5.2 ℓ_1-regularized logistic regression

Least square may be interpreted as maximum likelihood estimation (MLE) where each $b_i \in \mathbb{R}$ is the realization of a Normal random variable with mean $a_i^T x$ and variance 1. Here a_i^T denotes the ith row of A. For classification problems, each b_i takes the value of either 1 or 0 (instead of a continuum of values). Analogously, logistic regression corresponds to MLE where each $b_i \in \{1, 0\}$ is the realization of a random variable β_i with distribution

$$\mathrm{P}[\beta_i = 1] = \frac{1}{1 + e^{-a_i^T x}}, \qquad \mathrm{P}[\beta_i = 0] = 1 - \mathrm{P}[\beta_i = 1] = \frac{1}{1 + e^{a_i^T x}}.$$

The negative log-likelihood function works out to be $\ell(Ax)$, where

$$\ell(u) = \sum_{i=1}^{m} \log\left(1 + e^{u_i}\right) - b_i u_i. \tag{83}$$

To avoid overfitting and for variable/feature selection (each column of A may correspond to an input variable or a feature), we seek a sparse MLE solution by solving

$$\min_x \ell(Ax) + \tau \|x\|_1. \tag{84}$$

5.3 TV-regularized image denoising

Images recorded from distance (by satellites or telescopes) and medical images (X-ray or PET scan or ultrasound) have significant noise. How to denoise such a noisy image without oversmoothing the key features (outlines, sharp edges) is a fundamental problem in signal processing. One such approach, studied by Stan Osher and others, is to use total-variation (TV) regularization. Specifically, for a given noisy image $b : \Omega \to \mathbb{R}$, with $\Omega \subseteq \mathbb{R}^2$ the image domain, it solves

$$\min_u \frac{1}{2} \int_\Omega |u(x) - b(x)|^2 dx + \tau \int_\Omega \|\nabla u(x)\|_2 dx,$$

with $\tau > 0$ a user-chosen parameter that trades off between edge-preservation (large τ) and least-square fit (small τ).

To solve the above problem numerically, we discretize the image domain Ω. For simplicity, assume Ω is a square and we discretize it with an $N \times N$ grid of width $h > 0$. Letting u_{ij} to denote the u-value at the (i,j)th grid point, we use forward finite-difference to evaluate $\nabla u(x)$ there:

$$\nabla u(x)_{ij} \approx \begin{bmatrix} \begin{cases} \frac{u_{i,j+1} - u_{ij}}{h} & \text{if } j < N \\ 0 & \text{else} \end{cases} \\ \begin{cases} \frac{u_{i+1,j} - u_{ij}}{h} & \text{if } i < N \\ 0 & \text{else} \end{cases} \end{bmatrix}.$$

We use Rieman sum to evaluate the integrals, resulting in the discretized problem:

$$\min_{u \in \mathbb{R}^{N^2}} \frac{1}{2} \sum_{1 \le i,j \le N} |u_{ij} - b_{ij}|^2 h + \tau \sum_{1 \le i,j \le N} \left\| \frac{A^{ij} u}{h} \right\|_2 h,$$

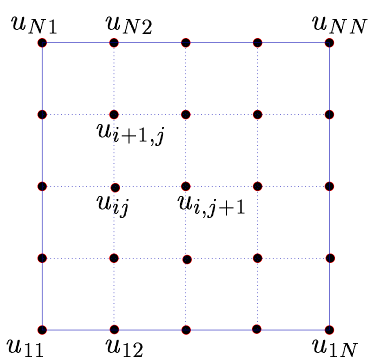
where $u = (u_{11}, u_{12}, \ldots, u_{NN})^T$ and we let

$$A^{ij}u = \begin{bmatrix} \begin{cases} u_{i,j+1} - u_{ij} & \text{if } j < N \\ 0 & \text{else} \end{cases} \\ \begin{cases} u_{i+1,j} - u_{ij} & \text{if } i < N \\ 0 & \text{else} \end{cases} \end{bmatrix}. \tag{85}$$

Dividing the objective function by h and letting $\tau' = \frac{\tau}{h}$, the discretized problem can be written more simply as

$$\min_{u \in \mathbb{R}^{N^2}} \quad \frac{1}{2}\|u - b\|_2^2 + \tau' \sum_{1 \le i,j \le N} \|A^{ij}u\|_2. \tag{86}$$

Here we view u as a vector for notational simplicity. For computation, it may be more convenient to represent u as an $N \times N$ matrix, so that forward finite-differencing can be implemented by a row and column shift and then differencing.



The objective function in (86) is convex, but $\|\cdot\|_2$ is not differentiable, which complicates its structure. It can be reformulated as an SOCP and solved by an interior-point method, but we will see faster methods for solving it. Specifically, the dual of this problem has a simpler structure that can be exploited. To get the dual, we introduce new variables $y^{ij} = A^{ij}u$ and rewrite (86) as

$$\begin{aligned} \min_{u,y} \quad & \frac{1}{2}\|u - b\|_2^2 + \tau' \sum_{i,j} \|y^{ij}\|_2 \\ \text{s.t.} \quad & y^{ij} = A^{ij}u \quad \forall i, j. \end{aligned}$$

We then form the Lagrangian, with Lagrange multipliers $x^{ij} \in \mathbb{R}^2$ associated with the constraints:

$$L(u, y, x) = \frac{1}{2}\|u - b\|_2^2 + \tau' \sum_{i,j} \|y^{ij}\|_2 + \sum_{i,j} \langle y^{ij} - A^{ij}u, x^{ij} \rangle.$$

Intuitively, x^{ij} acts as a variable penalty that penalizes violation of the constraint $y^{ij} = A^{ij}u$. In particular, the primal problem (86) is equivalent to $\min_{u,y} \max_{x} L(u, y, x)$, where $x = (x_{11}, x_{12}, \ldots, x_{NN})^T$. The dual problem is $\max_{x} \min_{u,y} L(u, y, x)$, which works out to be

$$\begin{aligned} \max_{x \in \mathbb{R}^{2N^2}} \quad & -\frac{1}{2}\left\| \sum_{1 \le i,j \le N} (A^{ij})^* x^{ij} \right\|_2^2 - \left\langle b, \sum_{1 \le i,j \le N} (A^{ij})^* x^{ij} \right\rangle \\ \text{s.t.} \quad & \|x^{ij}\|_2 \le \tau' \quad \forall 1 \le i, j \le N. \end{aligned} \tag{87}$$

The objective function is quadratic and the constraint is a Cartesian product of Euclidean balls. However, the problem size is large. For a 4 Mega-pixel image, the number of variables is 8 million!

5.4 Matrix rank minimization

A matrix analog of the compressed sensing problem (1) is that of rank minimization:

$$\begin{array}{ll} \min\limits_{X \in \mathbb{S}^n} & \operatorname{rank}(X) \\ \text{s.t.} & \mathcal{A}X = b, \end{array} \tag{88}$$

where $\mathcal{A}X = [\langle A_i, X\rangle]_{i=1}^m$ with $A_i \in \mathbb{S}^n$ and $b \in \mathbb{R}^m$. This has applications in control and systems theory, such as model reduction, minimum order control synthesis; see the work of Boyd, Fazel, Candès, and others. In the so-called matrix completion problem, we seek the lowest rank matrix with certain entries given. This problem, like (1), can be shown to be NP-hard. Here we consider $X \in \mathbb{S}^n$, but the discussions readily generalize to rectangular matrices $X \in \mathbb{R}^{p\times n}$.

A convex approximation of (88), analogous to (2), is

$$\begin{array}{ll} \min\limits_{X \in \mathbb{S}^n} & \|X\|_{\text{nuc}} \\ \text{s.t.} & \mathcal{A}X = b, \end{array} \tag{89}$$

where $\|X\|_{\text{nuc}} = \sum_i \sigma_i(X)$ ("nuclear" norm) and $\sigma_1(X), \ldots, \sigma_n(X)$ are the singular values of X. Thus, nuclear norm is simply the 1-norm of the singular values. Since $X \in \mathbb{S}^n$, we have

$$\sigma_i(X) = \sqrt{\lambda_i(X^2)} = |\lambda_i(X)|.$$

The problem (89) can be reformulated as an SDP by using a fact that

$$\begin{array}{lll} \|X\|_{\text{nuc}} & = & \min\limits_{W,Z} \ \frac{1}{2}(\operatorname{tr}[W] + \operatorname{tr}[Z]) \\ & & \text{s.t.} \ \begin{bmatrix} W & X \\ X & Z \end{bmatrix} \succeq 0. \end{array}$$

This can be shown by verifying that $W = Z = (X^2)^{1/2}$ is feasible for this problem, and that $\operatorname{Sign}(X)$ is feasible for its dual with the same objective function value, where $\operatorname{Sign}(X)$ is obtained from X by replacing the eigenvalues in its eigen-decomposition by their signs. The SDP can be solved by primal-dual interior-point method, but the work per iteration is $O(n^4)$ operations, which limits the size of problems solvable.

If b is noisy, then we consider, analogous to (82), the regularized least-square problem

$$\min_X \frac{1}{2}\|\mathcal{A}X - b\|_2^2 + \tau\|X\|_{\text{nuc}}, \tag{90}$$

with $\tau > 0$. This problem can be reformulated as a quadratic SDP.

6 First-Order Gradient Methods

Looking at the application problems of the previous section, we see that they mostly have the following form:

$$\min_x \quad f_P(x) := f(x) + P(x), \tag{91}$$

where $P : \mathbb{H} \to \mathbb{R} \cup \{\infty\}$ is proper, closed, convex, and $f : \mathbb{H} \to \mathbb{R}$ is differentiable, convex, and ∇f is Lipschitz continuous on dom$P$, i.e., there exists scalar $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \text{dom}P. \tag{92}$$

Recall that $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$, i.e., $\mathbb{H}$ is a Hilbert space (although the subsequent development readily extend to $\mathbb{H}$ being a real Banach space). Moreover, in the application problems, P is simply structured, which will be key to the efficient solution of (91). We discuss this in more detail below.

1. The lasso problem (82) corresponds to

$$\mathbb{H} = \mathbb{R}^n, \quad f(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad P(x) = \tau\|x\|_1.$$

   Thus dom$P = \mathbb{R}^n$ and, by the chain rule for differentiation, $\nabla f(x) = A^T(Ax - b)$, which is computable in $O(mn)$ operations (or less if A is sparse or structured, such as when Ax and $A^T u$ are computable by fast Fourier transform). Moreover,

$$\|\nabla f(x) - \nabla f(y)\|_2 = \|A^T(Ax - Ay)\|_2 \leq \lambda_{\max}(A^T A)\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n.$$

2. The regularized logistic regression problem (84) corresponds to

$$\mathbb{H} = \mathbb{R}^n, \quad f(x) = \ell(Ax), \quad P(x) = \tau\|x\|_1,$$

   with ℓ given by (83). Thus dom$P = \mathbb{R}^n$ and it can be verified that $\nabla \ell$ is Lipschitz continuous (with constant $\frac{1}{4}$, I think), so that ∇f is Lipschitz continuous (with constant $\frac{\lambda_{\max}(A^T A)}{4}$, I think). Moreover, $\nabla f(x)$ is computable in $O(mn)$ operations or less.

3. The dual TV-regularized problem (87) corresponds to

$$\mathbb{H} = \mathbb{R}^{2N^2}, \quad f(x) = \frac{1}{2}\left\|\sum_{i,j} A^{ij} x^{ij}\right\|_2^2 + \left\langle b, \sum_{i,j} A^{ij} x^{ij} \right\rangle, \quad P(x) = \begin{cases} 0 & \text{if } \|x^{ij}\|_2 \leq \tau' \quad \forall i, j \\ \infty & \text{else.} \end{cases}$$

   Notice that P is closed and convex since it is the indicator function for a closed convex set, namely, the Cartesian product of closed Euclidean balls. Also, using the sparsity structure of A^{ij} (see (85)), $\nabla f(x)$ is computable in $O(N^2)$ operations.

4. The regularized least-square problem (90) corresponds to

$$\mathbb{H} = \mathbb{S}^n, \quad f(X) = \frac{1}{2}\|\mathcal{A}X - b\|_2^2, \quad P(X) = \tau\|X\|_{\text{nuc}}.$$

   The work to compute $\nabla f(X)$ depends on $\mathcal{A}$ and $\mathcal{A}^*$.

6.1 Partial linear approximation

So (91) has nice large-scale applications. How to solve it? As always, we approximate a complex problem by a simpler problem. Here, we will exploit the properties that f is differentiable and P, though nondifferentiable, is simply structured. Specifically, we will approximate f locally to first order by its linearization at a given $x \in \mathrm{dom}P$:

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + o(\|y - x\|).$$

Adding $P(y)$ to both sides yields the approximation $f_P(y) = \ell(y; x) + o(\|y - x\|)$, where we let

$$\ell(y; x) = f(x) + \langle \nabla f(x), y - x \rangle + P(y). \tag{93}$$

(We do not approximate P since it is already simply structured.)

The above discussion suggests a simple method for solving (91): Given $x \in \mathrm{dom}P$, solve the (partial) linearization

$$\min_y \; \ell(y; x) \tag{94}$$

to obtain a new x_+ and re-iterate. What's wrong with this? For one, the minimum may not exist. This is certainly true if, say, $P \equiv 0$. But it's also true if P is coercive (i.e., its level set $\{x \mid P(x) \le \alpha\}$ is bounded). An example is $\min_{x \in \mathbb{R}} \frac{1}{2}x^2 + |x|$; see (82). At $x = 2$, $\ell(y; x) = 2 + 2(y - 2) + |y|$ has no minimum. Even if a minimum exists, the minimizing y may be far from x, so that $\ell(y; x)$ poorly approximates $f_P(y)$ (although this can sometimes be remedied by performing a line search on the line segment joining x and y).

How to ensure a minimizing y exists and is not far from x? We can add a proximity term between x and y to (94). The simplest such term is the quadratic $\frac{1}{2}\|y - x\|^2$. (An alternative is $\|y - x\|$, but it is not differentiable nor separable.) Scaling this by L yields a second-order approximation (since L is a bound on the rate of change in the gradient). This results in

$$\min_y \; \ell(y; x) + \frac{L}{2}\|y - x\|^2. \tag{95}$$

The objective function is strictly convex and coercive (due to the quadratic proximal term), so it has a unique minimizer. Let's see some examples of P for which the minimizer is easy to compute. Letting $g = \nabla f(x)$ and using (93), this simplifies to

$$\min_y \; \langle g, y \rangle + P(y) + \frac{L}{2}\|y - x\|^2. \tag{96}$$

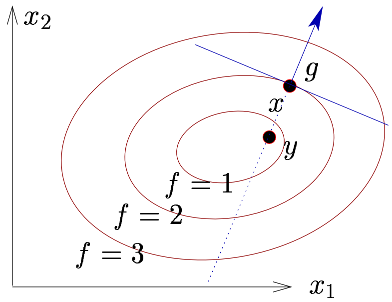
1. Suppose $P \equiv 0$. Then (96) reduces to

   $$\min_y \; \langle g, y \rangle + \frac{L}{2}\|y - x\|^2.$$

   The objective function is quadratic. By either completing the square or differentiating with respect to y and setting it to 0, we obtain that the minimizing y satisfies $g + L(y - x) = 0$ and hence

   $$y = x - \frac{g}{L}.$$

   This is Cauchy's steepest descent method, with stepsize $\frac{1}{L}$ (so a larger L means a smaller step).



2. Suppose $\mathbb{H} = \mathbb{R}^n$ and $P(y) = \tau\|y\|_1$ (as in (82) and (84)). Then (96) reduces to
$$\min_y \langle g, y\rangle + \tau\|y\|_1 + \frac{L}{2}\|y - x\|^2 \;=\; \min_y \sum_{i=1}^{n} g_i y_i + \tau|y_i| + \frac{L}{2}(y_i - x_i)^2.$$
The objective function is separable, so we can minimize over each y_i independently. By considering the three cases $y_i > 0$, $y_i < 0$, $y_i = 0$, it is straightforward to check that the minimizing y_i is given by the formula
$$y_i = \text{median}\left\{x_i - \frac{g_i + \tau}{L}, x_i - \frac{g_i - \tau}{L}, 0\right\}.$$
Then the minimizing y can be found in only $O(n)$ operations.

3. Suppose $\mathbb{H} = \mathbb{R}^{2n}$ and $P(y) = \begin{cases} 0 & \text{if } \|y_\ell\|_2 \le \tau \quad \forall \ell \\ \infty & \text{else} \end{cases}$, where $y = (y_\ell)_{\ell=1}^n$ with $y_\ell \in \mathbb{R}^2$ and $\tau > 0$ (as in (87)). Then (96) reduces to
$$\begin{array}{ll} \min_y & \langle g, y\rangle + \frac{L}{2}\|y - x\|_2^2 \\ \text{s.t.} & \|y_\ell\|_2 \le \tau \quad \forall \ell \end{array} \;=\; \begin{array}{ll} \min_y & \sum_\ell \langle g_\ell, y_\ell\rangle + \frac{L}{2}\|y_\ell - x_\ell\|_2^2 \\ \text{s.t.} & \|y_\ell\|_2 \le \tau \quad \forall \ell \end{array}$$
$$= \begin{array}{ll} \min_y & \sum_\ell \frac{L}{2}\|y_\ell - (x_\ell - \frac{g_\ell}{L})\|_2^2 + \cdots \\ \text{s.t.} & \|y_\ell\|_2 \le \tau \quad \forall \ell \end{array}$$
where the second equality is obtained by completing the square, and "$\cdots$" denotes terms independent of y. The objective function is separable, so we can minimize over each y_ℓ independently, yielding y_ℓ as the nearest-point projection of $x_\ell - \frac{g_\ell}{L}$ onto the Euclidean ball of radius τ. Thus the minimizing y_ℓ is given by the formula
$$y_\ell = \begin{cases} x_\ell - \frac{g_\ell}{L} & \text{if } \|x_\ell - \frac{g_\ell}{L}\| \le \tau \\ \tau \dfrac{x_\ell - \frac{g_\ell}{L}}{\|x_\ell - \frac{g_\ell}{L}\|_2} & \text{else.} \end{cases}$$
Then the minimizing y can be found in only $O(n)$ operations.

4. Suppose $\mathbb{H} = \mathbb{S}^n$ and $P(Y) = \tau\|Y\|_{\text{nuc}}$ (as in (89)). Then (96) reduces to
$$\min_Y \langle G, Y\rangle + \tau\|Y\|_{\text{nuc}} + \frac{L}{2}\|Y - X\|_F^2.$$
It can be shown that the minimizing Y can be computed from an eigen-decomposition of $X - \frac{G}{L}$ in $O(n^3)$ operations.