GRAPH NEURAL NETWORKS INSPIRED BY CLASSICAL ITERATIVE ALGORITHMS

Zepeng Zhang

June 14, 2022

This paper [1] proposes TWIRLS (together with iterative reweighted least squares), a simple integrated framework for combining propagation and attention layers anchored to the iterative descent of an objective function that is robust against edge uncertainty and oversmoothing.

First we consider the energy function:

$$\begin{aligned}\ell_Y(Y) &\triangleq \|Y - f(X;W)\|_{\mathcal{F}}^2 + \lambda \mathrm{tr}\left[Y^\top L Y\right] \\ &= \|Y - f(X;W)\|_{\mathcal{F}}^2 + \lambda \sum_{\{i,j\}\in\mathcal{E}} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 .\end{aligned} \tag{1}$$

Minimizing (1) is equivalent to maximum a posteriori (MAP) estimation via $P(Y \mid X) \propto p(X \mid Y)p(Y)$ with $p(X \mid Y) \propto \exp\left[-\frac{1}{2\lambda}\|Y - f(X;W)\|_{\mathcal{F}}^2\right]$ and $p(Y) \propto \prod_{\{i,j\}\in\mathcal{E}} \mathcal{N}\left(\mathbf{y}_i - \mathbf{y}_j \mid 0, I\right)$, where $p(Y)$ represents a structured Gaussian prior distribution with unit variance along each edge. Take the edge uncertainty into account, this paper extends the unit variances to a more flexible Gaussian scale mixture prior defined as

$$p(Y) = Z^{-1} \prod_{\{i,j\}\in\mathcal{E}} \int \mathcal{N}\left(\mathbf{y}_i - \mathbf{y}_j \mid 0, \gamma_{ij}^{-1} I\right) d\mu\left(\gamma_{ij}\right), \tag{2}$$

where μ is a positive measure over latent precision parameters γ_{ij} and Z is a standard partition function that ensures (2) sums to one. This extension introduces uncertainty into the allowable variance along each edge, and hence can contribute to robustness w.r.t. edge uncertainty.

Lemma 1. *For any $p(Y)$ expressible via* (2)*, we have*

$$-\log p(Y) = \pi(Y;\rho) \triangleq \sum_{\{i,j\}\in\mathcal{E}} \rho\left(\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2^2\right)$$

excluding irrelevant constants, where $\rho : \mathbb{R}^+ \to \mathbb{R}$ is a concave non-decreasing function that depends on μ.

Based on Lemma 1, we obtain a revised energy function with the Gaussian scale mixture prior as follows:

$$\ell_Y(Y;\rho) \triangleq \|Y - f(X;W)\|_{\mathcal{F}}^2 + \lambda \sum_{\{i,j\}\in\mathcal{E}} \rho\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right).$$

To build propagation layers based on this revised energy function, we introduce a variational decomposition of $\pi(Y;\rho)$:

$$\hat{\pi}\left(Y;\widetilde{\rho},\{\gamma_{ij}\}\right) \triangleq \sum_{\{i,j\}\in\mathcal{E}} \left[\gamma_{ij}\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2^2 - \widetilde{\rho}\left(\gamma_{ij}\right)\right],$$

where $\{\gamma_{ij}\}_{i,j\in\mathcal{E}}$ are redefined as a set of variational weights and $\tilde{\rho}$ is the concave conjugate of ρ. With this approximate penalty function, we form an alternative revised energy function:

$$\hat{\ell}_Y(Y;\Gamma,\widetilde{\rho}) \triangleq \|Y - f(X;W)\|_{\mathcal{F}}^2 + \lambda \mathrm{tr}\left[Y^\top \hat{L} Y\right] + f(\Gamma),$$

where Γ is a diagonal matrix with the variational parameters $\{\gamma_{ij}\}_{i,j\in\mathcal{E}}$ in the diagonal, $f(\Gamma) = -\sum_{\{i,j\}\in\mathcal{E}} \widetilde{\rho}\left(\gamma_{ij}\right)$, and $L = B^T \Gamma B$.

Lemma 2. *For all* $\{\gamma_{ij}\}_{i,j\in\mathcal{E}}$*, we have* $\hat{\ell}_Y(Y;\Gamma,\widetilde{\rho}) \geq \ell_Y(Y;\rho)$ *with equality iff*

$$\gamma_{ij} = \arg\min_{\{\gamma_{ij}>0\}} \widetilde{\pi}\left(Y;\widetilde{\rho},\{\gamma_{ij}\}\right) = \left.\frac{\partial\rho\left(z^2\right)}{\partial z^2}\right|_{z=\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|_2}.$$

Base on above results, we can minimize $\ell_Y(Y;\rho)$ using the MM method. With the initialization $Y^{(0)} = f(X;W)$, we update the variational parameters based on Lemma 2 in the majorization step, and we execute one (or possibly multiple) gradient steps on $\hat{\ell}_Y(Y;\Gamma,\widetilde{\rho})$ via

$$\begin{aligned} Y^{(k+1)} &= Y^{(k)} - \frac{\alpha}{2}\frac{\partial\hat{\ell}_Y(Y;\Gamma,\widetilde{\rho})}{\partial Y}\,|_{Y=Y^{(k)},\Gamma=\Gamma^{(k+1)}} \\ &= Y^{(k)} - \alpha\left[\left(\lambda\hat{L}+I\right)Y - f(X;W)\right], \end{aligned}$$

where $\frac{\alpha}{2}$ is the stepsize. To accelerate the convergence, the Jacobi preconditioning technique is used, which rescales the gradient using $(\lambda D+I)^{-1}$. After rearranging terms and defining the diagonal matrix $\tilde{D} = \lambda D + I$, the update becomes

$$Y^{(k+1)} = (1-\alpha)Y^{(k)} + \alpha(\tilde{D}^{(K+1)})^{-1}\left[\lambda A^{(k+1)}Y^{(k)} + f(X;W)\right],$$

where $A^{(k+1)}$ denotes the adjacency matrix A with edges weighted by $\Gamma^{(k+1)}$, similarly for $D^{(k+1)}$. The convergence of this algorithm is guaranteed with suitable α. The function ρ can be flexibly chose (e.g., truncated ℓ_p quasinorm), which can lead to different flavors of attention.

Experiment results show that TWIRLS performs well handling adversarial attacks, heterophily, or long-range dependencies.

References

[1] Yongyi Yang, Tang Liu, Yangkun Wang, Jinjing Zhou, Quan Gan, Zhewei Wei, Zheng Zhang, Zengfeng Huang, and David Wipf. Graph neural networks inspired by classical iterative algorithms. In *International Conference on Machine Learning*, pages 11773–11783. PMLR, 2021.