May 15, 2019

Conservative updates by backward-forward cycles

These notes describe an iterative scheme for computing the exact batch-conservative update for standard neural networks with arbitrary activation function f. The algorithm is in the spirit of Newton's root finding algorithm, where solutions are fixed points of the iteration scheme. However, these solutions are exactly conservative only in a local sense. When large parameter changes are required to accommodate a batch of data, the minimizer found by the algorithm may differ from the global minimizer.

The training task is to make the smallest 2-norm changes to the network parameters so that a batch of input data is exactly mapped to given output data. Such regression problems arise, for example, when batch-conservatively training an autoencoder.

For simplicity we consider networks without bias parameters.

Notation and overview

We use indices i and j for nodes and the notation $i \to j$ for an edge between nodes. A layered architecture is not assumed, and in fact the formulas are just as simple (or even simpler) without making reference to layers. However, the update computation does distinguish five kinds of nodes that every feed-forward network has, layered or not. First there are the input nodes I that hold the input data values. When successfully trained, the propagated input data will match corresponding output data that resides in the output nodes O. All nodes not in I or O comprise the hidden nodes H. We single out a subset H_- of the hidden nodes by the property that all nodes in this subset get input only from nodes in I. Similarly, the subset H_+ of hidden nodes sends outputs only to nodes in O. We use E for the set of all edges in the network.

Post and pre-activation node values are denoted x and y respectively, and the network weights are w. The equations for neuron $j \in H$ is the pair

$$y_j^k = \sum_{i \to j} x_i^k \, w_{i \to j} \tag{1a}$$

$$x_j^k = f(y_j^k), \tag{1b}$$

where $k \in K$ is the index for data items in the training batch. There is an activation function only at the hidden nodes, not the output nodes. Thus there are no x variables at the output nodes, just as there are no y variables at the input nodes. When trained, the network maps x_i^k, $i \in I$ to y_j^k, $j \in O$ for all data $k \in K$.

The initial network weights are denoted $w_{i\to j}^0$. These are updated in an "outer" optimization loop as

$$w_{i\to j}^0 \to w_{i\to j}^0 + \Delta w_{i\to j} \coloneqq w_{i\to j}. \tag{2}$$

Conservative learning aims to solve the batch-regression problem while minimizing

$$\sum_{i\to j\in E} \|\Delta w_{i\to j}\|^2. \tag{3}$$

The weight increments Δw are updated as

$$\Delta w_{i\to j} + \delta w_{i\to j} \to \Delta w'_{i\to j} \tag{4}$$

upon completion of an "inner" loop. The inner loop solves a linearized optimization problem wherein $\delta w_{i\to j}$ participates as a variable. Most of the work occurs in this inner loop, and as we will see, involves cycles of backward and forward propagation.

The inner loop is initialized by feeding-forward all items in the training batch K using the current weights $w_{i\to j}$ set by the outer loop. No approximations are made in the initialization. In the inner loop the initialized (feed-forward) node values, denoted $\tilde{x}$ and $\tilde{y}$, are treated as constants and the neuron equations (1) are linearized about these values:

$$\delta y_j^k = \sum_{i\to j} \left(\delta x_i^k \, w_{i\to j} + \tilde{x}_i^k \, \delta w_{i\to j}\right) \tag{5a}$$

$$\delta x_j^k = f'(\tilde{y}_j^k)\delta y_j^k. \tag{5b}$$

The derivatives $f'(\tilde{y}_j^k) \coloneqq df_j^k$ are also treated as constants in the inner loop.

Equation (5a) has two cases owing to the fact that $\delta x_i^k = 0$ for $i \in I$:

$$\forall\, (j \in H_-, k \in K): \quad \delta y_j^k = \sum_{i\to j} \tilde{x}_i^k \, \delta w_{i\to j} \tag{6a}$$

$$\forall\, (j \notin H_-, k \in K): \quad \delta y_j^k = \sum_{i\to j} \left(\delta x_i^k \, w_{i\to j} + \tilde{x}_i^k \, \delta w_{i\to j}\right) \tag{6b}$$

At nodes $j \in O$ we know what changes δy_j^k are required for the network to produce the given output data. Let these changes, determined by feed-forward of data in the outer optimization loop, be ϵ_j^k. Equation (5b) is then replaced by

$$\forall\, (j \in O, k \in K): \quad \epsilon_j^k = \delta y_j^k. \tag{7}$$

The linear equations in the inner optimization loop treat the output discrepancies ϵ_j^k as small. When the discrepancies are indeed small, the variables δx, δy and δw will be small as well and the the succession of linearizations performed by the outer loop have a chance of converging. Only in this case do we expect the net update Δw to be properly conservative. At the start of training, when the discrepancies often are not small, the updates might be "neo-conservative", that is, only a local minimum of (3). However, we can hope this will just be transient behavior and after a certain stage of training the discrepancies will always be sufficiently small so the updates Δw are conservative in a global sense.

Lagrangian

The constrained optimization for the batch-conservative update in the ("inner-loop") linear approximation is compactly defined by the stationary points of the following Lagrangian:

$$\mathcal{L} = \frac{1}{2} \sum_{i \to j \in E} \left(\Delta w_{i\to j} + \delta w_{i\to j}\right)^2 \tag{8a}$$

$$+ \sum_{\substack{k\in K\\ i\in H}} \beta_i^k \left(\delta x_i^k - df_i^k \, \delta y_i^k\right) \tag{8b}$$

$$+ \sum_{\substack{k\in K\\ j\in H_-}} \gamma_j^k \left(\delta y_j^k - \sum_{i\to j} \tilde{x}_i^k \, \delta w_{i\to j}\right) \tag{8c}$$

$$+ \sum_{\substack{k\in K\\ j\notin H_-\cup O}} \gamma_j^k \left(\delta y_j^k - \sum_{i\to j} \left(\delta x_i^k \, w_{i\to j} + \tilde{x}_i^k \, \delta w_{i\to j}\right)\right) \tag{8d}$$

$$+ \sum_{\substack{k\in K\\ j\in O}} \gamma_j^k \left(\epsilon_j^k - \sum_{i\to j} \left(\delta x_i^k \, w_{i\to j} + \tilde{x}_i^k \, \delta w_{i\to j}\right)\right). \tag{8e}$$

The variables in the unconstrained optimization are δw, δx, δy and the Lagrange multipliers β and γ. Stationarity with respect to the latter just reproduce equations (5b) and (6). For δx we obtain

$$\forall\, (i \in H, k \in K): \quad \beta_i^k = \sum_{i\to j} w_{i\to j} \, \gamma_j^k, \tag{9}$$

and stationarity with respect to δy implies

$$\forall\, (i \in H, k \in K): \quad \gamma_i^k = df_i^k \, \beta_i^k. \tag{10}$$

We can eliminate β between the last two equations:

$$\forall\,(i \in H, k \in K): \quad \gamma_i^k = df_i^k \sum_{i\to j} w_{i\to j}\,\gamma_j^k. \tag{11}$$

The final set of equations follow from stationarity with respect to δw:

$$\forall\,(i \to j \in E): \quad \Delta w_{i\to j} + \delta w_{i\to j} = \sum_{k\in K} \tilde{x}_i^k \gamma_j^k \tag{12}$$

$$\coloneqq \tilde{x}_i \cdot \gamma_j. \tag{13}$$

A special case of this applies to $j \in H_-$. Substituting $\delta w_{i\to j}$ for this case from (12) into (6a) we obtain

$$\forall\,(j \in H_-, k \in K): \quad \delta y_j^k = \sum_{i\to j} \tilde{x}_i^k\,\left(\tilde{x}_i \cdot \gamma_j - \Delta w_{i\to j}\right). \tag{14}$$

Making the same substitution but now into (6b) and using (5b), we obtain

$$\forall\,(j \notin H_-, k \in K): \quad \delta y_j^k = \sum_{i\to j} \left(df_i^k \delta y_i^k\, w_{i\to j} + \tilde{x}_i^k(\tilde{x}_i \cdot \gamma_j - \Delta w_{i\to j})\right). \tag{15}$$

Inner loop initialization

Each cycle of the linear optimization is initialized at the output layer:

$$\forall\,(j \in O, k \in K): \quad \epsilon_j^k = \sum_{i\to j} \left(df_i^k\,\delta y_i^k\,w_{i\to j} + \tilde{x}_i^k\,\delta w_{i\to j}\right). \tag{16}$$

Introducing an iteration counter $n = 0, 1, \ldots$ for the inner loop and the matrix notation

$$\mathcal{E}(n)_{kj} \coloneqq \epsilon_j^k - \sum_{i\to j} df_i^k\,w_{i\to j}\,\delta y(n)_i^k \tag{17}$$

$$\widetilde{X}_{ik} \coloneqq \tilde{x}_i^k \tag{18}$$

$$\delta W(n)_{ij} \coloneqq \delta w(n)_{i\to j}, \tag{19}$$

we can rewrite (16) in matrix form:

$$\mathcal{E}(n) = \widetilde{X}^T \delta W(n). \tag{20}$$

In the first cycle we set $\delta y(0) = 0$, in effect ignoring the accumulated effects of changes to the node values when resolving the output discrepancy. Since $\mathcal{E}(0)$ is just the known discrepancy ϵ_j^k from the feed-forward of data in the outer loop, by inverting (20) we can get an initial estimate of the weight changes $\delta W(0)$ on edges to the output nodes.

In the generic case of (20) we can only solve for $\delta W(n)$ when the dimensions of the $|H_+| \times |K|$ matrix $\widetilde{X}$ satisfy

$$|K| \leq |H_+|. \tag{21}$$

When the inequality is strict we get a unique solution by imposing, additionally, that $\delta W(n)$ has minimum 2-norm. But that is exactly what we seek in conservative learning. Assuming the batch size and network architecture satisfy (21), equation (20) is inverted by applying the pseudo-inverse:

$$\delta W(n) = \widetilde{X}(\widetilde{X}^T\widetilde{X})^{-1}\mathcal{E}(n). \tag{22}$$

Computing the inverse of the $|K| \times |K|$ matrix $\widetilde{X}^T\widetilde{X}$ is the only matrix inverse required by the algorithm and it is needed only once per iteration of the outer optimization loop.

Defining two more matrices, for $j \in O$,

$$\Delta W_{ij} \coloneqq \Delta w_{i\to j} \tag{23}$$

$$\Gamma(n)_{kj} \coloneqq \gamma(n)_j^k, \tag{24}$$

we can rewrite (12), for $j \in O$, in matrix form:

$$\Delta W + \delta W(n) = \widetilde{X}\,\Gamma(n). \tag{25}$$

Multiplying this by $(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T$, and using (22), we obtain a formula for the γ_j^k Lagrange multipliers for $j \in O$:

$$\Gamma(n) = (\widetilde{X}^T\widetilde{X})^{-1}\left(\widetilde{X}^T\Delta W + \mathcal{E}(n)\right). \tag{26}$$

Backward-forward cycle

One cycle of the inner optimization loop comprises four steps:

1. Initialize γ's at the output nodes.
2. Back-propagate γ's down to all the hidden nodes.

3. Initialize δy's at nodes adjacent to the input nodes.

4. Forward-propagate δy's up to the nodes adjacent to the output nodes.

Step 1 for cycle n is given by equation (26). Recall that the data for the first cycle, $\mathcal{E}(0)$, is just the set of discrepancies ϵ_j^k at the output nodes after the forward pass of the data in the training batch.

Step 2 is the evaluation of (11) by back-propagation. Propagation terminates at nodes $i \in H_-$ that only receive inputs from input nodes.

Step 3 is given by equation (14).

Step 4 is the evaluation of (15) by forward-propagation. Note that the γ's in this equation were determined in step 2 and propagation extends all the way to δy_i^k with $i \in H_+$. This defines $\mathcal{E}(n+1)$ via $\delta y(n+1)_i^k$ in equation (17) to start the next cycle.

We can use the matrix $\Gamma(n)$ that starts each cycle both as a criterion for terminating iterations and as a means for stabilizing/boosting convergence. For the former we monitor the norm

$$\|\Gamma(n+1) - \Gamma(n)\| \tag{27}$$

and terminate iterations when it falls below some threshold. For the latter we simply make the replacement

$$\Gamma(n+1) \leftarrow (1-r)\Gamma(n+1) + r\Gamma(n), \tag{28}$$

where r is a relaxation parameter and $0 < r < 1$ enhances stability. If stability is not a concern, one can try to accelerate convergence by over-relaxation, or $r < 0$.

By (12), from the converged γ's we obtain the weight changes of the inner (linear optimization) loop by

$$\forall\ (i \to j \in E): \quad \delta w_{i\to j} = \tilde{x}_i \cdot \gamma_j - \Delta w_{i\to j}. \tag{29}$$

When the inner loop is exited, the weights are updated by (4) and the same training data is fed back through the network to define $\tilde{x}$ and $\tilde{y}$ for the next round of backward-forward cycles. We can use the final $\mathcal{E}$ matrix of the inner loop to decide when to terminate the outer loop. By (22), when the converged $\|\mathcal{E}\|$ is small, so are the δw needed to fix the network outputs on the training batch.

Comparison with SGD

Because the inner optimization begins with backward propagation initialized by output discrepancies, we should not be surprised that limiting the inner loop to a single half-cycle reproduces a version of the stochastic gradient descent (SGD) algorithm.

Here is a summary of the weight update when we perform just a single outer-loop iteration and a half-cycle in the inner loop. In the first iteration of the outer loop, Δw (accumulated weight updates) is zero and (26) reduces for $n = 0$ (first inner-loop iteration) to

$$\forall\ (j \in O, k \in K): \quad \gamma_j^k = \sum_{k' \in K} (\widetilde{X}^T \widetilde{X})^{-1}_{kk'}\ \epsilon_j^{k'}. \tag{30}$$

We then use the back-propagation equation (11), again with $\Delta w = 0$, to determine the γ's on the hidden nodes:

$$\forall\ (i \in H, k \in K): \quad \gamma_i^k = f'(\tilde{y}_i^k) \sum_{i \to j} w^0_{i \to j}\, \gamma_j^k. \tag{31}$$

Finally, from (12) we get the weight updates:

$$\forall\ (i \to j \in E): \quad \delta w_{i \to j} = \sum_{k \in K} \tilde{x}_i^k \gamma_j^k. \tag{32}$$

Only (31) resembles the back-propagation of SGD, where information at the output nodes is propagated, independently for all items in the "mini-batch", to all the hidden nodes. However, already in the initialization (30) we see that there is "mixing" of the mini-batch discrepancies, and the weight updates in (32) are not a simple average but a weighted average of the propagated variables (γ's) over the mini-batch. On the other hand, we show next that for mini-batches of size 1, and up to indefiniteness in the step size, the equations above are the same as those in SGD.

To facilitate the comparison, we derive the SGD update in our notation. For the regression problem SGD computes the gradient of the loss

$$L = \frac{1}{2} \sum_{i \in O} (y_i - y_i^*)^2, \tag{33}$$

where the y^* are the given output values and we have simplified the notation for the case of a single training item. As before, $\tilde{x}$ and $\tilde{y}$ are post- and pre-activation node values when

the data is fed into the network with the current weights, w^0. The output discrepancies therefore satisfy

$$\forall\ (i \in O): \quad \tilde{y}_i + \epsilon_i = y_i^*. \tag{34}$$

By repeated application of the chain rule,

$$g_{i\to j} \coloneqq \frac{\partial L}{\partial w_{i\to j}} \tag{35a}$$

$$= \frac{\partial L}{\partial y_j}\,\frac{\partial y_j}{\partial w_{i\to j}} \tag{35b}$$

$$= \frac{\partial L}{\partial y_j}\,\tilde{x}_i \tag{35c}$$

$$= \frac{\partial L}{\partial x_j}\,\frac{\partial x_j}{\partial y_j}\,\tilde{x}_i \tag{35d}$$

$$= \frac{\partial L}{\partial x_j}\,f'(\tilde{y}_j)\,\tilde{x}_i. \tag{35e}$$

We will see that the quantity

$$\widetilde{\gamma}_j \coloneqq -\frac{\partial L}{\partial y_j} \tag{36}$$

is analogous to the Lagrange multiplier γ_j of the conservative update. For our loss function,

$$\forall\ (j \in O): \quad \widetilde{\gamma}_j = \epsilon_j. \tag{37}$$

This initializes the back-propagation.

Continued application of the chain rule to (35e) gives

$$g_{i\to j} = \frac{\partial L}{\partial y_j}\,\tilde{x}_i = f'(\tilde{y}_j)\,\tilde{x}_i \left(\sum_{j\to l} \frac{\partial L}{\partial y_l}\,\frac{\partial y_l}{\partial x_j}\right) \tag{38}$$

$$= f'(\tilde{y}_j)\,\tilde{x}_i \sum_{j\to l} \frac{\partial L}{\partial y_l}\,w^0_{j\to l}. \tag{39}$$

Using the definition (36), and assuming $\tilde{x}_i \neq 0$, we arrive at the back-propagation equation for the $\widetilde{\gamma}$'s:

$$\widetilde{\gamma}_j = f'(\tilde{y}_j) \sum_{j\to l} w^0_{j\to l}\,\widetilde{\gamma}_l. \tag{40}$$

This is identical to the γ-recursion (31) in conservative learning. In SGD, unlike conservative learning, the gradient step size is an empirical parameter s, the learning rate:

$$\delta w_{i\to j} = -s\,g_{i\to j} = s\,\tilde{x}_i\widetilde{\gamma}_j. \tag{41}$$

Soft output constraint

There are two reasons why one might want to soften the constraint that the network exactly produce the given output data. First, there may be noise in these data vectors and we want to avoid overfitting to this noise. Second, it turns out that softening this constraint eliminates the limit (21) on the size of the training batch.

In the Lagrangian (8) we replace the hard constraints on the outputs by a quadratic cost with stiffness parameter κ. In the limit $\kappa \to \infty$ we will recover the original update rule, when this is compatible with inequality (21).

The easiest way to introduce this change in the derivation above is to keep the Lagrangian $\mathcal{L}$ as is, while making the replacement

$$\epsilon_j^k \to \epsilon_j^k + \eta_j^k, \tag{42}$$

where the η's are new variables. Recall that the ϵ's are the discrepancies between the outputs (for the current weights) and the target outputs. We can think of the η variables as an attempt to reconstruct the noise in the targets, and we introduce (in $\mathcal{L}$) the penalty term

$$\frac{\kappa}{2} \sum_{\substack{k \in K \\ j \in O}} (\eta_j^k)^2 \tag{43}$$

to keep the reconstructed noise small. Since the only other place that η's appear in $\mathcal{L}$ — by (42) — is in (8e), stationarity of $\mathcal{L}$ with respect to the η variables produces the equations

$$\kappa\, \eta_j^k + \gamma_j^k = 0, \tag{44}$$

with solutions

$$\eta_j^k = -\frac{1}{\kappa} \gamma_j^k. \tag{45}$$

We keep the definition (17), but the replacement (42) and identity (45) have the effect that the matrix equation (20) is replaced by

$$\mathcal{E}(n) - \frac{1}{\kappa} \Gamma(n) = \widetilde{X}^T \delta W(n). \tag{46}$$

Since the companion matrix equation (25) is unchanged from its original form, we can solve for $\delta W(n)$ and substitute into (46) to arrive at

$$\left(\frac{1}{\kappa} + \widetilde{X}^T \widetilde{X} \right) \Gamma(n) = \widetilde{X}^T \Delta W + \mathcal{E}(n). \tag{47}$$

The $|K| \times |K|$ matrix multiplying $\Gamma(n)$ is the sum of the matrix we had originally, for hard output constraints, and a multiple of the identity. Without the latter we were limited in inverting the matrix $\widetilde{X}^T\widetilde{X}$ by the condition that its rank was $|K|$, that is, the inequality $|H_+| \geq |K|$ on the dimension of the $|H_+| \times |K|$ matrix factor $\widetilde{X}$. However, thanks to the identity term proportional to $1/\kappa$, the matrix is invertible (generically) even when $|H_+| < |K|$, that is, for arbitrarily large data batches. The upshot is that the only effect of the soft output constraint is to modify the Lagrange multiplier initialization:

$$\Gamma(n) = \left(\frac{1}{\kappa} + \widetilde{X}^T\widetilde{X}\right)^{-1} \left(\widetilde{X}^T\Delta W + \mathcal{E}(n)\right). \tag{48}$$

Taking the limit $\kappa \to \infty$ recovers the hard constraint initialization (26), except in the case when this limit is singular ($|H_+| < |K|$).