathbb{E}\left[\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^*\right\|_2^2\mid\mathscr{F}_t\right]+\frac{\eta_t}{2}Q^2$$

Taking expectation over the entire sequence $x_1,\ldots,x_t.$, we can have

$$\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right] \leq \frac{1}{2\eta_t}\mathbb{E}\left[\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^*\right\|_2^2\right]+\frac{\eta_t}{2}Q^2$$

For $\bar{x}^k = \sum_{t=1}^k \eta_t x^t / \sum_{t=1}^k \eta_t$, Jensen's inequality for f gives the result

$$\begin{aligned}\mathbb{E}\left[f\left(\bar{x}^k\right)-f\left(x^*\right)\right] &\leq \frac{1}{\sum_{t=1}^k\eta_t}\sum_{t=1}^k\mathbb{E}\left[\eta_t\left(f\left(x_t\right)-f\left(x^*\right)\right)\right]\\ &\leq \left(\frac{1}{2}\mathbb{E}\left[\left\|x^1-x^*\right\|_2^2\right]+\frac{Q^2}{2}\sum_{t=1}^k\eta_t^2\right)/\left(\sum_{t=1}^k\eta_t\right)\\ &= \frac{R^2+\left(\sum_{t=1}^k\eta_t^2\right)Q^2}{2\sum_{t=1}^k\eta_t}\end{aligned}$$

where $R = \mathbb{E}\left[\left\|x^1-x^*\right\|_2^2\right]$. For x_{best}^k in the notation (4.6), we also have the same results as $\bar{x}^k$,

$$\begin{aligned}\left(\sum_{t=1}^k\eta_t\right)\mathbb{E}\left[f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\right] &\leq \sum_{t=1}^k\eta_t\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right]\\ &\leq \frac{1}{2}\mathbb{E}\left[\left\|x^1-x^*\right\|_2^2\right]+\frac{Q^2}{2}\sum_{t=1}^k\eta_t^2\end{aligned}$$

and thus

$$\mathbb{E}\left[f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\right] \leq \frac{R^2+\left(\sum_{t=1}^k\eta_t^2\right)Q^2}{2\sum_{t=1}^k\eta_t}$$

Choosing non-adaptive constant step size $\eta_t = R/(Q\sqrt{k})$, we have our results

$$\mathbb{E}\left[f\left(\bar{x}^k\right)-f\left(x^*\right)\right] \leq \frac{RQ}{\sqrt{k}}$$
$$\mathbb{E}\left[f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\right] \leq \frac{RQ}{\sqrt{k}}$$

4.2.2 Convergence under stochastic Q-Lipschitz conditions and α-strong convexity

Starting from the key interface (4.9) and Q-Lipschitz, we have

$$\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right] \leq \frac{1}{2\eta_t}\mathbb{E}\left[\left(1-\alpha\eta_t\right)\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^*\right\|_2^2\right]+\frac{\eta_t}{2}Q^2 \quad (4.10)$$

As we have done before in the gradient descent setting, we choose $\eta_t = 2/(\alpha(t+1))$ to have a 'more cleaned' telescoping
$\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right] \leq \frac{\alpha}{4}\mathbb{E}\left[(t-1)\left\|x^t-x^*\right\|_2^2-(t+1)\left\|x^{t+1}-x^*\right\|_2^2\right]+\frac{Q^2}{\alpha(t+1)}$.
Denote $b_t = \mathbb{E}\left[\left\|x^t-x^*\right\|_2^2\right]$ as we have done before in gradient descent. We now have

$$t\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right] \leq \frac{\alpha}{4}\mathbb{E}\left[(t-1)tb_t-t(t+1)b_{t+1}\right]+\frac{2tQ}{\alpha(t+1)} \quad (4.11)$$

and thus

$$\sum_{t=1}^{k} t\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right] \leq \frac{\alpha}{4}\mathbb{E}\left[0-k(k+1)b_{k+1}\right]+\frac{2Q}{\alpha}\sum_{t=1}^{k}\frac{t}{t+1} \leq \frac{2kQ}{\alpha}$$

based on which we have

$$\mathbb{E}\left[f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\right] \leq \frac{2kQ^2}{\alpha k(k+1)} = \frac{2Q^2}{\alpha(k+1)}$$

From equation (4.11), using the fact that $\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\right] \geq 0$, we have

$$\frac{\alpha}{4}\mathbb{E}\left[(t-1)tb_t-t(t+1)b_{t+1}\right]+\frac{2Q}{\alpha(t+1)} \geq 0$$

Summing the above equation from 1 to k gives us

$$\frac{\alpha}{4}\left(0-\mathbb{E}\left[(k+1)kb_{k+1}\right]+\frac{2Q}{\alpha}\sum_{t=1}^{k}\frac{t}{t+1} \geq 0\right.$$

from which we conclude the decision variable convergence

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|_2^2\right] \leq \frac{8Q}{\alpha^2 k(k+1)} \sum_{t=1}^{k} \frac{t}{t+1} \leq \frac{8kQ}{k(k+1)\alpha^2} = \frac{8Q}{(k+1)\alpha^2}$$

4.2.3 Convergence under stochastic (L,σ)-smoothness condition

The condition of L-smoothness gives us the equation

$$\begin{aligned} f\left(x^{t+1}\right) &\leq f\left(x^t\right) + \left\langle \nabla f\left(x^t\right), x^{t+1}-x^t \right\rangle + \frac{L}{2}\left\|x^{t+1}-x^t\right\|_2^2 \\ &\leq f\left(x^t\right) - \eta_t \left\langle \nabla f\left(x^t\right), g\left(x^t,\xi^{t+1}\right)\right\rangle + \frac{L}{2}\eta_t^2 \left\|g\left(x^t,\xi_{t+1}\right)\right\|_2^2 \end{aligned}$$

Taking expectation on both sides and using equation (4.5) and unbiadness of $g\left(x^t,\xi^{t+1}\right)$, we have

$$f\left(x^{t+1}\right) - f\left(x^t\right) \leq -\eta_t \left\|\nabla f\left(x^t\right)\right\|_2^2 + \frac{L\eta_t^2}{2}\left(\left\|\nabla f\left(x^t\right)\right\|_2^2 + \sigma^2\right)$$

Assuming that $\eta_t \leq 1/L$, we further know that

$$\begin{aligned} f\left(x^{t+1}\right) - f\left(x^t\right) &\leq -\eta_t \left\|\nabla f\left(x^t\right)\right\|_2^2 + \frac{\eta_t}{2}\left(\left\|\nabla f\left(x^t\right)\right\|_2^2 + \sigma^2\right) \\ &= -\frac{\eta_t}{2}\left\|\nabla f\left(x^t\right)\right\|_2^2 + \frac{\eta_t}{2}\sigma^2 \end{aligned} \tag{4.12}$$

Remark. In the proof for gradient descent under L-smoothness condition, we had this equation

$$f\left(x^{t+1}\right) - f\left(x^t\right) \leq -\frac{\eta_t}{2}\left\|f\left(x^t\right)\right\|_2^2$$

Stochasticity only introduces an additional $\eta_t\sigma^2/2$ term compared to gradient descent! We will now plug (4.12) into the key interface (4.9) and get

$$\begin{aligned}
&\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\mid\mathscr{F}_t\right]\\
&\leq\frac{1}{2\eta_t}\mathbb{E}\left[\left(1-\alpha\eta_t\right)\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^t\right\|_2^2\mid\mathscr{F}_t\right]+\frac{\eta_t}{2}\mathbb{E}\left[\left\|g\left(x^t,\xi_{t+1}\right)\right\|_2^2\mid\mathscr{F}_t\right]\\
&\leq\frac{1}{2\eta_t}\mathbb{E}\left[\left(1-\alpha\eta_t\right)\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^t\right\|_2^2\mid\mathscr{F}_t\right]+\frac{\eta_t}{2}\left(\left\|\nabla f\left(x^t\right)\right\|_2^2+\sigma^2\right) \qquad (4.13)\\
&\leq\frac{1}{2\eta_t}\mathbb{E}\left[\left(1-\alpha\eta_t\right)\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^t\right\|_2^2\mid\mathscr{F}_t\right]+f\left(x^t\right)-f\left(x^{t+1}\right)+\frac{\eta_t}{2}\sigma^2 \qquad (4.14)
\end{aligned}$$

where (4.13) comes from (4.5) and (4.14) comes from (4.12).
From (4.14), we now have

$$\begin{aligned}
&\mathbb{E}\left[f\left(x^t\right)-f\left(x^*\right)\mid\mathscr{F}_t\right]\\
&\leq\frac{1}{2\eta_t}\mathbb{E}\left[\left(1-\alpha\eta_t\right)\left\|x^t-x^*\right\|_2^2-\left\|x^{t+1}-x^t\right\|_2^2\right]+\frac{\eta_t}{2}\sigma^2 \qquad (4.15)
\end{aligned}$$

In gradient descent, we don't have the $\eta_t\sigma^2/2$ term caused by stochastic noise and we can achieve $1/k$ convergence rate for $\alpha=0$. Here we can only expect to have $1/\sqrt{k}$ sub-linear rate because equation (4.15) has the same structure as equation (4.10) in Q-Lipschitz case with Q replaced by σ.
Using the same line of proof for Q-Lipschitz, we can have

$$\mathbb{E}\left[f\left(\bar{x}^k\right)-f\left(x^*\right)\right]\leq\frac{R^2+\left(\sum_{t=1}^k\eta_t^2\right)\sigma^2}{2\sum_{t=1}^k\eta_t}$$

and

$$\mathbb{E}\left[f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\right]\leq\frac{R^2+\left(\sum_{t=1}^k\eta_t^2\right)\sigma^2}{2\sum_{t=1}^k\eta_t}$$

For $k\geq\sigma^2/\left(L^2R^2\right)$, with non-adaptive step-size $\eta_t=R/(\sigma\sqrt{k})\leq 1/L$, we conclude that the function value convergence rate is $1/\sqrt{k}$, i.e.

$$\mathbb{E}\left[f\left(\bar{x}^k\right)-f\left(x^*\right)\right]\leq\frac{R\sigma}{\sqrt{k}},\text{ and }\mathbb{E}\left[f\left(x_{\text{best}}^k\right)-f\left(x^*\right)\right]\leq\frac{R\sigma}{\sqrt{k}}$$

4.2.4 *Convergence under stochastic (L,σ)-smoothness and strongly convexity conditions*

We can start with equation (4.15) for $\alpha>0$, which has the same structure with equation (4.10) for Q-Lipschitz and strongly convex case in gradient descent.

Because of the stochastic noise σ in equation (4.15), we can no longer achieve fast linear rate. The convergence rate will be similar to the $1/k$ rate in Q-Lipschitz and strongly convex case. We will try to choose a diminishing step size which can give a cleaned telescoping result, and also keep in mind that we made the assumption $\eta_t \leq 1/L$ earlier in our proof.
Let's assume we choose $\eta_t = s/(t+t_0)$ where s and t_0 are parameters to be determined later. From (4.15) we have

$$\frac{1}{\eta_t}\mathbb{E}\left[f\left(x^{t+1}\right) - f\left(x^*\right)\right] = \frac{1}{2}\left[\left(\frac{1}{\eta_t^2} - \frac{\alpha}{\eta_t}\right)b^t - \frac{1}{\eta_t}b^{t+1}\right] + \sigma^2 \tag{4.16}$$

where $b_t = \mathbb{E}\left[\left\|x^t - x^*\right\|_2^2\right]$ following the same notation as before. To make telescoping work, we need

$$\frac{1}{\eta_t^2} - \frac{\alpha}{\eta_{t+1}} = \frac{1}{\eta_t^2}$$

which is satisfied if and only if $s = 2/\alpha$.
We also need to find a t_0 such that $\eta_t \leq L$, or $2\alpha = s \leq L(t+t_0)$. Choosing $t_0 = 2\alpha/L$ can make sure $\eta_t \leq L$.

Our step size rule is then chosen as

$$\eta_t = \frac{2\alpha}{t + 2\alpha/L}, t \geq 1$$

Let's now sum (4.16) from $t = 1$ to $t = k$ and get

$$\sum_{t=1}^{k}\frac{1}{\eta_t}\mathbb{E}\left[f\left(x^{t+1}\right) - f\left(x^*\right)\right] \leq k\sigma^2$$

and thus
$\mathbb{E}\left[f\left(x_{\text{bets}}^k\right) - f\left(x^*\right)\right] \leq \frac{k\sigma^2}{\sum_{t=1}^k 1/\eta_t} = \frac{2k\sigma^2\alpha}{k(k+1)/2 + 2k\alpha/L} = \frac{4\sigma^2\alpha}{(k+1)+4\alpha/L}$.
We can also get the decision variable convergence rate using the same trick as before, i.e.

$$0 \leq \frac{1}{\eta_t}\mathbb{E}\left[f\left(x^{t+1}\right) - f\left(x^*\right)\right] = \frac{1}{2}\left[\left(\frac{1}{\eta_t^2} - \frac{\alpha}{\eta_t}\right)b^t - \frac{1}{\eta_t}b^{t+1}\right] + \sigma^2$$

which sums up to

$$0 \leq \frac{1}{2}\left[0 - \frac{1}{\eta_t}\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|_2^2\right] + k\sigma^2\right.$$

which implies the result

$$\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|_2^2\right] \leq 2k\eta_t\sigma^2 = \frac{4k\alpha\sigma^2}{t + 2\alpha/L}$$