Response to Pull Request:
https://github.com/scikit-learn/scikit-learn/pull/11116 by webdrone.

So I had some time today and decided to thoroughly go through the derivation of the gradient. I don't think there is a problem with the original gradient calculation, at least for the squared exponential.

Consider that the optimisation is over the space of θ, where $\theta_i = \ln(\lambda_i)$, and

$$k(x, y) = exp[-1/2(x-y)^\top M(x-y)],$$

where $M = \mathrm{diag}(\lambda^2)$. This is the right parameterisation since the kernel uses the hyperparameter 'length_scale' (λ) in this manner (it divides all input by λ to compute distances, making $M_{ii} = \lambda_i^2$).

Then

$$\frac{\partial k(x, y)}{\partial \lambda_i} = \frac{1}{\lambda_i^3}(x_i - y_i)^2 * k(x, y),$$

and so we compute every element of $\frac{\partial K}{\partial \lambda_i}$. Also expressed as:

$$\frac{\partial K}{\partial \lambda_i} = \frac{1}{\lambda_i^3} \mathrm{cdist}(X[:, i], X[:, i], \text{``sqeuclidean''}) * K,$$

where the last multiplication is element-wise.

Differentiating the marginal log-likelihood with respect to θ yields the result of R&W Chapter 5 , eq (5.9). Then with chain rule:

$$\begin{aligned}
\frac{\partial}{\partial \theta_i}\ell\ell &= \frac{1}{2}\mathrm{tr}(\alpha\alpha^\top - K^{-1})\frac{\partial K}{\partial \theta_i}) && (1)\\
&= \frac{1}{2}\mathrm{tr}((\alpha\alpha^\top - K^{-1})\frac{\partial K}{\partial \lambda_i}\frac{\partial \lambda_i}{\partial \theta_i}) && (2)\\
&= \frac{1}{2}\mathrm{tr}((\alpha\alpha^\top - K^{-1})\frac{\partial K}{\partial \lambda_i}\lambda_i) && (3)\\
&= \frac{1}{2}\mathrm{tr}((\alpha\alpha^\top - K^{-1})\frac{1}{\lambda_i^2}\mathrm{cdist}(X[:, i], X[:, i], \text{``sqeuclidean''}) * K). && (4)
\end{aligned}$$

Notice how the division is by λ_i^2, not λ_i^3, because of the λ_i factor from the chain rule since we optimise wrt $\theta = \ln \lambda$. This is the original operation being done before the change introduced in this PR.

I guess my problem really was too hard for the optimiser after all. I would be very happy if you spot an error in the above.