Algorithmic Foundations of Learning **Lecture 13**

The Lasso Estimator. Proximal Gradient Methods

Lecturer: Patrick Rebeschini *Version: December 8, 2021*

13.1 Introduction

In the last lecture we investigated the *statistical* performance of the estimator

$$W^0 := \operatorname*{argmin}_{w:\|w\|_0\le k} \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2$$

to recover the true sparse parameter $w^\star$ that generates the data via $Y = \mathbf{x}w^\star + \sigma\xi$, where ξ is the (unknown) noise vector, made of i.i.d. standard Gaussian components, and σ^2 is the variance of the noise. We derived bounds in expectation and in probability for the ℓ_2 estimation loss, i.e., for the quantity $\|W^0 - w^\star\|_2$.

In this lecture we focus on *computing* an estimator to find the parameter $w^\star$. The main difficulty with computing W^0 stems from the fact that the associated optimization problem is non-convex due to the presence of the ℓ_0 pseudo-norm. To obtain a convex problem, we could consider the estimator that is obtained by replacing the ℓ_0 norm with the ℓ_1 norm, namely,

$$W^1 := \operatorname*{argmin}_{w:\|w\|_1\le k} \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2.$$

So far in this course we only dealt with constrained problems (in the prediction setting) as a way to impose regularization. We now investigate another way of performing regularization, by means of adding a penalty term to the objective function. Namely, instead of considering the estimator W^1 defined above, we consider its *unconstrained penalized* version where the hard constraint $\|w\|_1 \le k$ is replaced by the penalty term $\lambda\|w\|_1$ added to the objective function, for a certain parameter $\lambda > 0$ to be tuned.[1]. Furthermore, we replace the function $w \to \|\mathbf{x}w - Y\|_2^2$ with a generic empirical loss function $R : \mathbb{R}^d \to \mathbb{R}_+$ (this is a random function, as it is function of the data). That is, we consider the following estimator:

$$\boxed{W^{p1} := \operatorname*{argmin}_{w\in\mathbb{R}^d} R(w) + \lambda\|w\|_1} \tag{13.1}$$

The choice $R(w) = \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2$ leads to the celebrated *Lasso estimator.*

We first investigate the statistical guarantees of the estimator W^{p1} and show that under appropriate conditions we recover the same statistical guarantees that we established for W^0. We then introduce a novel class of gradient methods that are specifically designed to solve the general type of "composed" convex programs associated to W^{p1}, where the objective function is made by the sum of a non-smooth function ($w \to \lambda\|w\|_1$) and a (possibly) smooth function ($w \to R(w)$).

[1] In the convex setting, due to the duality theory, the two problem formulations (constrained and penalized) are equivalent, in the sense that for any k there exists a λ such that the solutions of the two programs is the same. However, typically there is no explicit map from k to λ, and vice versa, and the penalized version is generally preferred in practice as it is more robust against misspecification of the penalty term

13.2 Convex Estimator: Lasso. Restricted Strong Convexity

As we saw last time, the restricted eigenvalue Assumption 12.2 is the key to establish Theorem 12.5. To establish an analogous result in the convex case, we need an assumption equivalent to Assumption 12.2 (which is stated with respect to the ℓ_0 pseudo-norm) with respect to the ℓ_1 norm. The right notion is given by restricted strong convexity over a cone constraint.

Assumption 13.1 (Restricted strong convexity) *Let $R : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $S = \operatorname{supp}(w^\star) := \{i \in [d] : w^\star_i \neq 0\}$ be the support of the vector $w^\star$, and let S^{C} denote its complement. Define the* cone set

$$\boxed{\mathcal{C} := \{w \in \mathbb{R}^d : \|w_{S^{\mathsf{C}}}\|_1 \leq 3\|w_S\|_1\}}$$

There exists $\alpha > 0$ such that

$$\boxed{R(w^\star + w) \geq R(w^\star) + \langle \nabla R(w^\star), w\rangle + \alpha\|w\|_2^2 \qquad \textit{for any } w \in \mathcal{C}} \tag{13.2}$$

Remark 13.2 (Connection with strong convexity) *The inequality* (13.2) *corresponds to the statement that the (random) function R restricted to the set $\mathcal{C}$ is lower-bounded by a quadratic function supported at $w^\star$. This property is* weaker *than the property of strong convexity of R on $\mathcal{C}$, as in the present case we only require the quadratic lower bound to be supported at a specific point $w^\star \in \mathcal{C}$, not for every point in $\mathcal{C}$.*

Remark 13.3 (Restricted strong convexity for the ℓ_2 norm) *If $R(w) = \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2$ then $\nabla R(w) = \frac{1}{n}\mathbf{x}^\top(\mathbf{x}w - Y)$ and Assumption 13.1 reads*

$$\boxed{\frac{1}{2n}\|\mathbf{x}w\|_2^2 \geq \alpha\|w\|_2^2 \qquad \textit{for any } w \in \mathcal{C}}$$

This is analogous to the bound in Assumption 12.2, with the key difference that in this case the inequality holds for any $w \in \mathcal{C}$, and $\mathcal{C}$ is defined via the ℓ_1 norm, while in the former case the inequality was taken to hold for any $w \in \mathbb{R}^d$ such that $\|w\|_0 \leq 2k$. Note that $\mathcal{C}$ is not convex, as it is a "two-sided" cone.

Assumption 13.1 yields the following result, which is the analogue of Theorem 12.5 for a convex estimator. A main difference, however, stems from the fact that the result below does *not* rely on the knowledge of an upper bound for the sparsity parameter ($\|w^\star\|_0 \leq k$) as in Theorem 12.5. This result shows how to tune the penalty parameter λ so that the estimator W^{p1} can provably achieve the mentioned guarantees.

Theorem 13.4 *Let W^{p1} be defined as in* (13.1)*. If Assumption 13.1 holds and $\lambda \geq 2\|\nabla R(w^\star)\|_\infty$, then*

$$\boxed{\|W^{p1} - w^\star\|_2 \leq \frac{3}{2}\frac{\lambda\sqrt{\|w^\star\|_0}}{\alpha}}$$

Proof: Let $\Delta = W^{p1} - w^\star$. The proof is divided into two parts.

Part 1: Prove that $\Delta \in \mathcal{C}$. By convexity of R we have

$$\begin{aligned} 0 &\leq R(W^{p1}) - R(w^\star) - \langle \nabla R(w^\star), \Delta\rangle \qquad (13.3)\\ &= R(W^{p1}) + \lambda\|W^{p1}\|_1 - \lambda\|W^{p1}\|_1 - R(w^\star) - \langle \nabla R(w^\star), \Delta\rangle \\ &\leq \lambda\|w^\star\|_1 - \lambda\|w^\star + \Delta\|_1 - \langle \nabla R(w^\star), \Delta\rangle, \end{aligned}$$

where the last inequality follows by using that, by the definition of W^{p1}, we have $R(W^{p1}) + \lambda\|W^{p1}\|_1 \le R(w^\star) + \lambda\|w^\star\|_1$. By Hölder's inequality and the fact that the ℓ_1 norm decomposes so that $\|w^\star + \Delta\|_1 = \|w^\star_S + \Delta_S\|_1 + \|w^\star_{S^\complement} + \Delta_{S^\complement}\|_1$, and $w^\star_{S^\complement} = 0$, we get

$$\begin{aligned} 0 &\le \lambda\|w^\star\|_1 - \lambda\|w^\star + \Delta\|_1 + \|\nabla R(w^\star)\|_\infty\|\Delta\|_1 \\ &\le \lambda\|w^\star\|_1 - \lambda\|w^\star_S + \Delta_S\|_1 - \lambda\|\Delta_{S^\complement}\|_1 + \|\nabla R(w^\star)\|_\infty\|\Delta\|_1. \end{aligned}$$

Using the assumption $\|\nabla R(w^\star)\|_\infty \le \frac{\lambda}{2}$ and the fact that the reverse triangle inequality yields $\|w^\star_S\|_1 - \|\Delta_S\|_1 \le \|w^\star_S + \Delta_S\|_1$, we get

$$0 \le \lambda\|\Delta_S\|_1 - \lambda\|\Delta_{S^\complement}\|_1 + \frac{\lambda}{2}\|\Delta\|_1 = \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^\complement}\|_1. \tag{13.4}$$

Rearranging this expression we obtain $3\|\Delta_S\|_1 \ge \|\Delta_{S^\complement}\|_1$, so $\Delta \in \mathcal{C}$.

Part 2: Prove the inequality. As $\Delta \in \mathcal{C}$, we can apply the restricted strong convexity assumption, Assumption 13.1, with $w = \Delta$ and we get

$$\alpha\|\Delta\|_2^2 \le R(W^{p1}) - R(w^\star) - \langle \nabla R(w^\star), \Delta\rangle,$$

which is analogous to (13.3) with 0 replaced by $\alpha\|\Delta\|_2^2$. Following the exact same steps as in Part 1, (13.4) now becomes

$$\alpha\|\Delta\|_2^2 \le \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^\complement}\|_1.$$

This yields, by the Cauchy-Schwarz's inequality,

$$\alpha\|\Delta\|_2^2 \le \frac{3\lambda}{2}\|\Delta_S\|_1 = \frac{3\lambda}{2}\langle \text{sign}(\Delta_S), \Delta_S\rangle \le \frac{3\lambda}{2}\sqrt{\|w^\star\|_0}\|\Delta_S\|_2 \le \frac{3\lambda}{2}\sqrt{\|w^\star\|_0}\|\Delta\|_2,$$

where we used that the cardinality of S is equal to $\|w^\star\|_0$, and that the ℓ_2 norm of a vector can only increase if we add non-zero coordinates to the vector. ■

As discussed in Remark 13.3, if $R(w) = \frac{1}{2n}\|\mathbf{x}w - Y\|^2$ then $\nabla R(w) = \frac{1}{n}\mathbf{x}^\top(\mathbf{x}w - Y)$ and $\nabla R(w^\star) = -\sigma\frac{\mathbf{x}^\top\xi}{n}$ so that $\|\nabla R(w^\star)\|_\infty = \sigma\frac{\|\mathbf{x}^\top\xi\|_\infty}{n}$. Therefore, Theorem 13.4 with the choice $\lambda = 2\|\nabla R(w^\star)\|_\infty$ yields

$$\|W^{p1} - w^\star\|_2 \le 3\frac{\sigma\sqrt{\|w^\star\|_0}}{\alpha}\frac{\|\mathbf{x}^\top\xi\|_\infty}{n}.$$

Recall that when the sparsity of the underlying vector is known, Theorem 12.5 with $k = \|w^\star\|_0$ yields

$$\|W^0 - w^\star\|_2 \le \sqrt{2}\frac{\sigma\sqrt{\|w^\star\|_0}}{\alpha}\frac{\|\mathbf{x}^\top\xi\|_\infty}{n},$$

which, modulo constant factors, is the same as the result above given by Theorem 13.4. In particular, this means that the bounds in expectation and in probability that we derived during the previous lecture for W^0 also apply for the convex estimator W^{p1}.

13.3 Restricted Strong Convexity: Sufficient Conditions

Restricted strong convexity, Assumption 13.1, is the key property that allows to establish Theorem 13.4. In general, given a certain problem instance (i.e., given a matrix $\mathbf{x} \in \mathbb{R}^{n\times d}$ and a certain function R), it is NP-hard to prove that this assumption holds. A sufficient condition that guarantees the restricted strong convexity for the ℓ_2 norm is given below. Checking if this sufficient condition holds is tractable if one knows $\|w^\star\|_0$. Let us recall the definition of the empirical second moment matrix:

$$\mathbf{c} = \frac{\mathbf{x}^\top \mathbf{x}}{n}.$$

Proposition 13.5 *For a matrix M, let $\|M\| := \max_{i,j} |M_{ij}|$. If*

$$\boxed{\|\mathbf{c} - I\| \leq \frac{1}{32\|w^\star\|_0}} \tag{13.5}$$

then Assumption 13.1 holds for $R(w) = \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2$ with $\alpha = \frac{1}{4}$. Namely, for any $w \in \mathcal{C}$, $\frac{1}{2n}\|\mathbf{x}w\|_2^2 \geq \frac{\|w\|_2^2}{4}$.

Proof: Let $w \in \mathcal{C}$. We have

$$\frac{1}{2n}\|\mathbf{x}w\|_2^2 = \frac{1}{2n} w^\top \mathbf{x}^\top \mathbf{x} w = \frac{1}{2} w^\top (\mathbf{c} - I) w + \frac{\|w\|_2^2}{2}.$$

Recall that Hölder's inequality gives $|a^\top b| \leq \|a\|_1 \|b\|_\infty$, or equivalently, $-\|a\|_1\|b\|_\infty \leq a^\top b \leq \|a\|_1 \|b\|_\infty$. Applying the lower bound we get, recalling the definition of the norm $\|\cdot\|$,

$$\frac{1}{2n}\|\mathbf{x}w\|_2^2 \geq \frac{\|w\|_2^2}{2} - \frac{\|w\|_1}{2}\|(\mathbf{c} - I)w\|_\infty \geq \frac{\|w\|_2^2}{2} - \frac{\|w\|_1^2}{2}\|\mathbf{c} - I\|.$$

As $w \in \mathcal{C}$, $\|w_{S^\complement}\|_1 \leq 3\|w_S\|_1$ and $S = \operatorname{supp}(w^\star) := \{i \in [d] : w_i^\star \neq 0\}$, by the Cauchy-Schwarz's inequality we have

$$\|w\|_1 = \|w_S\|_1 + \|w_{S^\complement}\|_1 \leq 4\|w_S\|_1 = 4\langle \operatorname{sign}(w_S), w_S \rangle \leq 4\sqrt{\|w^\star\|_0}\|w_S\|_2 \leq 4\sqrt{\|w^\star\|_0}\|w\|_2.$$

Hence, using the assumption of the proposition, we get

$$\frac{1}{2n}\|\mathbf{x}w\|_2^2 \geq \frac{\|w\|_2^2}{2} - 8\|w^\star\|_0\|w\|_2^2\|\mathbf{c} - I\| \geq \frac{\|w\|_2^2}{2} - \frac{\|w\|_2^2}{4} = \frac{\|w\|_2^2}{4}.$$

■

The quantity $\|\mathbf{c} - I\|$ is called the *incoherence parameter* in the compressed sensing literature. It is possible to show that (13.5) holds for many *random* ensembles, under mild assumptions. We now present the case of the Rademacher ensemble, which says that if the number of data n scales at least quadratically in the sparsity of the true parameter $\|w^\star\|_0$ and logarithmically in the dimension d, then the incoherence bound (13.5) can be made to hold with high probability.

Proposition 13.6 *Let $X \in \mathbb{R}^{n \times d}$ be a random matrix with independent Rademacher components, i.e., $\mathbf{P}(X_{ij} = 1) = \mathbf{P}(X_{ij} = -1) = 1/2$. If $n \geq 2048\tau\|w^\star\|_0^2 \log d$, then we have, for any $\tau \geq 2$,*

$$\boxed{\mathbf{P}\Big(\Big\|\frac{\mathbf{X}^\top \mathbf{X}}{n} - I\Big\| < \frac{1}{32\|w^\star\|_0}\Big) \geq 1 - \frac{2}{d^{\tau-2}}}$$

Proof: See Problem 4.1 in the Problem Sheets. ■

13.4 Beyond the Oracle Model. Proximal Gradient Methods

We now turn to the problem of *computing* the Lasso estimator, that is, solving a minimization problem of the following form:

$$\operatorname*{argmin}_{w\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2 + \lambda\|w\|_1. \tag{13.6}$$

Remark 13.7 (On convexity and multiple minima) *Note that the function $w \in \mathbb{R}^d \to H(w) := \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2 + \lambda\|w\|_1$ is convex, as $H(\delta w_1 + (1-\delta)w_2) \le \delta H(w_1) + (1-\delta)H(w_2)$ for any $\delta \in [0,1]$, $w_1, w_2 \in \mathbb{R}^d$. Note also that in the case $n < d$, in general the function H is not strictly convex and so does not have a unique minimum. However, if the entries of the matrix $\mathbf{x}$ are drawn from a continuous distribution, it is possible to show that the Lasso problem has a unique solution [4].*

For a given realization of the random variable Y, this problem has the following structure:

$$\boxed{\operatorname*{argmin}_{x\in\mathbb{R}^d} h(x) := f(x) + g(x)}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, and $g : \mathbb{R}^d \to \mathbb{R}$.

Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if at any point $x \in \mathbb{R}^d$ there exists a quadratic function with curvature β that uniformly upper-bounds f, namely,

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2 \qquad \text{for any } y \in \mathbb{R}^d. \tag{13.7}$$

In Lecture 9 we saw that in the smooth case gradient descent is the most natural algorithm we can think of: gradient descent is *defined* as the algorithm that at each time step moves to the point that maximizes the guaranteed local decrease given by the quadratic upper bound in (13.7), namely,

$$\operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2 \right\} = x - \frac{1}{\beta}\nabla f(x).$$

We can use the same idea if we want to minimize the function $f + g$ where f is β-smooth. In this case, the upper bound (13.7) given by smoothness yields, for any $x \in \mathbb{R}^d$,

$$f(y) + g(y) \le g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2 \qquad \text{for any } y \in \mathbb{R}^d$$

and the minimization of this upper bound can be written as follows

$$\begin{aligned}\operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2 \right\} &= \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ \frac{1}{\beta} g(y) + \frac{1}{2}\left\| y - \left(x - \frac{1}{\beta}\nabla f(x) \right) \right\|_2^2 \right\} \\ &= \mathtt{Prox}_{g/\beta}\left(x - \frac{1}{\beta}\nabla f(x) \right),\end{aligned}$$

where the *proximal operator* associated to a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\boxed{\mathtt{Prox}_\kappa(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ \kappa(y) + \frac{1}{2}\|y - x\|_2^2 \right\}}$$

This update defines the *proximal gradient method*, where the non-smooth component g is involved in the computations only through the associated proximal operator.

```
Algorithm 1: Proximal gradient method
Input: $x_1$, $\{\eta_s\}_{s\geq 1}$, stopping time $t$;
for $s = 1, \ldots, t$ do
    $x_{s+1} = \mathtt{Prox}_{\eta_s g}(x_s - \eta_s \nabla f(x_s))$
end
```

The proximal gradient method is a generalization of the gradient descent algorithm: when $g = 0$ we recover gradient descent with step size $\eta = \frac{1}{\beta}$. When $g(x) := 0$ for $x \in \mathcal{C}$ and $g(x) := +\infty$ for $x \notin \mathcal{C}$ we recover projected gradient descent on the set $\mathcal{C}$.

Proximal algorithms have recently attracted a lot of attention in machine learning due to their convergence rates and their ability to deal with large non-smooth convex problems. There are entire monographs devoted to them, such as [2, 3].

We will now see that if g is a convex function (possibly non-smooth), then the proximal gradient algorithm converges with a rate $O(1/t)$, which is better than the rate $O(1/\sqrt{t})$ we would obtain by applying the subgradient descent algorithm directly to minimize the function $h = f + g$, assuming that h is Lipschitz. As we discussed earlier, within the first order oracle model the rate $O(1/\sqrt{t})$ obtained by subgradient descent is optimal for a generic Lipschitz function. The proximal gradient algorithm can do better as it departs from the first order oracle model: in fact, to run this algorithm we assume to have access not only to the gradient of f, but also we assume to be able to solve the minimization problem that defines the proximal operator, for which we need *global* knowledge of the function g (it is not enough to know the gradient of g). Hence, the proximal gradient algorithm is an example of an algorithm that "goes beyond" the first order oracle model. There are many situations where we indeed have global knowledge of g. In the case of the Lasso estimator, for instance, we know that $g(x) = \lambda \|x\|_1$.

Theorem 13.8 (Proximal gradient method) *Let f be convex and β-smooth. Let g be convex. Assume $\|x_1 - x^\star\|_2 \leq b$. Then, the proximal gradient method applied to minimize $h = f + g$ with $\eta_s \equiv \eta = 1/\beta$ satisfies*

$$\boxed{h(x_t) - h(x^\star) \leq \frac{\beta b^2}{2(t-1)}}$$

The proof of Theorem 13.8 that we present relies on the following property.

Proposition 13.9 *For any $x, y \in \mathbb{R}^d$ we have*

$$h(y) - h(\rho(x)) \geq \frac{\beta}{2}\|\rho(x) - x\|_2^2 + \beta\langle x - y, \rho(x) - x\rangle,$$

where $\rho(x) := \mathtt{Prox}_{g/\beta}(x - \frac{1}{\beta}\nabla f(x))$.

Proof: For any $x, y \in \mathbb{R}^d$, let

$$\mu(x, y) := g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2,$$

$$\rho(x) := \operatorname*{argmin}_{y \in \mathbb{R}^d} \mu(x, y) = \mathtt{Prox}_{g/\beta}\left(x - \frac{1}{\beta}\nabla f(x)\right).$$

Fix any $x, y \in \mathbb{R}^d$. By smoothness of f we have $h(y) \le \mu(x,y)$ from which we have $h(\rho(x)) \le \mu(x,\rho(x))$ and

$$h(y) - h(\rho(x)) \ge h(y) - \mu(x,\rho(x)). \tag{13.8}$$

From the first order optimality conditions for the minimization problem $\min_{y\in\mathbb{R}^d} \mu(x,y)$ we known that $\rho(x)$ is the minimum if and only if there exists a subgradient of g evaluated at $\rho(x)$, $\bar{g} \in \partial g(\rho(x))$, such that

$$\bar{g} + \nabla f(x) + \beta(\rho(x) - x) = 0. \tag{13.9}$$

By convexity of both f and g we have

$$\begin{aligned} f(y) &\ge f(x) + \nabla f(x)^\top (y-x),\\ g(y) &\ge g(\rho(x)) + \bar{g}^\top (y - \rho(x)), \end{aligned}$$

so that, by summing, we find

$$h(y) \ge f(x) + \nabla f(x)^\top (y-x) + g(\rho(x)) + \bar{g}^\top (y-\rho(x)).$$

Plugging this lower bound for $h(y)$ into (13.8) and using the definition of $\mu(x,\rho(x))$, we find

$$h(y) - h(\rho(x)) \ge (\nabla f(x) + \bar{g})^\top (y - \rho(x)) - \frac{\beta}{2}\|\rho(x) - x\|_2^2.$$

By the optimality condition (13.9) we have $\nabla f(x) + \bar{g} = \beta(x - \rho(x))$ so that, by adding and subtracting x we find

$$h(y) - h(\rho(x)) \ge \beta(x-\rho(x))^\top (y-\rho(x)) - \frac{\beta}{2}\|\rho(x)-x\|_2^2 = \beta(x-\rho(x))^\top (y-x) + \frac{\beta}{2}\|\rho(x)-x\|_2^2.$$

■

Proof:[Proof of Theorem 13.8] Using Proposition 13.9 with $y = x^\star$ and $x = x_s$, noticing that $\rho(x_s) = x_{s+1}$, we find

$$\frac{2}{\beta}(h(x^\star) - h(x_{s+1})) \ge \|x_{s+1} - x_s\|_2^2 + 2\langle x_s - x^\star, x_{s+1} - x_s\rangle = \|x^\star - x_{s+1}\|_2^2 - \|x^\star - x_s\|_2^2,$$

which yields, summing from $s = 1$ to $s = t-1$,

$$\frac{2}{\beta}\Bigg((t-1)h(x^\star) - \sum_{s=1}^{t-1} h(x_{s+1})\Bigg) \ge \|x^\star - x_t\|_2^2 - \|x^\star - x_0\|_2^2. \tag{13.10}$$

Using again Proposition 13.9 with $y = x_s$ and $x = x_s$, we find

$$\frac{2}{\beta}(h(x_s) - h(x_{s+1})) \ge \|x_s - x_{s+1}\|_2^2,$$

which yields, multiplying it by $s-1$ and summing from $s = 1$ to $s = t-1$,

$$\frac{2}{\beta}\sum_{s=1}^{t-1}(s-1)(h(x_s) - h(x_{s+1})) \ge \sum_{s=1}^{t-1}(s-1)\|x_s - x_{s+1}\|_2^2.$$

The left hand side of this inequality can be rewritten as

$$\frac{2}{\beta}\sum_{s=1}^{t-1}((s-1)h(x_s) - sh(x_{s+1}) + h(x_{s+1})) = \frac{2}{\beta}\Bigg(- (t-1)h(x_t) + \sum_{s=1}^{t-1} h(x_{s+1})\Bigg),$$

which yields

$$\frac{2}{\beta}\bigg(-(t-1)h(x_t) + \sum_{s=1}^{t-1} h(x_{s+1}) \bigg) \geq \sum_{s=1}^{t-1}(s-1)\|x_s - x_{s+1}\|_2^2. \tag{13.11}$$

Summing (13.10) and (13.11) we find

$$\frac{2(t-1)}{\beta}(h(x^\star) - h(x_t)) \geq \|x^\star - x_t\|_2^2 - \|x^\star - x_0\|_2^2 + \sum_{s=1}^{t-1}(s-1)\|x_s - x_{s+1}\|_2^2 \geq -\|x^\star - x_0\|_2^2,$$

which yields the statement of the theorem. ■

As the gradient descent algorithm for smooth functions can be accelerated to yield a $O(1/t^2)$ convergence rate, also the proximal gradient method can be accelerated to yield a $O(1/t^2)$ convergence rate.

The improved rate of convergence of the proximal method over subgradient descent applied to $f+g$ comes with a price, however, as in general the minimization problem that we need to solve to compute the proximal operator can be a difficult problem by itself (note that the objective function to be minimized is strongly convex, so the solution is unique). When g is decomposable, i.e., $g(x) = \sum_{i=1}^d g_i(x_i)$ with $g_i : \mathbb{R} \to \mathbb{R}$, then the computation of the proximal operator involves solving d one-dimensional convex optimization problems. This is the case for the Lasso operator, as $g(x) = \lambda\|x\|_1 = \lambda\sum_{i=1}^d |x_i|$.

In general, if $\kappa(x) = \sum_{i=1}^d \kappa_i(x_i)$ with $\kappa_i : \mathbb{R} \to \mathbb{R}$ we have

$$\texttt{Prox}_\kappa(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \bigg\{ \sum_{i=1}^d \kappa_i(y_i) + \frac{1}{2}\sum_{i=1}^d (y_i - x_i)^2 \bigg\} = \begin{pmatrix} \operatorname{argmin}_{y\in\mathbb{R}}\{\kappa_1(y) + \frac{1}{2}(y-x_1)^2\} \\ \vdots \\ \operatorname{argmin}_{y\in\mathbb{R}}\{\kappa_d(y) + \frac{1}{2}(y-x_d)^2\} \end{pmatrix} \equiv \begin{pmatrix} \texttt{Prox}_{\kappa_1}(x_1) \\ \vdots \\ \texttt{Prox}_{\kappa_d}(x_d) \end{pmatrix}$$

where we implicitely defined the one-dimensional proximal operator. In some cases, the one-dimensional proximal operators can be computed analytically, as for the Lasso estimator.

13.5 Computing the Lasso Estimator

We now apply the proximal gradient method to solve the Lasso problem (13.6). In this case, the algorithm takes the name of *Iterative Shrinkage-Thresholding Algorithm* (ISTA) and its accelerated variant takes the name of *Fast Iterative Shrinkage-Thresholding Algorithm* (FISTA).

In the case of the Lasso, we have, for any $\theta > 0$ and $w \in \mathbb{R}$,

$$\iota(w;\theta) := \texttt{Prox}_{\theta|\,\cdot\,|}(w) = \operatorname*{argmin}_{y\in\mathbb{R}} \bigg\{ \theta|y| + \frac{1}{2}(y-w)^2 \bigg\} = \operatorname{sign}(w)\max\{|w|-\theta, 0\} = \begin{cases} w-\theta & \text{if } w > \theta \\ 0 & \text{if } -\theta \leq w \leq \theta \\ w+\theta & \text{if } w < -\theta \end{cases}$$

which is the so-called *soft-thresholding operator.*

If $f(w) = R(w) = \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2$ and $g(w) = \lambda\|w\|_1$, then $\nabla R(w) = \frac{1}{n}\mathbf{x}^\top(\mathbf{x}w - Y)$ and the proximal gradient method (ISTA) reads as follows (for a vector $w \in \mathbb{R}^d$, we write $\iota(w;\theta)$ to mean the vector $(\iota(w_1;\theta), \ldots, \iota(w_d;\theta))$).

```
Algorithm 2: Iterative Shrinkage-Thresholding Algorithm (ISTA)
Input: $W_1$, $\{\eta_s\}_{s\geq 1}$, stopping time $t$;
for $s = 1, \dots, t$ do
    $W_{s+1} = \mathtt{Prox}_{\eta_s g}(W_s - \eta_s \nabla R(W_s)) \equiv \iota\left(W_s - \frac{\eta_s}{n}\mathbf{x}^\top(\mathbf{x}W_s - Y); \lambda\eta_s\right)$
end
```

Note that the function R is β-smooth, with $\beta = \mu_{\max}(\frac{1}{n}\mathbf{x}^\top\mathbf{x})$ being the largest eigenvalue of the matrix $\nabla^2 R(w) = \frac{1}{n}\mathbf{x}^\top\mathbf{x} = \mathbf{c}$. Therefore, if we choose $\eta_s \equiv \eta = 1/\beta$, Theorem 13.8 guarantees the following: (recall that W^{p1} is by definition the solution of the Lasso optimization problem)

$$R(W_t) + \lambda\|W_t\|_1 - (R(W^{p1}) + \lambda\|W^{p1}\|_1) \leq \mu_{\max}\left(\frac{1}{n}\mathbf{x}^\top\mathbf{x}\right)\frac{\|W_1 - W^{p1}\|_2^2}{2(t-1)}. \tag{13.12}$$

Remark 13.10 *Note that the function R is* not *strongly convex, as if $n < d$ then the null space of the matrix $\mathbf{x}^\top\mathbf{x}$ is not-empty and thus $\alpha = \mu_{min}(\frac{1}{n}\mathbf{x}^\top\mathbf{x}) = 0$.*

Note that our ultimate goal is to bound the deviation $\|W_t - w^\star\|_2$, so in principle we would like to use the following inequality:

$$\boxed{\|W_t - w^\star\|_2 \leq \underbrace{\|W_t - W^{p1}\|_2}_{\text{optimization error}} + \underbrace{\|W^{p1} - w^\star\|_2}_{\text{statistics error}}}$$

and combine the guarantees given on the statistical error by Theorem 13.4 with the guarantees given for the optimization algorithm that we run. However, Theorem 13.8 only yields a bound on the error of the objective function $h(x_t) - h(x^\star)$—c.f. (13.12) for the Lasso problem—not an error on the iterate $\|x_t - x^\star\|_2$ themselves. In fact, it is possible to prove [1] that under the restricted strong convexity property the *iterates* of the ISTA algorithm applied to the Lasso problem convergence exponentially fast up to the statistical precision of the problem, namely,

$$\mathbf{P}\left(\underbrace{\|W_t - W^{p1}\|_2}_{\text{optimization error}} \lesssim e^{-ct} + \underbrace{\|W^{p1} - w^\star\|_2}_{\text{statistics error}}\right) \geq 1 - \delta,$$

where $\lesssim$ mean inequality up to constants ("constant" that do depend on the parameter of the model), for a given $\delta > 0$.

In the next lecture we will explore another way to approach this type of problems where we do not separate between statistics and optimization: the implicit regularization approach.

References

[1] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 10 2012.

[2] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, January 2012.

[3] Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014.

[4] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013.