Optimization Methods II
Nonlinear LS Methods and
Stochastic Optimization Methods

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

March 26, 2024

Second-Order Method

Consider

$$\text{minimize } f(x)$$

with assumption that $f : \mathbf{R}^n \to \mathbf{R}$ is twice differentiable.

Seecond-order Taylor expansion:

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x. \quad (1)$$

Derivative of right hand side with respect to Δx equal to zero:

$$\nabla^2 f(x) \Delta x = -\nabla f(x). \quad (2)$$

Newton's Method

In the convex case Newton's method uses as search direction $\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x)$ which leads to the iteration

$$x_{k+1} = x_k - t_k \nabla^2 f(x_k)^{-1} \nabla f(x_k), \quad k = 0, 1, 2, \ldots \tag{3}$$

where x_0 is a initial guess, and $t_k > 0$ is the step size at iteration k.

If x_k is sufficiently close to stationary point of ∇f, then full Newton step can be shown to yield descent provided that $\nabla^2 f$ is Lipschitz continuous. However, this is not always the case if x_k is far away from stationary point.

Variable Metric Methods

Cosider more general convex quadratic approximation $\hat{f}(\cdot|\cdot) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ given by

$$\hat{f}(y \,|\, x_k) = f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2}(y - x_k)^T B_k (y - x_k) \tag{4}$$

where $B_k \succ 0$. We have

$$\hat{f}(x_k \,|\, x_k) = f(x_k), \qquad \nabla \hat{f}(x_k \,|\, x_k) = \nabla f(x_k).$$

Define search direction as

$$\Delta x_k = \operatorname*{argmin}_{\Delta x} \hat{f}(x_k + \Delta x \,|\, x_k) = -B_k^{-1} \nabla f(x_k),$$

and iterations as

$$x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k). \tag{5}$$

Letting $B_k = I$ yields the gradient descent method and $B_k = \nabla^2 f(x_k)$ corresponds to Newton's method.

Nonlinear LS Problem

A general nonlinear LS problem can be written as

$$\text{minimize } f(x)$$

where

$$f(x) = \frac{1}{2}\|F(x)\|_2^2 = \frac{1}{2}\sum_{i=1}^{N} f_i(x)^2$$

with $f_i : \mathbf{R}^n \to \mathbf{R}$, $i \in \mathbf{N}_N$ being the nonlinear residuals and

$$F(x) = (f_1(x), \ldots, f_N(x))$$

Gauss-Newton (GN) Method

Model the residuals with first order Taylor series expansions

$$\hat{f}_i\left(y \mid x_k\right) = f_i(x_k) + \nabla f_i(x_k)^T \left(y - x_k\right), \qquad i \in \mathbf{N}_N$$

With

$$\hat{f}\left(y \mid x_k\right) = \frac{1}{2} \sum_{i=1}^{N} \hat{f}_i\left(y \mid x_k\right)^2$$

we obtain a quadratic approximation as in (4) where

$$\nabla f(x) = \frac{\partial F(x)^T}{\partial x} F(x) = \sum_{i=1}^{N} \nabla f_i(x) f_i(x)$$

and where

$$B_k = \frac{\partial F\left(x_k\right)^T}{\partial x} \frac{\partial F\left(x_k\right)}{\partial x^T} = \sum_{i=1}^{N} \nabla f_i\left(x_k\right) \nabla f_i\left(x_k\right)^T$$

Hence this a variable metric method.

Expected Performance

The Hessian of f is

$$\nabla^2 f(x) = \sum_{i=1}^{N} f_i(x)^2 \nabla^2 f_i(x) + \nabla f_i(x_k) \nabla f_i(x_k)^T$$

If residuals $f_i(x)^2$ are small, then search directions obtained from the GN method are close to the ones obtained from Newton's method.

If the residuals are small or have small curvature we can expect super-linear convergence for the GN method. In general the convergence rate is linear.

When the residuals $f_i(x_k)$ are not small or if the matrix B_k is poorly conditioned in the GN method, the method may not behave that well.

Levenberg-Marquardt (LM) Method

Add *damping* to the quadratic model

$$\hat{f}\left(y \mid x_k\right) = \frac{1}{2}\sum_{i=1}^{N} \hat{f}_i\left(y \mid x_k\right)^2 + \mu_k \left\|y - x_k\right\|_2^2$$

where $\mu_k \in \mathbf{R}_{++}$ is damping parameter. Then

$$B_k = \sum_{i=1}^{N} \nabla f_i\left(x_k\right) \nabla f_i\left(x_k\right)^T + \mu_k I$$

If μ^k very large the method will essentially behave as gradient method with step size roughly equal to $1/\mu^k$. If μ^k small we will recover the GN method.

Updating rule:

1. if $\left|f\left(x_{k+1}\right)\right|^2 < \left|f\left(x_k\right)\right|^2$, accept iterate and set $\mu_{k+1} = 0.8\mu_k$
2. otherwise increase μ and do no update, i.e. $\mu_{k+1} = 2\mu_k$ and $x_{k+1} = x_k$

Separable Nonlinear LS Problem

$$\text{minimize} \frac{1}{2}\|F(x,\alpha)\|_2^2 \tag{6}$$

with variables (x,α), where $F : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{R}^m$ is

$$F(x,\alpha) = A(\alpha)x - b(\alpha)$$

with $A : \mathbf{R}^p \to \mathbf{R}^{m\times n}$ and $b : \mathbf{R}^p \to \mathbf{R}^m$.

For fixed α linear LS problem.

Minimize with respect to x to obtain $x(\alpha) = \left(A^T A\right)^{-1} A^T b$ from the normal equations.

Reduced Nonlinear LS Problem

Back-substitution results in nonlinear LS problem

$$\text{minimize} \frac{1}{2}\|g(\alpha)\|_2^2$$

with variable α, where $g : \mathbf{R}^p \to \mathbf{R}^m$ defined as

$$g(\alpha) = F(x(\alpha), \alpha) = -P(\alpha)b(\alpha)$$

where $P : \mathbf{R}^p \to \mathbf{R}^{m \times m}$ is a projection matrix defined as $P = I - A\left(A^T A\right)^{-1} A^T$.

The method we will derive is often called the *variable projection method*, since it is based on minimizing the variable projection functional above.

Jacobian

The Jacobian of $g(\alpha)$:

$$\frac{\partial g}{\partial \alpha^T} = -\frac{\partial P}{\partial \alpha} b - P \frac{\partial b}{\partial \alpha^T}$$

where $\frac{\partial P}{\partial \alpha} b$ is defined to be the matrix with columns $\frac{\partial P}{\partial \alpha_k} b$, and where

$$\begin{aligned}
\frac{\partial P}{\partial \alpha_k} &= -\frac{\partial A}{\partial \alpha_k} \left(A^T A\right)^{-1} A^T \\
&\quad + A \left(A^T A\right)^{-1} \left(\frac{\partial A^T}{\partial \alpha_k} A + A^T \frac{\partial A}{\partial \alpha_k}\right) \left(A^T A\right)^{-1} A^T \\
&\quad - A \left(A^T A\right)^{-1} \frac{\partial A^T}{\partial \alpha_k} \\
&= -P \frac{\partial A}{\partial \alpha_k} \left(A^T A\right)^{-1} A^T - A \left(A^T A\right)^{-1} \frac{\partial A^T}{\partial \alpha_k} P
\end{aligned}$$

Approximation of Jacobian

Exact Jacobian given by

$$P\frac{\partial A}{\partial \alpha}\left(A^T A\right)^{-1} A^T b + A\left(A^T A\right)^{-1}\frac{\partial A^T}{\partial \alpha}Pb - P\frac{\partial b}{\partial \alpha^T}$$

Second term contains the factor $Pb = -g(\alpha)$, which is small for α close to optimum, and Kaufman suggested to drop it resulting in the Jacobian apporoximation

$$P\frac{\partial A}{\partial \alpha}x(\alpha) - P\frac{\partial b}{\partial \alpha^T}$$

Save about 25% in computational time and also provides better convergence properties

The Kaufman approximation can be used in any of the methods for nonlinear LS problems discussed above.

Coordinate Descent Methods

Consider

$$\text{minimize } f(x)$$

and coordinate descent iterations of the form

$$x_{k+1} = x_k - t_k[\nabla f(x_k)]_{i_k} e_{i_k}, \tag{7}$$

where $i_k \in \mathbf{N}_n$ is a coordinate index and $t_k > 0$ is a step size. Common index selection strategies include the cyclic order $i_k = (k \bmod n) + 1$, a randomized cyclic order where the order is reshuffled every n iterations, and a fully randomized order where the indices are selected uniformly at random.

The step size can be chosen in a similar way to the gradient method, *e.g.*, using some form of line search.

Iteration (7) can be shown to converge to a minimizer if f is convex and continuously differentiable with bounded sublevel sets.

Stochastic Optimization Methods

Consider the the *single stage stochastic optimization* problem

$$\text{minimize} \quad \mathbb{E}\, F(x, \xi) \tag{8}$$

where $F : \mathbf{R}^n \times \mathcal{D} \to \mathbf{R}$.

In statistical learning, ξ typically represents the problem data, *i.e.*, each realization of ξ corresponds to a random observation.

The problem is reduced to a deterministic problem if we let $f(x) = \mathbb{E}\, F(x, \xi)$.

However, the probability distribution associated with the random variable ξ is required to compute the expectation, and this distribution is typically not available in practice.

Sample Average Approximation (SAA)

Replace the expectation by

$$\mathbb{E}\, F(x, \xi) \approx \frac{1}{m} \sum_{i=1}^{m} F(x, \xi_i)$$

where $\xi_1, \ldots, \xi_m$ are m independent samples of the random variable ξ.

Resulting problem is deterministic:

$$\text{minimize} \quad \frac{1}{m} \sum_{i=1}^{m} f_i(x) \tag{9}$$

where $f_i(x) = F(x, \xi_i)$.

An example is *empirical risk minimization* where

$$f_i(x) = l(a_i^T x, b_i)$$

where $l : \mathbf{R}^n \to \mathbf{R}$ is a loss function and $(a_i, b_i) \in \mathbf{R}^n \times \mathbf{R}$ is one of m observations.

Stochastic Approximation Method

Instead we could try to find a solution to stationarity condition $\nabla f(x) = 0$, where $f(x) = \mathbb{E}[\, F(x, \xi)\,]$.

Stochastic Approximation (SA) method:

$$x_{k+1} = x_k - t_k\, g_k, \quad k = 0, 1, \ldots, \tag{10}$$

where x_0 is an initial guess, $t_k > 0$ is the step size at iteration k, and g_k is a realization of an estimator of G_k of $\nabla f(x_k)$.

Notice that G_k is *random variable* for each k, and hence (10) is a realization of a stochastic process

$$X_{k+1} = X_k - t_k\, G_k, \quad k = 0, 1, \ldots \tag{11}$$

where the random variable X_0 is the initial state.

Unbiased Estimate of Gradient

We consider unbiased estimator G_k of $\nabla f(x_k)$ with bounded variance, *i.e.*,

$$\mathbb{E}[\, G_k \mid X_k = x_k \,] = \nabla f(x_k) \tag{12}$$

$$\mathbb{E}\big[\, \|G_k - \nabla f(x_k)\|_2^2 \mid X_k = x_k \,\big] \leq c^2, \tag{13}$$

for all k and for some scalar $c \geq 0$.

$\nabla F(x, \xi)$ is an unbiased estimator of $\nabla f(x)$ if

$$\frac{\partial\, \mathbb{E}[\, F(x,\xi)\,]}{\partial x_i} = \mathbb{E}\left[\, \frac{\partial F(x,\xi)}{\partial x_i} \,\right], \quad i = 1, \ldots, n.$$

Then natural to choose $G_k = \nabla F(X_k, \xi_k)$, where ξ_k has same distribution as ξ for all $k \geq 0$, and where ξ_k and ξ_j, $j \neq k$ independent.

Convergence Analysis

Assume that f is bounded from below and β-smooth. Then $\exists\, \beta > 0$ such that

$$f(x + \Delta x) \leq f(x) + \nabla f(x)^T \Delta x + \frac{\beta}{2}\|\Delta x\|_2^2,$$

for all $x, \Delta x \in \mathbf{R}^n$. Hence

$$\mathbb{E}[\, f(X_k) \mid X_{k-1}\,] \leq$$

$$\mathbb{E}\left[f(X_{k-1}) - t_{k-1}\nabla f(X_{k-1})^T G_{k-1} + \frac{\beta t_{k-1}^2}{2}\| G_{k-1}\|_2^2 \mid X_{k-1} \right]. \tag{14}$$

It follows from (12) that

$$\mathbb{E}\Big[\, \nabla f(X_{k-1})^T G_{k-1} \mid X_{k-1} \Big] = \|\nabla f(X_{k-1})\|_2^2$$

and

$$\begin{aligned}\mathbb{E}\big[\, \|G_{k-1}\|_2^2 \mid X_{k-1}\,\big] &= \mathbb{E}\big[\, \|G_{k-1} - \nabla f(X_{k-1}) + \nabla f(X_{k-1})\|_2^2 \mid X_{k-1}\,\big] \\ &\leq c^2 + \|\nabla f(X_{k-1})\|_2^2.\end{aligned}$$

Convergence Analysis ctd.

Combining these results with (14), we arrive at the the upper bound

$$\begin{aligned}\mathbb{E}[\, f(X_k) \mid X_{k-1}\,] &\leq f(X_{k-1}) - t_{k-1}\|\nabla f(X_{k-1})\|_2^2 \\ &\quad + \frac{\beta t_{k-1}^2}{2}\left(c^2 + \|\nabla f(X_{k-1})\|_2^2\right) \\ &= f(X_{k-1}) - t_{k-1}(1 - \beta t_{k-1}/2)\|\nabla f(X_{k-1})\|_2^2 \\ &\quad + \frac{\beta t_{k-1}^2 c^2}{2}.\end{aligned}$$

Rearranging the terms leads to the inequality

$$\begin{aligned}t_{k-1}(1 - \beta t_{k-1}/2)\|\nabla f(X_{k-1})\|_2^2 &\leq f(X_{k-1}) - \mathbb{E}[\, f(X_k) \mid X_{k-1}\,] \\ &\quad + t_{k-1}^2 \frac{\beta c^2}{2},\end{aligned}$$

Convergence Analysis ctd.

Summing and taking expectation yields

$$\sum_{j=0}^{k-1} t_j(1-\beta t_j/2)\,\mathbb{E}\left[\,\|\nabla f(X_j)\|_2^2\,\right] \leq$$

$$\sum_{j=0}^{k-1}\left[\mathbb{E}[\,f(X_j)\,]-\mathbb{E}[\,f(X_{j+1})\,]\right]+\sum_{j=0}^{k-1} t_j^2\frac{\beta c^2}{2}. \quad (15)$$

The first sum on the right-hand side satisfies

$$\sum_{j=0}^{k-1}\left[\mathbb{E}[\,f(X_j)\,]-\mathbb{E}[\,f(X_{j+1})\,]\right]=\mathbb{E}[\,f(X_0)\,]-\mathbb{E}[\,f(X_k)\,]\leq \mathbb{E}[\,f(X_0)\,]-p^\star,$$

where $p^\star=\inf_x f(x)$, and combining this inequality with (15):

$$\sum_{j=0}^{k-1} t_j(1-\beta t_j/2)\min_{j=0,\ldots,k-1}\mathbb{E}\left[\,\|\nabla f(X_j)\|_2^2\,\right]\leq \mathbb{E}[\,f(X_0)\,]-p^\star$$

$$+\sum_{j=0}^{k-1} t_j^2\frac{\beta c^2}{2}.$$

Convergence Analysis ctd.

Equivalently, divide by $\sum_{j=0}^{k-1} t_j(1-\beta t_j/2)$ on both sides (assuming positive):

$$\min_{j=0,\ldots,k-1} \mathbb{E}\left[\, \|\nabla f(X_j)\|_2^2 \,\right] \leq$$

$$\frac{\mathbb{E}[\, f(X_0)\,] - p^\star}{\sum_{j=0}^{k-1} t_j(1-\beta t_j/2)} + \frac{\sum_{j=0}^{k-1} t_j^2}{\sum_{j=0}^{k-1} t_j(1-\beta t_j/2)} \frac{\beta c^2}{2}. \quad (16)$$

Sufficient condition for the right-hand side to vanish as $k \to \infty$ is

$$\lim_{k\to\infty} \sum_{j=0}^{k-1} t_j(1-\beta t_j/2) = \infty, \quad \lim_{k\to\infty} \frac{\sum_{j=0}^{k-1} t_j(1-\beta t_j/2)}{\sum_{j=0}^{k-1} t_j^2} = \infty,$$

or equivalently,

$$\lim_{k\to\infty} \sum_{j=0}^{k-1} t_j = \infty, \quad \lim_{k\to\infty} \frac{\sum_{j=0}^{k-1} t_j}{\sum_{j=0}^{k-1} t_j^2} = \infty. \quad (17)$$

Examples of Step Sizes

Examples of step size sequences that satisfy conditions:

$$t_k = \frac{t}{(k+1+\zeta)^\delta}, \qquad k = 0, 1, 2, \ldots, \tag{18}$$

where $t > 0$, $\zeta \geq 0$, and $\delta \in (0, 1]$ are fixed parameters. The parameter t scales the sequence, δ controls the asymptotic rate of decay, and ζ may be used to reduce the rate of decay in early iterations.

The value of t has no effect on the asymptotic behavior, but it typically has a strong effect on the non-asymptotic behavior. Step size sequences of the form (18) satisfy

$$\frac{1}{\sum_{j=0}^{k-1} t_j} \propto \frac{1}{t}, \qquad \frac{\sum_{j=0}^{k-1} t_j^2}{\sum_{j=0}^{k-1} t_j} \propto t.$$

Comparing with the right-hand side of (16), the choice of t presents a trade-off between the two terms: increasing t reduces the first term but increases the second and vise versa.

Analysis of Worst Case Bound

If $c = 0$ corresponding to ordinary gradient method, then sufficient that $t_k = t \in (0, 2/\beta)$ such that $\sum_{j=0}^{k-1} t_j(1 - \beta t_j/2) = O(k)$. Hence right-hand side of (16) decays as $O(1/k)$.

However, in the stochastic setting where $c > 0$, a constants step size sequence does not make the right-hand side of (16) vanish as $k \to \infty$.

Analysis of Worst Case Bound ctd.

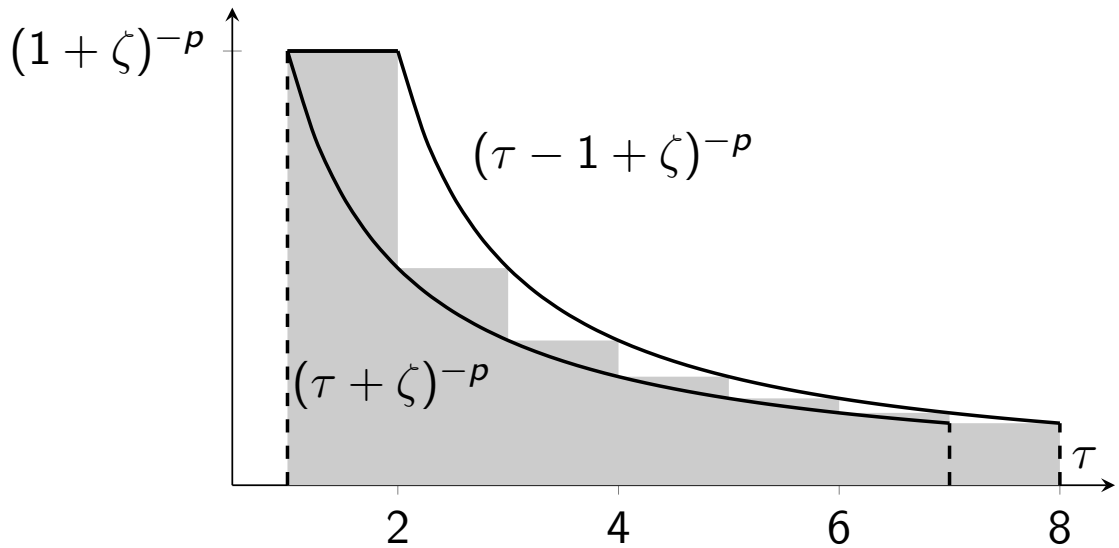
Consider (18) for different values of δ. We have

$$\sum_{j=0}^{k-1} t_j = t \sum_{j=1}^{k} (j+\zeta)^{-\delta}, \qquad \sum_{j=0}^{k-1} t_j^2 = t^2 \sum_{j=1}^{k} (j+\zeta)^{-2\delta},$$

both of which involve sums of the form

$$s_k = \sum_{j=1}^{k} \frac{1}{(j+\zeta)^p}, \quad p \in \mathbf{R}_+.$$

We bound s_k from above and below with $\int_1^k (\tau+\zeta)^{-p}\, d\tau$.



Analysis of Worst Case Bound ctd.

This leads to inequalies

$$\int_1^k (\tau+\zeta)^{-p}\,d\tau \le s_k \le (1+\zeta)^{-p} + \int_1^k (\tau+\zeta)^{-p}\,d\tau,$$

and using

$$\lim_{k\to\infty} \int_1^k (\tau+\zeta)^{-p}\,d\tau = \begin{cases} \frac{(1+\zeta)^{1-p}}{p-1}, & p>1, \\ \infty, & p\in[0,1], \end{cases}$$

we conclude that s_k converges when $p>1$ and diverges otherwise. In latter case, the upper and lower bounds on s_k gives

$$s_k \sim \begin{cases} k^{1-p}, & p\in[0,1), \\ \ln(k), & p=1. \end{cases}$$

This result may be used to derive upper bounds on the right-hand side of (16) for different values of δ, which are summarized in

Asymptotic Behaviour as Function of δ.

Parameter	$\sum_{j=0}^{k-1} t_j$	$\sum_{j=0}^{k-1} t_j^2$	Upper bound
$\delta > 1$	$\Theta(1)$	$\Theta(1)$	$O(1)$
$\delta = 1$	$\sim \ln(k))$	$\Theta(1)$	$O(1/\ln(k))$
$1/2 < \delta < 1$	$\sim k^{1-\delta}$	$\Theta(1)$	$O(1/k^{1-\delta})$
$\delta = 1/2$	$\sim \sqrt{k}$	$\sim \ln(k)$	$O(\ln(k)/\sqrt{k})$
$0 < \delta < 1/2$	$\sim k^{1-\delta}$	$\sim k^{1-2\delta}$	$O(1/k^{\delta})$
$\delta = 0$	$\sim k$	$\sim k$	$O(1)$

Note that asymptotically, the upper bound decays the fastest when $\delta = 1/2$. However, this choice is not necessarily the best one in practice.

Comments

The iteration (10) is often referred to as a *stochastic gradient* (SG) method or a *stochastic gradient descent* (SGD) method.

However, it is important to note that it is not a descent method in the deterministic sense, *i.e.*, the search direction $-g_k$ is not necessarily a descent direction, so the objective value may increase in some iteration.

We note that in some communities, the step size t_k is referred to as the *learning rate*.

Incremental Methods

The SG method is closely related to the class of *incremental methods*, and the two terms are often used synonymously.

Incremental methods solves (9) where objective function is a finite sum of components $f_i(x)$, and the iterative update is of the form

$$x_{k+1} = x_k - t_k \nabla f_{i_k}(x_k) \tag{19}$$

where $i_k \in \mathbf{N}_m$ is chosen according to some index selection rule.

Most common rules are the *cyclic* rule $i_k = (k \mod m) + 1$ and the *random* rule where each i_k is chosen uniformly at random from $\mathbf{N}_m$.

With the random index selection rule, the incremental gradient method may be viewed as a SG method applied to (8) if the random variable ξ is discrete with m equiprobable outcomes.

Batch Gradient Method

Consider again

$$\text{minimize} \quad \frac{1}{m}\sum_{i=1}^{m} f_i(x) \tag{20}$$

which has gradient

$$\nabla f(x_k) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_k).$$

If full gradient is used the gradient method is often referred to as *batch* gradient descent.

The *incremental* gradient method only uses

$$\nabla f_{i_k}(x_k)$$

in each iteration

Can be viewed as noisy approximation of full gradient and is not necessarily a descent direction.

Mini-Batch Gradient Method

Compromise:

$$x_{k+1} = x_k - \frac{1}{p} \sum_{i \in \mathcal{S}_k} \nabla f_i(x_k)$$

where $\mathcal{S}_k \subset \mathbf{N}_m$ and $|\mathcal{S}_k| = p$.

Adaptive Step Size

Diminishing step size sequence required for convergence.

Adaptive step size rules needed for practical convergence.

Prototype method:

$$x_{k+1} = x_k - B_k^{-1} g_k$$

where $B_k \in \mathbf{S}^n_{++}$ and g_k is stochastic gradient.

Interpretations: incremental/stochastic variant of variable metric method, or adaptively preconditioned stochastic gradient method.

Adaptive Gradient Method (AdaGrad)

Consider again

$$\text{minimize} \quad \mathbb{E}[F(x,\xi)] \tag{21}$$

with variable $x \in \mathbf{R}^n$ and where ξ is a random variable.

Assume $F(x,\xi) = G(x,\xi) + h(x)$ with $G(\cdot,\xi)$ and h closed and convex.

Let $g_k \in \partial G(x_k,\xi_k)$ where ξ_k a realization of ξ at iteration k and

$$\widehat{G}_k = \sum_{i=0}^{k} g_k (g_k)^T,$$

AdaGrad uses either $B_k = \gamma(\mathbf{diag}(\widehat{G}_k) + \varepsilon I)^{1/2}$ or $B_k = \gamma(\widehat{G}_k + \varepsilon I)^{1/2}$ where $\gamma > 0$ and $\varepsilon > 0$ are parameters.

Diagonal AdaGrad

Diagonal variant of AdaGrad:

$$
\begin{aligned}
v_k &= v_{k-1} + g_k \circ g_k \\
B_k &= \gamma \, \mathbf{diag}(v_k + \varepsilon \, I)^{1/2} \\
x_{k+1} &= \mathbf{prox}_h^{B_k}(x_k - B_k^{-1} g_k)
\end{aligned}
$$

where $v_{-1} = 0$ and x_0 are initial values.

Implementational Details

The matrix B_k can be expressed as

$$B_k = \sqrt{k+1}\left(\frac{1}{k+1}\,\mathbf{diag}\,\widehat{G}_k + \frac{\varepsilon}{k+1} I\right)^{1/2}$$

which may be implemented recursively as

$$\tilde{v}_k = \frac{k}{k+1}\tilde{v}_{k-1} + \frac{1}{k+1}(g_k \circ g_k)$$
$$B_k = \gamma\sqrt{k+1}\;\mathbf{diag}(\tilde{v}_k + \varepsilon/(k+1)\,\mathbf{1})^{1/2}.$$

Summary of AdaGrad

Then

$$\begin{aligned}
\tilde{v}_k &= \frac{k}{k+1}\tilde{v}_{k-1} + \frac{1}{k+1}(g_k \circ g_k) \\
B_k &= \gamma\sqrt{k+1}\,\mathbf{diag}(\tilde{v}_k + \varepsilon/(k+1)\,\mathbf{1})^{1/2} \\
\tilde{g}_k &= \mathbf{diag}(\tilde{v}_k + \varepsilon/(k+1)\,\mathbf{1})^{-1/2} g_k \\
x_{k+1} &= \mathbf{prox}_h^{B_k}(x_k - t_k \tilde{g}_k)
\end{aligned}$$

with $t_k = \frac{1}{\gamma\sqrt{k+1}}$, which shows that AdaGrad implicitly employs a diminishing step size sequence.

In the convex setting, AdaGrad can be shown to satisfy the worst-case bound

$$\mathbb{E}[f(x_k) - p^\star] \leq O(1/\sqrt{k}).$$

RMSprop

Consider again

$$\text{minimize} \quad \mathbb{E}[F(x,\xi)] \tag{22}$$

with variable $x \in \mathbf{R}^n$ and where ξ is a random variable. Assume $F(x,\xi)$ differentiable with gradient $g_k = \nabla F(x_k,\xi_k)$.

Stochastic gradient iteration using moving average

$$\begin{aligned} v_k &= \beta v_{k-1} + (1-\beta)(g_k \circ g_k) \\ B_k &= \gamma \, \mathbf{diag}(v_k + \varepsilon \mathbf{1})^{1/2} \\ x_{k+1} &= x_k - B_k^{-1} g_k \end{aligned}$$

where $\beta \in (0,1)$.

RMSprop does not implictly result in a diminishing step size sequence, and in fact, the method need not converge.

AdaDelta

Same optimization problem as before.

AdaDelta (similar to RMSprop but more averaging):

$$
\begin{aligned}
v_k &= \beta v_{k-1} + (1-\beta)(g_k \circ g_k) \\
B_k &= \mathbf{diag}(v_k + \varepsilon \mathbf{1})^{1/2}\, \mathbf{diag}(p_{k-1} + \varepsilon \mathbf{1})^{-1/2} \\
p_k &= \beta p_{k-1} + (1-\beta)\left(B_k^{-1} g_k \circ B_k^{-1} g_k\right) \\
x_{k+1} &= x_k - B_k^{-1} g_k.
\end{aligned}
$$

We note that AdaDelta need not converge.

Adaptive Moment Estimation (Adam)

Combines the adaptive scaling approach of RMSprop with *gradient aggregation* or *momentum*.

$$\begin{aligned}
\mu_k &= \beta_1 \mu_{k-1} + (1-\beta_1) g_k \\
v_k &= \beta_2 v_{k-1} + (1-\beta_2)(g_k \circ g_k) \\
H_k &= \gamma^{-1} \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k} \mathbf{diag}(\mu_k) \left(\mathbf{diag}(v_k)^{1/2} + \varepsilon I \right)^{-1} \\
x_{k+1} &= x_k - H_k \mu_k .
\end{aligned}$$

The method need not converge.

Variant called *AdaMax* obtained by replacing v_k and H_k by

$$u_k = \max(\beta_2 u_{k-1}, |g_k|), \qquad H_k = \gamma^{-1} \frac{1}{1-\beta_1^k} \mathbf{diag}(u_k)^{-1}$$

Nadam combines Adam with Nesterov-like acceleration.

AMSGrad

Addresses Adam's convergence issue by avoiding the possibility of increasing step sizes:

$$\begin{aligned}
\mu_k &= \beta_{1,k}\mu_{k-1} + (1-\beta_{1,k})g_k \\
v_k &= \beta_2 v_{k-1} + (1-\beta_2)(g_k \circ g_k) \\
\hat{v}_k &= \max(\hat{v}_{k-1}, v_k) \\
B_k &= \gamma\sqrt{k+1}\,\mathbf{diag}(\hat{v}_k)^{1/2} \\
x_{k+1} &= \mathbf{prox}_h^{B_k}(x_k - B_k^{-1}\mu_k)
\end{aligned}$$

where $\hat{v}_k = \max(\hat{v}_{k-1}, v_k)$ denotes the elementwise maximum of $\hat{v}_{k-1}$ and v_k.

WNGrad

Adaptive method based on technique called *weight normalization*.

Method can be used with both gradient method and stochastic gradient method.

$$x_{k+1} = x_k - t_k g_k$$
$$t_{k+1} = \frac{t_k}{1 + \|t_k g_k\|_2^2}$$

where $t_0 > 0$, and g_k is either stochastic gradient or gradient of f.

In batch setting (*i.e.*, $g_k = \nabla f(x_k)$), the method converges globally to stationary point if f is β-smooth, and in stochastic setting (*i.e.*, $g_k = \nabla F(x_k, \xi_k)$), convergence for convex function.