Microsoft Research

NIPS

On Decomposing the Proximal Map

Yao-Liang Yu

University of Alberta

December 6, 2013

Regularized loss minimization

Generic form for many ML problems:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \ell(\mathbf{w}) + f(\mathbf{w})$$

- ℓ is the loss function;
- f is the regularizer, usually a (semi)norm;

Special interest:

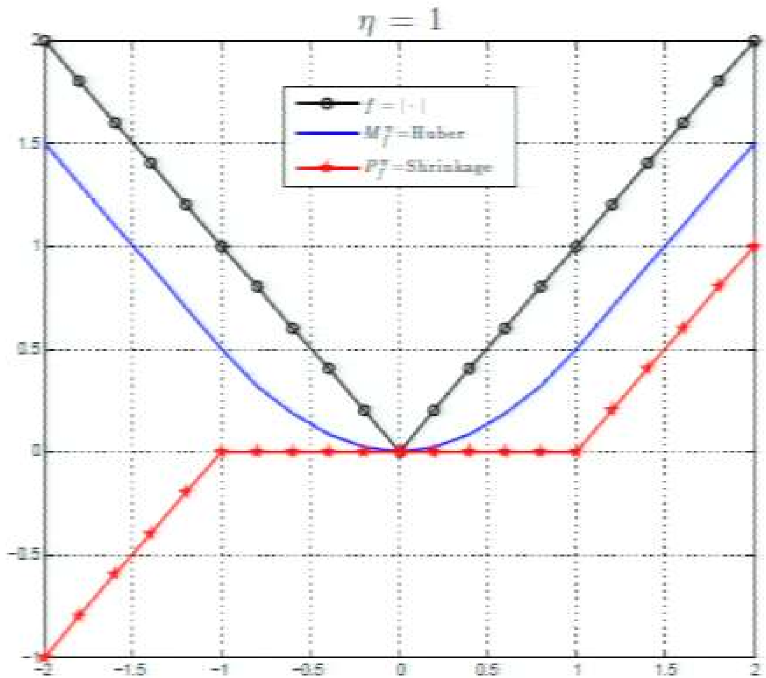
- sparsity;
- computational efficiency.

Moreau envelop and proximal map

Definition (Moreau'65)

$$\mathsf{M}_f(\mathbf{y}) = \min_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{w} - \mathbf{y}\|^2 + f(\mathbf{w})$$

$$\mathsf{P}_f(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{w} - \mathbf{y}\|^2 + f(\mathbf{w})$$



Proximal gradient (Fukushima & Mine'81)

$$\min_{\mathbf{w} \in \mathbb{R}^d} \ell(\mathbf{w}) + f(\mathbf{w})$$

1. $\mathbf{y}_t = \mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t)$;
2. $\mathbf{w}_{t+1} = \mathsf{P}_{\eta f}(\mathbf{y}_t)$.

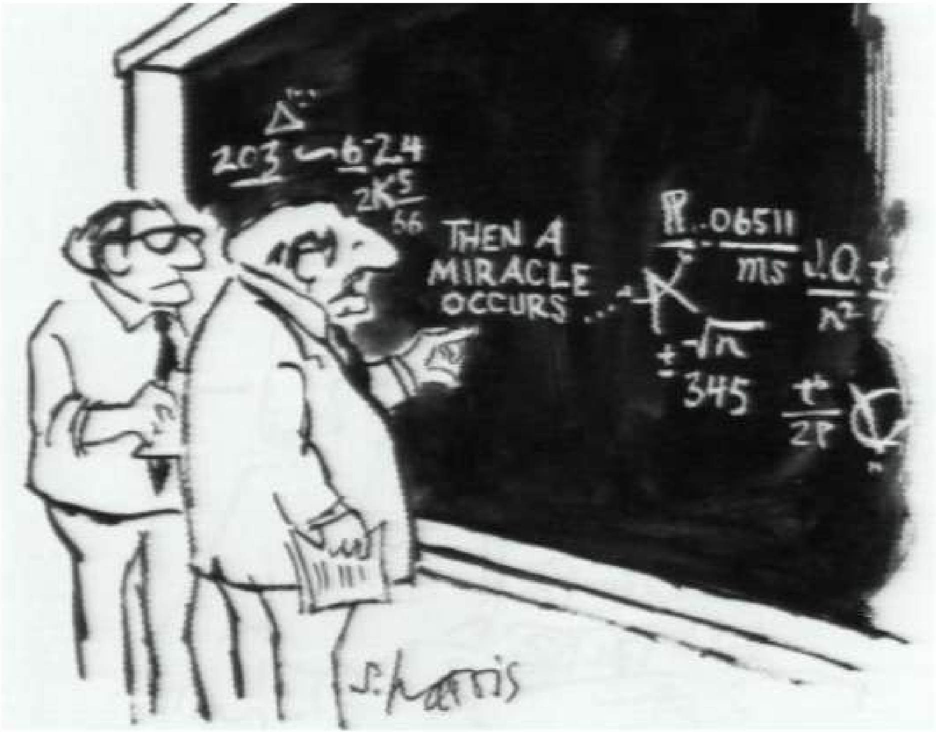
For $f = \|\cdot\|_1$, obtain the shrinkage operator

$$[\mathsf{P}_{\|\cdot\|_1}(\mathbf{y})]_i = \mathrm{sign}(y_i)(|y_i| - 1)_+.$$

- guaranteed convergence, can be accelerated;
- generalization of projected gradient: $f = \iota_C$;
- reveals the sparsity-inducing property.

Refs: Combettes & Wajs'05; Beck & Teboulle'09; Duchi & Singer'09; Nesterov'13; etc.

Then A Miracle Occurs...



"I think you should be more explicit here in step two."

from *What's so Funny about Science?* by Sidney Harris (1977)

Step 2: $\mathsf{P}_f(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2}\|\mathbf{y}-\mathbf{w}\|^2 + f(\mathbf{w})$

How to decompose?

- Typical structured sparse regularizers:

$$f(\mathbf{w}) = \sum_i f_i(\mathbf{w});$$

Theorem (Parallel Sum)

$$\mathsf{P}_{f+g} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id}).$$

- Not directly useful due to the inversion;
- Can numerically reduce to P_f and P_g (Combettes et al.'11);
- But a two-loop routine can be as slow as subgradient descent (Schmidt et.al'11; Villa et al.'13).

How to decompose?

- Typical structured sparse regularizers:

$$f(\mathbf{w}) = \sum_i f_i(\mathbf{w});$$

Theorem (Parallel Sum)

$$\mathsf{P}_{f+g} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id}).$$

- Not directly useful due to the inversion;
- Can numerically reduce to P_f and P_g (Combettes et al.'11);
- But a two-loop routine can be as slow as subgradient descent (Schmidt et.al'11; Villa et al.'13).

How to decompose?

- Typical structured sparse regularizers:

$$f(\mathbf{w}) = \sum_i f_i(\mathbf{w});$$

Theorem (Parallel Sum)

$$\mathsf{P}_{f+g} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id}).$$

- Not directly useful due to the inversion;
- Can numerically reduce to P_f and P_g (Combettes et al.'11);
- But a two-loop routine can be as slow as subgradient descent (Schmidt et.al'11; Villa et al.'13).

How to decompose?

- Typical structured sparse regularizers:

$$f(\mathbf{w}) = \sum_i f_i(\mathbf{w});$$

Theorem (Parallel Sum)

$$\mathsf{P}_{f+g} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id}).$$

- Not directly useful due to the inversion;
- Can numerically reduce to P_f and P_g (Combettes et al.'11);
- But a two-loop routine can be as slow as subgradient descent (Schmidt et.al'11; Villa et al.'13).

Two previous results

Theorem (Friedman et al.'07)

$$\mathsf{P}_{\|\cdot\|_1+\|\cdot\|_{\mathrm{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathrm{TV}}}, \quad \textit{where} \quad \|\mathbf{w}\|_{\mathrm{TV}} = \sum_{i=1}^{d-1} |w_i - w_{i+1}|.$$

Theorem (Jenatton et al.'11)

Assuming the groups $\{\mathrm{g}_i\}$ *form a laminar system* ($\mathrm{g}_i \cap \mathrm{g}_j \in \{\mathrm{g}_i, \mathrm{g}_j, \emptyset\}$)*, then, if appropriately ordered,*

$$\mathsf{P}_{\sum_{i=1}^k \|\cdot\|_{\mathrm{g}_i}} = \mathsf{P}_{\|\cdot\|_{\mathrm{g}_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{\mathrm{g}_k}},$$

where $\|\cdot\|_{\mathrm{g}_i}$ *is the restriction of* $\mathfrak{l}_p, p \in \{1, 2, \infty\}$ *to the group* g_i.

Generalization

$$\mathsf{P}_{f+g} \stackrel{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \stackrel{?}{=} \mathsf{P}_g \circ \mathsf{P}_f.$$

But, is it even sensible?

Bad news

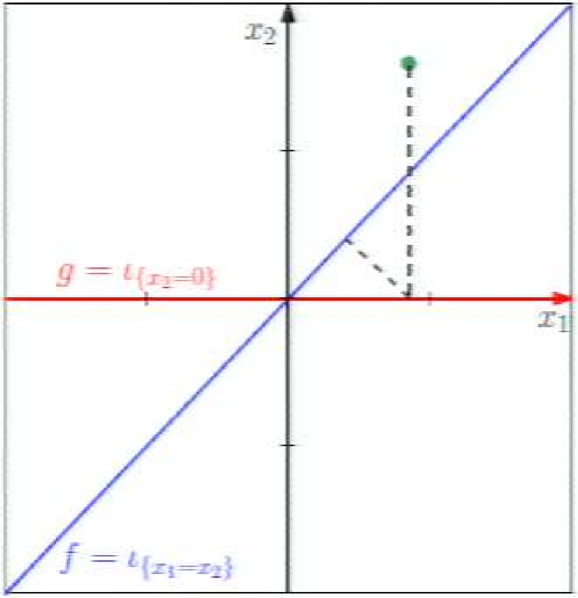
Theorem

On the real line, $\exists h$ such that $\mathsf{P}_h = \mathsf{P}_f \circ \mathsf{P}_g$.

- *Not necessarily $h = f + g$, though*

Example (A simple counterexample)

Consider $\mathbb{R}^2$, and let $f = \iota_{\{x_1 = x_2\}}, g = \iota_{\{x_2 = 0\}}$.



$$\mathsf{P}_f = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad \mathsf{P}_g = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

$$\text{But } \mathsf{P}_f \circ \mathsf{P}_g = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \end{bmatrix}$$

no h such that $\mathsf{P}_h = \mathsf{P}_f \circ \mathsf{P}_g$

Nevertheless

- Can ask the decomposition to hold for many but not all cases.
- Manipulating the optimality conditions:

$$\mathsf{P}_{f+g}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{w}\|^2 + (f+g)(\mathbf{w})$$
$$\mathsf{P}_{g}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{w}\|^2 + g(\mathbf{w})$$
$$\mathsf{P}_{f}(\mathsf{P}_{g}(\mathbf{z})) = \operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\mathsf{P}_{g}(\mathbf{z})-\mathbf{w}\|^2 + f(\mathbf{w}).$$

Theorem

A sufficient condition for $\mathsf{P}_{f+g}(\mathbf{z}) = \mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))$ *is*

$$\forall\, \mathbf{y} \in \operatorname{dom} g,\ \partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y}).$$

- Fails to be necessary at boundary points
- A special case appeared in a proof of (Zhou et al.'12)

Nevertheless

- Can ask the decomposition to hold for many but not all cases.
- Manipulating the optimality conditions:

$$\mathsf{P}_{f+g}(\mathbf{z}) - \mathbf{z} + \partial(f+g)(\mathsf{P}_{f+g}(\mathbf{z})) \ni 0$$
$$\mathsf{P}_{g}(\mathbf{z}) - \mathbf{z} + \partial g(\mathsf{P}_{g}(\mathbf{z})) \ni 0$$
$$\mathsf{P}_{f}(\mathsf{P}_{g}(\mathbf{z})) - \mathsf{P}_{g}(\mathbf{z}) + \partial f(\mathsf{P}_{f}(\mathsf{P}_{g}(\mathbf{z}))) \ni 0.$$

Theorem

A sufficient condition for $\mathsf{P}_{f+g}(\mathbf{z}) = \mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))$ *is*

$$\forall\, \mathbf{y} \in \operatorname{dom} g,\ \partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y}).$$

- Fails to be necessary at boundary points
- A special case appeared in a proof of (Zhou et al.'12)

Nevertheless

- Can ask the decomposition to hold for many but not all cases.
- Manipulating the optimality conditions:

$$P_{f+g}(\mathbf{z}) - \mathbf{z} + \partial(f+g)(P_{f+g}(\mathbf{z})) \ni 0$$
$$P_f(P_g(\mathbf{z})) - \mathbf{z} + \partial g(P_g(\mathbf{z})) + \partial f(P_f(P_g(\mathbf{z}))) \ni 0.$$

Theorem

A sufficient condition for $P_{f+g}(\mathbf{z}) = P_f(P_g(\mathbf{z}))$ *is*

$$\forall\, \mathbf{y} \in \operatorname{dom} g,\ \partial g(P_f(\mathbf{y})) \supseteq \partial g(\mathbf{y}).$$

- Fails to be necessary at boundary points
- A special case appeared in a proof of (Zhou et al.'12)

Nevertheless

- Can ask the decomposition to hold for many but not all cases.
- Manipulating the optimality conditions:

$$\mathsf{P}_{f+g}(\mathbf{z}) - \mathbf{z} + \partial(f+g)(\mathsf{P}_{f+g}(\mathbf{z})) \ni 0$$
$$\mathsf{P}_f(\mathsf{P}_g(\mathbf{z})) - \mathbf{z} + \partial g(\mathsf{P}_g(\mathbf{z})) + \partial f(\mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))) \ni 0.$$

Theorem

A sufficient condition for $\mathsf{P}_{f+g}(\mathbf{z}) = \mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))$ *is*

$$\forall\, \mathbf{y} \in \operatorname{dom} g,\ \partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y}).$$

- Fails to be necessary at boundary points
- A special case appeared in a proof of (Zhou et al.'12)

The rest is easy

- Find f and g that clinch our sufficient condition.

Start with "trivialities"

Theorem

Fix f. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ *for all* g *if and only if*

- $\dim(\mathcal{H}) \geq 2$; $f \equiv c$ *or* $f = \iota_{\{\mathbf{w}\}} + c$ *for some* $c \in \mathbb{R}$ *and* $\mathbf{w} \in \mathcal{H}$;
- $\dim(\mathcal{H}) = 1$ *and* $f = \iota_C + c$ *for some* $c \in \mathbb{R}$ *and set* C *that is closed and convex.*

Asymmetry.

Theorem

Fix g. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ *for all* f *if and only if* g *is continuous affine.*

- Reassuring the impossibility to always have $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$;
- Still hope to get interesting results!

Scaling Invariant ⇔ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

g positive homogeneous ⇔ $\forall \lambda > 0, \partial g(\lambda \mathbf{w}) = \partial g(\mathbf{w})$ ⇒ $\forall \mathbf{z}, \mathsf{P}_f(\mathbf{z}) \propto \mathbf{z}$

Scaling Invariant ⇔ Positive Homogeneous

$$\partial g(\mathsf{P}_f(y)) \supseteq \partial g(y)$$

g positive homogeneous ⇔ $\forall \lambda > 0, \partial g(\lambda \mathbf{w}) = \partial g(\mathbf{w}) \Rightarrow \forall \mathbf{z}, \mathsf{P}_f(\mathbf{z}) \propto \mathbf{z}$

Theorem

Fix f. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):

i).

ii).

iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv).

Scaling Invariant ⇔ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

g positive homogeneous ⇔ $\forall \lambda > 0, \partial g(\lambda \mathbf{w}) = \partial g(\mathbf{w}) \Rightarrow \forall \mathbf{z}, \mathsf{P}_f(\mathbf{z}) \propto \mathbf{z}$

Theorem

Fix f. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$*):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\mathbf{x} \perp \mathbf{y}$, $f(\mathbf{x} + \mathbf{y}) \geq f(\mathbf{y})$;

iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \mathrm{dom}\, f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* κ.

If $\dim(\mathcal{H}) = 1$, *only ii)* $\Longrightarrow$ *i) ceases to hold.*

Scaling Invariant ⇔ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

g positive homogeneous ⇔ $\forall \lambda > 0, \partial g(\lambda \mathbf{w}) = \partial g(\mathbf{w}) \Rightarrow \forall \mathbf{z}, \mathsf{P}_f(\mathbf{z}) \propto \mathbf{z}$

Theorem

Fix f. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\mathbf{x} \perp \mathbf{y}$, $f(\mathbf{x} + \mathbf{y}) \geq f(\mathbf{y})$;

iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* κ.

If $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

Some Implications

Theorem

Fix f. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\mathbf{x} \perp \mathbf{y}$, $f(\mathbf{x} + \mathbf{y}) \geq f(\mathbf{y})$;

iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv). $0 \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* κ.

If $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

i) $\iff$ ii)

- Characterizing representer theorem (Dinuzzo & Schölkopf'12);
- Now we have more.

Some Implications

Theorem

Fix f. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $x \perp y$, $f(x+y) \geq f(y)$;

iii). *For all* $z \in \mathcal{H}$, $P_f(z) = \lambda_z \cdot z$ *for some* $\lambda_z \in [0,1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $P_{f+\kappa} = P_f \circ P_\kappa$ *for all positive homogeneous* κ.

If $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

i) $\implies$ iv)

$$P_{\lambda\|\cdot\|^2+\kappa} = P_{\lambda\|\cdot\|^2} \circ P_\kappa = \frac{1}{\lambda+1} P_\kappa$$

- Double shrinkage;
- $\kappa = \|\cdot\|_1$: Elastic net (Zou & Hastie'05);
- Adding an ℓ_2-ish regularizer, computationally, is free.

Some Implications

Theorem

Fix f. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\mathbf{x} \perp \mathbf{y}$, $f(\mathbf{x} + \mathbf{y}) \geq f(\mathbf{y})$;

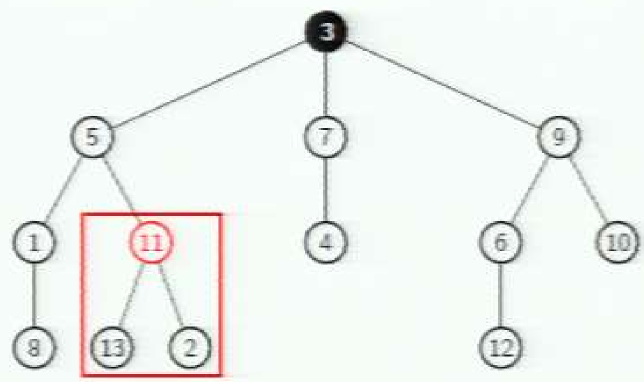
iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_\mathbf{z} \cdot \mathbf{z}$ *for some* $\lambda_\mathbf{z} \in [0, 1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* κ.

i) $\implies$ iv)

Tree-structured group norms (Jenatton et al.'11)

$$\mathsf{P}_{\sum_i \|\cdot\|_{\mathbf{g}_i}} = \mathsf{P}_{\|\cdot\|_{\mathbf{g}_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{\mathbf{g}_k}}.$$



Permutation Invariant $\Leftrightarrow$ Choquet Integral

$$\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

Theorem

Let f be permutation invariant and g be the Choquet integral of some submodular set function μ. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

Permutation Invariant ⇔ Choquet Integral

$$\partial g(\mathsf{P}_f(y)) \supseteq \partial g(y)$$

- f permutation invariant $\Rightarrow \mathsf{P}_f(\mathbf{y}) /\!/ \mathbf{y}$:

$$(y_i - y_j)([\mathsf{P}_f(\mathbf{y})]_i - [\mathsf{P}_f(\mathbf{y})]_j) \geq 0$$

- ∂g invariant to comonotone vectors
- Choquet integral (*a.k.a.* Lovász extension) of $\mu : 2^{[d]} \to \mathbb{R}$:

$$g(\mathbf{w}) := \int_0^\infty \mu([\![\mathbf{w} \geq t]\!])\,dt + \int_{-\infty}^0 [\mu([\![\mathbf{w} \geq t]\!]) - \mu([d])]\,dt$$

Theorem

Let f be permutation invariant and g be the Choquet integral of some submodular set function μ. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

Permutation Invariant $\Leftrightarrow$ Choquet Integral

$$\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

- f permutation invariant $\Rightarrow$ $\mathsf{P}_f(\mathbf{y}) /\!/ \mathbf{y}$:

$$(y_i - y_j)([\mathsf{P}_f(\mathbf{y})]_i - [\mathsf{P}_f(\mathbf{y})]_j) \geq 0$$

- ∂g invariant to comonotone vectors
- Choquet integral (*a.k.a.* Lovász extension) of $\mu : 2^{[d]} \to \mathbb{R}$:

$$g(\mathbf{w}) := \int_0^\infty \mu([\![\mathbf{w} \geq t]\!])\, \mathrm{d}t + \int_{-\infty}^0 [\mu([\![\mathbf{w} \geq t]\!]) - \mu([d])]\, \mathrm{d}t$$

Theorem

Let f be permutation invariant and g be the Choquet integral of some submodular set function μ. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

Permutation Invariant $\Leftrightarrow$ Choquet Integral

$$\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

- f permutation invariant $\Rightarrow$ $\mathsf{P}_f(\mathbf{y}) /\!/ \mathbf{y}$:

$$(y_i - y_j)([\mathsf{P}_f(\mathbf{y})]_i - [\mathsf{P}_f(\mathbf{y})]_j) \geq 0$$

- ∂g invariant to comonotone vectors
- Choquet integral (*a.k.a.* Lovász extension) of $\mu : 2^{[d]} \to \mathbb{R}$:

$$g(\mathbf{w}) := \int_0^\infty \mu([\![\mathbf{w} \geq t]\!])\,\mathrm{d}t + \int_{-\infty}^0 [\mu([\![\mathbf{w} \geq t]\!]) - \mu([d])]\,\mathrm{d}t$$

Theorem

Let f be permutation invariant and g be the Choquet integral of some submodular set function μ. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

Some Implications

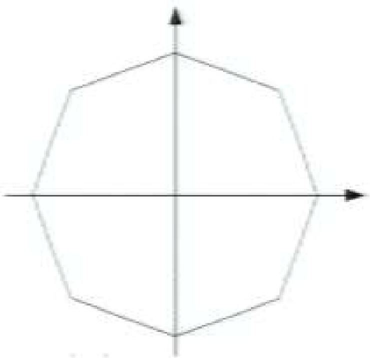
Theorem

Let f be permutation invariant and g be the Choquet integral of some submodular set function. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

- Special case $f = \|\cdot\|_1$ in (Bach'11);
- $\mathsf{P}_{\|\cdot\|_1 + \|\cdot\|_{\mathsf{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathsf{TV}}}$ (Friedman et al.'07);
- $\mathsf{P}_{\sum_{i=1}^k \|\cdot\|_{g_i}} = \mathsf{P}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{g_k}}$ (Jenatton et al.'11)

Some Implications

$$\|\mathbf{w}\|_{\texttt{oscar}} = \sum_{i<j} \max\{|w_i|, |w_j|\}.$$



- Feature grouping (Bondell & Reich'08)
- $\mathsf{P}_{\|\cdot\|_{\texttt{oscar}}}$ in (Zhong & Kwok'11)

Let

$$\kappa_i(\mathbf{w}) := \sum_{j:j<i} \max\{|w_i|, |w_j|\}.$$

- $\|\mathbf{w}\|_{\texttt{oscar}} = \sum_{i=2}^{d} \kappa_i(\mathbf{w})$
- $\mathsf{P}_{\|\cdot\|_{\texttt{oscar}}} = \mathsf{P}_{\kappa_d} \circ \cdots \circ \mathsf{P}_{\kappa_2}$
- Given P_{κ_i}, constant time for $\mathsf{P}_{\kappa_{i+1}}$.

Some Implications

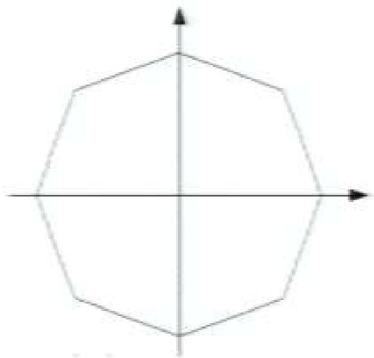
Theorem

Let f be permutation invariant and g be the Choquet integral of some submodular set function. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

- Special case $f = \|\cdot\|_1$ in (Bach'11);
- $\mathsf{P}_{\|\cdot\|_1 + \|\cdot\|_{\mathsf{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathsf{TV}}}$ (Friedman et al.'07);
- $\mathsf{P}_{\sum_{i=1}^k \|\cdot\|_{g_i}} = \mathsf{P}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{g_k}}$ (Jenatton et al.'11)

Some Implications

$$\|\mathbf{w}\|_{\texttt{oscar}} = \sum_{i<j} \max\{|w_i|, |w_j|\}.$$



- Feature grouping (Bondell & Reich'08)
- $\mathsf{P}_{\|\cdot\|_{\texttt{oscar}}}$ in (Zhong & Kwok'11)

Let

$$\kappa_i(\mathbf{w}) := \sum_{j:j<i} \max\{|w_i|, |w_j|\}.$$

- $\|\mathbf{w}\|_{\texttt{oscar}} = \sum_{i=2}^{d} \kappa_i(\mathbf{w})$
- $\mathsf{P}_{\|\cdot\|_{\texttt{oscar}}} = \mathsf{P}_{\kappa_d} \circ \cdots \circ \mathsf{P}_{\kappa_2}$
- Given P_{κ_i}, constant time for $\mathsf{P}_{\kappa_{i+1}}$.

Sufficient Condition Fails?

- (Martins et al.'11) showed that, under a shrinkage assumption, the prox-decomposition (even not true) can still be used in a subgradient-type algorithm
- (Yu'13a) showed that a simple linearization of the proximal map, i.e.

$$\mathsf{P}_{\sum_k f_k} \approx \sum_k \mathsf{P}_{f_k},$$

yields slightly faster convergence than the smoothing trick

Summary

- Posed the question: $\mathsf{P}_{f+g} \stackrel{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \stackrel{?}{=} \mathsf{P}_g \circ \mathsf{P}_f$;
- Presented a sufficient condition: $\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$;
- Identified two major cases;
- Immediately useful if plugged into PG;

Summary

- Posed the question: $\mathsf{P}_{f+g} \stackrel{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \stackrel{?}{=} \mathsf{P}_g \circ \mathsf{P}_f$;
- Presented a sufficient condition: $\partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$;
- Identified two major cases;
- Immediately useful if plugged into PG;

Thanks!

NIPS

Neural Information Processing Systems Foundation

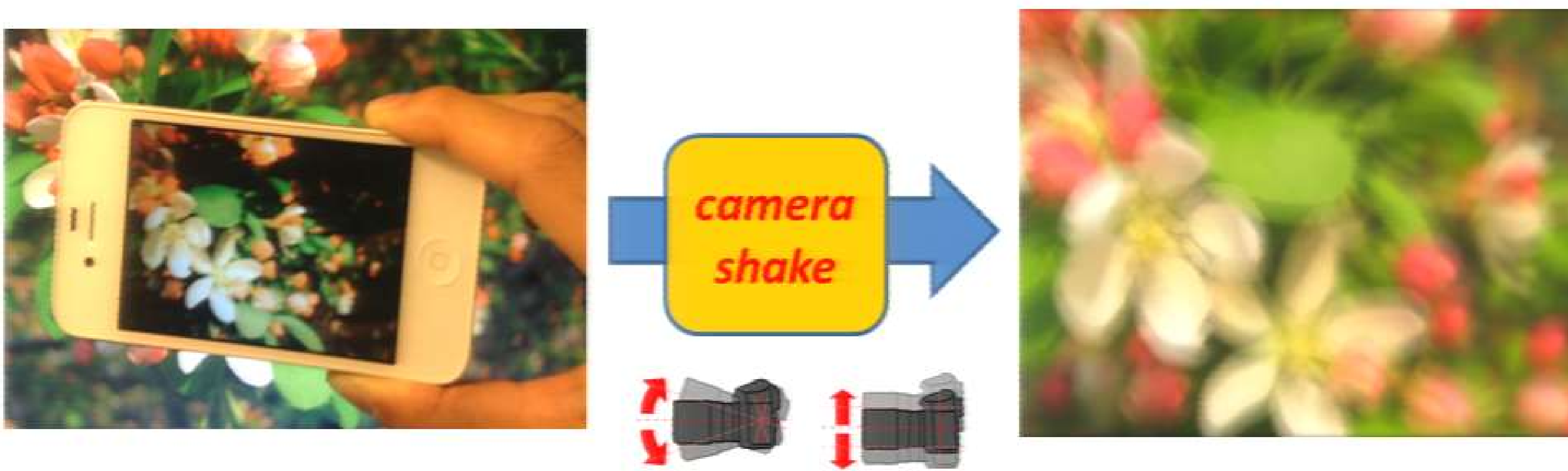
Non-Uniform Camera Shake Removal using a Spatially Adaptive Sparse Penalty

Haichao Zhang [1,2]   David Wipf [3]

NIPS, December 2013

Problem & Objective

- Problem
  - **Camera shake blur** caused by **relative movement** between camera and scene **during exposure**.



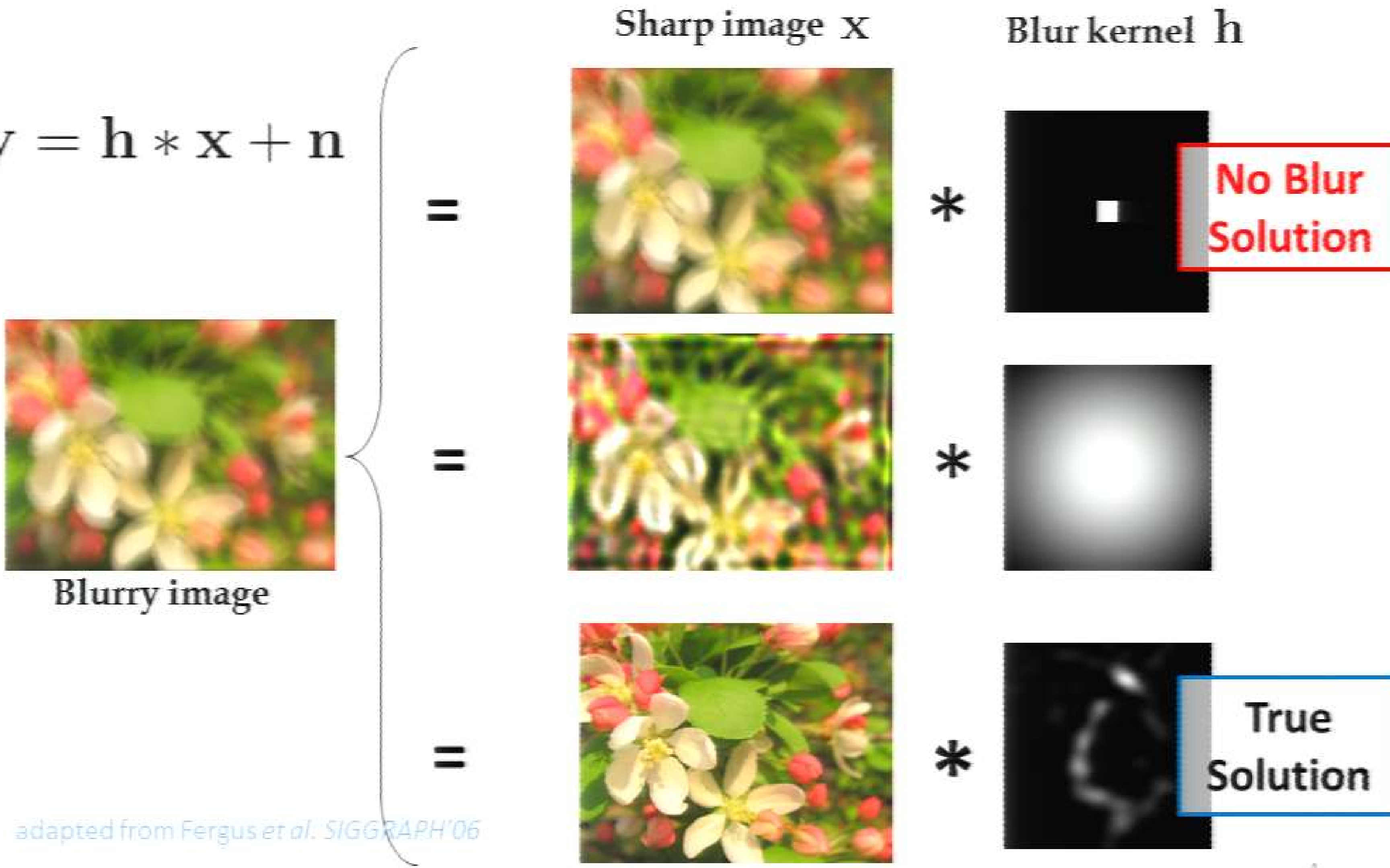
Problem & **Objective**

- Objective
  - Recover the **sharp image** from a single **blurry** image with **unknown** camera shake.

Challenge I

Ill-posed Problem: no unique solution

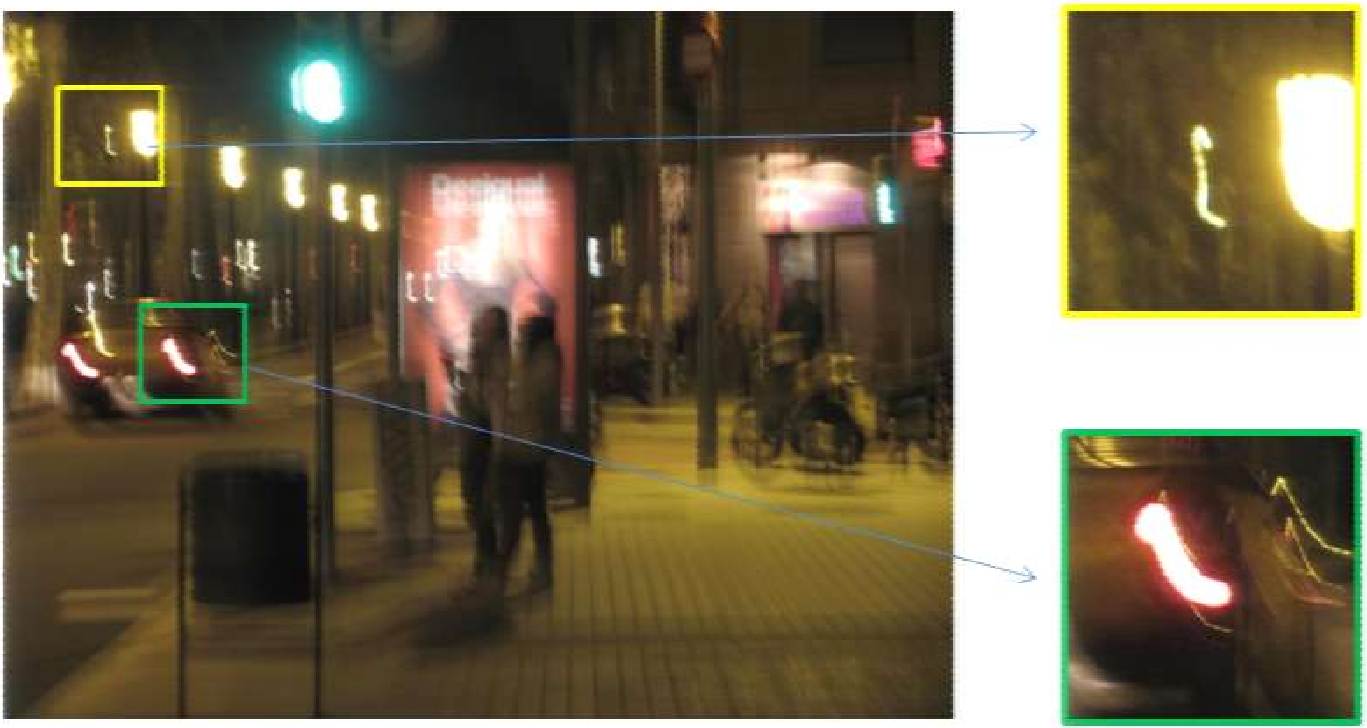
$$\mathbf{y} = \mathbf{h} * \mathbf{x} + \mathbf{n}$$



adapted from Fergus *et al.* *SIGGRAPH'06*

Challenge II

Real-World Camera Shake: *spatial-variant*

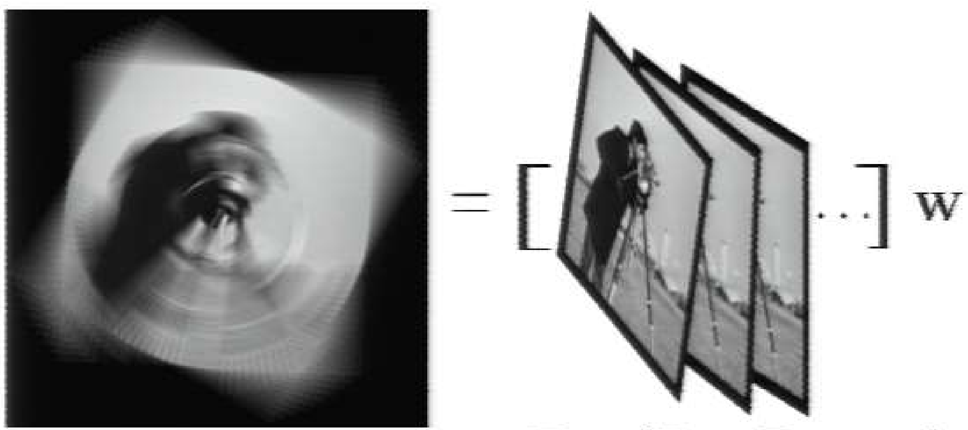


Non-Uniform Observation Model
... regarding challenge II

Observation Model [Projective Motion Path]

$$\mathbf{y} = \sum_j w_j \mathbf{P}_j \mathbf{x} + \mathbf{n} = \mathbf{D}\mathbf{w} + \mathbf{n} \qquad \mathbf{D} = [\mathbf{P}_1\mathbf{x}, \mathbf{P}_2\mathbf{x}, \cdots]$$

blurry image — blur vector — sharp image — noise



$$\mathbf{D} = [\mathbf{P}_1\mathbf{x}, \mathbf{P}_2\mathbf{x}, \cdots]$$

Tai et al. 2009
Hirsch et al. ICCV 2011

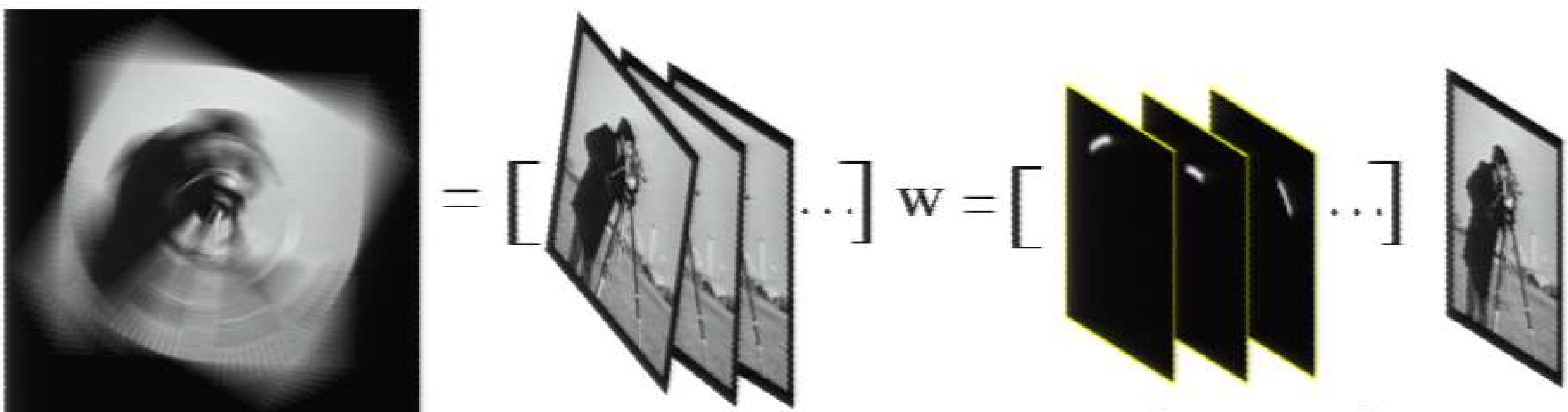
Non-Uniform Observation Model
... regarding challenge II

Observation Model [Projective Motion Path]

$$\mathbf{y} = \sum_j w_j \mathbf{P}_j \mathbf{x} + \mathbf{n} = \mathbf{D}\mathbf{w} + \mathbf{n} = \mathbf{H}\mathbf{x} + \mathbf{n}$$

$$\mathbf{D} = [\mathbf{P}_1\mathbf{x}, \mathbf{P}_2\mathbf{x}, \cdots] \qquad \mathbf{H} = \sum_j w_j \mathbf{P}_j$$

blurry image ($\mathbf{y}$), blur vector (w_j), sharp image ($\mathbf{x}$), noise ($\mathbf{n}$)



$$\mathbf{D} = [\mathbf{P}_1\mathbf{x}, \mathbf{P}_2\mathbf{x}, \cdots] \qquad \mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots]$$

Tai et al. 2009
Hirsch et al. ICCV 2011

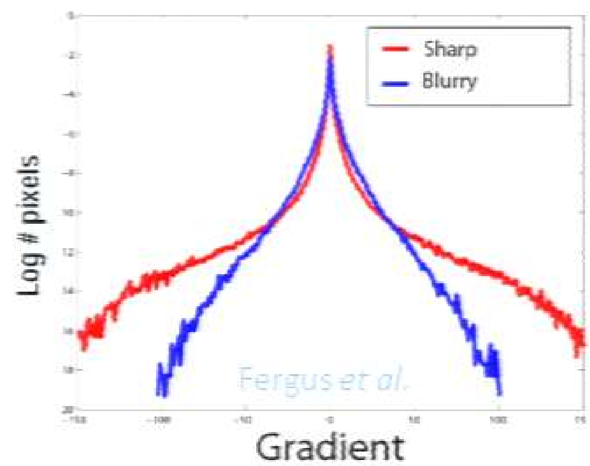
Sparse Image Prior
… regarding challenge I

Likelihood $$p(\mathbf{y}|\mathbf{x},\mathbf{w}) \propto \exp\left\{-\frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2\right\}$$

Image Prior [sparse gradient]

$$p(\mathbf{x}) \propto \exp\left[-\sum_i g(x_i)\right]$$

$g(x_i)$ is a concave function



Blur Prior $p(\mathbf{w})$

Work in derivative domain

$\mathbf{x}$ (vectorized) derivatives of the sharp image

$\mathbf{y}$ (vectorized) derivatives of the blurry image

Fergus et al. Removing camera shake from a single image, SIGGRAPH'06

Direct MAP ?

MAP Estimation

$$\max_{\mathbf{x},\mathbf{w}\geq 0} p(\mathbf{x},\mathbf{w}|\mathbf{y}) \equiv \min_{\mathbf{x},\mathbf{w}\geq 0} -\log[p(\mathbf{y}|\mathbf{x},\mathbf{w})p(\mathbf{x})p(\mathbf{w})]$$

⇕

$$\min_{\mathbf{x},\mathbf{w}\geq 0} \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \alpha \sum_i g(x_i) + \beta \sum_j f(w_j)$$

- Local minima and “no-blur” solution
- Empirical tricks
  - initialization, structure selection/prediction [Cho and Lee SIGGRAPH Asia’09, Xu & Jia ECCV’10, Hu et al. BMVC’12]

Type-II Estimation

Likelihood $p(\mathbf{y}|\mathbf{x},\mathbf{w}) \propto \exp\left\{-\frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2\right\}$

Image Prior $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{x};\mathbf{0},\mathbf{\Gamma})$ $\quad \mathbf{\Gamma} \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$

$$\max_{\gamma,\mathbf{w},\lambda\geq 0} \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x},\mathbf{w})p(\mathbf{x})d\mathbf{x}$$

uniform p(**w**) is used

Tipping et al. Sparse Bayesian Learning and Relevance Vector Machine, JMLR 2011
Levin et al. Efficient Marginal Likelihood Optimization in Blind Deconvolution, CVPR 2011
Wipf et al., Latent Variable Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Effective Cost Function

The Cost Function

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

$$\psi(u,\lambda) \triangleq \frac{2u}{u+\sqrt{4\lambda+u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda+u^2}\right) \quad u \geq 0$$

Wipf et al., Latent Variable Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Effective Cost Function

The Cost Function

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \boxed{\frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2} + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

$$\psi(u,\lambda) \triangleq \frac{2u}{u+\sqrt{4\lambda+u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda+u^2}\right) \quad u \geq 0$$

reconstruction error

Wipf et al., Latent Variable Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Effective Cost Function

The Cost Function

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

$$\psi(u,\lambda) \triangleq \frac{2u}{u+\sqrt{4\lambda+u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda+u^2}\right) \quad u \geq 0$$

sparse penalty function

$\psi(u)$ is a concave, non-decreasing function of u

Wipf et al., Latent Variable Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Effective Cost Function

The Cost Function

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2,\lambda) + (n-m)\log\lambda$$

$$\psi(u,\lambda) \triangleq \frac{2u}{u+\sqrt{4\lambda+u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda+u^2}\right) \quad u \geq 0$$

noise level penalty term

Wipf et al., Latent Variable Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Effective Cost Function

The Cost Function

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

$$\psi(u,\lambda) \triangleq \frac{2u}{u+\sqrt{4\lambda+u^2}} + \log\left(2\lambda+u^2+u\sqrt{4\lambda+u^2}\right) \quad u \geq 0$$

can be solved using the *majorization-minimization* technique

Wipf et al., Latent Variable Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Effective Cost Function

The Cost Function

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

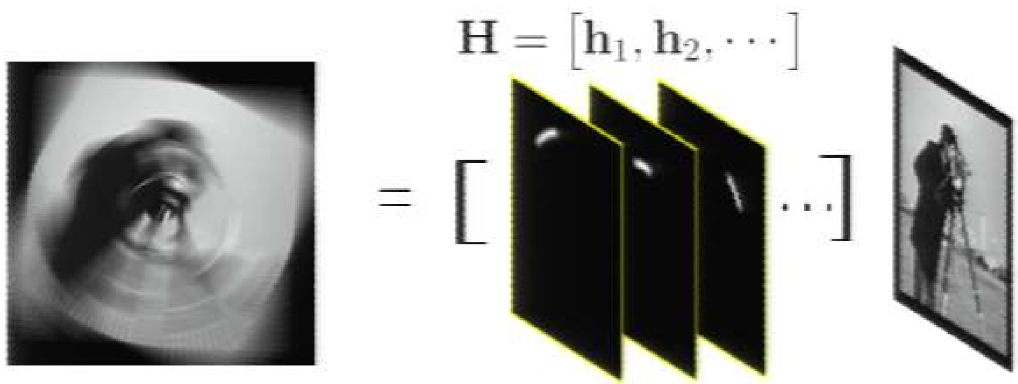
**looks similar, what's the real advantage ...
(over the regular MAP)?**

can be solved using the majorization-minimization technique

$$\min_{\mathbf{x},\mathbf{w}\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \alpha\sum_i g(x_i) + \beta\sum_j f(w_j)$$

Challenge II *Revisited*
Real-World Camera Shake: *spatially-variant*

- **Effect of Spatially-Variant Blur on H**
  - Imbalanced Column of **H** ($\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots]$)
    - Each column of **H** corresponds to a localized blur kernel
    - Large blur has smaller L2 norm ($h_i \geq 0, \sum h_i = 1$)
    - Columns of **H** have different L2 norms (local kernel norm $\|\mathbf{h}_i\|_2$)

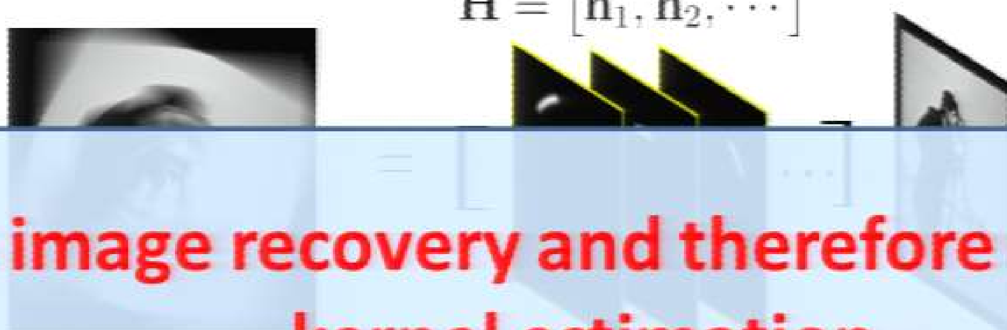


Challenge II *Revisited*

Real-World Camera Shake: *spatially-variant*

- **Effect of Spatially-Variant Blur on H**
  - Imbalanced Column of **H** ($\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots]$)
    - Each column of **H** corresponds to a localized blur kernel
    - Large blur has smaller L2 norm ($h_i \geq 0, \sum h_i = 1$)
    - Columns of **H** have different L2 norms (local kernel norm $\|\mathbf{h}_i\|_2$)

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots]$$



Bias image recovery and therefore affect the kernel estimation.

Model Properties
Automated Column-Normalization

- **Column-Normalized Sparse Estimation**

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

local kernel norm embedded

compensates for the spatial variance

Model Properties
Automated Column-Normalization

- **Column-Normalized Sparse Estimation**

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

$$z_i \triangleq x_i\|\mathbf{h}_i\|_2$$

$$\min_{\mathbf{z};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\tilde{\mathbf{H}}\mathbf{z}\|_2^2 + \sum_i \psi(|z_i|, \lambda) + (n-m)\log\lambda$$

$\tilde{\mathbf{H}}$ is the column-normalized $\mathbf{H}$

Model Properties
Automated Column-Normalization

- **Column-Normalized Sparse Estimation**

$$\min_{\mathbf{x};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \sum_i \psi(|x_i|\|\mathbf{h}_i\|_2, \lambda) + (n-m)\log\lambda$$

$$z_i \triangleq x_i\|\mathbf{h}_i\|_2$$

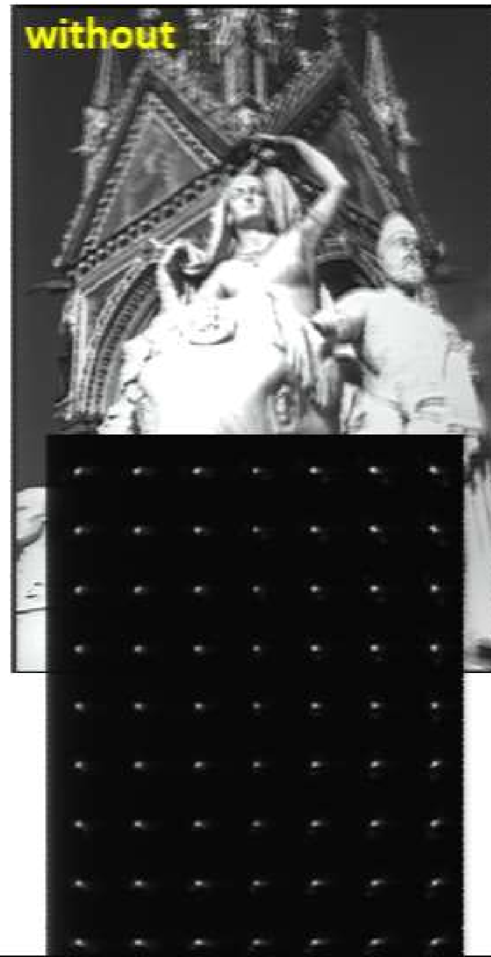
large structure, low blur region wil be naturally emphasized

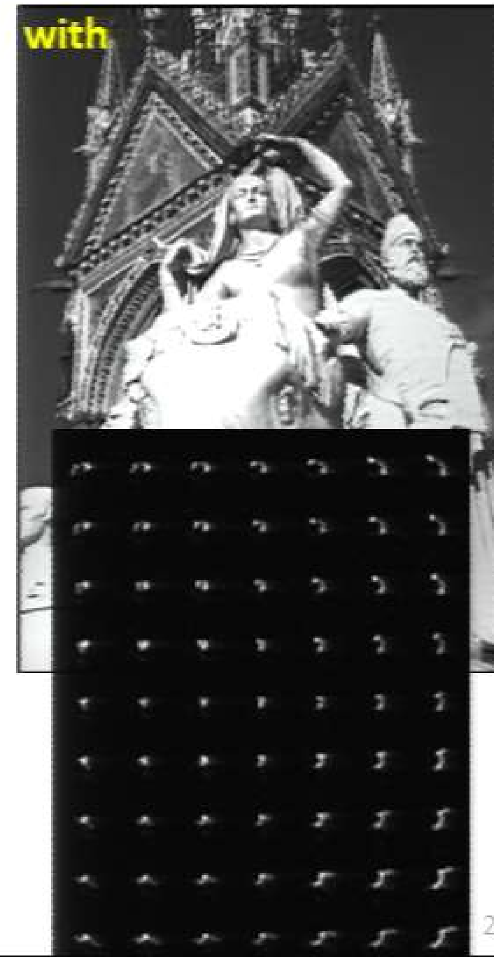
$$\min_{\mathbf{z};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\tilde{\mathbf{H}}\mathbf{z}\|_2^2 + \sum_i \psi(|z_i|, \lambda) + (n-m)\log\lambda$$

Avoids **premuture favoring of any one element of z over another (thus avoid biased image recovery)**

Model Properties
Automated Column-Normalization

- **Effects of Column-Normalization** (blind deblurring)



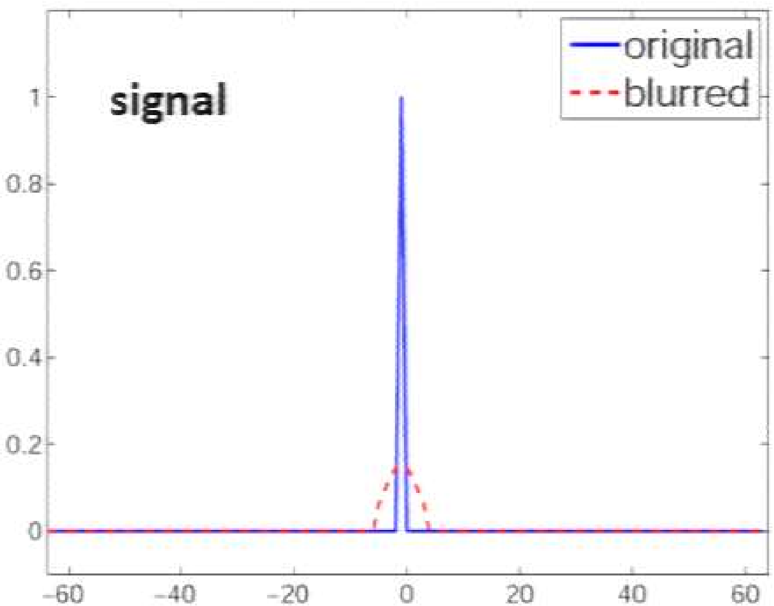


Blurry image from Harmeling et al.
NIPS 2010

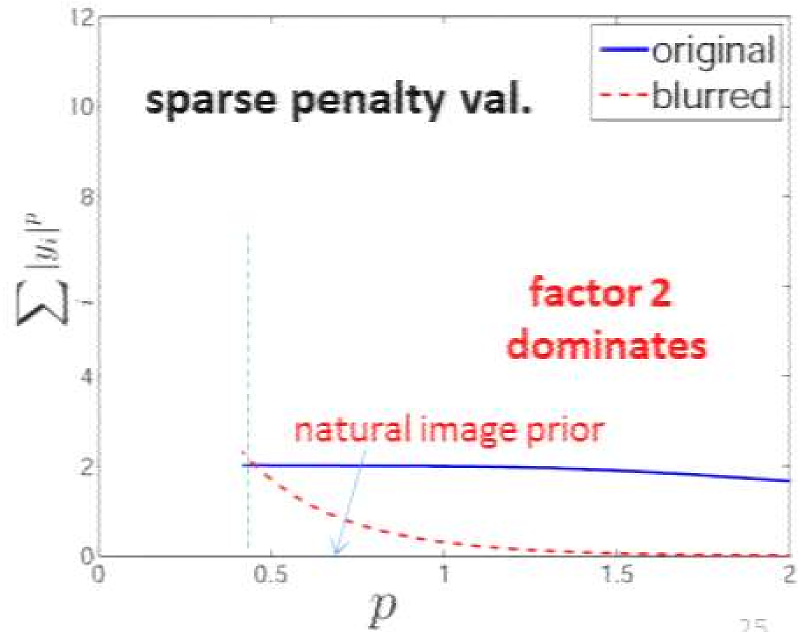
Challenge I *Revisited*

Ill-posed Problem: no unique solution

- **Two Effects of Blur on Sparsity Measure** (Lp-norm)
  1. Reduces sparsity $\sum_i |y_i|^p \nearrow$
  2. Reduces variance $\sum_i |y_i|^p \searrow$

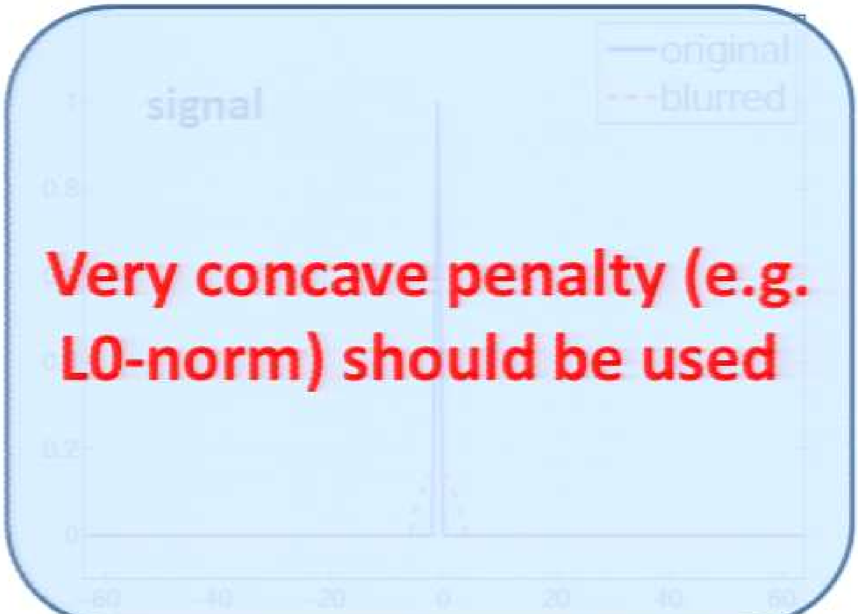


Levin et al., *CVPR* 2009

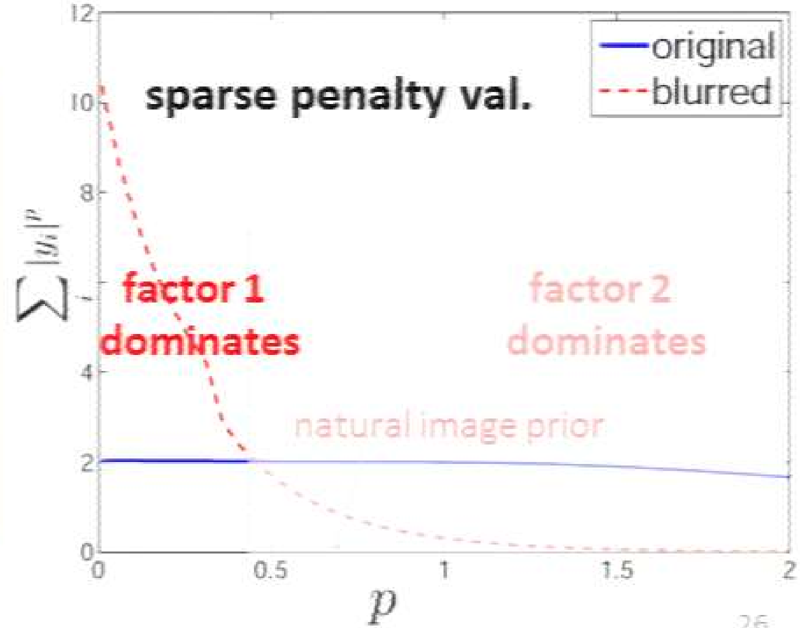


Challenge I *Revisited*
Ill-posed Problem: no unique solution

- **Two Effects of Blur on Sparsity Measure** (Lp-norm)
  1. Reduces sparsity $\sum_i |y_i|^p \nearrow$
  2. Reduces variance $\sum_i |y_i|^p \searrow$



Very concave penalty (e.g. L0-norm) should be used



Levin et al., *CVPR* 2009

Challenge I *Revisited*

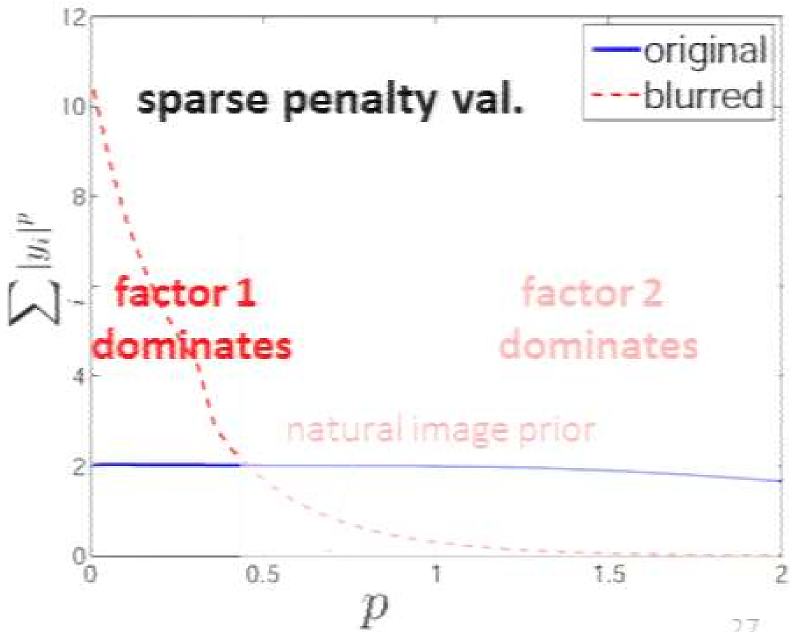
Ill-posed Problem: no unique solution

- **Two Effects of Blur on Sparsity Measure** (Lp-norm)
  1. Reduces sparsity $\sum_i |y_i|^p \nearrow$
  2. Reduces variance $\sum_i |y_i|^p \searrow$

Very concave penalty (e.g. L0-norm) should be used

⇩

Non-Convex Problem!!!



Levin et al., *CVPR* 2009

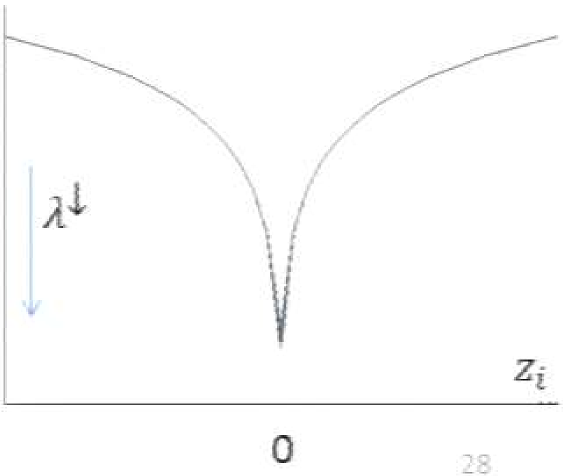
Model Properties
Noise Dependent Homotopy Continuation

- **The penalty function in the proposed model**
  - A qualified "very concave" sparse penalty

As $\lambda \longrightarrow 0$, $\sum \psi(|z_i|, \lambda) \longrightarrow C\|z\|_0$

no-blur solution avoidance

$$\psi(u,\lambda) \triangleq \frac{2u}{u+\sqrt{4\lambda+u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda+u^2}\right) \quad u \geq 0$$



Model Properties
Noise Dependent Homotopy Continuation

- **The penalty function in the proposed model**
  - A qualified "very concave" sparse penalty

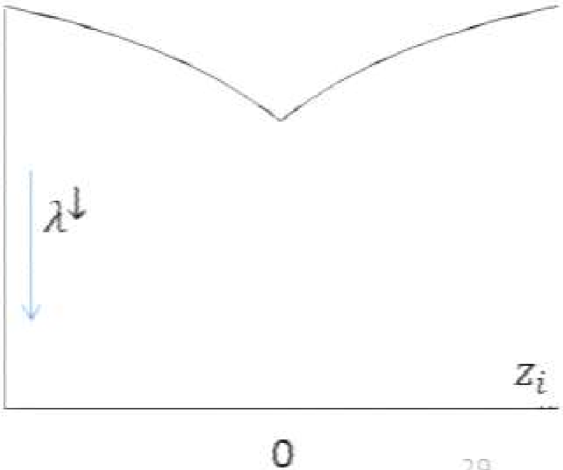
$$\textbf{As } \lambda \longrightarrow 0,\ \sum \psi(|z_i|, \lambda) \longrightarrow C\|z\|_0$$

no-blur solution avoidance

  - Adaptive penalty shape

$$\textbf{As } \lambda \textit{ is large},\ \sum \psi(|z_i|, \lambda) \longrightarrow 2\|z\|_1/\sqrt{\lambda}$$

$$\psi(u, \lambda) \triangleq \frac{2u}{u + \sqrt{4\lambda + u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda + u^2}\right) \quad u \geq 0$$



Model Properties
Noise Dependent Homotopy Continuation

- **The penalty function in the proposed model**
  - A qualified "very concave" sparse penalty

    As $\lambda \longrightarrow 0$, $\sum \psi(|z_i|, \lambda) \longrightarrow C\|z\|_0$

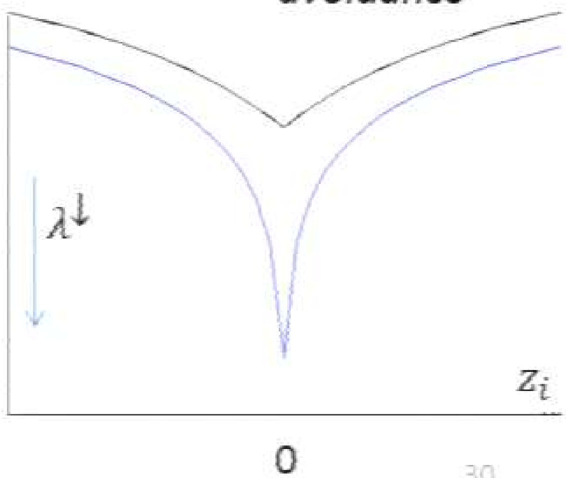
    no-blur solution avoidance

  - Adaptive penalty shape

    As λ *is large*, $\sum \psi(|z_i|, \lambda) \longrightarrow 2\|z\|_1/\sqrt{\lambda}$

    local-minia avoidance

    If $\lambda_1 < \lambda_2$, then $\psi(u, \lambda_1) < \psi(u, \lambda_2)$ for $u \geq 0$



$$\psi(u, \lambda) \triangleq \frac{2u}{u + \sqrt{4\lambda + u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda + u^2}\right) \quad u \geq 0$$

Model Properties
Noise Dependent Homotopy Continuation

- **The penalty function in the proposed model**
  - A qualified "very concave" sparse penalty

$$\text{As } \lambda \longrightarrow 0\,,\ \sum \psi(|z_i|, \lambda) \longrightarrow C\|z\|_0$$

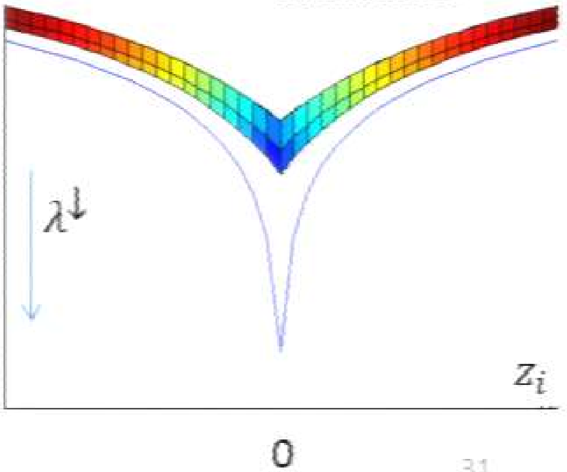
no-blur solution avoidance

  - Adaptive penalty shape

$$\text{As } \lambda \text{ is large},\ \sum \psi(|z_i|, \lambda) \longrightarrow 2\|z\|_1/\sqrt{\lambda}$$

local-minia avoidance

$$\text{If } \lambda_1 < \lambda_2, \text{ then } \psi(u, \lambda_1) \prec \psi(u, \lambda_2) \text{ for } u \geq 0$$



$$\psi(u, \lambda) \triangleq \frac{2u}{u + \sqrt{4\lambda + u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda + u^2}\right) \quad u \geq 0$$

Wipf et al., Latent Varianbel Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Model Properties
Noise Dependent Homotopy Continuation

- **The penalty function in the proposed model**
  - A qualified "very concave" sparse penalty

As $\lambda \longrightarrow 0$, $\sum \psi(|z_i|, \lambda) \longrightarrow C\|z\|_0$

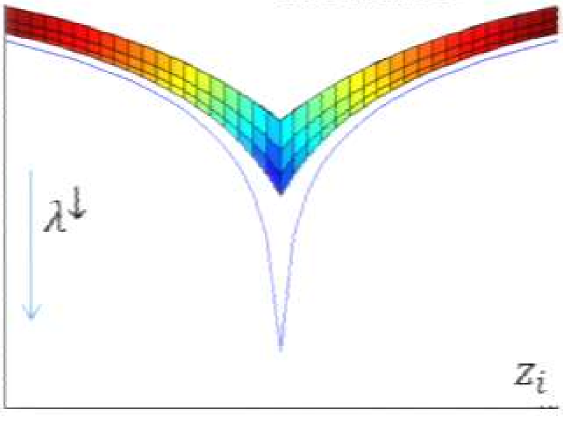
no-blur solution avoidance

  - Adaptive penalty shape

As λ *is large*, $\sum \psi(|z_i|, \lambda) \longrightarrow 2\|z\|_1/\sqrt{\lambda}$

local-minia avoidance

If $\lambda_1 < \lambda_2$, then $\psi(u, \lambda_1) \prec \psi(u, \lambda_2)$ for $u \geq 0$

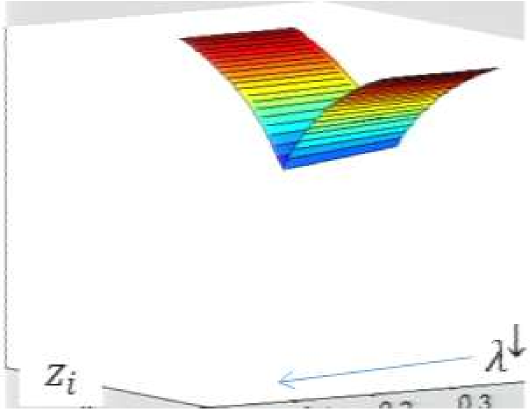


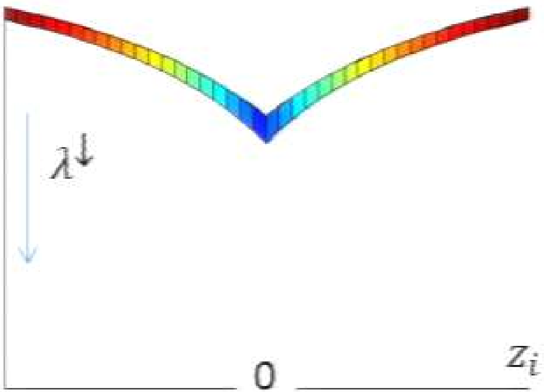
$$\psi(u, \lambda) \triangleq \frac{2u}{u + \sqrt{4\lambda + u^2}} + \log\left(2\lambda + u^2 + u\sqrt{4\lambda + u^2}\right) \quad u \geq 0$$

Wipf et al., Latent Varianbel Bayesian Model for Promoting Sparsity, IEEE TIT, 2011

Model Properties
Noise Dependent Homotopy Continuation

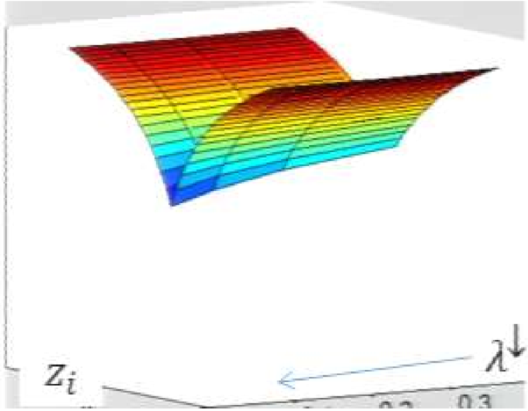
- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i ||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

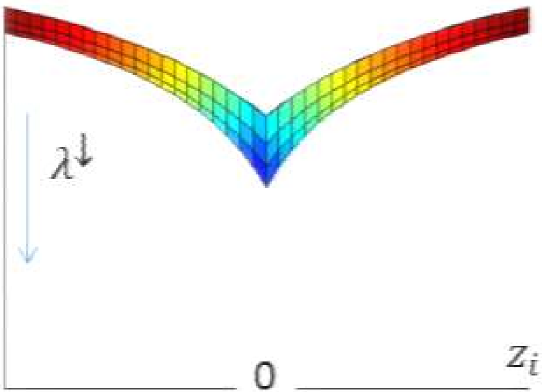




Model Properties
Noise Dependent Homotopy Continuation

- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i ||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

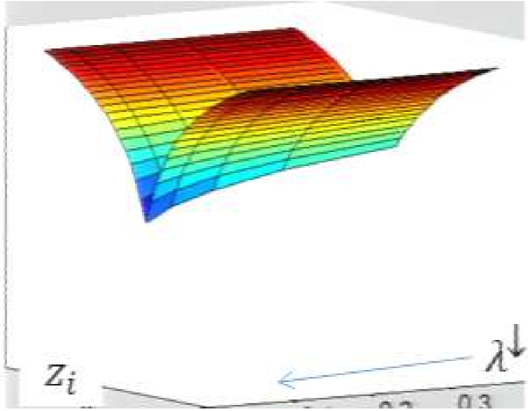


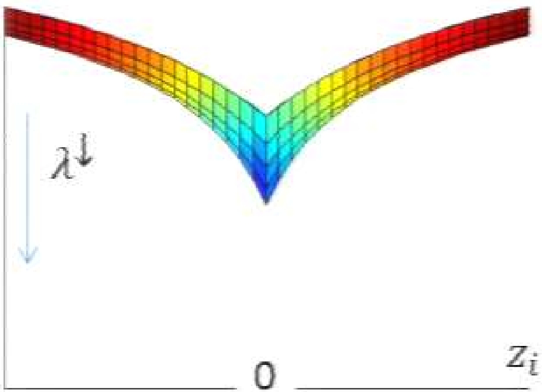


Model Properties

Noise Dependent Homotopy Continuation

- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i ||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

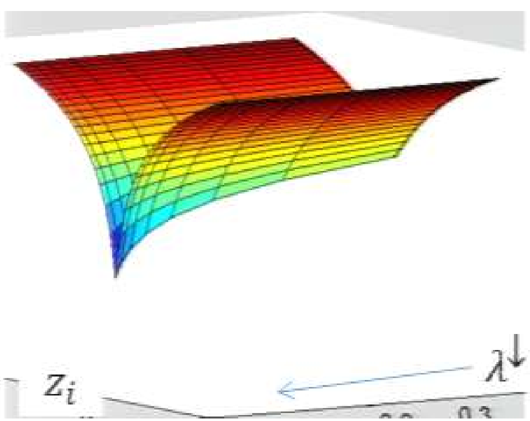


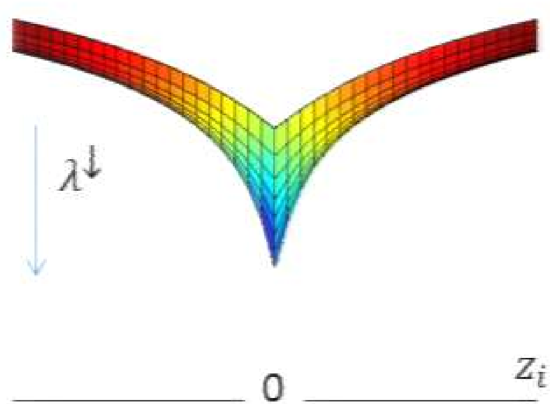


Model Properties

Noise Dependent Homotopy Continuation

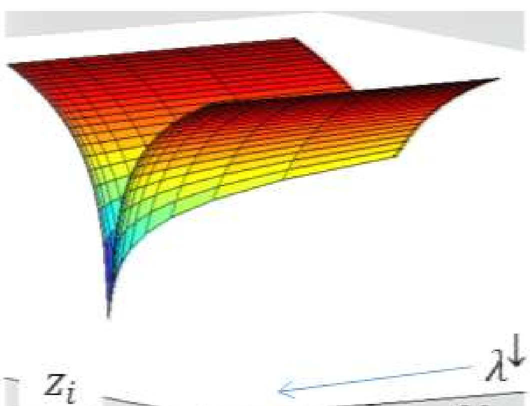
- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i ||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

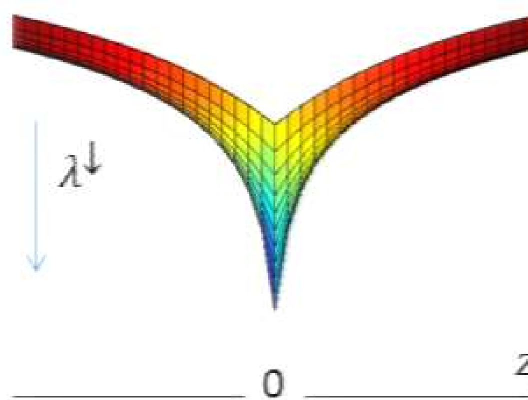




Model Properties
Noise Dependent Homotopy Continuation

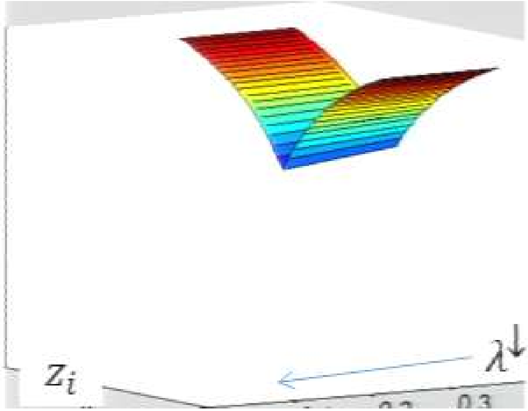
- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i \|h_i\|_2$
    - focus first on large structure (x_i large), low blur ($\|h_i\|_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

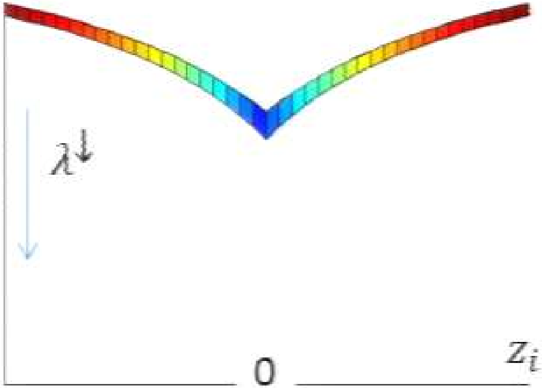




Model Properties
Noise Dependent Homotopy Continuation

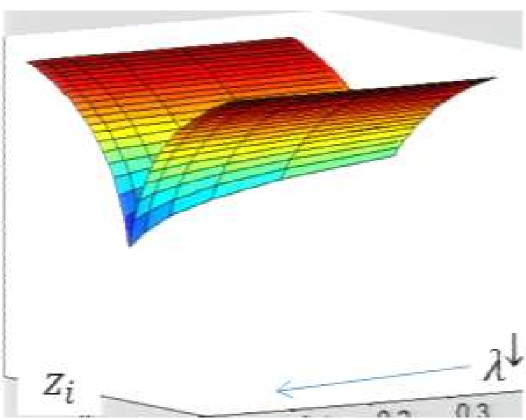
- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i \|h_i\|_2$
    - focus first on large structure (x_i large), low blur ($\|h_i\|_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

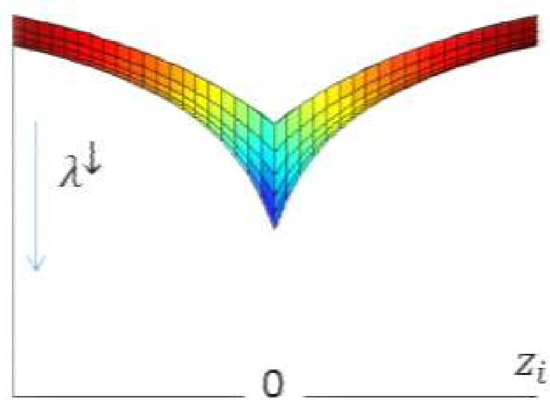




Model Properties
Noise Dependent Homotopy Continuation

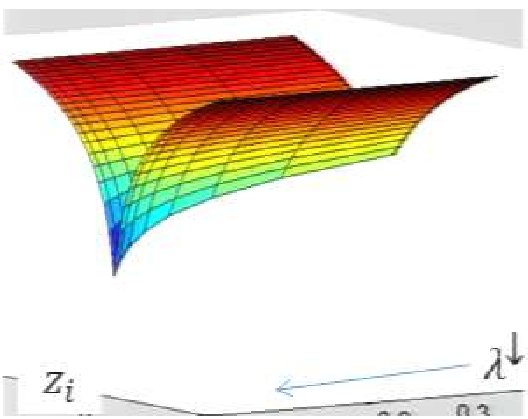
- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) z_i=$x_i||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

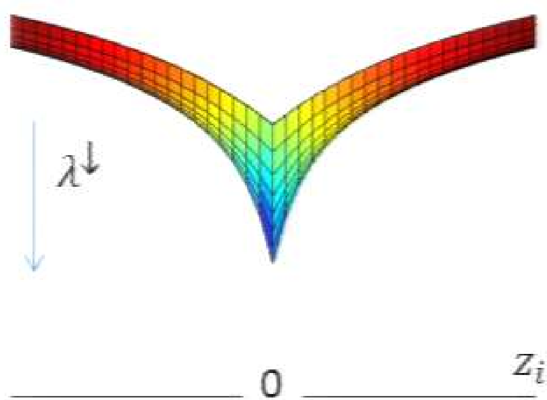




Model Properties
Noise Dependent Homotopy Continuation

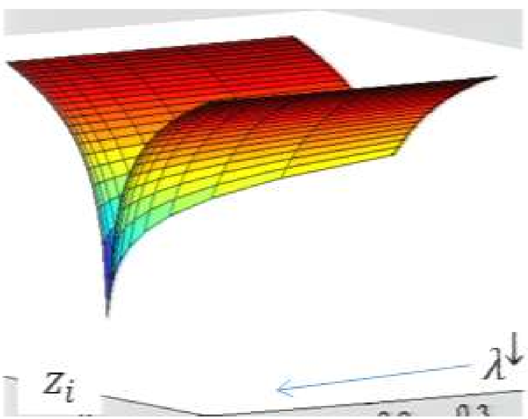
- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i ||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered

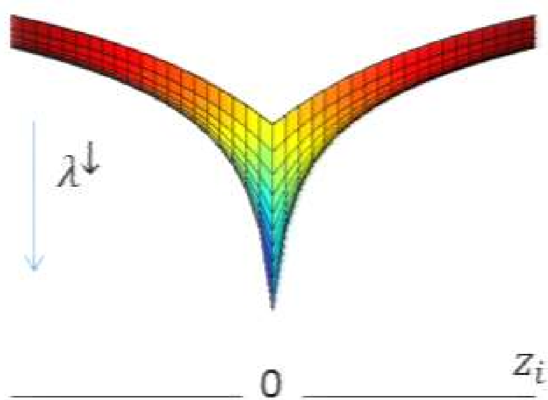




Model Properties
Noise Dependent Homotopy Continuation

- **Implications on Camera Shake Removal**
  - Initially, λ is large, penalty function is less concave
    - de-emphasize high blur regions (z_i small) $z_i = x_i ||h_i||_2$
    - focus first on large structure (x_i large), low blur ($||h_i||_2$ large) regions
  - Later, λ is reduced, relative concavity of ψ is increased, more fine details will be recovered





Model Properties
Tuning Parameter Free

- **The proposed cost function**

$$\min_{\mathbf{z};\gamma,\mathbf{w},\lambda\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\tilde{\mathbf{H}}\mathbf{z}\|_2^2+\sum_i \psi(|z_i|,\lambda)+(n-m)\log\lambda$$

  – Learning λ
  – Tuning parameter free

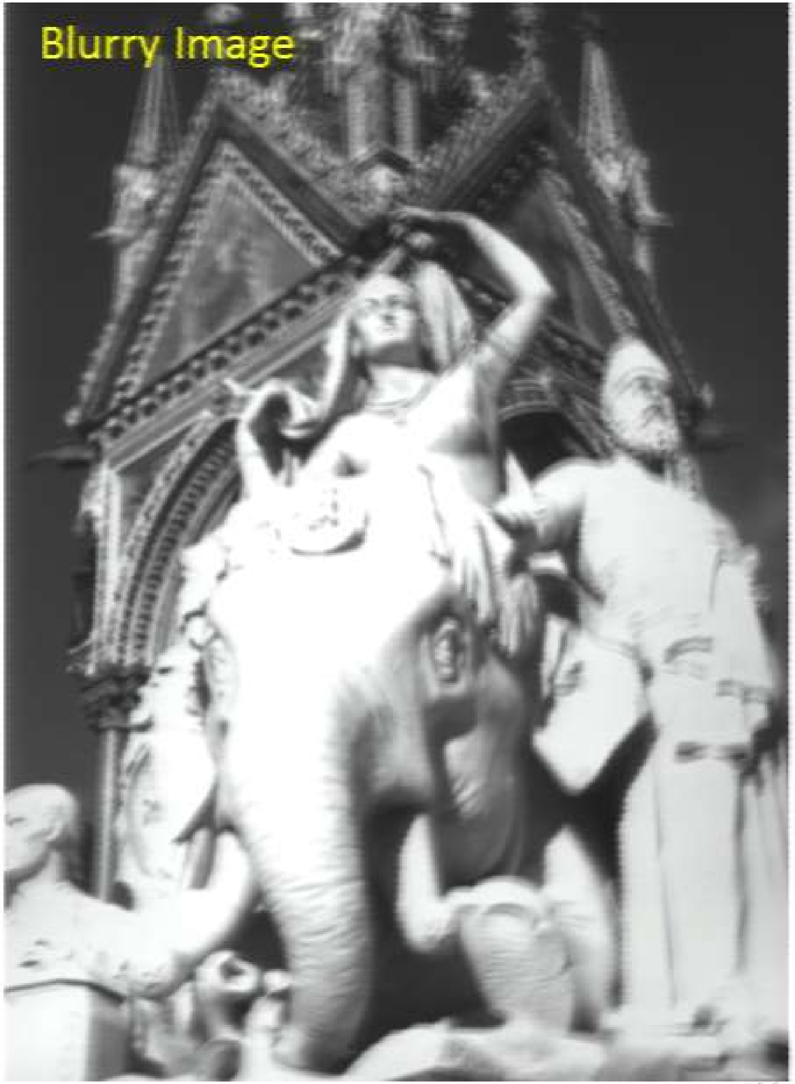
Experiments

- Test Images
  - Real-world blurry images from literature
- Compared Methods
  - Harmeling et al. *NIPS* 2010
  - Whyte et al. *CVPR* 2010
  - Gupta et al. *ECCV* 2010
  - Hirsch et al. *ICCV* 2011
  - Joshi et al. *SIGGRAPH* 2010 [hardware asisted]
  - Cho et al. *Pacific Graphics* 2012 [dual image]

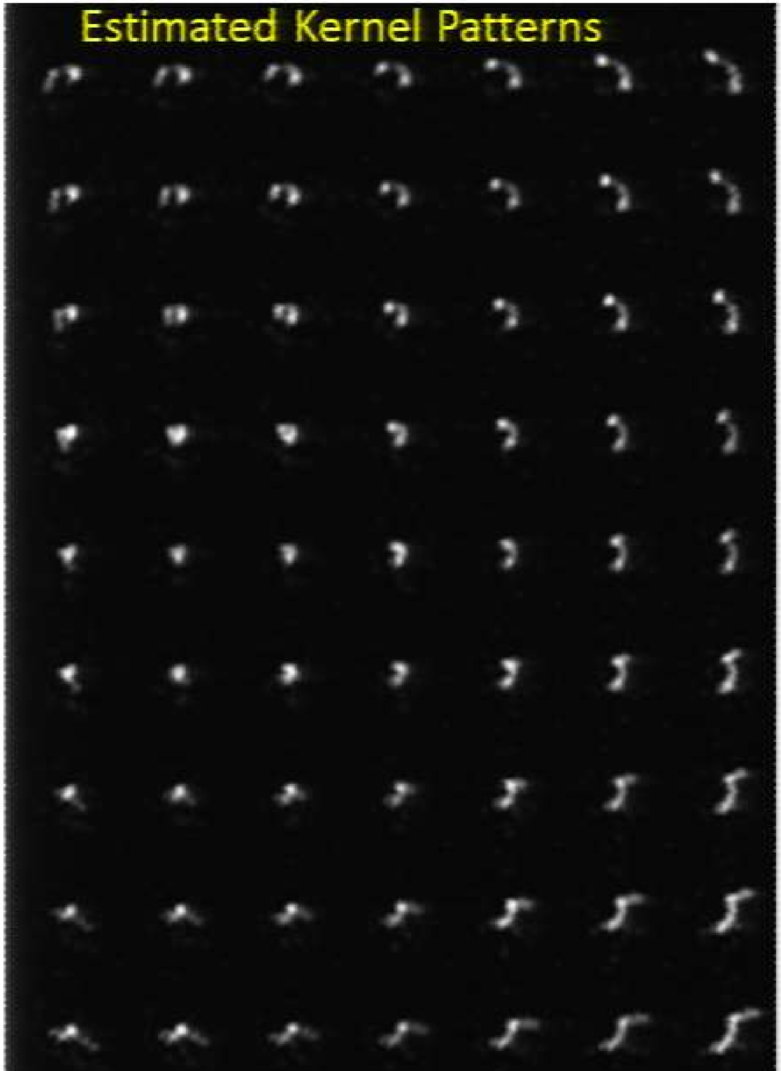
All the compared results are from the original authors

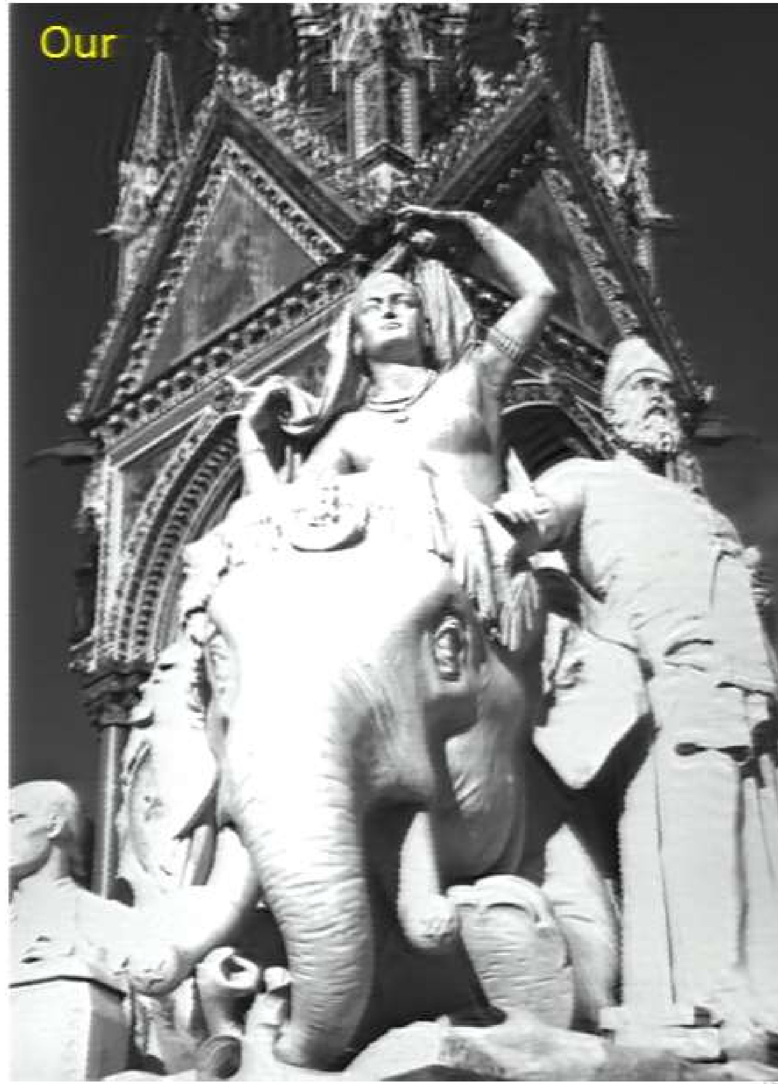
Experimental Results
An illustration

A test blurry image from Harmeling et al. , *NIPS* 2010.

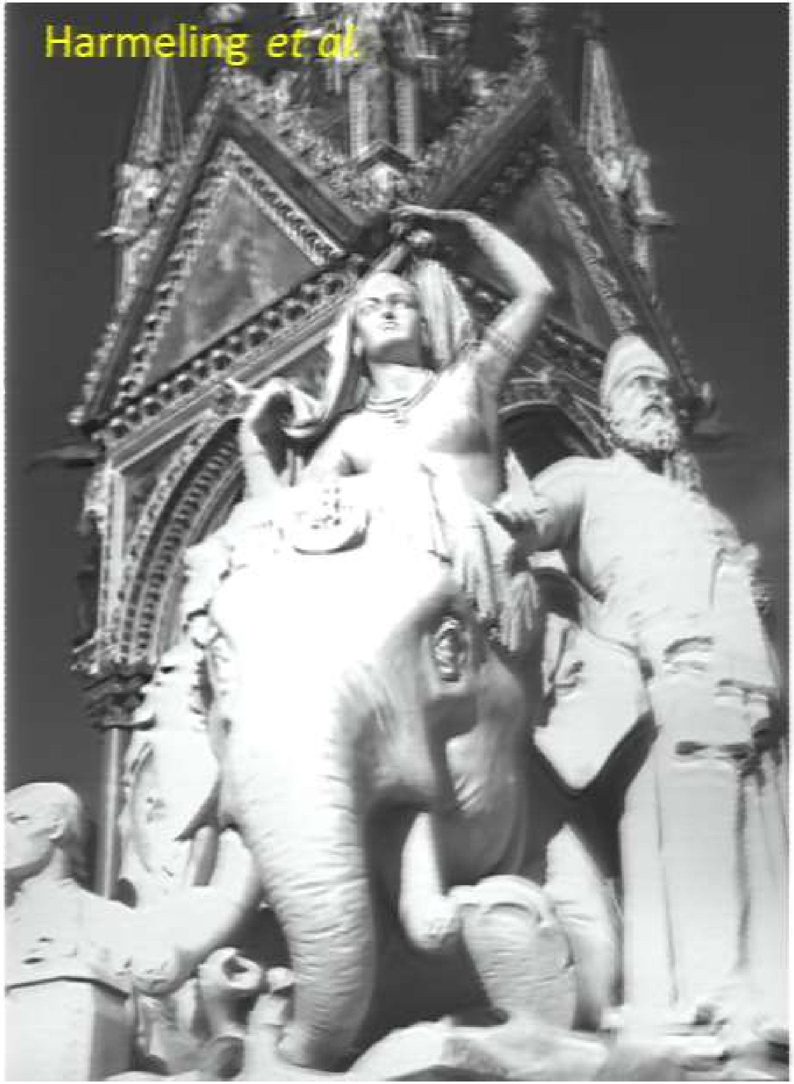


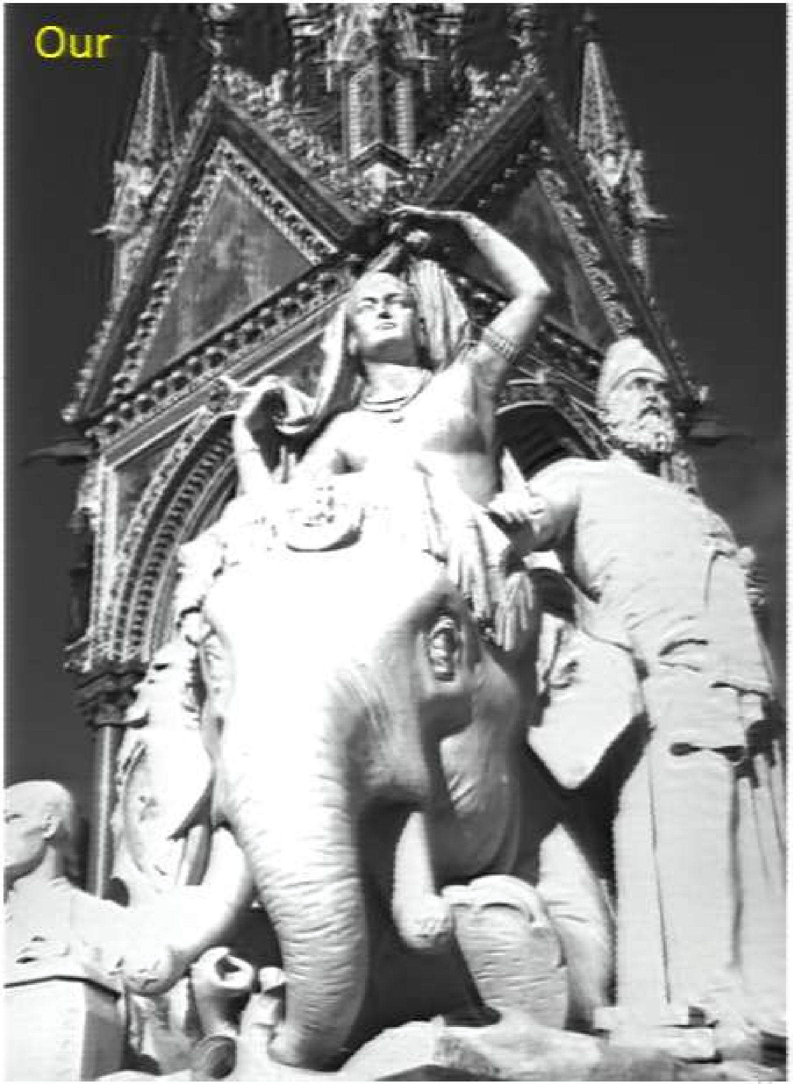
Experimental Results
An illustration





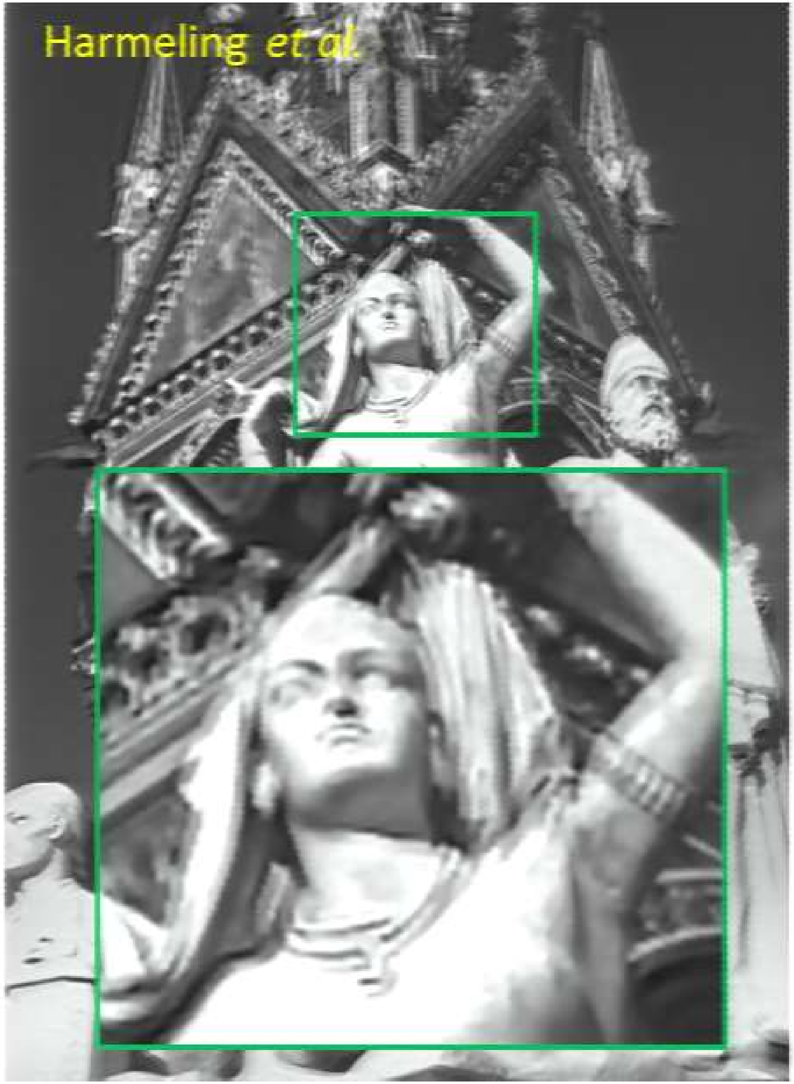
Experimental Results
comparison with Harmeling *et al.* NIPS'10

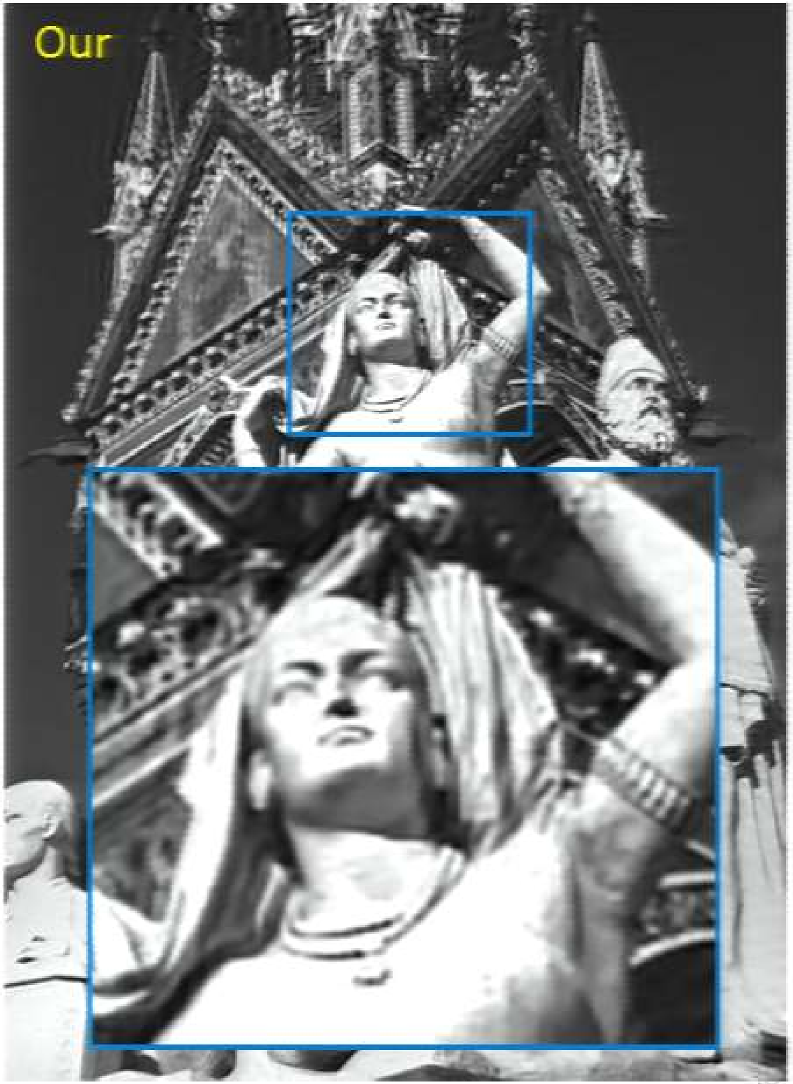




Experimental Results

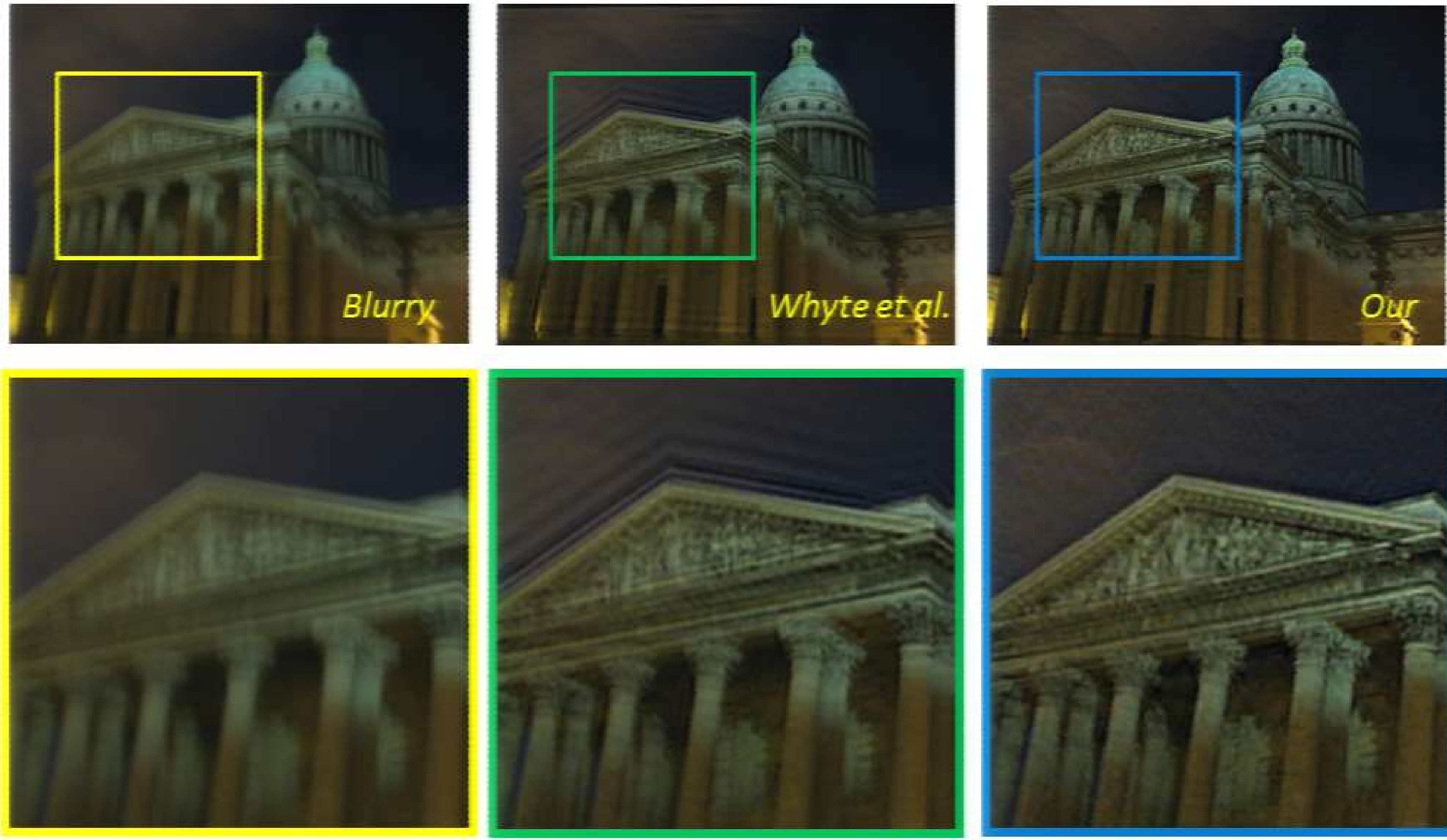
comparison with Harmeling *et al.* NIPS'10





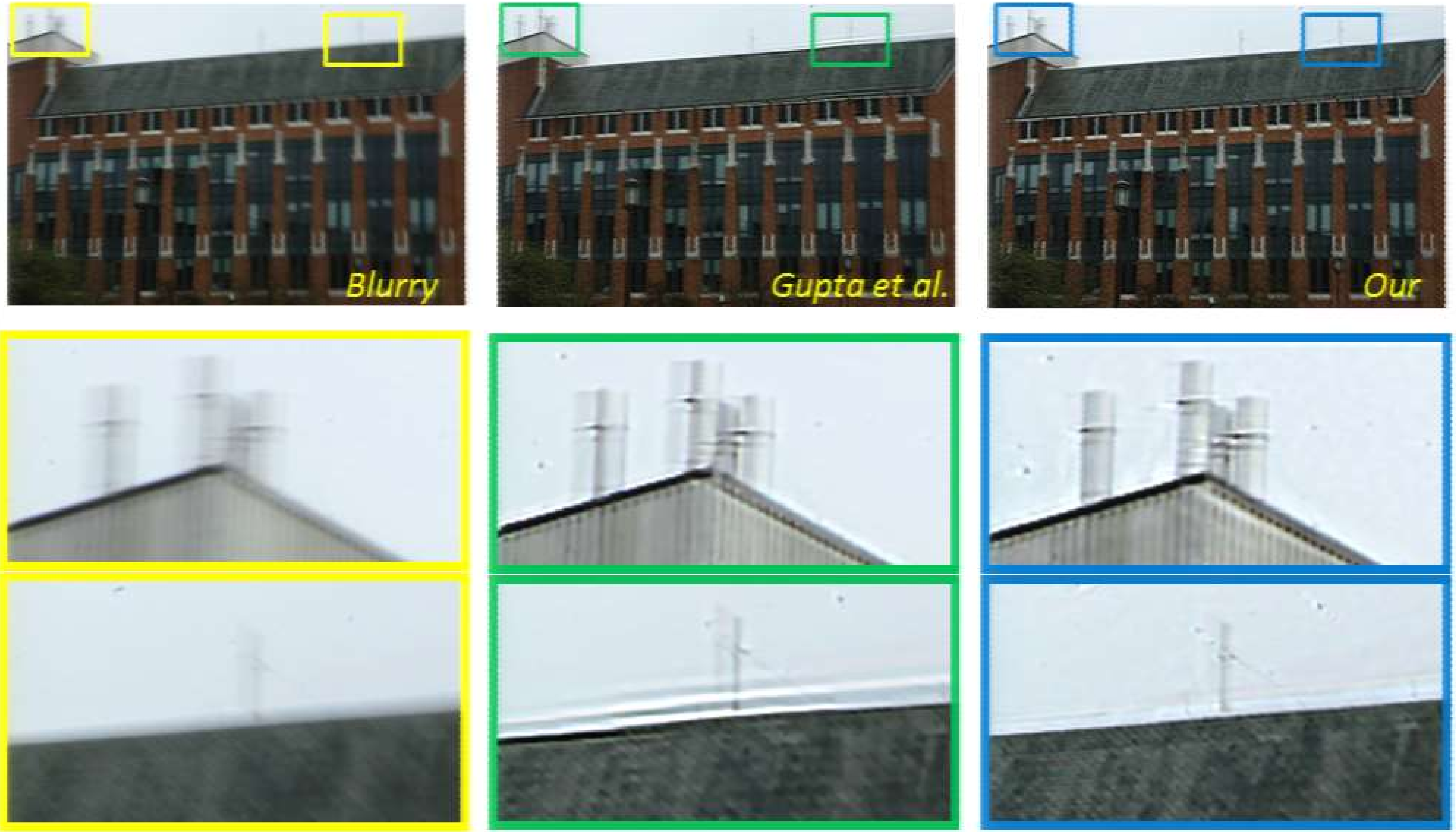
Experimental Results

comparison with Whyte *et al.* CVPR'10



O. Whyte et al. , *Non-uniform deblurring for shaken images*, CVPR, 2010.

Experimental Results
comparison with Gupta *et al.* ECCV'10

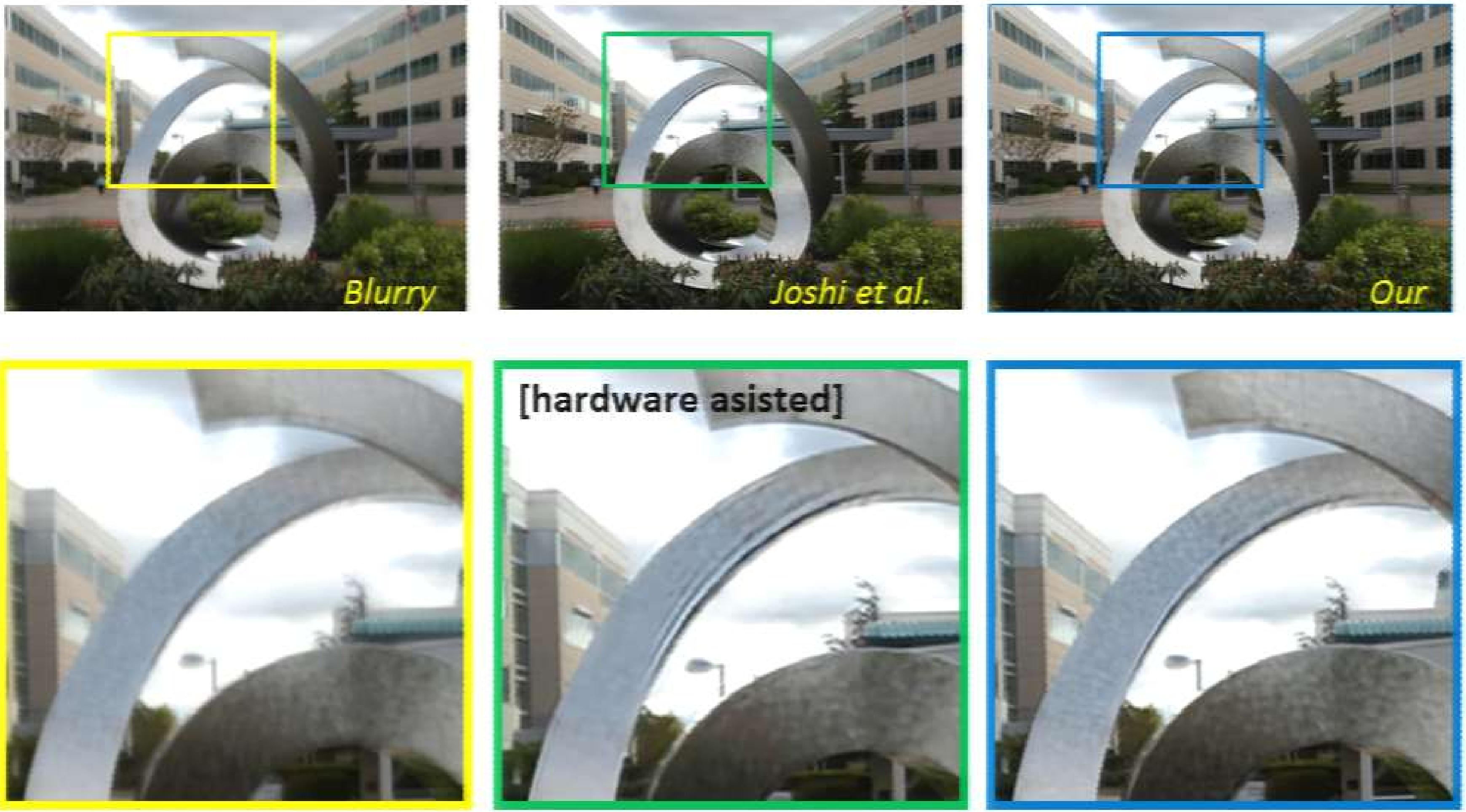


Gupta et al. , *Single image deblurring using motion density functions*, ECCV, 2010.

Experimental Results
comparison with Hirsch *et al.* ICCV'11

S. Hirsch et al. , *Fast removal of non-uniform camera shake*, ICCV, 2011.

Experimental Results
comparison with Joshi *et al.* SIGGRAPH'10



N. Joshi et al. , *Image deblurring using inertial measurement sensors*, SIGGRAPH, 2010.

Experimental Results
comparison with Cho *et al*. PG'12

Cho et al. , *Registration Based Non-uniform Motion Deblurring*, Pacific Graphics, 2012.

Experimental Results

comparison with Cho *et al*. PG'12

Summary

- An effective approach for camera shake removal
  - simple & clear cost function
- Model property analysis
  - automated column normalization (spatially adaptive sparsity): high-bur, low structure regions will be de-emphasized first, and emphasized progressively later
  - noise dependent homotopy continuation
  - tuning parameter free
- State-of-the-art performance on real-images
- Applicable to other problems (e.g., structured dictionary learning)

Non-Uniform Camera Shake Removal

Thank you!

Questions?

Welcome to our poster Fri26

NIPS

Experimental Results
comparison with Cho *et al.* PG'12

Summary

- An effective approach for camera shake removal
  - simple & clear cost function
- Model property analysis
  - automated column normalization (spatially adaptive sparsity): high-bur, low structure regions will be de-emphasized first, and emphasized progressively later
  - noise dependent homotopy continuation
  - tuning parameter free
- State-of-the-art performance on real-images
- Applicable to other problems (e.g., structured dictionary learning)

Thank you!

Questions?

Welcome to our poster Fri26

N I
P S