Notes on Conjugate-Gradient
Ling Li, Learning Systems Group, Caltech

Minimize the error function $E(w)$ with CG

- start: $w_0, d_0 = -g_0 = -\frac{\partial E}{\partial w_0}$;

- at time $t = 1, 2, \ldots$

$$
\begin{align*}
  w_t &= w_{t-1} + \alpha_{t-1}d_{t-1} \\
  g_t &= \frac{\partial E}{\partial w_t} \\
  d_t &= -g_t + \beta_t d_{t-1}
\end{align*}
$$

where $\alpha_t$ is determined by a line search, and

$$
\beta_t = \frac{g_t^T (g_t - g_{t-1})}{g_{t-1}^T g_{t-1}}.
$$

- $\beta_t$ is usually reset to 0 every several iterations.

- for neural nets, we do this in batch mode.

Questions

- Why do we do this?

- Why and when to reset $\beta$?

- In my experience, $\beta_t$ is usually around 1, and $g_t^T g_{t-1}$ is usually very small. Why?
Gradient descent

When we do the gradient descent,

\[ w_{t+1} = w_t - \eta \frac{\partial E}{\partial w_t}, \]

we approximate \( E(w) \) by a linear function

\[ E(w_t + \delta w) \approx E(w_t) + \frac{\partial E}{\partial w_t} \delta w, \]

with \( \|\delta w\| \) small enough.

If we can use the second-order approximation, the error should be decreased faster.

Quadratic problem

Assume \( A \) (the Hessian) is positive definite, and we want to

\[ \min_w E(w) = \frac{1}{2} w^T A w - b^T w. \]

If we have a conjugate base \( \{d_i\} \), i.e.,

\[ d_i^T A d_j = 0 \quad \text{if} \quad i \neq j, \]

then for \( w = \sum \alpha_i d_i \),

\[ E(w) = E \left( \sum \alpha_i d_i \right) = \sum \left( \frac{1}{2} \alpha_i^2 d_i^T A d_i - \alpha_i b^T d_i \right). \]

Minimizing \( E(w) \) is reduced to several line searches along those conjugate directions.

This is the base of the conjugate-gradient method.
Some properties

For quadratic problems, the conjugate-gradient uses

- start: \( w_0 = 0, \quad d_0 = -g_0 = -\frac{\partial E}{\partial w_0} = b; \)

- \( \alpha_t = \frac{g_t^T g_t}{d_t^T A d_t} \) and \( \beta_t = \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}} \) (Fletcher-Reeves).

We can prove (by induction) that

- \( g_i^T g_j = 0, \quad g_i^T d_j = 0, \quad \text{and} \quad d_i^T A d_j = 0, \) for \( i > j \). \((*)\)

- the conjugate-gradient method converges in \( n \) iterations, where \( n \) is the number of distinct eigenvalues of \( A \).

Nonlinear conjugate-gradient methods

For \( E(w) \) not perfectly quadratic, properties \((*)\) may not still hold due to many reasons (Hessian \( A \) may change, line search may not be perfect, \ldots).

A bad situation is where \( \|g_t\| \) or \( |g_t^T d_t| \) is very small, and the line search also gives a very small step \( \alpha_t \). Since now

\[
    w_{t+1} \approx w_t, \quad g_{t+1} \approx g_t, \quad \beta_{t+1} \approx 1, \quad d_{t+1} \approx d_t,
\]

the bad situation still holds at step \( t + 1 \).

Thus we need a restart — reset \( \beta_t \) to 0.

By analogy with \( g_t^T g_{t-1} = 0 \) in \((*)\), restart if

\[
    |g_t^T g_{t-1}| \geq \tau \|g_t\|^2, \quad \tau \text{ can be 0.1 }
\]
Variant (Polak-Ribiére)

Different formulas for $\beta_t$ give different conjugate-gradient methods.

$$\beta_t^{PR} = \frac{g_t^T (g_t - g_{t-1})}{g_{t-1}^T g_{t-1}}$$

is identical to $\beta_t^{FR} = \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}}$ when $E(w)$ is strongly convex quadratic and the line search is exact. When a bad situation such as $g_t \approx g_{t-1}$ happens, $\beta_t \approx 0$, corresponding to an automatic restart.

Even though examples have been designed which cause this method to display especially poor performance, it is generally better than the Fletcher-Reeves version in practice.

Nocedal and Wright suggested

$$\beta_t^{PR^+} = \max \left\{ \beta_k^{PR}, 0 \right\}$$

which guarantees descent and can be proved to be globally convergent.