# Gradient descents and inner products

(Parenthesis)

Guillaume Charpiat

# Gradient descent and inner products

- Gradient (definition)
- Usual inner product ($L^2$)
- Natural inner products
- Other inner products
- Examples

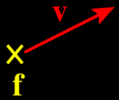
# Gradient (definition)

Energy $E$ depending on a variable $\mathbf{f}$ (vector or function)

- gradient $\nabla E(\mathbf{f}) = ?$

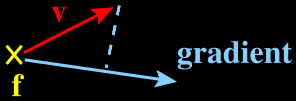
**Directional derivative:** consider a variation $\mathbf{v}$ of $\mathbf{f}$.

- $$\delta E(\mathbf{f})(\mathbf{v}) := \lim_{\varepsilon \to 0} \frac{E(\mathbf{f} + \varepsilon \mathbf{v}) - E(\mathbf{f})}{\varepsilon}$$




Hopefully, $\delta E$ is linear and continuous wrt $\mathbf{v}$. Riesz representation theorem $\implies$ **gradient:** variation $\nabla E(\mathbf{f})$ such that

- $$\forall \mathbf{v}, \;\; \delta E(\mathbf{f})(\mathbf{v}) = \langle \nabla E(\mathbf{f}) \,|\, \mathbf{v} \rangle$$




Usual **inner product** $\langle \cdot | \cdot \rangle$: $L^2$

## Gradient Descent Scheme

- Build minimizing path:

$$\mathbf{f}_{t=0} = \mathbf{f}_0$$

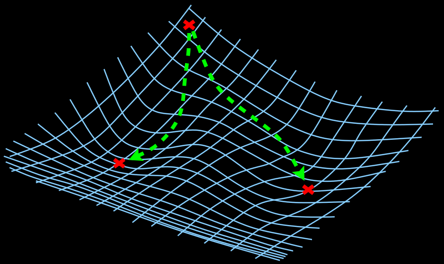
$$\frac{\partial \mathbf{f}}{\partial t} = -\nabla^P_{\mathbf{f}} E(\mathbf{f})$$

## Gradient Descent Scheme

- Build minimizing path:

$$\mathbf{f}_{t=0} = \mathbf{f}_0$$

$$\frac{\partial \mathbf{f}}{\partial t} = -\nabla^P_{\mathbf{f}} E(\mathbf{f})$$



- Change the inner product $P \implies$ change the minimizing path
- $-\nabla^P_{\mathbf{f}} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|^2_P \right\}$
- $P$ as a prior on the minimizing flow

**Example 1:** vectors, lists of parameters

$$E(\overrightarrow{\alpha}) = E(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_n)$$

- Usual $L^2$ gradient: $\nabla E(\overrightarrow{\alpha}) = \sum_i \partial_i E(\alpha_1, \alpha_2, ..., \alpha_n)\ \mathbf{e}_i$

  (supposes $\alpha_i$ independent)

## Example 1: vectors, lists of parameters

$$E(\overrightarrow{\alpha}) = E(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_n)$$

- Usual $L^2$ gradient: $\nabla E(\overrightarrow{\alpha}) = \sum_i \partial_i E(\alpha_1, \alpha_2, ..., \alpha_n)\ \mathbf{e}_i$ (supposes $\alpha_i$ independent)
- Should all parameters $\alpha_i$ have the same weights ?
- Consider a different parameterization $\overrightarrow{\beta} = P(\overrightarrow{\alpha})$: $\beta_1 = 10\ \alpha_1$ and $\beta_i = \alpha_i \ \ \forall i \neq 1$

$$F(\overrightarrow{\beta}) = E(P^{-1}(\overrightarrow{\beta})) = E(0.1\beta_1, \beta_2, ..., \beta_n)$$

**Example 1:** vectors, lists of parameters

$$E(\overrightarrow{\alpha}) = E(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_n)$$

- Usual $L^2$ gradient: $\nabla E(\overrightarrow{\alpha}) = \sum_i \partial_i E(\alpha_1, \alpha_2, ..., \alpha_n)\ \mathbf{e}_i$ (supposes $\alpha_i$ independent)
- Should all parameters $\alpha_i$ have the same weights ?
- Consider a different parameterization $\overrightarrow{\beta} = P(\overrightarrow{\alpha})$: $\beta_1 = 10\ \alpha_1$ and $\beta_i = \alpha_i \ \ \forall i \neq 1$

$$F(\overrightarrow{\beta}) = E(P^{-1}(\overrightarrow{\beta})) = E(0.1\beta_1, \beta_2, ..., \beta_n)$$

- $\forall i \neq 1, \ \ \partial_{\beta_i} F(\overrightarrow{\beta}) = \partial_{\alpha_i} E(\overrightarrow{\alpha})$

$$\partial_{\beta_1} F(\overrightarrow{\beta}) = 0.1\ \ \partial_{\alpha_1} E(\overrightarrow{\alpha})|_{\alpha=P^{-1}(\overrightarrow{\beta})}$$

$$E(\overrightarrow{\alpha}) = E(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_n)$$

$$\beta_1 = 10\ \alpha_1$$

$$\partial_{\beta_1} F(\overrightarrow{\beta}) = 0.1\ \ \partial_{\alpha_1} E(\overrightarrow{\alpha})|_{\alpha=P^{-1}(\overrightarrow{\beta})}$$

- gradient ascent wrt $\beta$: $\delta\beta_1 = \partial_{\beta_1} F(\overrightarrow{\beta})$
  $\implies \delta\beta_1 = 0.1\, \partial_{\alpha_1} E(\overrightarrow{\alpha})$
- gradient ascent wrt $\alpha$: $\delta\alpha_1 = \partial_{\alpha_1} E(\overrightarrow{\alpha})$
  $\implies \delta\beta_1 = 10\ \delta\alpha_1 = 10\ \partial_{\alpha_1} E(\overrightarrow{\alpha})$
- Difference between two approaches: factor 100 for the first parameter!
- Conclusion: $L^2$ inner product is bad (not intrinsic)

## Example 2: (not specific to) kernels and intrinsic gradients

$$\mathbf{f} = \sum_i \alpha_i k(x_i, \cdot)$$

$$E(\mathbf{f}) = \sum_i |\mathbf{f}(x_i) - y_i|^2 + \|\mathbf{f}\|_H^2$$

- $L^2$ gradient wrt $\overrightarrow{\alpha}$: not the best idea
- Choose a “natural”, intrinsic inner product:

$$\langle \delta\overrightarrow{\alpha}_a | \delta\overrightarrow{\alpha}_b \rangle_P = \langle \delta\mathbf{f}_a | \delta\mathbf{f}_b \rangle_H = \langle D\mathbf{f}(\delta\overrightarrow{\alpha}_a) | D\mathbf{f}(\delta\overrightarrow{\alpha}_b) \rangle_H = \left\langle \delta\overrightarrow{\alpha}_a \right| {}^t D\mathbf{f}\, D\mathbf{f} \left| \delta\overrightarrow{\alpha}_b \right\rangle_H$$

- In case of kernels this gives:

$$\langle \delta\overrightarrow{\alpha}_a | \delta\overrightarrow{\alpha}_b \rangle_P = \left\langle \delta\overrightarrow{\alpha}_a \right| K \left| \delta\overrightarrow{\alpha}_b \right\rangle$$

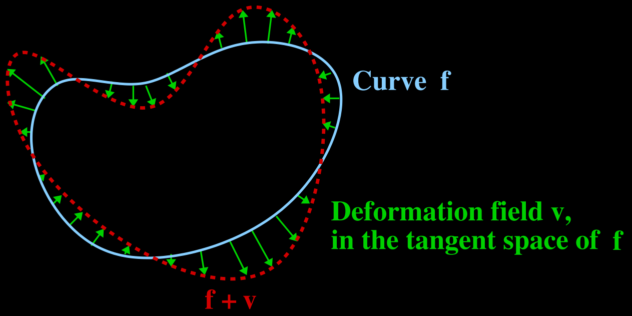
- Corresponding gradient:

$$\nabla^P_{\overrightarrow{\alpha}} E = \left({}^t D\mathbf{f}\, D\mathbf{f}\right)^{-1} \left(P_{adm}(\nabla^H_{\mathbf{f}} E)\right)$$

**Example 3:** parameterized shapes

- The same as for kernels
- Choose any representation for shapes (splines, polygon, level-set...)
- Intrinsic gradient $\Longrightarrow$ as close as possible to the "real" evolution, most independent as possible of the representation/parameterization

## Example 4: shapes, contours




Usual tangent space: $L^2(\mathbf{f})$:

$$\langle \mathbf{u} | \mathbf{v} \rangle_{L^2(\mathbf{f})} = \int_{\mathbf{f}} \mathbf{u}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}) \, d\mathbf{f}(\mathbf{x})$$

## Examples of inner products for shapes (or functions)

- $L^2$ inner product: $\langle \mathbf{u} | \mathbf{v} \rangle_{L^2(\mathbf{f})} = \int_{\mathbf{f}} \mathbf{u}(x) \cdot \mathbf{v}(x) \, d\mathbf{f}(x)$

## Examples of inner products for shapes (or functions)

- $L^2$ inner product: $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{L^2(\mathbf{f})} = \int_{\mathbf{f}} \mathbf{u}(x) \cdot \mathbf{v}(x) \, d\mathbf{f}(x)$
- $H^1$ inner product: $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{H^1} = \langle \mathbf{u} \,|\, \mathbf{v} \rangle_{L^2(\mathbf{f})} + \langle \partial_x \mathbf{u} \,|\, \partial_x \mathbf{v} \rangle_{L^2(\mathbf{f})}$

Interesting property of the $H^1$ gradient:

$$\nabla^{H^1} E = arg \inf_{\mathbf{u}} \left\| \mathbf{u} - \nabla^{L^2} E \right\|_{L^2}^2 + \|\partial_x \mathbf{u}\|_{L^2}^2$$

## Examples of inner products for shapes (or functions)

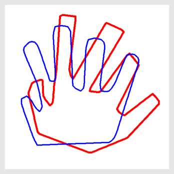
- $L^2$ inner product: $\langle \mathbf{u} | \mathbf{v} \rangle_{L^2(\mathbf{f})} = \int_{\mathbf{f}} \mathbf{u}(x) \cdot \mathbf{v}(x) \, d\mathbf{f}(x)$
- $H^1$ inner product: $\langle \mathbf{u} | \mathbf{v} \rangle_{H^1} = \langle \mathbf{u} | \mathbf{v} \rangle_{L^2(\mathbf{f})} + \langle \partial_x \mathbf{u} | \partial_x \mathbf{v} \rangle_{L^2(\mathbf{f})}$
- Set $S$ of prefered transformations (e.g. rigid motion)
  Projection on $S$: $Q$
  Projection orthogonal to $S$: $R \quad (Q + R = Id)$

$$\langle \mathbf{u} | \mathbf{v} \rangle_S = \langle Q(\mathbf{u}) | Q(\mathbf{v}) \rangle_{L^2} + \alpha \langle R(\mathbf{u}) | R(\mathbf{v}) \rangle_{L^2}$$

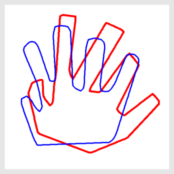
## Examples of inner products for shapes (or functions)

- $L^2$ inner product: $\langle \mathbf{u} | \mathbf{v} \rangle_{L^2(\mathbf{f})} = \int_{\mathbf{f}} \mathbf{u}(x) \cdot \mathbf{v}(x)\, d\mathbf{f}(x)$
- $H^1$ inner product: $\langle \mathbf{u} | \mathbf{v} \rangle_{H^1} = \langle \mathbf{u} | \mathbf{v} \rangle_{L^2(\mathbf{f})} + \langle \partial_x \mathbf{u} | \partial_x \mathbf{v} \rangle_{L^2(\mathbf{f})}$
- Set of prefered transformations (e.g. rigid motion)
- Example: minimizing the Hausdorff distance between two curves

$$\frac{\partial \mathbf{f}}{\partial t} = -\nabla_{\mathbf{f}}\, d_H(\mathbf{f}, \mathbf{f}_2)$$



usual



rigidified

## Examples of inner products for shapes (or functions)

- $L^2$ inner product: $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{L^2(\mathbf{f})} \; = \; \int_{\mathbf{f}} \mathbf{u}(x) \cdot \mathbf{v}(x) \, d\mathbf{f}(x)$
- $H^1$ inner product: $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{H^1} \; = \; \langle \mathbf{u} \,|\, \mathbf{v} \rangle_{L^2(\mathbf{f})} + \langle \partial_x \mathbf{u} \,|\, \partial_x \mathbf{v} \rangle_{L^2(\mathbf{f})}$
- Set of prefered transformations (e.g. rigid motion)
- Example: minimizing the Hausdorff distance between two curves
- Change an inner product for another one: linear symmetric positive definite transformation of the gradient

## Examples of inner products for shapes (or functions)

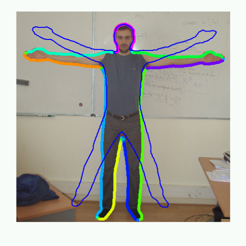
- $L^2$ inner product: $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{L^2(\mathbf{f})} = \int_{\mathbf{f}} \mathbf{u}(x) \cdot \mathbf{v}(x) \, d\mathbf{f}(x)$
- $H^1$ inner product: $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{H^1} = \langle \mathbf{u} \,|\, \mathbf{v} \rangle_{L^2(\mathbf{f})} + \langle \partial_x \mathbf{u} \,|\, \partial_x \mathbf{v} \rangle_{L^2(\mathbf{f})}$
- Set of prefered transformations (e.g. rigid motion)
- Example: minimizing the Hausdorff distance between two curves
- Change an inner product for another one: linear symmetric positive definite transformation of the gradient
- Gaussian smoothing of the $L^2$ gradient: symmetric positive definite

## Extension to non-linear criteria

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

## Extension to non-linear criteria

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$
- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \{ \delta E(\mathbf{f})(\mathbf{v}) + R_P(\mathbf{v}) \}$
- Example: semi-local rigidification



## Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

## Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \implies$ different gradient descents:

## Remark on Taylor series and Newton method

- $-\nabla^P_{\mathbf{f}} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|^2_P \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \quad \Longrightarrow \quad$ different gradient descents:

- $\arg\min_{\mathbf{v}} \{E(\mathbf{f})\}$

## Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \implies$ different gradient descents:

- $\arg\min_{\mathbf{v}} \{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) \}$

## Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_{P}^{2} \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \implies$ different gradient descents:

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_{P}^{2} \right\}$

# Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \implies$ different gradient descents:

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\} = -\nabla_{\mathbf{f}}^{P} E(\mathbf{f})$

  1st order + metric choice$^2$ = quadratic

## Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \implies$ different gradient descents:

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\} = -\nabla_{\mathbf{f}}^{P} E(\mathbf{f})$

  1st order + metric choice$^2$ = quadratic

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \delta^2 E(\mathbf{f})(\mathbf{v})(\mathbf{v}) \right\}$

## Remark on Taylor series and Newton method

- $-\nabla^P_{\mathbf{f}} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|^2_P \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v}) \implies$ different gradient descents:

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|^2_P \right\} = -\nabla^P_{\mathbf{f}} E(\mathbf{f})$

  1st order + metric choice$^2$ = quadratic

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \delta^2 E(\mathbf{f})(\mathbf{v})(\mathbf{v}) \right\}$

  quadratic : order-2 Newton

## Remark on Taylor series and Newton method

- $-\nabla_{\mathbf{f}}^{P} E(\mathbf{f}) = \arg\min_{\mathbf{v}} \left\{ \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\}$

Different approximations of $E(\mathbf{f}+\mathbf{v})$ $\implies$ different gradient descents:

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \|\mathbf{v}\|_P^2 \right\} = -\nabla_{\mathbf{f}}^{P} E(\mathbf{f})$

  1st order + metric choice$^2$ = quadratic

- $\arg\min_{\mathbf{v}} \left\{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + \frac{1}{2} \delta^2 E(\mathbf{f})(\mathbf{v})(\mathbf{v}) \right\}$

  quadratic : order-2 Newton

- $\arg\min_{\mathbf{v}} \{ E(\mathbf{f}) + \delta E(\mathbf{f})(\mathbf{v}) + R_P(\mathbf{v}) \}$

  1st order + any suitable criterion

(End of the parenthesis)