Convex Optimization and Modeling

Proximal Methods

12th lecture, 30.06.2010

Jun.-Prof. Matthias Hein

First Order Methods (continued):

- Fast Projected Gradient,
- Proximal Gradient Methods (ISTA, FISTA)

References:

- Bertsekas: Convex Optimization (2010) - Chapter 6.
- Beck, Teboulle: "Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems", SIAM J. Imaging Sciences, 2:183-202,(2009).
- Combettes, Pesquet: "Proximal Splitting Methods in Signal Processing", submitted, (2009).
- Tseng: "On Accelerated Proximal Gradient Methods for Convex-Concave Optimization, submitted, (2008).

Constrained Optimization:

$$\min_{x \in C} f(x)$$

- f is continuously differentiable and gradient has Lipschitz constant L,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| .$$

- C is convex and closed
- efficient if projection P_C onto C can be easily computed

Main Ingredient:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 .$$

Later: Projected Newton, Projected Subgradient.

Projected Gradient

Quadratic upper bound at current point x^k: For $t \leq \frac{1}{L}$,

$$\begin{aligned} x' &= \arg\min_{x \in C} \; f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2t} \left\| x - x^k \right\|^2 \\ &= \arg\min_{x \in C} \; \left\| x - \left(x^k - t \nabla f(x^k)\right) \right\|^2 \\ &= P_C\Big(x^k - t\,\nabla f(x^k)\Big) \end{aligned}$$

Next iterate:

$$x^{k+1} = P_C\Big(x^k - t^k\,\nabla f(x^k)\Big),$$

Stepsize t^k:

- constant $0 < t^k < \frac{1}{L}$,
- backtracking line search - finds Lipschitz constant (somehow).

Theorem 1. *The iterate $f(x^{(k)})$ of the projected gradient method satisfies*

- *for constant stepsize $0 < t \leq \frac{1}{L}$,*

$$f(x^{(k)}) - p^* \leq \frac{1}{2\,k\,t} \left\| x^{(0)} - x^* \right\|^2 .$$

- *for backtracking line-search,*

$$f(x^{(k)}) - p^* \leq \frac{1}{2\,k\,t_{\min}} \left\| x^{(0)} - x^* \right\|^2 ,$$

where $t_{\min} = \min_{i=1,\ldots,k} t^i$.

- much **slower** than the convergence rate of the gradient method for strictly convex functions with bounds on the Hessian,
- much **faster** than the $O(\frac{1}{\sqrt{k}})$ convergence rate of the subgradient method
- Projection does not influence the convergence rate.

Nesterov's accelerated projected gradient method Start with $x^{(0)} \in C$, set $y^{(0)} = x^{(0)}$

$$x^{(k+1)} = P_C\Big(y^{(k)} - t^{(k)} \nabla f(y^{(k)})\Big),$$

$$y^{(k+1)} = x^{(k+1)} + \frac{k}{k+3}\big(x^{(k+1)} - x^{(k)}\big).$$

Remarks:

- related to the "heavy-ball" method of Polyak.
- Other update factors than $\frac{k-1}{k+2}$ are possible,
- $y^{(k)}$ is not necessarily feasible.
- As $k \to \infty$ the step becomes the projected gradient step.
- $f(x^{(k)})$ is not monotonically decreasing.

Theorem 2. *The iterate $f(x^{(k)})$ of the fast projected gradient method satisfies*

- *for constant stepsize $0 < t \leq \frac{1}{L}$,*

$$f(x^{(k)}) - p^* \leq \frac{2}{(k+1)^2\, t} \left\| x^{(0)} - x^* \right\|^2 .$$

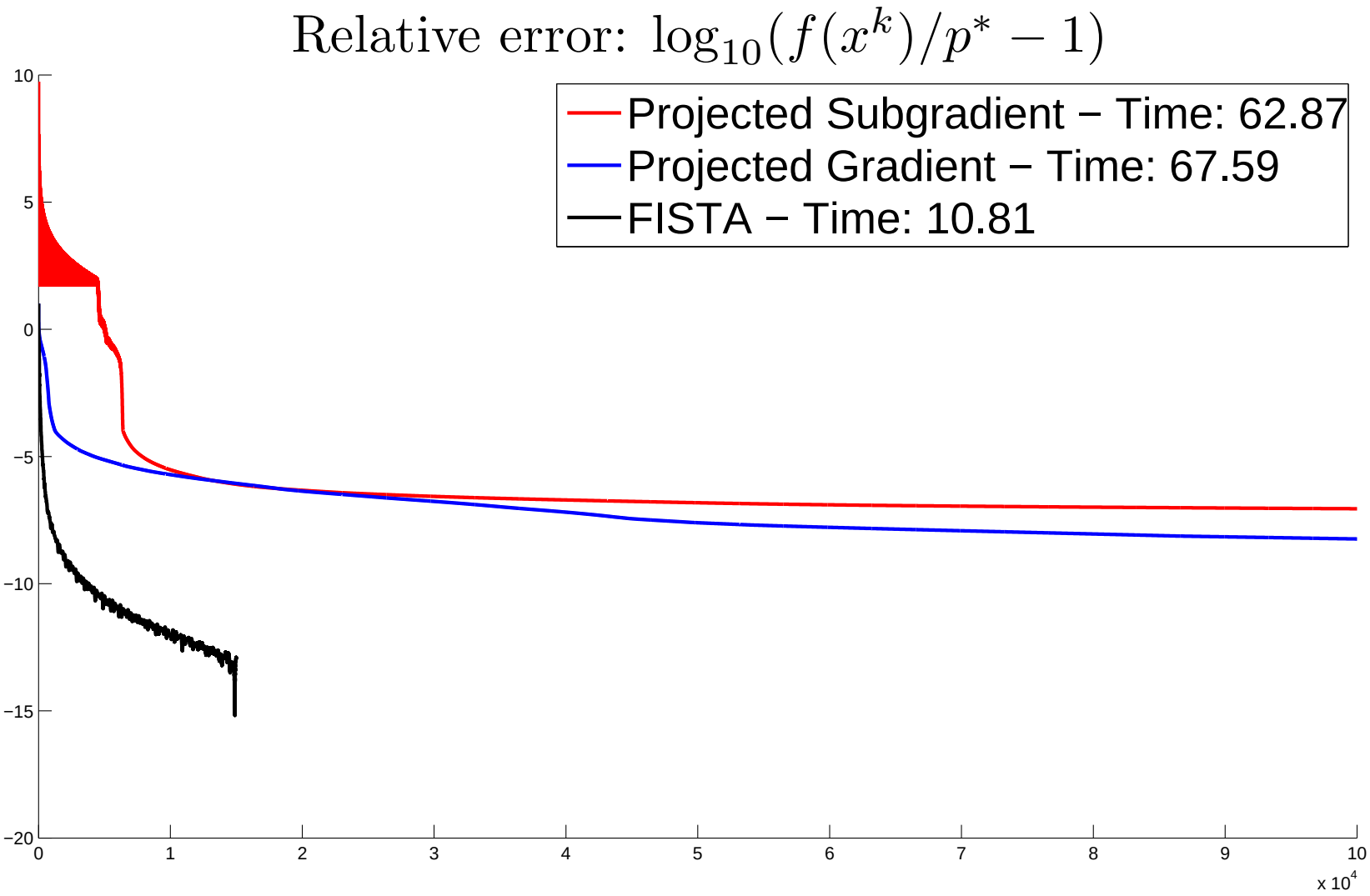
- *for backtracking line-search,*

$$f(x^{(k)}) - p^* \leq \frac{2}{(k+1)^2\, t_{\min}} \left\| x^{(0)} - x^* \right\|^2 ,$$

where $t_{\min} = \min_{i=1,\ldots,k} t^i$.

Discussion:

- Much faster than the projected gradient method !
  Requires $O\left(\sqrt{\frac{L}{\varepsilon}}\right)$ iterations instead of $O\left(\frac{L}{\varepsilon}\right)$ for the projected gradient method !


Relative error: $\log_{10}(f(x^k)/p^* - 1)$
Projected Subgradient − Time: 62.87
Projected Gradient − Time: 67.59
FISTA − Time: 10.81
10
5
0
−5
−10
−15
−20
0
1
2
3
4
5
6
7
8
9
10
x 10^4

Gradient map:

$$G_t(x) = \frac{1}{t}\Big(x - P_C\big(x - t\nabla f(x)\big)\Big).$$

Note that,

$$P_C\big(x - t\nabla f(x)\big) = x - t\, G_t(x).$$

and from the optimality condition of the Projection P_C,

$$\big\langle x - t\nabla f(x) - P_C\big(x - t\nabla f(x)\big), z - P_C\big(x - t\nabla f(x)\big)\big\rangle \;\leq\; 0, \quad \forall\, z \in C.$$

one obtains

$$\langle G_t(x) - \nabla f(x), z - x + tG_t(x)\rangle \;\leq\; 0, \quad \forall\, z \in C.$$

Projected Gradient - Proof

Lemma: We have for all $y \in C$ and $0 < t \leq \frac{1}{L}$,

$$f(P_C\big(x - t\nabla f(x)\big)) \leq f(y) + \langle G_t(x), x - y\rangle - \frac{t}{2}\left\|G_t(x)\right\|^2.$$

We get with $x^{(k+1)} = P_C\big(y^{(k)} - t\nabla f(y^{(k)})\big) = y^{(k)} - t\,G_t(y^{(k)})$,

$$f(x^{(k+1)}) \;\leq\; f(x^{(k)}) + \left\langle G_t(x), y^{(k)} - x^{(k)}\right\rangle - \frac{t}{2}\left\|G_t(y^{(k)})\right\|^2,$$

$$f(x^{(k+1)}) \;\leq\; f(x^{*}) + \left\langle G_t(x), y^{(k)} - x^{*}\right\rangle - \frac{t}{2}\left\|G_t(y^{(k)})\right\|^2,$$

and thus a convex combination yields,

$$f(x^{(k+1)}) \leq (1-\theta)f(x^{(k)}) + \theta f(x^*) + \left\langle G_t(y^{(k)}), y^{(k)} - (1-\theta)x^{(k)} - \theta x^*\right\rangle - \frac{t}{2}\left\|G_t(y^{(k)})\right\|^2.$$

Projected Gradient - Proof II

Denote by $\theta = \theta^{(k)}$ and $t = t^{(k)}$

$$f(x^{(k+1)}) \leq (1-\theta) f(x^{(k)}) + \theta f(x^*) + \left\langle G_t(y^{(k)}), y^{(k)} - (1-\theta)x^{(k)} - \theta x^* \right\rangle - \frac{t}{2}\left\| G_t(y^{(k)}) \right\|^2$$

$$= (1-\theta) f(x^{(k)}) + \theta f(x^*) + \theta \left\langle G_t(x), v^{(k)} - x^* \right\rangle - \frac{t}{2}\left\| G_t(y^{(k)}) \right\|^2$$

$$= (1-\theta) f(x^{(k)}) + \theta f(x^*) + \frac{\theta^2}{2t}\left(\left\| v^{(k)} - x^* \right\|^2 - \left\| v^{(k)} - \frac{t}{\theta} G_t(y^{(k)}) - x^* \right\|^2 \right)$$

where we defined, $v^{(k)} := \frac{1}{\theta^{(k)}}\left(y^{(k)} - (1-\theta^{(k)})x^{(k)}\right)$, $\quad v^{(k+1)} := v^{(k)} - \frac{t^{(k)}}{\theta^{(k)}} G_t(y^{(k)})$,

$$\frac{1}{\theta^2}(f(x^{(k+1)}) - f(x^*)) + \frac{1}{2t}\left\| v^{(k+1)} - x^* \right\|^2 \leq \frac{1-\theta}{\theta^2}(f(x^{(k)}) - f(x^*)) + \frac{1}{2t}\left\| v^{(k)} - x^* \right\|^2.$$

Using $\theta = \frac{2}{k+2}$,

$$\frac{1-\theta^{(k)}}{(\theta^{(k)})^2} = \frac{k(k+2)}{4} \leq \frac{1}{(\theta^{(k-1)})^2},$$

we get,

$$\frac{1}{(\theta^{(k)})^2}(f(x^{(k+1)}) - f(x^*)) + \frac{1}{2t^{(k)}}\left\| v^{(k+1)} - x^* \right\|^2 \leq \frac{1-\theta^{(0)}}{(\theta^{(0)})^2}(f(x^{(0)}) - f(x^*)) + \frac{1}{2t^{(0)}}\left\| v^{(0)} - x^* \right\|^2.$$

Projected Gradient - Proof II

$$\frac{1}{(\theta^{(k)})^2}(f(x^{(k+1)})-f(x^*))+\frac{1}{2t}\left\|v^{(k+1)}-x^*\right\|^2 \leq \frac{1-\theta^{(0)}}{(\theta^{(0)})^2}(f(x^{(0)})-f(x^*))+\frac{1}{2t^{(0)}}\left\|v^{(0)}-x^*\right\|^2.$$

Using

$$\theta^{(0)}=\frac{2}{2}=1 \quad \Longrightarrow \quad v^{(0)}=\frac{1}{\theta^{(0)}}\left(y^{(0)}-(1-\theta^{(0)})x^{(0)}\right)=x^{(0)},$$

we arrive at

$$f(x^{(k+1)})-f(x^*) \leq \frac{4}{(k+2)^2}\frac{1}{2t^{(0)}}\left\|x^{(0)}-x^*\right\|^2.$$

For the backtracking line search one has to replace $t^{(0)}$ with $\min_{i=1,\ldots,k} t^{(k)}$.

How does the update scheme enter the proof ?

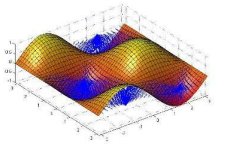
The update scheme enters by $x^{(k+1)} = y^{(k)} - t\, G_t(y^{(k)})$ and $\theta^{(k)} = \frac{2}{k+2}$ using

$$v^{(k)} := \frac{1}{\theta^{(k)}}\big(y^{(k)} - (1-\theta^{(k)})x^{(k)}\big), \qquad v^{(k+1)} := v^{(k)} - \frac{t^{(k)}}{\theta^{(k)}}\, G_t(y^{(k)}).$$

we obtain,

$$\begin{aligned}
y^{(k+1)} &= (1-\theta^{(k+1)})x^{(k+1)} + \theta^{(k+1)}v^{(k+1)} \\
&= \frac{k+1}{k+3}x^{(k+1)} + \frac{2}{k+3}\frac{1}{\theta^{(k)}}\left(y^{(k)} - (1-\theta^{(k)})x^{(k)} - \frac{t^{(k)}}{\theta^{(k)}}\, G_t(y^{(k)})\right) \\
&= \frac{k+1}{k+3}x^{(k+1)} + \frac{k+2}{k+3}\left(x^{(k+1)} - \frac{k}{k+2}x^{(k)}\right) \\
&= x^{(k+1)} + \frac{k}{k+3}\left(x^{(k+1)} - x^{(k)}\right)
\end{aligned}$$

Nesterov's Optimality Results



First Order Method: We define a first-order method as an iterative methods with

$$x^{(k+1)} \in x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k)})\}.$$

Optimality:

There is no first order method which has a better convergence rate than $O(\frac{1}{k^2})$ **uniformly** over the set of all functions with Lipschitz continuous gradient.

Motivation: FISTA

Optimization problems in:

- machine learning, data mining,
- signal processing, image processing

are often of the form:

$$\min_{x \in C} f(x) + g(x),$$

where

- f is convex, continuously differentiable and has Lipschitz-continuous gradient,
$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|.$$
- g is convex and is possibly non-differentiable.

Often $C = \mathbb{R}^d$.

Assumption: Problem is primal feasible.

Examples:

- Lasso:

$$\min_{w \in \mathbb{R}^n} \|\Phi w - Y\|^2 + \|w\|_1 ,$$

  or in combination with other loss functions.

- Total Variation Denoising:

$$\min_{f \in \mathbb{R}^{n \times m}, 0 \preceq f \preceq 1} \|f - Y\|^2 + \|Df\|_1 ,$$

  where D is the derivative operator on the grid (or a general graph).

- Matrix Completion:

$$\min_{M \in \mathbb{R}^{n \times n}, M \succeq 0} \|M - Y\|^2 + \|M\|_1 .$$

- and many more !

Definition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous and convex. For every $x \in \mathbb{R}^n$,*

$$\min_{y \in \mathbb{R}^n} \; f(y) + \frac{1}{2} \|x - y\|^2 ,$$

has a unique solution denoted by $\operatorname{prox}_f x$. The operator $\operatorname{prox}_f : \mathbb{R}^n \to \mathbb{R}$ is called the ***proxmity operator****.*

Properties:

- if f is the indicator function of a closed, convex set C, then prox_f is the projection onto C,
- $\|\operatorname{prox}_f x - \operatorname{prox}_f y\| \leq \|x - y\|$ (non-expansive).
- $x^* = \operatorname{prox}_f x^*$ if and only if x^* is a minimizer of f.
- Let $t > 0$, then $f_t(x) := \inf_y f(y) + \frac{1}{2t} \|x - y\|^2$ is continuously differentiable with $\frac{1}{t}$-Lipschitz gradient,

$$\nabla f_t(x) = \frac{1}{t}\big(x - \operatorname{prox}_{t\, f} x\big).$$

Examples of the proximity operator:

It can be easily computed when g is separable.

- $f(x) = |x|$, $\operatorname{prox}_{tf} x = \operatorname{sign}(x)(|x| - t)_+$.
- $f(x) = x^2$, $\operatorname{prox}_{tf} x = \frac{x}{1+2t}$,
- $f(x) = |x|^p$, $\operatorname{prox}_{tf} x = \operatorname{sign}(x)\rho$, where $\rho \geq 0$ and $\rho + t\,p\,\rho^{p-1} = |x|$,
- The Huber-loss:

$$f(x) = \begin{cases} \alpha x^2, & \text{if } |x| \leq \frac{\beta}{\sqrt{2\alpha}}, \\ \beta\sqrt{2\alpha}|x| - \frac{\beta^2}{2}, & \text{else.} \end{cases}$$

Then

$$\operatorname{prox}_{tf} x = \begin{cases} \frac{x}{1+2\alpha}, & \text{if } |x| \leq (1+2\alpha)\frac{\beta}{\sqrt{2\alpha}}, \\ x - \beta\sqrt{2\alpha}\operatorname{sign}(x), & \text{else.} \end{cases}$$

Backward-Forward Splitting:

$$x^{(k+1)} = \operatorname{prox}_{t^{(k)}\, g}\left(x^{(k)} - t^{(k)} \nabla f(x^{(k)})\right).$$

Note, that

$$\begin{aligned}
x^{(k+1)} &= \arg\min_y t^{(k)} g(y) + \frac{1}{2}\left\|x^{(k)} - t^{(k)}\nabla f(x^{(k)}) - y\right\|^2 \\
&= \arg\min_y t^{(k)}\left[f(x^{(k)}) + \frac{1}{t^{(k)}}\left\|x^{(k)} - t^{(k)}\nabla f(x^{(k)}) - y\right\|^2 + g(y)\right] \\
&= \arg\min_y f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), y - x^{(k)}\right\rangle + \frac{1}{2\,t^{(k)}}\left\|x^{(k)} - y\right\|^2 + g(y)
\end{aligned}$$

$\Rightarrow$ minimize quadratic approximation of f + non-smooth part g.

- becomes projected gradient for $g(x) = I_C(x)$.
- becomes iterative shrinkage thresholding algorithm (ISTA) for $f(x) = \|Ax - b\|^2$ and $g(x) = \lambda \|x\|_1$.

ISTA - Convergence Rate

Let $F(x) = f(x) + g(x)$.

Theorem 3. *The iterate $F(x^{(k)})$ of the backward forward splitting satisfies*

- *for constant stepsize $0 < t \leq \frac{1}{L}$,*

$$F(x^{(k)}) - p^* \leq \frac{1}{2\,k\,t} \left\| x^{(0)} - x^* \right\|^2.$$

- *for backtracking line-search,*

$$F(x^{(k)}) - p^* \leq \frac{1}{2\,k\,t_{\min}} \left\| x^{(0)} - x^* \right\|^2,$$

where $t_{\min} = \min_{i=1,\ldots,k} t^i$.

- L is the Lipschitz constant of ∇f
- Proof: replace projection by proximal map.

FISTA - accelerated proximal gradient method:

$$x^{(k+1)} = \text{prox}_{t^{(k)}\, g}\left(y^{(k)} - t^{(k)}\nabla f(y^{(k)})\right)$$

$$y^{(k+1)} = x^{(k+1)} + \frac{k}{k+3}\left(x^{(k+1)} - x^{(k)}\right).$$

Proposed by Beck and Teboulle (2009) - Nesterov in technical report(2007).

- Beck and Teboulle use $t^{(k)} = \frac{1}{L}$ and

$$y^{(k+1)} = x^{(k+1)} + \frac{s_k - 1}{s_{k+1}}\left(x^{(k+1)} - x^{(k)}\right),$$

with $s_0 = 1$ and $s_{k+1} = \frac{1+\sqrt{1+4\,s_k^2}}{2}$.

FISTA

Let $F(x) = f(x) + g(x)$.

Theorem 4. *The iterate $F(x^{(k)})$ of FISTA satisfies*

- *for constant stepsize $0 < t \leq \frac{1}{L}$,*

$$F(x^{(k)}) - p^* \leq \frac{2}{(k+1)^2\, t} \left\| x^{(0)} - x^* \right\|^2 .$$

- *for backtracking line-search,*

$$F(x^{(k)}) - p^* \leq \frac{2}{(k+1)^2\, t_{\min}} \left\| x^{(0)} - x^* \right\|^2 ,$$

  where $t_{\min} = \min_{i=1,\ldots,k} t^i$.

- Proof: replace projection by proximal map.

Total Variation Denoising:

- given image $Y \in \mathbb{R}^{n \times m} \rightarrow Y \in \mathbb{R}^{nm}$.
- E denotes the edge set of the $n \times m$-grid.
- λ regulates the amount of denoising.
- the total variation regularizer enforces a piecewise constant image - sharpening of edges.
- the box constraints enforce that I is again an image.

$$\min_{0 \leq I \leq 1} \frac{1}{2} \|I - Y\|^2 + \lambda \sum_{(i,j) \in E} w_{ij} |I_i - I_j|.$$

Total Variation Denoising:

- general graph-based point of view,
- image f as function on a graph (grid),
- total-variation penalizes function differences on edges

$$\min_{f \in C} \frac{1}{2} \|f - Y\|^2 + \lambda \sum_{i,j=1}^{n} w_{ij} |f_i - f_j|$$

$$= \min_{f \in C} \max_{\alpha \in \mathbb{R}^E \, \|\alpha\|_\infty \leq 1, \, \alpha_{ij} = -\alpha_{ji}} \frac{1}{2} \|f - Y\|^2 + \lambda \sum_{i,j=1}^{n} w_{ij} (f_i - f_j) \alpha_{ij}$$

$$= \max_{\alpha \in \mathbb{R}^E, \, \|\alpha\|_\infty \leq 1, \, \alpha_{ij} = -\alpha_{ji}} \min_{f \in C} \frac{1}{2} \|f - Y\|^2 + 2\lambda \sum_{i,j=1}^{n} w_{ij} \alpha_{ij} f_i.$$

Inner problem has solution: $f = P_C(Y - \lambda A\alpha)$, where $(A\alpha)_i := 2 \sum_{j=1}^{n} w_{ij} \alpha_{ij}$.

Transformed Problem:

$$\max_{\|\alpha\|_\infty \leq 1, \alpha_{ij}=-\alpha_{ji}} \Psi(\alpha),$$

where

$$\begin{aligned}
\Psi(\alpha) &:= \frac{1}{2}\|Y - P_C(Y - \lambda A\alpha)\|^2 + \lambda \langle A\alpha, P_C(Y - \lambda\, A\alpha)\rangle \\
&= \frac{1}{2}\|Y - \lambda A\alpha - P_C(Y - \lambda A\alpha)\|^2 + \lambda \langle A\alpha, P_C(Y - \lambda\, A\alpha)\rangle - \frac{\lambda^2}{2}\|A\alpha\|^2 \\
&\quad + \lambda \langle A\alpha, Y - P_C(Y - \lambda\, A\alpha)\rangle \\
&= \frac{1}{2}\|Y - \lambda A\alpha - P_C(Y - \lambda A\alpha)\|^2 + \lambda \langle A\alpha, Y\rangle - \frac{\lambda^2}{2}\|A\alpha\|^2
\end{aligned}$$

From the properties of the proximity operator:

$$\nabla \frac{1}{2}\|x - P_C(x)\|^2 = x - P_C(x).$$

$$\Longrightarrow \quad \nabla\Psi(\alpha) = -\lambda\, A^T P_C(Y - \lambda A\alpha)$$

Lipschitz constant of $\nabla\Psi$:

$$\begin{aligned}
\|\Psi(\alpha) - \Psi(\beta)\| &= \lambda \left\|A^T P_C(Y - \lambda A\alpha) - A^T P_C(Y - \lambda A\beta)\right\| \\
&\leq \lambda \left\|A^T\right\| \|P_C(Y - \lambda A\alpha) - P_C(Y - \lambda A\beta)\| \\
&\leq \lambda \left\|A^T\right\| \|(Y - \lambda A\alpha) - (Y - \lambda A\beta)\| \\
&= \lambda^2 \|A\|^2 \|\alpha - \beta\|
\end{aligned}$$

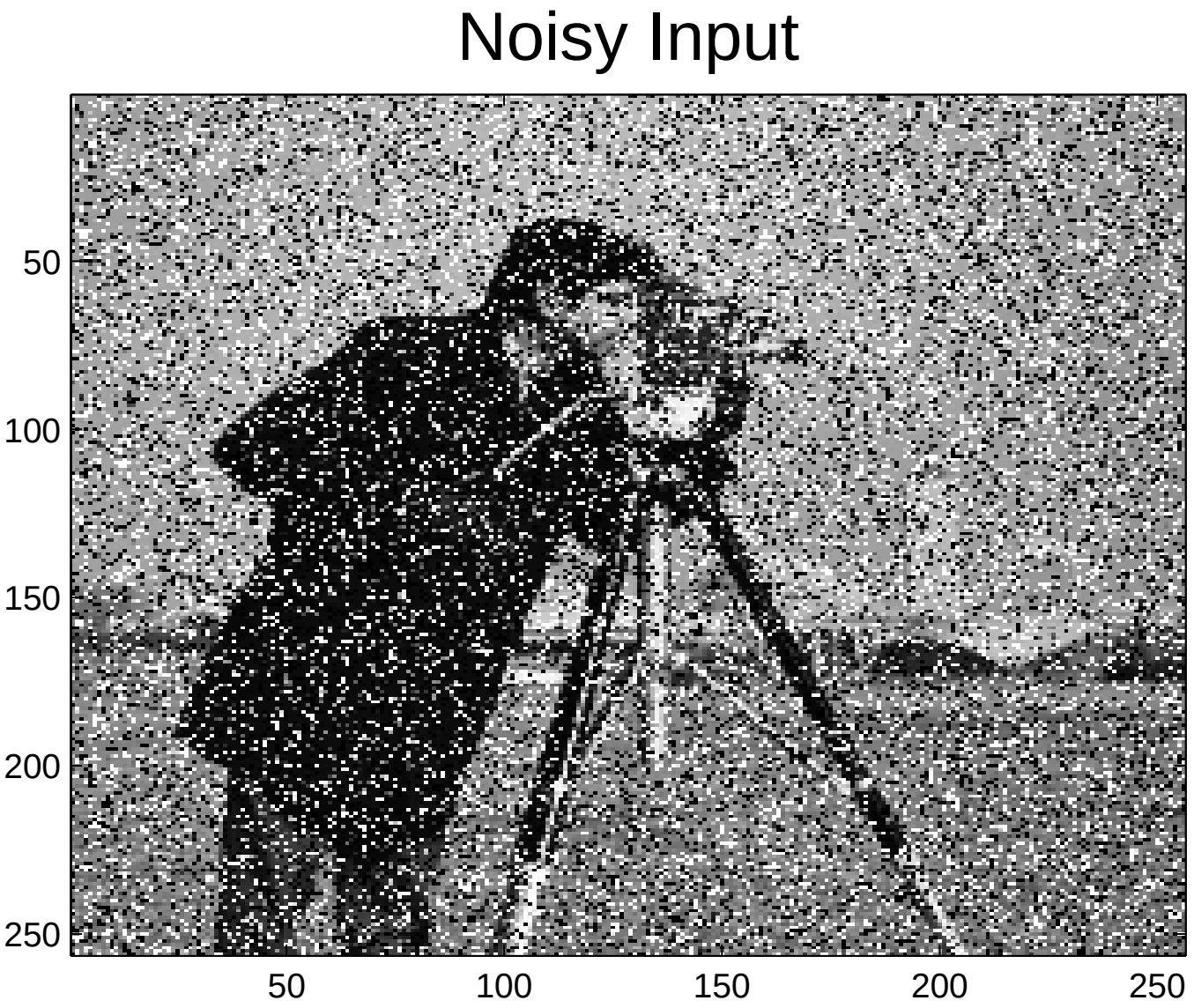
using the non-expansive property of projections and $\|A\| = \left\|A^T\right\|$. Finally,

$$\|A\|^2 \leq 4 \max_r \sum_{j=1}^{n} w_{rj}^2.$$

Total Variation - Original

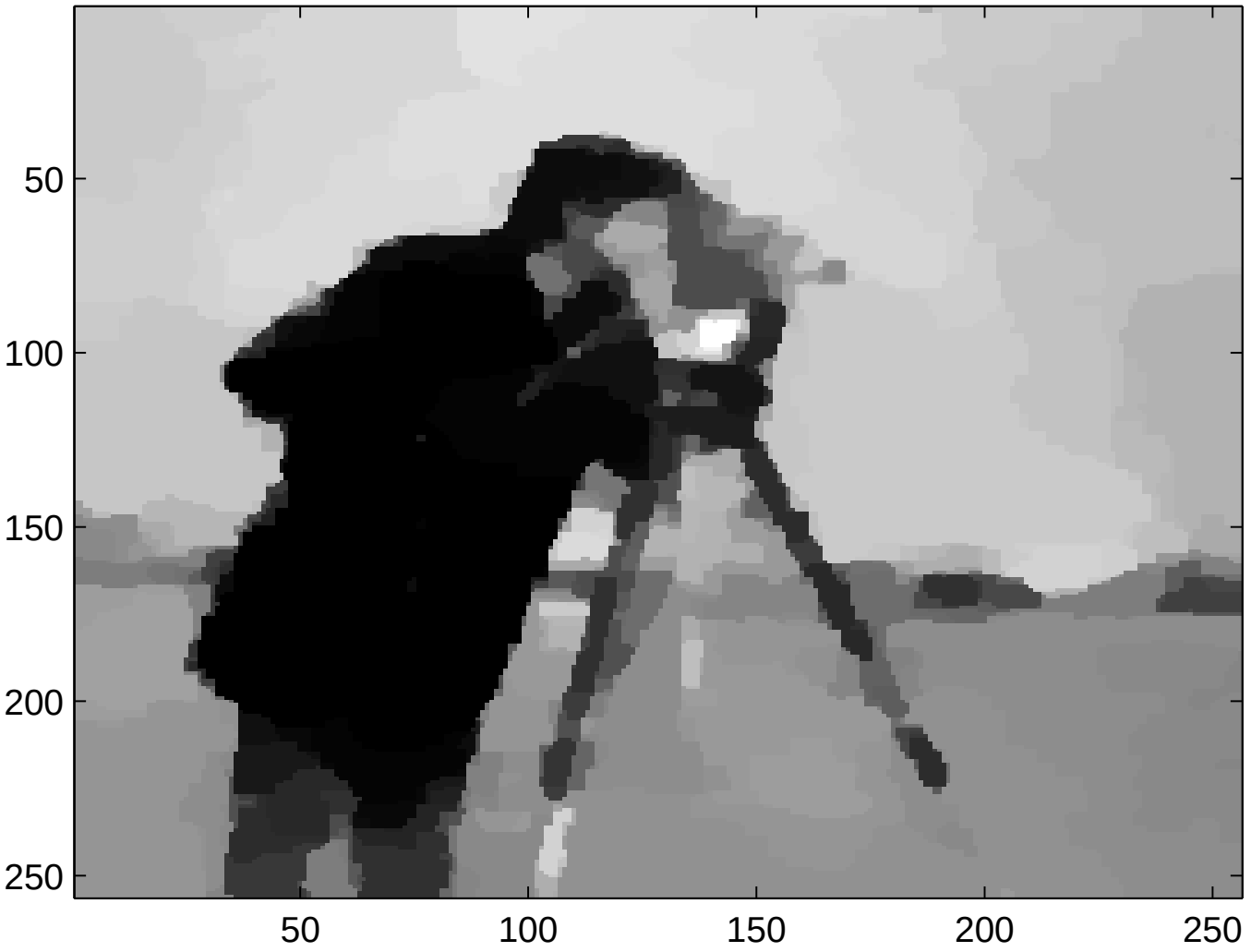
Original

Total Variation - Noisy Input



Total Variation - Denoised Image

λ=0.9



Total Variation - All