Score-Matching by Denoising

Edward T. Reehorst and P. Schniter
Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43202, USA.

***Abstract*—Recently, a family of plug-and-play image recovery algorithms were proposed by Romano, Elad, and Milanfar under the name of "Regularization by Denoising" (RED). The RED algorithms were originally described as minimizing an optimization objective with an explicit regularization term. It was later shown, however, that this interpretation holds only when the denoiser exhibits both Jacobian symmetry and local homogeneity, which is not the case for practical denoisers like non-local means, BM3D, TNRD, and DnCNN. To explain these RED algorithms, we propose a new framework called Score-Matching by Denoising (SMD). We show tight connections between SMD, Parzen windowing, and approximate minimum mean-squared error denoising. Also, we propose new RED/SMD algorithms with fast convergence and guaranteed convergence to a fixed point.**

I. Introduction

Consider the problem of recovering image $\boldsymbol{x}^0 \in \mathbb{R}^N$ from measurements $\boldsymbol{y}$ that are noisy and linearly (or non-linearly) transformed. Recently, Romano, Elad, and Milanfar [1] proposed a family of algorithms that seek a solution $\widehat{\boldsymbol{x}}$ to

$$\mathbf{0} = \nabla \ell(\widehat{\boldsymbol{x}}; \boldsymbol{y}) + \lambda(\widehat{\boldsymbol{x}} - \boldsymbol{f}(\widehat{\boldsymbol{x}})), \tag{1}$$

where $\ell(\cdot; \boldsymbol{y})$ is a differentiable loss function, $\boldsymbol{f}(\cdot)$ is an image denoiser like BM3D [2] or DnCNN [3], and $\lambda > 0$ is a constant. They called this approach *regularization by denoising* (RED) because they claimed that the solutions to (1) obey

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \big\{ \ell(\boldsymbol{x}; \boldsymbol{y}) + \lambda \rho(\boldsymbol{x}) \big\} \tag{2}$$

with an explicit regularizer $\rho(\cdot)$ of the form

$$\rho_{\mathsf{red}}(\boldsymbol{x}) = \tfrac{1}{2} \boldsymbol{x}^\top \big(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) \big). \tag{3}$$

It was shown in [4], however, that (2)-(3) match (1) only when $\boldsymbol{f}(\cdot)$ is both Jacobian symmetric and locally homogeneous, which is not the typical case in practice. Still, experiments show that the RED algorithms usually give excellent performance. So, there remains the question of how to interpret them.

II. Score Matching by Denoising

From the Bayesian perspective, $\widehat{\boldsymbol{x}}$ in (2) is the MAP estimate when $\ell(\boldsymbol{x}; \boldsymbol{y})$ is the negative log-likelihood and $\lambda \rho(\boldsymbol{x})$ is the negative log prior. The likelihood is often straightforward to choose, but what about the prior? Suppose we have access to a large corpus of training data $\{\boldsymbol{x}_t\}_{t=1}^T$, from which we build the empirical pdf

$$\widehat{p}(\boldsymbol{x}) \triangleq \tfrac{1}{T} \textstyle\sum_{t=1}^T \delta(\boldsymbol{x} - \boldsymbol{x}_t). \tag{4}$$

Now say we smooth $\widehat{p}$ via Parzen windowing to build the image prior

$$p_\nu(\boldsymbol{x}) \triangleq \tfrac{1}{T} \textstyle\sum_{t=1}^T \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_t, \nu \boldsymbol{I}) \tag{5}$$

with appropriately chosen $\nu > 0$. In this case, Tweedie's formula [5] says that $\nabla \ln p_\nu(\boldsymbol{x})$, known as the *score* of p_ν, takes the form

$$\nabla \ln p_\nu(\boldsymbol{x}) = \big(\boldsymbol{f}_{\widehat{p},\nu}(\boldsymbol{x}) - \boldsymbol{x} \big) / \nu, \tag{6}$$

where $\boldsymbol{f}_{\widehat{p},\nu}(\boldsymbol{r})$ is the MMSE estimator of $\boldsymbol{x} \sim \widehat{p}$ from the noisy measurement $\boldsymbol{r} = \boldsymbol{x} + \mathcal{N}(\mathbf{0}, \nu \boldsymbol{I})$. If we use the smoothed prior p_ν for MAP estimation, then, from (2) and (6), $\widehat{\boldsymbol{x}}$ must obey

$$\mathbf{0} = \nabla \ell(\widehat{\boldsymbol{x}}; \boldsymbol{y}) + \lambda \big(\widehat{\boldsymbol{x}} - \boldsymbol{f}_{\widehat{p},\nu}(\widehat{\boldsymbol{x}}) \big) \text{ for } \lambda = 1/\nu. \tag{7}$$

Since (7) recovers the RED fixed-point equation (1), it explains RED with MMSE denoisers like $\boldsymbol{f}_{\widehat{p},\nu}(\cdot)$. But what about other $\boldsymbol{f}(\cdot)$?

Above, we established that RED results from MAP estimation through Tweedie's formula with MMSE denoising, i.e.,

$$\boldsymbol{f}_{\widehat{p},\nu} = \arg\min_{\boldsymbol{f}} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \big\| \boldsymbol{x}_t - \boldsymbol{f}\big(\boldsymbol{x}_t + \mathcal{N}(\mathbf{0}, \nu \boldsymbol{I})\big) \big\|^2. \tag{8}$$

But, since $\boldsymbol{f}_{\widehat{p},\nu}(\cdot)$ is difficult to implement, it is usually approximated by some computationally efficient $\boldsymbol{f}(\cdot)$, which may not be Jacobian symmetric nor locally homogeneous. However, noting from (6) that

$$\big\| \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}_{\widehat{p},\nu}(\boldsymbol{x}) \big\|^2 \propto \big\| (\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}))/\nu - \nabla \ln p_\nu(\boldsymbol{x}) \big\|^2, \tag{9}$$

we see that denoiser approximation is equivalent to *score matching*, as defined in [6]. In summary, the fixed-point equation (1) results from approximating MAP estimation under a smooth prior p_ν, where the approximation is performed on the score $\nabla \ln p_\nu$. Thus, we claim that the algorithms solving (1) are actually performing *score-matching by denoising* (SMD), not regularization by denoising (RED).

III. Convergence and Acceleration

When $\boldsymbol{f}(\cdot)$ has a non-symmetric Jacobian, the right side of (1) is not the gradient of any cost function. So, for algorithms that solve (1), we consider convergence not to a cost minimizer but rather to a fixed point. Consider iterating, for $k = 1, 2, 3, \ldots$, the lines

$$\boldsymbol{x}_k = \arg\min_{\boldsymbol{x}} \big\{ \ell(\boldsymbol{x}; \boldsymbol{y}) + \tfrac{\lambda L}{2} \|\boldsymbol{x} - \boldsymbol{v}_k\|^2 \big\} \tag{10a}$$

$$\boldsymbol{z}_k = \boldsymbol{x}_k + \tfrac{t_{k-1}-1}{t_k} (\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \tag{10b}$$

$$\boldsymbol{v}_k = \tfrac{1}{L} \boldsymbol{f}(\boldsymbol{z}_k) - \tfrac{1-L}{L} \boldsymbol{z}_k, \tag{10c}$$

with $L > 1$. In the full paper [4], the authors prove that, when $t_k = 1 \; \forall k$ (i.e., the unaccelerated case), the iteration (10) converges to a fixed point when $\ell(\cdot; \boldsymbol{y})$ is convex and $\boldsymbol{f}(\cdot)$ is non-expansive.

Meanwhile, with the Nesterov-like acceleration

$$t_k = \frac{1 + \sqrt{1 + 4 t_{k-1}^2}}{2}, \tag{11}$$

although there is no convergence proof, numerical experiments in [4] show (10) converging about 3 times as fast as the "fixed point" algorithm from [1], which is the fastest algorithm proposed in [1].

References

[1] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[3] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[4] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," *arXiv:1806.02296*, 2018.

[5] B. Efron, "Tweedie's formula and selection bias," *J. Am. Statist. Assoc.*, vol. 106, no. 496, pp. 1602–1614, 2011.

[6] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, pp. 695–709, 2005.