1 Некорректные задачи в обработке изображений

1.1 Введение

Задача нахождения решения z операторного уравнения

$$Az = u$$

по исходным данным u на паре метрических пространств (Z, U) называется корректно поставленной (по Адамару), если выполняются следующие условия:

1. Решение существует: для всякого $u \in U$ существует решение $z \in Z$.

2. Решение единственно.

3. Задача устойчива на пространствах (Z, U).

Некорректно поставленная задача — это задача, для которой хотя бы одно из условий не выполнено.

1.2 Восстановление изображений

В основе класса задач восстановления изображений лежит обратная задача для операторного уравнения

$$Az = u \tag{1}$$

где $u \in U$ — наблюдаемое искажённое изображение из пространства изображений U, $z \in Z$ — искомое изображение из пространства Z, A — оператор искажения. В общем случае и оператор A, и исходное изображение u известны приближённо.

Пространства изображений Z и U — это конечномерные линейные пространства, элементы которых — сеточные функции, определяемые на равномерной сетке с узлами, называемыми пикселями

$$(ih_x, jh_y), \quad 0 \le i \le N, 0 \le j \le M,$$

где $(N+1) \times (M+1)$ — размер изображения, h_x и h_y — шаги сетки.

Значения пикселей могут быть как вещественными, так и комплексными, например, после применения преобразования Фурье. В случае мультиспектральных (например, цветных) изображений значением пикселя будет являться вектор из K цветовых компонент. Также стоит отметить, что в случае цифровых изображений пиксели могут принимать ограниченный набор значений, например, только целые значения из диапазона $[0, 255]$.

В общем случае размеры изображений, значения шагов сетки и типы пикселей в пространствах Z и U являются различными.

Не ограничивая общности, положим, что изображения представлены в градациях серого ($K = 1$), а множеством значений пикселей является пространство вещественных чисел $\mathbb{R}$.

Большинство задач восстановления изображений являются некорректно поставленными. Примерами таких задач являются: повышение резкости, подавление шума, повышение разрешения, многокадровое суперразрешение, заполнение пустот, реставрация изображений (inpainting), построение поля векторов движения между соседними кадрами на видео.

1.3 Математические модели и постановки задач восстановления изображений

Шумоподавление

Рассмотрим шум, характерный для сенсоров камер при получении изображений. Этот шум аддитивный и имеет распределение, близкое к распределению Пуассона. В свою очередь, этот тип шума хорошо аппроксимируется нормальным распределением.

В случае, когда единственным искажением изображения является шум, оператор A в (1) принимает вид

$$Az = u + n,$$

где n — шум.

В дальнейшем будем считать, что n — это гауссовский шум — изображение, каждый пиксель которого является независимой случайной величиной с нормальным распределением с нулевым средним и дисперсией σ_n^2, одинаковой в каждом пикселе.

Устранение размытия

Типичными источниками размытия при съёмке изображений могут являться: нахождение объекта вне плоскости фокусировки, движение объекта и камеры, атмосферные эффекты (дым, туман), несовершенство оптики, антиалиасинговый фильтр.

Как правило, размытие сочетается с шумом, поэтому при решении задачи восстановления размытых изображений одновременно решается и задача шумоподавления.

Оператор A обычно неизвестен, поэтому перед решением обратной задачи (1) требуется провести предварительный анализ по его определению. В общем случае размытие неодинаково в каждом пикселе, но на практике используют упрощённые модели размытия, для которых имеется возможность определения оператора A с достаточной точностью. Выбор конкретной модели размытия определяется задачей и классом используемых изображений.

Наиболее распространённой моделью размытия изображений является свёртка, предполагающая линейность и однородность размытия на изображении:

$$Az = z * H + n, \tag{2}$$
$$(f * g)_{i,j} = \sum_{s,t} f_{s,t} g_{i-s,j-t},$$

где H — ядро размытия, n — шум.

Повышение разрешения

Рассмотрим модель формирования изображения в цифровой камере с шагом сетки h. На матрицу сенсоров камеры проецируется изображение $f(x, y)$, заданное на непрерывной области определения F, далее каждый из сенсоров камеры суммирует интенсивность света в некоторой области, получая значения пикселей.

$$z = D_h H_h f + n,$$

где $f(x, y)$ — изображение, заданное на непрерывной области определения, H_h — оператор рассеяния точки (PSF — point spread function), n — шум, $D_h : F \to Z$ — оператор дискретизации:

$$(D_h f)_{i,j} = f(ih, jh).$$

Далее рассмотрим задачу формирования изображения z_{h_2} для шага сетки h_2 по имеющемуся изображению z_{h_1} с шагом сетки h_1. В соответствии с описанной моделью, данная задача может быть сформулирована в виде:

$$z_{h_2} = D_{h_2} H_{h_2} H_{h_1}^{-1} D_{h_1}^{-1} z_{h_1}. \tag{3}$$

Пиксели камеры измеряют интенсивность света не в точке, а в некоторой окрестности, определяемой размером сенсора и микролинзой, расположенной перед каждым из сенсоров. Также для снижения алиасинга перед матрицей сенсоров может располагаться плёнка, немного размывающая изображение — антиалиасинговый фильтр. Учитывая одинаковость всех сенсоров, оператор H_h может быть представлен в виде

свёртки. Хорошим приближением данного оператора является свёртка с фильтром Гаусса с параметром, зависящим от h:

$$H_h f = f * G_{\sigma_0 h},$$

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right).$$

где σ_0 обычно лежит в диапазоне $[0.3, 0.35]$.

Рассмотрим задачу понижения разрешения в s раз, $s \geq 1$, s — целое. Пусть $h_1 = h$, $h_2 = sh$. Тогда:

$$z_h = D_h(f * G_{\sigma_0 h}),$$
$$z_{sh} = D_{sh}(f * G_{\sigma_0 sh}) = (D_h(f * G_{\sigma_0 sh})) \downarrow_s,$$

где

$$(z \downarrow_s)_{i,j} = z_{si,sj}.$$

Для фильтра Гаусса выполняется соотношение

$$G_{\sigma_1} * G_{\sigma_2} = G_{\sqrt{\sigma_1^2 + \sigma_2^2}},$$

поэтому z_{sh} можно записать в виде

$$z_{sh} = (D_h(f * G_{\sigma_0 h} * G_{\sigma_0 h\sqrt{s^2-1}})) \downarrow_s .$$

Далее воспользуемся приближением

$$D_h(g * G_{\sigma h}) \approx (D_h g) * G_\sigma,$$

применяя которое для $g = f * G_{\sigma_0 h}$ и $\sigma = \sigma_0\sqrt{s^2 - 1}$, получим:

$$z_{sh} \approx (D_h(f * G_{\sigma_0 h}) * G_{\sigma_0\sqrt{s^2-1}}) \downarrow_s = (z_h * G_{\sigma_0\sqrt{s^2-1}}) \downarrow_s .$$

Таким образом, задача понижения разрешения изображений в s раз принимает вид:

$$Az = (z * G_{\sigma_0\sqrt{s^2-1}}) \downarrow_s,$$

а задача повышения разрешения ставится как обратная для данной задачи.

Многокадровое суперразрешение

Использование нескольких изображений низкого разрешения одного и того же статичного объекта, сделанных с небольшими сдвигами, позволяет добиться лучших результатов по сравнению с однокадровыми алгоритмами [1]. Основной источник информации здесь — субпиксельные сдвиги, за счёт которых дискретизация осуществляется в большем числе точек и, как следствие, эффективный шаг дискретизации h становится ниже.

Модель понижения разрешения изображений в s раз для многокадрового суперразрешения для k-го изображения выглядит следующим образом:

$$A_k z = ((F_k z) * G_{\sigma_0\sqrt{s^2-1}}) \downarrow_s, \quad k = 1, \ldots, K,$$

где F_k — оператор движения, в общем случае неизвестный, K — количество изображений.

Обратная задача — задача нахождения изображения высокого разрешения z по исходным изображениям низкого разрешения u_k — представляет собой систему уравнений:

$$A_k z = u_k, \quad k = 1, \ldots, K.$$

Оптический поток

Оптическим потоком называется трансформация изображения, вызванная движением $\vec{v}$, без изменения значений пикселей изображения:

$$f(x + \vec{v}_x(x,y), y + \vec{v}_y(x,y)) = g(x,y). \tag{4}$$

Оптический поток может использоваться в качестве оператора движения F_k для задачи многокадрового суперразрешения изображений.

Задача нахождения оптического потока — это задача нахождения функции $\vec{v}$ по известным изображениям f и g.

Предполагая, что вектора движения в каждом пикселе малы, разложим левую часть (4) в ряд Тейлора

$$\begin{gathered} f(x + \vec{v}_x(x,y), y + \vec{v}_y(x,y)) = \\ = f(x,y) + \vec{v}_x(x,y) f'_x(x,y) + \vec{v}_y(x,y) f'_y(x,y) + o(x,y) \end{gathered}$$

и приходим к системе линейных уравнений:

$$(\nabla f, \vec{v}) = g - f,$$

которая представляет собой задачу (1) в обозначениях $z = \vec{v}$, $u = g - f$, $Az = (\nabla f, z)$.

Заполнение пустот

В ретушировании изображений часто возникает задача восстановления информации в определённом множестве пикселей, информация о значениях которых отсутствует, например, в случае «битых» пикселей или же при оцифровке повреждённой фотоплёнки.

Пусть B — множество повреждённых пикселей. Определим оператор A следующим образом:

$$Az = M_B z, \quad [M_B z]_{i,j} = \begin{cases} z_{i,j}, & (i,j) \notin B, \\ 0, & (i,j) \in B. \end{cases}$$

На пространство изображений U при этом накладывается ограничение: если $u \in U$, тогда $u_{i,j} = 0$ для всех $(i,j) \in B$.

В случае, если на множестве B исходное изображение u имеет ненулевые значения, достаточно применить к нему оператор M_B.

2 Методы решения задач восстановления изображений

2.1 Обращение свёртки

Простейшим способом нахождения решения для задачи устранения размытия (2) при отсутствии шума является прямое обращение свёртки:

$$z = A^{-1}u = u * H^{-1}.$$

Данную операцию можно выполнить, используя теорему о свёртке — одно из свойств преобразования Фурье:

$$\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g),$$

откуда следует

$$\mathcal{F}(H^{-1}) = \frac{1}{\mathcal{F}(H)}$$

и

$$z = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(u)}{\mathcal{F}(H)}\right).$$

Точное восстановление исходного изображения возможно, если $\mathcal{F}(H)$ не обращается в ноль ни в одной точке.

При наличии шума формула обращения свёртки принимает вид

$$z = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(u) - \mathcal{F}(n)}{\mathcal{F}(H)}\right).$$

Размытие изображений, как правило, связано с ослаблением высокочастотной информации, что проявляется в виде малых значений $|\mathcal{F}(H)|$, соответствующих данным частотам. Обращение свёртки приводит к пропорциональному усилению высокочастотной информации.

Гауссовский шум n имеет одинаковую спектральную мощность (модуль преобразования Фурье) на всех частотах, в высокочастотной области его мощность может многократно превышать мощность размытого изображения u. Так как шум неизвестен, его устранение невозможно, и в результате обращения свёртки происходит значительное усиление шума, что исключает возможность практического использования метода обращения свёртки для решения данной задачи.

Одним из способом решения данной проблемы является применение низкочастотной фильтрации для устранения высокочастотного шума, например, фильтра Гаусса.

Если нам известны распределения мощностей шума $P(n)$ и сигнала $P(z)$ в частотной области, то в качестве такого фильтра может быть использован фильтр Винера. Целью винеровской фильтрации для обращения задачи свёртки

$$z = u * H + n$$

является нахождение фильтра G такого, чтобы $\hat{z} = z * G$ минимизировало среднеквадратичное отклонение от оригинального z. Данный фильтр может быть записан в явном виде:

$$\mathcal{F}(G) = \frac{1}{\mathcal{F}(H)} \left[\frac{|\mathcal{F}(H)|^2}{|\mathcal{F}(H)|^2 + \frac{P(n)}{P(z)}} \right].$$

Если известно только соотношение сигнал-шум (SNR), то винеровский фильтр принимает вид

$$\mathcal{F}(G) = \frac{1}{\mathcal{F}(H)} \left[\frac{|\mathcal{F}(H)|^2}{|\mathcal{F}(H)|^2 + \frac{1}{SNR}} \right]. \tag{5}$$

Недостатком фильтрации является невозможность разделения сигнала и шума. Это приводит к невозможности восстановления высокочастотной информации без возникновения артефактов, вызванных влиянием шума.

Также большой проблемой является низкая устойчивость обращения свёртки относительно ядра размытия. Даже в условиях полного отсутствия шума небольшие ошибки в определении ядра размытия могут привести к нежелательным артефактам при обращении свёртки.

2.2 Регуляризация

Одним из способов решения некорректно поставленных задач восстановления изображений является использование регуляризирующих методов, основоположником теории которых является А. Н. Тихонов [2].

Метод регуляризации Тихонова

Рассмотрим некорректно поставленную задачу

$$Az = u, \tag{6}$$

где Z и U — нормированные пространства.

В основе метода регуляризации Тихонова лежит идея нахождения приближённого решения задачи (6) с помощью регуляризирующего оператора

$$z_\alpha = R(u, \alpha),$$

обладающего свойством устойчивости и обеспечивающего стремление z_α к точному решению (6) при $\alpha \to 0$.

Одним из способов построения регуляризирующего оператора является задача минимизации регуляризирующего функционала

$$R(u, \alpha) = \arg\min_z \left(\|Az - u\|_U + \alpha\Omega[z] \right),$$

где оператор $\Omega[z]$ называется *стабилизатором* и удовлетворяет следующему условию: для всякого $d > 0$ множество $\{z \in Z_1 : \Omega[z] \leq d\}$ компактно на Z. Числовой параметр α называется *параметром регуляризации*, а выражение $\|Az - u\|_U$ — невязкой.

Три способа построения регуляризирующего оператора

Рассмотрим три задачи минимизации:

I. $z_\alpha = \arg\min\limits_{z \in Z_1} \left(\|Az - u\|_U + \alpha\Omega[z] \right), \ \alpha > 0$

II. $z_\delta = \arg\min\limits_{z \in Z_\delta(u)} \Omega[z], \quad Z_\delta(u) = \{z \in Z_1 : \|Az - u\|_U \leq \delta\}, \ \delta > 0$

III. $z_\tau = \arg\min\limits_{z \in Z_\tau(u)} \|Az - u\|_U, \quad Z_\tau(u) = \{z \in Z_1 : \Omega[z] \leq \tau\}, \ \tau > 0$

При определённых условиях эти постановки являются эквивалентным, что позволяет использовать постановку, более удобную для использования в конкретной задаче обработки изображений. В частности, эквивалентность достигается при следующих условиях [3]:

1. A — линейный оператор;
2. $\delta \leq \frac{1}{2}\|u\|_U$; $u \neq 0$;
3. $\Omega[0] = 0$;
4. Для любого $z \neq 0$ функция $\gamma(a) = \Omega[az]$ является строго возрастающей функцией аргумента $a \geq 0$.
5. $\Omega[z]$ — неотрицательный выпуклый функционал, т.е. для $0 \leq a \leq 1$ и любых z_1 и z_2 выполняется соотношение

$$\Omega[az_1 + (1-a)z_2] \leq a\Omega[z_1] + (1-a)\Omega[z_2].$$

Применение регуляризирующих методов для восстановления изображений

Поставим задачу нахождения регуляризированного решения уравнения (6) для обратной задачи (1) в следующем виде:

$$z_\alpha = \arg\min_Z \left(\|Az - u\|_U + \alpha\Omega[z]\right), \tag{7}$$

где Z и U — конечномерные нормированные пространства, A — линейный оператор, $\Omega[z]$ — стабилизирующий функционал, областью определения операторов A и $\Omega[z]$ является Z, $\alpha > 0$. Норма пространств Z и U определяется спецификой конкретной задачи восстановления изображений.

В данной постановке рассмотренные в разделе 1.3 задачи восстановления изображений могут быть сформулированы следующим образом:

1. Шумоподавление

$$z_\alpha = \arg\min_Z \left(\|z - u\|_U + \alpha\Omega[z]\right).$$

2. Устранение размытия

$$z_\alpha = \arg\min_Z \left(\|z * H - u\|_U + \alpha\Omega[z]\right).$$

3. Повышение разрешения

$$z_\alpha = \arg\min_Z \left(\|(z * G_{\sigma_0\sqrt{s^2-1}}) \downarrow_s - u\|_U + \alpha\Omega[z]\right).$$

4. Многокадровое суперразрешение

$$z_\alpha = \arg\min_Z \left(\sum_{k=1}^{K} \|((F_k z) * G_{\sigma_0\sqrt{s^2-1}}) \downarrow_s - u_k\|_U + \alpha\Omega[z]\right).$$

5. Заполнение пустот

$$z_\alpha = \arg\min_Z \left(\|M_B(z-u)\|_U + \alpha\Omega[z] \right).$$

6. Оптический поток

$$\vec{v} = \arg\min_{\vec{v}} \left(\|A[\vec{v}] - u\|_U + \alpha\Omega[\vec{v}] \right).$$

Основным сложностями в применении регуляризирующего метода для восстановления изображений являются:

1. Выбор параметра регуляризации α и стабилизатора $\Omega[z]$. Данные параметры должны быть выбраны как с учётом того, что исходное изображение u известно неточно, так и с учётом того, что сам оператор A может быть известен с ошибками.

2. Выбор норм пространств Z и U в соответствии со спецификой конкретной задачи восстановления изображений.

3. Эффективная минимизация регуляризирующего функционала с помощью численных методов.

2.3 Выбор стабилизатора

Стабилизирующий функционал обеспечивает отбор конкретного решения среди множества возможных приближённых решений. Его также можно рассматривать как штрафную функцию, ограничивающую определённые характеристики изображения.

Простейшие стабилизаторы

А. Н. Тихоновым было предложено использовать следующие стабилизаторы:

$$\Omega[z] = \|z\|_2^2,$$
$$\Omega[z] = \|\Delta z\|_2^2,$$

где Δz — оператор Лапласа.

В сочетании с квадратичной нормой невязки получаем задачу минимизации квадратичного функционала

$$z_\alpha = \arg\min_z \left(\|Az-u\|_2^2 + \alpha\|z\|_2^2 \right).$$

Производная данного функционала в точке минимума равна нулю:

$$2A^*(Az - u) + 2\alpha z = 0,$$

откуда получаем возможность построить решение явным образом:

$$z_\alpha = (A^*A + \alpha I)^{-1}A^*u.$$

Если A — это оператор свёртки $Az = z * H$, то регуляризирующий оператор может быть представлен в виде оператора свёртки G

$$z_\alpha = u * G,$$

где оператор G может быть найден, используя теорему о свёртке:

$$\mathcal{F}(G) = \frac{\mathcal{F}(H^*)}{\mathcal{F}(H^*)\mathcal{F}(H) + \alpha}.$$

Примечание: сопряжённый оператор для операции свёртки представляет собой свёртку с ядром, повёрнутым на 180°.

Изображения, восстанавливаемые данным регуляризирующим оператором, похожи на результаты винеровской фильтрации (5), а в случае, когда H является фильтром Гаусса, полностью совпадают при $\alpha = \frac{1}{SNR}$, так как для фильтра Гаусса $|\mathcal{F}(H)| = \mathcal{F}(H) = \mathcal{F}(H^*)$.

Аналогичным образом строится решение для стабилизатора $\|\Delta z\|_2^2$:

$$z_\alpha = (A^*A + \alpha\Delta^*\Delta)^{-1}A^*u.$$

Общим недостатком данных стабилизаторов является их слабая связь с характеристиками изображений, что приводит к невозможности восстановления высокочастотной информации на изображениях без внесения артефактов.

Полная вариация

Более эффективным в задачах обработки изображений является использование функционала полной вариации:

$$TV[f] = \int |\nabla f| d\mathbf{x}.$$

Полезность полной вариации в задачах обработки изображений заключается в её тесной связи с контурами на изображениях:

1. Значение полной вариации не зависит от наклона функции, а только от величины перепада интенсивности. Таким образом, ограничение полной вариации не будет приводить к нежелательному размытию контуров, в отличие от классических стабилизаторов.

2. Имеется связь с осцилляциями Гибсса, представляющим собой колебания возле резких перепадов интенсивности при низкочастотной фильтрации. Возникновение осцилляций Гиббса приводит к значительному росту полной вариации. Таким образом, ограничение полной вариации позволяет бороться с этим эффектом.

3. Ещё одна связь между полной вариацией и структурой изображения формулируется следующей теоремой:

Теорема 1. Если $TV[f] < \infty$, тогда полная вариация равна сумме длин линий уровня:

$$TV[f] = \int_{-\infty}^{+\infty} H^1(\partial\Omega_y)dy,$$

где $\Omega_y = \{(x_1, x_2) \in \mathbb{R}^2 : f(x_1, x_2) > y\}$, $\partial\Omega_y$ — граница множества (линии уровня), $H^1(\partial\Omega_y)$ — длина $\partial\Omega_y$. Формально H^1 является одномерной мерой Хаусдорфа.

Эта теорема может быть использована для выбора параметра регуляризации (подробнее об этом написано в разделе 2.5).

Полная обобщённая вариация

Недостатком использования полной вариации в качестве стабилизатора в регуляризирующих методах восстановления изображений является ступенчатость получаемого изображения. Особенно сильно этот эффект наблюдается при шумоподавлении.

Полная обобщённая вариация накладывает дополнительные ограничения на производные изображения высших порядков [4]. В обработке изображений обычно используется полная обобщённая вариация второго порядка, которая для дифференцируемых функций записывается в виде:

$$TGV^2[f] = \min_g \left(\alpha_1 \int |\nabla f - g| d\mathbf{x} + \alpha_0 \int |\nabla g| d\mathbf{x} \right), \tag{8}$$

где минимум вычисляется среди всевозможных элементов векторного поля g.

Стоит обратить внимание, что вместо одного параметра регуляризации α здесь используется пара значений (α_1, α_0).

Эффект от применения полной обобщённой вариации второго порядка для восстановления изображений состоит в том, что изображе-

ния стремятся к кусочно-линейному виду, а не к кусочно-постоянному, как в случае обычной полной вариации.

Использование TGV^2 в форме (8) на практике может быть затруднительным с вычислительной точки зрения. Вместо этого возможно использование упрощённых моделей, например, взвешенной суммы модулей первых и вторых производных [5]:

$$TGV^2[f] = \alpha_1 \int |\nabla f| d\mathbf{x} + \alpha_0 \int |\nabla^2 f| d\mathbf{x}, \tag{9}$$

либо использование модуля оператора Лапласа [6]:

$$TGV^2[f] = \alpha_1 \int |\nabla f| d\mathbf{x} + \alpha_0 \int |\Delta f| d\mathbf{x}.$$

2.4 Выбор норм пространств

Выбор норм пространств Z и U в задаче (7) обусловлен спецификой конкретной задачи восстановления изображений. Наиболее распространёнными являются следующие варианты:

$$\|u\|_1 = \sum_{i,j} |u_{i,j}|, \tag{10}$$

$$\|u\|_2^2 = \sum_{i,j} u_{i,j}^2. \tag{11}$$

Использование квадратичной нормы (11) обычно удобнее с вычислительной точки зрения, тогда как норма (10) имеет определённые преимущества для ряда задач обработки изображений.

Рассмотрим различия между этими нормами на примере задачи шумоподавления изображений с использованием полной вариации в качестве стабилизатора:

$$\text{TV-L1}: \quad z_\alpha = \arg\min_z \left(\|z - u\|_1 + \alpha TV[z] \right).$$

$$\text{TV-L2}: \quad z_\alpha = \arg\min_z \left(\|z - u\|_2^2 + \alpha TV[z] \right).$$

В случае TV-L1 при увеличении параметра регуляризации исчезают, в первую очередь, мелкие детали, тогда как в случае TV-L2 первыми пропадают слабоконтрастные детали. Этот факт можно использовать для выбора соответствующей модели в зависимости от типа шума на изображении: модель TV-L1 хорошо подходит для устранения импульсного шума, тогда как TV-L2 лучше справляется с гауссовским шумом.

Также стоит отметить, что в модели TV-L1 результат инвариантен относительно изменения контраста. Если z_α минимизирует TV-L1 для исходного изображения u, тогда $kz_\alpha + c$ будет минимизировать TV-L1 для исходного изображения $ku + c$.

2.5 Выбор параметра регуляризации

Выбор параметра регуляризации представляет собой отдельную проблему. Использование слишком малого значения параметра регуляризации приводит к возникновению артефактов, например, усилению шума, тогда как при чрезмерно большом значении параметра изображение становится сглаженным, а детали теряются.

Существует несколько способов выбора параметра регуляризации в зависимости от конкретной задачи восстановления изображений.

Априорный выбор при известной погрешности

Если известна либо имеется возможность оценить погрешность δ, с которой дано u, тогда можно использовать регуляризирующий оператор в постановке (II):

$$z_\delta = \arg\min_{z \in Z_\delta} \Omega[z], \quad Z_\delta = \{z : \|Az - u\|_U \leq \delta\}, \quad \delta > 0.$$

Этот способ хорошо подходит для задачи шумоподавления изображения: имеется прямая зависимость между среднеквадратичным отклонением и погрешностью δ.

Стоит отметить, что прямой связи между погрешностью оператора A и значением δ нет, поэтому для задачи восстановления размытых изображений с неточно заданным оператором A данный способ выбора параметра регуляризации может привести к существенным ошибкам.

Априорный выбор при известном стабилизаторе

Если можно оценить, какое значение должен иметь конкретный стабилизатор $\Omega[z]$ для восстановленного изображения, удобно использовать регуляризирующий оператор в постановке (III):

$$z_\tau = \arg\min_{z \in Z_\tau} \|Az - u\|_U, \quad Z_\tau = \{z : \Omega[z] \leq \tau\}, \quad \tau > 0.$$

Подобная оценка возможна в задаче повышения разрешения изображений для стабилизатора полной вариации $\Omega[z] = TV[z]$. Повышение

разрешения в s раз не должно приводить к возникновению новых деталей или исчезновению существующих, а все структуры должны быть масштабированы в s раз. Используя теорему о связи между полной вариацией и суммой длин линий уровня, приходим к выводу, что полная вариация тоже должна возрасти в s раз. Таким образом, для задачи повышения разрешения в s раз хорошей оценкой параметра является $\tau = sTV[u]$.

Априорный выбор с использованием баз изображений

При использовании постановки (I)

$$z_\alpha = \arg\min_z \left(\|Az - u\|_U + \alpha\Omega[z]\right)$$

в ряде задач обработки изображений наблюдается закономерность, показывающая близость значений оптимальных параметров регуляризации для изображений одних и тех же классов при схожих параметрах искажения, например, при одинаковом уровне шума.

Для данных задач оптимальные параметры можно найти путём составления пар изображений — эталон и результат моделирования искажений с заданными параметрами — с последующим применением регуляризирующих методов с оптимизацией метрики, показывающей близость результата к эталону, например, метрики PSNR.

В качестве эталонных изображений берутся изображения из класса обрабатываемых изображений, например, для фотографических изображений можно использовать стандартные изображения из общеупотребительных баз, таких как TID2013 [7]. Использование стандартных изображений позволяет более эффективно сравнивать различные алгоритмы обработки изображений между собой.

Этот подход применяется в задаче повышения резкости дефокусированных изображений, где возможно построить зависимость между радиусом дефокусировки, уровнем шума и параметром регуляризации.

Апостериорный выбор

В случае невозможности априорного выбора параметра регуляризации возможен его апостериорный подбор. Суть его заключается в использовании функции, оценивающей качество восстановленного изображения. Регуляризирующий метод при этом применяется итерационно с различными параметрами регуляризации и останавливается при достижении определённых условий.

3 Численные методы минимизации регуляризирующего функционала

В данном разделе будет рассмотрена задача минимизации регуляризирующего функционала в постановке (I):

$$f(z) = \arg\min_z \left(\|Az - u\|_U + \alpha\Omega[z]\right) \to \min.$$

Классическим методом минимизации функционалов является градиентный спуск — метод нахождения локального минимума или максимума функции с помощью движения вдоль градиента. Если функционал является строго выпуклым, то локальный минимум будет единственным и совпадающим с глобальным минимумом.

3.1 Дифференцирование функционалов

Вычисление производной для функционала полной вариации $TV[z]$ и производной невязки с нормой $\|\cdot\|_1$ не представляется возможным, так как данные функционалы не являются дифференцируемыми. В самом деле, функция $f(x) = |x|$ не имеет производной при $x = 0$. Одним из способов решения данной проблемы является приближение её гладкой функцией:

$$f_\varepsilon(x) = \sqrt{x^2 + \varepsilon^2}, \quad f'_\varepsilon(x) = \frac{x}{\sqrt{x^2 + \varepsilon^2}}, \quad \varepsilon > 0.$$

Недостатком данного подхода являются высокие вычислительные затраты для операций деления и вычисления квадратного корня. Также возникает сложность выбора параметра ε. Вместо этого используются понятия субградиента и субдифференциала.

Определение 1. Элемент g называется субградиентом выпуклого функционала $f(z) : Z \to \mathbb{R}$ в точке z_0, если для всех $z \in Z$ выполняется неравенство

$$f(z) - f(z_0) \geq g \cdot (z - z_0).$$

Определение 2. Множество всех субградиентов выпуклого функционала $f(z) : Z \to \mathbb{R}$ в точке z_0 называется субдифференциалом $\partial f(z_0)$.

Свойства субдифференциала выпуклого функционала $f(z)$:

1. $\partial f(z)$ — выпуклое множество, возможное пустое;
2. Линейность:

$$\partial(f_1(z) + f_2(z)) = \partial f_1(z) + \partial f_2(z)$$
$$\partial(\lambda f(z)) = \lambda \partial f(z), \quad \lambda > 0.$$

Здесь сумма понимается в смысле суммы Минковского:

$$\partial f_1 + \partial f_2 = \{g \mid g = g_1 + g_2,\, g_1 \in \partial f_1,\, g_2 \in \partial f_2\}.$$

3. Если $f(z)$ непрерывен в точке z_0, тогда $f(z)$ имеет в z_0 непустой субдифференциал, причём $\partial f(z_0)$ — компактное и выпуклое множество.

4. Функционал $f(z)$ дифференцируем в точке z_0 тогда и только тогда, когда его субдифференциал в этой точке состоит из единственного элемента и равен градиенту $\partial f(z_0) = \{\nabla f(z_0)\}$.

5. Функционал имеет локальный минимум в точке тогда и только тогда, когда ноль принадлежит субдифференциалу в этой точке.

На рис. 1 приведён пример, демонстрирующий определение субградиента. Выпуклая функция $f(z)$ имеет единственный субградиент g_1 в x_1, совпадающий с её производной. В точке x_2 функция недифференцируема, субдифференциал представляет собой отрезок $[g_2, g_3]$.

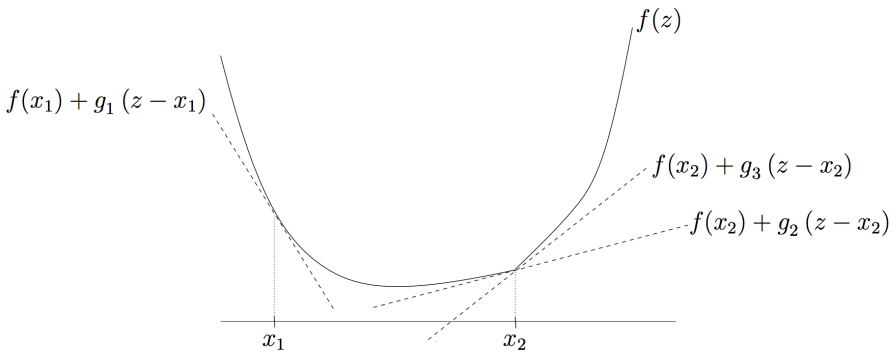

Рис. 1: Иллюстрация к определению субградиента

Функция $f(x) = |x|$ не является дифференцируемой при $x = 0$, но имеет в этой точке субдифференциал:

$$\partial|x| = \begin{cases} \{-1\}, & x < 0, \\ [-1, 1], & x = 0, \\ \{1\}, & x > 0. \end{cases} \tag{12}$$

Примеры

Приведём примеры вычисления производной и субдифференциала для операторов, используемых в регуляризирующем функционале в задачах восстановления изображений.

Сначала рассмотрим дифференцируемые операторы (A — линейный оператор):

$$\begin{aligned}
&\left(\|z\|_2^2\right)' = 2z,\\
&(Az)' = A^T,\\
&\left(\|Az-u\|_2^2\right)' = 2(Az-u)\circ A^T = 2A^TAz - A^Tu.
\end{aligned}$$

Если A — оператор свёртки $Az = z * H$, тогда $A^Tz = z * H^*$, где $H^*(x,y) = H(-x,-y)$. Стоит обратить внимание, что здесь происходит поворот ядра свёртки H на 180°. Частой ошибкой на практике является транспонирование ядра вместо поворота.

Если A — оператор понижения разрешения $Az = z\downarrow_s$, тогда

$$(A^Tz)_{i,j} = \begin{cases} z_{i/s,j/s}, & i,j \in \{0,s,2s,\ldots\},\\ 0, & \text{иначе.}\end{cases}$$

Если прямой оператор $\downarrow_s$ представляет собой прореживание пикселей изображения путём взятия каждого s-го пикселя в строке и столбце, то сопряжённый к нему оператор, наоборот, осуществляет разрежение изображения в s раз с дополнением нулями.

Далее приведём примеры вычисления субдифференциала для недифференцируемых операторов:

$$\begin{aligned}
&\partial\|z\|_1 = \{(g_1,\ldots,g_n)^T\}, \quad g_i \in \partial|z_i|,\\
&\partial\left(\|Az-u\|_1\right) = A_T\partial\|Az-u\|_1,\\
&\partial TV[z] = \partial\left(\|\nabla z\|_1\right) = \nabla^T\left(\partial\|\nabla z\|_1\right).
\end{aligned}$$

3.2 Дискретное представление функционала полной вариации

Для двумерных изображений полная вариация принимает вид:

$$TV[z] = \sum_{i,j=1}^{N,M} |\nabla z_{i,j}| = \sum_{i,j=1}^{N,M} \sqrt{[z'_x]^2_{i,j} + [z'_y]^2_{i,j}}.$$

Здесь возникает необходимость вычисления производной изображения разностными методами. Наиболее распространёнными являются следующие методы:

1. Обычная производная

$$\nabla z_{i,j} = \left(z_{i+1,j} - z_{i,j}, z_{i,j+1} - z_{i,j}\right).$$

2. Симметричная производная

$$\nabla z_{i,j} = (z_{i+1,j} - z_{i-1,j}, z_{i,j+1} - z_{i,j-1}).$$

Оба этих способа имеют проблемы. Обычная производная из-за несимметричности приводит к смазу изображения, в общем случае малозаметному, но при определённых условиях этот эффект может стать раздражающим. Симметричная производная лишена этого недостатка, но имеет другую проблему: центральный пиксель $z_{i,j}$ выпадает из вычисления вариации, что приводит к тому, что пиксели с чётной суммой координат $i + j$ не оказывают никакого влияния на пиксели с нечётной суммой. Это приводит к появлению на изображении регулярного шума с шахматным паттерном.

Преодолеть эти проблемы возможно, перейдя от вычисления градиента изображения к вычислению суммы перепадов интенсивности. Простейший способ — это сумма перепадов по горизонтали и вертикали:

$$TV[z] = \sum_{i,j} |z_{i+1,j} - z_{i,j}| + \sum_{i,j} |z_{i,j+1} - z_{i,j}|. \tag{13}$$

Данный способ имеет высокую вычислительную эффективность, но одновременно с этим и незначительный недостаток в виде анизотропии — неоднородности обработки изображения по различным направлениям.

Исправить этот недостаток можно с помощью билатеральной полной вариации — функционала, вычисляющего сумму перепадов по множеству направлений:

$$BTV[z] = \sum_{(x,y)\in Q} \frac{1}{\sqrt{x^2+y^2}} \sum_{i,j} |z_{i+x,j+y} - z_{i,j}|,$$

где Q — множество смещений, например:

$$Q_{TV} = \{(1,0),(0,1)\},$$
$$Q_1 = \{(1,0),(0,1),(1,1),(1,-1)\},$$
$$Q_2 = \{(x,y),\ x,y = 0,\pm 1,\pm 2,\ (x,y) \neq (0,0)\}.$$

При использовании $Q = Q_{TV}$ функционал $BTV[z]$ совпадает с функционалом полной вариации (13).

Чтобы вычислить производную от $BTV[z]$, запишем его сначала в операторном виде:

$$BTV[z] = \sum_{(x,y)\in Q} \frac{1}{\sqrt{x^2+y^2}} \|S_X^x S_Y^y z - z\|_1,$$

где S_X и S_Y — операторы сдвига на 1 пиксель по горизонтали и вертикали соответственно:

$$(S_X z)_{i,j} = z_{i+1,j},$$
$$(S_Y z)_{i,j} = z_{i,j+1}.$$

Сопряженные операторы для S_X и S_Y представляют собой сдвиг на 1 пиксель по горизонтали и вертикали соответственно в противоположном направлении: $S_X^T = S_X^{-1}$, $S_Y^T = S_Y^{-1}$.

Далее, применим правила вычисления производной составных функционалов и найдём субдифференциал от $BTV[z]$:

$$\partial BTV[z] = \sum_{(x,y)\in Q} \frac{1}{\sqrt{x^2+y^2}} (S_X^{-x} S_Y^{-y} - I)\partial \|S_X^x S_Y^y z - z\|_1.$$

Субдифференциал $\partial BTV[z]$ состоит из множества субградиентов. На практике для разрешения неоднозначности используют следующий субградиент:

$$BTV[z]' = \sum_{(x,y)\in Q} \frac{1}{\sqrt{x^2+y^2}} (S_X^{-x} S_Y^{-y} - I)\mathrm{sgn}(S_X^x S_Y^y z - z),$$

где

$$(\mathrm{sgn} z)_{i,j} = \begin{cases} -1, & z_{i,j} < 0, \\ 0, & z_{i,j} = 0, \\ 1, & z_{i,j} > 0. \end{cases}$$

Аналогичная разностная схема строится для функционала полной обобщенной вариации второго порядка (9):

$$BTV_2[z] = \alpha_1 BTV[z] + \alpha_2 \sum_{(x,y)\in Q} \frac{1}{\sqrt{x^2+y^2}} \|S_X^x S_Y^y z + S_X^{-x} S_Y^{-y} z - 2z\|_1.$$

3.3 Субградиентные методы

Нахождение минимума выпуклого субдифференцируемого функционала $f(x) : \mathbb{R}^n \to \mathbb{R}$ может быть осуществлено с помощью итерационного процесса — субградиентного метода:

$$x^{(k+1)} = x^{(k)} - \beta_k g^{(k)}. \tag{14}$$

Здесь $x^{(k)}$ — значение k-й итерации, $g^{(k)} \in \partial f(x^{(k)})$ — любой субградиент $f(x)$ в точке $x^{(k)}$, $\beta_k > 0$ — шаг на k-й итерации. Если $f(x)$

— дифференцируемая функция, тогда единственным возможным вариантом выбора для $g^{(k)}$ становится градиент функции $\nabla f(x^{(k)})$, и метод (14) сводится к классического методу градиентного спуска.

В отличие от метода градиентного спуска, где $-\nabla f(x^{(k)})$ является направлением спуска, это в общем случае не выполняется для субградиентного метода: возможен рост значения $f(x)$ при движении вдоль направления, противоположному субградиенту $f'(x)$. Другими словами, субградиентный метод не гарантирует монотонного уменьшения $f(x^{(k)})$. С учётом этого факта, обычно отслеживают наилучшую точку среди всех предыдущих итераций:

$$f_{best}^{(k)} = \min\left(f_{best}^{(k-1)}, f(x^{(k)})\right).$$

Так как последовательность $f_{best}^{(k)}$ является монотонно убывающей, а $f(x)$ ограничена снизу, то эта последовательность имеет предел. Представляют интерес способы выбора последовательности шагов, при которых этот предел совпадает с минимумом $f(x)$.

Будем рассматривать сходимость субградиентного метода, исходя из предположения, что субградиент $f(x)$ ограничен, т.е. существует G такое, что $\|g^{(k)}\|_2 \leq G$ для всех k. Это будет выполняться, например, когда $f(x)$ удовлетворяет условию Липшица:

$$|f(x_1) - f(x_2)| \leq G\|x_1 - x_2\|_2 \tag{15}$$

для любых x_1 и x_2.

Также предположим, что известно значение R — точность, с которой дано начальное приближение к точному решению x_T:

$$\|x^{(1)} - x_T\|_2 \leq R.$$

В [8] показано, что при данных условиях (15) справедливо неравенство

$$f_{best}^{(k)} - f(x_T) \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \beta_i^2}{2\sum_{i=1}^{k} \beta_i}.$$

Способы выбора шага

В субградиентных методах способы выбора шага отличаются от общепринятых в классических методах градиентного спуска. Основными способами выбора шага являются следующие:

1. Постоянный коэффициент β:

$$\beta_k = \beta, \quad \beta > 0.$$

2. Постоянная норма шага γ:

$$\beta_k = \frac{\gamma}{\|g^{(k)}\|_2}, \quad \gamma > 0.$$

Это означает, что $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$.

3. Несуммируемая последовательность коэффициентов, суммируемая в квадратах:

$$\beta_k > 0, \quad \sum_{k=1}^{\infty} \beta_k = \infty, \quad \sum_{k=1}^{\infty} \beta_k^2 < \infty.$$

Пример такой последовательности:

$$\beta_k = \frac{a}{b+k}, \quad a > 0, \quad b \geq 0.$$

4. Несуммируемая бесконечно убывающая последовательность коэффициентов:

$$\beta_k > 0, \quad \sum_{k=1}^{\infty} \beta_k = \infty, \quad \lim_{k\to\infty} \beta_k = 0.$$

Пример:

$$\beta_k = \frac{a}{\sqrt{k}}, \quad a > 0.$$

5. Несуммируемая бесконечно убывающая последовательность шагов:

$$\beta_k = \frac{\gamma_k}{\|g^{(k)}\|_2}, \quad \gamma_k > 0, \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \quad \lim_{k\to\infty} \gamma_k = 0.$$

Интересной особенностью вышеперечисленных способов является то, что последовательность шагов фиксируется до начала итерационного процесса, она не зависит от получаемых в процессе итераций промежуточных значений.

Оптимальная последовательность шагов

С практической точки зрения важно получить наилучший результат за наименьшее число итерации. Рассмотрим задачу выбора таких коэффициентов $\beta_1, \ldots, \beta_k$, при которых за k итераций субградиентного метода можно получить результат, наиболее близкий к точному, т.е.

$$\frac{R^2 + G^2 \sum_{i=1}^{k} \beta_i^2}{2 \sum_{i=1}^{k} \beta_i} \to \min.$$

Этот функционал выпуклый и симметричный относительно $\beta_1, \ldots, \beta_k$. Соответственно, оптимум достигается, когда все β_i равны одному значению β. Получаем задачу минимизации

$$\frac{R^2 + G^2 k \beta^2}{2k\beta} \to \min,$$

для которой минимум достигается при

$$\beta = (R/G)/\sqrt{k}.$$

При этом

$$f_{best}^{(k)} - f(x_T) \leq RG/\sqrt{k}. \tag{16}$$

Для случая функций, не удовлетворяющих условию Липшица (15), также есть вариант выбора параметров β_i [8], однако, скорость сходимости при его использовании настолько мала, что практической ценности он не имеет.

Практический выбор шага

Использование оптимальной последовательности шагов позволяет достичь гарантированной точности за определённое число итераций. Однако, верхняя оценка скорости сходимости (16) показывает, что с ростом желаемой точности количество итераций растёт квадратично.

Для реальных задач обработки изображений наблюдается гораздо более быстрая скорость сходимости субградиентного метода. В частности, может применяться следующий способ выбора последовательности шагов в виде геометрической прогрессии:

$$\beta_i = \beta_1 \left(\frac{\beta_k}{\beta_1} \right)^{\frac{i-1}{k-1}}$$

с тремя параметрами: β_1 — шаг на первой итерации, β_k — шаг на последней итерации, k — количество итераций. Параметры выбираются таким образом, чтобы на первой итерации изображение менялось довольно сильно, а к последней итерации изменение оказывалось нихе градации пикселя.

Ускорение субградиентого метода

Классический субградиентный метод (14) сходится достаточно медленно. Для решения этой проблемы используется ряд методов, получивших второе рождение в задачах оптимизации коэффициентов свёрточных нейронных сетей.

Метод моментов Идея метода моментов заключается в продолжении движения по направлению, полученному на предыдущей итерации, с поправкой на текущее значение субградиента:

$$v^{(k+1)} = \mu v^{(k)} - g^{(k)}, \quad v^{(0)} = 0, \quad 0 \leq \mu < 1,$$
$$x^{(k+1)} = x^{(k)} + \beta_{k+1} v^{(k+1)}.$$

Альтернативная формулировка:

$$v^{(k+1)} = \mu v^{(k)} - \beta_{k+1} g^{(k)}, \quad v^{(0)} = 0, \quad 0 \leq \mu < 1,$$
$$x^{(k+1)} = x^{(k)} + v^{(k+1)}.$$

При $\mu = 0$ метод становится обычным субградиентным методом.

Метод ускоренного градиента Нестерова В отличие от метода моментов, вместо того чтобы высчитывать градиент в текущей точке, в методе ускоренного градиента Нестерова используется градиент в точке, "предсказанной" на основании сдвига, расчитанного на предыдущем шаге:

$$g^{(k)} \in \partial f(x^{(k)} + \beta_k \mu v^{(k)}),$$
$$v^{(k+1)} = \mu v^{(k)} - g^{(k)},$$
$$x^{(k+1)} = x^{(k)} + \beta_{k+1} v^{(k+1)}.$$

Альтернативная формулировка:

$$g^{(k)} \in \partial f(x^{(k)} + \mu v^{(k)}),$$
$$v^{(k+1)} = \mu v^{(k)} - \beta_k g^{(k)},$$
$$x^{(k+1)} = x^{(k)} + v^{(k+1)}.$$

Практическое применение метод ускоренного градиента Нестерова для задачи обращения свёртки показывает, что для достижения результата достаточно качества достаточно использование 40 итераций при $0,8 \leq \mu \leq 0,9$ и выборе нормы шага $\gamma_0 = 25$, $\gamma_k = 0,25$.

Литература

[1] *Крылов А. С., Насонов А. В.* Регуляризирующие методы интерполяции изображений. — АРГАМАК-МЕДИА Москва, 2014. — 100 с.

[2] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1979. — С. 288.

[3] *Васин В. В.* О связи некоторых вариационных методов приближенного решения некорректных задач // *Матем. заметки.* — 1970. — Т. 7, № 3. — С. 265–272.

[4] *Bredies K., Kunisch K., Pock T.* Total generalized variation // *SIAM Journal on Imaging Sciences.* — 2010. — Vol. 3, no. 3. — Pp. 492–526.

[5] *Papafitsoros K., Schönlieb C.-B.* A combined first and second order variational approach for image reconstruction // *Journal of mathematical imaging and vision.* — 2014. — Vol. 48, no. 2. — Pp. 308–338.

[6] *Chan T. F., Esedoglu S., Park F.* A fourth order dual method for staircase reduction in texture extraction and image restoration problems // 2010 IEEE International Conference on Image Processing / IEEE. — 2010. — Pp. 4137–4140.

[7] Image database TID2013: Peculiarities, results and perspectives / N. Ponomarenko, L. Jin, O. Ieremeiev et al. // *Signal processing: Image communication.* — 2015. — Vol. 30. — Pp. 57–77.

[8] *Boyd S., Mutapcic A.* Subgradient Methods. Notes for EE364b // *Stanford University.* — Winter 2006-07.