Momentum-Net: Fast and convergent iterative neural network for inverse problems (Supplement)

Il Yong Chun, *Member, IEEE*, Zhengyu Huang*, Hongki Lim*, *Student Member, IEEE*, and Jeffrey A. Fessler, *Fellow, IEEE*

This supplementary material for [1] *a)* reviews the Block Proximal Extrapolated Gradient method using a Majorizer (BPEG-M) [2], [3], *b)* lists parameters of Momentum-Net, and summarizes selection guidelines or gives default values, *c)* compares the convergence properties between Momentum-Net and BCD-Net, and *d)* provides mathematical proofs or detailed descriptions to support several arguments in the main manuscript. We use the prefix "S" for the numbers in section, theorem, equation, figure, table, and footnote in the supplement.

S.1 BPEG-M: Review

This section explains *multi-(non)convex* optimization problems, and summarizes the state-of-the-art method for block multi-(non)convex optimization method, BPEG-M [2], [3], along with its convergence guarantees.

S.1.1 Multi-(non)convex optimization

In a block optimization problem, the variables of the underlying optimization problem are treated either as a single block or multiple disjoint blocks. In multi-(non)convex optimization, we consider the following problem:

$$\min_{u} \; F(u_1, \ldots, u_B) \triangleq f(u_1, \ldots, u_B) + \sum_{b=1}^{B} r_b(u_b) \tag{S.1}$$

where variable u is decomposed into B blocks $u_1, \ldots, u_B$ ($\{u_b \in \mathbb{R}^{n_b} : b = 1, \ldots, B\}$), f is assumed to be (continuously) differentiable, but functions $\{r_b : b = 1, \ldots, B\}$ are not necessarily differentiable. The function r_b can incorporate the constraint $u_b \in \mathcal{U}_b$, by allowing r_b's to be extended-valued, e.g., $r_b(u_b) = \infty$ if $u_b \notin \mathcal{U}_b$, for $b = 1, \ldots, B$. It is standard to assume that both f and $\{r_b\}$ are proper and closed, and the sets $\{\mathcal{U}_b\}$ are closed. We consider either that (S.1) has block-wise convexity (but (S.1) is jointly nonconvex in general) [2], [4] or that f, $\{r_b\}$, or $\{\mathcal{U}_b\}$ are not necessarily convex [3], [5]. Importantly, r_b can include (non)convex and nonsmooth ℓ^p (quasi-)norm, $p \in [0, 1]$. The next section introduces our optimization framework that solves (S.1).

The following sections review BPEG-M [2], [3], the state-of-the-art optimization framework for solving multi-(non)convex problems, when used with sufficiently sharp majorizers. BPEG-M uses block-wise extrapolation, majorization, and proximal mapping. By using a more general Lipschitz continuity (see Definition 1) for block-wise gradients, BPEG-M is particularly useful for rapidly calculating majorizers involved with large-scale problems, and successfully applied to some large-scale machine learning and computational imaging problems; see [2], [3], [6] and references therein.

S.1.2 BPEG-M

This section summarizes the BPEG-M framework. Using Definition 1 and Lemma 2, the proposed method, BPEG-M, is given as follows. To solve (S.1), we minimize majorizers of F cyclically over each block $u_1, \ldots, u_B$, while fixing the remaining blocks at their previously updated variables. Let $u_b^{(i+1)}$ be the value of u_b after its ith update, and define

$$f_b^{(i+1)}(u_b) \triangleq f\Big(u_1^{(i+1)}, \ldots, u_{b-1}^{(i+1)}, u_b, u_{b+1}^{(i)}, \ldots, u_B^{(i)}\Big),$$

- *Supplementary material dated November 22, 2020.*
- *The authors indicated by asterisks (*) contributed equally to this work.*
- *Il Yong Chun was with the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48019 USA, and is now with the Department of Electrical Engineering, the University of Hawai'i at Mānoa, Honolulu, HI 96822 USA (email: iychun@hawaii.edu). Zhengyu Huang, Hongki Lim, and Jeffrey A. Fessler are with the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48019 USA (email: zyhuang@umich.edu; hongki@umich.edu; fessler@umich.edu).*

Algorithm S.1 BPEG-M [2], [3]

Require: $\{u_b^{(0)} = u_b^{(-1)} : \forall b\}$, $\{w_b^{(i)} \in [0,1], \forall b, i\}$, $i = 0$
  while a stopping criterion is not satisfied **do**
    for $b = 1, \ldots, B$ **do**
      Calculate $\widetilde{M}_b^{(i+1)}$ by (S.4), and $E_b^{(i+1)}$ to satisfy (S.5) or (S.6)
      $\acute{u}_b^{(i+1)} = u_b^{(i)} + E_b^{(i+1)}\big(u_b^{(i)} - u_b^{(i-1)}\big)$
      $u_b^{(i+1)} = \mathrm{Prox}_{r_b}^{\widetilde{M}_b^{(i+1)}}\Big(\acute{u}_b^{(i+1)} - \big(\widetilde{M}_b^{(i+1)}\big)^{-1}\nabla f_b^{(i+1)}(\acute{u}_b^{(i+1)})\Big)$
    end for
    $i = i + 1$
  end while

for all b, i. At the bth block of the ith iteration, we apply Lemma 2 to functional $f_b^{(i+1)}(u_b)$ with a $M^{(i+1)}$-Lipschitz continuous gradient at the extrapolated point $\acute{u}_b^{(i+1)}$, and minimize a majorized function. In other words, we consider the updates

$$\begin{aligned} u_b^{(i+1)} &= \underset{u_b}{\mathrm{argmin}}\, \langle \nabla f_b^{(i+1)}(\acute{u}_b^{(i+1)}), u_b - \acute{u}_b^{(i+1)}\rangle + \frac{1}{2}\Big\|u_b - \acute{u}_b^{(i+1)}\Big\|^2_{\widetilde{M}_b^{(i+1)}} + r_b(u_b) \\ &= \mathrm{Prox}_{r_b}^{\widetilde{M}_b^{(i+1)}}\Big(\underbrace{\acute{u}_b^{(i+1)} - \big(\widetilde{M}_b^{(i+1)}\big)^{-1}\nabla f_b^{(i+1)}(\acute{u}_b^{(i+1)})}_{\textit{extrapolated gradient step using a majorizer of } f_b^{(i+1)}}\Big), \end{aligned} \tag{S.2}$$

where

$$\acute{u}_b^{(i+1)} = u_b^{(i)} + E_b^{(i+1)}\big(u_b^{(i)} - u_b^{(i-1)}\big), \tag{S.3}$$

the proximal operator is defined by (2), $\nabla f_b^{(i+1)}(\acute{u}_b^{(i+1)})$ is the block-partial gradient of f at $\acute{u}_b^{(i+1)}$, a *scaled majorization matrix* is given by

$$\widetilde{M}_b^{(i+1)} = \lambda_b \cdot M_b^{(i+1)} \succ 0, \qquad \lambda_b \geq 1, \tag{S.4}$$

and $M_b^{(i+1)} \in \mathbb{R}^{n_b \times n_b}$ is a symmetric positive definite *majorization matrix* of $\nabla f_b^{(i+1)}(u_b)$. In (S.3), the $\mathbb{R}^{n_b \times n_b}$ matrix $E_b^{(i+1)} \succeq 0$ is an *extrapolation matrix* that accelerates convergence in solving multi-convex problems [2]. We design it to satisfy conditions (S.5) or (S.6) below. In (S.4), $\{\lambda_b = 1 : \forall b\}$ and $\{\lambda_b > 1 : \forall b\}$, for multi-convex and multi-nonconvex problems, respectively.

For some $f_b^{(i+1)}$ having sharp majorizers, we expect that extrapolation (S.3) has no benefits in accelerating convergence, and use $\{E_b^{(i+1)} = 0 : \forall i\}$. Other than the blocks having sharp majorizers, one can apply some increasing momentum coefficient formula [7], [8] to the corresponding extrapolation matrices. The choice in [2]–[4] accelerated BPEG-M for some machine learning and data science applications. Algorithm S.1 summarizes these updates.

S.1.3 Convergence results

This section summarizes convergence results of Algorithm S.1 under the following assumptions:

- *Assumption S.1)* In (S.1), F is proper and lower bounded in $\mathrm{dom}(F) \triangleq \{u : F(u) < \infty\}$. In addition,
  for *multi-convex* (S.1), f is differentiable and (S.1) has a Nash point or block-coordinate minimizer[S.1] (see its definition in [4, (2.3)–(2.4)]);
  for *multi-nonconvex* (S.1), f is continuously differentiable, r_b is lower semicontinuous[S.2], $\forall b$, and (S.1) has a critical point $u^\star$ that satisfies $0 \in \partial F(u^\star)$.
- *Assumption S.2)* $\nabla f_b^{(i+1)}(u_b)$ is M-Lipschitz continuous with respect to u_b, i.e.,

$$\Big\|\nabla f_b^{(i+1)}(u) - \nabla f_b^{(i+1)}(v)\Big\|_{\left(M_b^{(i+1)}\right)^{-1}} \leq \|u - v\|_{M_b^{(i+1)}},$$

  for $u, v \in \mathbb{R}^{n_b}$, where $M_b^{(i+1)}$ is a bounded majorization matrix.
- *Assumption S.3)* The extrapolation matrices $E_b^{(i+1)} \succeq 0$ satisfy that

$$\text{for } \textit{multi-convex} \text{ (S.1)}, \ \big(E_b^{(i+1)}\big)^T M_b^{(i+1)} E_b^{(i+1)} \preceq \delta^2 \cdot M_b^{(i)}; \tag{S.5}$$

$$\text{for } \textit{multi-nonconvex} \text{ (S.1)}, \ \big(E_b^{(i+1)}\big)^T M_b^{(i+1)} E_b^{(i+1)} \preceq \frac{\delta^2(\lambda_b - 1)^2}{4(\lambda_b + 1)^2} \cdot M_b^{(i)}, \tag{S.6}$$

S.1. Given a feasible set $\mathcal{U}$, a point $u^\star \in \mathrm{dom}(F) \cup \mathcal{U}$ is a critical point (or stationary point) of F if the directional derivative $d^T \nabla F(u^\star) \geq 0$ for any feasible direction d at $u^\star$. If $u^\star$ is an interior point of $\mathcal{U}$, then the condition is equivalent to $0 \in \partial F(u^\star)$.

S.2. F is lower semicontinuous at point u_0 if $\liminf_{u \to u_0} F(u) \geq F(u_0)$.

with $\delta < 1, \forall b, i$.

Theorem S.1 (*Multi-convex* (S.1): A limit point is a Nash point [2]). *Under Assumptions S.1–S.3, let* $\{u^{(i+1)} : i \geq 0\}$ *be the sequence generated by Algorithm S.1. Then any limit point of* $\{u^{(i+1)} : i \geq 0\}$ *is a Nash point of (S.1).*

Theorem S.2 (*Multi-nonconvex* (S.1): A limit point is a critical point [3]). *Under Assumptions S.1–S.3, let* $\{u^{(i+1)} : i \geq 0\}$ *be the sequence generated by Algorithm S.1. Then any limit point of* $\{u^{(i+1)} : i \geq 0\}$ *is a critical point of (S.1).*

Remark S.3. Theorems S.1–S.2 imply that, if there exists a critical point for (S.1), i.e., $0 \in \partial F(u^\star)$, then any limit point of $\{u^{(i+1)} : i \geq 0\}$ is a critical point. One can further show global convergence under some conditions: if $\{u^{(i+1)} : i \geq 0\}$ is bounded and the critical points are isolated, then $\{u^{(i+1)} : i \geq 0\}$ converges to a critical point [2, Rem. 3.4], [4, Cor. 2.4].

S.1.4 Application of BPEG-M to solving multi-(non)convex problem (1)

For update (3), we do not use extrapolation, i.e., (S.3), since the corresponding majorization matrices are sharp, so one obtains tight majorization bounds in Lemma 2. See, for example, [3, §V-B]. For updates (3) and (5), we rewrite $\sum_{k=1}^K \|h_k * x - \zeta_k\|_2^2$ as $\|x - \sum_{k=1}^K \text{flip}(h_k^*) * \zeta_k\|_2^2$ by using the TF condition in §2.1 [3, §VI], [6].

S.2 Empirical measures related to the convergence of Momentum-Net using sCNN refiners

This section provides empirical measures related to Assumption 4 for Momentum-Net using single-hidden layer autoencoders (18); see Fig. S.1 below. We estimated the sequence $\{\epsilon^{(i)} : i = 2, \ldots, N_{\text{lyr}}\}$ in Definition 7, the sequence $\{\Delta^{(i)} : i = 2, \ldots, N_{\text{lyr}}\}$ in Definition 8, and the Lipschitz constants $\{\kappa^{(i)} : i = 1, \ldots, N_{\text{lyr}}\}$ of refining NNs $\{\mathcal{R}_{\theta^{(i)}} : \forall i\}$, based on a hundred sets of randomly selected training samples related with the corresponding bounds of the measures, e.g., u and v in (11) are training input to $\mathcal{R}_{\theta^{(i+1)}}$ and $\mathcal{R}_{\theta^{(i)}}$ in (Alg.1.1), respectively.

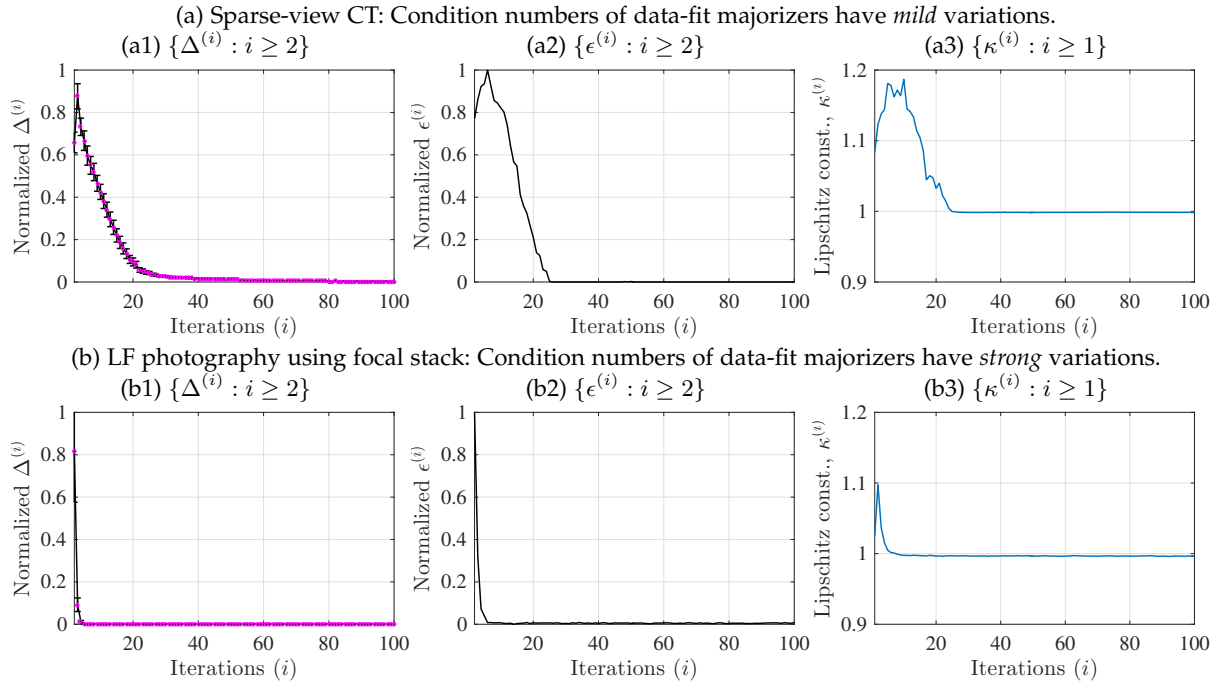

Fig. S.1. Empirical measures related to Assumption 4 for guaranteeing convergence of Momentum-Net using sCNN refiners (for details, see (18) and §4.2.1), in different applications. (a) The sparse-view CT reconstruction experiment used fan-beam geometry with 12.5% projections views. (b) The LF photography experiment used five detectors and reconstructed LFs consisting of 9×9 sub-aperture images. (a1, b1) For both the applications, we observed that $\Delta^{(i)} \to 0$. This implies that the $z^{(i+1)}$-updates in (Alg.1.1) satisfy the asymptotic block-coordinate minimizer condition in Assumption 4. (Magenta dots denote the mean values and black vertical error bars denote standard deviations.) (a2) Momentum-Net trained from training data-fits, where their majorization matrices have *mild* condition number variations, shows that $\epsilon^{(i)} \to 0$. This implies that paired NNs $(\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}})$ in (Alg.1.1) are asymptotically nonexpansive. (b2) Momentum-Net trained from training training data-fits, where their majorization matrices have *mild* condition number variations, shows that $\epsilon^{(i)}$ becomes close to zero, but does not converge to zero in one hundred iterations. (a3, b3) The NNs, $\mathcal{R}_{\theta^{(i+1)}}$ in (Alg.1.1), become nonexpansive, i.e., its Lipschitz constant $\kappa^{(i)}$ becomes less than 1, as i increases.

S.3 Probabilistic justification for the asymptotic block-coordinate minimizer condition in Assumption 4

This section introduces a useful result for an asymptotic block-coordinate minimizer $z^{(i+1)}$: the following lemma provides a *probabilistic* bound for $\|x^{(i)} - z^{(i+1)}\|_2^2$ in (12), given a subgaussian vector $z^{(i+1)} - z^{(i)}$ with independent and zero-mean entries.

Lemma S.4 (Probabilistic bounds for $\|x^{(i)} - z^{(i+1)}\|_2^2$)**.** *Assume that $z^{(i+1)} - z^{(i)}$ is a zero-mean subgaussian vector of which entries are independent and zero-mean subgaussian variables. Then, each bound in (12) holds with probability at least*

$$1 - \exp\left(\frac{-\left(\|z^{(i+1)} - z^{(i)}\|_2^2 + \Delta^{(i+1)}\right)^2}{8\rho \cdot \sigma^{(i+1)} \cdot \|\mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - x^{(i)}\|_2^2} \right),$$

where $\sigma^{(i+1)}$ is a subgaussian parameter for $z^{(i+1)} - z^{(i)}$, and a random variable is subgaussian with parameter σ if $\mathbb{P}\{|\cdot| \geq t\} \leq 2\exp(-\frac{t^2}{2\sigma})$ for $t \geq 0$.

Proof. First, observe that

$$\begin{aligned}
\left\|x^{(i)} - z^{(i+1)}\right\|_2^2 &= \left\|x^{(i)} - z^{(i)} - (z^{(i+1)} - z^{(i)})\right\|_2^2 \\
&= \left\|x^{(i)} - z^{(i)}\right\|_2^2 + \left\|z^{(i+1)} - z^{(i)}\right\|_2^2 - 2\langle x^{(i)} - z^{(i)}, z^{(i+1)} - z^{(i)}\rangle \\
&= \left\|x^{(i)} - z^{(i)}\right\|_2^2 + \left\|z^{(i+1)} - z^{(i)}\right\|_2^2 - 2\langle z^{(i+1)} - z^{(i)} + \rho(x^{(i)} - \mathcal{R}_{\theta^{(i+1)}}(x^{(i)})), z^{(i+1)} - z^{(i)}\rangle \quad \text{(S.7)} \\
&= \left\|x^{(i)} - z^{(i)}\right\|_2^2 - \left\|z^{(i+1)} - z^{(i)}\right\|_2^2 + 2\rho\langle \mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - x^{(i)}, z^{(i+1)} - z^{(i)}\rangle \quad \text{(S.8)}
\end{aligned}$$

where the inequality (S.7) holds by $x^{(i)} = \rho x^{(i)} - \rho\mathcal{R}_{\theta^{(i+1)}} + z^{(i+1)}$ via (Alg.1.1). We now obtain a probablistic bound for the third quantity in (S.8) via a concentration inequality. The concentration inequality on the sum of independent zero-mean subgaussian variables (e.g., [9, Thm. 7.27]) yields that for any $t^{(i+1)} \geq 0$

$$\mathbb{P}\left\{\langle \mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - x^{(i)}, z^{(i+1)} - z^{(i)}\rangle \geq t^{(i+1)}\right\} \leq \exp\left(-\frac{(t^{(i+1)})^2}{2\sigma^{(i+1)}\|\mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - x^{(i)}\|_2^2}\right) \quad \text{(S.9)}$$

where $\sigma^{(i+1)}$ is given as in Lemma S.4. Applying the result (S.9) with $t^{(i+1)} = \frac{1}{2\rho}(\|z^{(i+1)} - z^{(i)}\|_2^2 + \Delta^{(i+1)})$ to the bound (S.8) completes the proofs. □

Lemma S.4 implies that, given sufficiently large $\Delta^{(i+1)}$, or sufficiently small $\sigma^{(i+1)}$ (e.g., variance for a Gaussian random vector $z^{(i+1)} - z^{(i)}$) or $\|\mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - x^{(i)}\|_2^2$, bound (12) is satisfied with high probability, for each i. In particular, $\Delta^{(i+1)}$ can be large for the first several iterations; if paired operators $(\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}})$ in (Alg.1.1) map their input images to similar output images (e.g., the trained NNs $\mathcal{R}_{\theta^{(i+1)}}$ and $\mathcal{R}_{\theta^{(i)}}$ have good refining capabilities for $x^{(i)}$ and $x^{(i-1)}$), then $\sigma^{(i+1)}$ is small; if the regularization parameter γ in (Alg.1.3) is sufficiently large, then $\|\mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - x^{(i)}\|_2^2$ is small.

S.4 Proofs of Proposition 9

First, we show that $\sum_{i=0}^{\infty} \|x^{(i+1)} - x^{(i)}\|_2^2 < \infty$ for convex and nonconvex $F(x; y, z^{(i+1)})$ cases.

- *Convex $F(x; y, z^{(i+1)})$* case: Using Assumption 2 and $\{\widetilde{M}^{(i+1)} = M^{(i+1)} : \forall i\}$ for the convex case via (7), we obtain the following results for any $\mathcal{X}$:

$$\begin{aligned}
&F\left(x^{(i)}; y, z^{(i)}\right) - F\left(x^{(i+1)}; y, z^{(i+1)}\right) + \gamma\Delta^{(i+1)} \\
&\geq F\left(x^{(i)}; y, z^{(i+1)}\right) - F\left(x^{(i+1)}; y, z^{(i+1)}\right) \quad \text{(S.10)} \\
&\geq \frac{1}{2}\left\|x^{(i+1)} - \acute{x}^{(i+1)}\right\|_{M^{(i+1)}}^2 + \left(\acute{x}^{(i+1)} - x^{(i)}\right)^T M^{(i+1)}\left(x^{(i+1)} - \acute{x}^{(i+1)}\right) \quad \text{(S.11)} \\
&= \frac{1}{2}\left\|x^{(i+1)} - x^{(i)}\right\|_{M^{(i+1)}}^2 - \frac{1}{2}\left\|E^{(i+1)}\left(x^{(i)} - x^{(i-1)}\right)\right\|_{M^{(i+1)}}^2 \quad \text{(S.12)} \\
&\geq \frac{1}{2}\left\|x^{(i+1)} - x^{(i)}\right\|_{M^{(i+1)}}^2 - \frac{\delta^2}{2}\left\|x^{(i)} - x^{(i-1)}\right\|_{M^{(i)}}^2 \quad \text{(S.13)}
\end{aligned}$$

  where the inequality (S.10) uses the condition (12) in Assumption 4, the inequality (S.11) is obtained by using the results in [2, Lem. S.1], the equality (S.12) uses the extrapolation formula (Alg.1.2) and the symmetry of $M^{(i+1)}$, the inequality (S.13) holds by Assumption 3.
  Summing the inequality of $F(x^{(i)}; y, z^{(i)}) - F(x^{(i+1)}; y, z^{(i+1)}) + \gamma\Delta^{(i+1)}$ in (S.13) over $i = 0, \ldots, N_{\text{lyr}} - 1$, we obtain

$$\begin{aligned}
F\left(x^{(0)}; y, z^{(0)}\right) - F\left(x^{(N_{\text{lyr}})}; y, z^{(N_{\text{lyr}})}\right) + \gamma\sum_{i=0}^{N_{\text{lyr}}-1}\Delta^{(i+1)} &\geq \sum_{i=0}^{N_{\text{lyr}}-1}\frac{1}{2}\left\|x^{(i+1)} - x^{(i)}\right\|_{M^{(i+1)}}^2 - \frac{\delta^2}{2}\left\|x^{(i)} - x^{(i-1)}\right\|_{M^{(i)}}^2 \\
&\geq \sum_{i=0}^{N_{\text{lyr}}-1}\frac{1-\delta^2}{2}\left\|x^{(i+1)} - x^{(i)}\right\|_{M^{(i+1)}}^2
\end{aligned}$$

$$\geq \sum_{i=0}^{N_{\text{lyr}}-1} \frac{m_{F,\min}(1-\delta^2)}{2} \left\| x^{(i+1)} - x^{(i)} \right\|_2^2 \tag{S.14}$$

where the inequality (S.14) holds by Assumption 2. Due to the lower boundedness of $F(x; y, z)$ in Assumption 1 and the summability of $\{\Delta^{(i+1)} \geq 0 : \forall i\}$ in Assumption 4, taking $N_{\text{lyr}} \to \infty$ gives

$$\sum_{i=0}^{\infty} \left\| x^{(i+1)} - x^{(i)} \right\|_2^2 < \infty. \tag{S.15}$$

- *Nonconvex $F(x; y, z^{(i+1)})$ case*: Using Assumption 2, we obtain the following results without assuming that $F(x; y, z^{(i+1)})$ is convex:

$$\begin{aligned}
&F\Big(x^{(i)}; y, z^{(i)}\Big) - F\Big(x^{(i+1)}; y, z^{(i+1)}\Big) + \gamma\Delta^{(i+1)} \\
&\geq F\Big(x^{(i)}; y, z^{(i+1)}\Big) - F\Big(x^{(i+1)}; y, z^{(i+1)}\Big) && \text{(S.16)} \\
&\geq \frac{\lambda-1}{4} \left\| x^{(i+1)} - x^{(i)} \right\|_{M^{(i+1)}}^2 - \frac{(\lambda+1)^2}{\lambda-1} \left\| x^{(i)} - \acute{x}_b^{(i+1)} \right\|_{M^{(i+1)}}^2 && \text{(S.17)} \\
&= \frac{\lambda-1}{4} \left\| x^{(i+1)} - x^{(i)} \right\|_{M^{(i+1)}}^2 - \frac{(\lambda+1)^2}{\lambda-1} \left\| E^{(i+1)} \left(x^{(i)} - x^{(i-1)} \right) \right\|_{M^{(i+1)}}^2 && \text{(S.18)} \\
&\geq \frac{\lambda-1}{4} \left(\left\| x^{(i+1)} - x^{(i)} \right\|_{M^{(i+1)}}^2 - \delta^2 \left\| x^{(i)} - x^{(i-1)} \right\|_{M^{(i)}}^2 \right) && \text{(S.19)}
\end{aligned}$$

where the inequality (S.16) uses the condition (12) in Assumption 4, the inequality (S.17) use the results in [3, §S.3], the equality (S.18) holds by (Alg.1.2), the inequality (S.19) is obtained by Assumption 3.
Summing the inequality of $F(x^{(i)}; y, z^{(i)}) - F(x^{(i+1)}; y, z^{(i+1)}) + \gamma\Delta^{(i+1)}$ in (S.19) over $i = 0, \ldots, N_{\text{lyr}} - 1$, we obtain

$$\begin{aligned}
F\Big(x^{(0)}; y, z^{(0)}\Big) - F\Big(x^{(N_{\text{lyr}})}; y, z^{(N_{\text{lyr}})}\Big) + \gamma \cdot \sum_{i=0}^{N_{\text{lyr}}-1} \Delta^{(i+1)} &\geq \sum_{i=0}^{N_{\text{lyr}}-1} \frac{\lambda-1}{4} \left(\left\| x^{(i+1)} - x^{(i)} \right\|_{M^{(i+1)}}^2 - \delta^2 \left\| x^{(i)} - x^{(i-1)} \right\|_{M^{(i)}}^2 \right) \\
&\geq \sum_{i=0}^{N_{\text{lyr}}-1} \frac{(\lambda-1)(1-\delta^2)}{2} \left\| x^{(i+1)} - x^{(i)} \right\|_{M^{(i+1)}}^2 \\
&\geq \sum_{i=0}^{N_{\text{lyr}}-1} \frac{m_{F,\min}(\lambda-1)(1-\delta^2)}{2} \left\| x^{(i+1)} - x^{(i)} \right\|_2^2,
\end{aligned}$$

where we follow the arguments in obtaining (S.14) above. Again, using the lower boundedness of $F(x; y, z)$ and the summability of $\{\Delta^{(i+1)} \geq 0 : \forall i\}$, taking $N_{\text{lyr}} \to \infty$ gives the result (S.15) for nonconvex $F(x; y, z^{(i+1)})$.

Second, we show that $\sum_{i=0}^{\infty} \|z^{(i+1)} - z^{(i)}\|_2^2 < \infty$. Observe

$$\begin{aligned}
\left\| z^{(i+1)} - z^{(i)} \right\|_2^2 &= \left\| (1-\rho)\Big(x^{(i)} - x^{(i-1)} \Big) + \rho\Big(\mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - \mathcal{R}_{\theta^{(i)}}(x^{(i-1)}) \Big) \right\|_2^2 \\
&\leq (1-\rho) \left\| x^{(i)} - x^{(i-1)} \right\|_2^2 + \rho \left\| \mathcal{R}_{\theta^{(i+1)}}(x^{(i)}) - \mathcal{R}_{\theta^{(i)}}(x^{(i-1)}) \right\|_2^2 \\
&\leq \left\| x^{(i)} - x^{(i-1)} \right\|_2^2 + \rho\epsilon^{(i+1)}
\end{aligned} \tag{S.20}$$

where the first equality uses the image mapping formula in (Alg.1.1), the first inequality holds by applying Jensen's inequality to the (convex) squared ℓ^2-norm, the second inequality is obtained by using the asymptotically non-expansiveness of the paired operators $(\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}})$ in Assumption 4. Summing the inequality of $\|z^{(i+1)} - z^{(i)}\|_2^2$ in (S.20) over $i = 0, \ldots, N_{\text{lyr}} - 1$, we obtain

$$\sum_{i=0}^{N_{\text{lyr}}-1} \left\| z^{(i+1)} - z^{(i)} \right\|_2^2 \leq \sum_{i=0}^{N_{\text{lyr}}-2} \left\| x^{(i+1)} - x^{(i)} \right\|_2^2 + \rho \sum_{i=0}^{N_{\text{lyr}}-1} \epsilon^{(i+1)}, \tag{S.21}$$

where we used $x^{(0)} = x^{(-1)}$ as given in Algorithm 1. By taking $N_{\text{lyr}} \to \infty$ in (S.21), using result (S.15), and the summability of the sequence $\{\epsilon^{(i+1)} : i \geq 0\}$, we obtain

$$\sum_{i=0}^{\infty} \left\| z^{(i+1)} - z^{(i)} \right\|_2^2 < \infty. \tag{S.22}$$

Combining the results in (S.15) and (S.22) completes the proofs.

S.5 PROOFS OF THEOREM 10

Let $\bar{x}$ be a limit point of $\{x^{(i)} : i \geq 0\}$ and $\{x^{(i_j)}\}$ be the subsequence converging to $\bar{x}$. Let $\bar{z}$ be a limit point of $\{z^{(i)} : i \geq 0\}$ and $\{z^{(i_j)}\}$ be the subsequence converging to $\bar{z}$. The closedness of $\mathcal{X}$ implies that $\bar{x} \in \mathcal{X}$. Using the results in Proposition 9, $\{x^{(i_j+1)}\}$ and $\{z^{(i_j+1)}\}$ also converge to $\bar{x}$ and $\bar{z}$, respectively. Taking another subsequence if necessary, the subsequence $\{M^{(i_j+1)}\}$ converges to some $\bar{M}$, since $M^{(i+1)}$ is bounded by Assumption 2. The subsequences $\{\theta^{(i_j+1)}\}$ converge to some $\bar{\theta}$, since $x^{(i_j+1)} \to \bar{x}$, $z^{(i_j+1)} \to \bar{z}$, and $\{\theta^{(i+1)}\}$ is bounded via Assumption 4.

Next, we show that the convex proximal minimization (S.23) below is continuous in the sense that the output point $x^{(i_j+1)}$ continuously depends on the input points $\acute{x}^{(i_j+1)}$ and $z^{(i_j+1)}$, and majorization matrix $\widetilde{M}^{(i_j+1)}$:

$$\begin{aligned} x^{(i_j+1)} &= \underset{x\in\mathcal{X}}{\operatorname{argmin}} \langle \nabla F(\acute{x}^{(i_j+1)}; y, z^{(i_j+1)}), x - \acute{x}^{(i_j+1)} \rangle + \frac{1}{2}\Big\|x - \acute{x}^{(i_j+1)}\Big\|^2_{\widetilde{M}^{(i_j+1)}} \\ &= \operatorname{Prox}^{\widetilde{M}^{(i_j+1)}}_{\mathbb{I}_{\mathcal{X}}}\Big(\acute{x}^{(i_j+1)} - \big(\widetilde{M}^{(i_j+1)}\big)^{-1}\nabla F(\acute{x}^{(i_j+1)}; y, z^{(i_j+1)})\Big). \end{aligned} \tag{S.23}$$

where the proximal mapping operator $\operatorname{Prox}^{\widetilde{M}^{(i_j+1)}}_{\mathbb{I}_{\mathcal{X}}}(\cdot)$ is given as in (2). We consider the two cases of majorization matrices $\{M^{(i+1)}\}$ given in Theorem 10:

- For a sequence of diagonal majorization matrices, i.e., $\{M^{(i+1)} : i \geq 0\}$, one can obtain the continuity of the convex proximal minimization (S.23) with respect to $\acute{x}^{(i_j+1)}$, $z^{(i_j+1)}$, and $\widetilde{M}^{(i_j+1)}$, by extending the existing results in [10, Thm. 2.26], [11] with the separability of (S.23) to element-wise optimization problems.
- For a fixed general majorization matrix, i.e., $M = M^{(i+1)}, \forall i$, we obtain that the convex proximal minimization (S.24) below is continuous with respect to the input points $\acute{x}^{(i_j+1)}$ and $z^{(i_j+1)}$:

$$\begin{aligned} x^{(i_j+1)} &= \underset{x\in\mathcal{X}}{\operatorname{argmin}} \langle \nabla F(\acute{x}^{(i_j+1)}; y, z^{(i_j+1)}), x - \acute{x}^{(i_j+1)} \rangle + \frac{1}{2}\Big\|x - \acute{x}^{(i_j+1)}\Big\|^2_{\widetilde{M}} \qquad \text{(S.24)} \\ &= \operatorname{Prox}^{\widetilde{M}}_{\mathbb{I}_{\mathcal{X}}}\big(\acute{x}^{(i_j+1)} - \widetilde{M}^{-1}\nabla F(\acute{x}^{(i_j+1)}; y, z^{(i_j+1)})\big) \\ &= \big(\mathrm{Id} + \widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}}\big)^{-1}\big(\acute{x}^{(i_j+1)} - \widetilde{M}^{-1}\nabla F(\acute{x}^{(i_j+1)}; y, z^{(i_j+1)})\big) \qquad \text{(S.25)} \end{aligned}$$

  where $\hat{\partial} f(x)$ is the subdifferential of f at x and Id denotes the identity operator, and the proximal mapping of $\mathbb{I}_{\mathcal{X}}$ relative to $\|\cdot\|_{\widetilde{M}}$ is uniquely determined by the resolvent of the operator $\widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}}$ in (S.25).
  First, we obtain that the operator $\widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}}$ is monotone. For a convex extended-valued function $f_e : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$, observe that $\widetilde{M}^{-1}\hat{\partial} f_e$ is a monotone operator:

$$\langle \widetilde{M}^{-1}\hat{\partial} f_e(u) - \widetilde{M}^{-1}\hat{\partial} f_e(v), u - v\rangle = \langle \underbrace{\widetilde{M}^{-1}\widetilde{M}}_{=I}\hat{\partial} f_e(\widetilde{M}\tilde{u}) - \underbrace{\widetilde{M}^{-1}\widetilde{M}}_{=I}\hat{\partial} f_e(\widetilde{M}\tilde{v}), \widetilde{M}\tilde{u} - \widetilde{M}\tilde{v}\rangle \geq 0, \qquad \forall u, v, \tag{S.26}$$

  where the equality uses the variable change $\{u = \widetilde{M}\tilde{u}, v = \widetilde{M}\tilde{v}\}$, a chain rule of the subdifferential of a composition of a convex extended-valued function and an affine mapping [12, §7], and the symmetry of $\widetilde{M}$, and the inequality holds because the subdifferential of convex extended-valued function is a monotone operator [13, §4.2]. Because characteristic function of a convex set is extended-valued function, the result in (S.26) implies that the operator $\widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}}$ is monotone. Second, note that the resolvent of a monotone operator $\widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}}$ (with a parameter 1), i.e., $(\mathrm{Id}+\widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}})^{-1}$ in (S.25), is nonexpansive [10, §6] and thus continuous. We now obtain that the convex proximal minimization (S.24) is continuous with respect to the input points $\acute{x}^{(i_j+1)}$ and $z^{(i_j+1)}$, because the proximal mapping operator $(\mathrm{Id}+\widetilde{M}^{-1}\hat{\partial}\mathbb{I}_{\mathcal{X}})^{-1}$ in (S.25), the affine mapping $\widetilde{M}^{-1}$, and $\nabla F(x; y, z)$ are continuous with respect to their input points.

For the two cases above, using the fact that $x^{(i_j+1)} \to \bar{x}$, $\acute{x}^{(i_j+1)} \to \bar{x}$, $z^{(i_j+1)} \to \bar{z}$, and $M^{(i_j+1)} \to \bar{M}$ (or $\bar{M} = M$ for the $\{M^{(i+1)} = M\}$ case) as $j \to \infty$, (S.23) becomes

$$\bar{x} = \underset{x\in\mathcal{X}}{\operatorname{argmin}} \langle \nabla F(\bar{x}; y, \bar{z}), x - \bar{x}\rangle + \frac{1}{2}\|x - \bar{x}\|^2_{\bar{M}}. \tag{S.27}$$

Thus, $\bar{x}$ satisfies the first-order optimality condition of $\min_{x\in\mathcal{X}} F(x; y, \bar{z})$:

$$\langle \nabla F(\bar{x}; y, \bar{z}), x - \bar{x}\rangle \geq 0, \qquad \text{for any } x \in \mathcal{X},$$

and this completes the proof of the first result.

Next, note that the result in Proposition 9 imply

$$\left\| \mathcal{A}^{M^{(i+1)}}_{\mathcal{R}_{\theta^{(i+1)}}}\left(\begin{bmatrix} x^{(i)} \\ x^{(i-1)} \end{bmatrix}\right) - \begin{bmatrix} x^{(i)} \\ x^{(i-1)} \end{bmatrix} \right\|_2 \to 0. \tag{S.28}$$

Additionally, note that a function $\mathcal{A}^{M^{(i+1)}}_{\mathcal{R}_{\theta^{(i+1)}}} - I$ is continuous. To see this, observe that the convex proximal mapping in (Alg.1.3) is continuous (see the obtained results above), and $\mathcal{R}_{\theta^{(i+1)}}$ is continuous (see Assumption 4). Combining (S.28), the convergence of $\{M^{(i_j+1)}, \mathcal{R}_{\theta^{(i_j+1)}}\}$, and the continuity of $\mathcal{A}^{M^{(i+1)}}_{\mathcal{R}_{\theta^{(i+1)}}} - I$, we obtain $[\bar{x}^T, \bar{x}^T]^T = \mathcal{A}^{\bar{M}}_{\mathcal{R}_{\bar{\theta}}}([\bar{x}^T, \bar{x}^T]^T)$, and this completes the proofs of the second result.

S.6 PROOFS OF COROLLARY 11

To prove the first result, we use proof by contradiction. Suppose that $\mathrm{dist}(x^{(i)}, \mathcal{S}) \nrightarrow 0$. Then there exists $\epsilon > 0$ and a subsequence $\{x^{(i_j)}\}$ such that $\mathrm{dist}(x^{(i_j)}, \mathcal{S}) \geq \epsilon$, $\forall j$. However, the boundedness assumption of $\{x^{(i_j)}\}$ in Corollary 11 implies that there must exist a limit point $\bar{x} \in \mathcal{S}$ via Theorem 10. This is a contradiction, and gives the first result (via the result in Proposition 9). Under the isolation point assumption in Corollary 11, using the obtained results, $\|x^{(i+1)} - x^{(i)}\|_2 \to 0$ (via Proposition 9) and $\mathrm{dist}(x^{(i+1)}, \mathcal{S}) \to 0$, and the following the proofs in [4, Cor. 2.4], we obtain the second result.

S.7 MOMENTUM-NET VS. BCD-NET

This section compares the convergence properties of Momentum-Net (Algorithm 1) and BCD-Net (Algorithm 2). We first show that for convex $f(x; y)$ and $\mathcal{X}$, the sequence of reconstructed images generated by BCD-Net converges:

Proposition S.5 (Sequence convergence)**.** *In Algorithm 2, let $f(x; y)$ be convex and subdifferentiable, and $\mathcal{X}$ be convex. Assume that the paired operators $(\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}})$ are asymptotically contractive, i.e.,*

$$\|\mathcal{R}_{\theta^{(i+1)}}(u) - \mathcal{R}_{\theta^{(i)}}(v)\|_2 < \|u - v\|_2 + \epsilon^{(i+1)},$$

with $\sum_{i=0}^{\infty} \epsilon^{(i+1)} < \infty$ and $\{\epsilon^{(i+i)} \in [0, \infty) : \forall i\}$, $\forall u, v, i$. Then, the sequence $\{x^{(i+1)} : i \geq 0\}$ generated by Algorithm 2 is convergent.

Proof. We rewrite the updates in Algorithm 2 as follows:

$$\begin{aligned} x^{(i+1)} &= \underset{x \in \mathcal{X}}{\mathrm{argmin}} \; f(x; y) + \frac{\gamma}{2}\Big\|x - \mathcal{R}_{\theta^{(i+1)}}(x^{(i)})\Big\|_2^2 = \mathrm{Prox}^{\gamma I}_{f + \mathbb{I}_{\mathcal{X}}}\big(\mathcal{R}_{\theta^{(i+1)}}(x^{(i)})\big) \\ &= \big(\mathrm{Id} + \gamma^{-1}\hat{\partial}(f(x; y) + \mathbb{I}_{\mathcal{X}})\big)^{-1}\big(\mathcal{R}_{\theta^{(i+1)}}(x^{(i)})\big) \\ &=: \mathcal{A}^{(i+1)}(x^{(i)}). \end{aligned}$$

We first show that the paired operators $\{\mathcal{A}^{(i+1)}, \mathcal{A}^{(i)}\}$ is asymptotically contractive:

$$\begin{aligned} &\Big\|\mathcal{A}^{(i+1)}(u) - \mathcal{A}^{(i)}(v)\Big\|_2 \\ &= \Big\|\big(\mathrm{Id} + \gamma^{-1}\hat{\partial}(f(x; y) + \mathbb{I}_{\mathcal{X}})\big)^{-1}\big(\mathcal{R}_{\theta^{(i+1)}}(u)\big) - \big(\mathrm{Id} + \gamma^{-1}\hat{\partial}(f(x; y) + \mathbb{I}_{\mathcal{X}})\big)^{-1}\big(\mathcal{R}_{\theta^{(i)}}(v)\big)\Big\|_2 \\ &\leq \|\mathcal{R}_{\theta^{(i+1)}}(u) - \mathcal{R}_{\theta^{(i)}}(v)\|_2 \qquad \text{(S.29)} \\ &\leq L'\|u - v\|_2 + \epsilon^{(i+1)}\|u - v\|_2, \qquad \text{(S.30)} \end{aligned}$$

$\forall u, v$, where the inequality (S.29) holds because the subdifferential of the convex extended-valued function $f(x; y) + \mathbb{I}_{\mathcal{X}}$ (the characteristic function of a convex set $\mathcal{X}$, $\mathbb{I}_{\mathcal{X}}$, is convex, and the sum of the two convex functions, $f(x; y) + \mathbb{I}_{\mathcal{X}}$, is convex) is a monotone operator [13, §4.2], and the resolvent of a monotone relation with a positive parameter, i.e., $(\mathrm{Id} + \gamma^{-1}\hat{\partial}(f(x; y) + \mathbb{I}_{\mathcal{X}}))^{-1}$ with $\gamma^{-1} > 0$, is nonexpansive [13, §6], and the inequality (S.30) holds by $L' < 1$ via the contractiveness of the paired operators $(\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}})$, $\forall i$. Note that the inequality (S.29) does not hold for nonconvex $f(x; y)$ and/or $\mathcal{X}$. Considering that $L' < 1$, we show that the sequence $\{x^{(i+1)} : i \geq 0\}$ is Cauchy sequence:

$$\begin{aligned} \Big\|x^{(i+l)} - x^{(i)}\Big\|_2 &= \Big\|(x^{(i+l)} - x^{(i+l-1)}) + \ldots + (x^{(i+1)} - x^{(i)})\Big\|_2 \\ &\leq \Big\|x^{(i+l)} - x^{(i+l-1)}\Big\| + \ldots + \Big\|x^{(i+1)} - x^{(i)}\Big\|_2 \\ &\leq \big(L'^{l-1} + \ldots + 1\big)\Big\|x^{(i+1)} - x^{(i)}\Big\|_2 + \big(\epsilon^{(i+l)} + \ldots + \epsilon^{(i+1)}\big) \\ &\leq \frac{1}{1 - L'}\Big\|x^{(i+1)} - x^{(i)}\Big\|_2 + \sum_{i'=1}^{l} \epsilon^{(i+i')} \end{aligned}$$

where the second inequality uses the result in (S.30). Since the sequence $\{x^{(i+1)} : i \geq 0\}$ is Cauchy sequence, $\{x^{(i+1)} : i \geq 0\}$ is convergent, and this completes the proofs. □

In terms of guaranteeing convergence, BCD-Net has three theoretical or practical limitations compared to Momentum-Net:

- Different from Momentum-Net, BCD-Net assumes the asymptotic contractive condition for the paired operators $\{\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}}\}$. When image mapping operators in (Alg.2.1) are identical across iterations, i.e., $\{\mathcal{R}_\theta = \mathcal{R}_{\theta^{(i+1)}} : i \geq 0\}$, then $\mathcal{R}_\theta$ is assumed to be contractive. On the other hand, a mapping operator (identical across iterations) of Momentum-Net only needs to be nonexpansive. Note, however, that when $f(x; y) = \frac{1}{2}\|y - Ax\|_W^2$ with $A^H W A \succ 0$ (e.g., Example 5), BCD-Net can guarantee the sequence convergence with the asymptotically nonexpansive paired operators $(\mathcal{R}_{\theta^{(i+1)}}, \mathcal{R}_{\theta^{(i)}})$ (see Definition 7) [14].
- When one applies an iterative solver to (Alg.2.2), there always exist some numerical errors and these obstruct the sequence convergence guarantee in Proposition S.5. To guarantee sufficiently small numerical errors from iterative

methods solving (Alg.2.2) (so that one can find a critical point solution for the MBIR problem (Alg.2.2)), one needs to use sufficiently many inner iterations that can substantially slow down entire MBIR.

- BCD-Net does not guarantee the sequence convergence for nonconvex data-fit $f(x;y)$, whereas Momentum-Net guarantees convergence to a fixed-point for both convex $f(x;y)$ and nonconvex $f(x;y)$.

S.8 For the sCNN architecture (18), connection between convolutional training loss (P2) and its patch-based training loss

This section shows that given the sCNN architecture (18), the convolutional training loss in (P2) has three advantages over the patch-based training loss in [14], [15] that may use all the extracted overlapping patches of size R:

- The corresponding patch-based loss does not model the patch aggregation process that is inherently modeled in (18).
- It is an upper bound of the convolutional loss (P2).
- It requires about R times more memory than (P2).

We prove the benefits of (P2) using the following lemma.

Lemma S.6. *The loss function (P2) for training the residual convolutional autoencoder in (18) is bounded by the patch-based loss function:*

$$\frac{1}{2L}\sum_{s=1}^{S}\Big\|\hat{x}_s^{(i)}-\sum_{k=1}^{K}d_k\circledast\mathcal{T}_{\alpha_k}(e_k\circledast x_s^{(i)})\Big\|_2^2\le\frac{1}{2LR}\sum_{s=1}^{S}\big\|\widehat{X}_s^{(i)}-D\mathcal{T}_{\tilde{\alpha}}(EX_s^{(i)})\big\|_{\mathrm{F}}^2, \tag{S.31}$$

where the residual is defined by $\hat{x}_s^{(i)}\triangleq x_s-x_s^{(i)}$, $\{x_s,x_s^{(i)}\}$ are given as in (P2), $\widehat{X}_s\in\mathbb{R}^{R\times V_s}$ and $X_s\in\mathbb{R}^{R\times V_s}$ are the lth training data matrices whose columns are V_s vectorized patches extracted from the images $\hat{x}_s$ and x_s (with the circulant boundary condition and the "stride" parameter 1), respectively, $D\triangleq[d_1,\dots,d_K]\in\mathbb{C}^{R\times K}$ is a decoding filter matrix, and $E\triangleq[e_1^*,\dots,e_K^*]^H\in\mathbb{C}^{K\times R}$ is an encoding filter matrix. Here, the definition of soft-thresholding operator in (6) is generalized by

$$(\mathcal{T}_{\tilde{\alpha}}(u))_k\triangleq\begin{cases}u_k-\alpha_k\cdot\mathrm{sign}(u_k), & |u_k|>\alpha_k,\\ 0, & \textit{otherwise},\end{cases} \tag{S.32}$$

for $K=1,\dots,K$, where $\tilde{\alpha}=[\alpha_1,\dots,\alpha_K]^T$. See other related notations in (18).

Proof. First, we have the following reformulation [3, §S.1]:

$$\begin{bmatrix}e_1*u\\ \vdots\\ e_K*u\end{bmatrix}=\underbrace{P'\begin{bmatrix}EP_1\\ \vdots\\ EP_N\end{bmatrix}}_{\triangleq\widetilde{E}}u,\quad\forall u, \tag{S.33}$$

where $P'\in\mathbb{C}^{KN\times KN}$ is a permutation matrix, E is defined in Lemma (S.6), and $P_n\in\mathbb{C}^{R\times N}$ is the nth patch extraction operator for $n=1,\dots,N$. Considering that $\sum_{k=1}^{K}\mathsf{flip}(e_k^*)\circledast(e_k\circledast u)=\frac{1}{R}\widetilde{E}^H\widetilde{E}u$ via the definition of $\widetilde{E}$ in (S.33) (see also the reformulation technique in [3, §S.1]), we obtain the following reformulation result:

$$\sum_{k=1}^{K}\mathsf{flip}(e_k^*)\circledast\mathcal{T}_{\alpha_k}(e_k\circledast x_s^{(i)})=\frac{1}{R}\sum_{n=1}^{N}P_n^HE^H\mathcal{T}_{\tilde{\alpha}}\big(EP_nx_s^{(i)}\big) \tag{S.34}$$

where the soft-thresholding operators $\{\mathcal{T}_{\alpha_k}(\cdot):\forall k\}$ and $\mathcal{T}_{\tilde{\alpha}}(\cdot)$ are defined in (S.32) and we use the permutation invariance of the thresholding operator $\mathcal{T}_\alpha(\cdot)$, i.e., $\mathcal{T}_\alpha(P(\cdot))=P\mathcal{T}_\alpha(\cdot)$ for any α. Finally, we obtain the result in (S.31) as follows:

$$\begin{aligned}\frac{1}{2L}\sum_{s=1}^{S}\Big\|\hat{x}_s^{(i)}-\sum_{k=1}^{K}d_k\circledast\mathcal{T}_{\alpha_k}(e_k\circledast x_s^{(i)})\Big\|_2^2&=\frac{1}{2L}\sum_{s=1}^{S}\Big\|\hat{x}_s^{(i)}-\frac{1}{R}\sum_{n=1}^{N}P_n^HD\mathcal{T}_{\tilde{\alpha}}\big(EP_nx_s^{(i)}\big)\Big\|_2^2 &\text{(S.35)}\\ &=\frac{1}{2L}\sum_{s=1}^{S}\Big\|\frac{1}{R}\sum_{n=1}^{N}P_n^HP_n\hat{x}_s^{(i)}-\frac{1}{R}\sum_{n=1}^{N}P_n^HD\mathcal{T}_{\tilde{\alpha}}\big(EP_nx_s^{(i)}\big)\Big\|_2^2 &\text{(S.36)}\\ &=\frac{1}{2LR^2}\sum_{s=1}^{S}\Big\|\sum_{n=1}^{N}P_n^H\Big(\hat{x}_{l,n}^{(i)}-D\mathcal{T}_{\tilde{\alpha}}\big(Ex_{l,n}^{(i)}\big)\Big)\Big\|_2^2\\ &\le\frac{1}{2LR}\sum_{s=1}^{S}\sum_{n=1}^{N}\Big\|\hat{x}_{l,n}^{(i)}-D\mathcal{T}_{\tilde{\alpha}}\big(Ex_{l,n}^{(i)}\big)\Big\|_2^2 &\text{(S.37)}\\ &=\frac{1}{2LR}\sum_{s=1}^{S}\Big\|\widehat{X}_s^{(i)}-D\mathcal{T}_{\tilde{\alpha}}\Big(EX_s^{(i)}\Big)\Big\|_{\mathrm{F}}^2,\end{aligned}$$

where D is defined in Lemma S.6, $\{\hat{x}_{l,n}^{(i)} = P_n \hat{x}_s^{(i)} \in \mathbb{C}^R, x_{l,n}^{(i)} = P_n x_s^{(i)} \in \mathbb{C}^R : n = 1, \dots, N\}$ is a set of extracted patches, the training matrices $\{\widehat{X}_s^{(i)}, X_s^{(i)}\}$ are defined by $\widehat{X}_s^{(i)} \triangleq [\hat{x}_{l,n}^{(i)}, \dots, \hat{x}_{l,N}^{(i)}]$ and $X_s^{(i)} \triangleq [x_{l,1}^{(i)}, \dots, x_{l,N}^{(i)}]$. Here, the equality (S.35) uses the result in (S.34), the equality (S.36) holds by $\sum_{n=1}^N P_n^H P_n = R \cdot I$ (for the circulant boundary condition in Lemma S.6), and the inequality (S.37) holds by $\widetilde{P}\widetilde{P}^H \preceq R \cdot I$ with $\widetilde{P} \triangleq [P_1^H \cdots P_N^H]^H$. □

Lemma S.6 reveals that when the patch-based training approach extract all the R-size overlapping patches, *1)* the corresponding patch-based loss is an upper bound of the convolutional loss (P2); *2)* it requires about R-times larger memory than (P2) because $V_s \approx RN_s$ for $x \in \mathbb{R}^{N_s}$ and the boundary condition described in Lemma S.6, $\forall l$; and *3)* it misses modeling the patch aggregation process that is inherently modeled in (18) – see that the patch aggregation operator $\sum_{n=1}^N P_n^H(\cdot)_n$ is removed in the inequality (S.37) in the proof of Lemma S.6. In addition, different from the patch-based training approach [14], [15], i.e., training with the function on the right-hand side in (S.31), one can use different sizes of filters $\{e_k, d_k : \forall k\}$ in the convolutional training loss, i.e., the function on the left-hand side in (S.31).

S.9 Details of experimental setup

S.9.1 Majorization matrix designs for quadratic data-fit

For (real-valued) quadratic data-fit $f(x; y)$ in the form of $\frac{1}{2}\|y - Ax\|_W^2$, if a majorization matrix M exists such that $A^H W A \preceq M$, it is straightforward to verify that the gradient of quadratic data-fit $f(x; y)$ satisfies the M-Lipchitz continuity in Definition 1, i.e.,

$$\|\nabla f(u; y) - \nabla f(v; y)\|_{M^{-1}} = \|A^H W A u - A^H W A v\|_{M^{-1}} \leq \|u - v\|_M^2, \quad \forall u, v \in \mathbb{R}^N.$$

because the assumption $A^T W A \preceq M \Leftrightarrow M^{-1/2} A^T W A M^{-1/2} \preceq I$ implies that the eigenspectrum of $M^{-1/2} A^T W A M^{-1/2}$ lies in the interval $[0, 1]$, and gives the following result:

$$\left(M^{-1/2} A^T W A M^{-1/2}\right)^2 \preceq I \Leftrightarrow (A^T W A) M^{-1} (A^T W A) \preceq M.$$

Next, we review a useful lemma in designing majorization matrices for a wide class of quadratic data-fit $f(x; y)$:

Lemma S.7 ([2, Lem. S.3]). *For a (possibly complex-valued) matrix A and a diagonal matrix W with non-negative entries, $A^H W A \preceq \mathrm{diag}(|A^H||W||A|1)$, where $|A|$ denotes the matrix consisting of the absolute values of the elements of A.*

S.9.2 Parameters for MBIR optimization models: Sparse-view CT reconstruction

For MBIR model using EP regularization, we combined a EP regularizer $\sum_{n=1}^N \sum_{n' \in \mathcal{N}_n} \iota_n \iota_{n'} \varphi(x_n - x_{n'})$ and the data-fit $f(x; y)$ in §4.1.1, where $\mathcal{N}_n$ is the set of indices of the neighborhood, ι_n and $\iota_{n'}$ are parameters that encourage uniform noise [16], and $\varphi(\cdot)$ is the Lange penalty function, i.e., $\varphi(t) = \delta^2(|t/\delta| - \log(1 + |t/\delta|))$, with $\delta = 10$ in HU. We chose the regularization parameter (e.g., γ in (P0)) as $2^{15.5}$. We ran the relaxed linearized augmented Lagrangian method [17] with 100 iterations and 12 ordered-subsets, and initialized the EP MBIR algorithms with a conventional FBP method using a Hanning window.

For MBIR model using a learned convolutional regularizer [6, (P2)], we trained convolutional regularizer with filters of $\{h_k \in \mathbb{R}^R : R = K = 7^2\}$ via CAOL [3] in an unsupervised training manner; see training details in [3]. The regularization parameters (e.g., γ in (1)) were selected by applying the "spectral spread" based selection scheme in §3.2 with the tuned factor $\chi^\star = 167.64$. We selected the spatial-strength-controlling hard-thresholding parameter (i.e., α' in [6, (P2)]) as follows: for Test samples #1–2, we chose it is as 10^{-10} and 6^{-11}, respectively. We initialized the MBIR model using a learned regularizer with the EP MBIR results obtained above. We terminated the iterations if the relative error stopping criterion (e.g., [2, (44)]) is met before reaching the maximum number of iterations. We set the tolerance value as 10^{-13} and the maximum number of iterations to 4×10^3.

S.9.3 Parameters for MBIR optimization models: LF photography using a focal stack

For MBIR model using 4D EP regularization [18], we combined a 4D EP regularizer $\sum_{n=1}^N \sum_{n' \in \mathcal{N}_n} \varphi(x_n - x_{n'})$ and the data-fit $f(x; y)$ in §4.1.2, where $\mathcal{N}_n$ is the set of indices of the 4D neighborhood, and $\varphi(\cdot)$ is the hyperbola penalty function, i.e., $\varphi(t) = \delta^2(\sqrt{1 + |t/\delta|^2} - 1)$. We selected the hyperbola function parameter δ and regularization parameter (e.g., γ in (P0)) as follows: for Test samples #1–3, we chose them as $\{\delta = 10^{-4}, \gamma = 10^3\}$, $\{\delta = 10^{-1}, \gamma = 10^7\}$, and $\{\delta = 10^{-1}, \gamma = 5 \times 10^3\}$, respectively. We ran the conjugate gradient method with 100 iterations, and initialized the 4D EP MBIR algorithms with $A^T y$ rescaled in the interval $[0, 1]$.

S.9.4 Reconstruction accuracy and depth estimation accuracy of different MBIR methods

Tables S.1–S.3 below provide reconstruction accuracy numerics of different MBIR methods in sparse-view CT reconstruction and LF photography using a focal stack, and reports the SPO depth estimation [19] accuracy numerics on reconstructed LFs from different MBIR methods:

TABLE S.1
RMSE (HU) of different CT MBIR methods
(fan-beam geometry with 12.5% projections views and 10^5 incident photons)

	(a) FBP	(b) EP reg.	(c) Learned convolutional reg. [3], [6]	(d) Momentum-Net-sCNN	(e) **Momentum-Net-sCNN** w/ larger width	(f) Momentum-Net-dCNN
Test #1	82.8	40.8	35.2	19.9	**19.5**	19.8
Test #2	74.9	38.5	34.5	18.4	**17.7**	17.8

(c)'s convolutional regularizer uses $\{R=K=7^2\}$
(d)'s refining sCNNs are in the form of residual single-hidden layer convolutional autoencoder (18) with $\{R=K=7^2\}$.
(e)'s refining sCNNs are in the form of residual single-hidden layer convolutional autoencoder (18) with $\{R=7^2, K=9^2\}$.
This setup gives results in Fig. 8(d), as described in §4.2.1.
(f)'s refining dCNNs are in the form of residual multi-hidden layer CNN (19) with $\{L=4, R=3^2, K=64\}$.

TABLE S.2
PSNR (dB) of different LF MBIR methods
(LF photography systems with $C=5$ detectors obtain a focal stack of LFs consisting of $S=81$ sub-aperture images)

	(a) A^Ty	(b) 4D EP reg. [18]	(c) Momentum-Net-sCNN	(d) **Momentum-Net-dCNN**
Test #1	16.4	32.0	35.8	**37.1**
Test #2	21.1	28.1	30.7	**32.0**
Test #3	21.6	28.1	30.9	**31.7**

(c)'s refining sCNNs are in the form of residual single-hidden layer convolutional autoencoder with $\{R=5^2, K=32\}$.
(d)'s refining dCNNs are in the form of residual multi-hidden layer CNN (19) with $\{L=6, R=3^2, K=16\}$.
Momentum-Nets use refining CNNs in an epipolar-domain; see details in §4.2.1.

TABLE S.3
RMSE (in 10^{-2}, m) of estimated depth from reconstructed LFs with different LF MBIR methods
(LF photography systems with $C=5$ detectors obtain a focal stack of LFs consisting of $S=81$ sub-aperture images)

	(a) Ground truth LF	(b) Reconstructed LF by A^Ty	(c) Reconstructed LF by 4D EP reg. [18]	(d) Reconstructed LF by Momentum-Net-sCNN	(e) Reconstructed LF by **Momentum-Net-dCNN**
Test #1	4.7	41.0	13.8	8.0	**5.7**
Test #2	30.5	117.6	39.5	34.6	**31.9**
Test #3	n/a†	n/a†	n/a†	n/a†	n/a†

SPO depth estimation [19] was applied to reconstructed LFs.
(d)'s refining sCNNs are in the form of residual single-hidden layer convolutional autoencoder with $\{R=5^2, K=32\}$.
(e)'s refining dCNNs are in the form of residual multi-hidden layer CNN (19) with $\{L=6, R=3^2, K=16\}$.
Momentum-Nets use refining CNNs in an epipolar-domain; see details in §4.2.1.
†The ground truth depth map for Test sample #3 does not exist in the LF dataset [20].

S.9.5 Reconstructed images and estimated depths with noniterative analytical methods

This section provides reconstructed images by an analytical back-projection method in sparse-view CT reconstruction and LF photography using a focal stack (see the first two columns in Fig. S.2), and estimated depths from reconstructed LFs via the SPO depth estimation method [19] (see the third column in Fig. S.2(c)). Results in Fig. S.2 below are supplementary to Fig. 8, Fig. 9, and Fig. 10, and the first two columns visualize initial input images to INN methods.

S.10 How to choose parameters of image refining modules in soft-refining INNs?

In soft-refining INNs using iterative-wise refining NNs, one does not need to greatly increase parameter dimensions of refining NNs [14], [21]. The natural question then arises, "How one can choose between sCNN (18) and dCNN (19) refiners, and select their parameters (R, K, and L)?" The first answer to this question depends on some understanding of data-fit $f(x;y)$ in MBIR problem (P1), e.g., the regularization strength γ and the condition number variations across training data-fit majorizers. (An additional criteria could be general understandings between sample size/diversity and parameter dimension of NNs.)

For example, the sparse-view CT system in §4.1.1 needs moderate regularization strength ($\chi^\star = 167.64$) and the majorization matrices of its training data-fits have mild condition number variations (the standard deviation is 1.1). training data-fits have mild parameter variations across samples. Comparing results between Momentum-Net-sCNN and -dCNN in Fig. 5 and Table S.1 demonstrates that sCNN (18) seems suffice. Table S.1(d)–(e) shows that one can further improve the refining accuracy of sCNN (18) by increasing its width, i.e., K. The LF photography system using a limited focal stack in §4.1.2 needs a large γ value ($\chi^\star = 1.5$), and the majorization matrices of its training data-fits have large condition number variations (the standard deviation is 2245.5). Comparing results between Momentum-Net-sCNN and -dCNN in Fig. 7 and Table S.2 demonstrates that dCNN (19) yields higher PSNR than sCNN (18). For dCNN (19), we observed increasing its depth, i.e., L, up to a certain number is more effective than increasing its width, i.e., K, as briefly discussed in §4.2.1.

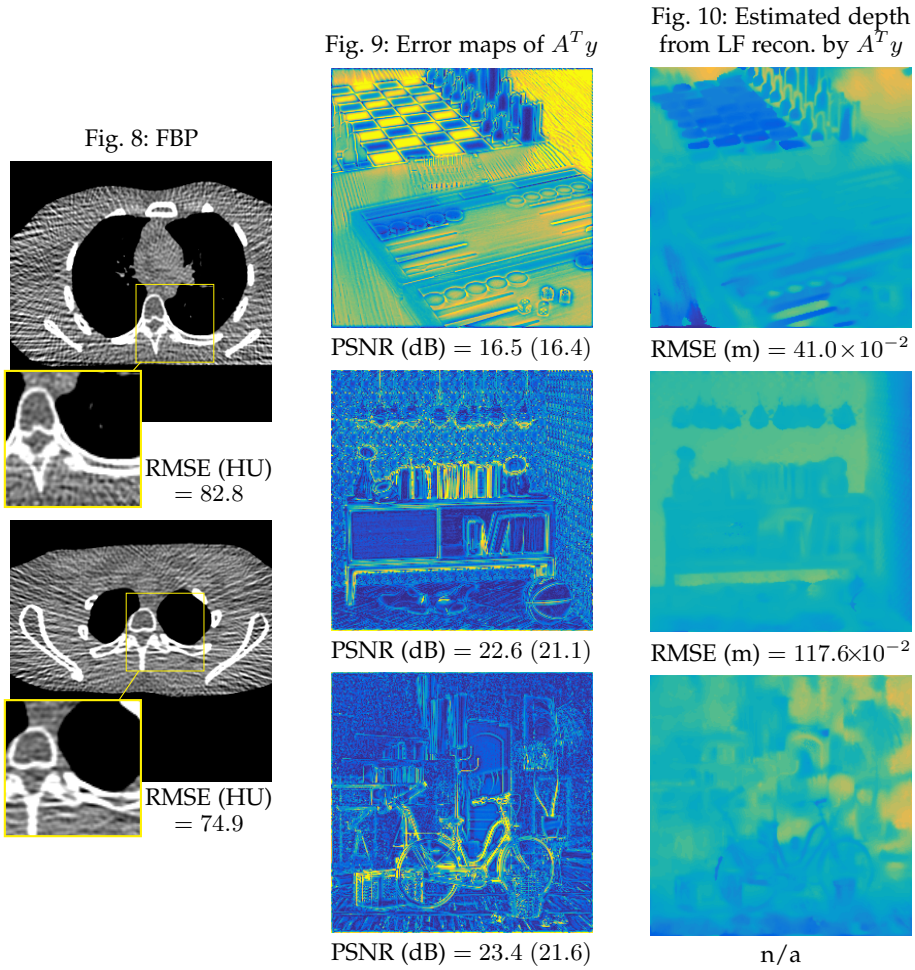

Fig. S.2. Reconstructed images from analytical back-projection methods. We used such results in the first two columns to initialize INN methods.

For choosing the relaxation parameter ρ in (Alg.1.1), we also suggest considering the regularization strength in (Alg.1.3). For an application that needs moderate regularization strength, e.g., sparse-view CT in §4.1.1, we suggest setting ρ to 0.5 so as to mix information between input and output of refining NNs, rather than $1-\varepsilon$ that does not mix input and output. For an application that needs strong regularization, e.g., LF photography using a limited focal stack in §4.1.2, we suggest using $\rho = 1-\varepsilon$ than $\rho = 0.5$. Results in the next section validate this suggestion.

Performance of Momentum-Net with different relaxation parameters ρ in (Alg.1.1)

Fig. S.3 below compares the performances of Momentum-Net-sCNN with different ρ values. The results in Fig. S.3 support the ρ selection guideline in §4.2.3. One can maximize the MBIR accuracy of Momentum-Net by properly selecting ρ.

Note that $\rho \in (0, 1)$ controls strength of inference from refining NNs in (Alg.1.1), but does not affect the convergence guarantee of Momentum-Net. Fig. S.3 illustrates that Momentum-Net appears to converge regardless of ρ values.

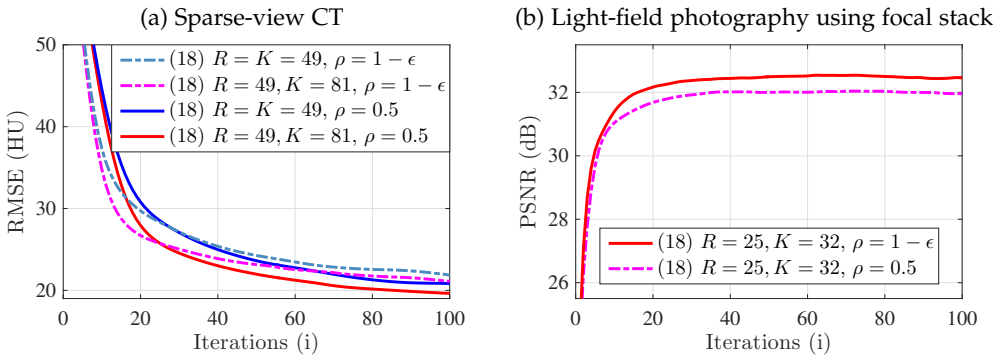

Fig. S.3. Convergence behavior of Momentum-Net-sCNN with different relaxation parameters, $\rho = 0.5$ and $\rho = 1 - \varepsilon$. For both applications (see their imaging setups in §4.1), PyTorch ver. 0.3.1 was used.

S.11 PARAMETERS OF MOMENTUM-NET

Table S.4 below lists parameters of Momentum-Net, and summarizes selection guidelines or default values. Similar to BCD-Net/ADMM-Net, the main tuning jobs to maximize the performance of Momentum-Net include selecting architectures of refining NNs $\{\mathcal{R}_{\theta^{(i)}} : \forall i\}$ in (Alg.1.1), and choosing a regularization parameter γ in (Alg.1.3) by tuning χ in §3.2. One can simplify the tuning process by using the selection guidelines in §S.10 for selecting architectures of $\{\mathcal{R}_{\theta^{(i)}} : \forall i\}$, and training χ in §3.2. Note that one designs majorization matrices $\{M^{(i)} : \forall i\}$ rather than tuning them: majorization matrices can be analytically designed, e.g., Lemma S.7 as used in §4.2.1; one can algorithmically design them [22]. Tighter majorization matrices are expected to further accelerate the convergence of Momentum-Net [2], [3].

TABLE S.4
Guidelines for choosing parameters of Momentum-Net

Param.	Module	Guidelines or default values
$\{\mathcal{R}_{\theta^{(i)}} : \forall i\}$	(Alg.1.1)	Trainable by §3.1. For selecting their architecture/param., see guideline §S.10.
$\rho \in (0, 1)$	(Alg.1.1)	Use regularization strength γ; see guideline in §S.10.
$\delta < 1$ in (8)–(9)	(Alg.1.2)	$1-\varepsilon$
$\{M^{(i)} : \forall i\}$	(Alg.1.3)	Designed off-line. For large-scale inverse problems with quadratic data-fit, use Lemma S.7.
$\lambda \geq 1$ in (7)	(Alg.1.3)	For convex $F(x; y, z^{(i+1)})$, $\lambda = 1$; for nonconvex $F(x; y, z^{(i+1)})$, $\lambda = 1+\varepsilon$.
$\gamma > 0$	(Alg.1.3)	Chosen by tuning/training χ in §3.2

All INN methods also must select a number of INN iterations, N_{iter}. One could determine it by using the convergence behavior of iteration-wise refiners in Fig. 2.

REFERENCES

[1] I. Y. Chun, Z. Huang, H. Lim, and J. A. Fessler, "Momentum-Net: Fast and convergent iterative neural network for inverse problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/TPAMI.2020.3012955.
[2] I. Y. Chun and J. A. Fessler, "Convolutional dictionary learning: Acceleration and convergence," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1697–1712, Apr. 2018.
[3] ——, "Convolutional analysis operator learning: Acceleration and convergence," *IEEE Trans. Image Process.*, vol. 29, pp. 2108–2122, 2020.
[4] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imaging Sci.*, vol. 6, no. 3, pp. 1758–1789, Sep. 2013.
[5] ——, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *J. Sci. Comput.*, vol. 72, no. 2, pp. 700–734, Aug. 2017.
[6] I. Y. Chun and J. A. Fessler, "Convolutional analysis operator learning: Application to sparse-view CT," in *Proc. Asilomar Conf. on Signals, Syst., and Comput.*, Pacific Grove, CA, Oct. 2018, pp. 1631–1635.
[7] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Program.*, vol. 140, no. 1, pp. 125–161, Aug. 2013.
[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.
[9] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. New York, NY: Springer, 2013.
[10] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Berlin: Springer Verlag, 2009, vol. 317.
[11] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optm.*, vol. 14, no. 5, pp. 877–898, Aug. 1976.
[12] B. S. Mordukhovich and N. M. Nam, "Geometric approach to convex subdifferential calculus," *Optimization*, vol. 66, no. 6, pp. 839–873, 2017.
[13] E. K. Ryu and S. Boyd, "Primer on monotone operator methods," *Appl. Comput. Math*, vol. 15, no. 1, pp. 3–43, Jan. 2016.
[14] I. Y. Chun, X. Zheng, Y. Long, and J. A. Fessler, "BCD-Net for low- dose CT reconstruction: Acceleration, convergence, and generalization," in *Proc. Med. Image Computing and Computer Assist. Interven.*, Shenzhen, China, Oct. 2019, pp. 31–40.
[15] I. Y. Chun and J. A. Fessler, "Deep BCD-net using identical encoding-decoding CNN structures for iterative image recovery," in *Proc. IEEE IVMSP Workshop*, Zagori, Greece, Jun. 2018, pp. 1–5.
[16] J. H. Cho and J. A. Fessler, "Regularization designs for uniform spatial resolution and noise properties in statistical image reconstruction for 3-D X-ray CT," *IEEE Trans. Med. Imag.*, vol. 34, no. 2, pp. 678–689, Feb. 2015.
[17] H. Nien and J. A. Fessler, "Relaxed linearized algorithms for faster X-ray CT image reconstruction," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 1090–1098, Apr. 2016.
[18] M.-B. Lien, C.-H. Liu, I. Y. Chun, S. Ravishankar, H. Nien, M. Zhou, J. A. Fessler, Z. Zhong, and T. B. Norris, "Ranging and light field imaging with transparent photodetectors," *Nature Photonics*, vol. 14, no. 3, pp. 143–148, Jan. 2020.
[19] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Und.*, vol. 145, pp. 148–159, Apr. 2016.
[20] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. ACCV*, Taipei, Taiwan, Nov. 2016, pp. 19–34.
[21] H. Lim, I. Y. Chun, Y. K. Dewaraja, and J. A. Fessler, "Improved low-count quantitative PET reconstruction with an iterative neural network," *IEEE Trans. Med. Imag.*, early access, May 2020, doi: 10.1109/TMI.2020.2998480.
[22] M. G. McGaffin and J. A. Fessler, "Algorithmic design of majorizers for large-scale inverse problems," *ArXiv preprint stat.CO/1508.02958*, Oct. 2015.