Total Variation Denoising:
Optimization, Algorithms, and Examples

Contents

1 Introduction

Total variation denoising (TVD) is a variational denoising method designed to remove noise while preserving sharp edges. In its simplest 1-D form, TVD assumes noisy observations

$$y_n = x_n + w_n, \quad n = 0, \dots, N-1,$$

where x is approximately piecewise constant and w is noise, often modeled as white Gaussian.

The TVD estimate is obtained as the solution of the convex optimization problem

$$\min_{x \in \mathbb{R}^N} \left\{ F(x) := \frac{1}{2}\|y - x\|_2^2 + \lambda \, \mathrm{TV}(x) \right\}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter and $\mathrm{TV}(x)$ is the (discrete) total variation.

This document:

- Reviews optimization tools: proximal gradient / ISTA, ADMM, and majorization–minimization (MM).
- Defines total variation and formulates TV denoising.
- Derives several algorithms for TVD, including a TVD-specific MM algorithm.
- Discusses optimality conditions, dual viewpoint, and extensions.

2 Basic notions and definitions

Here we collect one-line definitions for math terms beyond basic calculus that will be used frequently.

2.1 Linear algebra and norms

- **Vector space $\mathbb{R}^N$:** The set of N-dimensional real column vectors with standard addition and scalar multiplication.
- **Inner product $\langle u, v\rangle$:** For $u, v \in \mathbb{R}^N$, $\langle u, v\rangle := \sum_{n=0}^{N-1} u_n v_n$ is the usual Euclidean inner product.
- **Norm $\|x\|$:** A function measuring vector size that is nonnegative, homogeneous, and satisfies the triangle inequality.
- **ℓ_2 norm $\|x\|_2$:** $\|x\|_2 := \left(\sum_n x_n^2\right)^{1/2}$, the Euclidean norm.
- **ℓ_1 norm $\|x\|_1$:** $\|x\|_1 := \sum_n |x_n|$, the sum of absolute values; promotes sparsity.
- **Matrix transpose $A^\top$:** For a matrix A, $A^\top$ is the matrix whose entries are $[A^\top]_{ij} = A_{ji}$.
- **Positive definite matrix:** A symmetric matrix Q such that $x^\top Q x > 0$ for all nonzero x.
- **Tridiagonal matrix:** A matrix whose nonzero entries lie only on the main diagonal and the first upper and lower diagonals.

2.2 Convex analysis

- **Convex set:** A set C such that $\theta x + (1-\theta)y \in C$ for all $x, y \in C$ and $\theta \in [0, 1]$.
- **Convex function:** A function f is convex if $f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$ for all x, y and $\theta \in [0, 1]$.
- **Proper, lower semicontinuous (lsc) function:** A function f that never takes value $-\infty$, is not identically $+\infty$, and has closed epigraph; standard assumptions in convex optimization.
- **Subgradient:** For convex f, a vector g is a subgradient of f at x if

$$f(z) \ge f(x) + \langle g, z - x\rangle \quad \forall z.$$

- **Subdifferential $\partial f(x)$:** The set of all subgradients of f at x; $0 \in \partial f(x)$ is the optimality condition for unconstrained minimization.
- **Indicator function:** For a set C, $\iota_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise; used to encode constraints.
- **Lipschitz continuous gradient:** ∇g is L-Lipschitz if $\|\nabla g(x) - \nabla g(y)\|_2 \le L\|x - y\|_2$ for all x, y.

2.3 Proximal and splitting notions

- **Proximal operator:** For proper lsc convex f, $\operatorname{prox}_{\tau f}(v) := \arg\min_x \left\{ f(x) + \frac{1}{2\tau}\|x - v\|_2^2 \right\}$ is a generalized projection.
- **Soft-thresholding (shrinkage):** The scalar map

$$S_\alpha(t) := \operatorname{sign}(t) \max(|t| - \alpha, 0)$$

  is the proximal operator of $\alpha|\cdot|$.

- **Argmin:** $\arg\min_x f(x)$ denotes the set (or element) of minimizers of f.
- **Augmented Lagrangian:** A Lagrangian with an additional quadratic penalty term to stabilize constrained optimization.

3 Proximal gradient and ISTA

3.1 Composite optimization setup

Consider the composite convex optimization problem

$$\min_{x\in\mathbb{R}^N} F(x) := g(x) + h(x), \tag{2}$$

where:

- $g : \mathbb{R}^N \to \mathbb{R}$ is convex and differentiable with L-Lipschitz continuous gradient.
- $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is convex, possibly nonsmooth, but with an easily computable proximal operator.

3.2 Proximal gradient (forward–backward) iteration

The proximal gradient method (a.k.a. forward–backward splitting) iterates:

$$x^{k+1} = \operatorname{prox}_{\alpha_k h}\Big(x^k - \alpha_k \nabla g(x^k)\Big), \tag{3}$$

where $\alpha_k > 0$ is a step size, typically $\alpha_k \in (0, 2/L)$.

Intuition:

- The term $x^k - \alpha_k \nabla g(x^k)$ is a gradient descent step on the smooth part g.
- The proximal operator $\operatorname{prox}_{\alpha_k h}$ performs a "nonsmooth regularization" step controlled by h.

3.3 Iterative Shrinkage-Thresholding (ISTA)

For the classical ℓ_1-regularized least-squares problem

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1,$$

we have $g(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $h(x) = \lambda\|x\|_1$. Then $\nabla g(x) = A^\top(Ax - b)$, and the proximal map of $\lambda\|\cdot\|_1$ is componentwise soft-thresholding:

$$\operatorname{prox}_{\alpha\lambda\|\cdot\|_1}(v)_i = S_{\alpha\lambda}(v_i).$$

ISTA is exactly the proximal gradient method (3) for this choice:

$$x^{k+1} = S_{\alpha_k\lambda}\big(x^k - \alpha_k A^\top(Ax^k - b)\big). \tag{4}$$

Remark (convergence). If $0 < \alpha_k < 2/\|A^\top A\|_2$, then ISTA converges to a minimizer of the objective (sublinear rate $O(1/k)$ in function value).

4 Alternating Direction Method of Multipliers (ADMM)

4.1 General form

ADMM is a splitting method for constrained problems of the form

$$\min_{x,z} f(x) + g(z) \quad \text{s.t.} \quad Ax + Bz = c, \tag{5}$$

where f and g are convex, and the linear constraint couples x and z.

4.2 Augmented Lagrangian and scaled form

The augmented Lagrangian for (5) is

$$\mathcal{L}_\rho(x, z, u) = f(x) + g(z) + u^\top(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2,$$

where u is the dual variable and $\rho > 0$ is the penalty parameter.

ADMM updates (in the "scaled" form with u replaced by a scaled dual variable) are

$$x^{k+1} := \arg\min_x \left\{ f(x) + \frac{\rho}{2}\|Ax + Bz^k - c + u^k\|_2^2 \right\}, \tag{6}$$

$$z^{k+1} := \arg\min_z \left\{ g(z) + \frac{\rho}{2}\|Ax^{k+1} + Bz - c + u^k\|_2^2 \right\}, \tag{7}$$

$$u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c. \tag{8}$$

ADMM is attractive when the x- and z-subproblems are easier to solve than the original problem.

5 Majorization–Minimization (MM)

5.1 Majorizer

Given an objective $F : \mathbb{R}^N \to \mathbb{R}$, a function $G_k : \mathbb{R}^N \to \mathbb{R}$ is a *majorizer* of F at x^k if:

$$G_k(x) \geq F(x) \quad \forall x, \tag{9}$$

$$G_k(x^k) = F(x^k). \tag{10}$$

Intuitively, G_k lies above F everywhere and touches it at x^k.

5.2 MM iteration

With such a sequence of majorizers, MM iterates

$$x^{k+1} := \arg\min_x G_k(x). \tag{11}$$

For convex F, under mild assumptions, x^k converges to a minimizer of F. Quadratic majorizers are especially convenient because they lead to linear systems.

5.3 Example: majorizing $|t|$

For scalar $f(t) = |t|$ and a current point $t^k \neq 0$, one quadratic majorizer is

$$g(t; t^k) = \frac{1}{2|t^k|} t^2 + \frac{1}{2}|t^k|. \tag{12}$$

One checks easily that $g(t; t^k) \geq |t|$ for all t, with equality at $t = t^k$.

This scalar majorization extends to vector ℓ_1 norms and underpins the MM algorithm for TVD.

6 Discrete total variation and TV denoising

6.1 Discrete gradient and TV

For a 1-D signal $x = (x_0, \ldots, x_{N-1})^\top$, define the first-order difference operator $D \in \mathbb{R}^{(N-1)\times N}$ via

$$(Dx)_n = x_{n+1} - x_n, \quad n = 0, \ldots, N-2.$$

Concretely, D has the form

$$D = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -1 & 1 \end{bmatrix}.$$

Discrete total variation (anisotropic 1-D) is

$$\mathrm{TV}(x) := \|Dx\|_1 = \sum_{n=0}^{N-2} |x_{n+1} - x_n|. \tag{13}$$

It measures the sum of absolute jumps between consecutive samples.

Remark (2-D TV). For images, x is 2-D and D stacks horizontal and vertical finite differences, leading to 2-D anisotropic or isotropic TV, but we focus on 1-D for clarity.

6.2 TVD optimization problem

In vector form, the TVD problem is

$$\min_{x \in \mathbb{R}^N} \left\{ F(x) = \frac{1}{2}\|y - x\|_2^2 + \lambda \|Dx\|_1 \right\}. \tag{14}$$

Here:

- The data fidelity term $\frac{1}{2}\|y - x\|_2^2$ comes from a Gaussian noise model.
- The regularizer $\lambda\|Dx\|_1$ enforces sparsity of the discrete derivative, i.e., encourages piecewise constant x.
- $\lambda > 0$ tunes the tradeoff: larger $\lambda \Rightarrow$ stronger smoothing (fewer jumps).

6.3 TVD as a proximal operator

Define the functional $R(x) := \mathrm{TV}(x) = \|Dx\|_1$. Then (14) can be written as

$$x^\star = \arg\min_x \left\{ \frac{1}{2}\|x - y\|_2^2 + \lambda R(x) \right\}.$$

Hence $x^\star$ is the *proximal mapping* of λR at y:

$$x^\star = \mathrm{prox}_{\lambda R}(y). \tag{15}$$

This is conceptually useful: TVD is just a particular proximal operator. However, unlike separable ℓ_1-penalties, the prox of $R(x) = \|Dx\|_1$ is nontrivial, which motivates specialized algorithms.

7 MM algorithm for 1-D TVD

We now derive a TVD-specific MM algorithm that exploits the structure of D.

7.1 Quadratic majorizer for $\|Dx\|_1$

Set $v = Dx$ and $v^k = Dx^k$. Using the scalar majorizer of $|t|$ applied componentwise,

$$|v_n| \le \frac{1}{2|v_n^k|} v_n^2 + \frac{1}{2}|v_n^k|,$$

and summing over n we obtain a quadratic majorizer of $\|v\|_1$:

$$\|v\|_1 \le \frac{1}{2} v^\top \Lambda_k^{-1} v + \frac{1}{2}\|v^k\|_1,$$

where

$$\Lambda_k := \mathrm{diag}(|v^k|) = \mathrm{diag}(|Dx^k|).$$

Substituting $v = Dx$ gives

$$\|Dx\|_1 \le \frac{1}{2} x^\top D^\top \Lambda_k^{-1} D x + \frac{1}{2}\|Dx^k\|_1. \tag{16}$$

7.2 Majorizer for the full TVD objective

Add the data fidelity term to both sides of (16):

$$\frac{1}{2}\|y - x\|_2^2 + \lambda\|Dx\|_1 \le \frac{1}{2}\|y - x\|_2^2 + \lambda\left(\frac{1}{2} x^\top D^\top \Lambda_k^{-1} D x + \frac{1}{2}\|Dx^k\|_1\right).$$

Define

$$G_k(x) := \frac{1}{2}\|y - x\|_2^2 + \frac{\lambda}{2} x^\top D^\top \Lambda_k^{-1} D x + \frac{\lambda}{2}\|Dx^k\|_1. \tag{17}$$

Then G_k is a majorizer of F at x^k:

$$G_k(x) \ge F(x)\ \forall x, \quad G_k(x^k) = F(x^k).$$

7.3 MM update

The MM step is

$$x^{k+1} := \arg\min_x G_k(x). \tag{18}$$

Ignoring the additive constant $\frac{\lambda}{2}\|Dx^k\|_1$, we minimize the strictly convex quadratic

$$Q_k(x) := \frac{1}{2}\|y - x\|_2^2 + \frac{\lambda}{2} x^\top D^\top \Lambda_k^{-1} D x.$$

Setting the gradient to zero:

$$0 = \nabla Q_k(x) = (x - y) + \lambda D^\top \Lambda_k^{-1} D x,$$

which yields the linear system

$$\left(I + \lambda D^\top \Lambda_k^{-1} D\right) x = y. \tag{19}$$

Assuming invertibility (which holds because the matrix is symmetric positive definite), the update is

$$x^{k+1} = \left(I + \lambda D^\top \Lambda_k^{-1} D\right)^{-1} y. \tag{20}$$

7.4 Avoiding Λ_k^{-1} blow-up

As k increases, some entries of Dx^k approach zero, so some entries of Λ_k^{-1} blow up. To avoid explicit inversion of Λ_k, one can use the matrix inverse lemma (a.k.a. Woodbury identity), whose one form is

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(C^{-1} + DA^{-1}B\right)^{-1} DA^{-1}.$$

Applying this identity with a suitable choice of A, B, C, D leads to an equivalent expression for the inverse of $I + \lambda D^\top \Lambda_k^{-1} D$ that only involves Λ_k (not its inverse). One convenient expression is:

$$x^{k+1} = y - D^\top \left(\tfrac{1}{\lambda}\Lambda_k + DD^\top\right)^{-1} Dy, \quad \Lambda_k = \mathrm{diag}(|Dx^k|). \tag{21}$$

Banded structure. The matrix $DD^\top$ is tridiagonal, hence $\frac{1}{\lambda}\Lambda_k + DD^\top$ is tridiagonal (banded) and can be solved efficiently in $O(N)$ storage and $O(N)$ time using banded linear solvers.

7.5 Algorithm summary (MM for TVD)

Given y, $\lambda > 0$, and iterations K:

1. Initialize $x^0 := y$.
2. For $k = 0, 1, \ldots, K-1$:
   (a) Compute $v^k := Dx^k$ and $\Lambda_k := \mathrm{diag}(|v^k|)$.
   (b) Form the tridiagonal matrix
   $$M_k := \tfrac{1}{\lambda}\Lambda_k + DD^\top.$$
   (c) Solve $M_k z = Dy$ for z.
   (d) Set $x^{k+1} := y - D^\top z$.

Under standard conditions, x^k converges to the TVD solution $x^\star$.

8 ADMM for TVD

Another important algorithmic route is ADMM. We use a variable splitting formulation.

8.1 Splitting formulation

Rewrite TVD (14) as

$$\min_{x,z} \frac{1}{2}\|y - x\|_2^2 + \lambda\|z\|_1 \quad \text{s.t.} \quad z = Dx. \tag{22}$$

Here:

- x is the denoised signal.
- z is an auxiliary variable representing the discrete gradient Dx.

This matches the ADMM form (5) with

$$f(x) = \frac{1}{2}\|y - x\|_2^2, \quad g(z) = \lambda\|z\|_1, \quad A = D,\ B = -I,\ c = 0.$$

8.2 Augmented Lagrangian

Introduce a scaled dual variable u and penalty parameter $\rho > 0$. The scaled augmented Lagrangian is

$$\mathcal{L}_\rho(x, z, u) = \frac{1}{2}\|y - x\|_2^2 + \lambda\|z\|_1 + \frac{\rho}{2}\|Dx - z + u\|_2^2 - \frac{\rho}{2}\|u\|_2^2,$$

where the last term is a constant w.r.t. x, z for fixed u.

8.3 ADMM updates

ADMM proceeds by alternating minimization over x and z, followed by dual ascent on u.

x-update.

$$x^{k+1} = \arg\min_x \left\{ \frac{1}{2}\|y - x\|_2^2 + \frac{\rho}{2}\|Dx - z^k + u^k\|_2^2 \right\}.$$

This is a strictly convex quadratic; the optimality condition yields

$$\left(I + \rho D^\top D\right) x^{k+1} = y + \rho D^\top (z^k - u^k),$$

so

$$x^{k+1} = \left(I + \rho D^\top D\right)^{-1} \left[y + \rho D^\top (z^k - u^k)\right]. \tag{23}$$

Again $I + \rho D^\top D$ is tridiagonal and can be solved efficiently.

z-update.

$$z^{k+1} = \arg\min_z \left\{ \lambda\|z\|_1 + \frac{\rho}{2}\|Dx^{k+1} - z + u^k\|_2^2 \right\}.$$

This is the proximal operator of $\lambda\|\cdot\|_1$ applied to $Dx^{k+1} + u^k$, i.e.,

$$z^{k+1} = S_{\lambda/\rho}\left(Dx^{k+1} + u^k\right) \quad \text{(componentwise soft-thresholding)}. \tag{24}$$

u-update (dual variable).

$$u^{k+1} = u^k + Dx^{k+1} - z^{k+1}. \tag{25}$$

8.4 Algorithm summary (ADMM-TV)

Given y, $\lambda > 0$, $\rho > 0$, and iterations K:

1. Initialize $x^0 := y$, $z^0 := Dx^0$, $u^0 := 0$.
2. For $k = 0, \dots, K-1$:
   (a) Solve (23) for x^{k+1}.
   (b) Compute z^{k+1} via soft-thresholding (24).
   (c) Update u^{k+1} via (25).

ADMM-TV and MM-TV both exploit the banded structure of $D^\top D$; ADMM has the additional advantage of decoupling the nonsmooth ℓ_1 term.

9 Proximal gradient viewpoint for TVD

Direct application of proximal gradient to (14) is slightly awkward because the nonsmooth part is $h(x) = \lambda\|Dx\|_1$, whose proximal operator is not separable in the canonical basis.

9.1 Naive forward–backward split

Take

$$g(x) = \frac{1}{2}\|y - x\|_2^2, \quad h(x) = \lambda\|Dx\|_1.$$

Then

$$\nabla g(x) = x - y, \quad L = 1 \text{ (Lipschitz constant for } \nabla g).$$

The proximal gradient step reads

$$x^{k+1} = \operatorname{prox}_{\alpha\lambda\|D\cdot\|_1}\Big(x^k - \alpha(x^k - y)\Big) = \operatorname{prox}_{\alpha\lambda\|D\cdot\|_1}\Big((1-\alpha)x^k + \alpha y\Big).$$

To implement this, one must compute the TV proximal operator:

$$\operatorname{prox}_{\tau\|D\cdot\|_1}(v) := \arg\min_x \left\{\frac{1}{2}\|x - v\|_2^2 + \tau\|Dx\|_1\right\},$$

which is exactly a TVD problem again. Thus naive ISTA for TVD is essentially a nested TVD-in-TVD scheme; not attractive algorithmically. This is why MM, ADMM, and dual algorithms that exploit structure are more common for TVD.

10 Optimality conditions and dual characterization

10.1 Subgradient optimality

The TVD objective is

$$F(x) = \frac{1}{2}\|y - x\|_2^2 + \lambda\|Dx\|_1.$$

Its subdifferential is

$$\partial F(x) = x - y + \lambda D^\top p,$$

where $p \in \partial\|Dx\|_1$ and

$$p_n \in \partial|(Dx)_n| = \begin{cases} \{\mathrm{sign}((Dx)_n)\}, & (Dx)_n \neq 0, \\ [-1, 1], & (Dx)_n = 0. \end{cases}$$

The condition for $x^\star$ to minimize F is

$$0 \in \partial F(x^\star) \quad \Leftrightarrow \quad y - x^\star \in \lambda D^\top p^\star, \quad p^\star \in \partial\|Dx^\star\|_1. \tag{26}$$

10.2 Cumulative-sum characterization (1-D)

Define the "discrete antiderivative" operator $S \in \mathbb{R}^{N\times(N-1)}$ as the strict lower-triangular matrix of ones:

$$S = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

Then $DS = I_{N-1}$, i.e., S is a discrete integration operator (left inverse of D).

Let

$$r := y - x, \quad s := S^\top r.$$

The vector s is the cumulative sum of residuals:

$$s_n = \sum_{k=0}^{n}(y_k - x_k).$$

One can show that the optimality conditions for TVD are equivalent to the following:

$$|s_n| \leq \lambda, \quad \forall n, \tag{27}$$

$$\text{and} \quad \begin{cases} (Dx)_n > 0 \Rightarrow s_n = +\lambda, \\ (Dx)_n < 0 \Rightarrow s_n = -\lambda, \\ (Dx)_n = 0 \Rightarrow |s_n| < \lambda. \end{cases} \tag{28}$$

In words:

- The cumulative sum of residuals stays within the tube $[-\lambda, \lambda]$.
- Whenever the signal has a positive jump, s_n hits the upper boundary; for a negative jump, it hits the lower boundary; in flat regions, s_n remains strictly inside the tube.

This gives a geometric picture of the TVD solution and underlies fast direct algorithms in 1-D.

11 Examples and qualitative behavior

11.1 Piecewise-constant signal

Consider a "blocky" signal with two levels:

$$x_n = \begin{cases} 0, & 0 \le n < N/2, \\ 1, & N/2 \le n < N. \end{cases}$$

Add Gaussian noise $w_n \sim \mathcal{N}(0, \sigma^2)$ to get $y_n = x_n + w_n$.

Qualitative behavior of TVD as λ changes:

- Very small λ: $x^\star$ follows noisy fluctuations, only slightly smoothed; edges are preserved but noise remains.
- Moderate λ: $x^\star$ becomes nearly piecewise constant, with a well-localized jump near the true edge; noise in flat regions is strongly suppressed.
- Very large λ: $x^\star$ collapses towards a constant (the global mean of y); edges are oversmoothed and disappear.

11.2 Staircasing on non-blocky signals

For non-piecewise-constant signals (e.g., ramps or sinusoids), TVD often exhibits "staircasing":

- Slowly varying regions are approximated by a sequence of flat segments separated by small jumps.
- This is because TV explicitly penalizes the ℓ_1 norm of discrete derivatives, favoring exact zeros in the derivative (flat), rather than small nonzero slopes.

This motivates higher-order TV (penalizing second-order differences) for signals with smooth trends.

12 Extensions

12.1 2-D TV for images

For an image X on a 2-D grid, define horizontal and vertical finite differences:

$$(D_x X)_{i,j} = X_{i,j+1} - X_{i,j}, \quad (D_y X)_{i,j} = X_{i+1,j} - X_{i,j}.$$

- **Anisotropic TV:** $\mathrm{TV}_{\mathrm{aniso}}(X) = \sum_{i,j} \big(|(D_x X)_{i,j}| + |(D_y X)_{i,j}| \big)$.
- **Isotropic TV:** $\mathrm{TV}_{\mathrm{iso}}(X) = \sum_{i,j} \sqrt{(D_x X)_{i,j}^2 + (D_y X)_{i,j}^2}$.

Both forms lead to image denoising problems analogous to (14). Algorithms like ADMM and primal–dual methods generalize straightforwardly.

12.2 Higher-order TV

To reduce staircasing, one can penalize higher-order differences, e.g., second-order TV:

$$\mathrm{TV}_2(x) = \|D_2 x\|_1,$$

where D_2 is a discrete second-difference operator, $(D_2 x)_n = x_{n+1} - 2x_n + x_{n-1}$. This encourages piecewise-linear signals instead of piecewise-constant signals.

12.3 Other noise models and data terms

TVD is easily adapted to other data terms:

- Poisson noise $\Rightarrow$ Kullback–Leibler-type data fidelity.
- Laplacian noise $\Rightarrow$ ℓ_1 data fidelity term.
- Deconvolution and inpainting $\Rightarrow$ Hx in place of x in the fidelity term for some linear operator H.

The same splitting and MM strategies apply, with modified linear systems and proximal steps.

13 Summary

Total variation denoising is the solution of a convex variational problem that combines a quadratic data fidelity term with an ℓ_1 penalty on discrete gradients. The TV prior encodes piecewise-constant structure and preserves edges, unlike linear low-pass filters.

We have:

- Presented proximal gradient / ISTA, ADMM, and MM as core tools from convex optimization.
- Formulated TVD in 1-D as
$$\min_x \frac{1}{2}\|y - x\|_2^2 + \lambda\|Dx\|_1,$$
with D the first-difference operator.
- Derived a TVD-specific MM algorithm based on a quadratic majorizer of $\|Dx\|_1$ leading to a banded linear system per iteration.
- Derived an ADMM algorithm for TVD using the splitting $z = Dx$, yielding linear system + soft-thresholding updates.
- Described optimality conditions via cumulative sums of residuals, giving intuition and enabling fast 1-D algorithms.
- Briefly discussed staircasing, higher-order TV, and extensions to 2-D and other data terms.

These tools form the basic mathematical toolkit for understanding and implementing total variation denoising in both 1-D signal and 2-D image settings.