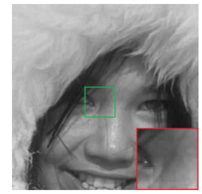
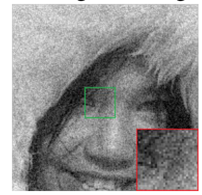
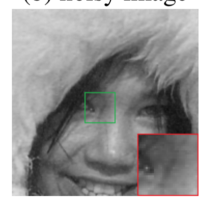
We would like to thank all reviewers for their careful reading, thoughtful comments, and overall positive assessment! We address the questions raised by reviewers below.



(a) original image



(b) noisy image



(c) denoised by $\texttt{GD}_\phi^{12}(E_\theta(X,\cdot))$

Real-world application [Reviewer #1, #2, #3, #4]. To show the real world applicability of our theoretical framework, we consider the **local adaptive image denoising** task, where the noise levels in different parts of the images can be different. More specifically,

(i) Dataset. We split BSD500 dataset (400 images) [1] into a training set (100 images) and a test set (300 images). Gaussian noises are added to each pixel with noise levels depending on image local smoothness, making the noise levels on edges lower than non-edge regions. The task is to restore the original image from the noisy version $X \in [0,1]^{180\times 180}$.
(ii) Architecture. We designed a hybrid architecture $\texttt{Alg}_\phi^k(E_\theta(X,\cdot))$ where $\texttt{Alg}_\phi^k$ is a k-step unrolled minimization algorithm to the ℓ_2-regularized reconstruction objective $E_\theta(X,Y) := \frac{1}{2}\|Y + g_\theta(X) - X\|_F^2 + \frac{1}{2}\sum_{i,j}|[f_\theta(X)]_{i,j}Y_{i,j}|^2$, and the residual $g_\theta(X)$ and position-wise regularization coefficient $f_\theta(X)$ are both DnCNN networks as in [2]. The optimization objective, $E_\theta(X,Y)$, is quadratic in Y, which follows the settings our theory focused on.
(iii) Generalization gap. We instantiate the hybrid architecture into different models using GD and NAG algorithms with different unrolled steps k. Each model is trained with 3000 epochs, and the *generalization gaps* between training and test errors are reported in Fig. 1. The results also show good consistency with our theory, where stabler algorithm (GD) can generalize better given *over-/under*-parameterized neural module, and for the *about-right* parameterization case, the generalization gap behaviors similar to *Stab*$(k) *$ *Cvg*(k). We will conduct more experimental trials and provide figures with smoother curves and error bars in our revised paper.
(iv) Visualization. To show that the learned hybrid model has a good performance in this real application, we include a visualization of the original, noisy, and denoised images.

Generality of problem setting [Reviewer #1, #2, #3, #4]. We acknowledged that our theoretical analysis is performed under a simplified problem setting, but we'd like to clarify a few points to avoid confusion.

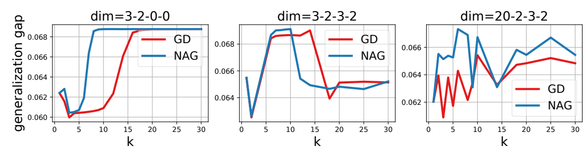


Figure 1: Generalization gap. **Each k corresponds to a separately trained model**. *Left* (under-parameterized): f_θ is a DnCNN with 3 channels and 2 hidden layers and $g_\theta = 0$. *Middle* (about-right): both f_θ and g_θ have 3 channels and 2 hidden layers. *Right (over-parameterized)*: f_θ has 20 channels.

• We assume $E_\theta(\boldsymbol{x},\boldsymbol{y})$ is quadratic in $\boldsymbol{y}$ but it can depend on *any way* in the input $\boldsymbol{x}$ (i.e., Q_θ can be any neural network). This can cover many real applications. For example, the above image denoising task, and many other data reconstruction problems can be cast into quadratic energy minimization.
• Even though in the paper we only present the results for using GD and NAG algorithms as the reasoning module, the main theorems which state the relation between the learning behavior and algorithm properties can be applied to *any optimization algorithm* as long as corresponding properties of the algorithm are provided. **[Reviewer #3]**.
• Our analysis framework can be extended to more general settings where the neural network module outputs a suitable strongly convex energy function. In fact the key component of our approximation and generalization analysis is the Lipschitz stability of the maps between the energy function and the exact minimizer, which can be ensured for general convex optimization problems if suitable regularization terms are introduced in the energy functions. We aim to analyze this general setting in future research.

Other questions.
• Space shrinks to a single function. **[Reviewer #2]** We sincerely thank Reviewer #2 for his/her very detailed comments. Regarding why $\{\texttt{Alg}_\phi^\infty : \phi \in \Phi\}$ contains only a single function, this comes from the convergence guarantee of the algorithms. That is, when the step size ϕ is in the stable region Φ, the optimization error will decrease in each iteration and gradually decrease to 0 when $k \to \infty$. Therefore, for every step size $\phi \in \Phi$, $\texttt{Alg}_\phi^\infty = \texttt{Opt}$, the exact minimizer.
• Generalized to DP? **[Reviewer #3]** Thanks Reviewer #3 for bringing in this interesting question, which is also what we want to address in our future work. Our analysis for RNN/GNN can potentially be adapted to the case where RNN/GNN are used to learn problems requiring DP, since RNN/GNN can present those operations in DP. However, as explained in our theory, one may not able to obtain as a tight bound as GD/NAG due to the difficulty of analyzing RNN. Furthermore, in the case of DP, the notion of convergence with respect the number of step k is different: the k is a fixed number of stages needed to run the DP iterations to solve the optimization. This will require more research.

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.
[2] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep51cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.