Лекция 8. Общее решение для недиагональной регуляризации

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

План лекции

Дифференцирование скалярной функции

- Пусть $f(\boldsymbol{x})$ — некоторая скалярная функция, зависящая от вектора $\boldsymbol{x} \in \mathbb{R}^n$. Тогда ее производная по вектору по определению есть

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right) = \nabla f(\boldsymbol{x})$$

- Пусть $f(A)$ — скалярная функция, зависящая от матрицы $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. Производной функции по матрице назовем матрицу производных по соответствующим элементам A:

$$\frac{\partial f(A)}{\partial A} = \left(\frac{\partial f(A)}{\partial a_{ij}} \right) \in \mathbb{R}^{n \times n}$$

Примеры дифференцирования по матрице

- Производная следа матрицы

$$\frac{\partial\,\mathrm{tr}(AB)}{\partial A} = B^T, \quad \frac{\partial\,\mathrm{tr}(A^TB)}{\partial A} = B$$

- Выведем производную определителя матрицы $\frac{\partial \det(A)}{\partial A}$. Для этого распишем определитель по строке

$$\det(A) = \sum_{i=1}^{n} a_{ij}M_{ij},$$

где $M_{ij} = (-1)^{i+j-1}\det(A^{ij})$ — алгебраическое дополнение, а A^{ij} — матрица, полученная из A путем вычеркивания i-ой строки и j-го столбца. Тогда, учитывая, что M_{ik} не зависит от a_{ij} для любых $k \neq j$, получаем

$$\frac{\partial \det(A)}{\partial a_{ij}} = \frac{\partial \sum_{i=1}^{n} a_{ij}M_{ij}}{\partial a_{ij}} = M_{ij}.$$

Каждый элемент матрицы A^{-1} выражается через алгебраические дополнения матрицы A как $a_{ij}^{-1} = \frac{1}{\det(A)}M_{ji}$, отсюда

$$\frac{\partial \det(A)}{\partial A} = \det(A)A^{-T}$$

Дифференцирование многомерной функции

- Пусть $\boldsymbol{f}(x) = (f_1(x), \ldots, f_m(x))^T$ — некоторая векторная функция от скалярной переменной $x \in \mathbb{R}$. Тогда ее производная по аргументу по определению есть

$$\frac{\partial \boldsymbol{f}(x)}{\partial x} = \left(\frac{\partial f_1(x)}{\partial x}, \ldots, \frac{\partial f_m(x)}{\partial x}\right)^T$$

- Пусть $A(x) = (a_{ij}(x)) \in \mathbb{R}^{n \times n}$ — квадратная матрица, зависящая от параметра x. Тогда ее производная по параметру по определению равна

$$\frac{\partial A(x)}{\partial x} = \left(\frac{\partial a_{ij}(x)}{\partial x}\right)$$

- В частности, выписывая выражения покоординатно можно показать, что

$$\frac{\partial AB}{\partial x} = \frac{\partial A}{\partial x}B + A\frac{\partial B}{\partial x} \qquad \frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$$

$$\frac{\partial \log \det(A)}{\partial x} = \mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

План лекции

Постановка задачи

- Имеется X ресурсов, которые нужно распределить по m целям. Эффект от выделения x_j ресурсов на цель j выражается целевой функцией $h_j(x_j)$
- Как правило предполагается, что целевые функции неубывающие
- Задача максимизировать суммарный эффект

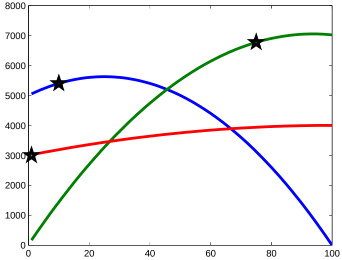
$$\begin{cases} \sum_{j=1}^m h_j(x_j) \to \max_{x_1,\dots,x_m} \\ \sum_{j=1}^m x_j = X, \quad x_j \geq 0 \end{cases}$$

Лемма Гиббса

Лемма. Пусть $\boldsymbol{x}^0$ — оптимальное решение задачи распределения ресурсов и все целевые функции дифференцируемы на участке $[0, X]$. Тогда найдется такое число λ, что выполнено следующее необходимое условие

$$\begin{cases} h_j'(x_j^0) = \lambda, & x_j^0 > 0 \\ h_j'(x_j^0) \le \lambda, & x_j^0 = 0 \end{cases}$$

Замечание. Если все целевые функции вогнуты, то это необходимое условие является и достаточным



План лекции

Регуляризация обобщенных линейных моделей

Регрессия	Классификация
Выборка $\{\boldsymbol{x}_n, t_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^d$, $t_n \in \mathbb{R}$	Выборка $\{\boldsymbol{x}_n, t_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^d$, $t_n \in \{-1, +1\}$
Решающее правило: $y(\boldsymbol{x}) = \sum_{j=1}^M w_j\phi_j(\boldsymbol{x})$	Решающее правило: $y(\boldsymbol{x}) = \mathrm{sign}(f(\boldsymbol{x})) = \mathrm{sign}\left(\sum_{j=1}^M w_j\phi_j(\boldsymbol{x})\right)$
Функция правдоподобия: $p(t_n\|\boldsymbol{x}_n) = \mathcal{N}(t_n\|y(\boldsymbol{x}_n), \sigma^2)$	Функция правдоподобия: $p(t_n\|\boldsymbol{x}_n) = \frac{1}{1+\exp(-t_n f(\boldsymbol{x}_n))}$
Поиск весов: $\log p(\boldsymbol{t}\|X, \boldsymbol{w}) + \log p(\boldsymbol{w}\|A) \to \max_{\boldsymbol{w}}$	Поиск весов: $\log p(\boldsymbol{t}\|X, \boldsymbol{w}) + \log p(\boldsymbol{w}\|A) \to \max_{\boldsymbol{w}}$
Регуляризация: $p(\boldsymbol{w}\|A) = \mathcal{N}(\boldsymbol{w}\|\boldsymbol{0}, A)$	Регуляризация: $p(\boldsymbol{w}\|A) = \mathcal{N}(\boldsymbol{w}\|\boldsymbol{0}, A)$
Поиск A: $p(\boldsymbol{t}\|X, A) = \int p(\boldsymbol{t}\|X, \boldsymbol{w})p(\boldsymbol{w}\|A)d\boldsymbol{w} \to \max_A$	Поиск A: $p(\boldsymbol{t}\|X, A) = \int p(\boldsymbol{t}\|X, \boldsymbol{w})p(\boldsymbol{w}\|A)d\boldsymbol{w} \to \max_A$

Виды гауссовской регуляризации

Регуляризатор:

$$p(\boldsymbol{w}|A) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, A) = \frac{\sqrt{\det A}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right)$$

- Байесовская линейная/логистическая регрессия:

$$A \in \{A | A = \alpha I,\ \alpha \geq 0\}$$

- Метод релевантных векторов

$$A \in \{A | A = \mathrm{diag}(\alpha_1, \dots, \alpha_M),\ \alpha_j \geq 0\}$$

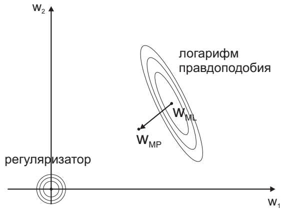
- Недиагональная регуляризация

$$\exists Q : Q^T = Q^{-1},\ QHQ^T = \Lambda_1 = \mathrm{diag}(h_1, \dots, h_M),\ QAQ^T = \Lambda_2 = \mathrm{diag}(\lambda_1, \dots, \lambda_M)$$

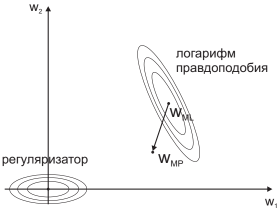
- Общее решение

$$A \in \{A | A = A^T, A \succ 0\}$$

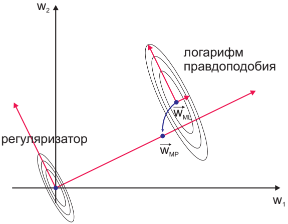
Иллюстрация видов гауссовской регуляризации



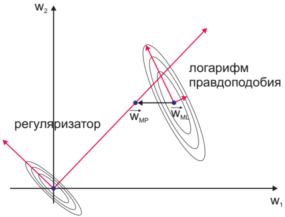
Байесовская регрессия



Метод релевантных векторов



Недиагональная регуляризация



Общее решение

Общая постановка задачи

- Очевидно, что ни байесовская логистическая регрессия, ни метод релевантных векторов не покрывают все возможные гауссовские априорные распределения на множество весов $\boldsymbol{w}$
- Рассмотрим задачу поиска наиболее обоснованного распределения во всем классе нормальных распределений

$$A = \arg\max_{A\in\mathcal{A}} p(\boldsymbol{t}|X, A) = \arg\max_{A\in\mathcal{A}} \int p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w}|A)d\boldsymbol{w},$$

где $\mathcal{A} = \{A | A^T = A,\ A \succeq 0\}$

Приближение Лапласа для правдоподобия

- Используем метод Лапласа для того, чтобы приблизить правдоподобие гауссианой
- Пусть $H = \nabla\nabla - \log p(\boldsymbol{t}|X,\boldsymbol{w})|_{\boldsymbol{w}_{ML}}$ — отрицательный гессиан логарифма правдоподобия, взятый в точке максимума, тогда

$$p(\boldsymbol{t}|X,\boldsymbol{w}) \approx$$

$$\hat{p}(\boldsymbol{t}|X,\boldsymbol{w}) = p(\boldsymbol{t}|X,\boldsymbol{w}_{ML})\exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{ML})^T H(\boldsymbol{w}-\boldsymbol{w}_{ML})\right)$$

- Обозначим

$$Q(\boldsymbol{w}) = \hat{p}(\boldsymbol{t}|X,\boldsymbol{w})p(\boldsymbol{w}|A) = \hat{p}(\boldsymbol{t}|X,\boldsymbol{w})\frac{\sqrt{\det(A)}}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w}\right),$$

тогда легко показать (Упр.), что выражение для обоснованности принимает вид

$$E(A) \approx \frac{Q(\boldsymbol{w}_{MP})(2\pi)^{m/2}}{\sqrt{\det(-\nabla\nabla\log Q(\boldsymbol{w})|_{\boldsymbol{w}_{MP}})}}$$

Вид оптимизируемого функционала

- Для упрощения выкладок, перейдем к рассмотрению логарифма обоснованности, очевидно, что

$$A = \arg\max_{A\in\mathcal{A}} E(A) = \arg\max_{A\in\mathcal{A}} \log E(A)$$

- Выражение для логарифма обоснованности имеет вид

$$\log E(A) \approx \log \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{MP}) - 0.5\boldsymbol{w}_{MP}^T A \boldsymbol{w}_{MP} + \\ + 0.5 \log\det\left((H+A)^{-1}A\right) + C \to \max_{A\in\mathcal{A}}$$

Задача поиска оптимальной матрицы в классе неотрицательно определенных (semi-definite programming) является нетривиальной и проблема разработки эффективного численного метода решения на настоящий момент является открытой

Компонента $\log\det(A)$ возникает из плотности $p(\boldsymbol{w}|A)$, являющейся множителем $Q(\boldsymbol{w})$, а $\det(H+A)$ — это определитель гессиана $\det(-\nabla\nabla \log Q(\boldsymbol{w})|_{\boldsymbol{w}_{ML}})$

План лекции

Выражение $\boldsymbol{w}_{MP}$ через $\boldsymbol{w}_{ML}$

- Обоснованность зависит от точки максимума регуляризованного правдоподобия $\boldsymbol{w}_{MP}$, которая на момент поиска наилучшего регуляризатора неизвестна
- Учитывая, что $\boldsymbol{w}_{MP}$ зависит от выбранной матрицы регуляризации A, получим явный вид этой зависимости

$$Q(\boldsymbol{w}) = \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{ML})^T H(\boldsymbol{w}-\boldsymbol{w}_{ML})\right) \times \frac{\sqrt{\det(A)}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right)$$

$$\log Q(\boldsymbol{w}) = -0.5\left[(\boldsymbol{w}-\boldsymbol{w}_{ML})^T H(\boldsymbol{w}-\boldsymbol{w}_{ML}) + \boldsymbol{w}^T A\boldsymbol{w} - \log\det(A))\right] + \log\hat{p}(\boldsymbol{t}|X,\boldsymbol{w}_{ML}) - \frac{m}{2}\log(2\pi)$$

$$\frac{\partial \log Q(\boldsymbol{w})}{\partial \boldsymbol{w}} = -H(\boldsymbol{w}-\boldsymbol{w}_{ML}) - A\boldsymbol{w} = -(H+A)\boldsymbol{w} + H\boldsymbol{w}_{ML}$$

- В точке $\boldsymbol{w} = \boldsymbol{w}_{MP}$ производная регуляризованного правдоподобия равна нулю, отсюда

$$\boldsymbol{w}_{MP} = (H+A)^{-1} H \boldsymbol{w}_{ML}.$$

Выражение обоснованности через точку максимума правдоподобия I

- Подставим формулу для $\boldsymbol{w}_{MP}$ в выражение для обоснованности

$$\log E(A) = 0.5 \log \det((H+A)^{-1}A) - \frac{m}{2}\log(2\pi) + \log \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) -$$

$$-0.5 \left[(\boldsymbol{w}_{MP} - \boldsymbol{w}_{ML})^T H (\boldsymbol{w}_{MP} - \boldsymbol{w}_{ML}) + \boldsymbol{w}_{MP}^T A \boldsymbol{w}_{MP}\right]$$

- Учитывая, что матрицы H и $(H+A)$ симметричные, $\boldsymbol{w}_{MP}^T = \boldsymbol{w}_{ML}^T H (H+A)^{-1}$
- Разность $\boldsymbol{w}_{MP} - \boldsymbol{w}_{ML}$ может быть записана в матричном виде как

$$\boldsymbol{w}_{MP} - \boldsymbol{w}_{ML} = \left((H+A)^{-1}H - I\right)\boldsymbol{w}_{ML}$$

Выражение обоснованности через точку максимума правдоподобия II

Результат подстановки

$$-0.5\boldsymbol{w}_{ML}^T\left[\left\{H(H+A)^{-1}-I\right\}H\left\{(H+A)^{-1}H-I\right\}+ H(H+A)^{-1}A(H+A)^{-1}H\right]\boldsymbol{w}_{ML}=$$

$$-0.5\boldsymbol{w}_{ML}^T\left[H(H+A)^{-1}H(H+A)^{-1}H-2H(H+A)^{-1}H+H+ H(H+A)^{-1}A(H+A)^{-1}H\right]\boldsymbol{w}_{ML}=$$

$$-0.5\boldsymbol{w}_{ML}^T\left[H(H+A)^{-1}(H+A)(H+A)^{-1}H-2H(H+A)^{-1}H+H\right]\boldsymbol{w}_{ML}=$$

$$-0.5\boldsymbol{w}_{ML}^T\left[H(H+A)^{-1}H-2H(H+A)^{-1}H+H\right]\boldsymbol{w}_{ML}=$$

$$-0.5\boldsymbol{w}_{ML}^T\left[-H(H+A)^{-1}H+H\right]\boldsymbol{w}_{ML}.$$

Матричная хитрость

- Воспользуемся следующим матричным тождеством

$$H - H(H+A)^{-1}H = H(H+A)^{-1}((H+A)-H) = H(H+A)^{-1}A$$

- Тогда выражение для логарифма обоснованности (не забыв добавить $0.5 \log \det((H+A)^{-1}A)$) можно переписать

$$\log E(A) = \log \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) - \frac{m}{2}\log(2\pi) +$$

$$0.5[-\boldsymbol{w}_{ML}^T H(H+A)^{-1}A\boldsymbol{w}_{ML} + \log\det\left((H+A)^{-1}A\right)]$$

- Но и в таком виде оптимизация по A крайне затруднительна

Еще одна матричная хитрость

- Сделаем замену переменной $M = H(H+A)^{-1}A$, тогда

$$\log E(A) = 0.5[-\boldsymbol{w}_{ML}^T M \boldsymbol{w}_{ML} + \log\det\left((H+A)^{-1}A\right)] + C$$

- Используя свойство определителя произведения, перепишем второе слагаемое

$$\log\det\left((H+A)^{-1}A\right) = \log\det(M) - \log\det(H)$$

- Учитывая, что H не зависит от A, получаем

$$\log E(A) = 0.5[-\boldsymbol{w}_{ML}^T M \boldsymbol{w}_{ML} + \log\det(M)] + C_1,$$

но такое выражение имеет гораздо более простой вид!

Оптимизационная задача

После преобразований мы получили следующую задачу условной матричной оптимизации

$$
\begin{aligned}
&-\boldsymbol{w}_{ML}^T M \boldsymbol{w}_{ML} + \log\det M \to \max_M \\
&H^{-1} \preceq M^{-1} = M^{-T}
\end{aligned}
$$

где $M = H(H+A)^{-1}A$, или эквивалентно $M^{-1} = H^{-1} + A^{-1}$.
Это задача т.н. полуопределенного программирования (semi-definite programming), для которой, в общем случае, не существует эффективных методов решения.
К счастью, наша задача является частным случаем, для которого решение может быть получено в явном виде

План лекции

Идея решения I

Сделаем замену переменной $R = H^{-1/2}MH^{-1/2}$ и получим следующую оптимизационную задачу

$$\boldsymbol{v}^T R\boldsymbol{v} - \log\det R \to \min_R,$$
$$0 \preceq R = R^T \preceq I$$

где $\boldsymbol{v} = H^{1/2}\boldsymbol{w}_{ML}$.
Приведем R к диагональному виду с помощью ортогонального преобразования: $R = Q\Lambda Q^T$, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m)$, $Q^T = Q^{-1}$. Пусть $\boldsymbol{x} = Q^T\boldsymbol{v}$. Тогда оптимизационная задача принимает вид

$$\boldsymbol{x}^T\Lambda\boldsymbol{x} - \log\det\Lambda = \sum_{j=1}^m \left(x_j^2\lambda_j - \log\lambda_j\right) \to \min_{\boldsymbol{x},\boldsymbol{\lambda}},$$
$$0 \le \lambda_j \le 1,$$
$$\|\boldsymbol{x}\| = \|\boldsymbol{v}\|.$$

Оптимизируя функционал по $\boldsymbol{\lambda}$, получаем $\lambda_j^* = \min\left(1, \frac{1}{x_j^2}\right)$

Идея решения II

Подставляя $\boldsymbol{\lambda}^*$ обратно получаем

$$\sum_{j=1}^{m}\left[\min(x_j^2, 1) - \min\left(0, \log\frac{1}{x_j^2}\right)\right] \to \min_{\boldsymbol{x}},$$

$$\|\boldsymbol{x}\| = \|\boldsymbol{v}\|.$$

Но это же задача об оптимальном распределении ресурсов!

$$\sum_{j=1}^{m} h(y_j) \to \min_{\boldsymbol{y}}$$

$$\sum_{j=1}^{m} y_j = \|\boldsymbol{v}\|^2$$

здесь $y_j = x_j^2$, а целевая функция

$$h(y) = \begin{cases} y, & 0 \le y \le 1, \\ 1 + \log(y), & y \ge 1. \end{cases}$$

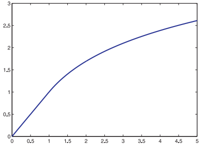
Идея решения III

Задача распределения ресурсов

$$\sum_{j=1}^m h(y_j) \to \min_{\boldsymbol{y}}$$
$$\sum_{j=1}^m y_j = \|\boldsymbol{v}\|^2$$

Целевая функция

$$h(y) = \begin{cases} y, & 0 \le y \le 1, \\ 1 + \log(y), & y \ge 1. \end{cases}$$



При $\|\boldsymbol{v}\| \ge 1$ все имеющиеся ресурсы выделяются единственной целевой функции. Не ограничивая общности (все целевые функции одинаковые), можно считать $x_1^2 = y_1 = \|\boldsymbol{v}\|^2$

При $\|\boldsymbol{v}\| < 1$ функционал $\sum_{j=1}^m h(y_j)$ остается постоянным и y_j может принимать любые значения между 0 и 1 с учетом ограничения $\sum_{j=1}^m y_j = \|\boldsymbol{v}\|^2$. Таким образом, все $x_j < 1$ и $\lambda_j^* = 1$.

Оптимальный регуляризатор

В первом случае легко видеть, что $\lambda_1^* = \|\boldsymbol{v}\|^{-2}$, $\lambda_j^* = 1$, $j \neq 1$. Вектор $\boldsymbol{x}$ является собственным для матрицы Λ, т.е. $\boldsymbol{v}$ является собственным для R.

$$R = H^{-1/2}MH^{-1/2} = H^{-1/2}(H^{-1} + A^{-1})^{-1}H^{-1/2} = (I + H^{1/2}A^{-1}H^{1/2})^{-1}$$

$$Q^T H^{1/2}A^{-1}H^{1/2}Q = \Lambda^{-1} - I = \mathrm{diag}(\|\boldsymbol{v}\|^2 - 1, 0, \ldots, 0) = \frac{\|\boldsymbol{v}\|^2 - 1}{\|\boldsymbol{v}\|^2}\boldsymbol{x}\boldsymbol{x}^T$$

$$A^{-1} = H^{-1/2}Q(\Lambda - I)Q^T H^{-1/2} = \frac{\|\boldsymbol{v}\|^2 - 1}{\|\boldsymbol{v}\|^2} H^{-1/2}Q\boldsymbol{x}\boldsymbol{x}^T Q^T H^{-1/2} =$$

$$= \frac{\|\boldsymbol{v}\|^2 - 1}{\|\boldsymbol{v}\|^2} H^{-1/2}\boldsymbol{v}\boldsymbol{v}^T H^{-1/2} = \frac{\boldsymbol{w}_{ML}^T H \boldsymbol{w}_{ML} - 1}{\boldsymbol{w}_{ML}^T H \boldsymbol{w}_{ML}} \boldsymbol{w}_{ML}\boldsymbol{w}_{ML}^T$$

Во втором случае все $\lambda_j^* = 1$, т.е. $R = I$, тогда

$$Q^T H^{1/2}A^{-1}H^{1/2}Q = \Lambda^{-1} - I = O$$

Отсюда $A^{-1} = O$, т.е. получается бесконечно большой регуляризатор, приводящий к нулевому решающему правилу. Физически это отвечает ситуации, когда в обучающей выборке не содержится значимых закономерностей в рамках выбранной системы базисных функций

Оптимальные веса решающего правила

Условие вырожденности решения:

$$\|\boldsymbol{v}\|^2 \leq 1 \Leftrightarrow \boldsymbol{w}_{ML}^T H \boldsymbol{w}_{ML} \leq 1$$

Если выполнено условие вырожденности, то решающее правило будет нулевым и для байесовской регрессии, и для метода релевантных векторов, и для недиагональной регуляризации.

Оптимальные веса решающего правила:

$$\boldsymbol{w}_{MP}(A_*) = \begin{cases} \frac{\boldsymbol{w}_{ML}^T H \boldsymbol{w}_{ML} - 1}{\boldsymbol{w}_{ML}^T H \boldsymbol{w}_{ML}} \boldsymbol{w}_{ML}, & \boldsymbol{w}_{ML}^T H \boldsymbol{w}_{ML} > 1 \\ \boldsymbol{0}, \text{иначе} \end{cases}$$

При общем решении $\boldsymbol{w}_{MP}$ пропорционален $\boldsymbol{w}_{ML}$, т.е. оптимальные направления регуляризации состоят из $\boldsymbol{w}_{ML}$ и ортогонального дополнения, вдоль которого бесконечно большая регуляризация.