IE 539: Convex Optimization **KAIST, Fall 2024**
Lecture #15: ISTA and FISTA for LASSO November 6, 2024
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- regularization for linear regression,
- lasso: least absolute shrinkage and selection operator,
- ISTA and FISTA for LASSO,
- accelerated proximal gradient descent.

2 Linear Regression

In this section, we consider linear regression in the context of smoothness and strong convexity. We assume that the relationship between the vector $x \in \mathbb{R}^d$ of features and the response variable $y \in \mathbb{R}$ is modeled using a linear equation given by

$$y = \theta_{\text{true}}^\top x + \epsilon$$

where:

- $\theta_{\text{true}} \in \mathbb{R}^d$ is the coefficient vector,
- $\epsilon \in \mathbb{R}$ is the noise term representing unexplained variation.

We infer the true coefficient vector θ_{true} using the method of least squares, which minimizes the average of squared differences between the observed and predicted values of y. Namely, given a set of n data $(x_1, y_1), \ldots, (x_n, y_n)$, we solve

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \theta^\top x_i\right)^2 \quad = \quad \min_{\theta} \quad \frac{1}{n}\|Y - X\theta\|_2^2. \tag{15.1}$$

Here, Y denotes the vector whose components are $y_1, \ldots, y_n$, and X denotes the matrix whose rows are $x_1^\top, \ldots, x_n^\top$. Note that

$$f(\theta) := \frac{1}{n}\|Y - X\theta\|_2^2 = \frac{1}{n}\theta^\top X^\top X\theta - \frac{2}{n}Y^\top X\theta + \frac{1}{n}Y^\top Y.$$

Since $X^\top X$ is positive semidefinite, it follows that the MSE loss $f(\theta)$ is convex. Moreover, $f(\theta)$ is α-strongly convex and β-smooth in the ℓ_2-norm with

$$\alpha = \frac{1}{n}\lambda_{\min}(X^\top X) \quad \text{and} \quad \beta = \frac{1}{n}\lambda_{\max}(X^\top X)$$

where $\lambda_{\min}(X^\top X)$ and $\lambda_{\max}(X^\top X)$ denote the minimum and maximum eigenvalues of $X^\top X$. As long as X is a nonzero matrix, we have $\lambda_{\max}(X^\top X) > 0$. However, we can have $\lambda_{\min}(X^\top X) = 0$ when the rank of $X^\top X$ is lower than the number of features d.

Data-Rich Regime Recall that n is the number of data and d is the number of features. When $n \geq d$, then it is possible that X is of full column rank, in which case $X^\top X$ is invertible. If $X^\top X$ is invertible, it is positive definite, and therefore, we have $\alpha = \lambda_{\min}(X^\top X)/n > 0$. In this case, the MSE loss $f(\theta)$ is indeed strongly convex. Another Remark is that if $X^\top X$ is invertible, then

$$\theta_{\text{opt}}^{\text{rich}} := \text{argmin}_\theta \frac{1}{n}\|Y - X\theta\|_2^2 = (X^\top X)^{-1}X^\top y$$

because

$$\nabla f(\theta) = \frac{2}{n}X^\top(X\theta - y).$$

Data-Poor Regime When $n < d$, then the rank of X is less than d, which means that $X^\top X$ is not of full rank and thus $X^\top X$ is not invertible. In this case, we have $\alpha = \lambda_{\min}(X^\top X)/n = 0$, and therefore, the MSE loss $f(\theta)$ is not strongly convex. When $X^\top X$ is not invertible, we have

$$\theta_{\text{opt}}^{\text{poor}} := \text{argmin}_\theta \frac{1}{n}\|Y - X\theta\|_2^2 = (X^\top X)^{\dagger}X^\top y$$

where $(X^\top X)^\dagger$ denotes the Moore-Penrose pseudo-inverse of $X^\top X$.

2.1 Gradient Descent for Minimizing the MSE Loss

We generated a random instance with 75 feature variables and 100 data samples. To consider a data-poor regime, we randomly selected 30 samples from the data set. Recall that θ_{true} denotes the true coefficient vector in the linear model $y = \theta_{\text{true}}^\top x + \epsilon$.

The following figures map loss convergence patterns under the data-rich and data-poor regimes. The figures show that gradient descent quickly minimizes the MSE loss under both regimes.

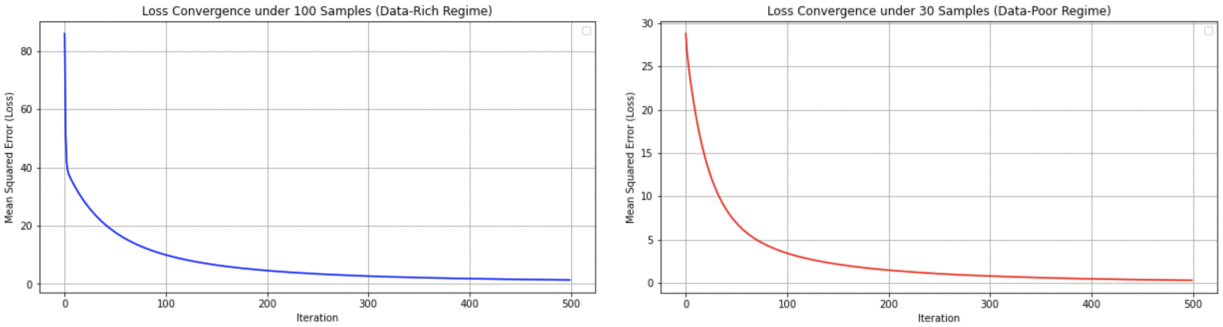

Figure 15.1: Loss convergence patterns under the data-rich regime (Left) and the data-poor regime (Right)

Let us verify whether gradient descent returns solutions that converge to the optimal model minimizing the MSE loss. Recall that $\theta_{\text{opt}}^{\text{rich}} = (X^\top X)^{-1}X^\top y$ is the model minimizing the MSE loss under the data-rich regime while $\theta_{\text{opt}}^{\text{poor}} = (X^\top X)^\dagger X^\top y$ is the model minimizing the MSE loss under the data-poor regime. Figure 15.2 reports the distances between models θ generated by gradient descent and the optimal model under each regime. Here, the purple line shows the squared norm of $\theta_{\text{opt}}^{\text{rich}}$ and that of $\theta_{\text{opt}}^{\text{poor}}$, given by $\|\theta_{\text{opt}}^{\text{rich}}\|_2^2$ and $\|\theta_{\text{opt}}^{\text{poor}}\|_2^2$, respectively. The red line depicts $\|\theta - \theta_{\text{opt}}^{\text{poor}}\|_2^2$ under the data-poor regime, while the blue one shows $\|\theta - \theta_{\text{opt}}^{\text{rich}}\|_2^2$ under the data-rich regime. Figure 15.2 shows that the solution deduced by gradient descent under the data-rich regime indeed

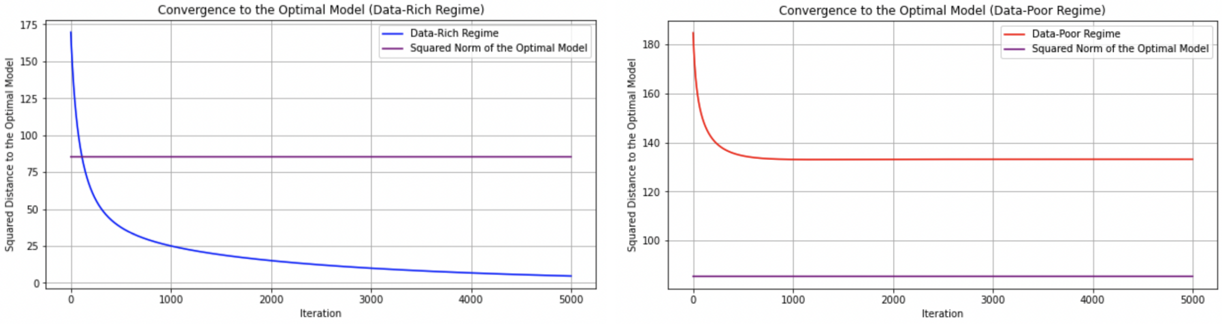

Figure 15.2: Convergence to the optimal model under the data-rich regime (Left) and the data-poor regime (Right)

seems to converge to the optimal vector minimizing the MSE loss, but that under the data-poor regime does not. This is because the MSE loss is no longer strongly convex under the data-poor regime.

In Figure 15.3, we report the distances between each model θ generated by gradient descent and the true coefficient vector θ_{true}. Here, the green line shows the squared norm of θ_{true}, given by $\|\theta_{\text{true}}\|_2^2$. The red line depicts $\|\theta - \theta_{\text{true}}\|_2^2$ under the data-poor regime, while the blue one shows $\|\theta - \theta_{\text{true}}\|_2^2$ under the data-rich regime. Figure 15.3 shows that the solution deduced by gradient descent under the data-rich regime indeed seems to converge to the actual true coefficient vector, but that under the data-poor regime does not.

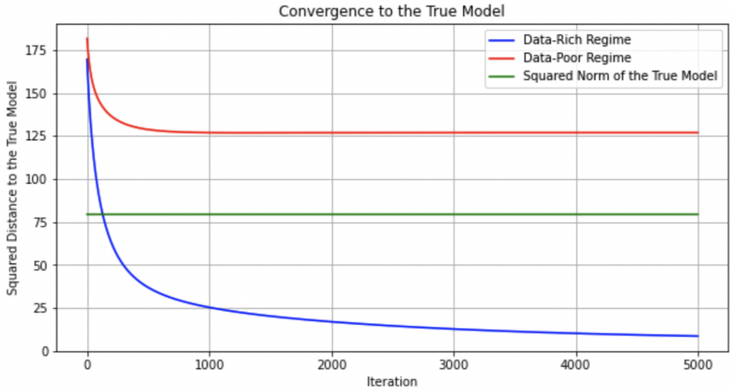

Figure 15.3: Convergence to the true model

2.2 ℓ_2-Regularized Least Squares

We discussed that the MSE loss under the data-poor regime is not strongly convex. In practice, it is often desirable to add an ℓ_2-regularization term, which makes the resulting loss function strongly convex. To be more precise, we consider

$$\min_{\theta} \quad \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2 \tag{15.2}$$

for some positive λ. Note that the regularized loss is α-strongly convex and β-smooth in the ℓ_2-norm with

$$\alpha = \frac{1}{n}\lambda_{\min}(X^\top X) + \lambda \quad \text{and} \quad \beta = \frac{1}{n}\lambda_{\max}(X^\top X) + \lambda.$$

Hence, as long as $\lambda > 0$, the regularized loss is strongly convex. As $X^\top X + \alpha I$ is positive definite, the model minimizing the regularized loss is given by

$$\theta_{\text{opt}} := \text{argmin}_\theta \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2 = (X^\top X + \lambda I)^{-1}X^\top y.$$

In Figure 15.4, we report the distances between each model θ generated based on the regularized loss and the true coefficient vector θ_{true}. Here, the green line shows the squared norm of θ_{true}, given by $\|\theta_{\text{true}}\|_2^2$. The orange line depicts $\|\theta - \theta_{\text{true}}\|_2^2$ for the regularized loss, while the red one shows $\|\theta - \theta_{\text{true}}\|_2^2$ for the original MSE loss. We see that

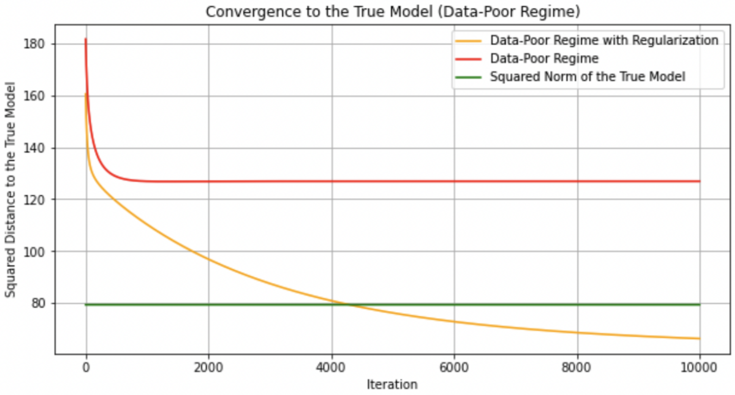

Figure 15.4: Convergence to the true model under regularization

Let us also check convergence to the optimal model minimizing the regularized loss. Recall that $\theta_{\text{opt}} = (X^\top X + \lambda I)^{-1}X^\top y$ is the model minimizing the regularized loss. Here, the purple line shows the squared norm of θ_{opt} given by $\|\theta_{\text{opt}}\|_2^2$. The orange line depicts $\|\theta - \theta_{\text{opt}}\|_2^2$ for the regularized loss, while the red one shows $\|\theta - \theta_{\text{opt}}\|_2^2$ for the original MSE loss.

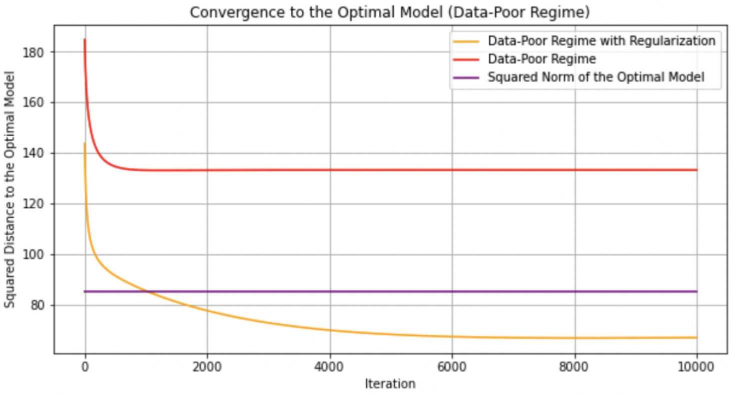

Figure 15.5: Convergence to the optimal model under regularization

3 LASSO: Least Absolute Shrinkage and Selection Operator

Recall the formulation of LASSO, given by

$$\min_{\theta} \quad \frac{1}{n}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1.$$

Here, the objective function is non-differentiable because of the ℓ_1-regularization term $\lambda\|\theta\|_1$, and therefore, it is non-smooth. On the other hand, the objective is convex, and we have a characterization of the subdifferential of $\|\theta\|_1$, so we can simply apply the subgradient method. To bound the additive error by ϵ, the subgradient method requires $O(1/\epsilon^2)$ iterations.

As discussed in the last lecture, we know that the first part is smooth, and the other part is something whose subdifferential is well understood. Hence, we may apply proximal gradient descent with $h(\theta) = \lambda\|\theta\|_1$ whose associated prox operator is given by

$$\mathrm{prox}_{\eta\lambda\|\cdot\|_1}(\theta) = \left(\underbrace{\max\left\{0, |\theta_i| - \eta\lambda\right\}}_{\text{shirinkage operator}} \cdot \mathrm{sign}(\theta_i)\right)_{i\in[d]}$$

The proximal gradient algorithm applies to this composite problem proceeds with the following update rule.

$$\theta_{t+1} = \mathrm{prox}_{\eta h}(\theta_t - \eta\nabla g(\theta_t)).$$

Proximal Gradient Descent applied to LASSO is referred to as Iterative Shrinkage-Thresholding Algorithm (ISTA).

4 Nesterov's Acceleration and FISTA

We observed that proximal gradient descent achieves a convergence rate of $O(1/T)$, and therefore, ISTA solves LASSO with a convergence rate of $O(1/T)$. In fact, we may deduce a faster convergence rate based on Nesterov's acceleration. We mentioned that Nesterov's accelerated gradient descent guarantees a convergence rate of $O(1/T^2)$ for smooth convex minimization. We will show that an accelerated version of proximal gradient descent achives a rate of $O(1/T^2)$ for the composite convex minimization where g is smooth and convex.

Recall that proximal gradient descent for minimizing $g + h$ where g is β-smooth and convex and h is convex follows the update rule of

$$x_{t+1} = \mathrm{prox}_{h/\beta}\left(x_t - \frac{1}{\beta}\nabla g(x_t)\right)$$

from a given point x_t. Instead of applying the gradient descent update to x_t, we move a bit further from x_t along the momentum direction that we took from x_{t-1} to x_t. Let $\gamma_t > 0$ be a weight, and

$$y_t = x_t + \gamma_t(x_t - x_{t-1}).$$

Then we apply the primal gradient descent update on y_t to obtain the next point x_{t+1}, as follows.

$$x_{t+1} = \mathrm{prox}_{h/\beta}\left(y_t - \frac{1}{\beta}\nabla g(y_t)\right).$$

Algorithm 1 summarizes the accelerated version of proximal gradient descent that we just explained.

To provide a convergence result of the accelerated proximal gradient descent method, we need the following lemma.

Algorithm 1 Accelerated Proximal Gradient Descent

```
Initialize $x_1 \in \mathbb{R}^d$.
Set $x_0 = x_1$.
for $t = 1, \ldots, T$ do
    $y_t = x_t + \gamma_t(x_t - x_{t-1})$ for some $\gamma_t > 0$.
    $x_{t+1} = \mathrm{prox}_{h/\beta}\left(y_t - \frac{1}{\beta}\nabla g(y_t)\right)$.
end for
Return $x_{T+1}$.
```

Lemma 15.1. *Let $u, v \in \mathbb{R}^d$. Then for all $z \in \mathbb{R}^d$,*

$$\frac{1}{\eta}(\mathrm{prox}_{\eta h}(x) - x)^\top (z - \mathrm{prox}_{\eta h}(x)) + h(z) \geq h(\mathrm{prox}_{\eta h}(x)).$$

Proof. Note that

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

By the optimality condition, it follows that for any $z \in \mathbb{R}^d$ and $g \in \partial h(\mathrm{prox}_{\eta h}(x))$,

$$\left(g + \frac{1}{\eta}\left(\mathrm{prox}_{\eta h}(x) - x\right)\right)^\top (z - \mathrm{prox}_{\eta h}(x)) \geq 0.$$

This implies that

$$\frac{1}{\eta}(\mathrm{prox}_{\eta h}(x) - x)^\top (z - \mathrm{prox}_{\eta h}(x)) + g^\top (z - \mathrm{prox}_{\eta h}(x)) \geq 0.$$

Here, since h is convex, we have

$$h(z) \geq h(\mathrm{prox}_{\eta h}(x)) + g^\top (z - \mathrm{prox}_{\eta h}(x)).$$

Adding the two inequalities, we prove the desired bound of this lemma. □

Theorem 15.2. *Let $f = g + h$ where g is a β-smooth convex function in the ℓ_2 norm and h is convex. We set η and γ_t as*

$$\eta = \frac{1}{\beta}, \quad \gamma_t = \frac{t-2}{t+1}.$$

Then x_{T+1} returned by Accelerated Proximal Gradient Descent (Algorithm 1) satisfies

$$f(x_{T+1}) - f(x^*) \leq \frac{2\beta}{(T+1)^2}\|x_1 - x^*\|_2^2$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Proof. Note that Algorithm 1 is equivalent to

$$\begin{aligned} y_t &= (1 - \lambda_t)x_t + \lambda_t v_t \\ x_{t+1} &= \mathrm{prox}_{h/\beta}\left(y_t - \frac{1}{\beta}\nabla g(y_t)\right) \\ v_{t+1} &= x_t + \frac{1}{\lambda_t}(x_{t+1} - x_t) \end{aligned}$$

where

$$\lambda_t = \frac{2}{t+1}.$$

This is because $y_t = x_t + \lambda_t(v_t - x_t)$ and

$$\lambda_t(v_t - x_t) = \lambda_t\left(\left(\frac{1}{\lambda_{t-1}} - 1\right)x_t + \left(1 - \frac{1}{\lambda_{t-1}}\right)x_{t-1}\right) = \frac{\lambda_t(1-\lambda_{t-1})}{\lambda_{t-1}}(x_t - x_{t-1}) = \gamma_t(x_t - x_{t-1}).$$

Moreover, we have $\lambda_1 = 1$, and for $t \geq 2$,

$$\frac{1-\lambda_t}{\lambda_t^2} \leq \frac{1}{\lambda_{t-1}^2}.$$

First, as g is β-smooth,

$$g(x_{t+1}) \leq g(y_t) + \nabla g(y_t)^\top (x_{t+1} - y_t) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2.$$

Next, Lemma 15.1 implies that for any $z \in \mathbb{R}^d$,

$$\begin{aligned} h(x_{t+1}) &\leq h(z) + \beta\left(x_{t+1} - y_t + \frac{1}{\beta}\nabla g(y_t)\right)^\top (z - x_{t+1}) \\ &= h(z) + \nabla g(y_t)^\top (z - x_{t+1}) + \beta\left(x_{t+1} - y_t\right)^\top (z - x_{t+1}). \end{aligned}$$

Adding these two inequalities, we deduce that

$$\begin{aligned} f(x_{t+1}) &\leq h(z) + g(y_t) + \nabla g(y_t)^\top (z - y_t) + \beta\left(x_{t+1} - y_t\right)^\top (z - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \\ &\leq f(z) + \beta\left(x_{t+1} - y_t\right)^\top (z - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \end{aligned}$$

where the second inequality follows from convexity of g. By setting $z = x^*$ and $z = x_t$, we have

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq \beta\left(x_{t+1} - y_t\right)^\top (x^* - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \\ f(x_{t+1}) - f(x_t) &\leq \beta\left(x_{t+1} - y_t\right)^\top (x_t - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2. \end{aligned}$$

Summing up the first inequality multiplied by λ_t and the second one multiplied by $(1 - \lambda_t)$, we get

$$\begin{aligned} &f(x_{t+1}) - f(x^*) - (1-\lambda_t)(f(x_t) - f(x^*)) \\ &\leq \beta\left(x_{t+1} - y_t\right)^\top (\lambda_t x^* + (1-\lambda_t)x_t - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \\ &= \frac{\beta}{2}(x_{t+1} - y_t)^\top (2\lambda_t x^* + 2(1-\lambda_t)x_t - x_{t+1} - y_t) \\ &= \frac{\beta}{2}\|y_t - (1-\lambda_t)x_t - \lambda_t x^*\|_2^2 - \frac{\beta}{2}\|x_{t+1} - (1-\lambda_t)x_t - \lambda_t x^*\|_2^2 \\ &= \frac{\beta\lambda_t^2}{2}\|v_t - x^*\|_2^2 - \frac{\beta\lambda_t^2}{2}\|v_{t+1} - x^*\|_2^2. \end{aligned}$$

This implies that

$$
\begin{aligned}
\frac{1}{\lambda_t^2}(f(x_{t+1}) - f(x^*)) + \frac{\beta}{2}\|v_{t+1} - x^*\|_2^2 &\le \frac{1-\lambda_t}{\lambda_t^2}(f(x_t) - f(x^*)) + \frac{\beta}{2}\|v_t - x^*\|_2^2 \\
&\le \frac{1}{\lambda_{t-1}^2}(f(x_t) - f(x^*)) + \frac{\beta}{2}\|v_t - x^*\|_2^2 \\
&\quad \vdots \\
&\le \frac{1}{\lambda_1^2}(f(x_2) - f(x^*)) + \frac{\beta}{2}\|v_2 - x^*\|_2^2 \\
&\le \frac{1-\lambda_1}{\lambda_1^2}(f(x_1) - f(x^*)) + \frac{\beta}{2}\|v_1 - x^*\|_2^2 \\
&= \frac{\beta}{2}\|v_1 - x^*\|_2^2 \\
&= \frac{\beta}{2}\|x_1 - x^*\|_2^2.
\end{aligned}
$$

Therefore, it follows that

$$f(x_{T+1}) - f(x^*) \le \frac{\beta\lambda_T^2}{2}\|x_1 - x^*\|_2^2 = \frac{2\beta}{(T+1)^2}\|x_1 - x^*\|_2^2,$$

as required. □

Hence, the convergence rate is $O(1/T^2)$, which matches the oracle lower bound. The number of required iterations to bound the error by ϵ is $O(1/\sqrt{\epsilon})$.

FISTA stands for Fast ISTA, that is an accelerated version of ISTA. Basically, FISTA is the accelerated proximal gradient descent method applied to LASSO. ISTA requires $O(1/\epsilon)$ iterations, while FISTA needs $O(1/\sqrt{\epsilon})$ iterations to converge to an ϵ-approximate solution.

We generated a random instance with 300 feature variables and 100 data samples. The figure

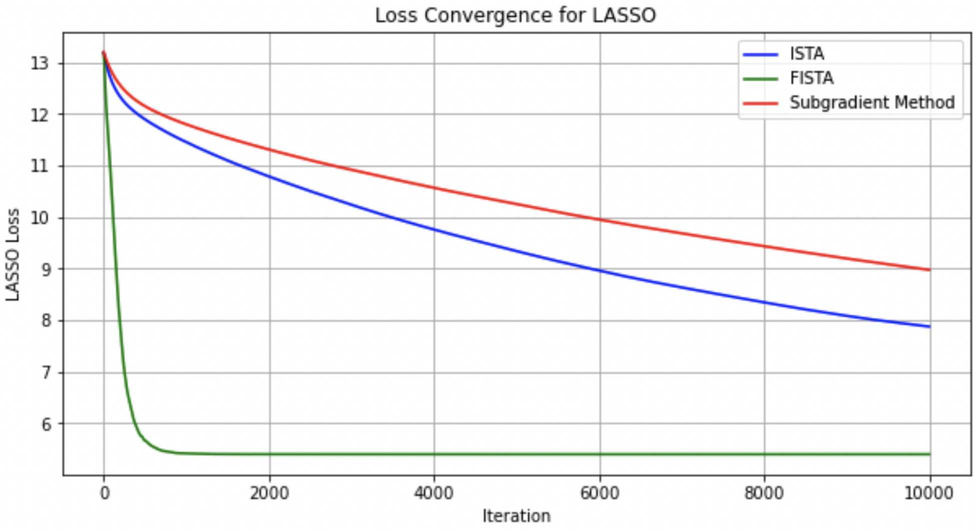

Figure 15.6: Comparing the subgradient method, ISTA, and FISTA for LASSO

compares the subgradient method, ISTA, and FISTA for the random LASSO instance. We can see that FISTA has the fastest rate of convergence while ISTA is also faster than the subgradient method.