Lecture 13. Proximal Gradient Descent

13.1 Proximal operator and proximal gradient descent

So far we only consider minimizing differentiable functions. If the objective function is not differentiable, clearly neither the gradient descent nor the Newton's method works. In this lecture, we focus on how to solve the optimization problem for much more families of convex functions. We will generalize a method of gradient descent named the *proximal gradient descent* for nondifferentiable functions.

Recall the gradient descent iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)$, where we let $\boldsymbol{x}_{k+1}$ be the minimum point of

$$\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 .$$

Now we assume $f(\boldsymbol{x})$ is not differentiable, but $f(\boldsymbol{x})$ can be divided into two parts:

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x})$$

where $g(\boldsymbol{x})$ is convex and differentiable, and $h(\boldsymbol{x})$ is convex but not necessarily differentiable. Then we define

$$\hat{g}(\boldsymbol{x}) = g(\boldsymbol{x}_k) + \langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2$$

to approximate $g(\boldsymbol{x})$ and let

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \hat{g}(\boldsymbol{x}) + h(\boldsymbol{x})$$

to approximate the minimum point of $f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x})$.

Note that

$$\begin{aligned}
\langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 &= \frac{1}{2\eta} \langle \boldsymbol{x} - \boldsymbol{x}_k + \eta \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k + \eta \nabla g(\boldsymbol{x}_k) \rangle - \frac{\eta}{2} \|\nabla \\
&= \frac{1}{2\eta} \|\boldsymbol{x} - (\boldsymbol{x}_k - \eta \nabla g(\boldsymbol{x}_k))\|^2 - \frac{\eta}{2} \|\nabla g(\boldsymbol{x}_k)\|^2
\end{aligned}$$

where $\frac{\eta}{2}\|\nabla g(\boldsymbol{x}_k)\|^2$ is a constant not depending on $\boldsymbol{x}$ (assuming that $\boldsymbol{x}_k$ and η are fixed). So we obtain

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \hat{g}(\boldsymbol{x}) + h(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \frac{1}{2\eta}\|\boldsymbol{x} - (\boldsymbol{x}_k - \eta\nabla g(\boldsymbol{x}_k))\|^2 + h(\boldsymbol{x})\,.$$

Here $\boldsymbol{x}_k - \eta\nabla g(\boldsymbol{x}_k)$ is the gradient descent iteration if we would like to optimize only $g(\boldsymbol{x})$. So roughly speaking, after adding $h(\boldsymbol{x})$, we hope $\boldsymbol{x}_{k+1}$ locate near the local minimum of g, and not make h large.

Now we define the *proximal operator* as follows.

Definition (*Proximal operator*)

Given $\boldsymbol{y} \in \mathbb{R}^n$, let

$$\mathrm{prox}_h(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + h(\boldsymbol{x})\,.$$

Then we can rewrite the proximal gradient descent method as

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}} \frac{1}{2\eta}\|\boldsymbol{x} - (\boldsymbol{x}_k - \eta\nabla g(\boldsymbol{x}_k))\|^2 + h(\boldsymbol{x}) \\ &= \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - (\boldsymbol{x}_k - \eta\nabla g(\boldsymbol{x}_k))\|^2 + \eta h(\boldsymbol{x}) \\ &= \mathrm{prox}_{\eta h}(\boldsymbol{x}_k - \eta\nabla g(\boldsymbol{x}_k))\,.\end{aligned}$$

Tip

Another viewpoint of the proximal operator is a discretization of the gradient flow. Recall the gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = -\nabla f(X(t))\,.$$

The forward discretization (the *forward Euler method*) is the gradient descent, where

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k)\,.$$

Similarly, we can also try to discretize it in a backward form (the *backward Euler method*):

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_{k+1})\,.$$

However, the iteration becomes difficult since we need to find the point $\boldsymbol{x}_{k+1}$ satisfying above equations. Actually, this is what the proximal operator is doing. Let

$$\boldsymbol{x}_{k+1} = \operatorname{prox}_{\eta f}(\boldsymbol{x}_k) = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|^2 + \eta f(\boldsymbol{x})\,.$$

Then we have $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k + \eta\,\nabla f(\boldsymbol{x}_{k+1}) = \boldsymbol{0}$, since

$$\mathsf{LHS} = \nabla\Big(\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|^2 + \eta f(\boldsymbol{x})\Big)\Big|_{\boldsymbol{x}=\boldsymbol{x}_{k+1}} = \boldsymbol{0}\,.$$

13.2 LASSO

The key is that we need to decompose $f(x) = g(x) + h(x)$ properly. Clearly we can set $g(x) \equiv 0$ and $h(x) \equiv f(x)$. But this decomposition is meaningless since wo do not know how to compute the proximal operator at all.

Fortunately for some important problems we have a "good" decomposition. For example, we consider the problem of linear progression avoiding overfitting. Suppose we have a data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ and we assume that $y_i = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x}_i$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$.

Let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathsf{T}} \end{pmatrix} \in \mathbb{R}^{m\times n} \qquad \text{and} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m\,.$$

We can compute coefficients $\boldsymbol{\beta}^*$ by solving the following optimization problem (*least square method*)

$$\boldsymbol{\beta}^* = \arg\min_{\beta\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

However, there are usually some redundant coefficients that may cause *overfitting*. So we hope to add some constraints, such as, the number of nonzero entries in $\boldsymbol{\beta}^*$ is at most k. Unfortunately this problem is not a convex optimization after adding this constraint (why?). We still need to approximate it. One idea is to use the

following approximation

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

where $\lambda > 0$ is a parameter. This method is called *LASSO* (*least absolute shrinkage and selection operator*).

The objective function $f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ has a clear decomposition: $g(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ and $h(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$. The advantage is that the proximal operator is easy to compute with respect to this h.

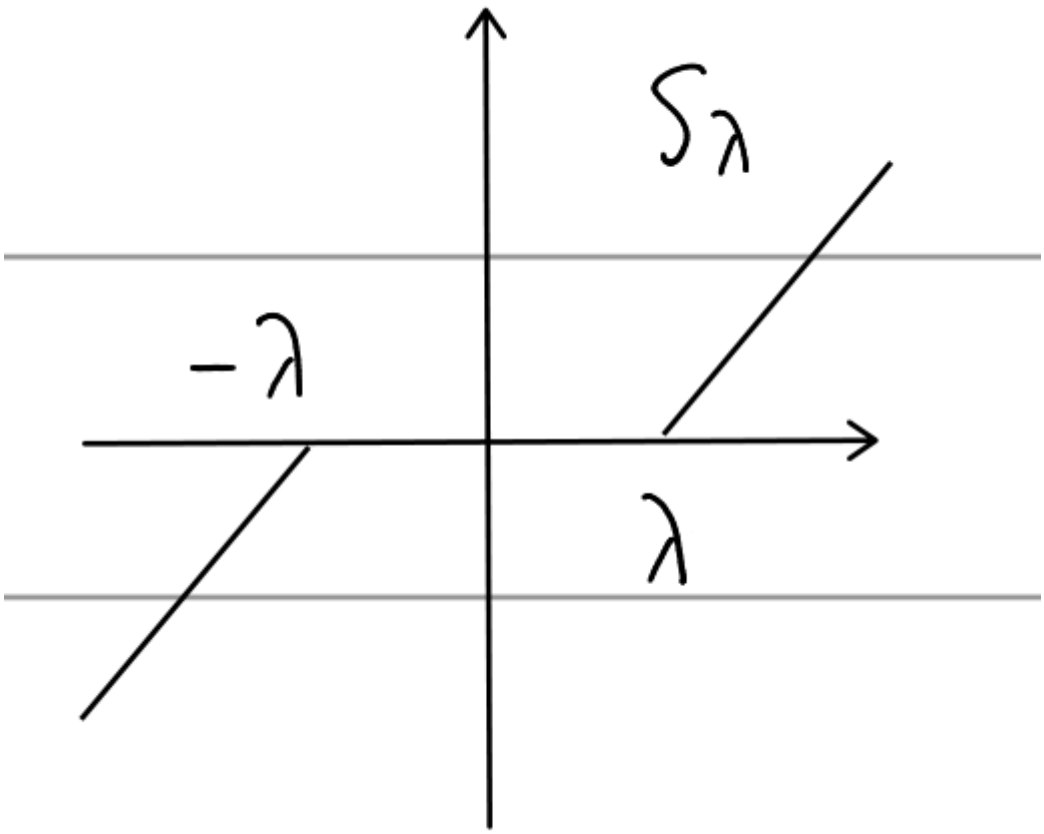
Given $\boldsymbol{\gamma} \in \mathbb{R}^n$, the proximal mapping is

$$\begin{aligned}
\operatorname{prox}_h(\boldsymbol{\gamma}) &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\left((\beta_1 - \gamma_1)^2 + \cdots + (\beta_n - \gamma_n)^2\right) + \lambda\left(|\beta_1| + \cdots + |\beta_n|\right).
\end{aligned}$$

Note that in this optimization, all entries in $\boldsymbol{\beta}$ are independent, so we can solve this optimization by solving the following optimization for each entry:

$$\begin{aligned}
\left(\operatorname{prox}_h(\boldsymbol{\gamma})\right)_i &= \arg\min_{\beta_i\in\mathbb{R}} \frac{1}{2}(\beta_i - \gamma_i)^2 + \lambda|\beta_i| \\
&= \arg\min_{\beta_i} \frac{1}{2}(\beta_i - \gamma_i)^2 + \lambda \cdot \begin{cases} \beta_i & \beta_i \geq 0 \\ -\beta_i & \beta_i < 0 \end{cases} \\
&= \arg\min_{\beta_i} \beta_i^2 - 2\beta_i \cdot \begin{cases} \gamma_i - \lambda & \beta_i \geq 0 \\ \gamma_i + \lambda & \beta_i < 0 \end{cases} \\
&= \mathcal{S}_\lambda(\gamma_i) \triangleq \begin{cases} \gamma_i - \lambda & \gamma_i > \lambda \\ 0 & |\gamma_i| \leq \lambda \\ \gamma_i + \lambda & \gamma_i < -\lambda \end{cases},
\end{aligned}$$

where $\mathcal{S}_\lambda$ is called the soft thresholding operator.



Finally, recall that $g(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$, so the iteration of the proximal gradient descent is

$$\boldsymbol{\beta}_{k+1} = \mathrm{prox}_{\eta h}(\boldsymbol{\beta}_k - \eta \nabla g(\boldsymbol{\beta}_k)) = \mathcal{S}_{\lambda\eta}\big(\boldsymbol{\beta}_k - \eta\left(-\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k)\right)\big)\,.$$

This algorithm is called the *ISTA* (*iterative soft-thresholding algorithm*).

13.3 Correctness and convergence

We now show the correctness and convergence of the proximal gradient descent.

Assume that g is L-smooth and set $\eta \leq 1/L$. Sometimes we would write the proximal gradient descent as the following form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta G_\eta(\boldsymbol{x}_k)\,,$$

where

$$G_\eta(\boldsymbol{x}) = \frac{\boldsymbol{x} - \mathrm{prox}_{\eta h}(\boldsymbol{x} - \eta \nabla g(\boldsymbol{x}))}{\eta}\,.$$

We first show that, $G_\eta(\boldsymbol{x}_k) = \mathbf{0}$ if and only if $\boldsymbol{x}_k$ is an minimum point of $f(\boldsymbol{x})$. A trivial example is $h \equiv 0$. Then we have $\mathrm{prox}_{\eta h}(\boldsymbol{x}) = \boldsymbol{x}$ and thus $G_\eta(\boldsymbol{x}) = \nabla g(\boldsymbol{x})$. So $G_\eta(\boldsymbol{x}) = \mathbf{0}$ if and only if $\boldsymbol{x}$ is a minimizer of $f = g$. The following theorem asserts the general cases.

Theorem

In the proximal gradient descent iterations, $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ iff $f(\boldsymbol{x}_k) = f^*$, where f^* is the minimum value of f.

The "only if" direction is easy, since we have

$$\hat{f}(\boldsymbol{x}) = \hat{g}(\boldsymbol{x}) + h(\boldsymbol{x}) = g(\boldsymbol{x}_k) + \langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_k\|^2 + h(\boldsymbol{x}) \geq g(\boldsymbol{x}) + h(\boldsymbol{x}) =$$

by L-smoothness of g. It yields that

$$f(\boldsymbol{x}_{k+1}) \leq \hat{f}(\boldsymbol{x}_{k+1}) = \min_{\boldsymbol{x}}\ \hat{f}(\boldsymbol{x}) \leq \hat{f}(\boldsymbol{x}_k) = f(\boldsymbol{x}_k)\,.$$

If $f(\boldsymbol{x}_k) = f^*$, all inequalities are tight, so $\hat{f}(\boldsymbol{x}_k) = \min\ \hat{f}(\boldsymbol{x})$. However, because h is convex and $\|\boldsymbol{x} - \boldsymbol{x}_k\|^2$ is strictly convex, $\hat{f}$ is also strictly convex, and thus has a unique minimizer $\boldsymbol{x}_{k+1}$, which gives that $\boldsymbol{x}_k = \boldsymbol{x}_{k+1}$.

Now we would like to show that if $\boldsymbol{x}_k = \boldsymbol{x}_{k+1}$ then $f(\boldsymbol{x}_k) = f^*$. In other words, for all $\boldsymbol{x} \in \mathbb{R}^n$, $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_k)$. How can we show this? A naive idea is to show that $g(\boldsymbol{x}) \geq g(\boldsymbol{x}_k)$ and $h(\boldsymbol{x}) \geq h(\boldsymbol{x}_k)$. However this idea looks too good to be true.

A reasonable method is to use convexity. Note that
$g(\boldsymbol{x}) \geq g(\boldsymbol{x}_k) + \langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle$. So we hope

$$h(\boldsymbol{x}) \geq h(\boldsymbol{x}_k) - \langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle\,.$$

This inequality looks like the first order condition for h. Here $-\nabla g$ performs like the gradient of h. Since h is not differentiable, we would introduce the notion of *subgradients*.

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $\boldsymbol{x} \in \mathbb{R}^n$. We say $\boldsymbol{v} \in \mathbb{R}^n$ is a *subgradient* of f at $\boldsymbol{x}$, denoted by $\boldsymbol{v} \in \partial f(\boldsymbol{x})$, if $\forall\, \boldsymbol{y} \in \mathbb{R}^n$,
$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle$.

If f is convex, then subgradients always exist (but may not unique). Just consider the supporting hyperplane of the epigraph of f.

Lemma

If $\boldsymbol{z} = \operatorname{prox}_{\eta h}(\boldsymbol{y})$, then $\frac{\boldsymbol{y}-\boldsymbol{z}}{\eta} \in \partial h(\boldsymbol{z})$.

This lemma immediately implies that if $\boldsymbol{x}_k = \boldsymbol{x}_{k+1} = \operatorname{prox}_{\eta h}(\boldsymbol{x}_k - \eta \nabla g(\boldsymbol{x}_k))$, then $-\nabla g(x_k) \in \partial h(x_k)$. Thus

$$\forall\, \boldsymbol{x} \in \mathbb{R}^n, \quad f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x}) \geq g(\boldsymbol{x}_k) + \langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle + h(\boldsymbol{x}_k) - \langle \nabla g(x_k),\, \boldsymbol{x} - \boldsymbol{x}_k$$

Proof of the Lemma

By the definition of $\boldsymbol{z}$, we have that for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$\frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + h(\boldsymbol{x}) \geq \frac{1}{2\eta}\|\boldsymbol{z} - \boldsymbol{y}\|^2 + h(\boldsymbol{z})\,.$$

Our goal is to show that for all $\boldsymbol{x} \in \mathbb{R}^n$, $h(\boldsymbol{x}) \geq h(\boldsymbol{z}) + \frac{1}{\eta}\langle \boldsymbol{y} - \boldsymbol{z},\, \boldsymbol{x} - \boldsymbol{z} \rangle$.
For the sake of contradiction, assume that there exists $\boldsymbol{w} \in \mathbb{R}^n$ such that $h(\boldsymbol{w}) < h(\boldsymbol{z}) + \frac{1}{\eta}\langle \boldsymbol{y} - \boldsymbol{z},\, \boldsymbol{w} - \boldsymbol{z} \rangle$. Define

$$\delta = \frac{h(\boldsymbol{z}) - h(\boldsymbol{w}) + \frac{1}{\eta}\langle \boldsymbol{y} - \boldsymbol{z}\,, \boldsymbol{w} - \boldsymbol{z} \rangle}{\|\boldsymbol{w} - \boldsymbol{z}\|} > 0\,.$$

Then $h(\boldsymbol{w}) = h(\boldsymbol{z}) + \frac{1}{\eta}\langle \boldsymbol{y} - \boldsymbol{z}, \boldsymbol{w} - \boldsymbol{z} \rangle - \delta\|\boldsymbol{w} - \boldsymbol{z}\|$.
Let $\theta \in (0, 1)$ and $\boldsymbol{x} = \theta \boldsymbol{w} + (1-\theta)\boldsymbol{z}$. By convexity we have

$$\begin{aligned} h(\boldsymbol{x}) &\leq \theta h(\boldsymbol{w}) + (1-\theta) h(\boldsymbol{z}) \\ &= h(\boldsymbol{z}) + \frac{\theta}{\eta}\langle \boldsymbol{y} - \boldsymbol{z}\,, \boldsymbol{w} - \boldsymbol{z} \rangle - \theta\delta\|\boldsymbol{w} - \boldsymbol{z}\| \\ &= h(\boldsymbol{z}) + \frac{1}{\eta}\langle \boldsymbol{y} - \boldsymbol{z}\,, \boldsymbol{x} - \boldsymbol{z} \rangle - \delta\|\boldsymbol{x} - \boldsymbol{z}\|\,. \end{aligned}$$

Since

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x} - \boldsymbol{z}\|^2 + \|\boldsymbol{z} - \boldsymbol{y}\|^2 + 2\langle \boldsymbol{x} - \boldsymbol{z},\, \boldsymbol{z} - \boldsymbol{y} \rangle\,,$$

it follows that

$$\begin{aligned}\frac{1}{2\eta}\|\boldsymbol{x}-\boldsymbol{y}\|^2+h(\boldsymbol{x}) &= \frac{1}{2\eta}\|\boldsymbol{x}-\boldsymbol{z}\|^2+\frac{1}{2\eta}\|\boldsymbol{y}-\boldsymbol{z}\|^2+\frac{1}{\eta}\langle \boldsymbol{x}-\boldsymbol{z},\,\boldsymbol{z}-\boldsymbol{y}\rangle+h(\boldsymbol{x}) \\ &\le \frac{1}{2\eta}\|\boldsymbol{x}-\boldsymbol{z}\|^2+\frac{1}{2\eta}\|\boldsymbol{y}-\boldsymbol{z}\|^2+h(\boldsymbol{z})-\delta\|\boldsymbol{x}-\boldsymbol{z}\| \\ &< \frac{1}{2\eta}\|\boldsymbol{z}-\boldsymbol{y}\|^2+h(\boldsymbol{z})\end{aligned}$$

if $\|\boldsymbol{x}-\boldsymbol{z}\|$ is sufficiently small ($<2\delta\eta$). This contradicts to the definition of $\boldsymbol{z}$.

Finally, we present the results of convergence rate.

Theorem

Suppose g is L-smooth and we set $\eta \le 1/L$. Then we have

$$f(\boldsymbol{x}_T)-f^* \le \frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{2\,T\eta}\,.$$

If g is further μ-strongly convex, then

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 \le (1-\mu\eta)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2$$

If L-smoothness is unknown, we can use the exact/backtracking line search, and the results to the convergence rate are also the same as the rate of gradient descent.