Lecture 8

Last time

- Quasi-Newton methods
- BFGS

Today

- Gauss-Newton
- Levenberg-Morrison-Marquardt

Gauss-Newton

Today we will cover one more family of preconditioners that tackle, a specific family known as "nonlinear least squares. These, are problems like

$$(\heartsuit) \qquad \min \underbrace{\frac{1}{2}\|r(w)\|^2}_{f(w)} \qquad r: \mathbb{R}^d \to \mathbb{R}^n \text{ differentiable}$$

Example: Imagine we want to learn a map $F: \mathbb{R}^m \to \mathbb{R}$ and we observe input-output pairs $(x_1, y_1), \dots, (x_n, y_n)$, where $y_i = F(x_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$.

If we have a parametric family of functions $\{F_w(\cdot) \mid w \in \mathbb{R}^d\}$, a natural strategy (MLE) is to

$$\min_w \frac{1}{2n} \sum_{j=1}^{n} (F_w(x_i) - y_i)^2 = \frac{1}{2} \left\| \underbrace{\frac{1}{\sqrt{n}} \begin{pmatrix} F_w(x_1) - y_1 \\ \vdots \\ F_w(x_n) - y_n \end{pmatrix}}_{r(w)} \right\|_2^2$$

Our idea to derive this method will also be to find a good model to minimize

$$w_{k+1} \leftarrow \operatorname{argmin}_w \left\{ \frac{1}{2} \| \underbrace{r(w_k) + \nabla r(w_k)(w - w_k)}_{\text{Taylor approximation}} \|^2 \right\}.$$

(Jacobian: $\nabla r(w_k)$)

This model is always convex, so although there might be more than one minimizer, we can use first order optimality,

$$0 = \nabla r(w_k)^T (r(w_k) + \nabla r(w_k)(w_{k+1} - w_k))$$

$$\Leftarrow$$

$$w_{k+1} = w_k - \underbrace{(\nabla r(w_k)^T \nabla r(w_k))^{-1} \nabla r(w_k)^T}_{\nabla r(w_k)^+} r(w_k).$$

(whenever defined)

There is a unique solution when $\nabla r(w_k)^T \nabla r(w_k)$ is invertible.

A slightly different interpretation

Let's compute the Hessian of $f(\cdot)$

$$\nabla^2 f(w) = \nabla r(w)^T \nabla r(w) + \sum_{j=1}^{n} \nabla^2 r_j(w)\, r_j(w)$$

ouch

The second term is horrible. Fortunately it should be close go to zero as $r(w) \to 0$, which is approximately true for least squares. Let's show that this method emulates Newton's.

Theorem (U): Consider (*) and suppose the following holds in a ball of radius δ near a minimizer w^*.

(1) We have that $\nabla r(w)^T \nabla r(w) \geq \sigma^2 I$.

(2) The Jacobian ∇r is β-Lipschitz.

(3) The residual at the solution is zero.

(4) $\delta < \frac{2\sigma}{\beta}$

Then, the iterates of (GN) give

$$\|w_{k+1} - w^*\| \leq \frac{\beta}{2\sigma} \|w_k - w^*\|^2 \quad \forall k.$$

provided $\|w_0 - w^*\| \leq \delta$.

Proof: Define $e_k = w_k - w^*$. Assume that $\|e_k\| < \delta$. We shall prove that $\|e_{k+1}\| < \delta$.

Expanding,

$$e_{k+1} = e_k - (\nabla r(w_k)^T \nabla r(w_k)^{-1} \nabla r(w_k)\, r(w_k)$$

$J^TJ = I$ ↷

$$= \nabla r(w_k)^\dagger (\nabla r(w_k)\, e_k - r(w_k)).$$

Using the Taylor bound from HW1

$$\|\underbrace{r(w^*)}_{0} - r(w_k) + \nabla r(w_k)\, e_k\| \leq \frac{\beta}{2} \|e_k\|^2$$

Combining these two gives

$$\|e_{k+1}\| \leq \frac{\beta}{2} \|\nabla r(w_k)^\dagger\| \, \|e_k\|^2$$

$\|\nabla r(w_k)^\dagger\| \leq \frac{1}{\sigma}$ (why?) ↷

$$\leq \frac{\beta}{2\sigma} \|e_k\|^2.$$

□

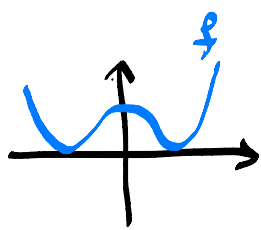
If we don't impose (3), you can easily prove that the convergence is linear.

Lemma: Suppose that $f = \frac{1}{2}\|r(\cdot)\|^2$ is μ-strongly convex and $r(w^*) = 0$. Then (1) holds with $\sigma = \sqrt{\mu/2}$. +

A very important issue with this method is that it can be super unstable when $\nabla r(w_k)^T \nabla r(w_k)$ is nearly singular.

Example: Consider the problem

$$f(t) = \frac{1}{2}\underbrace{(t^2-1)}_{r(t)}^2$$



Then $(\nabla r(t)^T \nabla r(t))^{-1} = \frac{1}{4t^2}$ so

$$t_{k+1} \leftarrow t_k - \frac{t_k^2-1}{2t_k} = \frac{t_k^2+1}{2t_k} = \frac{1}{2}\left(t_k + \frac{1}{t_k}\right)$$

Hence,

$$\lim_{t_1 \downarrow 0} t_1 = +\infty.$$ +

This motivates putting a cap on how big

the spectrum of $(\nabla r(w_k)^T \nabla r(w_k))^{-1}$ can be.

Levenberg - Morrison - Marquardt

To do so we add a regularizer to our model, namely

$$(LMM)\quad w_{k+1} \leftarrow \arg\min \frac{1}{2}\| r(w_k) + \nabla r(w_k)(w - w_k)\|^2 + \frac{\lambda_k}{2}\|w - w_k\|^2$$

This has close connections with ridge regression and 1st order optimality gives us

$$w_{k+1} \leftarrow w_k - (\nabla r(w_k)^T \nabla r(w_k) + \lambda_k I)^{-1} \nabla r(w_k)^T r(w_k).$$

In turn, the way set λ_k is very important. The following result captures this fact. We assuming a cap in the damping coefficient

$$(\Theta) \qquad \lambda_k \le \bar{\lambda}.$$

Theorem: Consider (Θ) and suppose the following holds in a ball of radius

δ near a minimizer w^*.

(1) We have that $\nabla r(w)^T \nabla r(w) \geq \sigma^2 I$.

(2) The Jacobian ∇r is β-Lipschitz.

(3) The residual at the solution is zero.

(4) We set $\lambda_k \leq \bar{\lambda}$.

(5) $\delta < 2\left(1 - \frac{\bar{\lambda}}{\sigma^2 + \bar{\lambda}}\right) \cdot \frac{\sigma}{\beta}$.

Then, the iterates of (LLM) give

$$\|w_{k+1} - w^*\| \leq \frac{\lambda_k}{\sigma^2 + \lambda_k} \|w_k - w^*\| + \frac{\beta}{2\sigma} \|w_k - w^*\|^2 \quad \forall k,$$

provided $\|w_0 - w^*\| \leq \delta$. ⊣

The proof is similar to that of Theorem (U). This theorem teaches us that at least under these assumptions, it makes sense to set

$$\lambda_k \approx \|w_k - w^*\|.$$

← Unrealistic

In practice people often use

$$\lambda_k = \|r(w_k)\|,$$ ← observable

One can extend the previous theorem to prove a guarantee for $\|r(w_k)\|$, which motivates this choice. (think how!)

History Trivia

People often call the (LMM) method, Levenberg-Marquardt. Philippe Toint told me that we met Morrison and we was a bit sour about. Let's give the guy some historic credit back.

A **method** for the **solution** of **certain non-linear problems** in **least squares**
K Levenberg - Quarterly of applied mathematics, 1944 - ams.org
... The standard **method** for solving **least squares problems** which lead to **non-linear** normal **equations** depends upon ... taken about an initial or trial **solution** for the parameters.2 If the usual ...
☆ Save 99 Cite Cited by 19709 Related articles All 8 versions

[PDF] An **algorithm** for **least-squares estimation** of **nonlinear parameters**

DW Marquardt - Journal of the society for Industrial and Applied …, 1963 - SIAM

… In this paper attention is confined to the **algorithm** for obtaining **leastsquares estimates**. For discussion of the broader statistical aspects of **nonlinear estimation** and applicationto specific …

☆ Save 99 Cite Cited by 44065 Related articles All 6 versions Web of Science: 25736

[CITATION] Methods for nonlinear least squares problems and convergence proofs

DD Morrison - JPL SEMINAR PROC. ON TRACKING PROGRAMS …, 1960 - ntrs.nasa.gov

Methods for **nonlinear least squares problems** and **convergence proofs** - NASA Technical Reports Server (NTRS) …

☆ Save 99 Cite Cited by 105 Related articles »