Balanced excitation and inhibition could help estimate gradients

Local synaptic plasticity improves downstream outcomes even when those outcomes are multiple synapses away. How neurons solve this "credit assignment problem" is a longstanding open problem in theoretical neuroscience. One proposed way credit assignment could work is if feedback connections provide "targets" and neurons are adjusted locally in the direction of their targets. This is the idea behind Target Propagation (TP) and Difference Target Propagation (DTP), two well-known strategies for credit assignment which share ties to predictive coding. While DTP outperforms TP, it requires multiple feedforward and feedback passes to compute each target, reducing its biological plausibility. Here, we begin by rederiving TP and DTP using a Bayesian approach and show that the two terms in DTP correspond to the gradients of conditional and prior log probabilities of each layer's activity. Next, we propose a new rule for gradient estimation that we call the Balanced Inhibition Gradient (BIG). BIG recruits local inhibition to estimate the gradient of the prior term. This can be derived by interpreting tightly-balanced excitation (E) and inhibition (I) as a local denoising circuit, analogous to how diffusion models estimate probability gradients. This change not only eliminates the additional forward/backward passes in DTP, but also provides a novel explanation for the tight E/I balance seen empirically in cortex. Compared with DTP, BIG has appealing properties from the perspective of biological plausibility: excitatory neurons' plasticity relates to the $E-I$ difference (i.e. the postsynaptic membrane potential), inhibition is local and recurrent (non-projecting), and local inhibition is trained to maintain a tight E/I balance. In artificial neural network simulations, we show initial experiments suggesting that BIG can effectively assign credit through multilayer networks. This approach thus promises to draw a throughline from the computational goal of credit assignment to its implementation in known physiology and plasticity rules.

Additional detail

We adopt the normative perspective that the role of neural plasticity is to improve some objective or loss function $\mathcal{L}$ by approximating gradient descent. Our aim is to derive an implementation of gradient-based learning that has ties to known properties of learning in biological neurons. In particular, we identify a link between gradient calculations and the facts that (i) plasticity at cortical synapses is steered by postsynaptic voltage, i.e. by brief excesses of excitation relative to inhibition in the postsynaptic cell; and (ii) how over longer periods of time, inhibition and excitation are tightly balanced (Hennequin, Agnes, & Vogels, 2017).

Originally proposed by (Bengio, 2014), we adopt a probabilistic interpretation of credit assignment which, as we show next, provides an alternative interpretation of TP and DTP. Many common learning objectives such as mean squared error or cross entropy can be viewed as a log conditional probability on a label Y^* given the hidden activities $\mathbf{h}^l$ in any layer l:

$$\mathcal{L} = -\log \mathrm{p}(\mathrm{Y}^*|\mathbf{h}^l)\,.$$

Applying Bayes' rule, we can rewrite this as $\mathcal{L} = -\log \mathrm{p}(\mathbf{h}^l|\mathrm{Y}^*) - \log \mathrm{p}(\mathrm{Y}^*) + \log \mathrm{p}(\mathbf{h}^l)$. The gradient of the loss can then be written

$$-\nabla_{\mathbf{h}^l}\mathcal{L} = \nabla_{\mathbf{h}^l}\log \mathrm{p}(\mathbf{h}^l|\mathrm{Y}^*) - \nabla_{\mathbf{h}^l}\log \mathrm{p}(\mathbf{h}^l) \qquad (1)$$

The loss gradient with respect to $\mathbf{h}^l$ is thus a difference between two score functions. The first term is the *score of the conditional* $\mathrm{p}(\mathbf{h}^l|\mathrm{Y}^*)$. The second term is the *score of the prior* $\mathrm{p}(\mathbf{h}^l)$.

Various algorithms can be derived as approximations to equation (1), clarifying the links between several previous proposals for credit assignment. In TP, as well as in many proposals related to predictive coding (Whittington & Bogacz, 2017), one introduces a learned feedback network $g_l(\mathrm{Y}^*)$ that aims to predict $\mathbf{h}^l$ given Y^*. Specifically, it if one uses an isotropic Gaussian approximation $\mathrm{p}(\mathbf{h}^l|\mathrm{Y}^*) \approx \mathrm{q}(\mathbf{h}^l|\mathrm{Y}^*) = \mathcal{N}(\mathrm{g_l}(\mathrm{Y}^*), \mathbf{I})$, then the conditional score gradient in (1) is simply

$$\nabla_{\mathbf{h}^l}\log \mathrm{p}(\mathbf{h}^l|\mathrm{Y}^*) \approx \nabla_{\mathbf{h}^l}\log \mathrm{q}(\mathbf{h}^l|\mathrm{Y}^*) = \mathbf{t}^l - \mathbf{h}^l \qquad (2)$$

where $\mathbf{t}^l \equiv g_l(\mathrm{Y}^*)$ are the **targets** for layer l. Further, if the parameters of the feedback network g_l are trained to minimize $||\mathbf{t}^l - \mathbf{h}^l||_2^2$, this can be interpreted as a variational fit of q to p.

In practice, TP and DTP use a layer-wise estimate for targets, so $g_l(\mathrm{Y}^*)$ is replaced with $g_l(\mathbf{t}^{l+1})$, and the targets for the top layer are set using the true gradient with step size 1: $\mathbf{t}^N = \mathbf{h}^N - \nabla_{\mathbf{h}^N}\mathcal{L}$. Classic TP uses (2) alone as an estimate of $\nabla_{\mathbf{h}^l}\mathcal{L}$, which is equivalent to *dropping the prior* in (1), which results in a bias.

DTP improves on TP by adjusting the targets $\mathbf{t}_l$ with an additional forward/backward pass,

$$\mathbf{t}^l = g_l(\mathbf{t}^{l+1}) - \left[g_l(f_{l+1}(\mathbf{h}^l)) - \mathbf{h}^l\right], \qquad (3)$$

where $\mathbf{h}^{l+1} = f_{l+1}(\mathbf{h}^l)$ is a single layer of the forward model. The result of this new target is that the gradient estimate is $-\nabla_{\mathbf{h}^l}\mathcal{L} \approx g_l(\mathbf{t}^{l+1}) - g_l(f_{l+1}(\mathbf{h}^l))$. The term in brackets in (3) can be seen as an approximation to the score of the prior in (1) (Bengio, 2014). This works because each f_{l+1}, g_l pair forms a denoising autoencoder, which provides an estimate of the score (Bengio, 2014; Kadkhodaie & Simoncelli, 2020); this is the same principle used by generative diffusion models.

The drawback of DTP in terms of biological plausibility is that it requires that feedback happens twice – once

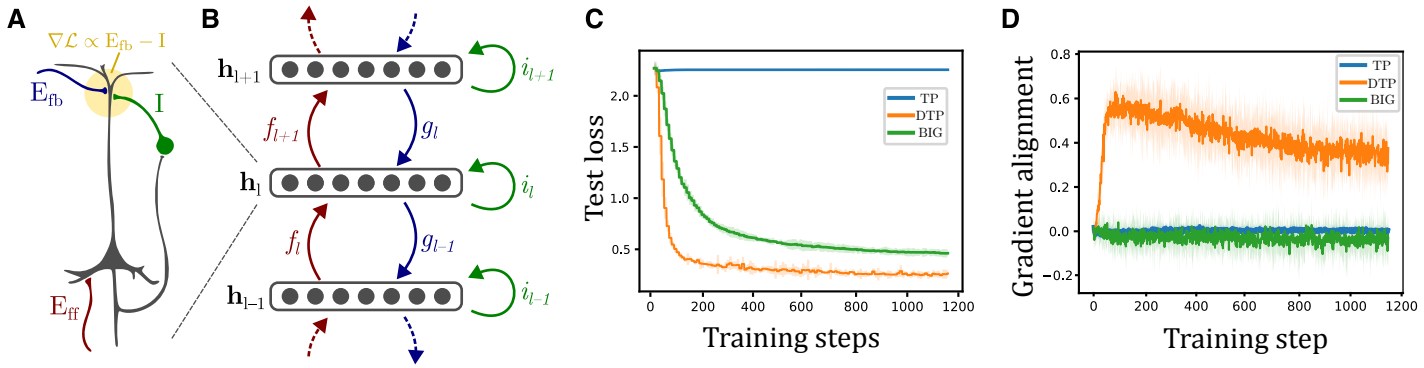

Figure 1: **A)** Sketch of a pyramidal cell (gray) with excitatory feedback connections ($\mathrm{E_{fb}}$ blue), excitatory feedforward connections ($\mathrm{E_{ff}}$, red), and local inhibitory connections (I, green). In the BIG algorithm, the loss gradient is proportional to the difference between $\mathrm{E_{fb}}$ and I (yellow highlight). **B)** We implemented TP, DTP, and BIG in a multilayer perception trained to solve MNIST. Each layer has a feedforward (f), feedback (g), and local inhibition (i) model, with colors as in (A). **C)** BIG can effectively train a 5-layer MLP, far outperforming TP though underperforming DTP. Shaded area is the variance of 10 random seeds. **D)** For discussion we wish to present a mystery: despite decreasing the loss, the estimated gradients in BIG do not align with the true gradients.

using $\mathbf{h}^{l+1}$, and again using $\mathbf{t}^{l+1}$. These pass through the same connections, and thus must sequential in time.

In the BIG algorithm, we propose using local circuits to estimate the score of the prior rather than feedback. Specifically, we introduce a local operation at each layer $\hat{\mathbf{h}}^l = i_l(\mathbf{h}^l + \varepsilon)$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is zero-mean noise. This local operation is trained as a denoising autoencoder by minimizing $||\hat{\mathbf{h}}^l - \mathbf{h}^l||_2^2$. An estimate of the marginal score is then $\nabla_{\mathbf{h}^l} \log \mathrm{p}(\mathbf{h}^l) \approx (i_l(\mathbf{h}^l) - \mathbf{h}_l)/\sigma^2$. Overall, we obtain for the BIG algorithm:

$$\mathbf{t}^l = \mathbf{h}^l\sigma^{-2} + g_l(\mathbf{t}^{l+1}) - i_l(\mathbf{h}^l)\sigma^{-2}, \tag{4}$$

which resembles DTP in (3) when $\sigma^2 = 1$.

To draw a connection between BIG and biology (Figure 1A,B), we interpret f_l as feedforward excitation ($\mathrm{E_{ff}}$), g_l as feedback excitation ($\mathrm{E_{fb}}$), and i_l as local inhibition (I). Because g_l and i_l are both trained to reconstruct $\mathbf{h}^l$, $\mathrm{E_{fb}} \approx \mathrm{E_{ff}} \approx \mathrm{I}$, so local inhibition is in **tight balance** with excitation after learning. Further, in BIG, plasticity is proportional to the local $\mathrm{E_{fb}} - \mathrm{I}$ **difference**, as is thought to be the case in biology (Artola, Bröcher, & Singer, 1990).

This perspective leads to the following model of gradient-based learning in cortical projecting neurons. First, a set of local inhibitory neurons exists that learns to cancel ongoing excitation, resulting in tight E/I balance. This estimates the score of the prior term in (1). Then, excitatory feedback from downstream areas also learns to predict somatic activity (Urbanczik & Senn, 2014). This estimates the score of the conditional term in (1). Finally, local plasticity (basal synapses, putatively) is proportional to the difference of these two inputs, $\mathrm{E_{fb}} - \mathrm{I}$, and this estimates the total gradient. If downstream activity is perturbed towards a more helpful direction, this will result in the propagation of effective credit assignment.

Our model is highly reminiscent of (Sacramento, Ponte Costa, Bengio, & Senn, 2018), and may provide an alternate theoretical interpretation of this work. However, some aspects differ; no feedback is copied to inhibitory neurons in our model, for instance.

References

Artola, A., Bröcher, S., & Singer, W. (1990). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*(6288), 69–72.

Bengio, Y. (2014). How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv*.

Hennequin, G., Agnes, E. J., & Vogels, T. P. (2017). Inhibitory plasticity: balance, control, and codependence. *Annual review of neuroscience*, *40*, 557–579.

Kadkhodaie, Z., & Simoncelli, E. P. (2020). Solving linear inverse problems using the prior implicit in a denoiser. *arXiv*.

Sacramento, J., Ponte Costa, R., Bengio, Y., & Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. *NeurIPS*, *31*.

Urbanczik, R., & Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron*, *81*(3), 521–528.

Whittington, J. C. R., & Bogacz, R. (2017, 05). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, *29*(5), 1229-1262.