Supplementary Material
Parametric Majorization for Data-Driven Energy Minimization Methods

Jonas Geiping Michael Moeller
Department of Electrical Engineering and Computer Science, University of Siegen
{jonas.geiping, michael.moeller}@uni-siegen.de

This document contains the appendix for the submission 'Parametric Majorization for Data-Driven Energy Minimization Methods'. It contains proofs that were omitted from the main paper and further details on the experimental setups. We will occasionally repeat equations from the main paper, but the equation numbering will always be identical. For the exact implementation of the proposed experiments, refer to project's github page at https://github.com/JonasGeiping/ParametricMajorization.

1. Convex Analysis in Section 3

1.1. Details for Derivation of Eqs. (11), (12)

Eq. (11) in the main paper describes the application of Bregman duality:

$$D^0_{E_\theta}(x_i^*, x_i(\theta)) = D^{x_i^*}_{E_\theta^*}(0, q_i) \quad q_i \in \partial E(x_i^*, y_i, \theta), \quad (11)$$

which is a common application of the following identity [6, 5]:

Lemma 1 (Bregman Identity)**.** *Consider a convex lsc. function $E : \mathbb{R}^n \to \mathbb{R}$ with a subgradient $p \in \partial E(y)$. Then, the following identity holds:*

$$D^p_E(x, y) = D^x_{E^*}(p, q), \quad q \in \partial E(x)$$

Proof. This property follows from equality (Fenchel's identity) in the Fenchel-Young inequality $E(x) + E^*(p) = \langle p, x\rangle \iff p \in \partial E(x)$. To see this we write

$$D^p_E(x, y) = E(x) - \langle p, x\rangle - E(y) + \langle p, y\rangle$$

and apply Fenchel's identity for p, y to find

$$D^p_E(x, y) = E(x) - \langle p, x\rangle + E^*(p)$$

We then introduce any $q \in \partial E(x)$ by writing $\langle p, x\rangle = \langle p - q + q, x\rangle$ and apply Fenchel's identity again:

$$D^p_E(x, y) = E^*(p) - E^*(q) - \langle x, p - q\rangle = D^x_{E^*}(p, q)$$

□

The step from Eq. (11) to Eq.(12) is simply the first step of this derivation:

$$\begin{aligned} D_{E_\theta}(x_i^*, x_i(\theta)) &= E(x_i^*, y_i, \theta) - \langle 0, x_i^*\rangle + E^*(0, y_i, \theta) \\ =D^{x_i^*}_{E_\theta^*}(0, q_i) &= E(x_i^*, y_i, \theta) + E^*(0, y_i, \theta) \end{aligned} \quad (12)$$

as $p_i = 0$ is a subgradient of E at $x_i(\theta)$ and q_i at x_i^*.

1.2. Details for Derivation of Eq. (14) -> (15)

A crucial subtlety of Lemma 1 is that this identity holds for any $q \in \partial E(x)$ and the choice of subgradients is irrelevant, the Bregman distance is equal for all choices. This motivates the introduction of the W-function $W_E(p, x) = E^*(p) + E(x) - \langle p, x\rangle$. This function is convex in either p or x and always non-negative. It can be understood as measuring the deviation of p from subgradients of x as a direct implementation of the Fenchel-Young inequality. As such it is 0 exactly if $p \in \partial E(x)$. Previous usage of this function can be found for example in [7, 18]. For Legendre functions [1], i.e. functions where both E and E^* are (essentially) smooth, the connection to Bregman distances is immediate:

$$W_E(p, x) = D^p_E(x, \nabla E^*(p)),$$

for non-smooth functions this is also a part of the proof of Lemma 1, replacing $\nabla E^*(p)$ by $y \in \partial E^*(p)$. As such, we can write Eq. (12) as

$$D^{x_i^*}_{E^*}(0, q_i) = W_{E_\theta}(0, x_i^*). \quad (12)$$

The introduction of this function then allows us to show that

$$W_E(0, x_i^*) = \min_z W_{E_1,\theta}(-z, x_i^*) + W_{E_2,\theta}(z, x_i^*) \quad (15)$$

under the assumption in Eq.(13), that E can be written as $E_1 + E_2$, with both functions convex. We recognize this as the clear extension of the infimal convolution property $E^*(0) = \min_z E_1^*(-z) + E_2^*(z)$ (which itself can be understood as Fenchel's duality theorem applied to E_1, E_2) to

these functions, in the smooth setting this could be written via

$$D^{x_i^*}_{E^*}(0, \nabla E(x_i^*)) = \min_z \; D_{E_1^*}(-z, \nabla E_1(x_i^*)) + D_{E_2^*}(z, \nabla E_2^*(x_i^*)).$$

We arrive at Eq. (15) from Eq. (14) by rewriting E in Eq.(14):

$$\begin{aligned} &\min_z E_1(x_i^*, y_i, \theta) + E_2(x_i^*, y_i, \theta) \\ &+ E_1^*(-z, y_i, \theta) + E_2^*(z, y_i, \theta) \\ =&\min_z E_1(x_i^*, y_i, \theta) + E_2(x_i^*, y_i, \theta) + \langle z, x_i^* \rangle \\ &+ E_1^*(-z, y_i, \theta) + E_2^*(z, y_i, \theta) - \langle z, x_i^* \rangle \\ =&\min_z W_{E_1,\theta}(-z, x_i^*) + W_{E_2,\theta}(z, x_i^*). \end{aligned} \tag{14}$$

(15)

1.3. Proof of Proposition 2

Proposition 2 (Ordering of parametric majorizers)**.** *Assuming the condition $l(x, z) \leq D_{E_\theta}(x, z)$ from Eq. (8), we find that the presented parametric majorizers can be ordered in the following way:*

$$\begin{aligned} l(x_i^*, x(\theta)) &\leq D^0_{E_\theta}(x_i^*, x_i(\theta)) = D^{x_i^*}_{E_\theta^*}(0, q_i) \\ &\leq \min_{z \in \partial E_2(x_i^*)} W_{E_1}(-z, x_i^*) \\ &\leq \frac{1}{m(\theta, y)} ||q_i||^2 \quad \text{s.t. } q_i \in \partial E(x_i^*, y, \theta). \end{aligned}$$

The Bregman surrogate (10) majorizes the original loss function and is in turn majorized by the partial surrogate (16) which is majorized by the gradient penalty (17) under the assumption of $m(\theta, y)$ - strong convexity of E_1.

Proof. The first inequality follows directly by the assumption $l(x, z) \leq D_{E_\theta}(x, z)$. The second inequality is the application of Bregman Duality discussed in Lemma 1. From Eq.(15) we now see that $D^{x_i^*}_{E_\theta^*}(0, q_i)$, $q_i \in \partial E(x_i^*, y_i, \theta)$ can be written as a minimum over z. Clearly choosing a non-optimal z yields an upper bound to this minimal value. Without loss of generality, we choose $z \in \partial E_2(x_i^*)$ so that $W_{E_2,\theta}(z, x_i^*)$ is equal to zero.

Now we assume that E is $m(\theta, y)$-strongly convex. We subsume this strong convexity term in E_1 again without loss of generality so that E_1 is strongly convex. By convex duality [2], this implies that E_1^* is $m(\theta, y)$ strongly smooth, i.e. $D^x_{E_1^*}(p, q) \leq \frac{1}{2m(\theta,y)}||p - q||^2$. Following Eq.(12), we write

$$\begin{aligned} W_{E_1^*}(-z, x_i^*) &= D^{x_i^*}_{E_1^*}(-z, r) \quad z \in \partial E_2(x_i^*, y_i, \theta), \\ &\qquad\qquad\qquad\quad r \in \partial E_1(x_i^*, y_i, \theta) \\ &\leq \frac{1}{2m(\theta, y)}|| - z - r||^2 \\ &= \frac{1}{2m(\theta, y)}||q_i||^2 \qquad\qquad q_i \in \partial E(x_i^*, y_i, \theta), \end{aligned}$$

under mild assumptions on the additivity of subgradients of E_1 and E_2. □

1.4. Derivation of the surrogate functions for the example in subsection 3.3

Section 3.3 discusses the non-smooth bi-level problem given in Eqs. (18) and (19):

$$\min_{\theta \in \mathbb{R}} \frac{1}{2}|x^* - x(\theta)|^2, \tag{18}$$

$$\text{subject to} \qquad x(\theta) = \arg\min_x \frac{1}{2}|x - y|^2 + \theta|x|. \tag{19}$$

for both $x^*, y \in \mathbb{R}$. In this setting, the 'primal' formulation of the Bregman surrogate is given by

$$\min_\theta \max_x \frac{1}{2}|x^* - y|^2 - \frac{1}{2}|x - y|^2 + \theta\left(|x^*| - |x|\right) \quad \text{(10 ex.)}$$

whereas the 'dual' formulation is given by

$$\min_\theta \min_{|z| \leq \theta} \frac{1}{2}|x^* - y|^2 + \theta|x^*| + \frac{1}{2}|z - y|^2. \quad \text{(12 ex.)}$$

Note that this problem is convex in z, θ as the epigraph constraint $|z| \leq \theta$ is convex. Both (equivalent!) variants are visualized in Figure 1. We see that the saddle-point of the primal formulation and the minimizer of the dual formulation correctly coincide with the optimal θ.

Moving forward, we set $E_1(x, y) = \frac{1}{2}|x - y|^2$ and $E_2(x, \theta) = \theta|x|$ to compute the two partial surrogates. Firstly $W_{E_1,\theta}(-z, x^*), z \in \partial E_2(x^*)$ leads to

$$\min_\theta \frac{1}{2}|x^* - y + q|^2, \quad q \in \partial|x^*|, \tag{16 ex.1}$$

where we take $q = \text{sign}(x^*)$ as $x^* \neq 0$ in our example. As E_1 is a quadratic function, this is also equivalent to the gradient penalty in Eq. (17). The second partial surrogate, $W_{E_2,\theta}(z, x^*), z \in \partial E_1(x^*)$ can be written as

$$\begin{aligned} &\min_\theta \theta|x^*| + I_{|\cdot| \leq \theta}(x^* - y) - \langle x^*, x^* - y \rangle \quad \text{(16 ex.2)} \\ =&\min_{|x^* - y| \leq \theta} \theta|x^*| + C. \end{aligned}$$

Figure 1 here and Figure 1 in the main paper both arise from the data point $x^* = 0.3, y = 1.5$.

To give some more details on the fact that the Bregman surrogate is exactly identical with the original loss function in the vicinity of the optimal value, note that this is caused by the special structure of the Bregman distance of the absolute value, $D_{|\cdot|}(x, y)$ as $D_{E_\theta}(x, y)$ decomposes into $\frac{1}{2}|x - y|^2 + \theta D_{|\cdot|}(x, y)$. This function is equal to the higher-level loss function as soon as the signs of x^* and $x(\theta)$ coincide and as such the majorizer is exact, even if it is much easier to compute.

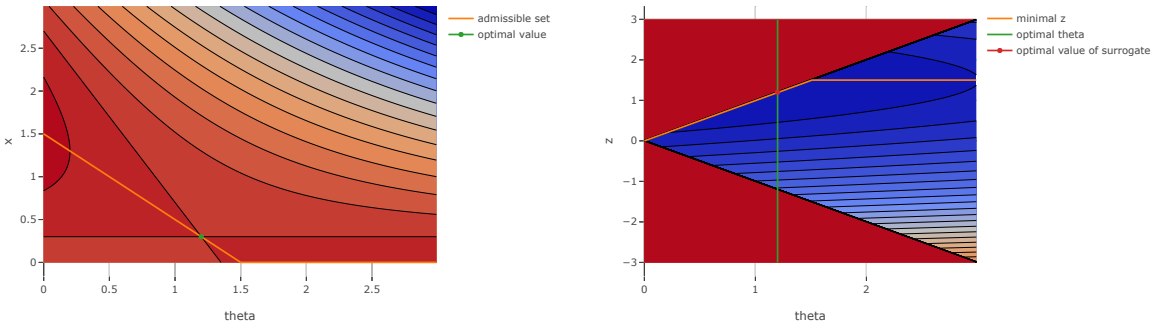


Figure 1. Visualization of the Bregman surrogate problem in primal formulation (left) and dual formulation (right). The problem in visualized over all (x, θ), respectively (z, θ). The admissible $x(\theta)$ are marked in orange in the left contour plot and the optimal $z(\theta)$ one the right. The optimal value in θ is marked in green in both plots.

1.5. Proof of Proposition 4

Section 3.4 describes an iterative procedure for repeated application of the majorization strategies discussed in section 3.2. This scheme was based on the result of Proposition 3:

$$l(x, y) \leq l(x, z) + \langle \nabla_z l(x, z), y - z \rangle + D_E(z, y), \quad (20)$$

inserting $x = x_i^*, y = x_i(\theta), z = x_i(\theta^k)$ leads to

$$l(x_i^*, x_i(\theta)) \leq l(x_i^*, x_i(\theta^k)) + D_{E_\theta}(x_i(\theta^k), x_i(\theta)) + \langle \nabla l(x_i^*, x_i(\theta^k)), x_i(\theta) - x_i(\theta^k) \rangle. \quad (20b)$$

Eq.(20), respectively (20b), lead to a monotone descent of the higher-level loss, as shown in Proposition 4:

Proposition 4 (Descent Lemma)**.** *The iterative procedure given by*

$$\theta^{k+1} = \arg\min_\theta \sum_{i=1}^N l(x_i^*, x_i(\theta^k)) + \langle \nabla l(x_i^*, x_i(\theta^k)), x_i(\theta) - x_i(\theta^k) \rangle + D^0_{E_\theta}(x_i(\theta^k), x_i(\theta))$$

is guaranteed to be stable, i.e. not to increase the bi-level loss:

$$\sum_{i=1}^N l\left(x_i^*, x_i(\theta^{k+1})\right) \leq \sum_{i=1}^N l\left(x_i^*, x_i(\theta^k)\right) \quad (23)$$

Proof of Proposition 4. θ^{k+1} is a minimizer of the iterative scheme. Therefore, evaluating the iteration at θ^{k+1} leads to a lower value than evaluating at θ^k:

$$\begin{aligned}
&\sum_{i=1}^N l(x_i^*, x_i(\theta^k)) + \langle \nabla l(x_i^*, x_i(\theta^k)), x_i(\theta^{k+1}) - x_i(\theta^k) \rangle \\
&\quad + D^0_{E_{\theta^{k+1}}}(x_i(\theta^k), x_i(\theta^{k+1})) \\
\leq &\sum_{i=1}^N l(x_i^*, x_i(\theta^k)) + \langle \nabla l(x_i^*, x_i(\theta^k)), x_i(\theta^k) - x_i(\theta^k) \rangle \\
&\quad + D^0_{E_{\theta^k}}(x_i(\theta^k), x_i(\theta^k)) \\
= &\sum_{i=1}^N l(x_i^*, x_i(\theta^k))
\end{aligned}$$

Now the left-hand-side is also equivalent to Eq. (20b) evaluated at θ^{k+1}. Applying the inequality in (20b) for all $i = 1, \ldots, N$ we find

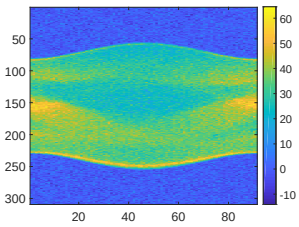
$$\sum_{i=1}^N l(x_i^*, x_i(\theta^{k+1})) \leq \sum_{i=1}^N l(x_i^*, x_i(\theta^k)).$$

□

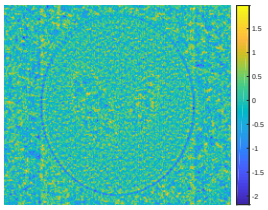
Remark. The iterative scheme given in Eq.(22), i.e.

$$\theta^{k+1} = \arg\min_\theta \sum_{i=1}^N E^*\left(\nabla l(x_i^*, x_i(\theta^k)), y_i, \theta\right) + E\left(x(\theta^k), y_i, \theta\right). \quad (22)$$

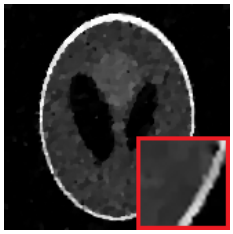
is an over-approximation of the iterative scheme discussed in Proposition 4. As such we expect the results of Proposition 4 to hold only approximately as stated in the main paper.



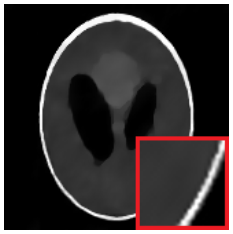
Noisy sinogram y



correction term $\mathcal{N}(\theta, y_i)$



Huber-TV, PSNR 23.9



Learned correction, PSNR 25.8

Figure 2. Illustrate our results for learning a linear correction term for a Huber-reguralized CT reconstruction problem. In reference to Figure 2 in the main paper we also visualize input data and the learned linear correction map. The predicted linear correction term can be visualized and inspected, and its influence can easily be quantified or explicitly scaled via a parameter.

2. Experimental Setup

This section will add additional details to the experiments presented in the paper[1].

2.1. CT - Additional Details

The implementation of the CT example in section 4.1 is straightforward. We generate pairs (y_i^*, x_i^*) of noisy sinograms and ground truth images and optimize

$$\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \|A^* A x_i^* - A^* y_i + \beta \nabla R(x_i^*) + \mathcal{N}(\theta, y_i)\|_2^2.$$

We test our model on the widely-used Shepp-Logan phantom, comparing the learned model with a pure Huber-TV solution, for which we found the optimal parameter β by grid search. This setup was implemented in Matlab. To visualize the linear correction term, we repeat an extended version of Figure 2.

2.2. Segmentation - Additional Details

The segmentation experiment shown in Figure 3 of the main paper shows the results of training the variational model in Eq.(25), which corresponds to an augmented cross-entropy term, as discussed in section 4.2.

The partial surrogate implemented in Figure 3 is a direct application of Eq.(16) to the segmentation setting, giving

$$\min_{\theta} \sum_{i=1}^{N} \min_{p_i \in \partial ||D x_i^*||} D_h \left(x_i^*, \nabla h^* \left(\mathcal{N}(\theta, y_i) - D^T p_i \right) \right),$$

where the computation of the auxiliary variable p_i is simplified. Note further that the gradient penalty cannot be applied in this setting, as the segmentation energy E is not strongly convex. Similarly, the iterative approach can be computed to be

$$\min_{\theta} \sum_{i=1}^{N} \min_{||p_i|| \leq 1} h^* \left(\frac{x_i^*}{x_i(\theta^k)} + \mathcal{N}(\theta, y_i) - D^T p_i \right) - \left\langle \mathcal{N}(\theta, y_i), x_i(\theta^k) \right\rangle$$

which is still convex in $\mathcal{N}(\theta, y)$, but the input arguments now take previous solutions into account.

To emphasize the convexity of the setup, we choose $\mathcal{N}(\theta, y_i)$ as a linear convolutional network of $3x3x3$ filters for each target class. We accordingly optimize the resulting convex minimization problems by an optimal convex optimization method, namely FISTA [4]. To solve the inference problem in Eq. (25) we apply usual strategies and optimize via a primal-dual algorithm [8] - to increase the speed we adapt a recent variant [9] and consider the Bregman-Proximal operator in the primal sub-problem for which we use the entropy function h described in the paper, paralleling [3, 16].

We draw four images and their corresponding segmentations from the `cityscapes` data set [12] and implement the proposed procedures in PyTorch [17]. For Figure 3 we drew the first four images, which we resized to 128x256 pixels. To visualize the improvement over the iterations, we initialize the subsequent iterations of the iterative scheme again with the initial value of θ, so that the training accuracy curves in Figure 3 are comparable. This is of course not strictly necessary and θ could be initialized with the current estimate in every iteration. We also point out that we visualize the actual training accuracy in Figure 3, meaning the percentage of successfully segmented pixels after *hard argmax* of the results of the algorithms.

2.3. Analysis Operators - Additional Details

For this experiment we considered the task of learning an 'analysis operator' $D(\theta)$, i.e. a set of convolutional filters θ^k so that $D(\theta) = \sum_{k=1}^{K} \theta_k * x$ for a set of K filters. Due to anisotropy, we can write the resulting minimization problem as

$$x(\theta) = \arg\min_x \frac{1}{2} ||x - y||^2 + \sum_{k=1}^{K} ||\theta_k * x||_1.$$

We repeat the experimental setup of [11] and train this model on image pairs x^*, y of noise-free and noisy image patches, to learn filters that result in a convex denoising

[1] Refer also to the implementations hosted on https://github.com/JonasGeiping/ParametricMajorization

model [10, 11]. To do so we draw a batch of 200 $64x64$ image patches from the training set of the Berkeley Segmentation data set [15], convert the images to gray-scale and add Gaussian noise. To compare with [11] and [20] we do not clip the noisy images and use Matlab's `rgb2gray` routine to generate this data. Further, as in [11], we do not optimize directly for the convolutional filters, but instead decompose each filter into a DCT-II basis, where we learn the weight of each basis function, excluding the constant basis function [13]. Before training we initialize these weights by orthogonal initialization [19] with a factor of 0.01, respectively 0.001 for the larger 9x9 filters.

To solve the training problem we minimize Eq. (33) in the paper jointly in $\theta, \{p_i\}_{i=1}^N$. We do this efficiently by taking steps toward the optimal weights with the 'Adam' optimization procedure [14] with a step size $\tau = 0.1$ (although gradient descent with momentum or FISTA [4] are also valid options). We use a standard accelerated primal-dual algorithm [8] to solve the convex inference problem. For the iterative procedure we repeat this process, computing $x(\theta^k)$ after every minimization of Eq.(33), inserting it as a factor into E^* and repeating the optimization. If the iterative procedure increases the loss value, we reduce the step size τ of the majorizing problem and repeat the step. If reducing the step size does not successfully improve the result for several iterations, we terminate the algorithm.

We implement this setup in PyTorch [17] and refer to our reference implementation for further details.

For total variation denoising, which corresponds to choosing $D(\theta)$ as the gradient operator with appropriate scaling, $\alpha\nabla$, we use grid search to find the optimal scaling parameter α.

We report execution times for a single minimization of Eq.(33) for different filter sizes in Table 1 in the paper as well as total time for an iterative procedure. These timings are reported for a single *GeForce RTX 2080Ti* graphics card.

References

[1] Heinz H Bauschke and Jonathan (Jon Borwein. Legendre Functions and the Method of Random Bregman Projections. *Journal of Convex Analysis*, 4(1):27–67, May 1997. 1

[2] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, New York, NY, 2011. 2

[3] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, May 2003. 4

[4] Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, Jan. 2009. 4, 5

[5] Martin Benning and Martin Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, May 2018. 1

[6] Martin Burger. Bregman Distances in Inverse Problems and Partial Differential Equations. In *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications, pages 3–33. Springer International Publishing, Cham, 2016. 1

[7] Dan Butnariu and Gabor Kassay. A Proximal-Projection Method for Finding Zeros of Set-Valued Operators. *SIAM J. Control Optim.*, 47(4):2096–2136, Jan. 2008. 1

[8] Antonin Chambolle and Thomas Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J Math Imaging Vis*, 40(1):120–145, May 2011. 4, 5

[9] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, Sept. 2016. 4

[10] Yunjin Chen, Rene Ranftl, and Thomas Pock. A bi-level view of inpainting - based image compression. In *Computer Vision Winter Workshop*. ., 2014. 5

[11] Yunjin Chen, René Ranftl, and Thomas Pock. Insights Into Analysis Operator Learning: From Patch-Based Sparse Models to Higher Order MRFs. *IEEE Transactions on Image Processing*, 23(3):1060–1072, Mar. 2014. 4, 5

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 4

[13] Jinggang Huang and D. Mumford. Statistics of natural images and models. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 541–547 Vol. 1, June 1999. 5

[14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, May 2015. 5

[15] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proceedings of 8th International Conference on Computer Vision*, volume 2, pages 416–423, July 2001. 5

[16] Peter Ochs, René Ranftl, Thomas Brox, and Thomas Pock. Techniques for Gradient-Based Bilevel Optimization with Non-smooth Lower Level Problems. *J Math Imaging Vis*, 56(2):175–194, Oct. 2016. 4

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS 2017 Autodiff Workshop*, Long Beach, CA, 2017. 4, 5

[18] Simeon Reich and Shoham Sabach. Existence and Approximation of Fixed Points of Bregman Firmly Nonexpansive Mappings in Reflexive Banach Spaces. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer Optimization and Its Applications, pages 301–316. Springer, New York, NY, 2011. 1

[19] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120 [cond-mat, q-bio, stat]*, Dec. 2013. 5

[20] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, July 2017. 5