28 Gauss-Newton

Goal

Composite minimization, Gauss-Newton, nonlinear least squares, Levenberg-Marquardt regularization, prox-linear

Alert 28.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 28.2: Composite minimization

We are interested in solving the composite minimization problem:

$$\min_{\mathbf{w}} \ f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) := \varphi\big(\mathbf{s}(\mathbf{w})\big), \tag{28.1}$$

$\mathbf{s}$ is a smooth (vector-valued) function, and φ is a (possibly nonsmooth) merit function (that measures the goodness of $\mathbf{s}(\mathbf{w})$).

Algorithm 28.3: Gauss-Newton

Gauss-Newton is a popular iterative algorithm for solving the composite problem (28.1). In details, given $\mathbf{w}_t$, we linearize the inner function $\mathbf{s}$ and proceed to minimize the outer function φ:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big). \tag{28.2}$$

However, it may happen that $f(\mathbf{w}_{t+1}) > f(\mathbf{w}_t)$, since our linearization only holds locally around $\mathbf{w}_t$ while there is no guarantee that $\mathbf{w}_{t+1}$ will remain close to $\mathbf{w}_t$.

This algorithm is attributed to Gauss due to the least-squares Example 28.4, and it is also attributed to Newton because the mapping $\mathbf{s}$ is often the gradient of some function so $\mathbf{s}'$ is the Hessian (see Remark 28.12).

Example 28.4: Nonlinear least squares through a sequence of linear least squares

Often we need to find a solution to some nonlinear equation, i.e. $\mathbf{s}(\mathbf{w}) = \mathbf{0}$. Operationally, it is preferred to solve the nonlinear least-squares reformulation:

$$\min_{\mathbf{w}} \ \tfrac{1}{2}\|\mathbf{s}(\mathbf{w})\|_2^2, \quad \text{where} \quad \varphi = \tfrac{1}{2}\|\cdot\|_2^2. \tag{28.3}$$

While directly solving the above problem may be challenging, we can reduce it to a sequence of linear least squares problems:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \tfrac{1}{2}\|\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\|_2^2.$$

However, we note that the above approach typically worsens the condition number of our problem: Even when $\mathbf{s}$ is linear, the least-squares reformulation squares the condition number!

Taking square root we arrive at an equivalent reformulation:

$$\min_{\mathbf{w}} \ \|\mathbf{s}(\mathbf{w})\|_2, \quad \text{where} \quad \varphi = \|\cdot\|_2. \tag{28.4}$$

However, we remind that an ϵ-minimizer of the smooth problem (28.3) is merely a $\sqrt{\epsilon}$-minimizer of the nonsmooth problem (28.4)! There is simply no free squaring.

Algorithm 28.5: Prox-linear (Levenberg 1944; Marquardt 1963)

A natural idea, due to Levenberg (1944) and rediscovered by Marquardt (1963), is to add regularization as in proximal methods:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \underbrace{\varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w}-\mathbf{w}_t)\big)}_{\tilde f_t(\mathbf{w}) = \tilde f(\mathbf{w};\mathbf{w}_t)} + \tfrac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_t\|_2^2, \quad \textit{i.e.}, \quad \mathbf{w}_{t+1} = \mathrm{P}_{\tilde f_t}^{\eta_t}(\mathbf{w}_t) \tag{28.5}$$

(We could also turn the implicit regularization into an explicit constraint, resulting in the so-called trust region methods. In practice, regularization is often preferred since it does not introduce extra constraints.)

When the outer function φ is convex, the regularized problem (28.5) is strongly convex, while the original function $f = \varphi \circ \mathbf{s}$ may not even be convex. Moreover, it follows from Theorem 17.22 that the increment $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2$ is (continuous) increasing w.r.t. η_t while $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2/\eta_t$ is (continuous) decreasing w.r.t. η_t.

Levenberg, K. (1944). “A method for the solution of certain non-linear problems in least squares”. *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168.

Marquardt, D. W. (1963). “An Algorithm for Least-Squares Estimation of Nonlinear Parameters”. *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441.

Remark 28.6: Making sense of prox-linear

As pointed out by Levenberg (1944), for sufficiently small η_t, $\mathbf{w}_{t+1}$ will remain close to $\mathbf{w}_t$ so that decreasing the surrogate function $\tilde f$ leads to decrease in the original function f as well. Indeed,

$$\frac{\mathrm{d}f(\mathbf{w}_{t+1})}{\mathrm{d}\eta_t} = f'(\mathbf{w}_{t+1})\frac{\mathrm{d}\mathbf{w}_{t+1}}{\mathrm{d}\eta_t} = f'(\mathbf{w}_{t+1})[-(\mathrm{Id} + \eta_t \tilde f_t''(\mathbf{w}_{t+1}))^{-1}\tilde f_t'(\mathbf{w}_{t+1})], \tag{28.6}$$

where we differentiated the optimality condition of $\mathbf{w}_{t+1}$ w.r.t. η_t in the last step:

$$\eta_t \tilde f_t'(\mathbf{w}_{t+1}) + \mathbf{w}_{t+1} - \mathbf{w}_t = 0. \tag{28.7}$$

Noting that $\mathbf{w}_{t+1} \to \mathbf{w}_t$ if $\eta_t \downarrow 0$, under mild continuity assumptions (e.g. φ and $\mathbf{s}$ are sufficiently smooth or convex), we have

$$\frac{\mathrm{d}f(\mathbf{w}_{t+1})}{\mathrm{d}\eta_t}\Big|_{\eta_t=0} = -\|f'(\mathbf{w}_t)\|_2^2.$$

Thus, provided that η_t is sufficiently small and $\mathbf{w}_t$ is not already stationary, the prox-linear update (28.5) strictly decreases the original function f!

Let us also examine the fixed point of the prox-linear update (28.5): If $\mathbf{w}_{t+1} = \mathbf{w}_t = \mathbf{w}$, then clearly $\mathbf{w}$ is a stationary point of $\tilde f_t$. Equating the derivative of $\tilde f_t$ at $\mathbf{w}$ to zero, we obtain $\mathbf{0} \in \partial f(\mathbf{w})$, i.e. any fixed point $\mathbf{w}$, if it exists at all, is indeed a stationary point of the original function!

Levenberg, K. (1944). “A method for the solution of certain non-linear problems in least squares”. *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168.

Exercise 28.7: The power of composition

- Let $\tilde{\mathbf{s}}(\mathbf{w}) = (\mathbf{s}(\mathbf{w}), \mathbf{w})$ and $\tilde\varphi(\mathbf{z}, \mathbf{w}) = \varphi(\mathbf{z}) + r(\mathbf{w})$. Show that

  $$\tilde\varphi(\tilde{\mathbf{s}}(\mathbf{w})) = \varphi(\mathbf{s}(\mathbf{w})) + r(\mathbf{w}), \tag{28.8}$$

  and the Gauss-Newton update (28.2) for the left-hand side of (28.8) reduces to:

  $$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w}-\mathbf{w}_t)\big) + r(\mathbf{w}),$$

which was studied by Salzo and Villa (2012) for minimizing the right-hand side of (28.8) with $\varphi = \frac{1}{2}\|\cdot\|_2^2$, in which case the update for $\mathbf{w}_{t+1}$ reduces to a proximal step w.r.t. r (under the metric induced by $\mathbf{s}'(\mathbf{w}_t)^\top \mathbf{s}'(\mathbf{w}_t)$).

- Find $\mathbf{s}$ and φ so that the Gauss-Newton update (28.2) for $\varphi \circ \mathbf{s}$ reduces to the generalized conditional gradient update for $\ell + r$.
- Find $\mathbf{s}$ and φ so that the prox-linear update (28.5) for $\varphi \circ \mathbf{s}$ reduces to the gradient update for $\ell + r$.
- Find $\mathbf{s}$ and φ so that the prox-linear update (28.5) for $\varphi \circ \mathbf{s}$ reduces to the proximal gradient update for $\ell + r$, with a forward step for ℓ and a backward step for r.

The above special cases also indicate corresponding lower bounds on the convergence rates of Gauss-Newton and prox-linear for certain class of functions.

Salzo, S. and S. Villa (2012). "Convergence analysis of a proximal Gauss-Newton method". *Computational Optimization and Applications*, vol. 53, pp. 557–589.

Remark 28.8: How to tune the step size η_t

It is clear that there is some trade-off in choosing the step size η_t: the larger it is the more we move away from the current iterate and hence more likely to invalidate the linear approximation, while the smaller it is the less we move away from the current iterate and hence more iterations are likely needed.

The following rules are typical for setting the step size η_t:

- Cauchy's rule, where we find η_t to minimize $f(\mathbf{w}_{t+1})$, or we simply set the derivative in (28.6) to 0. Needless to say, the resulting problem is very complicated to solve.
- Levenberg's rule, where we linearize $f(\mathbf{w}_{t+1})$ (as a function of η) and recall that $\mathbf{w}_{t+1} = \mathbf{w}_t$ if $\eta_t = 0$:

$$f(\mathbf{w}_{t+1}) \approx f(\mathbf{w}_t) + \frac{\mathrm{d}f(\mathbf{w}_{t+1})}{\mathrm{d}\eta_t}|_{\eta_t=0}\eta_t = f(\mathbf{w}_t) - \eta_t\|f'(\mathbf{w}_t)\|_2^2.$$

  Equating the left-hand side to 0 we can solve

$$\eta_t = \frac{f(\mathbf{w}_t)}{\|f'(\mathbf{w}_t)\|_2^2},$$

  which is essentially a Newton step at $\eta = 0$ for the nonlinear equation $f(\mathbf{w}) = 0$.
- Amijo's backtracking, where we find the largest η_t so that $f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t)$. In case η_t grows exceedingly large, Marquardt (1963) also proposed a damping factor:

$$\tilde{\mathbf{w}}_{t+1} = \mathbf{w}_t - \lambda_t(\mathbf{w}_t - \mathbf{w}_{t+1}) = (1 - \lambda_t)\mathbf{w}_t + \lambda_t\mathbf{w}_{t+1},$$

  where we fix η_t (and hence $\mathbf{w}_{t+1}$) and backtrack λ_t so that $f(\tilde{\mathbf{w}}_{t+1}) \le f(\mathbf{w}_t)$.

Marquardt, D. W. (1963). "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441.

Alert 28.9: Affine equivariance

Let us consider a change-of-variable $\mathbf{w} = A\mathbf{z}$ for some invertible linear map A and see how the Gauss-Newton update changes:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big)$$

$$\mathbf{z}_{t+1} = \operatorname*{argmin}_{\mathbf{z}} \ \varphi\big(\mathbf{s}(A\mathbf{z}_t) + \mathbf{s}'(A\mathbf{z}_t)A(\mathbf{z} - \mathbf{z}_t)\big).$$

Clearly, $\mathbf{w}_{t+1} = A\mathbf{z}_{t+1}$ for all t if $\mathbf{w}_0 = A\mathbf{z}_0$, i.e. the Gauss-Newton update is affine equivariant.

The prox-linear update, on the other hand, is not affine equivariant. However, if we adapt the metric properly, we can restore affine equivariance:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big) + \tfrac{1}{2\eta_t}\|\mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\|_2^2$$
$$\mathbf{z}_{t+1} = \operatorname*{argmin}_{\mathbf{z}} \ \varphi\big(\mathbf{s}(A\mathbf{z}_t) + \mathbf{s}'(A\mathbf{z}_t)A(\mathbf{z} - \mathbf{z}_t)\big) + \tfrac{1}{2\eta_t}\|\mathbf{s}'(A\mathbf{z}_t)A(\mathbf{z} - \mathbf{z}_t)\|_2^2.$$

If we are only interested in equivariance w.r.t. a diagonal linear map A, then we can simply use $\operatorname{diag}(\mathbf{s}'(\mathbf{w}_t))$ to reweigh the norm, which was already mentioned by Levenberg (1944).

Levenberg, K. (1944). "A method for the solution of certain non-linear problems in least squares". *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168.

Remark 28.10: Connection to gradient descent

The prox-linear update (28.5) is closely related to the gradient update, as pointed out by Marquardt (1963). Indeed, let us examine the (normalized) correlation between the prox-linear increment and the gradient:

$$\frac{\langle \mathbf{w}_t - \mathbf{w}_{t+1}, f'(\mathbf{w}_t)\rangle}{\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2 \cdot \|f'(\mathbf{w}_t)\|_2}. \tag{28.9}$$

Applying the optimality condition (28.7) for $\mathbf{w}_{t+1}$ and dropping (nonnegative) factors that do not depend on η_t, we arrive at the relevant quantity

$$c(\eta_t) := \frac{\langle \mathbf{w}_t - \mathbf{w}_{t+1}, f'(\mathbf{w}_t)\rangle}{\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2} = \frac{\Big\langle \mathbf{w}_t - \mathbf{w}_{t+1}, \tilde f_t'(\mathbf{w}_t) - \tilde f_t'(\mathbf{w}_{t+1})\Big\rangle}{\eta_t\|\tilde f_t'(\mathbf{w}_{t+1})\|_2} + \|\tilde f_t'(\mathbf{w}_{t+1})\|_2,$$

which is nonnegative if φ and hence $\tilde f_t$ is convex. Taking derivative w.r.t. η_t:

$$\begin{aligned}\dot c(\eta_t) &\propto \left\langle -\dot{\mathbf{w}}_{t+1}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + (\mathbf{w}_t - \mathbf{w}_{t+1})\left\langle \dot{\mathbf{w}}_{t+1}, \mathbf{w}_t - \mathbf{w}_{t+1}\right\rangle, f'(\mathbf{w}_t)\right\rangle \\ &\propto \Big\langle \tilde f_t'(\mathbf{w}_{t+1}), \tilde f_t'(\mathbf{w}_{t+1})\Big\rangle \cdot \Big\langle A\tilde f_t'(\mathbf{w}_{t+1}), \tilde f_t'(\mathbf{w}_t)\Big\rangle - \Big\langle \tilde f_t'(\mathbf{w}_{t+1}), \tilde f_t'(\mathbf{w}_t)\Big\rangle \cdot \Big\langle A\tilde f_t'(\mathbf{w}_{t+1}), \tilde f_t'(\mathbf{w}_{t+1})\Big\rangle,\end{aligned}$$

where $A := \big(\mathrm{Id} + \eta_t \tilde f_t''(\mathbf{w}_{t+1})\big)^{-1}$. If we take $\varphi = \frac{1}{2}\|\cdot\|_2^2$, then we can verify that

$$\tilde f_t'(\mathbf{w}_{t+1}) = A\tilde f_t'(\mathbf{w}_t) =: \mathbf{z} \implies \dot c(\eta_t) \propto \|\mathbf{z}\|_2^4 - \left\langle A^{-1}\mathbf{z}, \mathbf{z}\right\rangle \langle A\mathbf{z}, \mathbf{z}\rangle \leq 0. \tag{28.10}$$

In other words, the (normalized) correlation (28.9) between the prox-linear increment and the gradient is nonnegative and decreasing w.r.t. η_t. In particular, when $\eta \downarrow 0$, $c(\eta) \to \|f'(\mathbf{w}_t)\|_2$ and the normalized prox-linear update reduces to the gradient update. Needless to say, when $\eta \uparrow \infty$, prox-linear reduces to Gauss-Newton. Thus, the step size η_t in prox-linear offers an interpolation between gradient descent and Gauss-Newton.

Marquardt, D. W. (1963). "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441.

Exercise 28.11: Cauchy-Schwarz

Prove the first equality and the last inequality in (28.10).

Remark 28.12: Connection to Newton's update

Consider the nonlinear equation $f'(\mathbf{w}) = 0$, and its least-squares reformulation:

$$\min_{\mathbf{w}} \ \tfrac{1}{2}\|f'(\mathbf{w})\|_2^2.$$

Applying the prox-linear Algorithm 28.5 we obtain:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}} \ \tfrac{1}{2}\|f'(\mathbf{w}_t) + f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\|_2^2 + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 \\ &= \mathbf{w}_t - \big(\tfrac{1}{\eta_t}\mathrm{Id} + [f''(\mathbf{w}_t)]^2\big)^{-1} f''(\mathbf{w}_t) f'(\mathbf{w}_t) \\ &= \mathbf{w}_t - \big(\tfrac{1}{\eta_t}[f''(\mathbf{w}_t)]^{-1} + f''(\mathbf{w}_t)\big)^{-1} f'(\mathbf{w}_t).\end{aligned}$$

If we (only) change the regularization to $\|\mathbf{w}\|_{f''_t}^2 := \langle f''(\mathbf{w}_t)\mathbf{w}, \mathbf{w}\rangle$, then we obtain the familiar update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \big(\tfrac{1}{\eta_t}\mathrm{Id} + f''(\mathbf{w}_t)\big)^{-1} f'(\mathbf{w}_t).$$

Both updates reduce to the Newton update when $\eta_t \to \infty$, while for a finite η_t they make Newton's update more stable (e.g. when the Hessian f'' is close to singular).

Theorem 28.13: Sublinear convergence rate of prox-linear (Nesterov 2007)

TBD ■

Nesterov, Y. (2007). "Modified Gauss-Newton scheme with worst case guarantees for global performance". *Optimization Methods and Software*, vol. 22, no. 3, pp. 469–483.

Theorem 28.14: Linear convergence rate of prox-linear (Drusvyatskiy and Lewis 2018)

TBD ■

See also Lewis and Wright (2016).

Drusvyatskiy, D. and A. S. Lewis (2018). "Error Bounds, Quadratic Growth, and Linear Convergence of Proximal Methods". *Mathematics of Operations Research*, vol. 43, no. 3, pp. 919–948.

Lewis, A. S. and S. J. Wright (2016). "A proximal method for composite minimization". *Mathematical Programming*, vol. 158, pp. 501–546.

Algorithm 28.15: Stochastic and variance reduced prox-linear

When the function $\varphi \circ \mathbf{s}$ can be written as an expectation:

$$\varphi \circ \mathbf{s}(\mathbf{w}) := \mathbb{E}\varphi \circ \mathbf{s}(\mathbf{w}, \boldsymbol{\xi}),$$

we may apply the same idea as in stochastic gradient (Drusvyatskiy and Paquette 2019). If the expectation is over a discrete distribution, then we may further reduce the variance (Zhang and Xiao 2021).

Drusvyatskiy, D. and C. Paquette (2019). "Efficiency of minimizing compositions of convex functions and smooth maps". *Mathematical Programming*, vol. 178, pp. 503–558.

Zhang, J. and L. Xiao (2021). "Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization". *Mathematical Programming*.

Algorithm 28.16: Quadratic Gauss-Newton (Bolte et al. 2020)

Instead of linearizing the inner function, we could also apply a quadratic approximation directly:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|^2\big),$$

where the norm can incorporate a linear map (say, to induce different scaling on different components).

Bolte, J., Z. Chen, and E. Pauwels (2020). "The multiproximal linearization method for convex composite problems". *Mathematical Programming*, vol. 182, pp. 1–36.