Preconditioning for stochastic gradient descent

Jacob Hilton

August 7, 2019

Preconditioning is used in gradient descent to adjust for different curvature in different directions. In this note we argue that preconditioning for stochastic gradient descent should be thought of as having a second purpose, which may be just as important: adjusting for different signal-to-noise ratios in different directions. This is an old idea, but worth emphasising as the two purposes are sometimes conflated.

Gradient descent is usually justified as follows. Given a loss function $L(\boldsymbol{\theta})$, we would like to change the parameter vector $\boldsymbol{\theta}$ by some small amount in a way that most efficiently decreases the loss. Assuming L is differentiable, $L(\boldsymbol{\theta} - \alpha\mathbf{v}) \approx L(\boldsymbol{\theta}) - \alpha\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{v}$ for sufficiently small α, and so we solve the optimization problem

$$\underset{\mathbf{v}}{\text{minimize}} \quad L(\boldsymbol{\theta}) - \alpha\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{v} \quad \text{subject to} \quad \mathbf{v}^{\mathsf{T}}\mathbf{v} \leq 1,$$

the solution of which is to take $\mathbf{v} \propto \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. (Here $\propto$ denotes proportionality, i.e. collinearity.)

In stochastic gradient descent, we do not have access to the gradient of L, only to a stochastic version

$$\mathbf{g}_{\mathrm{s}} := \mathbf{g}_{\mathrm{t}} + \boldsymbol{\varepsilon}, \qquad \text{where} \qquad \mathbf{g}_{\mathrm{t}} := \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

and $\boldsymbol{\varepsilon}$ is zero-mean noise. (Here "s" stands for "stochastic" and "t" stands for "true".) Usually we simply make do with this stochastic version, but let us see if we can somehow do better by pre-multiplying $\mathbf{g}_{\mathrm{s}}$ by a diagonal matrix $\mathbf{T}$. In other words, let us solve the optimization problem

$$\underset{\mathbf{T}}{\text{minimize}} \quad \mathbb{E}\left[L(\boldsymbol{\theta}) - \alpha\mathbf{g}_{\mathrm{t}}^{\mathsf{T}}(\mathbf{T}\mathbf{g}_{\mathrm{s}})\right] \quad \text{subject to} \quad \mathbb{E}\left[(\mathbf{T}\mathbf{g}_{\mathrm{s}})^{\mathsf{T}}(\mathbf{T}\mathbf{g}_{\mathrm{s}})\right] \leq 1 \quad \text{and} \quad \mathbf{T} = \operatorname{diag}(\mathbf{T})$$

which forces us to take into account the effect of noise on how far we expect to move. In the absence of noise, we cannot do better than taking $\mathbf{T}$ to be a multiple of the identity matrix (since $\mathbf{T}\mathbf{g}_{\mathrm{s}} \propto \mathbf{g}_{\mathrm{t}}$ is optimal if $\mathbf{T}$ is not constrained to be diagonal), but in the prescence of noise, the solution becomes

$$\boxed{\mathbf{T} \propto \operatorname{diag}\left(\mathbf{g}_{\mathrm{t}}\mathbf{g}_{\mathrm{t}}^{\mathsf{T}}\right)\operatorname{diag}\left(\mathbb{E}\left[\mathbf{g}_{\mathrm{s}}\mathbf{g}_{\mathrm{s}}^{\mathsf{T}}\right]\right)^{-1}}.$$

Since $\mathbb{E}\left[\mathbf{g}_{\mathrm{s}}\mathbf{g}_{\mathrm{s}}^{\mathsf{T}}\right] = \mathbf{g}_{\mathrm{t}}\mathbf{g}_{\mathrm{t}}^{\mathsf{T}} + \operatorname{Var}[\mathbf{g}_{\mathrm{s}}]$, this causes us to not move as far in directions in which there is a low signal-to-noise ratio. Of course, we do not have access to this value of $\mathbf{T}$, but we may be able to approximate it using long-running averages, as long as it does not change too quickly. For example, the denominator in the Adam optimizer [Kingma and Ba, 2014] is an approximation of $\operatorname{diag}\left(\mathbb{E}\left[\mathbf{g}_{\mathrm{s}}\mathbf{g}_{\mathrm{s}}^{\mathsf{T}}\right]\right)^{1/2}$. This suggests that Adam is not best viewed as purely adjusting for curvature, as is sometimes implied (though Adam also works well in low noise settings, and there are reasons to expect curvature and noise to be related).

We obtain a corresponding result for Newton's method. This uses the second-order expansion $L(\boldsymbol{\theta} - \alpha\mathbf{v}) \approx L(\boldsymbol{\theta}) - \alpha\mathbf{g}_{\mathrm{t}}^{\mathsf{T}}\mathbf{v} + \frac{1}{2}\alpha^2\mathbf{v}^{\mathsf{T}}\mathbf{H}\mathbf{v}$, where $\mathbf{H}$ is the Hessian of $L(\boldsymbol{\theta})$, which we assume to be positive definite. Taking $\alpha = 1$, this expression is minimized by taking $\mathbf{v} = \mathbf{H}^{-1}\mathbf{g}_{\mathrm{t}}$, but if we instead solve the optimization problem

$$\underset{\mathbf{T}}{\text{minimize}} \quad \mathbb{E}\left[L(\boldsymbol{\theta}) - \mathbf{g}_{\mathrm{t}}^{\mathsf{T}}\left(\mathbf{H}^{-1}\mathbf{T}\mathbf{g}_{\mathrm{s}}\right) + \frac{1}{2}\left(\mathbf{H}^{-1}\mathbf{T}\mathbf{g}_{\mathrm{s}}\right)^{\mathsf{T}}\mathbf{H}\left(\mathbf{H}^{-1}\mathbf{T}\mathbf{g}_{\mathrm{s}}\right)\right] \quad \text{subject to} \quad \mathbf{T} = \operatorname{diag}(\mathbf{T}),$$

then we again obtain $\mathbf{T} = \operatorname{diag}\left(\mathbf{g}_{\mathrm{t}}\mathbf{g}_{\mathrm{t}}^{\mathsf{T}}\right)\operatorname{diag}\left(\mathbb{E}\left[\mathbf{g}_{\mathrm{s}}\mathbf{g}_{\mathrm{s}}^{\mathsf{T}}\right]\right)^{-1}$. Thus we see that preconditioning may be used to adjust for both curvature and signal-to-noise ratios independently.

A corresponding result may also be obtained for natural gradient descent, with the Hessian replaced by the Fisher information matrix.

We may interpret the diagonal components of $\mathbf{T}$ as "signal-to-noisy-signal ratios". We may also write $\mathbf{T} = \frac{1}{1+{}^1\!/\mathbf{S}}$, where $\mathbf{S} = \operatorname{diag}\left(\mathbf{g}_{\mathrm{t}}\mathbf{g}_{\mathrm{t}}^{\mathsf{T}}\right) \operatorname{diag}\left(\operatorname{Var}\left[\mathbf{g}_{\mathrm{s}}\right]\right)^{-1}$, whose diagonal components are signal-to-noise ratios. Note that $\mathbf{T} \approx \mathbf{S}$ when $\mathbf{S} \ll 1$, the so-called small-batch regime in which large-scale deep learning often takes place.

Note that preconditioning for curvature and for signal-to-noise ratios can be done in either order, in the sense that the solution to the optimization problem

$$\underset{\mathbf{T}}{\operatorname{minimize}} \quad \mathbb{E}\left[L\left(\boldsymbol{\theta}\right) - \mathbf{g}_{\mathrm{t}}^{\mathsf{T}}\left(\mathbf{T}\mathbf{H}^{-1}\mathbf{g}_{\mathrm{s}}\right) + \frac{1}{2}\left(\mathbf{T}\mathbf{H}^{-1}\mathbf{g}_{\mathrm{s}}\right)^{\mathsf{T}}\mathbf{H}\left(\mathbf{T}\mathbf{H}^{-1}\mathbf{g}_{\mathrm{s}}\right)\right] \quad \text{subject to} \quad \mathbf{T} = \operatorname{diag}\left(\mathbf{T}\right)$$

is $\mathbf{T} = \operatorname{diag}\left(\tilde{\mathbf{g}}_{\mathrm{t}}\tilde{\mathbf{g}}_{\mathrm{t}}^{\mathsf{T}}\right) \operatorname{diag}\left(\mathbb{E}\left[\tilde{\mathbf{g}}_{\mathrm{s}}\tilde{\mathbf{g}}_{\mathrm{s}}^{\mathsf{T}}\right]\right)^{-1}$, where $\tilde{\mathbf{g}}_{\mathrm{t}} = \mathbf{H}^{-1}\mathbf{g}_{\mathrm{t}}$ and $\tilde{\mathbf{g}}_{\mathrm{s}} = \mathbf{H}^{-1}\mathbf{g}_{\mathrm{s}}$.

Acknowledgements

Thanks to Sam McCandlish for introducing me to the idea that Adam is adjusting for signal-to-noise ratios.

References

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.