Robust Data-Driven Accelerated Mirror Descent

Tan, Hong Ye; Mukherjee, Subhadip; Tang, Junqi; Hauptmann, Andreas; Schönlieb, Carola-Bibiane

Link to publication on Research at Birmingham portal

ROBUST DATA-DRIVEN ACCELERATED MIRROR DESCENT

Hong Ye Tan[⋆]   *Subhadip Mukherjee*[⋆†]   *Junqi Tang*[⋆]
Andreas Hauptmann[‡§]   *Carola-Bibiane Schönlieb*[⋆]

[⋆] University of Cambridge, DAMTP, Cambridge CB3 0WA, U.K.
[†] Department of Computer Science, University of Bath, U.K.
[‡] University of Oulu, Research Unit of Mathematical Sciences, P.O.Box 8000, 90014 University of Oulu
[§] University College London, Department of Mathematics, 25 Gordon St, London WC1H 0AY, U.K.

ABSTRACT

Learning-to-optimize is an emerging framework that leverages training data to speed up the solution of certain optimization problems. One such approach is based on the classical mirror descent algorithm, where the mirror map is modelled using input-convex neural networks. In this work, we extend this functional parameterization approach by introducing momentum into the iterations, based on the classical accelerated mirror descent. Our approach combines short-time accelerated convergence with stable long-time behavior. We empirically demonstrate additional robustness with respect to multiple parameters on denoising and deconvolution experiments.

Index Terms— Mirror descent, accelerated optimization, learning-to-optimize, input-convex neural networks

1. INTRODUCTION

Learning-to-optimize is a general framework of methods that seek to minimize some objective functions quickly (in as few iterations as possible). This has been proposed in various settings such as unsupervised learning or using primal-dual methods, with applications such as imaging inverse problems and neural network training [1, 2, 3]. A recent work focuses on introducing a learned functional parameterization into the classical mirror descent (MD) algorithm, obtaining faster convergence rates with approximate convergence guarantees [4].

In this work, we propose a learned convex optimization approach based on the classical accelerated mirror descent (AMD) scheme. Our aim is to minimize a family of convex functions f in some function class $\mathcal{F}$ of qualitatively similar problems, such as variational image denoising. In this work, we will study and demonstrate convergence and robustness of our proposed learned scheme, and compare it to existing learned and classical optimization methods. Section 2 outlines the AMD scheme, and section 3 contains various experimental results on the robustness of our proposed method.

1.1. Mirror Descent

Mirror descent is a generalization of gradient descent (GD), first introduced by Nemirovsky and Yudin [5]. By exploiting the geometry of the cost function, MD achieves competitive convergence rate bounds for certain problems, including on-line learning and tomographic reconstruction [6, 7]. We first outline the method as presented by Beck and Teboulle [8].

Let Ψ be a convex function on a closed convex set $\mathcal{X} \subseteq \mathbb{R}^n$. Let $(\mathbb{R}^n)^*$ denote the corresponding dual space of $\mathbb{R}^n$. We denote by $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ the extended real line. Recall that the *convex conjugate* (or *Fenchel conjugate*) of Ψ, denoted $\Psi^* : (\mathbb{R}^n)^* \to \bar{\mathbb{R}}$, is given by $\Psi^*(y) = \sup_{x \in \mathcal{X}}(\langle y, x \rangle - f(x))$. Ψ induces a *Bregman divergence* $B_\Psi : \mathcal{X} \times \mathcal{X} \to \bar{\mathbb{R}}$, defined as:

$$B_\Psi(x, y) = \Psi(x) - \Psi(y) - \langle \nabla\Psi(y), x - y \rangle.$$

Definition 1 (Mirror potential) *We say* $\Psi : \mathcal{X} \to \mathbb{R}$ *is a **mirror potential** if it is continuously differentiable and strongly convex. We call the gradient* $\nabla\Psi : \mathcal{X} \to (\mathbb{R}^n)^*$ *a **mirror map**.*

If Ψ is a mirror potential, then the convex conjugate Ψ^* is everywhere differentiable, and moreover satisfies $\nabla\Psi^* = (\nabla\Psi)^{-1}$ [8, 9]. The standard MD update rule for minimizing convex f with initialization $x_0 \in \mathcal{X}$ is as follows [8]:

$$y_k = \nabla\Psi(x_k) - t_k\nabla f(x_k), \ \ x_{k+1} = \nabla\Psi^*(y_k). \quad (1)$$

For convex Lipschitz f, MD is able to attain $\mathcal{O}(1/\sqrt{k})$ convergence rate. Similarly to GD, algorithmic extensions such as stochasticity, ergodicity or acceleration are available for MD [10, 11, 12, 13]. MD can be additionally interpreted as natural GD, with a Riemannian metric induced by the mirror map [14, 15]. While the inclusion of the mirror map makes this method more flexible than GD, current applications are limited by the requirement of a hand-crafted mirror map. For efficient computation, such a mirror map is further restricted by the requirement of a closed-form convex conjugate, which is generally unavailable.

1.2. Learned MD

One way to make the MD algorithm learnable and alleviate the requirement of a hand-crafted mirror map is to replace $\nabla\Psi$ and $\nabla\Psi^*$ with learnable maps ∇M_θ and ∇M_ϑ^*. The learned mirror potential M_θ is parameterized as an input-convex neural network (ICNN) to enforce convexity [16]. While a closed-form convex conjugate is generally unavailable for an ICNN M_θ, this method learns another function M_ϑ^* to approximate the convex conjugate.

To maintain a MD structure for this scheme, the learned mirror maps are enforced to be close inverses of each other, $\nabla M_\vartheta^* \circ \nabla M_\theta \approx I$. We will refer to the distance between $\nabla M_\vartheta^* \circ \nabla M_\theta$ and the identity I as the *forward-backward error*. Using such a parameterization results in the learned MD (LMD) method with step-sizes $(t_k)_{k\geq 0}$ detailed in [4]:

$$\tilde{x}_{k+1} = \nabla M_\vartheta^*(\nabla M_\theta(\tilde{x}_k) - t_k\nabla f(\tilde{x}_k)). \tag{2}$$

The goal of LMD is to accelerate convergence on a class of functions, typically with qualitatively similar attributes, such as image denoising or inpainting. LMD has been shown to outperform methods such as GD and Adam for various convex problem classes. Formally, fix a function class $\mathcal{F}$ consisting of convex functions $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$. The goal is to minimize functions in this class $f \in \mathcal{F}$ quickly on average via LMD, with initializations $\tilde{x}_0 = x$ distributed possibly depending on f. For training, one possible loss function is to minimize the function values up to the N-th iteration for some fixed N, while also penalizing the distance of the mirror maps from being inverses of each other. This enforces the resulting LMD method to optimize quickly on the function class, and stay close to a proper MD method for convergence guarantees. This can be expressed as a minimization problem over the mirror potential parameters θ, ϑ:

$$\min_{\theta,\vartheta} \mathbb{E}_{f,x}\left[\sum_{k=1}^{N} r_k f(\tilde{x}_k) + s_k\|(\nabla M_\vartheta^* \circ \nabla M_\theta - I)(\tilde{x}_k)\|\right]. \tag{3}$$

2. ACCELERATED MD

The faster convergence rates of the Nesterov accelerated GD method can be extended to MD [17]. This can be done by considering the dynamics of certain ODEs corresponding to MD and Nesterov accelerated methods, and combining them in a suitable manner to derive an accelerated MD ODE [5, 10, 18]. Discretizing the resulting ODE and applying a small modification results in a family of accelerated MD (AMD) methods, which we summarize in the following section.

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, and $f : \mathcal{X} \to \bar{\mathbb{R}}$ be a proper C^1 convex function. Suppose further that ∇f is L_f-Lipschitz (w.r.t. a norm $\|\cdot\|$), and let f^* be the minimum of f on $\mathcal{X}$. Assume that Ψ^* is L_{Ψ^*}-smooth with respect to $\|\cdot\|_*$, the dual norm of $\|\cdot\|$ on the dual space. Let R be a

regularization function such that there exists $0 < \ell_R \leq L_R$ such that for all $x, x' \in \mathcal{X}$, $\frac{\ell_R}{2}\|x - x'\|^2 \leq R(x, x') \leq \frac{L_R}{2}\|x - x'\|^2$. The resulting AMD algorithm with initialization $\tilde{x}^{(0)} = x_0, \tilde{z}^{(0)} = \nabla\Psi^*(x_0)$ is as follows:

$$\lambda_k = \frac{r}{r + k}, \tag{4a}$$

$$x^{(k+1)} = \lambda_k\tilde{z}^{(k)} + (1 - \lambda_k)\tilde{x}^{(k)}, \tag{4b}$$

$$\tilde{z}^{(k+1)} = \nabla\Psi^*(\nabla\Psi(\tilde{z}^{(k+1)}) - \frac{kt}{r}\nabla f(x^{(k+1)})), \tag{4c}$$

$$\tilde{x}^{(k+1)} = \arg\min_{\tilde{x}\in\mathcal{X}} \gamma t\langle\nabla f(x^{(k+1)}), \tilde{x}\rangle + R(\tilde{x}, x^{(k+1)}). \tag{4d}$$

2.1. Learned AMD

We propose to make the AMD method learnable by replacing the mirror maps $\nabla\Psi$ and $\nabla\Psi^*$ with learned mirror maps ∇M_θ and ∇M_ϑ^*, respectively, as explained in Section 1.2. We additionally learn the step-sizes t_k for the first $N = 10$ iterations, similar to the adaptive LMD training in [4]. For simplicity, we fix $r = 3$ and $\gamma = 1$. We take the regularizer $R(x, x') = \frac{1}{2}\|x - x'\|_2^2$, turning (4d) into a GD step. The resulting learned iterates are as in Algorithm 1.

Algorithm 1: Learned AMD (LAMD)

Data: Mirror potential Ψ, step-sizes $(t_k)_{k=1}^N > 0$, parameter $r \geq 3$
Initialize $\tilde{x}^{(0)} = x_0, \tilde{z}^{(0)} = x_0$.
for $1 \leq k \leq N$ **do**
    $\lambda_k = \frac{r}{r+k}$.
    $x^{(k+1)} = \lambda_k\tilde{z}^{(k)} + (1 - \lambda_k)\tilde{x}^{(k)}$
    $\tilde{z}^{(k+1)} = \nabla M_\vartheta^*(\nabla M_\theta(\tilde{z}^{(k+1)}) - \frac{kt_k}{r}\nabla f(x^{(k+1)}))$
    $\tilde{x}^{(k+1)} = x^{(k+1)} - \gamma t_k\nabla f(x^{(k+1)})$
end

2.2. Training procedure

The first function class that we train on is for TV denoising of noisy ellipse phantoms in X-ray CT [19, 20]. The ellipse phantoms dataset was generated using the Deep Inversion Validation Library (DIVal) [21]. A sinogram is first created from an ellipse phantom of size 128^2 by first applying bilinear upsampling to 400^2 to avoid inverse crime [22], then applying a parallel-beam ray transform with 30 angles and 183 measurements per angle, and finally adding 10% Gaussian noise to the sinogram. A noisy phantom is created by taking the filtered back-projection (FBP) of this sinogram and downsampling it to 128^2. Denoting the noisy phantom thus generated by y, the resulting function class of TV-regularized variational losses is:

$$\mathcal{F} = \left\{ f(x) = \|x - y\|_2^2 + \lambda\|\nabla x\|_1 \right\}, \tag{5}$$

where λ is a regularization parameter, the gradient is taken pixel-wise, and the norms are taken over the image space. This setup allows us to train on realistic noise artifacts without introducing an expensive forward operator into training. We chose $\lambda = 0.15$ by visually comparing the TV-based reconstructions after running GD for 1000 iterations. The initializations were chosen to be the noisy phantoms $x_0 = y$.

For an initialization x_0, let $(x^{(k)})_{k=1}^N$ be the iterates generated by LAMD Algorithm 1. The loss function that we wish to minimize for training will be of the form (3) trained for $N = 10$ iterations, to minimize the function values at the iterates $f(x^{(k)})$, as well as to promote good mirror map inversion around the iterates and the ground-truths.

3. ROBUSTNESS STUDY OF LAMD

When using LMD or LAMD as a potential alternative for optimizing convex functions, the method should converge after running for many iterations. In the following section, we investigate the convergence behavior of the learned methods with different step-sizes, and generality with respect to the objective function class.

3.1. Extension of learned step-sizes

We first extend the LMD and LAMD iterations past the learned $N = 10$ iterations and investigate the long-time behavior of the iterations. This extension is done by continuing the corresponding MD update steps with the same learned mirror maps M_θ, M_ϑ^*, and by extending the step-sizes t_k using various methods. The extension methods used are based on the learned step-sizes $t_1, ..., t_{10}$. In particular, we use the maximum, mean, and minimum of the learned step-sizes, the final learned step-size $t_k = t_{10}$, and a "reciprocal" step-size. The reciprocal step-size is $t_k = c/k$, which was observed to be close to the learned step-sizes. Here, c was taken to fit the first $N = 10$ iterations, i.e. $c = \frac{1}{N} \sum_{k=1}^N k t_k$.

We observe in Figure 1 that the non-accelerated LMD method suffers from instability when taking small step-size extensions, and poor performance for large step-size extensions. While approximate convergence guarantees are available in this setting, a necessary condition is that the forward-backward error is uniformly bounded on the iterates, which is enforced for the trained iterates. However, this condition is violated when going past the trained number of iterates, as there is no guarantee that the forward-backward error stays low. In contrast, LAMD does well when taking small step-size extensions, as shown by the reciprocal extensions attaining roughly $\mathcal{O}(1/k)$ convergence rate. For the reciprocal extension, the iterates appear to continue decreasing the loss long after the learned number of iterations, making this a clear choice of step-size for further experiments.
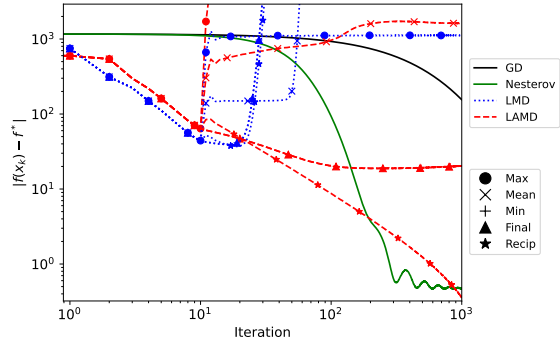

Fig. 1: Plot of $f(x^{(k)}) - f^*$ with various step-size extensions, where $f^* = \min f$. Extending step-sizes past the 10 learned iterations with the choice $t_k = c/k$ gives the best convergence. The minimum learned step-size here is the final one.

3.2. Domain transfer

We propose to transfer the LMD and LAMD methods away from denoising ellipse phantoms, introducing various structural changes to the objective function class. In particular, we will consider changing the ray transform, using a different phantom dataset, and using a convolution forward operator.

Changing the ray transform changes the noise distribution on the phantoms, introducing different artifacts. In particular, the parallel-beam ray transform in the noisy phantom generation is changed to a cone-beam ray transform with 60 angles and 400 measurements per angle. The parallel- and cone-beam transform experiments were also run on the LoDoPaB dataset, a benchmark for low dose CT reconstruction [23].

With the reciprocal step-sizes $t_k = c/k$, Figures 2a and 2b demonstrate that LMD continues to have good convergence for the earlier iterations, while still showing divergence as the iterations progress. LAMD continues to have roughly $\mathcal{O}(1/k)$ convergence rate in the later iterations, while still outperforming GD and the Nesterov accelerated scheme for the earlier iterations. We note that in Figure 2b, LMD appears to diverge later for the LoDoPaB dataset than with the ellipses dataset.

We also experiment with changing the forward operator. For the denoising experiments, the forward operator was taken to be the identity. Let A be a Gaussian convolution kernel with kernel size of 7px and standard deviation of 3px. We generate blurred phantoms y by applying the convolution A and adding 10% additive Gaussian noise. By changing the forward operator to the convolution A, we obtain a new function class corresponding to a deconvolution inverse problem:

$$\mathcal{F}_{\text{conv}} = \left\{ \|Ax - y\|^2 + \lambda\|\nabla x\|_1 : \text{blurred phantoms } y \right\}. \tag{6}$$

Figure 3 demonstrates the effect of applying LMD and LAMD with reciprocal step-sizes to the deconvolution function class, with mirror maps trained on either deconvolution or denoising problems. We see that LMD trained on deconvolution and denoising have very similar behavior. Moreover,
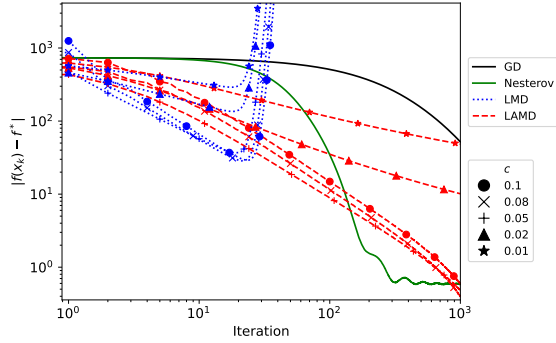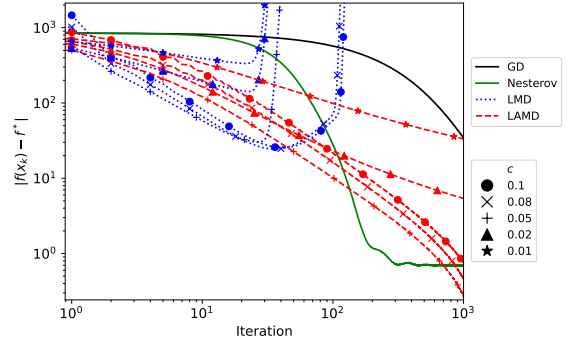
(a) Ellipses dataset.   (b) LoDoPaB dataset.

Fig. 2: Plot of $f(x^{(k)}) - f^*$ with various reciprocal step-sizes and cone beam transform. LMD and LAMD both generalize to denoising LoDoPaB phantoms, with LAMD achieving better convergence up to 10^3 iterations.
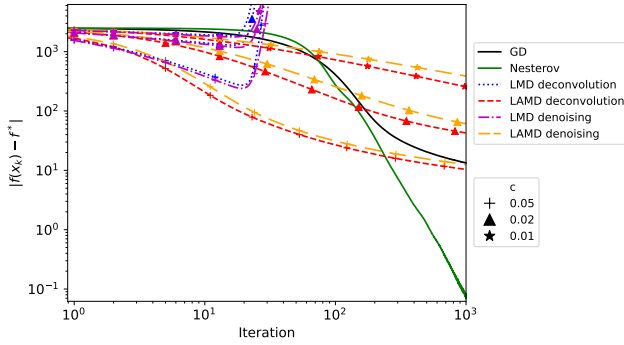

Fig. 3: Plot of $f(x^{(k)}) - f^*$ for deconvolution problem with various step-sizes, optimized using LMD/LAMD learned on deconvolution and on denoising. The similar performance of LMD and LAMD suggests a general learned mirror map which can be transferred across similar forward operators.
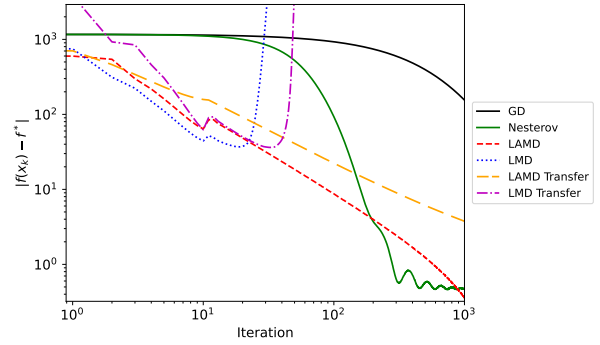
Fig. 4: Plot of $f(x^{(k)}) - f^*$ for LMD and LAMD for denoising with swapped mirror maps. Adding momentum and using the mirror map of LMD (yellow) gives long-term stability. Training with LAMD also adds stability to the mirror maps when applied without momentum (purple).

for LAMD, the optimizer trained on denoising performs only marginally worse than that trained on deconvolution for the three choices of c shown. This suggests generality of the mirror map, as applying the LAMD method trained on denoising to a different problem class still results in a stable and convergent method.

3.3. Effect of momentum on training

In this subsection, we investigate the difference between the mirror maps of the LMD and LAMD approaches. As the iterates are computed in a different manner, the training dynamics will also be different. Intuitively, we would expect the trained mirror maps to fit the geometry of the problem class.

We empirically compare the mirror maps by transferring the learned mirror maps between the LMD and LAMD methods. By transferring the mirror map learned using LMD to the LAMD scheme, or equivalently adding momentum to the LMD, we get the "LAMD Transfer" scheme. Transferring the maps learned using LAMD to the LMD scheme yields the "LMD Transfer" scheme. In Figure 4, we observe that adding acceleration to the LMD method still results in reasonable convergence rates, albeit worse than the LAMD method. If

momentum is removed from the LAMD method, the resulting iteration again performs worse than the trained LMD. However, the map trained using LAMD is more stable, diverging after some later iterations. This suggests that using momentum in the training process encourages the resulting mirror maps to be more stable, even with different training dynamics. Moreover, including momentum in the testing phase is a simple but effective method for adding stability to the iterates.

4. CONCLUSIONS

In this work, we propose Learned Accelerated Mirror Descent, which outperforms methods such as GD and Nesterov accelerated GD [4]. The LAMD scheme empirically results in stable extensions to further step-sizes, as well as more consistent mirror maps. The experiments also suggest that mirror maps trained on certain forward operators generalize to similar forward operators, such as from the identity to a convolution operator. A future avenue of research would be to extend convergence results for AMD to the approximate case where the inverse mirror map is not known exactly, or investigate robustness with respect to the momentum parameter.

5. REFERENCES

[1] Ke Li and Jitendra Malik, "Learning to optimize," 2017, International Conference on Learning Representations (ICLR).

[2] Sebastian Banert, Axel Ringh, Jonas Adler, Johan Karlsson, and Ozan Öktem, "Data-driven nonsmooth optimization," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 102–131, 2020.

[3] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin, "Learning to optimize: A primer and a benchmark," *Journal of Machine Learning Research*, vol. 23, pp. 1–59, 2022.

[4] Hong Ye Tan, Subhadip Mukherjee, Junqi Tang, and Carola-Bibiane Schönlieb, "Data-driven mirror descent with input-convex neural networks," 2022.

[5] Arkadij Semenovič Nemirovskij and David Borisovich Yudin, "Problem complexity and method efficiency in optimization," 1983.

[6] Francesco Orabona, Koby Crammer, and Nicolo Cesa-Bianchi, "A generalized online mirror descent with applications to classification and regression," *Machine Learning*, vol. 99, no. 3, pp. 411–435, 2015.

[7] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski, "The ordered subsets mirror descent optimization method with applications to tomography," *SIAM Journal on Optimization*, vol. 12, no. 1, pp. 79–108, 2001.

[8] Amir Beck and Marc Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[9] R. Tyrrell Rockafellar and Roger J.-B. Wets, *Variational Analysis*, Springer Verlag, Heidelberg, Berlin, New York, 1998.

[10] Walid Krichene, Alexandre Bayen, and Peter L Bartlett, "Accelerated mirror descent in continuous and discrete time," *Advances in neural information processing systems*, vol. 28, 2015.

[11] Filip Hanzely, Peter Richtarik, and Lin Xiao, "Accelerated bregman proximal gradient methods for relatively smooth convex optimization," 2018.

[12] John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan, "Ergodic mirror descent," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1549–1578, 2012.

[13] Ryan D'Orazio, Nicolas Loizou, Issam Laradji, and Ioannis Mitliagkas, "Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize," 2021.

[14] Garvesh Raskutti and Sayan Mukherjee, "The information geometry of mirror descent," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1451–1457, 2015.

[15] Suriya Gunasekar, Blake Woodworth, and Nathan Srebro, "Mirrorless mirror descent: A natural derivation of mirror descent," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2305–2313.

[16] Brandon Amos, Lei Xu, and J. Zico Kolter, "Input convex neural networks," in *Proceedings of the 34th International Conference on Machine Learning*. 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 146–155, PMLR.

[17] Yurii Nesterov, "A method for solving the convex programming problem with convergence rate $o(1/k^2)$," *Proceedings of the USSR Academy of Sciences*, vol. 269, pp. 543–547, 1983.

[18] Weijie Su, Stephen Boyd, and Emmanuel Candes, "A differential equation for modeling nesterov's accelerated gradient method: Theory and insights," *Advances in neural information processing systems*, vol. 27, pp. 2510–2518, 2014.

[19] Yang Wang and Haomin Zhou, "Total variation wavelet-based medical image denoising," *International Journal of Biomedical Imaging*, vol. 2006, 2006.

[20] Germana Landi and E Loli Piccolomini, "An efficient method for nonnegatively constrained total variation-based denoising of medical images corrupted by poisson noise," *Computerized Medical Imaging and Graphics*, vol. 36, no. 1, pp. 38–46, 2012.

[21] Johannes Leuschner, Maximilian Schmidt, and David Erzmann, "Deep inversion validation library," *Software available from https://github. com/jleuschn/dival*, 2019.

[22] Jari Kaipio and Erkki Somersalo, *Statistical and computational inverse problems*, vol. 160, Springer Science & Business Media, 2006.

[23] Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass, "LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction," *Scientific Data*, vol. 8, no. 1, apr 2021.