Minimizing the Mean of a Random Variable with One Real Parameter

Eric Boesch

May 26, 2010

1 Naive minimization

First consider upper bounds on the typical performance of algorithms whose only assumption about the relationship between the mean of a random variable being sampled and its parameter x is that the mean as a function of x has only one local minimum within the range of interest. (Some assumption of that kind is necessary to permit drawing conclusions about the mean at x values that have not yet been sampled.) I will call such algorithms *naive* to indicate that they do not make assumptions about the existence or nature of derivatives of the mean with respect to x.

Naive algorithms can only obtain bounds on the range of x values that may have the minimum expected value by identifying three subsets covering disjoint x ranges such that the sample mean of the middle set is shown, to within desired confidence bounds, to be less than the minimum of the mean of the left set and the mean of the right set. In that case, it can be concluded that the x value at which the mean is minimal is probably greater than the least x value in the left set and less than the greatest x value in the right set.

(Additional assumptions are needed in order to set confidence bounds, because sampling cannot provide confidence bounds on the mean of an unknown distribution. There is always some chance that, as in a lottery, the mean is heavily influenced by large outliers that are very unlikely to be observed. However, the question of which assumptions the confidence test makes when comparing the means of two sample sets is independent of which assumptions the algorithm makes about the nature of the mean of the random variable as a function of its parameter. For simplicity's sake, I will assume for now that the random variable is normal at any fixed x value.)

The best deterministic algorithm cannot outperform the best nondeterministic one. Under the assumption of normality (and even under most other

reasonable assumptions I can think of), the best nondeterministic naive algorithm is one that invokes an oracle to tell which three points should be sampled in order to obtain the best bounds within a given number of samples. Sampling more than three points only decreases the power of the test.

1.1 Primary defect of naive minimization: convergence typically no better than $O(n^{-1/4})$

Suppose $R(x)$ is our random variable, so we wish to find x_0 such that the expected value $y_0 = E(R(x_0))$ is minimized. As the number of samples n increases, the accuracy of the x_0 estimate obtainable with naive minimization typically grows quite slowly.

If $\mathrm{var}(R(x))$ has positive lower bound in an interval about x_0, then the sample error of estimates of y_0 varies with $O(\frac{1}{\sqrt{n}})$, where n equals the number of samples taken. That is the typical best possible performance for sampling parameterless random variables, so typical best possible performance for parameterized random variables cannot be better than that.

To detect a significant difference between the means of two sample sets Y_1 and Y_2, one needs

$$O\left(\left(\frac{\mathrm{sdev}(Y_1) + \mathrm{sdev}(Y_2)}{|\bar{Y_1} - \bar{Y_2}|}\right)^2\right)$$

samples, where sdev is the sample standard deviation. If $\frac{\partial^2}{\partial x^2}E(R(x))$ is defined and positive at x_0 (that is, $E(R(x))$ has roughly parabolic shape near the minimum) then the minimum distance from x_0 necessary to detect a significant difference in y is proportional to the square root of the difference in y, yielding a convergence rate for $x_{0_{est}}$ towards the true minimum x_0 that is also the square root of the convergence rate for y. That is,

$$|x_{0_{est}} - x_0| \in O(n^{-\frac{1}{4}})$$

It may be possible for practical deterministic naive algorithms to exactly match this best big-O performance for x_0 and y_0, and at worst they can come very close (exceeding it by a ratio that is $O(\ln n)$).

The x_0 convergence for naive algorithms is not very good compared to what is possible with more aggressive interpolation ($O(n^{-\frac{1}{3}})$ for nondeterministic quadratic interpolation, with deterministic algorithms presumably coming quite close to that, and probably better for interpolations of higher degree). But to repeat, the big-O convergence towards y_0 for nondeterministic naive algorithms is typically ideal, and the big-O performance achievable by deterministic naive algorithms is either ideal or nearly so.

The convergence requirements for non-naive, interpolating algorithms may not be met during the initial stages of a search even if $E(R(x))$ is roughly parabolic near the minimum. For that reason, naive optimization may be useful during the early stages even if other methods are used later.

1.2 Fmin variant (naive algorithm)

- does not require any assumptions about smoothness of $E(R(x))$
- can achieve typically-minimal $O(\frac{1}{\sqrt{n}})$ convergence to the minimal y value under a wide variety of conditions
- may achieve $O(\frac{1}{\sqrt{n}})$ convergence to the minimal x value if slope is discontinuous at x_0, but typically does not achieve ideal convergence in x (see below) – but then, I am not aware of any algorithms that *do* achieve best convergence rates in x under typical conditions.
- Very sensitive to poor initial settings – if x_0 is the starting position and δ is the initial step size, then the fmin variant performs quite badly if it is very expensive to try to distinguish between $R(x_0 - \delta)$, $R(x_0)$, and $R(x_0 + \delta)$ by sampling.

Fmin loses its extreme simplicity when adapted to random variables with one parameter. Extra pivot values (comparing only two points at a time can create problems), bookkeeping, backtracking, an extrapolation phase, and significance testing for comparisons between points were needed.

2 Alternatives

As Rémi Coulom alludes to in his slideshow on QLR, performing a best fit to a curve that provides a second-order fit makes it possible to do better than $O(n^{-1/4})$ convergence to x_0 for variables that have defined nonzero second derivative in a neighborhood of the minimum.

Greedy minimization is not effective for performing an accurate second-order fit. An interval just large enough to discern the sign of the slope between two points is not enough to actually measure the slope with reasonable accuracy – that requires samples over a larger interval. (An actual greedy algorithm will have sampled some points farther away from the minimum during previous steps, but as the rate of progress slows so drastically as one gets very close to the minimum, the great majority of samplings every taken will be within a factor of two or three of this greedy minimum distance – not enough to perform a second-degree fit with sufficient accuracy.)

2.1 $O(n^{-1/3})$ convergence to x_0 typically possible with second-order fits

That is, the second-order fit does in fact improve asymptotic convergence towards x_0 as compared to greedy minimization.

2.1.1 Sampling error

Suppose we have sampled the three points $\{x_1 = x_2 - \delta, x_2, x_3 = x_2 + \delta\}$ n times apiece.

To estimate error, I will assume that $\text{var}(R(x)) < v$ for $x \in [x_1, x_3]$, and that $f^{(3)}(x)$ exists within the interval. Assuming the third derivative and variance equal these maximum values throughout the entire interval allows us to set an upper bound on the standard error of the second-degree fit.

$$R(x) = c_3(x - x_2)^3 + c_2(x - x_2)^2 + c_1(x - x_2) + c_0 + V$$

where V is a random variable such that $mean(V) = 0$ and $\text{var}(V) = v$.

Let y_1, y_2, and y_3 be the three sample means. A quadratic interpolation of these points yields

$$x_{0_{est}} = x_2 - \frac{\delta(y_3 - y_1)}{2(y_3 + y_1 - 2y_2)}$$

The standard error of each of the y_i estimates of $E(R(x_i))$ is

$$(\text{var}(R(x_i))/n)^{1/2} \leq \sqrt{\frac{v}{n}}$$

The standard error of $y_3 - y_1$ is $\sqrt{2}$ times that amount. If the standard error of the denominator is much less than its value; that is, if

$$y_3 + y_1 - 2y_2 \gg \sqrt{\frac{6v}{n}}$$

then even if $E(R(x))$ really is a quadratic polynomial (i.e. higher-order fit error does not exist), we have

$$\textit{standard sampling error} \text{ of } x_{0_{est}} \simeq \frac{\sqrt{\frac{2v}{n}}}{4c_2\delta}$$

So if the variance remains within the same bounds as the interval size is increased, then larger intervals reduce the impact of sampling error upon our estimate of x_0.

2.1.2 Higher-order errors from inaccuracy of second-order fit

Next, compute the estimate error attributable to a nonzero third derivative:

$$\begin{aligned} E(R(x)) &= c_3(x-x_2)^3 + c_2(x-x_2)^2 + c_1(x-x_2) + c_0 \\ y_1 &= -c_3\delta^3 + c_2\delta^2 - c_1\delta + c_0 \\ y_2 &= c_0 \\ y_3 &= c_3\delta^3 + c_2\delta^2 + c_1\delta \\ y_3 - y_1 &= 2c_3\delta^3 + 2c_1\delta \\ y_1 + y_3 - 2y_2 &= 2c_2\delta^2 \end{aligned}$$

Plugging these y values into the quadratic interpolation of

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$$

yields the approximation

$$x_{0_{est}} \simeq x_2 - \frac{c_3\delta^2 + c_1}{2c_2}$$

while in fact

$$x_0 = x_2 - \frac{c_2 - \sqrt{{c_2}^2 - 3c_1c_3}}{3c_3}$$

If the middle of our interval coincides with the actual minimum, that is, if $x_2 = x_0$, then $c_1 = 0$ (the slope is zero at the minimum) and the error in our second-degree fit equals

$$\textit{higher-order error} \simeq \frac{|c_3|\delta^2}{2c_2}$$

This error estimate will be approximately correct if the interval is roughly centered about the minimum and small enough that second-order error dominates third- and higher-order error terms within that interval – specifically, provided that

$$|c_1| \ll \frac{{c_2}^2}{|c_3|}$$

If $E(R(x))$ is really parabolic at the minimum, then c_2 is nonzero there. Also, c_1 is the slope at x_2, which should be your previous estimate of the minimum x value, so if your estimate is good, c_1 will be small. Finding a first interval over which these requirements hold may be tricky, but once they

do, they will continue to hold during later iterations if your estimates do not worsen.

2.1.3 Minimizing the sum of sample and higher-order error

These two error sources are independent, so the total standard error is the square root of the sum of their squares. Minimizing the standard error with respect to δ yields

$$\delta_{best} = \left(\frac{v}{4nc_3{}^2}\right)^{\frac{1}{6}}$$

(Aside: since one error level rises with δ while the other falls, a pretty good approximation of the minimum – one that provides a standard error no greater than $\sqrt{2}$ times the true minimum – is to set the two error terms equal, thus minimizing the maximum of the two terms to be squared and then summed. I don't do that here, but I do that in the not-quite-working-yet hopefully-practical version.)

If $c_3 \simeq 0$ near the minimum, then one crudely pessimistic solution is to let $c_3 = \delta c_4$ in the previous equation and re-solve, yielding

$$\delta_{best} = \left(\frac{v}{4nc_4{}^2}\right)^{\frac{1}{8}}$$

if this yields a smaller δ value. I won't talk about that case much, though; the anomalous case of a zero third derivative at the minimum can be left until after the ordinary case is in better shape.

Plugging δ_{best} into the total standard error equation yields

$$\textit{standard error} \simeq \frac{k}{c_2}(\frac{c_3 v}{n})^{\frac{1}{3}}$$

where

$$\begin{aligned}
k &= \sqrt{2^{-10/3} + 2^{-7/3}} \simeq 0.546 \\
v &\simeq \max\{\mathrm{var}(R(x)) : x \in [x_0 - \delta, x_0 + \delta]\} \\
c_2 &= \sqrt{|\frac{\partial^2}{\partial x^2} E(R(x_0))|} \\
|c_3| &\simeq \max \begin{cases} \max\{|\frac{\partial}{\partial x^3} E(R(x))|/6 : x \in [x_0 - \delta, x_0 + \delta]\}, \\ \frac{1}{\delta^3}\sqrt{\frac{v}{n}} \end{cases}
\end{aligned}$$

(This is a bit circular, in that δ is a dependent variable. However, if the limit of the variance in a neighborhood around the minimum is defined and nonzero, and if the third derivative is nonzero, then it is possible to estimate these values before δ is known. Extreme accuracy is not critical. If one's estimate of both c_3 and v is half the correct value, then at worst the total error is increased by 60%.)