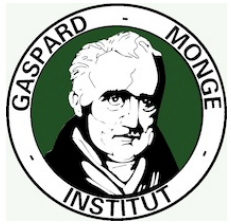
Cut Pursuit : Fast algorithms to learn piecewise constant functions on general weighted graphs

Guillaume Obozinski

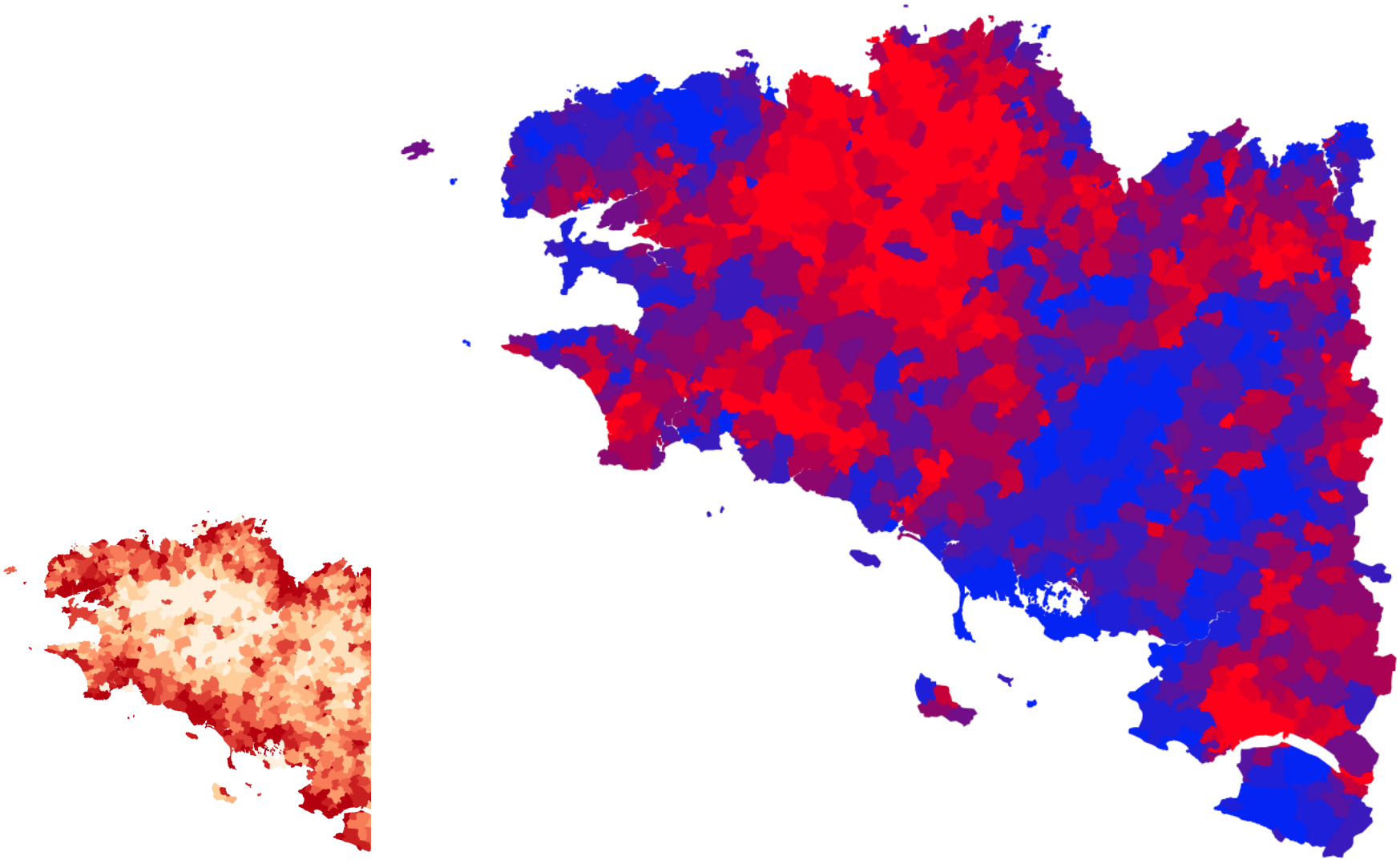
Ecole des Ponts - ParisTech
Université Paris-Est



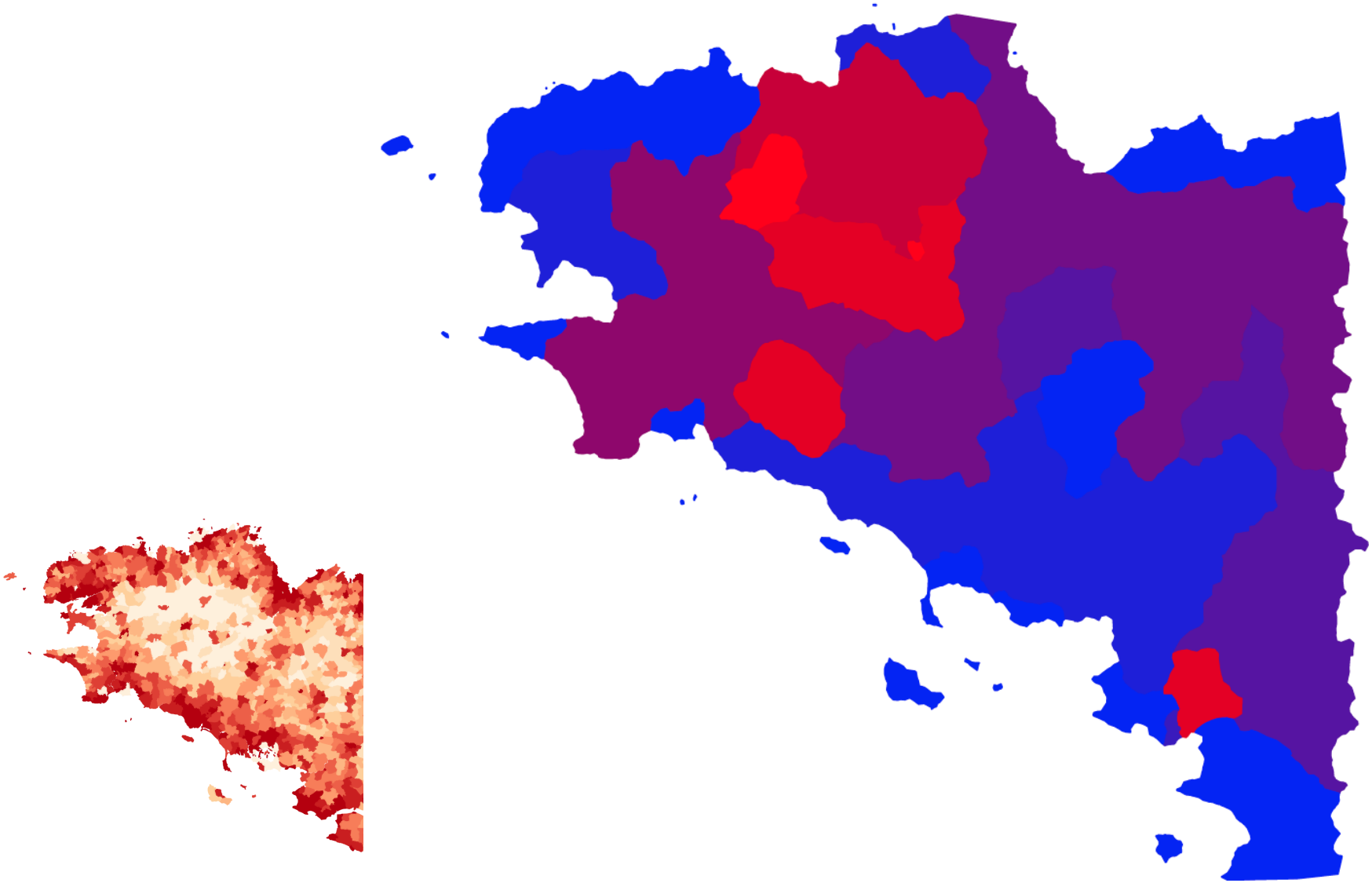
Joint work with Loïc Landrieu

SESO, June 3rd 2016

Piecewise constant models



Piecewise constant models



TV regularization on a weighted graph

Let $G = (V, E, w)$ be a weighted graph with

- $n := |V|$ nodes
- $m := |E|/2$ undirected edges modeled by oriented edges in both direction in E.
- $w_{ij} = w_{ji}$, with $w_{ij} = 0$ for $(i, j) \notin E$.

$$w(A, B) := \sum_{(i,j) \in A \times B} w_{ij}$$

For $x \in \mathbb{R}^n$

$$\mathrm{TV}(x) := \frac{1}{2} \sum_{(i,j) \in E} w_{ij} |x_i - x_j|$$

With f convex differentiable consider

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \mathrm{TV}(x)$$

Properties of the Total Variation on a graph

The Total Variation is the Lovász extension of

$$F : B \mapsto w(B, B^c)$$

and F is

- a submodular function
- measure of boundary size/*perimeter* of B

Moreover if $s \in \mathbb{R}^n$ and $s(B) := \sum_{i \in B} s_i$ then

$$\min_B s(B) + \lambda w(B, B^c)$$

is a max-flow/min cut problem (Picard and Ratliff, 1975)

Optimization with TV

- **Active contours** (Aubert et al., 2003)
- **Level-set approach :**
  Osher and Sethian (1988); Tsai and Osher (2005)
- **Proximal operator splitting :**
  Combettes and Pesquet (2008); Chambolle and Pock (2011); Couprie et al. (2013); Raguet et al. (2013); Lorenz and Pock (2015); Raguet and Landrieu (2015)
- **ROF (TV prox) as a parametric max-flow problem :**
  Chambolle and Darbon (2009); Goldfarb and Yin (2009)
- **Connections with submodularity :**
  Bach (2011); Jegelka et al. (2013); Kumar and Bach (2015)

Idea : An optimization problem over partitions of V

Write

$$x = \sum_{i=1}^{k} c_i \mathbf{1}_{A_i}$$

with

$$\Pi = \{A_1, \cdots, A_k\}$$

a partition of V into k connected components.

Let $Q(x) = f(x) + \lambda \mathrm{TV}(x)$ and define

$$x_\Pi = \underset{z \in \mathrm{span}(\Pi)}{\arg\min}\ Q(z).$$

Then TV minimization can be cast as the problem of finding an *optimal partition*

$$\Pi^\star = \underset{\Pi \in \mathcal{C}}{\arg\min}\ Q(x_\Pi).$$

Working set algorithms

- Consists in introducing variables as needed
- Know as *Column generation* in the linear programming literature
- Particularly relevant for problems regularized by sparsity inducing regularizers.

→ Way to exploit sparsity computationally

- Used in the sparsity literature :
  - Glmnet (Friedman et al., 2010)
  - Group Lasso (Obozinski et al., 2006; Roth and Fischer, 2008)
  - See also (Bach et al., 2012)
- Related to Frank-Wolfe and simplicial methods (Jaggi, 2013; Bach, 2013; Harchaoui et al., 2015)
- Exact/approximate regularization paths algorithm
  - Using warm starts
  - Exact homotopy algorithms (e.g. LARS algorithm of Efron et al., 2004)

Decomposing the objective

Appropriate notion of support : set of active edges

$$S(x) := \{(i,j) \in E \mid x_i \neq x_j\}.$$

$$\begin{cases} Q_S(x) & = f(x) + \frac{1}{2}\lambda \sum_{(i,j)\in S} w_{ij}\,|x_i - x_j|, \\ \mathrm{TV}_{|S^c}(x) & = \frac{1}{2}\lambda \sum_{(i,j)\in S^c} w_{ij}\,|x_i - x_j|. \end{cases}$$

- Q_S is differentiable
- $\mathrm{TV}_{|S^c}$ is the total variation on the graph without active edges.

TV directional derivative along cuts

Consider descent directions of the form :

$$u_B = \gamma_B \mathbf{1}_B - \gamma_{B^c} \mathbf{1}_{B^c}$$

with $\|u_B\|_2 = 1$.

Proposition

For $x \in \mathbb{R}^n$, if we set $S = S(x)$ then the directional derivative in the direction of $\mathbf{1}_B$ is

$$Q'(x, \mathbf{1}_B) = \langle \nabla Q_S(x), \mathbf{1}_B \rangle + \lambda w_{S^c}(B, B^c).$$

Moreover if $\langle \nabla f(x), \mathbf{1}_B \rangle = 0$ then

$$Q'(x, u_B) = (\gamma_B + \gamma_{B^c})\, Q'(x, \mathbf{1}_B).$$

Steepest binary cut

Define a *steepest binary cut* as any (B_Π, B_Π^c) such that

$$B_\Pi \in \arg\min_{B \subset V} \langle \nabla Q_S(x_\Pi), \mathbf{1}_B \rangle + \lambda w_{S^c}(B, B^c). \tag{1}$$

Note that since

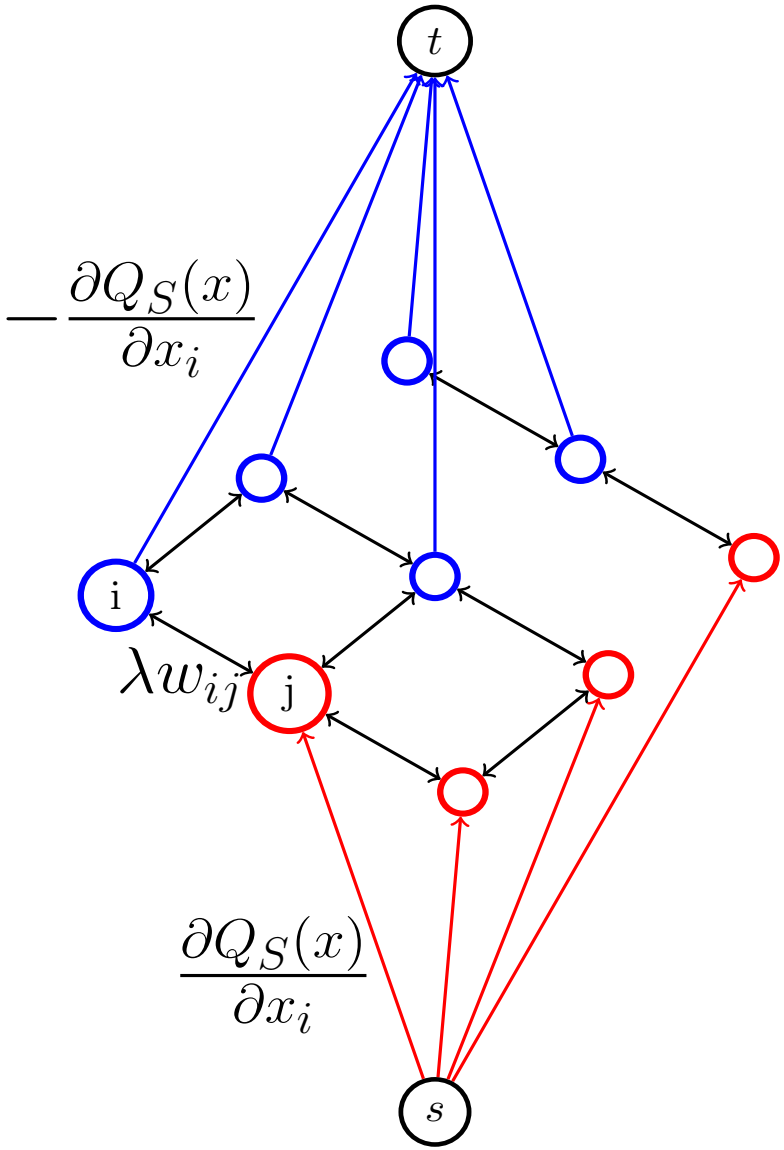
$$Q'(x, \mathbf{1}_\varnothing) = 0,$$

we have

$$\min_{B \subset V} Q'(x, \mathbf{1}_B) \leq 0.$$

If $\varnothing$ is a solution to (1), we set $B_\Pi = \varnothing$.

Max flow formulation (Picard and Ratliff, 1975)



nodes in ∇_-

nodes in ∇_+

$\longleftrightarrow$ edge in S^c

where
$\nabla_+ = \{i \in V \mid \nabla_i Q_S(x) > 0\}$ and
$\nabla_- = V \backslash \nabla_+$.

Proposition

$(C, V_{flow} \backslash C)$ is a min cut in G_{flow} if and only if B and $V \backslash B$ are minimizers of

$$B \mapsto Q'(x, \mathbf{1}_B),$$

with $B := C \backslash \{s\}$.

Characterisation of optimality via cuts

Proposition

If $\langle \nabla f(x), 1_V \rangle = 0$ *the we have that,*

$$x = \arg\min_{z \in \mathbb{R}^n} Q(z)$$

if and only if

$$\min_{B \subset V} Q'(x, \mathbf{1}_B) = 0.$$

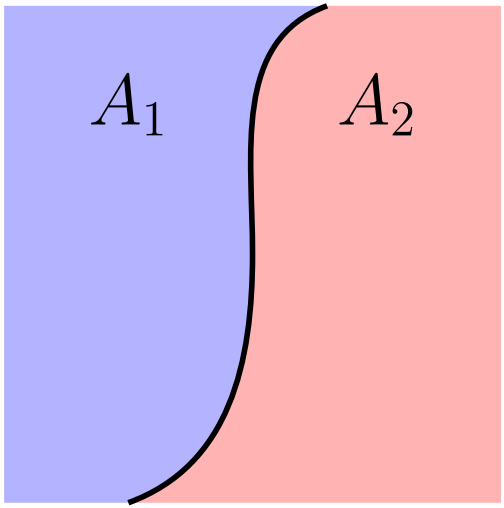
Partition update and new subspace

$$\text{Maintaining} \qquad x = \sum_{i=1}^{k} c_i \mathbf{1}_{A_i} \quad \text{with} \quad \Pi := \{A_1, \ldots, A_k\}$$

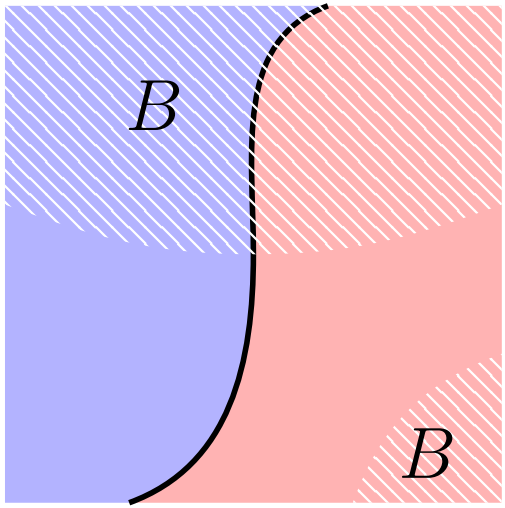
After adding $\mathbf{1}_B$, we have

$$x \in \text{span}(\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}, \mathbf{1}_B)$$

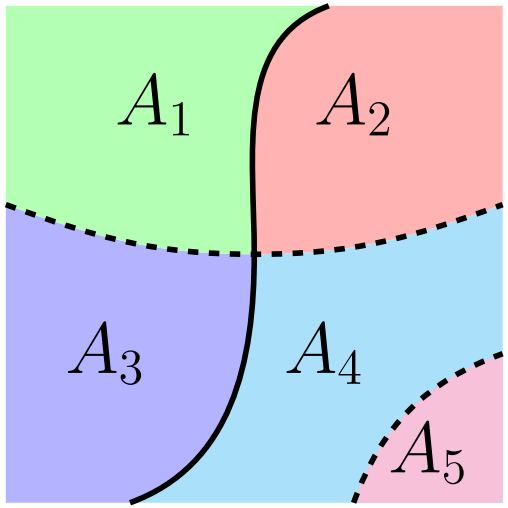
Partition update



Π



Binary cut



Π_{new}

The largest subspace $\mathcal{X}$ such that for all $x \in \mathcal{X}$

$$S(x) = S_{new} \quad \text{with} \quad S_{new} := S \cup (B \times B^c).$$

is $\operatorname{span}\big(\{\mathbf{1}_C \mid C \in \Pi_{new}\}\big)$ with

$\Pi_{new} := \{C \mid \exists A \in \Pi \text{ s.t. } C \text{ is a connected comp. of } A \cap B \text{ or } A \cap B^c\}$

Cut-Pursuit

Algorithm 1: Cut Pursuit

Initialize $\Pi \leftarrow \{V\}$

$x_\Pi \in \arg\min_{z=c\mathbf{1}_V, c\in\mathbb{R}} Q(z)$

while $\min_{B\subset V} Q'(x_\Pi, \mathbf{1}_B) < 0$ **do**

  Pick $B \in \arg\min_{B'\subset V} Q'(x_\Pi, \mathbf{1}_{B'})$

  $\Pi \leftarrow \{B \cap A\}_{A\in\Pi} \cup \{B^c \cap A\}_{A\in\Pi}$

  $\Pi \leftarrow$ connected components of elements of Π

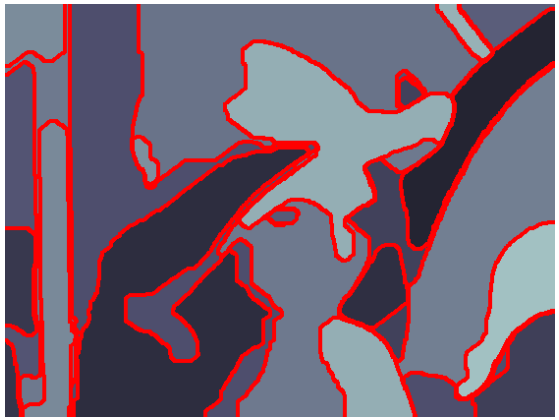
  Pick $x_\Pi \in \arg\min_{z\in \mathrm{span}(\Pi)} Q(z)$

return (Π, x_Π)

Illustration of the algorithm on Lena

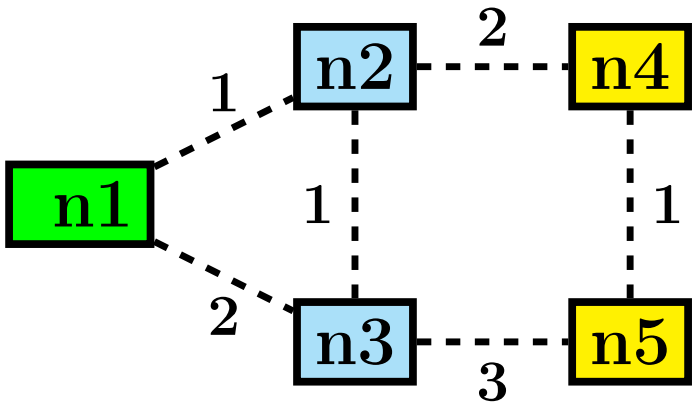




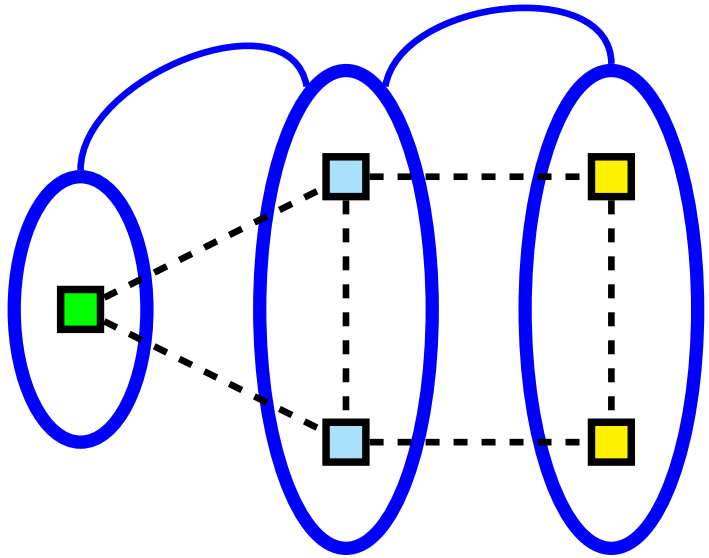


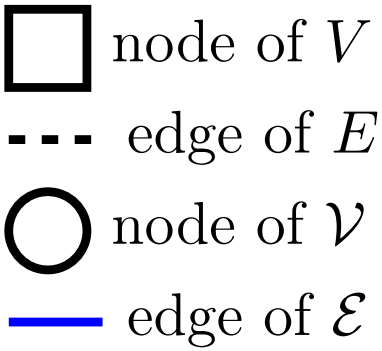
Reduced graph

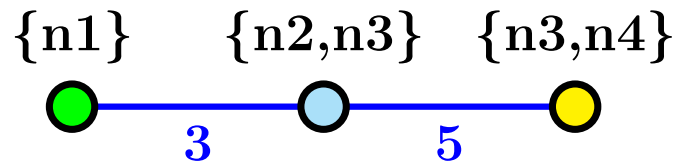
Original graph G



Reduced graph $\mathcal{G}$







Reduced graph

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with

$$\begin{cases} \mathcal{V} = \Pi \\ \mathcal{E} = \{(A, B) \in \mathcal{V}^2 \mid \exists (i, j) \in (A \times B) \cap E\} \end{cases}$$

Proposition

For $x = \sum_{A \in \Pi} c_A \mathbf{1}_A$ *we have* $\mathrm{TV}_G(x) = \mathrm{TV}_{\mathcal{G}}(c)$ *with*

$$\mathrm{TV}_{\mathcal{G}}(c) := \frac{1}{2} \sum_{(A,B) \in \mathcal{E}} w(A, B) \, |c_A - c_B|.$$

The case of deblurring

With H the blur operator,

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2} \left\| Hx - y \right\|^2 + \lambda \mathrm{TV}_G(x)$$

With $K = [\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}]$ and $x = Kc$ we then solve

$$\min_{c \in \mathbb{R}^k} \ \frac{1}{2} \left\| HKc - y \right\|^2 + \lambda \mathrm{TV}_{\mathcal{G}}(c)$$

And

$$\nabla_c \big(\tfrac{1}{2} \left\| HKc - y \right\|^2 \big) = K^\intercal H^\intercal HKc - K^\intercal Hy.$$

$K^\intercal H^\intercal HK$ and $K^\intercal Hy$ can be computed in $\mathcal{O}(k^2 \, n \log n)$ time using FFTs.

Complexity

(a) Cost of solving the **min cut/max flow problem** to obtain a steep binary cut. The algorithm of Boykov et al. (2001) has worst case exponential complexity but scales in practice linearly with the graph.

(b) Cost of building the **reduced graph**. Requires

- Computing connected components
- All weights $w(A, B)$

Done in $O(m + n)$ iterations.

(c) Cost of solving a **reduced problem** with k nodes. For deblurring :

- Cost of computation of the reduced Hessian : $O(k^2 n \log n)$
- Cost of computation of the reduced gradient : $O(k^2)$
- GFB reached a ε primal gap in $O(1/\varepsilon)$ iterations.

(d) **Number of global iterations** needed. In the worse case n iterations. In practice the partition grows exponentially with the number of cuts.

Experiments

Comparing

PGFB Preconditioned Generalized Forward-Backward of Raguet and Landrieu (2015). PGFB improves over GFB (Raguet et al., 2013) which was shown to outperform Chambolle and Pock (2011) and Lorenz and Pock (2015) on deblurring problems.

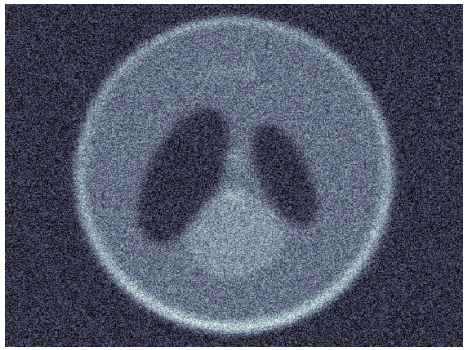
FB+ Forward backward with TV proximal operator computed as the solution of a parametric max-flow using the code of Chambolle and Darbon (2009).

CP Cut Pursuit

CPFW Cut Pursuit with *steepest cut* replaced by the FW direction of Harchaoui et al. (2015). Equivalent to fully corrective Frank-Wolfe in the generalization of Harchaoui et al. (2015).

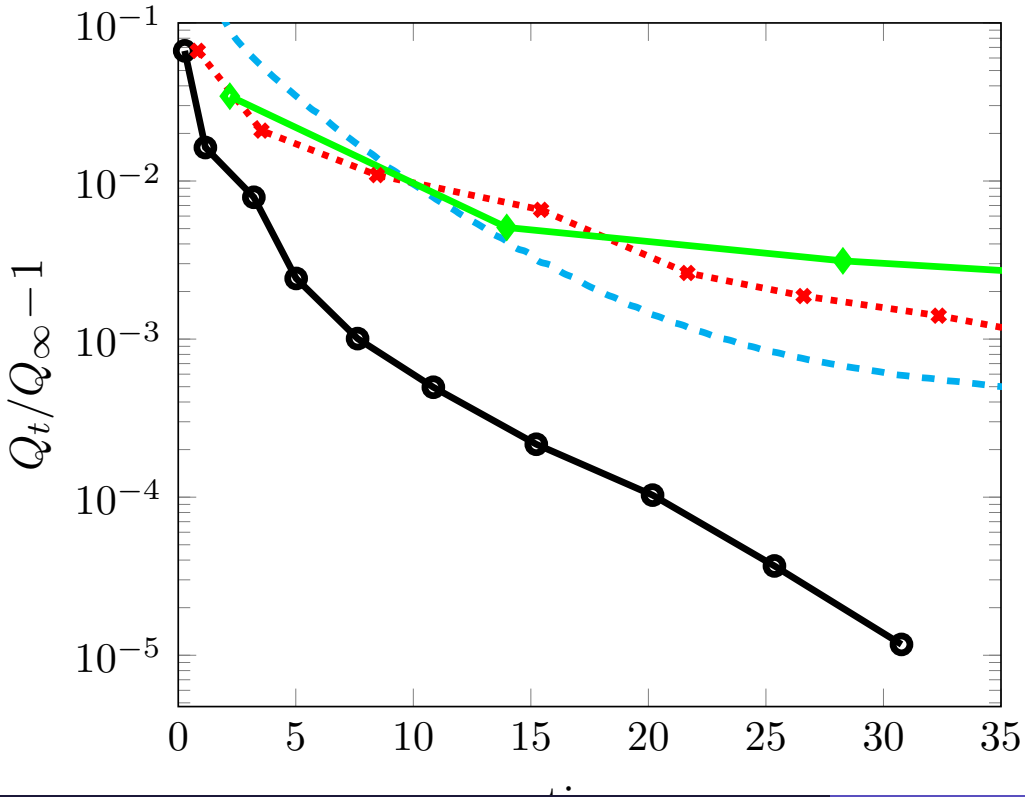
Deblurring experiments

Original

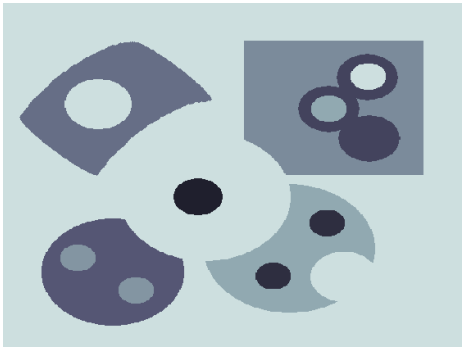


PSNR : 12.1

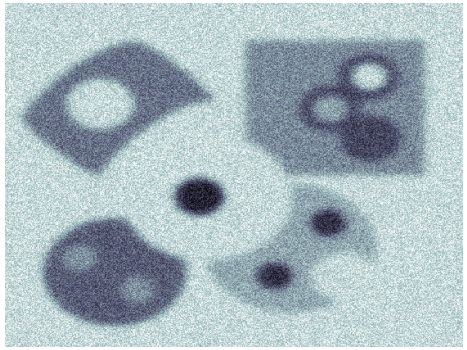
PSNR : 20.1



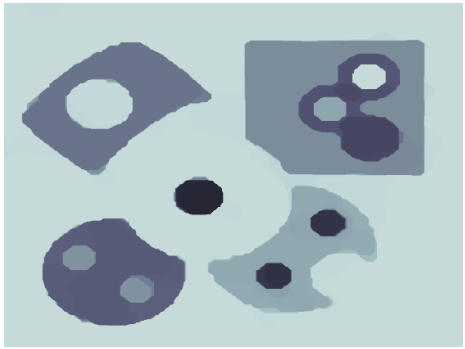
Deblurring experiments



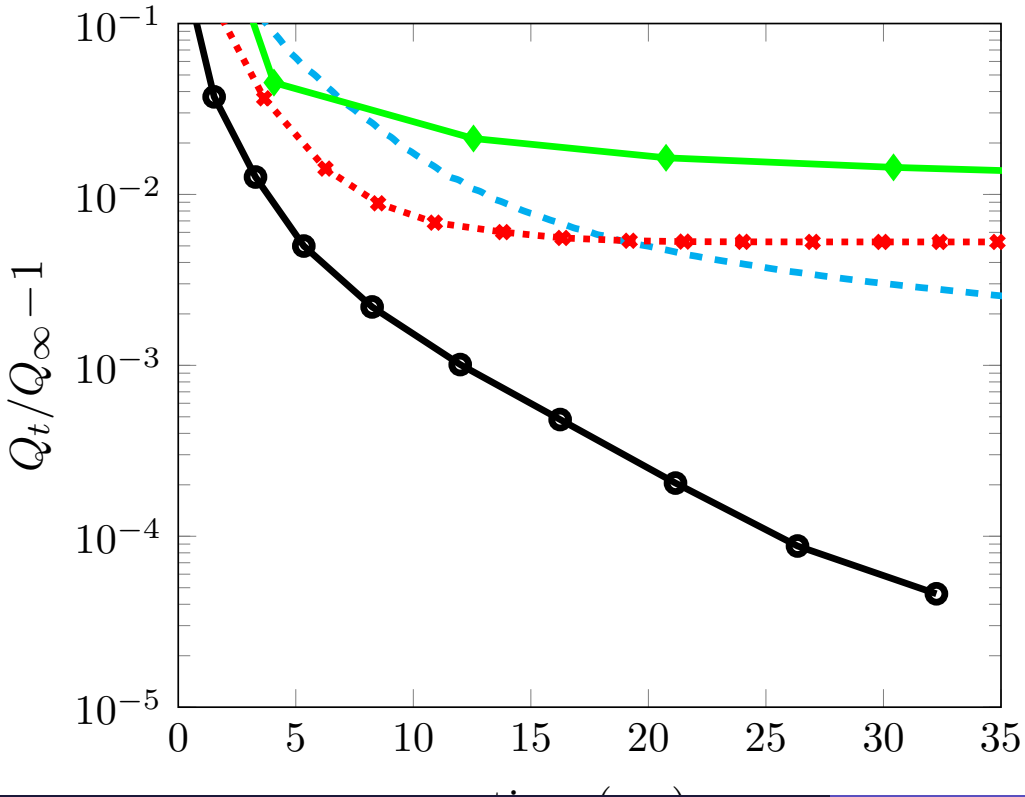
Original

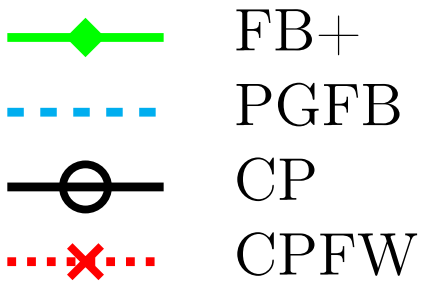


PSNR : 15.9



PSNR : 27.2



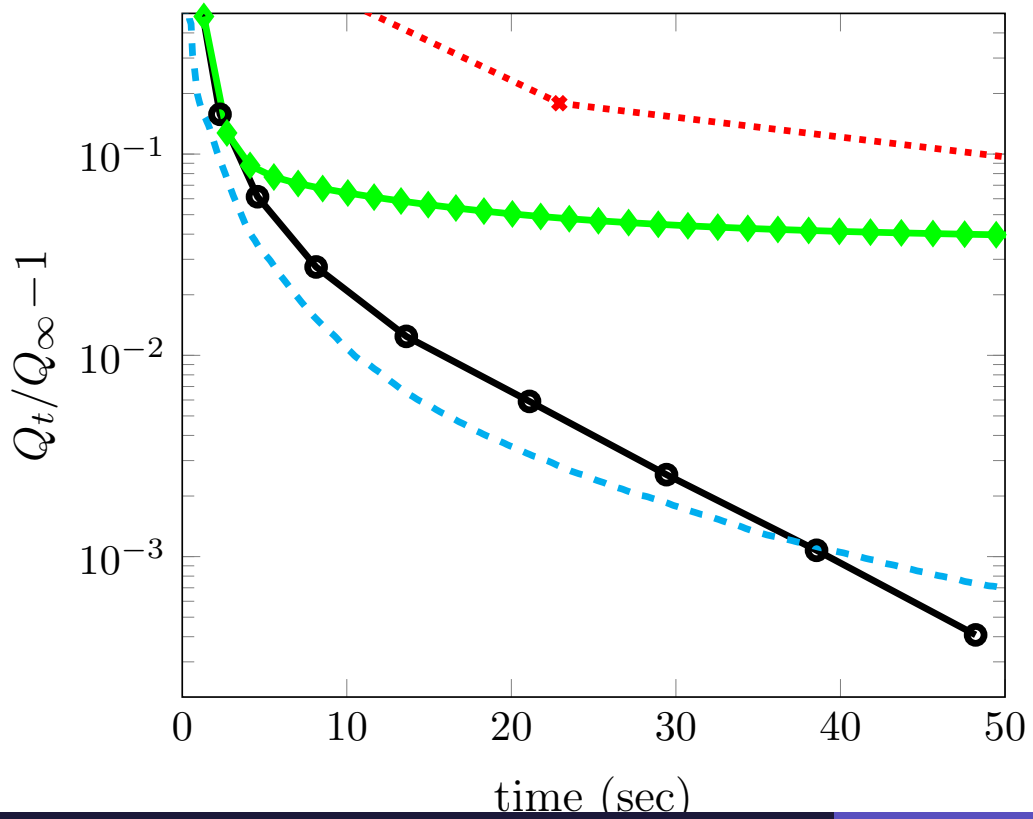


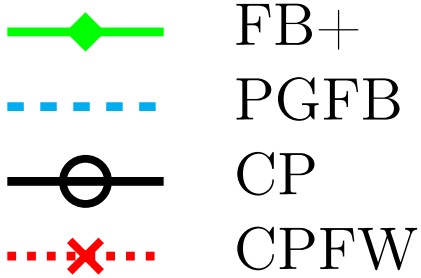
Deblurring experiments

Original

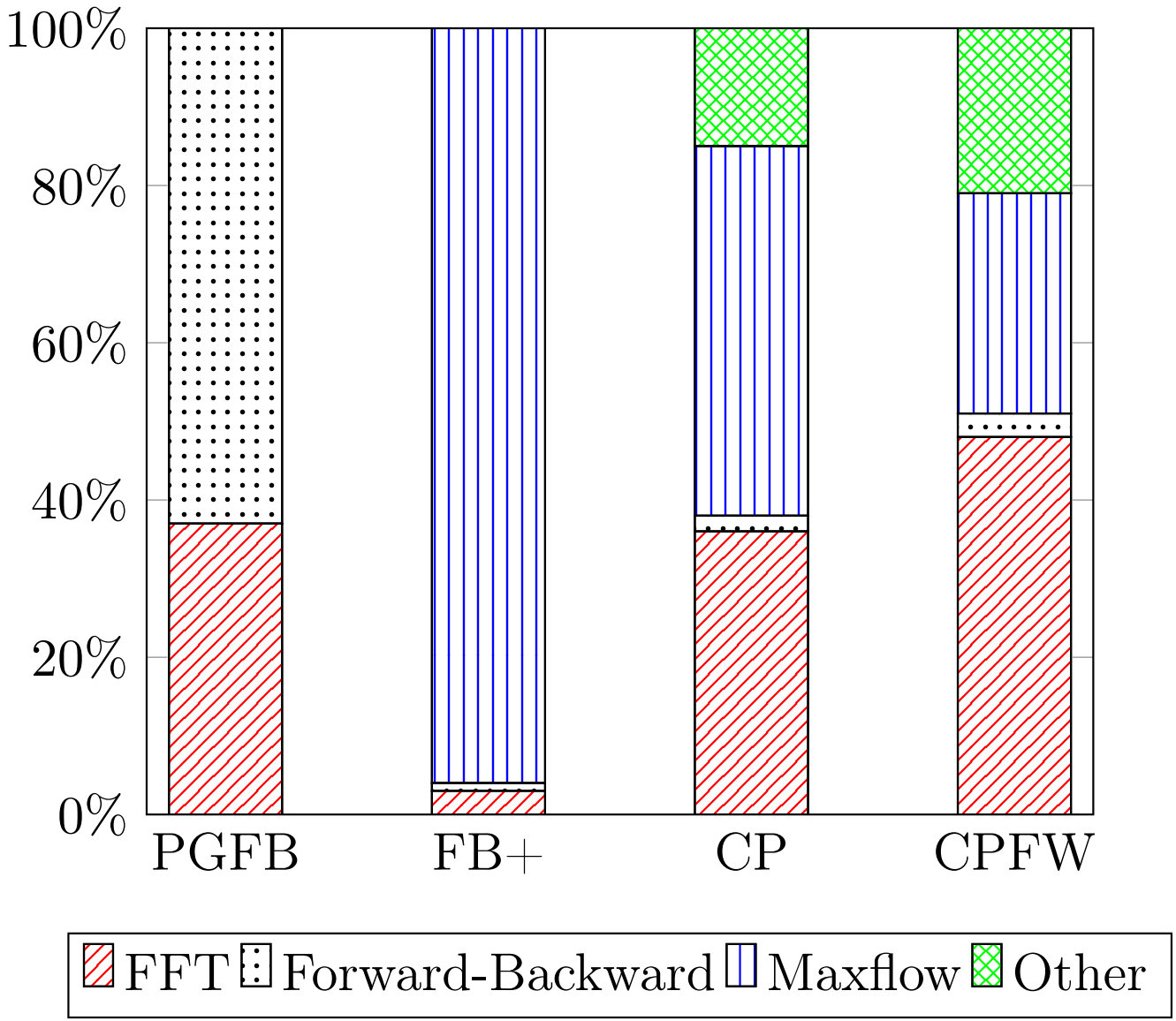
PSNR : 23.3

PSNR : 24.5

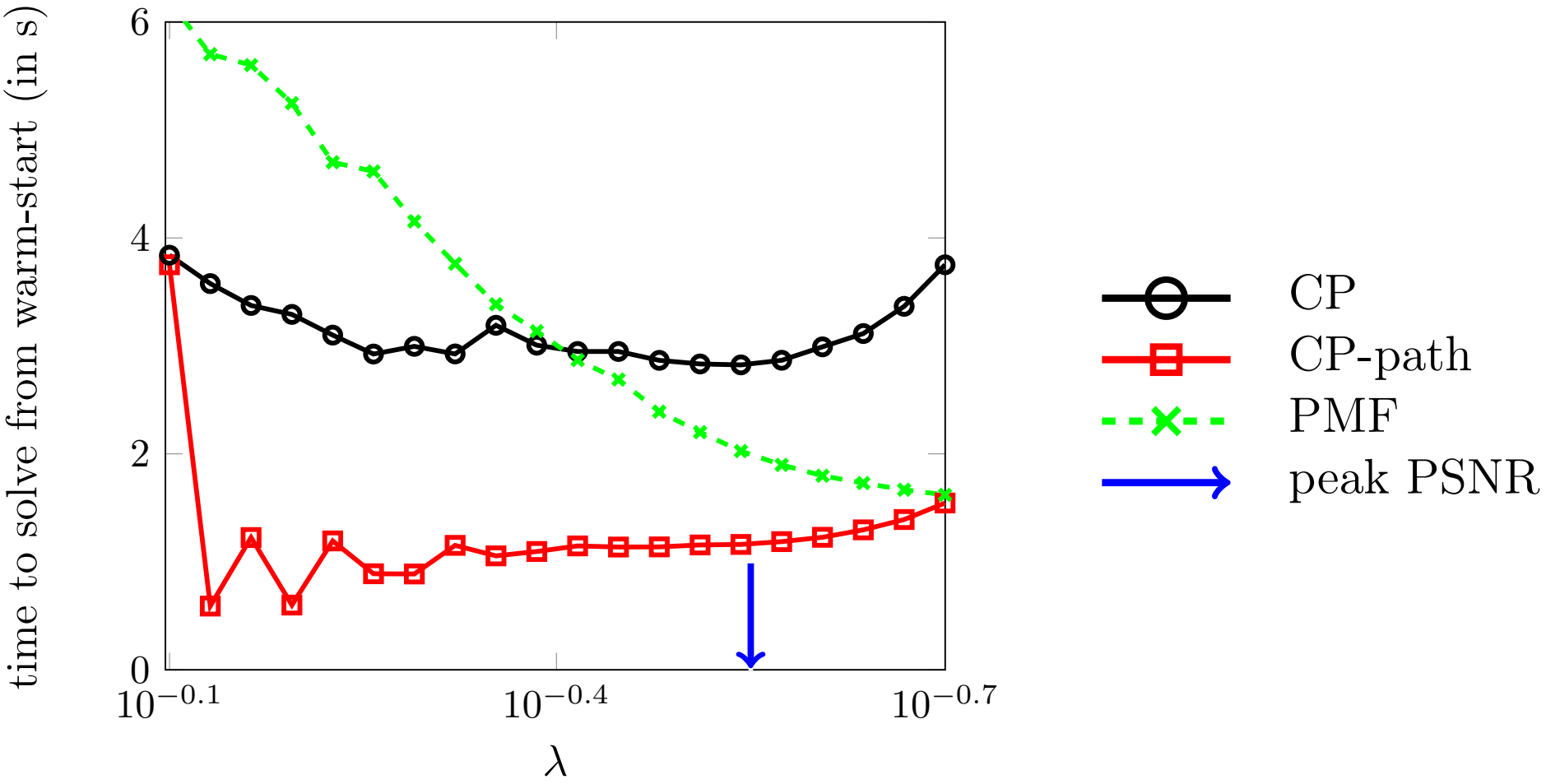




Time breakdown



Approximate regularization paths



Generalized minimal partition problem

Generalized minimal partition problem

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \Gamma(x),$$

with

$$\begin{cases} f(x) = \sum_{i \in V} f_i(x_i) \\ \Gamma(x) = \sum_{(i,j) \in S(x)} w_{ij} \end{cases}$$

with $f_i : \mathbb{R} \mapsto \mathbb{R}$ continuously differentiable and convex

Greedy algorithm to solve the regularized ℓ_0 problem

OMP, Orthogonal least squares (OLS), FoBa, and CoSamp implicitely tackle

$$\min_x f(x) \text{ s.t. } \|x\|_0 \leq k$$

For the regularized problem

$$\min_x f(x) + \lambda \|x\|_0$$

- Single best replacement (SBR) by Soussen et al. (2011)
- Single Maximum Likelihood Replacement (SMLR) of Kormylo and Mendel (1982)

Adds or removes the variable that decrease the most the objective after solving the OLS problem.

Forward step : greedy split by binary cut

Consider x of the form $x = h\mathbf{1}_B + h'\mathbf{1}_{B^c}$. Since

$$\Gamma(h\mathbf{1}_B + h'\mathbf{1}_{B^c}) = \Gamma(\mathbf{1}_B) = w(B, B^c),$$

we need to solve a problem of the form

$$\min_{B \subset V} \min_{h,h' \in \mathbb{R}} \sum_{i \in B} f_i(h) + \sum_{i \in B^c} f_i(h') + \lambda w(B, B^c)$$

Minimization w.r.t. B is obtained as the solution of a max-flow problem in the graph $(V \cup \{s, t\}, E_{flow})$ with

$$E_{flow} = \begin{cases} (s, i), \forall i \in \nabla_+, & \text{with} \quad c_{si} = f_i(h) - f_i(h'), \\ (i, t), \forall i \in \nabla_-, & \text{with} \quad c_{it} = f_i(h') - f_i(h), \\ (i, j), \forall (i, j) \in E, & \text{with} \quad c_{ij} = \lambda w_{ij}, \end{cases}$$

where $\nabla_+ = \{i \in V \mid f_i(h) > f_i(h')\}$ and $\nabla_- = V \backslash \nabla_+$.

Backward step

Let $\Pi_-(A,B) := \Pi \setminus \{A,B\} \cup \{A \cup B\}$ and

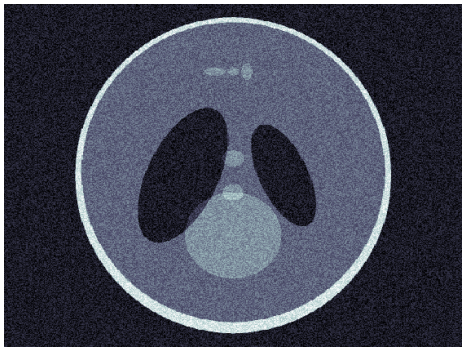
$$\delta_-(A,B) := f(x_\Pi) - f(x_{\Pi_-(A,B)}) + \lambda w(A,B).$$

Simple merge

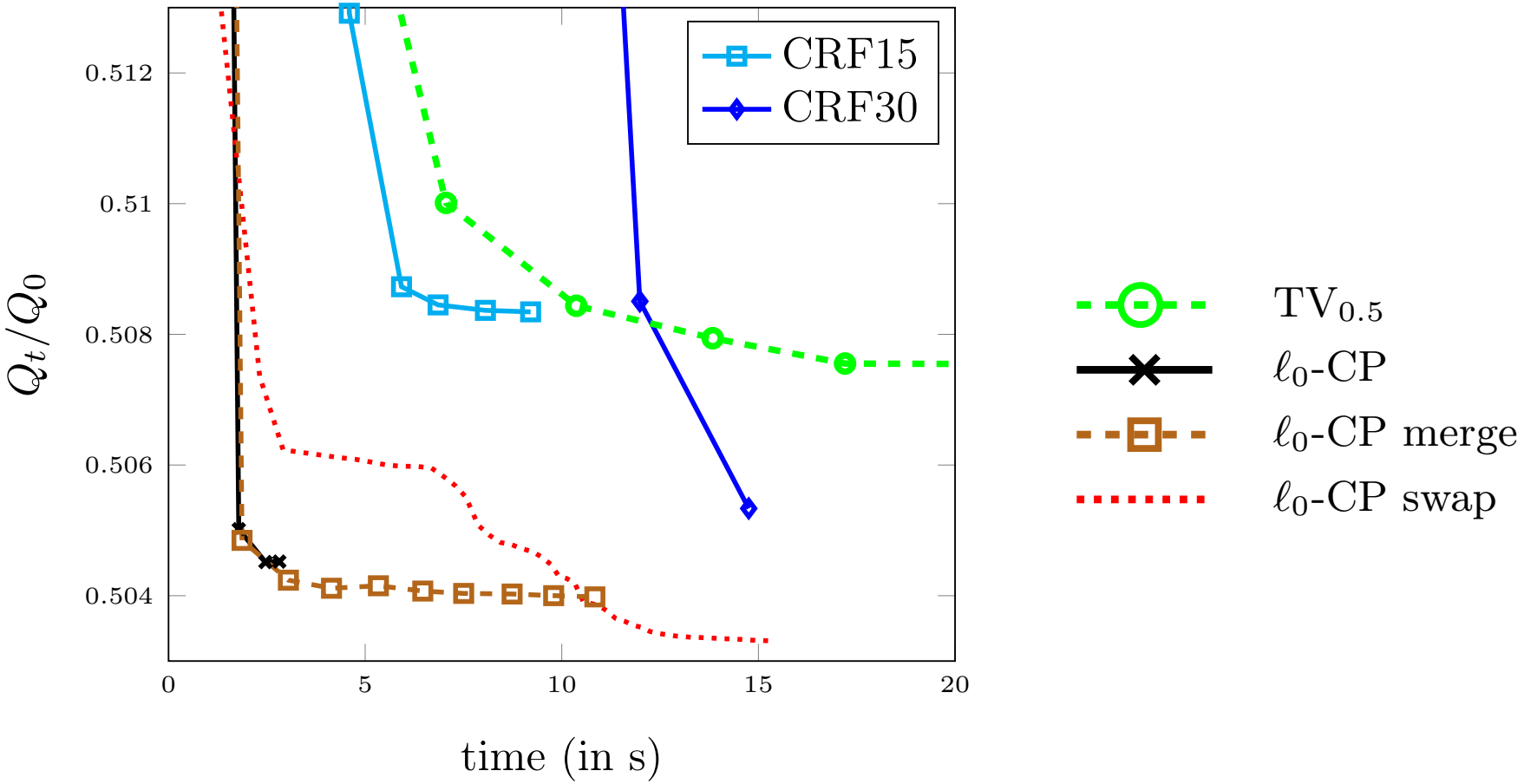
1. Take the pair (A,B) with maximal $\delta_-(A,B)$
2. If $\delta_-(A,B) \geq 0$ then set $\Pi_{new} = \Pi_-(A,B)$.

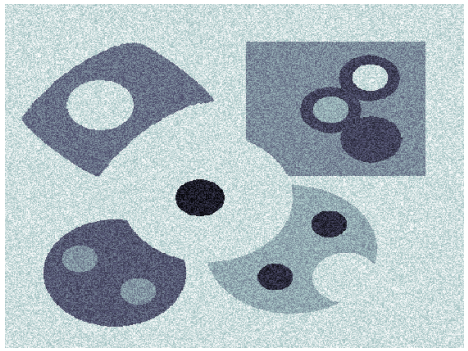
Merge-resplit

1. $$C \leftarrow \operatorname*{arg\,min}_{C \subset A \cup B} \sum_{i \in C} f_i(x_A) + \sum_{i \in A \cup B \setminus C} f_i(x_B) + \lambda\, w(C, A \cup B \setminus C)$$
2. Replace A and B in Π by the the connected components of C

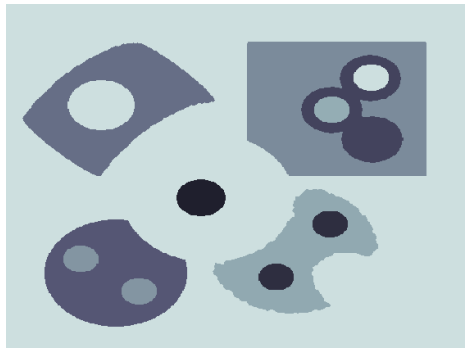


Noisy (PSNR : 24.8) ℓ_0-CPm (PSNR : 38.1)

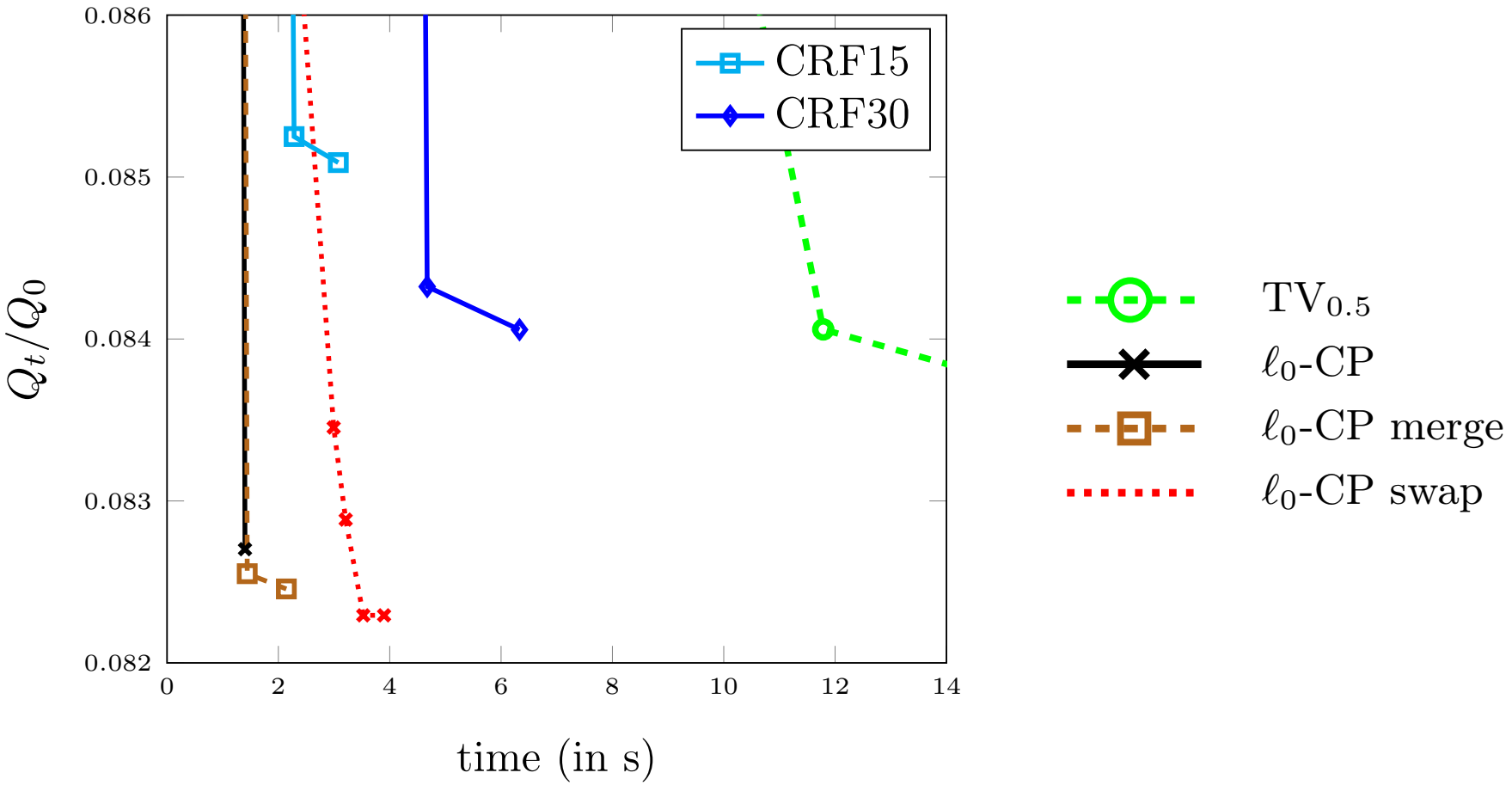


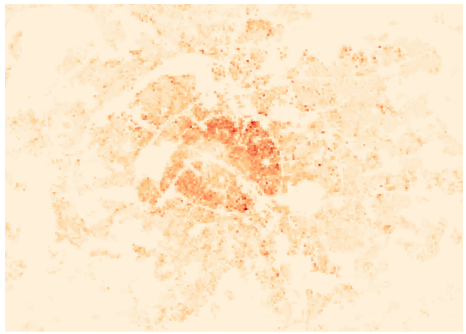


Noisy (PSNR : 18.8)

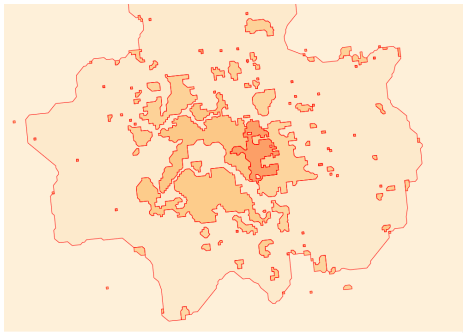


ℓ_0-CPm (PSNR = 34.8)

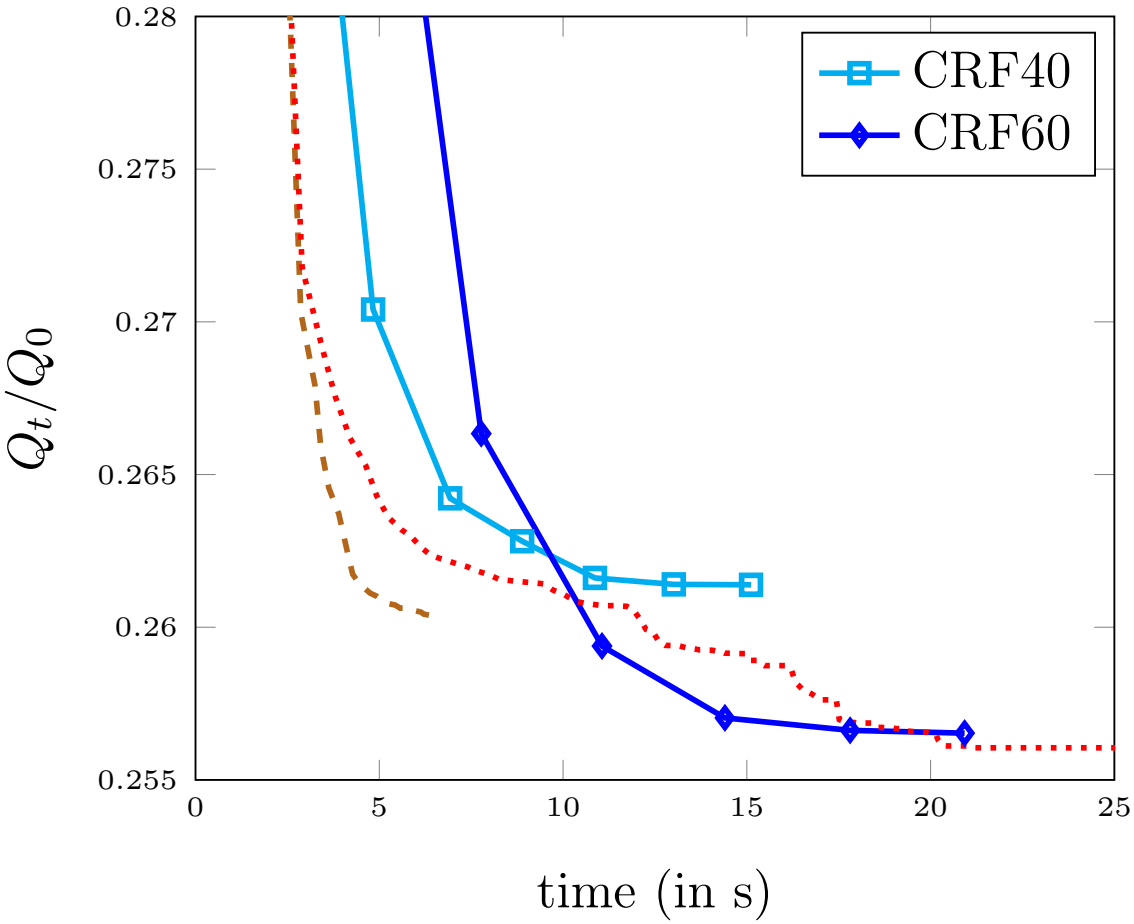




Population density



ℓ_0-CPm



ℓ_0-CP merge

ℓ_0-CP swap

Conclusions

Exploiting the relation between two forms of sparsity

- Short total perimeter/boundary size
- Coarse partition : small number of level sets

Improves over previous approaches by

- Solving a reduced problem on a reduced graph
- Choose cuts optimally based on the directional derivative
- Removing coupling between atoms

Allows warm-starts and approximate regularization path computations.

Future work

- Extension to other submodular/combinatorial functions
- Extension to ℓ_2-TV and multivariate TV ?
- Guarantees on the number of iterations under SNR + graph structure assumptions ?
- Guarantees like α-expansions in the greedy case ?

References I

Aubert, G., Barlaud, M., Faugeras, O., and Jehan-Besson, S. (2003). Image segmentation using active contours : calculus of variations or shape gradients ? *SIAM Journal on Applied Mathematics*, 63(6) :2128–2154.

Bach, F. (2010). Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126.

Bach, F. (2013). Learning with submodular functions : a convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3) :145–373.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1) :1–106.

Bach, F. R. (2011). Shaping level sets with submodular functions. In *Advances in Neural Information Processing Systems*, pages 10–18.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Efficient approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12) :1222–1239.

Chambolle, A. and Darbon, J. (2009). On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3) :288–307.

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1) :120–145.

References II

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6) :805–849.

Combettes, P. L. and Pesquet, J.-C. (2008). A proximal decomposition method for solving convex variational inverse problems. *Inverse problems*, 24(6) :65014–65040.

Couprie, C., Grady, L., Najman, L., Pesquet, J.-C., and Talbot, H. (2013). Dual constrained tv-based regularization on graphs. *SIAM Journal on Imaging Sciences*, 6(3) :1246–1273.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2) :407–499.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22.

Goldfarb, D. and Yin, W. (2009). Parametric maximum flow algorithms for fast total variation minimization. *SIAM Journal on Scientific Computing*, 31(5) :3712–3743.

Harchaoui, Z., Juditsky, A., and Nemirovski, A. (2015). Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1–2) :75–112.

Jaggi, M. (2013). Revisiting Frank-Wolfe : projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435.

References III

Jegelka, S., Bach, F., and Sra, S. (2013). Reflection methods for user-friendly submodular optimization. In *Advances in Neural Information Processing Systems*, pages 1313–1321.

Kormylo, J. J. and Mendel, J. M. (1982). Maximum likelihood detection and estimation of Bernoulli-Gaussian processes. *IEEE Transactions on Information Theory*, 28(3) :482–488.

Kumar, K. and Bach, F. (2015). Active-set methods for submodular optimization. *arXiv preprint arXiv :1506.02852*.

Lorenz, D. A. and Pock, T. (2015). An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2) :311–325.

Obozinski, G. and Bach, F. (2012). Convex relaxation for combinatorial penalties. *arXiv preprint arXiv :1205.1240*.

Obozinski, G., Taskar, B., and Jordan, M. (2006). Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep.*

Osher, S. and Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed : algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1) :12–49.

Picard, J.-C. and Ratliff, H. D. (1975). Minimum cuts and related problems. *Networks*, 5(4) :357–370.

Raguet, H., Fadili, J., and Peyré, G. (2013). A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3) :1199–1226.

References IV

Raguet, H. and Landrieu, L. (2015). Preconditioning of a generalized forward-backward splitting and application to optimization on graphs. *SIAM Journal on Imaging Sciences*, 8(4) :2706–2739.

Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear models : uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM.

Soussen, C., Idier, J., Brie, D., and Duan, J. (2011). From Bernoulli–Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10) :4572–4584.

Tsai, Y.-H. R. and Osher, S. (2005). Total variation and level set methods in image science. *Acta Numerica*, 14 :509–573.